PSL
UNIVERSITÉ PARIS

THÈSE DE DOCTORAT
DE L'UNIVERSITÉ PSL

Préparée à l'École Normale Supérieure

Entraînement de réseaux de neurones profonds : modèles mathématiques et résultats de convergence

Soutenue par
Raphaël Barboni
Le 8 Octobre 2025

École doctorale n°386
École doctorale de Sciences Mathématiques de Paris Centre

Spécialité
Mathématiques

Préparée au
Département de Mathématiques et Applications (DMA)

Composition du jury :

Claire BOYER Professeure, Université Paris-Saclay	*Présidente du jury*
Joan BRUNA Professeur, New York University	*Rapporteur*
Mikaela IACOBELLI Professeure, ETH Zürich	*Rapportrice*
Eric VANDEN-EIJNDEN Professeur, New York University	*Examinateur*
Borjan GESHKOVSKI Chargé de recherche, INRIA Paris	*Examinateur*
Gabriel PEYRÉ Directeur de recherche, CNRS	*Directeur de thèse*
François-Xavier VIALARD Professeur, Université Gustave Eiffel	*Directeur de thèse*

1794
ENS
ÉCOLE NORMALE SUPÉRIEURE

PSL
RESEARCH UNIVERSITY PARIS

Remerciements

La vie de thésard en mathématiques ressemble par bien des aspects à une course d'obstacles. Je me dois donc de commencer par remercier Gabriel et François-Xavier, qui m'ont fait prendre le départ, il y a de cela plus de 4 ans, et qui m'ont accompagné jusqu'à la ligne d'arrivée aujourd'hui. Je les remercie pour leur écoute, toujours attentive, et leurs conseils, toujours avisés. Si je m'en vais désormais vers d'autres horizons scientifiques et géographiques, j'espère que nos chemins se recroiseront aussi souvent que possible.

Je voudrais ensuite remercier tous ceux dont la présence quotidienne a égayé mes journées de travail. Merci Sibylle, d'avoir été une partenaire de bureau toujours présente dans les bons comme dans les mauvais moments. Un de mes regrets restera de partir avant de t'avoir vu affronter l'ensemble du laboratoire au bras de fer. Merci Jules pour toutes les distractions qui m'ont permis d'allonger toujours un peu plus mes pauses midi. J'espère pouvoir continuer à m'entraîner pour avoir un jour le même niveau que toi au pedantle. Merci Othmane, d'avoir été mon grand-frère de thèse. Merci Pierre pour les parties de GeoGuessr. Merci Zaccharie pour les pauses goûter à 14h chez Bruno Solques et merci à Bruno Solques pour les croissants et les pains au chocolat. Merci Étienne pour les *random lunch*. S'il est difficile de n'oublier personne, je pense évidemment à l'ensemble de mes collègues du CSD, à Geert, Francisco, Michaël, Valérie, Anna, Romain, Gauthier, Samuel, Jérémie, Thibault, et bien d'autres encore... Plus généralement c'est l'ensemble du DMA qui m'a permis d'étudier puis de travailler pendant plus de sept ans dans un cadre toujours stimulant et convivial. J'ai une pensée pour Cyril qui a su me tutorer avec attention et bienveillance pendant ma scolarité. J'ai enfin une pensée spéciale pour Fabienne, sans qui tant de choses ne seraient pas possibles. Grâce à elle j'ai pu aller aux quatre coins du monde : à Marseille, Berlin, Grenade et même à la Nouvelle-Orléans. C'est aussi grâce à son aide que l'organisation de la soutenance a été possible.

Ce manuscrit est également dédié aux amis qui me suivent depuis de nombreuses années. Si la vie prend des chemins parfois sinueux, je sais que je peux compter sur eux pour être là derrière chaque virage. Merci Adrien, Théo, Lancelot et Fehri d'avoir été toujours présents, et ce depuis le collège. Avec vous j'ai partagé un nombre incalculable de soirées à regarder des films en mangeant des pizzas, de discussions jusqu'à pas d'heure, de débats sans fin et encore plein d'autres souvenirs. Merci Victor pour m'avoir accompagné dans tant d'aventures, y compris celles où je t'ai embarqué contre ton gré. Après avoir partagé l'effort, la promiscuité et deux années de colocation, toi seul sais désormais que, en caleçon et à 23h30, c'est bien moi le plus grand râpeur. Merci Thomas d'être autant force de propositions et d'avoir organisé tous ces concerts, qui ont été autant de découvertes musicales, et toutes ces soirées, qui ont animé mes week-ends. Merci aussi pour les voyages, un des moments les plus forts de ma vie restant l'ouverture d'une noix de coco sur une plage de Martinique. Merci Lucas pour les discussions du midi au DMA, et avant cela pour les sorties escalade. C'est avec toi que j'ai partagé certains de mes meilleurs

souvenirs de montagne : au Chili ou bien à Ailefroide. Enfin Cyril, merci de m'avoir accompagné pendant toutes ces années, à l'entraînement comme en compétition. Certains de ces moments ont été des moments de souffrance, comme les côtes du parc de Sceaux peuvent en attester, d'autres des moments de joie et d'autres simplement d'absurdité, comme les séances de perche dans la neige ou bien les entraînements le 31 décembre. Si notre rivalité fut digne des annales de l'athlétisme, ce que je retiens, ce sont avant tout ces souvenirs.

Puisque le doctorat marque la fin de ma vie d'étudiant, je veux aussi rendre hommage à ces personnes qui, à un moment, ont été mes professeurs et resteront pour moi bien plus encore. Merci Carlos, de m'avoir appris à toujours faire les choses avec passion. Merci Thomas, de m'avoir appris à franchir les obstacles pour mieux repartir ensuite. Enfin, merci à Romain Vareille et Sébastien Kerner pour m'avoir lancé dans la voie des mathématiques en suscitant toujours ma curiosité.

Pour finir, si bien des choses ont changé tout au long de ces années de thèse, il y a aussi des personnes qui, fort heureusement, savent rester toujours fidèles à elles-mêmes. Mes derniers remerciements vont donc à ma famille. Je pense à mes oncles, Paul et Frédéric. Je pense à mon frère, Alexandre, qui depuis toujours a été une source intarissable d'admiration et d'inspiration. Si l'expérience m'a prouvé qu'il ne faut pas toujours le croire sur parole (surtout quand, en montagne, il dit que "T'inquiète on peut prendre ce chemin, c'est sûr que ça passe"), il reste bien souvent une de mes boussoles. Je pense à mes parents, pour leur soutien que je sais être indéfectible. Par bien des aspects, ce manuscrit est aussi le fruit de leur travail et de leur affection. Enfin, mes dernières pensées vont à Mamie et Manou. Quoi que je fasse et où que j'aille à l'avenir, elles seront toujours là.

Training deep neural networks: mathematical models and convergence results

Raphaël Barboni

8 Octobre 2025

Résumé

Les progrès récents des modèles d'apprentissage profond dans de nombreuses applications ont mis en lumière la nécessité d'une meilleure compréhension de leurs dynamiques d'entraînement. Dans cette thèse, nous contribuons à l'étude théorique des algorithmes de descente de gradient pour l'entraînement de réseaux de neurones surparamétrés. Des travaux récents ont en effet montré que, pour des architectures peu profondes, il est possible d'obtenir de bonnes garanties de convergence en relaxant le problème d'optimisation dans l'espace des distributions de paramètres.

Nous prolongeons cette approche au cas des architectures profondes en étudiant des limites de champ-moyen de *réseaux de neurones résiduels (ResNets)*. Ces modèles sont paramétrés par des distributions sur le produit de l'espace des couches et d'un espace de paramètres, avec la contrainte d'une marginale uniforme sur l'espace des couches. Dans ce cadre, nous proposons de modéliser l'apprentissage comme un flot de gradient pour une distance de *Transport Optimal Conditionnel (TOC)*, une variante du transport optimal classique incorporant cette contrainte de marginale. En nous appuyant sur la théorie des flots de gradient dans les espaces métriques, nous démontrons l'existence et la cohérence de ce flot avec l'entraînement des ResNets de largeur finie. Ce travail est également l'occasion d'explorer plus en détail les propriétés du TOC et de sa formulation dynamique.

Nous étudions ensuite le comportement asymptotique des flots de gradient en nous appuyant sur des inégalités de type *Polyak-Łojasiewicz locales*. Nous montrons que ces inégalités sont génériquement satisfaites par les ResNets profonds, et établissons des résultats de convergence pour certains exemples d'architectures et d'initialisations : si le nombre de neurones est fini mais suffisamment grand, et si le risque est suffisamment faible à l'initialisation, alors le flot de gradient converge vers un minimiseur global.

Enfin, afin d'étudier l'émergence de *représentations* non-linéaires durant l'apprentissage, nous considérons le cas de réseaux à une seule couche cachée avec une fonction de perte quadratique. Pour ce problème d'optimisation non convexe et de grande dimension, les résultats existants sont souvent qualitatifs, ou fondés sur une analyse par le *neural tangent kernel*, dans laquelle les représentations des données restent figées. Exploitant le fait qu'il s'agit d'un *problème quadratique non-linéaire séparable*, nous analysons un algorithme de *Variable Projection (VarPro)* ou d'*apprentissage à deux vitesses* qui permet d'éliminer les variables linéaires et de réduire le problème d'apprentissage à l'entraînement des paramètres non-linéaires. Dans un cadre "enseignant-élève", nous montrons que, dans la limite d'une régularisation nulle, la dynamique de la distribution des représentations est décrite par une équation de *weighted ultra-fast diffusion*, permettant ainsi d'établir un taux de convergence linéaire pour l'échantillonnage de la distribution enseignante.

Le code pour reproduire les résultats numériques présentés est en open source.

Mots clés : Théorie de l'apprentissage, Apprentissage profond, Optimisation, EDOs neuronales, Flots de gradient Wasserstein

Abstract

The recent successes of deep learning models across a wide range of applications have underscored the need for a deeper understanding of their training dynamics. This research is ultimately motivated by the design of more efficient architectures and learning algorithms.

In this PhD work, we contribute to the theoretical understanding of the dynamics of gradient-based methods for the training of neural networks by studying the case of overparameterized models. Indeed, a recent line of work has proven that, for shallow architectures, good convergence guarantees can be obtained by relaxing the training problem in the space of parameter distributions.

We extend this analysis to the case of deep architectures by studying mean-field models of *deep Residual Neural Networks (ResNets)*. These are parameterized by distributions over a product set of layers and parameter space, with a uniform marginal condition on the set of layers. We then propose to model training with a gradient flow w.r.t. the *Conditional Optimal Transport distance*: a restriction of the classical Optimal Transport distance which enforces the marginal condition. Relying on the theory of gradient flows in metric spaces, we show the well-posedness of the gradient flow equation and its consistency with the training of ResNets at finite width. In addition, this is an opportunity to study in more detail the Conditional Optimal Transport distance, particularly its dynamic formulation.

We then study the asymptotic behavior of gradient flow curves by relying on *local Polyak-Łojasiewicz inequalities*. We show such inequalities are generically satisfied by deep ResNets and prove convergence for well-chosen examples of architectures and initializations: if the number of neurons is finite but sufficiently large and the risk is sufficiently small at initialization, then gradient flow converges to a global minimizer of the training risk at a linear rate.

Finally, to study the learning of nonlinear *features* during training with gradient descent we consider the case of shallow single-hidden-layer neural networks with square loss. For this high-dimensional and non-convex optimization problem, most known convergence results are either qualitative or rely on a *neural tangent kernel* analysis where hidden representations of the data are fixed. Using that this problem belongs to the class of separable nonlinear least squares problems, we consider a *Variable Projection (VarPro)* or *two-timescale learning* algorithm, thereby eliminating the linear variables and reducing the learning problem to the training of nonlinear features. In a "teacher-student" scenario, we show that, in the limit where the regularization strength vanishes, the training dynamic on the feature distribution corresponds to a *weighted ultra-fast diffusion equation*. This provides a linear convergence rate for the sampling of the teacher distribution.

The code for reproducing the numerical results presented in this thesis is open-sourced.

Keywords : Machine learning theory, Deep learning, Optimization, Neural ODEs, Wasserstein gradient flows

Contents

Résumé i

Abstract iii

Table of contents v

Introduction en français 1

Introduction 17

1 Supervised learning: algorithms, architectures and mathematical models . . 17
1.1 The supervised learning framework . . . 18
1.2 Neural network architectures . . . 20
1.3 Scaling neural networks in the infinite width regime . . . 22
1.4 Scaling neural networks in the infinite depth regime . . . 25
1.5 Training and gradient descent algorithms . . . 27
2 Contributions . . . 31
I Training of infinitely deep and wide residual architectures . . . 31
II Convergence in the training of residual architectures . . . 34
III Feature learning in shallow architectures . . . 36

I Training of infinitely deep and wide residual architectures **39**

I.1 Introduction . . . 39
I.1.1 Mean-field models of neural networks . . . 41
I.1.2 Mean-field NODEs . . . 42
I.1.3 Related works and contributions . . . 44
I.2 Metric structure of the parameter set $\mathcal{P}_2^{\mathrm{Leb}}([0,1]\times\Theta)$ . . . 45
I.2.1 Conditional Optimal Transport distance . . . 45
I.2.2 Dynamical formulation of Conditional Optimal Transport . . . 49
I.3 Gradient flow dynamics . . . 55
I.3.1 Backward equation and adjoint variables . . . 55
I.3.2 The gradient flow equation . . . 57
I.3.3 Gradient flows as curves of maximal slope . . . 59
I.3.4 Existence, uniqueness, and stability of gradient flow curves . . . 68
Appendices . . . 75
I.A Well-posedness of the gradient flow equation for SHL residuals . . . 75

II Convergence in the training of residual architectures **79**

II.1 Introduction . . . 79
II.1.1 Related works and contributions . . . 81
II.2 Polyak-Łojasiewicz property and convergence of gradient flow . . . 83

II.2.1 The Polyak-Łojasiewicz property in Hilbert spaces . . . 83
II.2.2 The Polyak-Łojasiewicz property in metric spaces . . . 86
II.3 Convergence for general architectures . . . 89
II.3.1 Conditioning of the tangent kernel implies the P-Ł property . . . 89
II.3.2 Expressivity and functional properties of the set of residuals . . . 91
II.4 Linear parameterization of the residuals . . . 92
II.4.1 Gradient flow equation in the case of RKHS residuals . . . 96
II.4.2 Convergence of RKHS-NODE . . . 99
II.4.3 Convergence with finite width . . . 101
II.5 The case of SHL residuals . . . 104
II.5.1 Comparison with the case of a linear parameterization . . . 106
II.5.2 Convergence of NODEs with SHL residuals . . . 107
II.5.3 Examples of activations and quantitative convergence results . . . 108
II.6 Ensuring convergence with lifting and scaling . . . 111
II.7 Numerical results . . . 112
II.7.1 Experiments on MNIST . . . 113
II.7.2 Experiments on CIFAR10 . . . 116
II.8 Conclusion . . . 118

III Feature learning in shallow architectures **119**
III.1 Introduction . . . 119
III.1.1 Mean-field neural networks and two-timescale learning . . . 120
III.1.2 Contributions and related works . . . 125
III.2 Reduced risk associated to the VarPro algorithm . . . 127
III.2.1 Primal formulation of the reduced risk . . . 128
III.2.2 Partial minimization on the space of measures . . . 129
III.2.3 Dual formulation of the reduced risk . . . 130
III.2.4 Kernel learning in the case of quadratic regularization . . . 132
III.3 Properties of minimizers of the reduced risk . . . 132
III.3.1 Existence and uniqueness of minimizers . . . 132
III.3.2 Convergence of minimizers . . . 134
III.4 Training with gradient flow . . . 135
III.4.1 Wasserstein gradient flows in the case $\lambda > 0$ . . . 136
III.4.2 Wasserstein gradient flows in the case $\lambda = 0$ and ultra-fast diffusions 140
III.5 Convergence of gradient flow . . . 143
III.5.1 Algebraic convergence rate . . . 143
III.5.2 Convergence to ultra-fast diffusion. . . . 145
III.6 Numerics . . . 148
III.6.1 Single-hidden-layer neural networks with 1-dimensional feature space 148
III.6.2 VarPro for image classification on CIFAR10 . . . 155
III.7 Conclusion . . . 159
Appendices . . . 160
III.A Positive definite kernels and RKHS . . . 160
III.B Radial basis function neural network on the 2-dimensional torus . . . 162

Conclusion **167**

List of publications **171**

References **173**

Introduction en français : algorithmes, architectures et modèles mathématiques pour l'apprentissage supervisé

Table des matières

1 Apprentissage supervisé **2**
2 Architectures de réseaux de neurones **4**
3 Mise à l'échelle des réseaux de neurones dans le régime de largeur infinie **7**
3.1 Régime *Neural Tangent Kernel* 8
3.2 Modèles champ-moyen de réseaux de neurones 8
3.3 Expressivité et propriétés fonctionnelles des réseaux de neurones 9
4 Mise à l'échelle des réseaux de neurones dans le régime de grande profondeur **10**
4.1 Réseaux de neurones résiduels 10
4.2 Équations différentielles ordinaires neuronales 11
5 Apprentissage et algorithmes de descente de gradient **12**
5.1 Descente de gradient stochastique et variantes 13
5.2 Apprentissage à deux échelles de temps et projection de la variable 13
5.3 Flots de gradient de Wasserstein et transport optimal 14

Au cours des dernières années, l'apprentissage profond a connu des succès remarquables dans un large panel d'applications, allant de la génération d'images et de textes au calcul scientifique, et plus récemment à des tâches de raisonnement telles que la résolution de problèmes mathématiques complexes. Cependant, bien qu'un nombre croissant de travaux cherchent à apporter une meilleure compréhension des systèmes d'IA et à améliorer leur conception, ces succès dépassent souvent notre compréhension des mécanismes mathématiques sous-jacents.

D'un point de vue mathématique, l'entraînement des réseaux de neurones soulève de nombreuses questions théoriques. D'une part, les problèmes d'optimisation en jeu sont généralement non-convexes et en très grande dimension. Pourtant, des algorithmes simples, tels que la descente de gradient stochastique, obtiennent d'excellentes performances en pratique. D'autre part, les réseaux de neurones sont capables d'interpoler de larges ensembles de données tout en généralisant efficacement. Cela va ainsi à l'encontre de certains principes statistiques fondamentaux, tels que le *compromis biais–variance* ou la *malédiction de la dimension*. Ces phénomènes soulignent donc la nécessité de nouveaux cadres mathématiques capables de mieux décrire les dynamiques d'entraînement des réseaux de neurones et leurs interaction avec les architectures et les structures de données. En par-

ticulier, des travaux récents suggèrent que les outils issus de l'analyse des équations aux dérivées partielles (ÉDP) et du transport optimal peuvent apporter un éclairage précieux sur ces dynamiques d'entraînement.

Dans ce manuscrit, nous adoptons un point de vue mathématique sur l'entraînement des réseaux de neurones, fondé sur des outils d'optimisation et de théorie des équations aux dérivées partielles. Dans une limite d'architectures larges d'une part, la dynamique d'entraînement des réseaux peut être décrite par des modèles champ-moyen issus des systèmes de particules en interaction, correspondant à des flots de gradient dans des espaces de distributions de probabilité. D'autre part, les architectures résiduelles sont étudiées dans leur limite de grande profondeur, laquelle conduit à des paysages d'optimisation plus réguliers et à des dynamiques d'entraînement plus stables. Ces deux régimes asymptotiques, grande largeur et grande profondeur, ne sont pas de simples constructions théoriques, mais reflètent la structure des architectures modernes telles que les *ResNets* ou les *Transformers*, qui sont fortement surparamétrées et au cœur des modèles d'IA les plus performants actuellement.

1 Apprentissage supervisé

Dans l'ensemble de ce manuscrit, nous considérons un cadre d'*apprentissage supervisé* englobant un grand nombre de tâches classiques en apprentissage automatique. Nous commençons par en décrire les principaux éléments constitutifs, avant de détailler plus précisément les modèles et algorithmes étudiés dans ce manuscrit.

Jeu de données En apprentissage supervisé, la machine dispose d'un jeu de données $\mathcal{D} \subset \mathcal{X} \times \mathcal{Y}_{targ}$ constitué de paires de données d'entrée $x \in \mathcal{X}$ et de réponses cibles associées $y \in \mathcal{Y}_{targ}$. Ces données d'entrée et de sortie peuvent prendre des formes très variées :

- **Entrées :** En raison de la grande flexibilité des méthodes d'apprentissage automatique, les données d'entrée peuvent être de nature diverse : images, sons, vidéos, textes ou encore séries temporelles financières. Un exemple d'application que nous considérerons aux Chapter II et Chapter III est la classification d'images, où les données d'entrée sont des images numériques encodées sous forme de tableaux d'entiers sur 8 bits de taille $n_c \times n_w \times n_h$, où n_w et n_h désignent respectivement le nombre de pixels en largeur et en hauteur, et n_c le nombre de canaux (généralement $n_c = 1$ pour les images en niveaux de gris et $n_c = 3$ pour les images en couleur). Mathématiquement, ces images peuvent être modélisées comme des vecteurs dans l'espace vectoriel $\mathcal{X} = \mathbb{R}^{n_c \times n_w \times n_h}$.

- **Cibles :** On distingue en général deux grandes catégories de tâches d'apprentissage supervisé : la classification et la régression. En classification, l'objectif est d'associer chaque donnée d'entrée à l'une des classes d'un ensemble fini, représenté par des *étiquettes* ou *labels* dans $\mathcal{Y}_{targ} = \{1, \dots, C\}$, où $C \geq 1$ est le nombre de classes. Ces étiquettes peuvent également être encodées sous forme de vecteurs *one-hot* dans $\mathcal{Y}_{targ} = \{0, 1\}^C$. À l'inverse, en régression, l'objectif est de prédire un signal vectoriel dans $\mathcal{Y}_{targ} = \mathbb{R}^{d_{out}}$.

Dans la suite, les espaces de données d'entrée $\mathcal{X}$ et de données cibles $\mathcal{Y}_{targ}$ seront toujours supposés être des sous-ensembles d'espaces vectoriels réels de dimension finie. Il est alors standard de voir chaque paire $(x, y) \in \mathcal{X} \times \mathcal{Y}_{targ}$ comme la réalisation d'une variable aléatoire dont la loi sera également notée $\mathcal{D}$.

Fonction de perte L'objectif de la machine est d'apprendre, à partir des exemples de $\mathcal{D}$, une *fonction de prédiction* ou *prédicteur* $F : \mathcal{X} \to \mathcal{Y}_{out}$, associant à chaque entrée $x \in \mathcal{X}$ une prédiction de la réponse cible $y_{targ} \in \mathcal{Y}_{targ}$. L'espace des sorties $\mathcal{Y}_{out}$ est un espace vectoriel qui n'est pas nécessairement identique à celui des cibles $\mathcal{Y}_{targ}$. Pour évaluer la qualité de ses prédictions, la machine dispose d'une *fonction de perte* $\ell : \mathcal{Y}_{out} \times \mathcal{Y}_{targ} \to \mathbb{R}$. Nous considérerons deux exemples fondamentaux :

- **Régression :** Dans un problème de régression, les espaces des sorties et de cibles coïncident : $\mathcal{Y}_{out} = \mathcal{Y}_{targ} = \mathbb{R}^{d_{out}}$. Cet espace est muni de la géométrie euclidienne standard, et une mesure naturelle de l'erreur entre une prédiction y_{out} et une cible y_{targ} est donnée par la *perte quadratique* :

$$\ell(y_{out}, y_{targ}) = \frac{1}{2}|y_{out} - y_{targ}|^2_{\mathbb{R}^{d_{out}}} . \tag{1}$$

- **Classification :** Dans un problème de classification à C classes, la machine produit en général des sorties dans $\mathcal{Y}_{out} = \mathbb{R}^C$ représentant des estimations des log-probabilités a posteriori de chaque classe donnée l'entrée. Une prédiction $y_{out} \in \mathcal{Y}_{out}$ est alors comparée à une étiquette cible $y_{targ} \in \mathcal{Y}_{targ} = \{1, \dots, C\}$ à l'aide de la *fonction d'entropie croisée* :

$$\ell(y_{out}, y_{targ}) = -\log\left(\frac{\exp(y_{out}[y_{targ}])}{\sum_{i=1}^{C}\exp(y_{out}[i])}\right) . \tag{2}$$

Le problème de minimisation du risque Étant donné un jeu de données $\mathcal{D} \subset \mathcal{X} \times \mathcal{Y}_{targ}$ et une fonction de perte $\ell : \mathcal{Y}_{out} \times \mathcal{Y}_{targ} \to \mathbb{R}$, la qualité d'une fonction de prédiction $F : \mathcal{X} \to \mathcal{Y}_{out}$ peut être évaluée en moyennant la perte sur l'ensemble des exemples de $\mathcal{D}$. La stratégie de l'apprentissage automatique consiste à rechercher le meilleur prédicteur au sein d'une classe de fonctions paramétriques $\mathcal{F} = \{F_\theta \mid \theta \in \Theta\}$, où Θ désigne l'espace des paramètres. Dans le cas des réseaux de neurones, Θ correspond à l'espace des poids du réseau, généralement un espace vectoriel de grande dimension muni de la métrique euclidienne. Pour chaque paramètre $\theta \in \Theta$, on définit le *risque d'entraînement* par :

$$\mathcal{R}(\theta) \coloneqq \frac{1}{\#\mathcal{D}} \sum_{(x,y)\in\mathcal{D}} \ell(F_\theta(x), y) . \tag{3}$$

L'entraînement du modèle paramétrique F_θ consiste alors à résoudre un *problème de minimisation du risque* :

$$\text{Trouver} \quad \theta^* \in \arg\min_{\theta\in\Theta} \mathcal{R}(\theta) . \tag{4}$$

En pratique, cette optimisation est souvent réalisée à l'aide d'algorithmes itératifs du premier ordre, comme la *descente de gradient*. Partant d'un paramètre initial $\theta_0 \in \Theta$, les paramètres sont mis à jour selon la règle suivante :

$$\forall k \geq 0, \quad \theta_{k+1} = \theta_k - \tau\nabla_\theta\mathcal{R}(\theta_k) ,$$

où $\tau > 0$ désigne le *pas de gradient*. En apprentissage profond, pour permettre l'entraînement sur de grands jeux de données et améliorer la généralisation, le risque est souvent estimé

à chaque itération $k \geq 0$ sur un sous-ensemble $\mathcal{D}_k \subset \mathcal{D}$ de données échantillonnées aléatoirement. On obtient ainsi l'algorithme de *descente de gradient stochastique* :

$$\forall k \geq 0, \quad \theta_{k+1} = \theta_k - \tau \nabla_\theta \mathcal{R}_k(\theta_k),, \quad \text{où} \quad \mathcal{R}_k(\theta) \coloneqq \frac{1}{\#\mathcal{D}_k} \sum_{(x,y)\in\mathcal{D}_k} \ell(F_\theta(x), y)\,.$$

Dans les deux cas, le choix du modèle paramétrique ainsi que des hyperparamètres (tels que τ ou la taille des mini-lots) influence fortement la dynamique d'entraînement et les performances de généralisation du modèle appris. Dans la suite de cette introduction, nous détaillerons les architectures de réseaux de neurones et les procédures d'entraînement qui constituent le cœur de cette thèse.

Aspect statistique de l'apprentissage Bien que ce manuscrit adopte une approche centrée sur l'optimisation, en se concentrant sur la minimisation du risque d'entraînement, il est important de rappeler que l'objectif final de l'apprentissage supervisé est de construire une fonction de prédiction performante sur des exemples nouveaux. Dans le cadre statistique classique, les points de données (x, y) du jeu d'entraînement $\mathcal{D}$ sont supposés indépendants et identiquement distribués selon une loi inconnue $\mathcal{D}_{test}$ sur $\mathcal{X}\times\mathcal{Y}_{targ}$. L'objet central d'intérêt est alors l'*erreur de test*, définie par

$$\mathcal{E}_{test}(\theta) \coloneqq \mathbb{E}_{(x,y)\sim\mathcal{D}test}\left[\ell_{test}(F_\theta(x), y)\right]\,,$$

où la perte de test ℓ_{test} peut différer de la perte d'entraînement ℓ. Le problème de minimisation du risque d'entraînement $\mathcal{R}$ sert ainsi d'approximation à celui du risque de test $\mathcal{E}_{test}$, le principal défi résidant dans le fait que la distribution $\mathcal{D}_{test}$ est inconnue et que l'apprentissage doit s'effectuer à partir du nombre fini d'exemples contenus dans $\mathcal{D}$. Si la question des capacités de généralisation des modèles entraînés dépasse le cadre principal de cette thèse, elle motive néanmoins de nombreux choix de modélisation et d'algorithmes présentés dans ce manuscrit.

Apprentissage auto-supervisé Enfin, bien que ce manuscrit se concentre sur les tâches d'apprentissage supervisé, il convient de noter que de nombreux systèmes modernes d'apprentissage automatique sont entraînés de manière *auto-supervisé*, c'est à dire où le signal cible est dérivé directement des données d'entrée. Cette approche peut être vue comme un cas particulier d'apprentissage supervisé, dans lequel les cibles sont construites à partir de données non-annotées. Les exemples principaux sont les problèmes de *next token prediction* en modélisation du langage, ou l'entraînement de *modèles de diffusion* pour la génération d'images. Ces méthodes se sont révélées particulièrement efficaces pour exploiter de vastes ensembles de données non-étiquetées et préentraîner des modèles destinés à des tâches ultérieures.

2 Architectures de réseaux de neurones

La famille de modèles paramétriques que nous considérerons dans ce manuscrit est celle des *réseaux de neurones*. Ces modèles consistent en la composition successive de couches, chacune étant elle-même une transformation paramétrique élémentaire. Un réseau de neurones de profondeur $D \geq 1$ est ainsi un modèle paramétré par $\theta \in \Theta = \prod_{d=1}^{D} \Theta_d$ qui, pour une entrée $x \in \mathcal{X}$, renvoie :

$$F_\theta(x) = F_{\theta_D} \circ \cdots \circ F_{\theta_1}(x)\,,$$

où, pour chaque $d \in 1, \dots, D$, la d-ième couche F_{θ_d} est elle-même un (petit) réseau de neurones paramétré par $\theta_d \in \Theta_d$. Considérant $\mathcal{X} = \mathbb{R}^{d_{in}}$ et $\mathcal{Y}_{out} = \mathbb{R}^{d_{out}}$ pour $d_{in}, d_{out} \geq 1$, nous commençons par décrire ici quelques exemples et propriétés d'architectures $F_\theta : \mathbb{R}^{d_{in}} \to \mathbb{R}^{d_{out}}$ dites "peu profondes". Elles constituent les briques élémentaires des architectures plus profondes que nous présenterons ultérieurement.

- **Couches linéaires :**

  Les couches linéaires, ou *fully-connected*, réalisent des multiplications matrice–vecteur. Étant donnée une entrée $x \in \mathbb{R}^{d_{in}}$, la sortie est donnée par :

  $$F_W(x) = W \cdot x \,,$$

  où le paramètre $W \in \mathbb{R}^{d_{out} \times d_{in}}$ est une matrice dite de "poids". Ces transformations linéaires constituent les blocs de base de la plupart des architectures de réseaux de neurones. En pratique, les modèles modernes d'apprentissage profond sont généralement construits en composant ces applications linéaires avec des fonctions non-linéaires simples.

- **Couches convolutionnelles :**

  Les couches convolutionnelles sont un cas particulier de couches linéaires dans lesquelles la matrice de poids est contrainte à avoir une structure spécifique, celle d'une matrice de convolution. Introduites par LeCun et al. [LeCun, 1989] pour la reconnaissance de chiffres manuscrits, ces architectures, en raison de leur structure équivariante par translation, sont devenues omniprésentes dans les applications de traitement d'images [LeCun, 2015]. Une couche convolutionnelle est paramétrée par un ensemble de filtres W et, pour une image d'entrée x, renvoie :

  $$F_W(x) = W \star x \,, \tag{5}$$

  où $\star$ désigne l'opérateur de convolution discrète. Par exemple, si $x \in \mathbb{R}^{c_{in} \times d_w \times d_h}$ est une image comportant c_{in} canaux d'entrée et si $W \in \mathbb{R}^{c_{out} \times c_{in} \times k \times k}$ est un filtre convolutionnel de taille $k \times k$ avec c_{out} canaux de sortie, le résultat de la convolution discrète s'écrit :

  $$(W \star x)[c, i, j] = \sum_{1 \leq k_1, k_2 \leq k} \sum_{1 \leq c' \leq c_{in}} W[c, c', k_1, k_2], x[c', i + k_1, j + k_2] \,. \tag{6}$$

  Nous utiliserons les réseaux de neurones convolutionnels aux Chapter II et Chapter III pour résoudre des problèmes de classification d'images.

- **Modèles linéaires dans l'espace des paramètres :**

  Une classe importante de modèles d'apprentissage est celle des modèles linéaires en leurs paramètres mais non nécessairement linéaires en leurs entrées. C'est par exemple le cas des méthodes à noyau [Schölkopf, 2002; Steinwart, 2008] ou des modèles à "représentations aléatoires" [Rahimi, 2007]. Ces modèles possèdent un espace de paramètres $\Theta = \mathcal{H}^{d_{out}}$, où $\mathcal{H}$ est un espace de Hilbert de "représentations", et calculent, pour un paramètre $\theta \in \Theta$ et une entrée $x \in \mathbb{R}^{d_{in}}$:

  $$F_\theta(x) = \begin{pmatrix} \langle \theta_1, \phi(x) \rangle_{\mathcal{H}} \\ \vdots \\ \langle \theta d_{out}, \phi(x) \rangle_{\mathcal{H}} \end{pmatrix} , \tag{7}$$

où $\phi : \mathcal{X} \to \mathcal{H}$ est une application associant à chaque entrée une représentation dans $\mathcal{H}$. Alors que les réseaux de neurones classiques sont non-linéaires à la fois en leurs entrées et en leurs paramètres, ces modèles présentent l'avantage d'être linéaires dans l'espace des paramètres, ce qui facilite leur analyse théorique. Nous étudierons cette classe de modèles dans la section Section II.4, comme étape préliminaire à l'analyse d'architectures plus complexes.

- **Couches perceptron :**

  Le *perceptron* est sans doute l'un des exemples les plus simples d'architecture de réseau de neurones non-linéaire à la fois en ses entrées et en ses paramètres. Initialement introduit par Rosenblatt [Rosenblatt, 1958] pour reproduire certaines capacités visuelles et perceptuelles humaines, il peut être vu comme la composition de deux couches entièrement connectées séparées par une fonction non-linéaire. Un perceptron à deux couches, ou réseau à une seule couche cachée (SHL) de largeur $M \geq 1$, est paramétré par deux matrices de poids U et W de dimensions respectives $d_{out} \times M$ et $d_{in} \times M$, ainsi qu'un biais $b \in \mathbb{R}^M$. Pour une entrée $x \in \mathbb{R}^{d_{in}}$, il renvoie :

$$F_{(U,W,b)}(x) = U\,\sigma(W^\top x + b)\,, \tag{8}$$

  où $\sigma : \mathbb{R} \to \mathbb{R}$ est une fonction non-linéaire, appelée "fonction d'activation", appliquée composante par composante. Les fonctions d'activation les plus utilisées incluent la tangente hyperbolique tanh et l'"Unité Linéaire Rectifiée" (*Rectified Linear Unit* ou *ReLU*). Nous étudierons cette classe de modèles dans les sections Section II.5 et Chapter III.

- **Couches d'attention :**

  Le mécanisme d'*attention* [Bahdanau, 2014; Vaswani, 2017] est au cœur des architectures de type *Transformers*, qui se sont imposées comme modèles de référence en vision par ordinateur [Dosovitskiy, 2020], en *traitement du langage naturel (NLP)* [Devlin, 2019], ainsi que dans d'autres tâches de traitement ou de génération de séquences. Une "tête d'attention" (*attention head*) est paramétrée par trois matrices $Q, K, V \in \mathbb{R}^{d_{in} \times d_{in}}$ et, pour une séquence d'entrée de "jetons" (*tokens*) $\mathbf{x} = (x_i)1 \leq i \leq N \in (\mathbb{R}^{d_{in}})^N$ de longueur N, renvoie :

$$\text{Attention}_{Q,K,V}(\mathbf{x}) = \left(\sum_{j=1}^{N} \frac{e^{\langle Qx_i, Kx_j \rangle}}{\sum_{j=1}^{N} e^{\langle Qx_i, Kx_j \rangle}} V x_j \right)_{1 \leq i \leq N} \in (\mathbb{R}^{d_{in}})^N\,. \tag{9}$$

  Dans le cadre du NLP, ces jetons représentent des plongements de mots ou de syllabes, sur lesquels les modèles sont entraînés de manière auto-supervisée à prédire les prochains jetons. Dans les grands modèles de langage modernes, tels que les *Generative Pretrained Transformers (GPTs)* [Radford, 2018], des perceptrons multicouches sont empilés avec des couches d'attention à plusieurs têtes, où plusieurs opérations d'attention sont effectuées en parallèle.

- **Couches non-paramètrées :**

  Dans les architectures modernes de réseaux de neurones, les couches paramétriques sont souvent combinées à diverses opérations non-paramétriques, conçues pour améliorer l'expressivité et la stabilité de l'entraînement. La composition avec une fonction d'activation non-linéaire peut par exemple être vue comme une forme simple de couche sans paramètres. De plus, bien que nous les omettions pour la clarté de la

présentation, les architectures modernes incluent fréquemment des couches de *pooling*, qui réduisent les dimensions des représentations, ainsi que des couches de normalisation, connues pour faciliter l'entraînement des réseaux profonds [Ioffe, 2015; Ba, 2016]. Dans le contexte du NLP, les opérations sans paramètres incluent également des "encodages positionnels", qui injectent une information d'ordre dans les représentations de séquences, ainsi que des mécanismes de "masquage" (*masking*), qui contraignent le flux d'information (par exemple pour préserver la causalité dans les modèles autorégressifs) [Vaswani, 2017].

3 Mise à l'échelle des réseaux de neurones dans le régime de largeur infinie

La dernière décennie a vu une augmentation exponentielle de la taille des architectures de réseaux de neurones, les modèles modernes comptant ainsi des milliards, voire des milliers de milliards de paramètres [Villalobos, 2022]. Cela révèle pourtant un phénomène contre-intuitif : de nombreux modèles opèrent dans un régime surparamétré, où le nombre de paramètres entraînables dépasse le nombre de points de données disponibles. En statistique classique, une telle situation conduirait typiquement à un surapprentissage et à une mauvaise généralisation. Pourtant, en pratique, les réseaux de neurones surparamétrés généralisent souvent remarquablement bien [Belkin, 2019; Zhang, 2021]. D'importants efforts théoriques ont donc été consacrés à la compréhension du comportement des réseaux de neurones dans le régime de largeur infinie, c'est-à-dire lorsque le nombre de neurones par couche tend vers l'infini. Au-delà de leur intérêt théorique, ces analyses asymptotiques présentent également des bénéfices pratiques, notamment pour le choix et le transfert d'hyperparamètres entre architectures de largeurs différentes [Yang, 2021; Bordelon, 2025], conduisant à d'importantes économies de calcul dans l'entraînement de modèles de grande taille [OpenAI, 2023].

La plupart des architectures de réseaux de neurones présentées précédemment peuvent être représentées comme des applications de la forme :

$$F_{(\theta_i)_{1\leq i\leq M}} : x \in \mathbb{R}^{d_{in}} \mapsto \alpha_M \sum_{i=1}^{M} \psi(\theta_i, x) \in \mathbb{R}^{d_{out}}\,, \tag{10}$$

où Θ désigne l'espace des paramètres, $\psi : \Theta \times \mathbb{R}^{d_{in}} \to \mathbb{R}^{d_{out}}$ est une fonction élémentaire, et $\alpha_M \in \mathbb{R}$ est un facteur d'échelle dépendant de la largeur du réseau M. Par exemple, un perceptron à deux couches correspond au cas où $\Theta = \mathbb{R}^{d_{out}} \times \mathbb{R}^{d_{in}} \times \mathbb{R}$ et où ψ est donnée par :

$$\psi : ((u, w, b), x) \in \Theta \times \mathbb{R}^{d_{in}} \mapsto u\sigma(w^\top x + b)\,, \tag{11}$$

avec $\sigma : \mathbb{R} \to \mathbb{R}$ une fonction d'activation.

Les paramètres du modèle sont généralement initialisés aléatoirement et d'ordre 1, et des travaux récents ont mis en évidence le rôle crucial joué par le choix du facteur d'échelle α_M dans la dynamique d'entraînement des modèles de grande largeur M. Différents choix de mise à l'échelle conduisent à des comportements asymptotiques distincts lorsque M tend vers l'infini. Deux cadres théoriques principaux ont ainsi émergé : le régime champ-moyen (*mean-field*), qui capture l'apprentissage de représentations non-linéaires, et le régime *Neural Tangent Kernel (NTK)*, qui décrit une dynamique d'entraînement linéarisée autour de l'initialisation aléatoire.

3.1 Régime *Neural Tangent Kernel*

Un premier cadre asymptotique pour l'analyse des réseaux de neurones dans la limite de largeur infinie est celui du Neural Tangent Kernel (NTK) [Jacot, 2018]. Ce régime correspond à une mise à l'échelle des paramètres de l'ordre $\alpha_M = 1/\sqrt{M}$ pour un réseau de largeur M, sous laquelle l'évolution du réseau pendant la descente de gradient peut être approchée par une linéarisation autour de son initialisation aléatoire $\theta_0 \in \Theta$. Dans ce régime linéarisé, le modèle est linéaire par rapport à ses paramètres, comme dans Eq. (7). Le réseau de neurones se ramène alors à une méthode à noyau [Schölkopf, 2002; Steinwart, 2008] dont le noyau associé :

$$K(x, x') = \mathrm{D}_\theta F_{\theta_0}(x) \cdot \mathrm{D}_\theta F_{\theta_0}(x')^\top , \tag{12}$$

appelé *Neural Tangent Kernel*, converge vers une limite déterministe dans la limite de largeur infinie. Ce cadre conduit à des résultats théoriques forts : on peut montrer que la descente de gradient converge vers un minimum global du risque empirique à une vitesse linéaire, gouvernée par les propriétés spectrales du NTK [Allen-Zhu, 2019; Du, 2019; Lee, 2019; Zou, 2020]. Nous étudierons plus en détail le conditionnement du NTK associé aux perceptrons à deux couches dans la Section II.5.

Cependant, le régime NTK présente d'importantes limitations. Notamment, il induit une forme d'"apprentissage paresseux" (*lazy training*) [Chizat, 2019], dans lequel les paramètres du réseau se déplacent très peu par rapport à leur initialisation, et où les représentations apprises évoluent peu au cours de l'entraînement. En conséquence, le modèle ne parvient pas à extraire de représentations non-linéaires des données et se comporte essentiellement comme une méthode à noyau. À l'inverse, les réseaux de neurones bénéficient en pratique de capacité d'apprentissage hiérarchiques ou spécifiques à certaines tâches, conduisant à de meilleures capacités de généralisation [Bach, 2017a; Ghorbani, 2019; Ghorbani, 2020].

3.2 Modèles champ-moyen de réseaux de neurones

Un autre cadre asymptotique est le régime de champ-moyen (*mean-field*), correspondant à un facteur de mise à l'échelle de la sortie du modèle de $1/M$ pour une largeur M. L'une des caractéristiques essentielles de ce régime, contrairement au cadre NTK, est sa capacité à capturer l'apprentissage de représentations non-linéaires [Yang, 2021].

Avec le facteur d'échelle $\alpha_M = 1/M$, le réseau peut être interprété comme une intégrale sur une distribution de paramètres. En effet, pour une famille de paramètres $(\theta_i)_{1\leq i\leq M} \in \Theta^M$, en considérant la mesure empirique $\hat{\mu} = \frac{1}{M}\sum_{i=1}^M \delta_{\theta_i}$, l'équation Eq. (10) s'écrit :

$$\forall x \in \mathbb{R}^{d_{in}} , \quad F_{(\theta_i)1\leq i\leq M}(x) = \frac{1}{M}\sum_{i=1}^{M} \psi(\theta_i, x) = \int_\Theta \psi(\theta, x)\mathrm{d}\hat{\mu}(\theta) = F_{\hat{\mu}}(x) ,$$

où, pour toute mesure de probabilité μ sur l'espace des paramètres Θ, on définit :

$$F_\mu : x \in \mathbb{R}^{d_{in}} \mapsto \int_\Theta \psi(\theta, x)\mathrm{d}\mu(\theta) \ \in \mathbb{R}^{d_{out}} . \tag{13}$$

Cette représentation englobe donc les réseaux de neurones de largeur finie (quand μ est une mesure empirique), mais décrit aussi, lorsque $M \to \infty$, une limite de champ-moyen où μ peut être une mesure de probabilité arbitraire [Rotskoff, 2018; Chizat, 2018; Mei, 2019; Sirignano, 2020].

D'un point de vue mathématique, au-delà de l'élimination de la dépendance en M, la représentation champ-moyen de Eq. (13) permet de capturer naturellement l'interchangeabilité des neurones. En effet, l'invariance par permutation de l'indice $i \in 1, ..., M$ dans Eq. (10) induit des symétries dans le paysage du risque, ce qui complique son analyse. Elle permet également de relaxer le problème de minimisation du risque Eq. (4) dans l'espace des mesures, menant ainsi à un paysage d'optimisation plus simple [Chizat, 2018; Rotskoff, 2019].

Enfin, la représentation champ-moyen permet d'étudier la dynamique d'entraînement à travers le prisme des flots de gradient dans l'espace $\mathcal{P}(\Theta)$ des mesures de probabilité sur Θ [Ambrosio, 2008b; Santambrogio, 2017]. Cela conduit à des équations aux dérivées partielles non-locales dont la convergence peut être étudiée qualitativement [Chizat, 2018; Rotskoff, 2019] ou quantitativement, à condition que le risque vérifie certaines inégalités fonctionnelles [Mei, 2019; Chizat, 2022; Nitanda, 2022]. Cependant, bien que le régime champ-moyen permette une approximation plus fidèle de comportements d'apprentissage réalistes, les résultats de convergence existants restent principalement qualitatifs : ils ne fournissent ni taux explicites de convergence, ni caractérisations complètes des performances de généralisation. Cela constitue un axe de recherche important, et le cœur de nos contributions au Chapter III.

3.3 Expressivité et propriétés fonctionnelles des réseaux de neurones

Bien que définis par des structures compositionnelles simples, le succès des réseaux de neurones reposent sur de puissantes propriétés d'expressivité. Les propriétés fonctionnelles de l'espace des applications représentables par un réseau de neurones jouent également un rôle essentiel dans ses performances d'apprentissage et de généralisation. Ces propriétés dépendent à la fois de l'architecture et de la structure métrique de l'espace des paramètres, et seront au cœur de notre analyse au Chapter II.

Un exemple important est la famille des perceptrons à deux couches de largeur arbitraire avec des fonctions d'activation non-linéaires. Un résultat fondateur de Cybenko [Cybenko, 1989] a établi que de tels réseaux sont denses dans l'espace des fonctions continues pour la topologie uniforme sur les compactes. Plus tard, Barron [Barron, 1993] a fourni des bornes d'approximation quantitatives en norme L^2, montrant qu'une large classe de fonctions peut être approximée à un taux $\mathcal{O}(1/\sqrt{M})$, où M désigne la largeur de la couche cachée. Fait remarquable, ce taux ne dépend pas de la dimension des données d'entrée, suggérant que les réseaux de neurones ne sont, en théorie, pas impactés par la malédiction de la dimension. Cependant, ces résultats sont non constructifs : ils garantissent l'existence d'approximations précises sans fournir de méthode pratique pour les obtenir. Cette limitation souligne le rôle central des algorithmes d'optimisation en pratique, car il faut s'appuyer sur eux pour découvrir de bonnes approximations.

Dans le cas de l'activation ReLU, l'espace des fonctions représentées par des perceptrons à deux couches est décrit par l'espace de Barron [E, 2021; E, 2022] :

$$\mathcal{B} := \left\{ F : x \in \mathbb{R}^{d_{in}} \mapsto \int u \,\mathrm{ReLU}(w^\top x + b) \mathrm{d}\mu(u, w, b)\,,\ \mu \in \mathcal{P}(\mathbb{R} \times \mathbb{R}^{d_{in}} \times \mathbb{R}) \right\}.$$

Lorsque l'ensemble des poids est muni de la métrique euclidienne standard, il est naturellement associé à une norme d'espace de Banach :

$$\forall f \in \mathcal{B}\,, \quad |F|_{\mathcal{B}} := \inf \left\{ \int |u|(|w| + |b|)\mathrm{d}\mu\,,\ \mu \in \mathcal{P}_2(\mathbb{R} \times \mathbb{R}^{d_{in}} \times \mathbb{R})\,,\ f = F_\mu \right\}.$$

Cet espace peut être caractérisé comme le plus petit espace de Banach de fonctions approximables efficacement par des perceptrons à deux couches. Il contient par exemple tous les espaces de Sobolev de régularité suffisante [E, 2022].

4 Mise à l'échelle des réseaux de neurones dans le régime de grande profondeur

En dépit de l'expressivité déjà remarquable des architectures peu profondes, les avancées récentes en apprentissage automatique reposent largement sur la composition de fonctions. Dans de nombreuses tâches classiques d'apprentissage supervisé, les modèles à l'état de l'art s'appuient désormais sur des réseaux de neurones "profonds", qui prennent la forme :

$$F_\theta(x) = F_{\theta_D} \circ \cdots \circ F_{\theta_1}(x)\,,$$

où chaque F_{θ_d} désigne un sous-réseau plus simple (par exemple un perceptron, une couche convolutionnelle, un mécanisme d'attention, une couche de normalisation, etc...), et où la profondeur D est généralement très grande.

Si cette augmentation de profondeur accroît considérablement l'expressivité de la classe de modèles [Montufar, 2014], elle introduit également d'importants défis en terme d'optimisation. En particulier, il a été observé que l'erreur d'entraînement des réseaux convolutionnels profonds tend à se dégrader lorsque la profondeur dépasse un certain seuil [Srivastava, 2015; He, 2016a]. De plus, l'apprentissage de réseaux très profonds souffre fréquemment d'instabilités numériques, telles que des problèmes d'evanescence/explosion des gradients, où les gradients deviennent respectivement trop faibles ou trop grands dans les premières couches, compromettant ainsi l'efficacité de l'apprentissage [Bengio, 1994; Glorot, 2010]. Ces difficultés ont motivé le développement d'architectures spécifiques facilitant l'entraînement de réseaux très profonds. Parmi celles-ci, les architectures dites "résiduelles" ont rencontré d'importants succès.

4.1 Réseaux de neurones résiduels

Les réseaux de neurones résiduels (*ResNets*) sont une classe d'architectures de réseaux de neurones introduite par He et al. [He, 2016a; He, 2016b] pour des applications en classification d'images. L'idée fondamentale des ResNets consiste à paramétrer chaque couche comme une petite perturbation, appelée "résidu", de l'application identité. Concrètement, cette idée se traduit par la présence de connexions "saute-couche" (*skip connections*) qui permettent de réinjecter le signal entre des couches successives.

Un ResNet de profondeur $D \geq 1$, recevant une entrée $x \in \mathcal{X}$, produit une sortie x_D, où les données sont traitées de manière récursive selon :

$$\forall d \in \{0, \dots, D-1\}\,, \quad x_{d+1} = \underbrace{x_d}_{\text{connexion saute-couche}} + \underbrace{F_{\theta_d}(x_d)}_{\text{résidu}}, \quad \text{avec} \quad x_0 = x\,. \tag{14}$$

Une illustration d'une architecture ResNet est présentée en Fig. 1. Les applications résiduelles F_{θ_d} correspondent à de petites sous-architectures de réseaux de neurones, adaptées au type des données considérées. Par exemple, on utilise généralement des couches convolutionnelles pour le traitement d'images [He, 2016a; He, 2016b] ou des couches à base d'attention dans les Transformers pour le traitement du langage [Vaswani, 2017]. Notons que, bien que l'équation Eq. (14) contraigne les dimensions de sortie de chacune des couches à être identiques, les architectures ResNet incluent en pratique plusieurs couches de sous-échantillonnage qui réduisent progressivement la dimension du signal.

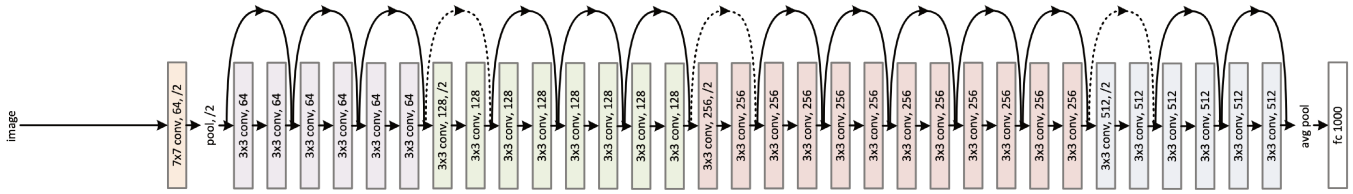


Figure 1: Illustration de l'architecture ResNet-34 [He, 2016a].

Les ResNets ont démontré des performances empiriques remarquables, atteignant l'état de l'art sur des jeux de données de référence tels que CIFAR-10 [Krizhevsky, 2009] et ImageNet [Deng, 2009]. L'une de leurs principales innovations réside dans le mécanisme de connexion saute-couche, qui permet de réduire les problèmes d'évanescence ou d'explosions des gradients [Raiko, 2012; Szegedy, 2017], et rend possible l'entraînement de réseaux comportant plusieurs centaines, voire plusieurs milliers de couches [He, 2016b]. Cette augmentation de la profondeur a soulevé de nouvelles questions théoriques concernant le comportement des algorithmes d'apprentissage dans les réseaux profonds. Alors que la majorité des analyses théoriques existantes portent sur des architectures peu profondes [Chizat, 2018], le succès des ResNets a motivé le développement d'une compréhension mathématique des dynamiques d'entraînement à très grande profondeur.

4.2 Équations différentielles ordinaires neuronales

Une question centrale dans l'analyse des ResNets profonds concerne le choix d'un facteur d'échelle approprié pour les branches résiduelles lorsque la profondeur du réseau augmente. Afin de garantir un entraînement stable dans la limite de grande profondeur, on introduit un facteur d'échelle dépendant de la profondeur, noté β_D, ce qui conduit à la formule suivante pour la propagation du signal :

$$\forall d \in \{0, \dots, D-1\}\ , \quad x_{d+1} = x_d + \beta_D F_{\theta_d}(x_d)\,. \tag{15}$$

Le modèle d'équations différentielles ordinaires neuronales (*Neural Ordinary Differential Equations* ou *NODEs*), introduit par Chen et al. [Chen, 2018], correspond au choix particulier $\beta_D = 1/D$, pour lequel Eq. (15) peut être interprétée comme une discrétisation d'Euler explicite d'une équation différentielle ordinaire (ÉDO). Dans la limite où la profondeur tend vers l'infini, le modèle possède un continuum de paramètres $\theta \in \Theta^{[0,1]}$ et traite une donnée d'entrée $x \in \mathcal{X}$ en résolvant l'ÉDO :

$$\forall s \in [0,1]\,, \quad \frac{\mathrm{d}}{\mathrm{d}s} x(s) = F_{\theta(s)}(x(s)), \quad x(0) = x\,, \tag{16}$$

où les champs de vecteurs paramétriques $F_{\theta(s)}$, aussi appelés résidus, sont appris.

Applications À l'origine, l'introduction des NODEs était motivée par la possibilité de calculer les gradients via la méthode de l'adjoint (*adjoint state method*), plus économe en mémoire que l'algorithme classique de rétropropagation, lequel nécessite de stocker l'ensemble des activations et devient très coûteux en mémoire à grande profondeur [Chen, 2018]. De plus, la formulation continue en termes d'ÉDO permet d'utiliser des méthodes d'intégration adaptatives. Par conception, les NODEs modélisent des systèmes dynamiques, ce qui a conduit à leur application dans des domaines tels que la modélisation physique ou les séries temporelles financières [Oh, 2025]. Leur capacité à implémenter des

flots continus de difféomorphismes en fait également un outil naturel pour la conception de modèles génératifs [Rezende, 2015; Kobyzev, 2020]. Plus largement, les NODEs ont inspiré le développement de nouvelles architectures de réseaux de neurones [Sander, 2021] ainsi que de nouveaux algorithmes d'entraînement [Chen, 2018; Vialard, 2020].

Analyse théorique Sur le plan théorique, les NODEs offrent un cadre mathématique commode, celui des ÉDO, pour analyser la dynamique d'entraînement et les performances des réseaux de neurones très profonds. En particulier, leur formulation en temps continu permet d'exploiter les outils de la théorie du contrôle optimal pour étudier les questions d'entraînement et de généralisation [E, 2019; E, 2021]. Cette formulation conduit en outre à un paysage de risque bien conditionné, permettant d'obtenir des garanties de convergence pour les méthodes d'optimisation par gradient [Sander, 2022b; Marion, 2023b]. Dans ce manuscrit, nous exploitons le formalisme des NODEs pour étudier, au Chapter I et au Chapter II, les dynamiques d'entraînement des ResNets à la fois profonds et larges.

5 Apprentissage et algorithmes de descente de gradient

Comme expliqué précédemment, dans le cadre classique de l'apprentissage supervisé, la phase d'entraînement consiste généralement à minimiser une fonction de risque. L'objectif est de trouver une paramétrisation $\theta^* \in \arg\min_{\theta\in\Theta} \mathcal{R}(\theta)$, où $\mathcal{R}$ désigne le risque d'entraînement défini en Eq. (3). Ce problème d'optimisation est en pratique résolu à l'aide de méthodes d'optimisation du premier ordre, dont le cas le plus simple est l'algorithme de "descente de gradient", défini par :

$$\forall k \geq 0\,, \quad \theta_{k+1} = \theta_k - \tau \nabla_\theta \mathcal{R}(\theta_k)\,, \tag{17}$$

où $\theta_0 \in \Theta$ est une initialisation et $\tau > 0$ un "pas d'apprentissage" (*learning rate*). Dans la limite où le pas τ tend vers zéro, cette dynamique discrète peut être modélisée par une dynamique en temps continu. Ce "flot de gradient" s'écrit :

$$\forall t \geq 0\,, \quad \frac{\mathrm{d}}{\mathrm{d}t}\theta_t = -\nabla_\theta \mathcal{R}(\theta_t)\,. \tag{18}$$

Dans ce manuscrit, nous analyserons les propriétés de convergence de ces dynamiques pour l'apprentissage de plusieurs architectures de réseaux de neurones : les réseaux résiduels profonds au Chapter I et au Chapter II, et les modèles à une couche cachée au Chapter III. Une telle analyse, tout en contribuant à la compréhension des performances des modèles modernes d'apprentissage automatique, présente des défis mathématiques importants, le risque $\mathcal{R}$ étant une fonction non-convexe d'un nombre généralement très élevé de paramètres.

Algorithme de rétropropagation En pratique, le gradient du risque par rapport aux paramètres est calculé à l'aide de l'algorithme de "rétropropagation des gradients" (*back-propagation*), correspondant à une différentiation automatique par accumulation inverse (*reverse mode automatic differentiation*) [Baydin, 2018]. En appliquant systématiquement la règle de la chaîne à travers le réseau, la rétropropagation propage les dérivées de la couche de sortie vers les couches internes, avec un coût de calcul comparable à celui de l'évaluation du modèle.

La capacité de passage à l'échelle de cette approche a été un facteur clé des percées récentes de l'apprentissage profond. Les implémentations efficaces de la différenti-

ation automatique par accumulation inverse, combinées aux capacités de calcul parallèle du matériel informatique moderne (GPU, TPU), ont permis l'entraînement de modèles fortement surparamétrés sur de grands ensembles de données. Ces implémentations sont notamment disponibles dans les bibliothèques d'apprentissage profond telles que *PyTorch* [Paszke, 2017], que nous utilisons pour valider nos résultats en Section II.7 et en Section III.6.

5.1 Descente de gradient stochastique et variantes

Bien que nous concentrions notre analyse sur la dynamique de la descente de gradient classique (Eq. (17)), il convient de rappeler qu'en pratique, l'apprentissage sur de grands ensembles de données repose sur une approximation du risque, calculée sur de petits sous-ensembles (*mini-batch*) de données. Cela conduit à l'algorithme de "descente de gradient stochastique" (SGD) :

$$\forall k \geq 0\,, \quad \theta_{k+1} = \theta_k - \tau \nabla_\theta \mathcal{R}_k(\theta_k)\,, \tag{19}$$

où, à l'itération k, le risque approché $\mathcal{R}_k$ s'écrit :

$$\mathcal{R}_k(\theta) = \frac{1}{\#\mathcal{D}_k} \sum_{(x,y)\in\mathcal{D}_k} \ell(F_\theta(x), y)\,,$$

avec $\mathcal{D}_k \subset \mathcal{D}$ un sous-ensemble de données échantillonné à partir du jeu de données $\mathcal{D}$. Outre la réduction du coût de calcul par itération, la SGD introduit une stochasticité dans le processus d'optimisation, qui agit comme un régularisateur implicite et permet souvent d'éviter le surapprentissage, conduisant ainsi à une meilleure généralisation [Hardt, 2016b].

En complément, un terme de "moment" est souvent ajouté, ce qui conduit à la formule suivante :

$$\forall k \geq 0\,, \quad \left\{ \begin{array}{rcl} b_{k+1} & = & \mathrm{m} b_k + (1-\mathrm{m}) \nabla_\theta \mathcal{R}_k(\theta_k)\,, \\ \theta_{k+1} & = & \theta_k - \tau b_{k+1}\,, \end{array} \right. \tag{20}$$

où $\mathrm{m} \in [0, 1]$ est le paramètre de moment. Introduites initialement par Polyak [Polyak, 1964], les méthodes à moments sont connues pour accélérer la convergence de la descente de gradient dans le cas d'objectifs lisses et fortement convexes. Des raffinements ultérieurs, tels que la méthode du gradient accéléré de Nesterov [Nesterov, 1983], atteignent des vitesses de convergence optimales dans le cas convexe. En apprentissage profond, l'ajout d'un terme de moment améliore également la stabilité de l'entraînement [Sutskever, 2013].

Enfin, de nombreuses autres techniques ont été développées pour faciliter l'apprentissage à grande échelle, notamment le *dropout*, le *weight decay*, ainsi que des méthodes d'optimisation adaptatives telles que *Adam* [Kingma, 2014] ou *RMSprop* [Hinton, 2012]. Ces méthodes jouent un rôle crucial dans l'amélioration de la stabilité, de la vitesse de convergence et des performances de généralisation [Bottou, 2018].

5.2 Apprentissage à deux échelles de temps et projection de la variable

Le choix des hyperparamètres, en particulier du pas d'apprentissage τ, joue un rôle essentiel dans le comportement asymptotique de la dynamique d'apprentissage. En pratique, les pas d'apprentissage peuvent différer selon les paramètres [Yang, 2021]. Dans le Chapter III, nous distinguerons deux types de paramètres :

- **Paramètres linéaires :** il s'agit généralement des poids de la dernière couche du réseau, pour lesquels la sortie dépend linéairement des paramètres. Lorsque les autres paramètres sont fixés, l'apprentissage de ces paramètres revient à la résolution d'un problème d'optimisation convexe.
- **Paramètres non-linéaires :** ils correspondent aux paramètres des couches internes du réseau, liés de manière non-linéaire à la sortie. Ils permettent l'extraction de représentations non-linéaires des données et jouent un rôle central dans la capacité de généralisation des réseaux de neurones. Leur apprentissage constitue toutefois un problème d'optimisation non-convexe.

L'espace des paramètres se décompose ainsi en $\Theta = \Theta^l \times \Theta^{nl}$, où Θ^l et Θ^{nl} désignent respectivement les sous-espaces des paramètres linéaires et non-linéaires. Attribuer des taux d'apprentissage distincts à ces deux sous-ensembles conduit à une descente de gradient à deux échelles de temps :

$$\forall k \geq 0, \quad \left\{ \begin{array}{rcl} \theta^l_{k+1} & = & \theta^l_k - \eta\tau \nabla_{\theta^l} \mathcal{R}(\theta^l_k, \theta^{nl}_k)\,, \\ \theta^{nl}_{k+1} & = & \theta^{nl}_k - \tau \nabla_{\theta^{nl}} \mathcal{R}(\theta^l_k, \theta^{nl}_k)\,, \end{array} \right. \tag{21}$$

où $\tau > 0$ est un pas d'apprentissage et $\eta > 0$ un hyperparamètre contrôlant la vitesse relative des mises à jour. Lorsque $\eta < 1$, les paramètres linéaires θ^l sont appris plus lentement que les paramètres non-linéaires θ^{nl}, et inversement lorsque $\eta > 1$.

La limite asymptotique de grandes échelles de temps correspond à une optimisation partielle des paramètres linéaires : c'est l'algorithme de "projection de la variable" (*Variable Projection* ou *VarPro*), introduit initialement par Golub and Pereyra [Golub, 1973] pour la minimisation de problèmes non-linéaires séparables. En effet, lorsque $\eta \to +\infty$, à chaque étape on a $\theta^l_k \in \arg\min_{\theta^l \in \Theta^l} \mathcal{R}(\theta^l, \theta^{nl}_k)$. D'après le théorème de l'enveloppe, la dynamique sur les paramètres non-linéaires s'écrit alors :

$$\forall k \geq 0\,, \quad \theta^{nl}_{k+1} = \theta^{nl}_k - \tau \nabla_{\theta^{nl}} \mathcal{R}(\theta^l_k, \theta^{nl}_k) = \theta^{nl}_k - \tau \nabla_{\theta^{nl}} \mathcal{L}(\theta^{nl}_k)\,, \tag{22}$$

où, pour tout $\theta^{nl} \in \Theta^{nl}$, le "risque réduit" $\mathcal{L}(\theta^{nl})$ est défini par :

$$\mathcal{L}(\theta^{nl}) := \inf_{\theta^l \in \Theta^l} \mathcal{R}(\theta^l, \theta^{nl})\,. \tag{23}$$

En pratique, dans le cas de la régression avec fonction de perte quadratique, cette étape d'optimisation partielle peut être effectuée numériquement en résolvant un système linéaire.

Ainsi, en séparant l'apprentissage des représentations (paramètres non-linéaires, lents) de celui de l'ajustement prédictif (paramètres linéaires, rapides), l'apprentissage à deux échelles de temps et la projection de variables fournissent un cadre conceptuel solide pour comprendre l'apprentissage des représentations dans les réseaux de neurones. De telles approches ont récemment suscité un intérêt croissant dans la communauté de la théorie de l'apprentissage automatique [Marion, 2023a; Berthier, 2024; Bietti, 2023; Takakura, 2024]. Nous étudierons dans le Chapter III les propriétés de convergence de VarPro pour l'entraînement de modèles champ-moyen de réseaux de neurones.

5.3 Flots de gradient de Wasserstein et transport optimal

Nous nous intéressons dans ce manuscrit à l'apprentissage d'architectures de réseaux de neurones surparamétrés, décrits par une distribution de paramètres sur un espace de paramètres Θ (Eq. (13)). Pour de tels modèles champ-moyen, le risque d'entraînement

défini en Eq. (3) devient une fonctionnelle $\mathcal{R} : \mathcal{P}(\Theta) \to \mathbb{R}$ définie sur l'espace $\mathcal{P}(\Theta)$ des mesures de probabilité sur Θ. En particulier, pour un réseau de largeur finie M, la distribution des paramètres est donnée par la mesure empirique $\mu = \frac{1}{M}\sum_{i=1}^{M} \delta_{\theta^i}$. Lorsque les paramètres $(\theta_t^i)_{1\leq i\leq M}$ suivent la dynamique de flot de gradient Eq. (18), la distribution associée $\mu_t = \frac{1}{M}\sum i = 1^M \delta_{\theta_t^i}$ évolue selon l'équation de continuité :

$$\partial_t \mu_t - \operatorname{div}\left(\mu_t \nabla \frac{\delta \mathcal{R}}{\delta \mu}[\mu_t]\right) = 0\,, \quad \text{sur } [0,+\infty)\times\Theta. \tag{24}$$

Ici, pour une mesure $\mu \in \mathcal{P}(\Theta)$, le champ de potentiel $\frac{\delta \mathcal{R}}{\delta \mu}[\mu]$ désigne la première variation (ou différentielle de Fréchet) de $\mathcal{R}$. Dans le cas général où μ_t n'est pas nécessairement une mesure empirique, cette ÉDP peut être interprétée comme un flot de gradient métrique pour la distance de Wasserstein sur $\mathcal{P}(\Theta)$ [Ambrosio, 2008b; Santambrogio, 2015].

La distance de Wasserstein découle du problème de transport optimal entre mesures de probabilité [Villani, 2009; Santambrogio, 2015]. En supposant que Θ est un espace de Hilbert, la p-distance de Wasserstein $\mathcal{W}_p$, pour $p \geq 1$, est définie entre deux mesures boréliennes $\mu, \mu' \in \mathcal{P}(\Theta)$ par :

$$\mathcal{W}_p(\mu,\mu') \coloneqq \left(\inf_{\gamma\in\Gamma(\mu,\mu')} \int_{\Theta\times\Theta} |\theta-\theta'|^p \,\mathrm{d}\gamma(\theta,\theta')\right)^{1/p}, \tag{25}$$

où $\Gamma(\mu,\mu')$ désigne l'ensemble des couplages entre μ et μ', c'est-à-dire l'ensemble des mesures de probabilité sur $\Theta\times\Theta$ dont les marginales sont respectivement μ et μ' :

$$\Gamma(\mu,\mu') \coloneqq \left\{\gamma \in \mathcal{P}(\Theta\times\Theta) \;:\; \pi^1_{\#}\gamma = \mu\,,\, \pi^2_{\#}\gamma = \mu'\right\}. \tag{26}$$

Ainsi, $\mathcal{W}_p$ muni $\mathcal{P}_p(\Theta)$, l'espace des mesures de probabilité à p-moment fini, d'une structure d'espace métrique complet et séparable. Il est notamment connu depuis les travaux de Jordan, Kinderlehrer, and Otto [Jordan, 1998] que plusieurs ÉDP linéaires ou non-linéaires, telles que les équations de Fokker–Planck ou les équations des milieu poreux, peuvent être interprétées comme des flots de gradient pour cette métrique. De même que le flot de gradient Eq. (18) s'obtient comme la limite $\tau \to 0^+$ de la descente de gradient discrète Eq. (17), le flot de gradient de Wasserstein Eq. (24) peut être approché par un "schéma JKO" correspondant à une discrétisation implicite :

$$\forall k \geq 0\,, \quad \mu_{k+1} \in \operatorname*{arg\,min}_{\mu\in\mathcal{P}(\Theta)} \mathcal{R}(\mu) + \frac{1}{2\tau}\mathcal{W}_2(\mu,\mu_k)^2\,.$$

En apprentissage automatique, des ÉDP de la forme Eq. (24) ont été utilisées dans de nombreux travaux pour étudier la dynamique d'entraînement des réseaux de neurones peu profonds [Chizat, 2018; Rotskoff, 2019; Mei, 2019; Chizat, 2022; Nitanda, 2022] ou profonds [Lu, 2020; Ding, 2021; Isobe, 2023]. Outre l'élégance de ce formalisme pour décrire l'apprentissage dans un régime de grande largeur, la relaxation du risque dans l'espace des mesures permet également de simplifier le paysage d'optimisation, notamment en éliminant les points critiques qui ne sont pas des minimiseurs.

Introduction

Contents

1 Supervised learning: algorithms, architectures and mathematical models **17**
1.1 The supervised learning framework 18
1.2 Neural network architectures 20
1.3 Scaling neural networks in the infinite width regime 22
1.4 Scaling neural networks in the infinite depth regime 25
1.5 Training and gradient descent algorithms 27
2 Contributions **31**
I Training of infinitely deep and wide residual architectures 31
II Convergence in the training of residual architectures 34
III Feature learning in shallow architectures 36

1 Supervised learning: algorithms, architectures and mathematical models

In recent years, deep learning has achieved remarkable empirical successes across a wide range of applications — from image and text generation to scientific computing, and more recently in reasoning tasks such as the resolution of complex mathematical problems. However, while a rapidly growing body of research has emerged to interpret the behavior of AI systems and guide their design, these successes often outpace our understanding of the underlying mathematical mechanisms.

From a mathematical perspective, the training of neural networks presents a number of challenging questions. On the one hand, the optimization problems involved are typically high-dimensional and non-convex, yet simple algorithms like stochastic gradient descent often perform surprisingly well in practice. On the other hand, neural networks are capable of interpolating large datasets while still generalizing effectively, seemingly defying foundational statistical intuitions such as the bias–variance trade-off or the curse of dimensionality. These phenomena point to a need for new mathematical frameworks capable of capturing the dynamical behavior of neural network training and its interaction with model architecture and data structure. In particular, a recent line of work suggests that tools from the analysis of partial differential equations and optimal transport can offer valuable insights into these dynamics.

In this manuscript, we adopt a mathematical perspective on neural network training based on tools from optimization and the theory of partial differential equations. In the limit of large width, the training dynamic of neural networks can be described as mean-field models of interacting particle systems and corresponds to gradient flows in spaces of probability measures. On the other hand, residual architectures are studied for their

infinite-depth limit, which gives rise to smoother optimization landscapes and more stable training dynamics. These two asymptotic regimes — large width and large depth — are not merely theoretical constructs, but reflect the structure of modern architectures such as ResNets or Transformers, which are highly overparameterized and central to current state-of-the-art models.

1.1 The supervised learning framework

We will consider throughout this manuscript a *supervised learning* framework encompassing a large number of classical machine learning tasks. We start by describing the main ingredients of this framework, before turning in more details to the description of the models and algorithms.

Dataset In supervised learning the machine is provided with a dataset $\mathcal{D} \subset \mathcal{X} \times \mathcal{Y}_{targ}$ constituted of pairs of input data $x \in \mathcal{X}$ and associated target response $y \in \mathcal{Y}_{targ}$. These input and target data can have various forms:

- **Inputs:** Due to the versatile nature of machine learning methods, these can be virtually anything ranging from images, sounds, videos to text or financial time series. An example of application we will consider in Chapters II and III is image classification, where the input data are numerical images encoded as 8-bit arrays of shape $n_c \times n_w \times n_h$ where n_w and n_h are respectively the number of pixels in the width and height of the image and n_c is the number of channels used to encode the image, usually $n_c = 1$ for gray-scale images and $n_c = 3$ for color images. Mathematically, these images can then be modeled by vectors in $\mathcal{X} = \mathbb{R}^{n_c \times n_w \times n_h}$.

- **Targets:** One can generally distinguish between two categories of supervised learning tasks that are *classification* and *regression*. In classification, the objective is to classify data into a finite set of classes which are usually represented by *labels* in $\mathcal{Y}_{targ} = \{1, ..., C\}$, with $C \geq 1$ the number of classes. These labels can also be encoded as *one-hot* vectors in $\mathcal{Y}_{targ} = \{0, 1\}^C$. In contrast, in regression, the objective is usually to predict a vector-valued signal in $\mathcal{Y}_{targ} = \mathbb{R}^{d_{out}}$.

In the following, the spaces of inputs $\mathcal{X}$ and targets $\mathcal{Y}_{targ}$ will always be subsets of real finite dimensional vector spaces. It is then standard to see $(x, y) \in \mathcal{X} \times \mathcal{Y}_{targ}$ as random variables whose distribution we will also denote by $\mathcal{D}$.

Loss The objective for the machine is to learn from the examples in $\mathcal{D}$ a *prediction function* or *predictor* $F : \mathcal{X} \to \mathcal{Y}_{out}$, giving for inputs $x \in \mathcal{X}$ predictions of the target response $y_{targ} \in \mathcal{Y}_{targ}$. The space of outputs $\mathcal{Y}_{out}$ is a vector space which is not necessarily the same as the space of targets $\mathcal{Y}_{targ}$ and, to evaluate the quality of its predictions, the machine is provided with a *loss function* $\ell : \mathcal{Y}_{out} \times \mathcal{Y}_{targ} \to \mathbb{R}$. We will consider two fundamental examples:

- **Regression:** Usually in regression, the space of outputs and response is the same vector space $\mathcal{Y}_{out} = \mathcal{Y}_{targ} = \mathbb{R}^{d_{out}}$. This space is equipped with the standard Euclidean geometry and a natural notion of error between a prediction y_{out} and a target signal y_{targ} is the *square loss*:

$$\ell(y_{out}, y_{targ}) = \frac{1}{2} \left\| y_{out} - y_{targ} \right\|^2_{\mathbb{R}^{d_{out}}} . \tag{27}$$

- **Classification:** In a classification problem with C classes, the machine usually outputs predictions in $\mathcal{Y}_{out} = \mathbb{R}^C$ representing estimates of the posterior log-probabilities of each classes given the input. A prediction $y_{out} \in \mathcal{Y}_{out}$ is then compared to a target label $y_{targ} \in \mathcal{Y}_{targ} = \{1, ..., C\}$ by forming the *cross-entropy*:

$$\ell(y_{out}, y_{targ}) = -\log\left(\frac{\exp(y_{out}[y_{targ}])}{\sum_{i=1}^{C}\exp(y_{out}[i])}\right). \tag{28}$$

Risk minimization Given a dataset $\mathcal{D} \subset \mathcal{X}\times\mathcal{Y}_{targ}$ and a loss function $\ell : \mathcal{Y}_{out} \times \mathcal{Y}_{targ} \to \mathbb{R}$, the quality of a prediction function $F : \mathcal{X} \to \mathcal{Y}_{out}$ can be assessed by averaging the loss incurred over the dataset. The strategy developed in machine learning is to search for the best predictor in a class of parametric function $\mathcal{F} = \{F_\theta \,:\, \theta \in \Theta\}$, where Θ denotes a parameter space. For instance, in the case of neural networks, Θ corresponds to the space of the network's weights, which is usually a high-dimensional vector space equipped with the Euclidean metric. For each parameter $\theta \in \Theta$, the *training risk* is then defined by:

$$\mathcal{R}(\theta) \coloneqq \frac{1}{\#\mathcal{D}}\sum_{(x,y)\in\mathcal{D}} \ell(F_\theta(x), y)\,. \tag{29}$$

Training the parametric model F_θ then consists in solving the *risk minimization problem*:

$$\text{Find} \quad \theta^* \in \operatorname*{arg\,min}_{\theta\in\Theta} \mathcal{R}(\theta)\,. \tag{30}$$

In practice, this optimization is often performed using first-order iterative algorithms, with *gradient descent* being a canonical example. Starting from an initial parameter $\theta_0 \in \Theta$, the parameters are updated according to:

$$\forall k \geq 0\,, \quad \theta_{k+1} = \theta_k - \tau\nabla_\theta\mathcal{R}(\theta)\,,$$

where $\tau > 0$ denotes the *step size* or *learning rate*. In deep learning, to allow for training on large datasets and improve generalization, the risk is usually computed at each step $k \geq 0$ on a smaller batch $\mathcal{D}_k \subset \mathcal{D}$ of freshly sampled data. This results in the *stochastic gradient descent* algorithm:

$$\forall k \geq 0\,, \quad \theta_{k+1} = \theta_k - \tau\nabla_\theta\mathcal{R}_k(\theta)\,, \quad \text{where} \quad \mathcal{R}_k(\theta) \coloneqq \frac{1}{\#\mathcal{D}_k}\sum_{(x,y)\in\mathcal{D}_k} \ell(F_\theta(x), y)\,.$$

For both gradient descent and stochastic gradient descent, the choice of parametric model, as well as the selection of hyperparameters (such as τ or the batch size) has a significant impact on both the training dynamics and the generalization performance of the learned model. In the remainder of this introduction, we describe in more detail the neural network architectures and training procedures that will be the focus of this thesis.

Statistical learning perspective While this manuscript adopts an optimization-oriented viewpoint — focusing on the minimization of the training risk — it is important to recall that the ultimate goal of supervised learning is to construct a prediction function that performs well on unseen examples. In a standard statistical learning paradigm, data points (x, y) in the training dataset $\mathcal{D}$ are assumed to be independent and identically distributed according to an unknown distribution $\mathcal{D}_{test}$ over $\mathcal{X} \times \mathcal{Y}_{targ}$. The central object of interest is then the *test error*, defined as

$$\mathcal{E}_{test}(\theta) \coloneqq \mathbb{E}_{(x,y)\sim\mathcal{D}_{test}}\left[\ell_{test}(F_\theta(x), y)\right],$$

where the *test loss* ℓ_{test} can also differ from the training loss ℓ. Minimizing the training risk $\mathcal{R}$ thus serves as a proxy for minimizing $\mathcal{E}_{test}$, the fundamental challenge lying in the fact that $\mathcal{D}_{test}$ is unknown, and learning must proceed solely from the finite number of examples in $\mathcal{D}$. While the question of the generalization abilities of the trained models lie beyond the primary scope of this thesis, it motivates and justifies many of the modeling and algorithmic choices made in the following chapters.

Self-supervised learning Finally, while this manuscript focuses on supervised learning tasks, it is worth noting that many modern machine learning systems are trained under the *self-supervised learning* paradigm, where the target signal is derived from the input data itself. This approach can be seen as a special case of supervised learning in which the targets are constructed from unlabeled data using pretext tasks. Prominent examples include next-token prediction in language modeling or score-based training in generative modeling. These methods have proven effective in leveraging large amounts of unlabeled data to pretrain models for downstream tasks.

1.2 Neural network architectures

The family of parametric models we will consider in this manuscript is the one of *neural networks*. These consist in the successive composition of *layers* which are themselves smaller parametric transformations. A neural network of depth $D \geq 1$, is thus a model parameterized by $\theta \in \Theta = \prod_{d=1}^{D} \Theta_d$ which on input $x \in \mathcal{X}$ returns:

$$F_\theta(x) = F_{\theta_D} \circ \cdots \circ F_{\theta_1}(x)$$

where, for each $d \in \{1, ..., D\}$, the d-th layer F_{θ_d} is a (smaller) neural network parameterized by $\theta_d \in \Theta_d$. Considering $\mathcal{X} = \mathbb{R}^{d_{in}}$ and $\mathcal{Y}_{out} = \mathbb{R}^{d_{out}}$ for some $d_{in}, d_{out} \geq 1$, we start here by describing examples and properties of common *shallow* architectures $F_\theta : \mathbb{R}^{d_{in}} \to \mathbb{R}^{d_{out}}$. These constitute the building blocks of deeper architectures we will describe later-on.

- **Linear layers:** Linear *fully-connected* layers compute matrix-vector multiplications. Given an input $x \in \mathbb{R}^{d_{in}}$ the output is:

  $$F_W(x) = W \cdot x \,,$$

  where the parameter $W \in \mathbb{R}^{d_{out} \times d_{in}}$ is some weight matrix. Such linear transformations are the basic building blocks of most neural network architectures. In practice, modern deep learning models are typically constructed by composing these linear maps with simple nonlinear functions.

- **Convolutional layers:** Convolutional layers are a particular instance of linear layers where, in contrast with fully-connected layers, the weight matrices are constrained to have particular shape, namely to be convolution matrices. This kind of architecture was introduced by LeCun et al. [LeCun, 1989] for digit recognition and, owing to their translation-equivariant structure, has since become ubiquitous in image processing applications [LeCun, 2015]. A convolutional layer is parameterized by filters W and, for an input image x, computes:

  $$F_W(x) = W \star x \,, \tag{31}$$

where $\star$ denotes the discrete convolution operator. For example, if $x \in \mathbb{R}^{c_{in} \times d_w \times d_h}$ is an image with c_{in} input channels and $W \in \mathbb{R}^{c_{out} \times c_{in} \times k \times k}$ is a convolutional filter of size $k \times k$ with c_{out} output channels the result of the discrete convolution reads:

$$(W \star x)[c, i, j] = \sum_{1 \leq k_1, k_2 \leq k} \sum_{1 \leq c' \leq c} W[c, c', k_1, k_2]\, x[c', i + k_1, j + k_2]\,. \tag{32}$$

We will consider convolutional neural networks in Chapters II and III for solving image classification problems.

- **Linear models in parameter space:** An important class of machine learning models is the one of models that are linear in their parameters but not necessarily in their inputs. This is for example the case of kernel methods [Schölkopf, 2002; Steinwart, 2008] or of random feature models [Rahimi, 2007]. These models have a parameter space $\Theta = \mathcal{H}^{d_{out}}$ where $\mathcal{H}$ is some Hilbert space of features, and compute for a parameter $\theta \in \Theta$ and an input $x \in \mathbb{R}^{d_{in}}$:

$$F_\theta(x) = \begin{pmatrix} \langle \theta_1, \phi(x) \rangle_{\mathcal{H}} \\ \vdots \\ \langle \theta_{d_{out}}, \phi(x) \rangle_{\mathcal{H}} \end{pmatrix}, \tag{33}$$

where $\phi : \mathcal{X} \to \mathcal{H}$ is a map associating to each input a *feature representation* in $\mathcal{H}$. While standard neural networks are nonlinear in both their inputs and their parameters, such models offer the advantage of being linear in parameter space, which facilitates theoretical analysis. We will study this class of models in Section II.4, as a preliminary step towards understanding more complex architectures.

- **Perceptron layers:** The *perceptron* is arguably one the simplest instance of a neural network architecture that is nonlinear in both its input and its parameters. It was originally introduced by Rosenblatt [Rosenblatt, 1958] to emulate human visual and perceptual capacities and can actually be seen as the composition of two fully-connected layers with a nonlinear function. A 2-layer or *single-hidden-layer (SHL)* perceptron of width $M \geq 1$ is parameterized by two weight matrices U, W of respective shape $d_{out} \times M$ and $d_{in} \times M$ and a bias term $b \in \mathbb{R}^M$. For an input $x \in \mathbb{R}^{d_{in}}$ it computes:

$$F_{(U,W,b)}(x) = U\sigma(W^\top x + b)\,, \tag{34}$$

where the $\sigma : \mathbb{R} \to \mathbb{R}$ is a nonlinear function, called *activation*, applied component wise. Popular examples of activations are for example the hyperbolic tangent tanh or the *Rectified Linear Unit (ReLU)* activation. We will study this class of models in Section II.5 and in Chapter III.

- **Attention layers:** Attention mechanisms [Bahdanau, 2014; Vaswani, 2017] is at the heart of Transformers architectures which have emerged as state of the art models in computer vision [Dosovitskiy, 2020], *Natural Language Processing (NLP)* [Devlin, 2019] as well as other sequence processing or generation tasks. An attention head is parameterized by matrices $Q, K, V \in \mathbb{R}^{d_{in} \times d_{in}}$ and, for an input sequence of *tokens* $\mathbf{x} = (x_i)_{1 \leq i \leq N} \in (\mathbb{R}^{d_{in}})^N$ of length N, returns:

$$\text{Attention}_{Q,K,V}(\mathbf{x}) = \left(\sum_{j=1}^{N} \frac{e^{\langle Qx_i, Kx_j \rangle}}{\sum_{j=1}^{N} e^{\langle Qx_i, Kx_j \rangle}} Vx_j \right)_{1 \leq i \leq N} \in (\mathbb{R}^{d_{in}})^N\,. \tag{35}$$

In the context of NLP, these tokens represent embeddings of words or syllables on which models are trained in a self-supervised manner to perform a *next-token prediction* task. In modern large-scale language models, such as *Generative Pretrained Transformers (GPTs)* [Radford, 2018], multi-layer perceptrons are actually stacked with *multi-head attention layers* where several attention operations are computed in parallel.

- **Parameter-free layers:** In modern neural network architectures, parametric layers are composed with several other parameter-free operations engineered to enhance the expressivity and trainability of the models. Composition with a non-linear activation can for example be seen as a simple form of parameter-free layer. Also, while for the sake of simplicity we omit them in the rest of this manuscript, common parameter-free layers usually include *pooling layers*, which reduce the spatial or temporal dimensions of feature maps, or *normalization layers*, which have been shown to facilitate the training of deep neural networks [Ioffe, 2015; Ba, 2016]. In the context of NLP, parameter-free operations include *positional encodings*, which inject order information into sequence representations, and *attention masking* mechanisms, which constrain the flow of information (e.g., to preserve causality in autoregressive models) [Vaswani, 2017].

1.3 Scaling neural networks in the infinite width regime

The past decade has witnessed an exponential increase in the scale of neural network architectures, with modern models comprising billions, and in some cases even trillions, of parameters [Villalobos, 2022]. However, a striking and somewhat counterintuitive phenomenon has emerged: many of these models operate in an overparameterized regime, where the number of trainable parameters exceeds the number of available data points. In classical statistics, such settings would typically lead to overfitting and poor generalization. Yet, in practice, overparameterized neural networks often generalize remarkably well [Belkin, 2019; Zhang, 2021]. Significant theoretical efforts have thus been devoted to understanding neural networks in the infinite-width regime — that is, when the number of neurons (or channels) per layer tends to infinity. Beyond their theoretical value, these asymptotic analyses also offer practical benefits, particularly in guiding hyperparameter selection and enabling hyperparameter transfer across architectures of different widths [Yang, 2021; Bordelon, 2025] leading to important computational savings in the training of large models [OpenAI, 2023].

Many of the above presented neural network architectures can be represented as mappings of the form

$$F_{(\theta_i)_{1\leq i\leq M}} : x \in \mathbb{R}^{d_{in}} \mapsto \alpha_M \sum_{i=1}^{M} \psi(\theta_i, x) \in \mathbb{R}^{d_{out}}\,, \tag{36}$$

where Θ denotes a parameter space, $\psi : \Theta \times \mathbb{R}^{d_{in}} \to \mathbb{R}^{d_{out}}$ is a basis function, and $\alpha_M \in \mathbb{R}$ is a scaling factor that depends on the network width M. For example, the 2-layer perceptron model corresponds to the case where $\Theta = \mathbb{R}^{d_{out}} \times \mathbb{R}^{d_{in}} \times \mathbb{R}$ and ψ is given by:

$$\psi : ((u, w, b), x) \in \Theta \times \mathbb{R}^{d_{in}} \mapsto u\sigma(w^\top x + b)\,, \tag{37}$$

where $\sigma : \mathbb{R} \to \mathbb{R}$ is the activation function.

Parameters of the model are usually initialized randomly of order 1 and recent research has then highlighted the crucial role played by the choice of scaling α_M in shaping

the training dynamics for models of large width M. Different scalings lead to different asymptotic behaviors when M tends to infinity. Two main theoretical frameworks have emerged: the mean-field regime, which captures nonlinear feature learning, and the neural tangent kernel regime, which describes a linearized training dynamic around the random initialization.

1.3.1 Neural Tangent Kernel regime

A first asymptotic framework for analyzing neural networks in the infinite-width limit is the Neural Tangent Kernel (NTK) regime [Jacot, 2018]. This regime corresponds to a scaling of the network parameters of order $\alpha_M = 1/\sqrt{M}$ for a network of width M, under which the evolution of the network during gradient descent can be closely approximated by a linearization around its random initialization $\theta_0 \in \Theta$. In this linearized regime, the model is linear in its parameters, as in Eq. (33). The neural network thus reduces to a kernel method [Schölkopf, 2002; Steinwart, 2008] whose associated kernel:

$$K(x, x') = \mathrm{D}_\theta F_{\theta_0}(x) \cdot \mathrm{D}_\theta F_{\theta_0}(x')^\top , \tag{38}$$

called NTK, becomes deterministic in the infinite-width limit. This leads to powerful theoretical results: it can be shown that gradient descent converges to a global minimizer of the empirical risk at a linear rate, governed by the spectral properties of the NTK [Allen-Zhu, 2019; Du, 2019; Lee, 2019; Zou, 2020]. We will study in more detail conditioning of the NTK associated to 2-layer perceptrons in Section II.5.

However, the NTK regime comes with intrinsic limitations. Most notably, it induces a form of "lazy training" [Chizat, 2019], in which the network parameters barely move from their initialization, and the feature representations do not evolve significantly over the course of training. As a result, the model fails to capture nonlinear data-dependent features in a meaningful way, instead behaving like a kernel method. In contrast, neural networks benefit from hierarchical or task-specific feature learning behaviors, leading to improved generalization [Bach, 2017a; Ghorbani, 2019; Ghorbani, 2020].

1.3.2 Mean-field models of neural networks

An alternative asymptotic framework is the mean-field regime, corresponding to a scaling of the output of order $1/M$ for width M. One of the key features of this regime — in contrast with the NTK setting — is its ability to capture nonlinear feature learning [Yang, 2021].

Under the $\alpha_M = 1/M$ scaling, the network can be interpreted as the integration over a distribution of parameter. Indeed, for a family of parameters $(\theta_i)_{1\le i\le M} \in \Theta^M$, considering the empirical measure $\hat{\mu} = \frac{1}{M}\sum_{i=1}^M \delta_{\theta_i}$, Eq. (36) can be written:

$$\forall x \in \mathbb{R}^{d_{in}}, \quad F_{(\theta_i)_{1\le i\le M}}(x) = \frac{1}{M}\sum_{i=1}^{M} \psi(\theta_i, x) = \int_\Theta \psi(\theta, x)\mathrm{d}\hat{\mu}(\theta) = F_{\hat{\mu}}(x) ,$$

where for every probability measure μ on the parameter space Θ we define:

$$F_\mu : x \in \mathbb{R}^{d_{in}} \mapsto \int_\Theta \psi(\theta, x)\mathrm{d}\mu(\theta) \ \in \mathbb{R}^{d_{out}} . \tag{39}$$

This representation thus encompasses neural networks of arbitrary finite width when μ is an empirical measure but also describes, when the width M tends to infinity, a mean-field

limit in which μ can be any probability measure [Rotskoff, 2018; Chizat, 2018; Mei, 2019; Sirignano, 2020].

From a mathematical perspective, beyond eliminating the dependence on the width M, the mean-field limit representation in Eq. (39) conveniently captures the interchangeability of neurons. Indeed, the invariance under permutations of the index $i \in \{1, ..., M\}$ in Eq. (36) induces symmetries in the risk landscape, which can complicate its analysis. It also enables relaxation of the risk minimization problem Eq. (30) in the space of measures, which has proven to lead to a simplified optimization landscape [Chizat, 2018; Rotskoff, 2019].

Finally, the mean-field representation allows one to study the training dynamics through the lens of gradient flows in the space $\mathcal{P}(\Theta)$ of probability measures over Θ [Ambrosio, 2008b; Santambrogio, 2017]. This results in non-local evolution PDEs whose convergence can be analyzed qualitatively [Chizat, 2018; Rotskoff, 2019] or quantitatively at the condition that the risk satisfies appropriate functional inequalities [Mei, 2019; Chizat, 2022; Nitanda, 2022]. However, while the mean-field scaling enables a more faithful approximation of realistic and desirable training behaviors, existing convergence results are predominantly qualitative: they do not provide explicit convergence rates, nor do they fully characterize the generalization performance of the learned models. Addressing this gap is an active area of research, and constitutes the focus of our contributions in Chapter III.

Remark 1.1. *Note that, while the mean-field representation Eq. (39) has encountered significant interest, it is not the only way to represent the infinite width limit of neural networks with the $\alpha_M = 1/M$ scaling. Several other representations have for example been proposed by E and Wojtowytsch [E, 2022], among which representations with signed measures (also proposed in [Bach, 2017a]) or with indexed particle systems.*

1.3.3 Expressivity and functional properties of neural networks

Though defined by simple compositional structures, the success and the versatility of neural networks owes to powerful expressivity properties. Functional properties of the set of maps that can be represented by a neural network also play an important role for the training and generalization performances. Such properties are determined by the architecture as well as the metric structure of the set of parameters and will be at the core of our analysis in Chapter II.

An important example is the class of 2-layer perceptrons of arbitrary width with non-linear activation functions. A seminal result by Cybenko [Cybenko, 1989] established that such networks are dense in the space of continuous functions with respect to the compact-open topology. Later, Barron [Barron, 1993] provided quantitative approximation bounds in the L^2 norm, showing that a large class of functions can be approximated at a rate of $\mathcal{O}(1/\sqrt{M})$, where M denotes the width of the hidden layer. Remarkably, this rate does not depend on the input dimension, suggesting that neural networks can, in principle, overcome the *curse of dimensionality*. However, these results are non-constructive: they state the existence of accurate approximations without providing a practical method to find them with computational guarantees. This limitation highlights the central role of training algorithms in practice, as one must rely on optimization procedures to discover good approximations.

In the case of the ReLU activation, the space of function represented by 2-layer perceptrons is described by the *Barron space* [E, 2021; E, 2022]:

$$\mathcal{B} := \left\{ F : x \in \mathbb{R}^{d_{in}} \mapsto \int u \,\mathrm{ReLU}(w^\top x + b) \mathrm{d}\mu(u, w, b)\,,\ \mu \in \mathcal{P}(\mathbb{R} \times \mathbb{R}^{d_{in}} \times \mathbb{R}) \right\}.$$

When the set of weights is equipped with the standard Euclidean metric, it is naturally provided with a Banach space norm:

$$\forall f \in \mathcal{B}\,, \quad \|F\|_{\mathcal{B}} \coloneqq \inf\left\{ \int |u|(\|w\| + |b|)\mathrm{d}\mu\,,\ \mu \in \mathcal{P}_2(\mathbb{R} \times \mathbb{R}^{d_{in}} \times \mathbb{R})\,,\ f = F_\mu \right\}.$$

This space can be characterized as the smallest Banach space of function efficiently approximable by 2-layer perceptrons and for example contains all Sobolev spaces of sufficient regularity [E, 2022].

Note that similar results hold for shallow convolutional architectures which are dense in the class of translation equivariant functions [Petersen, 2020; Yarotsky, 2022]. Also, attention-based architectures — and particularly Transformer models — are dense in the space of permutation equivariant sequence-to-sequence functions [Yun, 2020].

1.4 Scaling neural networks in the infinite depth regime

Despite the already high expressivity of shallow architectures, recent breakthroughs in machine learning have relied on the power of function composition. In numerous classical supervised learning tasks, state-of-the-art models are now based on *deep neural networks*, which take the general form:

$$F_\theta(x) = F_{\theta_D} \circ \cdots \circ F_{\theta_1}(x)\,,$$

where each F_{θ_d} denotes a simpler neural network (e.g., a perceptron, convolutional layer, attention mechanism, normalization layer, etc.), and the depth D is typically very large. While this increased depth greatly enhances the expressivity of the model class [Montufar, 2014], it also introduces significant optimization challenges. In particular, it has been observed that the training error of deep convolutional networks can degrade as depth increases beyond a certain point [Srivastava, 2015; He, 2016a]. Moreover, training deep networks often suffers from numerical instabilities such as the *vanishing* and *exploding gradient* problems, where gradients become too small or too large in early layers, impairing effective learning [Bengio, 1994; Glorot, 2010]. These difficulties have motivated the development of specialized architectures that ease the training of very deep networks. A particularly successful design is the class of *residual neural networks*.

1.4.1 Residual Neural Networks

Residual Neural Networks (ResNets) is a class of neural network architectures introduced by He et al. [He, 2016a; He, 2016b] for application in image classification. The idea behind ResNets is to parameterize each layer as a small perturbation, called *residual*, of the identity mapping. In practice, this idea materializes by the presence of *skip connections* whose function is to reinject the signal in-between successive layers. A ResNet of depth $D \geq 1$, with input $x \in \mathcal{X}$, outputs x_D where the data is processed recursively according to:

$$\forall d \in \{0, ..., D-1\}\,, \quad x_{d+1} = \underbrace{x_d}_{\text{skip connection}} + \underbrace{F_{\theta_d}(x_d)}_{\text{residual}}, \quad \text{with} \quad x_0 = x\,. \tag{40}$$

An illustration of a ResNet architecture is depicted in Fig. 2. The residual mappings F_{θ_d} are smaller neural network architectures which can be tailored to the application, typical examples are convolutional layers for image processing tasks [He, 2016a; He, 2016b] or attention-based layers in Transformers for natural language processing [Vaswani, 2017].

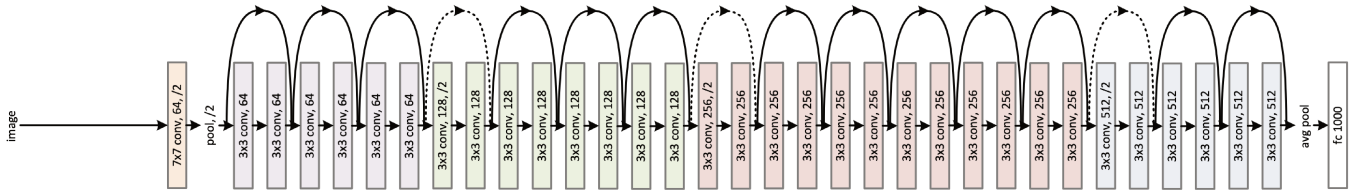


Figure 2: Illustration of ResNet-34 architecture from [He, 2016a].

Note that while Eq. (40) constrains the output of each layer to have the same dimension, in practice ResNets architectures include several downsampling layers reducing the dimension of the signal.

ResNets have demonstrated remarkable empirical performance, achieving state-of-the-art results on benchmarks such as CIFAR-10 [Krizhevsky, 2009] and ImageNet [Deng, 2009]. A key innovation is the skip connection mechanism, which alleviates the vanishing and exploding gradient problems [Raiko, 2012; Szegedy, 2017] and enables the effective training of networks with hundreds or even thousands of layers [He, 2016b]. This increase in depth has spurred new theoretical questions about the behavior of gradient-based optimization in deep networks. While much of the existing theory focuses on shallow architectures [Chizat, 2018], the success of ResNets has motivated the need for developing a mathematical understanding of training dynamics at very large depth.

1.4.2 Neural Ordinary Differential Equations

A central question in the analysis of deep ResNets concerns the appropriate scaling of residual branches as the network depth increases. To ensure effective training in the limit of large depth, a depth-dependent scaling factor β_D is introduced, leading to the modified update rule:

$$\forall d \in \{0, ..., D-1\}\ , \quad x_{d+1} = x_d + \beta_D F_{\theta_d}(x_d)\,. \tag{41}$$

The *Neural Ordinary Differential Equation (NODE)* model introduced by Chen et al. [Chen, 2018] corresponds to the specific scaling $\beta_D = 1/D$, under which Eq. (41) can be interpreted as an explicit Euler discretization of a continuous-time ODE. In the limit as depth tends to infinity, the model has a continuum of parameters $\theta \in \Theta^{[0,1]}$ and processes input data $x \in \mathcal{X}$ by solving the ODE:

$$\forall s \in [0,1]\,, \quad \frac{\mathrm{d}}{\mathrm{d}s} x(s) = F_{\theta(s)}(x(s)), \quad x(0) = x\,, \tag{42}$$

where the parametric vector fields $F_{\theta(s)}$, still referred to as *residuals*, are learned.

Applications At first, the introduction of NODEs was motivated by the possibility of computing gradient through a memory-efficient *adjoint state method* instead of the classical backpropagation algorithm, which requires storing the activations and becomes very memory intensive at large depth [Chen, 2018]. Moreover, a continuous formulation through ODEs also allows considering adaptative integration methods. By design, NODEs model dynamical systems, which has led to their application in areas such as physical modeling and financial time series [Oh, 2025]. Their ability to implement continuous flows of diffeomorphisms has also made them a natural fit for generative modeling with *normalizing flows* [Rezende, 2015; Kobyzev, 2020]. More broadly, NODEs have spurred the development of new neural architectures [Sander, 2021] as well as novel training algorithms [Chen, 2018; Vialard, 2020].

Theoretical analysis From a theoretical standpoint, NODEs offer a convenient mathematical framework — namely, that of ordinary differential equations — for analyzing the training dynamics and performance of very deep neural networks. In particular, their continuous-time formulation enables the use of tools from *optimal control* theory to study questions of training and generalization [E, 2019; E, 2021]. Moreover, this formulation leads to a well-behaved loss landscape, which allows for convergence guarantees when training with gradient-based methods [Sander, 2022b; Marion, 2023b]. In this manuscript, we leveraged the NODE formalism to study in Chapters I and II the training dynamics of both deep and wide ResNets.

Different scalings of residual branches The choice of the scaling factor β_D for residual branches — together with the initialization scheme — plays a crucial role in shaping the training dynamics of ResNets. As in the case of width-dependent scaling in wide neural networks, an appropriate depth scaling can improve convergence behavior and enable significant computational savings, for instance by allowing the transfer of hyperparameters across architectures of varying depth [Yang, 2023; Bordelon, 2025; OpenAI, 2023].

Several scaling regimes have been proposed and studied. The NODE scaling $\beta_D = 1/D$, when paired with smooth (e.g., zero) initialization, leads to stable training and ensures a finite contribution from residuals in the infinite-depth limit. In contrast, under a more standard random initialization of weights, a larger scaling of order $\beta_D = 1/\sqrt{D}$ is required to obtain a non-trivial limiting behavior as depth increases [Cohen, 2021; Marion, 2025]. This latter regime has also been associated with improved feature learning properties [Yang, 2023], although its optimization landscape remains less well understood, and theoretical guarantees for convergence under gradient descent are still lacking.

1.5 Training and gradient descent algorithms

As explained above, in a classical supervised learning framework, the training phase usually consists in the minimization of a risk functional. The objective is to find a parameterization $\theta^* \in \arg\min_{\theta\in\Theta} \mathcal{R}(\theta)$, where $\mathcal{R}$ is the training risk defined in Eq. (29). This optimization problem is usually solved using first order optimization methods whose simplest example is the *gradient descent algorithm* which reads:

$$\forall k \geq 0\,, \quad \theta_{k+1} = \theta_k - \tau \nabla_\theta \mathcal{R}(\theta_k)\,, \tag{43}$$

where $\theta_0 \in \Theta$ is some initialization and $\tau > 0$ is some *step-size* or *learning rate*. In the limit where the step-size $\tau > 0$ vanishes, this discrete dynamic can be modeled by a continuous-time dynamic. Such a *gradient flow* reads:

$$\forall t \geq 0\,, \quad \frac{\mathrm{d}}{\mathrm{d}t}\theta_t = -\nabla_\theta \mathcal{R}(\theta_t)\,. \tag{44}$$

In this manuscript we will analyze the convergence properties of the above dynamics for the training of several neural network architectures, for deep ResNets in Chapters I and II and for single-hidden-layer perceptrons models in Chapter III. While contributing to the understanding of the performances of modern machine learning models, such an analysis presents significant mathematical challenges, the training risk $\mathcal{R}$ being a non-convex function of a usually high number of parameters.

Backpropagation algorithm In practice, the gradient of the risk with respect to the parameter is computed using the *backpropagation algorithm*, corresponding to *reverse-mode automatic differentiation* [Baydin, 2018]. By systematically applying the chain rule

through the computational graph, backpropagation propagates derivatives from the output layer back to the inner layers with a computational cost comparable to that of evaluating the function itself.

The scalability of this approach has been a key enabler of modern deep learning breakthroughs. Efficient implementations of reverse-mode automatic differentiation — combined with the parallel processing capabilities of modern hardware such as GPUs and TPUs — has made it possible to train highly overparameterized models on large datasets. These implementations are made available in popular deep learning libraries such as Pytorch [Paszke, 2017], which we used to validate our result in Section II.7 and Section III.6.

1.5.1 Stochastic gradient descent and variants

While we will focus in this manuscript on the analysis of the simple gradient descent dynamic (Eq. (43)), one should keep in mind that, in practice, training on large datasets is made possible by replacing the full training risk with an approximation computed on a small subset (or *mini-batch*) of data. This leads to the *Stochastic Gradient Descent (SGD)* algorithm, defined by

$$\forall k \geq 0\,, \quad \theta_{k+1} = \theta_k - \tau \nabla_\theta \mathcal{R}_k(\theta_k)\,, \tag{45}$$

where, at iteration k, $\mathcal{R}_k$ is given by:

$$\mathcal{R}_k(\theta) = \frac{1}{\#\mathcal{D}_k} \sum_{(x,y)\in\mathcal{D}_k} \ell(F_\theta(x), y)\,,$$

with $\mathcal{D}_k \subset \mathcal{D}$ a mini-batch of data sampled independently from the dataset $\mathcal{D}$. Besides reducing the computational cost per iteration, SGD introduces stochasticity into the optimization process, which can act as an implicit regularizer and help avoid overfitting, often leading to improved generalization [Hardt, 2016b].

In conjunction with stochastic gradients, a momentum term is often incorporated into the gradient updates, resulting in the update rule:

$$\forall k \geq 0\,, \quad \left\{ \begin{array}{rcl} b_{k+1} & = & \mathrm{m} b_k + (1-\mathrm{m})\nabla_\theta \mathcal{R}_k(\theta_k)\,, \\ \theta_{k+1} & = & \theta_k - \tau b_{k+1}\,, \end{array} \right. \tag{46}$$

where $\mathrm{m} \in [0,1]$ is the momentum parameter. Originally introduced by Polyak [Polyak, 1964], momentum methods are known to accelerate convergence of gradient descent in the case of smooth and strongly convex objectives. Subsequent refinements, such as Nesterov's accelerated gradient method [Nesterov, 1983], achieve optimal convergence rates in the convex setting. In deep learning, incorporating momentum has been shown to lead to improves stability during training [Sutskever, 2013].

Finally, numerous additional techniques have been developed to facilitate the training of large-scale machine learning models, including *dropout*, *weight decay*, *learning rate scheduling*, and *adaptive gradient methods* such as *Adam* [Kingma, 2014] or *RMSprop* [Hinton, 2012]. These methods play a crucial role in improving training stability, convergence speed, and generalization performance [Bottou, 2018].

1.5.2 Two-timescale learning and variable projection

The choice of hyperparameters, and typically of the learning-rate τ for gradient descent, plays a crucial role in the asymptotic behavior of the training dynamic. In particular, the learning rates need not be the same for all of the parameters [Yang, 2021]. In Chapter III, we will distinguish between two types of parameters:

- **Linear parameters:** these are usually the weights of the last layer of the network and are parameters w.r.t. which the output is linear. Everything else being fixed, training these parameters is equivalent to the training of a linear model, i.e. a convex optimization problem for which convergence properties are well-known.
- **Nonlinear parameters:** these are the parameters of the inner layers of the network, which are in a nonlinear relation with the output. These allow extraction of nonlinear representations of the data and play a key role in the generalization abilities of neural networks. However, training these parameters constitutes a non-convex optimization problem.

The parameter space thus decomposes as $\Theta = \Theta^l \times \Theta^{nl}$, where Θ^l and Θ^{nl} denote the subspaces of linear and nonlinear parameters, respectively. Assigning different learning rates for linear and nonlinear parameters thus leads to the *two-timescale* gradient descent:

$$\forall k \geq 0, \quad \left\{ \begin{array}{rcl} \theta^l_{k+1} & = & \theta^l_k - \eta\tau \nabla_{\theta^l} \mathcal{R}(\theta^l_k, \theta^{nl}_k)\,, \\ \theta^{nl}_{k+1} & = & \theta^{nl}_k - \tau \nabla_{\theta^{nl}} \mathcal{R}(\theta^l_k, \theta^{nl}_k)\,, \end{array} \right. \tag{47}$$

where $\tau > 0$ is some step-size and $\eta > 0$ a timescale hyperparameter controlling the relative speed of updates. When $\eta < 1$ the linear parameters θ^l are learned more "slowly" than the nonlinear parameters θ^{nl} and conversely, when $\eta > 1$ the linear parameters are learned more "quickly" than the nonlinear ones.

The asymptotic limit of large timescales corresponds to a partial optimization of the linear parameters, a *Variable Projection (VarPro)* algorithm originally introduced by Golub and Pereyra [Golub, 1973] for the minimization of separable nonlinear least square problems. Indeed, as $\eta \to +\infty$, we have at each step $\theta^l_k \in \arg\min_{\theta^l \in \Theta^l} \mathcal{R}(\theta^l, \theta^{nl}_k)$. Then, by the envelope theorem, the dynamic on the nonlinear parameters reads:

$$\forall k \geq 0, \quad \theta^{nl}_{k+1} = \theta^{nl}_k - \tau \nabla_{\theta^{nl}} \mathcal{R}(\theta^l_k, \theta^{nl}_k) = \theta^{nl}_k - \tau \nabla_{\theta^{nl}} \mathcal{L}(\theta^{nl}_k)\,, \tag{48}$$

where for $\theta^{nl} \in \Theta^{nl}$, the *reduced risk* $\mathcal{L}(\theta^{nl})$ is obtained by:

$$\mathcal{L}(\theta^{nl}) \coloneqq \inf_{\theta^l \in \Theta^l} \mathcal{R}(\theta^l, \theta^{nl})\,. \tag{49}$$

In practice, in the case of regression with square loss, such a partial optimization step can be efficiently performed for a moderate number of neurons by solving a linear system.

Thus, isolating feature learning (slow, nonlinear parameters) from prediction refinement (fast, linear parameters), two-timescale learning and variable projection provide a principled framework for understanding feature learning in neural networks. For this reason, such approaches have recently attracted the interest of the machine learning theory community [Marion, 2023a; Berthier, 2024; Bietti, 2023; Takakura, 2024]. In turn, we will study in Chapter III the convergence properties of VarPro for the training of mean-field models of neural networks.

1.5.3 Wasserstein gradient flows and optimal transport

We focus in this manuscript on the training of overparameterized neural network architectures, which — as in Eq. (39) — are described by a distribution of parameters on a parameter space Θ. For such "mean-field" models, the training risk of Eq. (29) is a functional $\mathcal{R} : \mathcal{P}(\Theta) \to \mathbb{R}$ defined on the space $\mathcal{P}(\Theta)$ of probability distributions on Θ. In particular, for neural networks of finite width M, the distribution of parameters is the

empirical measure $\mu = \frac{1}{M}\sum_{i=1}^{M} \delta_{\theta^i}$. Then, when the parameters $(\theta_t^i)_{1\leq i\leq M}$ follow the gradient flow dynamic Eq. (44), the associated distribution of parameters $\mu_t = \frac{1}{M}\sum_{i=1}^{M} \delta_{\theta_t^i}$ evolves according to the continuity equation:

$$\partial_t \mu_t - \operatorname{div}\left(\mu_t \nabla \frac{\delta \mathcal{R}}{\delta \mu}[\mu_t]\right) = 0\,, \quad \text{on } [0,+\infty)\times\Theta, \tag{50}$$

where for a distribution $\mu \in \mathcal{P}(\Theta)$ the potential field $\frac{\delta \mathcal{R}}{\delta \mu}[\mu]$ is the *first variation* or *Fréchet differential* of $\mathcal{R}$. In the general case where μ_t is not necessarily an empirical probability measure, this PDE can be understood as a metric gradient flow with respect to the *Wasserstein metric* on $\mathcal{P}(\Theta)$ [Ambrosio, 2008b; Santambrogio, 2015].

The Wasserstein distance arises from the problem of optimal transportation of probability measures [Villani, 2009; Santambrogio, 2015]. Assuming Θ is some Hilbert space, the Wasserstein distance $\mathcal{W}_p$, for $p \geq 1$, is defined between two Borel probability measures $\mu, \mu' \in \mathcal{P}(\Theta)$ by:

$$\mathcal{W}_p(\mu,\mu') \coloneqq \left(\inf_{\gamma\in\Gamma(\mu,\mu')} \int_{\Theta\times\Theta} \|\theta-\theta'\|^p \,\mathrm{d}\gamma(\theta,\theta')\right)^{1/p}, \tag{51}$$

where $\Gamma(\mu,\mu')$ is the set of couplings between μ and μ', i.e. the set of probability measures on $\Theta\times\Theta$ whose marginals are respectively μ and μ':

$$\Gamma(\mu,\mu') \coloneqq \left\{\gamma \in \mathcal{P}(\Theta\times\Theta) \;:\; \pi^1_\#\gamma = \mu\,,\, \pi^2_\#\gamma = \mu'\right\}. \tag{52}$$

Then, $\mathcal{W}_p$ provides $\mathcal{P}_p(\Theta)$ — the space of probability measure with finite p-th moment — with a structure of complete separable metric space. In particular, it is known since the work of Jordan, Kinderlehrer, and Otto [Jordan, 1998] that several linear or nonlinear evolution PDEs such as *Fokker-Planck* or *porous medium* equations, can be interpreted as metric gradient flows for this metric. Indeed, in a similar manner than the gradient flow Eq. (44) can be obtained as the limit when $\tau \to 0^+$ of gradient descent Eq. (43), the Wasserstein gradient flow Eq. (50) can be obtained as the limit when $\tau \to 0^+$ of a "JKO scheme" corresponding to its implicit discretization. For an initialization $\mu_0 \in \mathcal{P}(\Theta)$ and a step-size $\tau > 0$, such a JKO scheme reads:

$$\forall k \geq 0\,, \quad \mu_{k+1} \in \operatorname*{arg\,min}_{\mu\in\mathcal{P}(\Theta)} \mathcal{R}(\mu) + \frac{1}{2\tau}\mathcal{W}_2(\mu,\mu_k)^2\,.$$

In machine learning, evolution PDEs of the form Eq. (50) have been used by several authors to study the training dynamics of shallow [Chizat, 2018; Rotskoff, 2019; Mei, 2019; Chizat, 2022; Nitanda, 2022] or deep neural networks [Lu, 2020; Ding, 2021; Isobe, 2023]. Indeed, in addition to providing an elegant formalism for studying the training of neural networks at large width, the relaxation of the risk in the space of measures also benefits from a simplified optimization landscape, for example by eliminating spurious critical points.

2 Contributions

We are now in position to detail the contributions of this work. These contributions are based on three papers that have been written in the context of this PhD and are listed in the list of publications. In all these works, the code for reproducing numerical results is freely available at: `https://github.com/rbarboni`.

I Training of infinitely deep and wide residual architectures

We saw that important effort has been put in developing a mathematical theory for studying the training of overparameterized neural network models. Convergence properties of gradient descent are now partially understood in some linearization regimes — such as the Neural Tangent Kernel regime [Jacot, 2018] — or for some architectures — such as shallow perceptrons with a mean-field scaling [Chizat, 2018]. Yet in applications, recent breakthrough have been made by very deep architectures such as ResNets [He, 2016a] or Transformers [Vaswani, 2017] whose training is permitted by the use of skip-connections.

Our first contribution in Chapter I, is thus to propose a mathematical framework for studying the training of ResNets of both infinite depth and arbitrary width. In this purpose we consider a *mean-field NODE* model, that is a NODE of the form Eq. (42) whose residuals are mean-field models of neural networks of the form Eq. (39). We consider the input space and output space are the same Euclidean space $\mathcal{X} = \mathcal{Y}_{out} = \mathbb{R}^d$, for some $d \geq 1$, and the basis function in Eq. (39) is of the form $\psi : \Theta \times \mathbb{R}^d \to \mathbb{R}^d$. Definition I.1 is as follows:

Definition (Mean-field NODE)**.** *For a family of probability measures $\mu = \{\mu(.|s)\}_{s\in[0,1]} \in \mathcal{P}(\Theta)^{[0,1]}$ and input x, we define the output of the* NODE *model as* $\mathrm{NODE}_\mu(x) \coloneqq x_\mu(1)$ *where $(x_\mu(s))_{s\in[0,1]}$ satisfies the* forward ODE*:*

$$\forall s \in [0,1]\,, \quad \frac{\mathrm{d}}{\mathrm{d}s} x_\mu(s) = F_{\mu(.|s)}(x_\mu(s))\,, \quad \textit{with} \quad x_\mu(0) = x\,. \tag{53}$$

We propose to parameterize this model over the set of probability measures on $[0,1] \times \Theta$ whose marginal is the Lebesgue measure on $[0,1]$. The family of probability measure $\{\mu(.|s)\}_{s\in[0,1]} \in \mathcal{P}(\Theta)^{[0,1]}$ is then obtained by disintegration. This space of parameterizations is defined by:

$$\mathcal{P}_2^{\mathrm{Leb}}([0,1]\times\Theta) \coloneqq \left\{\mu \in \mathcal{P}_2([0,1]\times\Theta) \;:\; \pi^1_\#\mu = \mathrm{Leb}([0,1])\right\}.$$

Conditional optimal transport

When training ResNets, gradient of the risk is computed with respect to the Euclidean metric on the space of parameters at each layer. For our mean-field model of NODE, this corresponds to a layer-wise Wasserstein-2 distance, which we interpret as a *Conditional Optimal Transport (COT)* distance, i.e. a restriction of the classical OT distance which preserves the marginal condition. We define and study properties of the COT distance in Section I.2. In this purpose, we assume Θ is some Euclidean space $\mathbb{R}^p$ for $p \geq 1$. Note that similar topologies on the set of probability measure on product spaces have found other applications, for examples in the study of evolution PDEs with heterogeneities [Peszek, 2023], of Bayesian inverse problems [Hosseini, 2025] or of Bayesian flow matching [Chemseddine, 2024].

The COT or Conditional Wasserstein distance $\mathcal{W}_2^{\mathrm{COT}}$ corresponds to a L^2-Wasserstein distance and is defined for $\mu, \mu' \in \mathcal{P}_2^{\mathrm{Leb}}([0,1] \times \Theta)$ by:

$$\mathcal{W}_2^{\mathrm{COT}}(\mu, \mu') \coloneqq \left(\int_0^1 \mathcal{W}_2 \left(\mu(.|s), \mu'(.|s) \right)^2 \mathrm{d}s \right)^{1/2} .$$

In particular, the space $\mathcal{P}_2^{\mathrm{Leb}}([0,1] \times \Theta)$ equipped with the distance $\mathcal{W}_2^{\mathrm{COT}}$ is a complete metric space (Propositions I.2.1 and I.2.3). As for the classical Wasserstein distance in Eq. (51), we show in Proposition I.2.2 that the distance $\mathcal{W}_2^{\mathrm{COT}}$ can be obtained as the optimal value of a convex minimization problem over the space of couplings. However, in contrast with Eq. (51), this problem has to be restricted to a set of "conditional" couplings. Namely for $\mu, \mu' \in \mathcal{P}_2^{\mathrm{Leb}}([0,1] \times \Theta)$ we have:

$$\mathcal{W}_2^{\mathrm{COT}}(\mu, \mu')^2 = \min_{\gamma \in \Gamma^{\mathrm{Leb}}(\mu, \mu')} \int_{([0,1] \times \Theta)^2} \|\theta - \theta'\|^2 \mathrm{d}\gamma(s, \theta, s', \theta') \,,$$

where $\Gamma^{\mathrm{Leb}}(\mu, \mu')$ is the set of probability measures γ on $[0,1] \times \Theta \times \Theta$ s.t. its first marginal is the Lebesgue measure on $[0,1]$ and s.t. $\gamma(.|s) \in \Gamma(\mu(.|s), \mu'(.|s))$ for $\mathrm{d}s$-a.e. $s \in [0,1]$. As a consequence, the Conditional Wasserstein topology is stronger (and in fact strictly stronger, cf. Remark I.2.1) than the Wasserstein topology.

In the case of the Wasserstein distance, it is a well-established result that absolutely continuous curves in the Wasserstein topology are characterized as solution to linear continuity equations [Ambrosio, 2008b, Thm. 8.3.1]. Generalizing on this result, we show in Theorem I.1 that absolutely continuous curves for the Conditional Wasserstein topology admit a similar dynamic characterization. Precisely, for an interval $I \subset \mathbb{R}$, a curve $(\mu_t)_{t \in I}$ in $\mathcal{P}_2^{\mathrm{Leb}}([0,1] \times \Theta)$ is absolutely continuous if and only if it is solution (in the weak sense) to the continuity equation:

$$\partial_t \mu_t + \mathrm{div}((0, v_t)\mu_t) = 0 \quad \text{on } I \times [0,1] \times \Theta, \tag{54}$$

for some Borel velocity field $v : I \times [0,1] \times \Theta \to \Theta$ such that $\|v_t\|_{L^2(\mu_t)} \in L^1(I)$.

Training NODEs with Conditional Wasserstein gradient flow

We then study the training of the mean-field NODE model. We assume the space of targets is some Euclidean space $\mathcal{Y}_{targ} = \mathbb{R}^{d'}$ for some $d' \geq 1$. Provided with a finite training dataset $\mathcal{D} \subset \mathbb{R}^d \times \mathbb{R}^{d'}$ and a loss function $\ell : \mathbb{R}^d \times \mathbb{R}^{d'} \to \mathbb{R}$ the training risk is then defined for a parameterization $\mu \in \mathcal{P}_2^{\mathrm{Leb}}([0,1] \times \Theta)$ as:

$$\mathcal{R}(\mu) \coloneqq \frac{1}{\#\mathcal{D}} \sum_{(x,y) \in \mathcal{D}} \ell(\mathrm{NODE}_\mu(1), y) \,.$$

In the case of NODEs of finite width, the original method proposed by [Chen, 2018] to compute the gradient is to rely on an *adjoint sensitivity analysis*. In addition to solving the forward equation Eq. (53), the gradient is obtained by solving a *backward ODE*, modeling the computations made by the backpropagation algorithm. For data $(x, y) \in \mathcal{D}$ and parameterization $\mu \in \mathcal{P}_2^{\mathrm{Leb}}([0,1] \times \Theta)$, the adjoint variable $(p_{\mu,x,y}(s))_{s \in [0,1]}$ is solution to:

$$\forall s \in [0,1], \quad \frac{\mathrm{d}}{\mathrm{d}s} p_{\mu,x,y}(s) = -\mathrm{D}_x F_{\mu(.|r)}(x_\mu(s))^\top p_{\mu,x,y}(s) \,, \tag{55}$$

with $p_{\mu,x,y}(1) = \nabla_x \ell(x_\mu(1), y)$. The gradient velocity field $\nabla\mathcal{R}[\mu] : [0,1] \times \Theta \to \Theta$ is then defined as:

$$\nabla\mathcal{R}[\mu](s,\theta) \coloneqq \frac{1}{\#\mathcal{D}} \sum_{(x,y)\in\mathcal{D}} \mathrm{D}_\theta \psi(\theta, x_\mu(s))^\top p_{\mu,x,y}(s).$$

This gives rise to the following Definition I.3, defining gradient flow curves for the training risk $\mathcal{R}$ as solution to a continuity equation of the form Eq. (54) with the gradient field $\nabla\mathcal{R}$. In particular, it has been proven that, in the case of residuals of fixed finite width, the parameter distribution of a ResNet trained with gradient flow converges locally-uniformly in time to a solution of this PDE when the depth of the ResNet tends to infinity [Marion, 2023b]. Eq. (56) generalizes here this gradient flow dynamics to infinitely deep ResNets of arbitrary (finite or infinite) width.

Definition (Gradient flow)**.** *Let $I \subset \mathbb{R}$ be an interval. A locally absolutely continuous curve $t \in I \mapsto \mu_t \in \mathcal{P}_2^{\mathrm{Leb}}([0,1] \times \Theta)$ is a* gradient flow *for $\mathcal{R}$ if it is solution to the continuity equation:*

$$\partial_t \mu_t - \operatorname{div}\left((0, \nabla\mathcal{R}[\mu_t])\, \mu_t\right) = 0 \quad \text{on } I \times [0,1] \times \Theta. \tag{56}$$

In contrast with Eq. (56), gradient flow curves in metric spaces are usually defined as solution to variational problems. We retain in Definition I.5, the notion of *curve of maximal slope* [Ambrosio, 2008b, Def.1.3.2].

Definition (Curve of maximal slope)**.** *Let $I \in \mathbb{R}$ be an interval. Then a locally absolutely continuous curve $(\mu_t)_{t\in I}$ in $\mathcal{P}_2^{\mathrm{Leb}}([0,1]\times\Theta)$ is a* curve of maximal slope *for the risk $\mathcal{R}$ if the map $t \mapsto \mathcal{R}(\mu_t)$ is non-increasing and for $\mathrm{d}t$-a.e. $t \in I$ the following* Energy Dissipation Inequality (EDI) *holds:*

$$\frac{\mathrm{d}}{\mathrm{d}t}\mathcal{R}(\mu_t) \leq -\frac{1}{2}\left(\left|\frac{\mathrm{d}}{\mathrm{d}t}\mu_t\right|^2 + |\nabla\mathcal{R}|^2(\mu_t)\right),$$

where $\left|\frac{\mathrm{d}}{\mathrm{d}t}\mu_t\right|$ is the metric derivative *and $|\nabla\mathcal{R}|$ is an* upper gradient *for $\mathcal{R}$ (Definition I.4).*

Relying on our characterization of absolutely continuous curves for the $\mathcal{W}_2^{\mathrm{COT}}$-topology on $\mathcal{P}_2^{\mathrm{Leb}}([0,1] \times \Theta)$, we show the two above definitions of *gradient flow* and *curve of maximal slope* of the risk $\mathcal{R}$ coincide. This is the content of Theorem I.2:

Theorem. *Let $I \subset \mathbb{R}$ be an open interval. Then a curve $(\mu_t)_{t\in I}$ in $\mathcal{P}_2^{\mathrm{Leb}}([0,1] \times \Theta)$ is a* gradient flow *for the risk $\mathcal{R}$ if and only if it is a* curve of maximal slope *for $\mathcal{R}$.*

In turn, this identification allows us to use results on the existence and uniqueness of curves of maximal slope to deduce corresponding statements for gradient flow curves. These results, presented in Section I.3.4 hold under mild regularity and growth assumptions on the basis function ψ and the special case of shallow perceptrons is treated in Section I.A. Our existence and uniqueness result is the following:

Theorem. *Let $\mu_0 \in \mathcal{P}_2^{\mathrm{Leb}}([0,1] \times \Theta)$ be some parameter initialization. Then there exists a unique curve of maximal slope / gradient flow $(\mu_t)_{t\in[0,+\infty)}$ for the risk $\mathcal{R}$ starting from μ_0. In particular, such a gradient flow curve is defined for every time $t \geq 0$.*

II Convergence in the training of residual architectures

Relying on the mathematical framework developed in Chapter I, we focus in Chapter II on the asymptotic analysis of the gradient flow dynamic Eq. (56) for the training of deep ResNets or NODEs. Considering standard examples of residual architectures such as *random feature models* [Rahimi, 2007] or *single-hidden-layer perceptrons*, we show a convergence result: for proper initializations of the parameters, the gradient flow converges at a linear rate to a parameterization that is a global minimizer of the training risk. In contrast, other convergence results for the training of deep and wide ResNets either state optimality of the parameterization under a convergence assumption [Lu, 2020; Ding, 2022], or convergence towards a first order critical point which is not necessarily a global optimizer [Isobe, 2023]. In the end, our theoretical results are supported by numerical experiments on image classification datasets. The code is available at: `https://github.com/rbarboni/FlowResNets`.

ResNets and Polyak-Łojasiewicz property

Our proof strategy to obtain convergence of gradient flow is to show that the training risk $\mathcal{R}$ satisfies a *Polyak-Łojasiewicz (P-Ł) inequality* around appropriate initializations. Initially introduced by Polyak [Polyak, 1963] for studying the convergence of gradient based optimization algorithms, such an inequality has been observed by several authors to hold for the risk associated to the training of neural networks [Oymak, 2019; Chatterjee, 2022; Marion, 2023b]. We review in Section II.2 the local versions of the P-Ł inequality we will use as well as the local convergence results it implies for gradient descent and gradient flow.

For the risk $\mathcal{R}$ associated to the training of our mean-field NODE model, the local P-Ł inequality here takes the form:

$$\|\nabla\mathcal{R}[\mu]\|^2_{L^2(\mu)} \geq m\mathcal{R}(\mu)\,, \tag{57}$$

where $m > 0$ is the *P-Ł constant* and μ is any parameterization in the neighbourhood of some initialization $\mu_0 \in \mathcal{P}_2^{\mathrm{Leb}}([0,1]\times\Theta)$. In particular, Eq. (57) implies that every critical point of $\mathcal{R}$ in a neighbourhood of μ_0 is actually a global minimizer. Also, applying results from [Dello Schiavo, 2024], it allows to conclude to the convergence of gradient flow curves to an optimal parameterization at a linear rate when the risk at initialization is already sufficiently small.

In the context of deep ResNets, we show in Lemma II.3.1 that a P-Ł property is generically satisfied by our mean-field NODE model and detail in Section II.3.2 how the P-Ł constant depends on the residuals architecture and on the approximation properties of the associated functional space. Precisely, Eq. (II.15) shows that the P-Ł constant can be expressed in terms of the conditioning of the residuals Neural Tangent Kernel (NTK). In the case of mean-field residuals of the form Eq. (39), the NTK depends on the parameterization $\mu \in \mathcal{P}(\Theta)$ at each layer and is given by:

$$K[\mu](x,x') \coloneqq \int_\Theta \mathrm{D}_\theta\psi(\theta,x)\mathrm{D}_\theta\psi(\theta,x')^\top \mathrm{d}\mu(\theta)\,. \tag{58}$$

The NTK in particular evolves with the parameterization during training but, assuming it stays well-conditioned, one can show convergence of gradient flow to a global minimizer of the training risk (Corollary II.3.1). In turn, we show this assumption can be satisfied for standard architectures in Sections II.4 and II.5.

Convergence for RKHS residuals

We first investigate in Section II.4 the case of a linear parameterization of the residuals. For a Hilbert space of features $\mathcal{H}$ and a parameter $\theta \in \Theta = \mathcal{H}^d$, these are residuals of the form:

$$F_\theta(x) = \begin{pmatrix} \langle \theta_1, \phi(x) \rangle_{\mathcal{H}} \\ \vdots \\ \langle \theta_d, \phi(x) \rangle_{\mathcal{H}} \end{pmatrix},$$

where $\phi = \mathbb{R}^d \to \mathcal{H}$ is some *feature map*. This for example encompasses the case of linear layers or of 2-layer perceptrons of arbitrary width with fixed hidden layer such as *random feature models* [Rahimi, 2007]. Our interest in this architecture is motivated by the fact that the space of residuals $\mathcal{F} \coloneqq \{F_\theta \,:\, \theta \in \Theta\}$ can then be provided with a structure of *Reproducing Kernel Hilbert Space (RKHS)*. This has two main interests:

- Equipped with this RKHS metric, the space of residual maps is isometric to the space of parameters $\Theta = \mathcal{H}^d$ equipped with its standard Hilbert metric. This allows seeing the NODE as a nonparametric RKHS-NODE model defined in Definition II.4 by the integration of an ODE with nonparametric time-dependent residuals. Moreover, the gradient flow dynamic Eq. (44) can be projected onto a gradient flow on the space of residuals (Proposition II.4.3).
- For this architecture, the NTK is the kernel naturally associated with the RKHS structure. In particular, it does not depend on the parameterization and stays constant during training. Choosing functional spaces with good approximation properties, we then obtain a convergence result for gradient flow in Theorem II.4 and for gradient descent in Theorem II.5. Moreover these functional spaces can be approximated by random feature models of sufficiently large width. As a consequence, we also obtain in Theorem II.6 convergence result for deep ResNets whose width is polynomial in the number of data samples.

Convergence for SHL residuals

We then turn in Section II.5 to the more realistic case of residuals which are two-layer perceptrons of arbitrary width with trained hidden layers. These are mean-field models of the form Eq. (39) with a parameter space $\Theta = \mathbb{R}^d \times \mathbb{R}^d \times \mathbb{R}$ and a basis function $\psi : \Theta \times \mathbb{R}^d \to \mathbb{R}^d$ of the form:

$$\forall (u, w, b) \in \Theta\,,\ \forall x \in \mathbb{R}^d\,, \quad \psi((u, w, b), x) = u\sigma(w^\top x + b)\,,$$

where σ is some nonlinear activation function such as any smooth approximation of ReLU. In this case, leveraging the partial linearity of ψ with respect to its parameters, the NTK in Eq. (58) decomposes as a sum of two positive kernels:

$$\forall x, x' \in \mathbb{R}^d\,, \quad K[\mu](x, x') = k^1[\mu](x, x')\mathrm{Id} + K^2[\mu](x, x')\,,$$

where $k^1[\mu]$ is a scalar kernel corresponding to gradients w.r.t. the linear parameter u and depending only on the marginal of μ w.r.t. (w, b), or *feature distribution*, and $K^2[\mu]$ corresponds to gradients w.r.t. the nonlinear parameters (w, b).

In case the feature distribution is fixed, that is ψ only has linear parameters, k^1 is the kernel associate to the random feature model previously studied in Section II.4. Spectral

properties of this type of kernel have been studied for different type of activations [Bach, 2017a; Cho, 2009] but, generically, strict positivity of the NTK is ensured as soon as the feature distribution has a dense support (Proposition II.5.1). Moreover, in the special case of a trigonometric activation and a uniform bias distribution, a lower bound on the conditioning of the NTK can be obtained by leveraging results from the theory of radial basis function interpolation [Schaback, 1995].

However, in contrast with Section II.4, the kernel k^1 may evolve during training. Still, using that the conditioning of k^1 is a Lipschitz continuous function of the parameter distribution (Lemma II.5.1), we are able to obtain a convergence result in Theorem II.7. Moreover, our convergence assumptions are precisely quantified with respect to the number of data samples in Corollary II.5.1, for special cases of activations and of initializations.

III Feature learning in shallow architectures

In Chapter II, we have shown convergence of gradient descent and gradient flow for the training of deep ResNets with different choices of residual architectures. In case residuals are single-hidden-layer (SHL) perceptrons of the form Eq. (34), convergence requires a sufficiently spread distribution of *features*, i.e. weights in the inner layer. However, while this assumption can be ensured at initialization, our analysis in Chapter II is unable to track the evolution of the feature distribution during training. This restriction is arguably the pitfall of many convergence results for the training of neural networks which are unable to describe the evolution of nonlinear parameters, even though *feature learning* is expected to be at the core of approximation and generalization capabilities of neural networks [Chizat, 2018; Rotskoff, 2019; Allen-Zhu, 2019; Du, 2019; Lee, 2019; Zou, 2020].

To tackle this problem, we study in Chapter III the training of a mean-field model of neural network for solving a univariate regression task in a *teacher-student scenario* where the target signal is given by a neural network. In this setting, we consider the *Variable Projection (VarPro)* algorithm described in Eq. (48) and show convergence of the student feature distribution to the teacher feature distribution. In addition, in a certain regime of small regularization, we are able to establish a linear convergence rate for the feature distribution by comparing the training dynamic to the solution of a *weighted ultra-fast diffusion equation* [Iacobelli, 2019b]. Our result are to be compared with the results of Chizat and Bach [Chizat, 2018] and Rotskoff et al. [Rotskoff, 2019], which establish qualitative convergence results for the learning of the feature distribution in the training of shallow neural networks with gradient descent. In contrast we study a two-timescale variant of gradient descent and establish a linear convergence rate.

In the end, these theoretical results are supported by numerical experiments. We show on low-dimensional problems with synthetic data that, for a suitable choice of hyperparameters, the evolution of the feature distribution during training can indeed be faithfully modeled by an ultra-fast diffusion equation. Moreover, we also show with experiments on the CIFAR10 dataset [Krizhevsky, 2009] that the VarPro algorithm can be adapted for solving large-scale machine learning problems. The code is available at: `https://github.com/rbarboni/VarPro`.

Variable Projection and reduced risk

We first study properties of the VarPro algorithm. In Chapter III, we consider mean-field models of neural networks of the form Eq. (39), with a basis function ψ that is partially linear with respect to its parameters. Precisely, we consider that the parameter set is of

the form $\Theta = \mathbb{R} \times \Omega$, where Ω is some space of features, and ψ is given by:

$$\psi : ((u, \omega), x) \in \Theta \times \mathbb{R}^d \mapsto u\phi(\omega, x) \in \mathbb{R}\,,$$

where $\phi : \Omega \times \mathbb{R}^d \to \mathbb{R}$ is some *feature map*. This setting for example encompasses the case of SHL perceptron models of Eq. (34) where the feature map is of the form $\phi : ((w, b), x) \mapsto \sigma(w^\top x + b)$. Then, given a feature distribution $\mu \in \mathcal{P}(\Omega)$, the linear parameters can be considered as a function $u \in L^1(\mu)$ and the neural network's output reads:

$$\forall x \in \mathbb{R}^d, \quad F_{\mu,u}(x) = \int_\Omega u(\omega)\phi(\omega, x)\mathrm{d}\mu(\omega)\,. \tag{59}$$

We consider a univariate regression problem with square loss. For a regularization strength λ, the training risk for a feature distribution $\mu \in \mathcal{P}(\Omega)$ and outer weights $u \in L^2(\mu)$ reads:

$$\mathcal{R}^\lambda(\mu, u) = \frac{1}{\#\mathcal{D}} \sum_{(x,y)\in\mathcal{D}} \frac{1}{2} |F_{\mu,u}(x) - y|^2 + \lambda \int_\Omega \|u\|^2 \mathrm{d}\mu\,.$$

Leveraging the partial linearity of ψ, one can distinguish in Eq. (59) between nonlinear parameters, which are encoded in the feature distribution $\mu \in \mathcal{P}(\Omega)$, and linear parameters, which are encoded in $u \in L^2(\mu)$. In particular, for a fixed feature distribution μ, minimization of $\mathcal{R}^\lambda$ with respect to u is a ridge regression problem which can be solved analytically in closed form and performed efficiently by numerically solving a linear system. As described in Eq. (48), the VarPro algorithm — or two-timescale limit of gradient descent — then consists in performing partial optimization over u before taking a gradient step on the non linear parameters. Equivalently, it can be seen as a gradient descent over a *reduced risk* defined for every feature distribution $\mu \in \mathcal{P}(\Omega)$ by:

$$\mathcal{L}^\lambda(\mu) = \inf_{u\in L^2(\mu)} \frac{1}{\lambda}\mathcal{R}^\lambda(\mu, u) = \inf_{u\in L^2(\mu)} \frac{1}{\#\mathcal{D}} \sum_{(x,y)\in\mathcal{D}} \frac{1}{2\lambda} |F_{\mu,u}(x) - y|^2 + \int_\Omega \|u\|^2 \mathrm{d}\mu\,.$$

In the case $\lambda = 0$, this reduced risk is the value of a constrained optimization problem:

$$\mathcal{L}^0(\mu) = \inf_{\substack{u\in L^2(\mu) \\ F_{\mu,u}=Y}} \int_\Omega \|u\|^2 \mathrm{d}\mu\,.$$

We consider a "teacher-student" scenario described by Assumption III.1 in which the target signal is represented by a teacher network with some teacher feature distribution $\bar{\mu} \in \mathcal{P}(\Omega)$. In this scenario, we show in Section III.2 that, for $\lambda > 0$, the reduced risk $\mathcal{L}^\lambda$ can be interpreted as an infimal convolution between two types of statistical distances: a *Maximum Mean Discrepancy (MMD)* distance, which arises from the convolution with the feature map ϕ, and a χ^2-divergence, which arises from the regularization term. In particular, in the limit where $\lambda \to 0^+$, $\mathcal{L}^0$ corresponds to the χ^2-divergence between the teacher feature distribution $\bar{\mu}$ and the student feature distribution μ. We then show in Lemma III.3.3 that the functional $\mathcal{L}^\lambda$ Γ-converges on $\mathcal{P}(\Omega)$ to this χ^2-divergence, implying for example convergence of sequences of minimizers of $\mathcal{L}^\lambda$ to the teacher distribution (Proposition III.3.1).

Convergence and ultra-fast diffusion regime

We show in Section III.4 that, in the limit of small learning rates, the evolution of the feature distribution μ under the VarPro algorithm corresponds to a Wasserstein gradient flow for the reduced risk $\mathcal{L}^\lambda$. Following Eq. (50), this evolution takes the form:

$$\partial_t \mu_t - \operatorname{div}\left(\mu_t \nabla \frac{\delta \mathcal{L}^\lambda}{\delta \mu}[\mu_t]\right) = 0\,, \quad \text{on } [0,\infty)\times\Omega, \tag{60}$$

where the potential $\frac{\delta \mathcal{L}^\lambda}{\delta \mu}[\mu]$ is the first variation of $\mathcal{L}^\lambda$. We show in Theorem III.1, that the above equation is well-posed in the case $\lambda > 0$. Moreover, we show in Theorem III.4 that, provided solutions stay sufficiently smooth (e.g., with bounded log-density) the reduced risk $\mathcal{L}^\lambda$ converges to 0 with a convergence rate of order $\mathcal{O}(1/t)$. This in particular implies that, in the limit where $t \to +\infty$, the student feature distribution converges to the teacher's.

In the case $\lambda = 0$, we explain in Section III.4 how the wasserstein gradient flow of the reduced risk $\mathcal{L}^0$ can be interpreted as a *weighted ultra-fast diffusion equation* of form:

$$\partial_t \mu - \operatorname{div}\left(\bar{\mu}\nabla\left(\frac{\mu}{\bar{\mu}}\right)^{-1}\right) = 0\,, \quad \text{on } [0,\infty)\times\Omega, \tag{61}$$

where $\bar{\mu} \in \mathcal{P}(\Omega)$ is the teacher feature distribution. In particular, well-posedness of such a *weighted ultra-fast diffusion* has been shown in the case where Ω is the n-dimensional flat torus, or a bounded convex domain of $\mathbb{R}^n$ and Neumann boundary conditions are imposed [Iacobelli, 2019b]. Moreover, in this case, the solutions converge in L^2 to the teacher feature distribution $\bar{\mu}$ at a linear rate, i.e. a convergence rate of order $\mathcal{O}(e^{-Ct})$ for some constant $C > 0$. In turn, taking the limit $\lambda \to 0^+$, we show in Theorem III.5 that (sufficiently regular) solutions of Eq. (60) converge locally-uniformly in time to solutions of the weighted ultra-fast diffusion Eq. (61).

In Section III.6, we show these theoretical predictions are supported by numerical experiments on simple settings reproducing our assumptions. We observe that, for a sufficiently low regulariation strength $\lambda > 0$, single-hidden-layer neural networks trained with VarPro indeed enter an ultra-fast diffusion regime where the teacher feature distribution is recovered with a linear convergence rate.

Chapter I

Training of infinitely deep and wide residual architectures: mathematical models and Conditional Optimal Transport

Contents

I.1 Introduction **39**
I.1.1 Mean-field models of neural networks 41
I.1.2 Mean-field NODEs 42
I.1.3 Related works and contributions 44
I.2 Metric structure of the parameter set $\mathcal{P}_2^{\mathrm{Leb}}([0,1]\times\Theta)$ **45**
I.2.1 Conditional Optimal Transport distance 45
I.2.2 Dynamical formulation of Conditional Optimal Transport 49
I.3 Gradient flow dynamics **55**
I.3.1 Backward equation and adjoint variables 55
I.3.2 The gradient flow equation 57
I.3.3 Gradient flows as curves of maximal slope 59
I.3.4 Existence, uniqueness, and stability of gradient flow curves 68
Appendices **75**
I.A Well-posedness of the gradient flow equation for SHL residuals **75**

I.1 Introduction

Understanding the training dynamics of neural networks is an important problem in Machine Learning as it brings the hope of understanding the good performances of these models. This training is however an involved optimization problem, usually solved by performing (stochastic) gradient descent for the training risk, an optimization procedure which, though simple, often manages to find a global minimum of the risk despite its non-convexity. This phenomenon is now correctly understood in some simple cases such as the one of linear networks [Hardt, 2016a; Bartlett, 2018; Zou, 2019; Bah, 2022]. In the more realistic case of non-linear architectures, most works have focused on *Multi-Layer Perceptrons (MLP)* [Li, 2017; Du, 2019; Allen-Zhu, 2019; Zou, 2020; Lee, 2019; Chen, 2020; Nguyen, 2021] and convergence towards a minimizer of the risk can be obtained with great

probability over a random initialization provided that the network is sufficiently wide, a regime referred to as "overparameterization". Taking the limit of infinite width, many works have also studied the convergence of gradient descent for the training of neural networks in the limit of an infinite number of parameters [Chizat, 2018; Mei, 2018; Javanmard, 2020; Wojtowytsch, 2020; Nguyen, 2023]. In those works, the neural network is trained by modeling the parameters as a probability measure over the parameter space and performing a Wasserstein gradient flow over the set of probability measures. Notably, Chizat and Bach [Chizat, 2018] establish a result of optimality at convergence: if the gradient flow converges then its limit is a global minimizer of the training risk.

We will focus in the first two chapters of the thesis on the case of the *Residual Neural Network (ResNet)* architecture which we presented in Section 1.4. ResNets were first introduced by He et al. [He, 2016a] for applications in computer vision but the architecture has since distinguished itself by obtaining state-of-the-art results in several other machine-learning applications. A key feature of ResNets is the extensive use of *skip connections* (Eq. (40)): each layer consists of the addition of a perturbation (called *residual*) to the output of the previous layer. The presence of skip connections has indeed been identified to ease the training of deeper neural networks [Raiko, 2012; Szegedy, 2017] by mitigating the *vanishing / exploding gradient* phenomena, a common problem encountered when training deep neural networks [Bengio, 1994; Glorot, 2010]. The ResNet architecture has thus permitted the training of neural networks of almost arbitrary depth [He, 2016b]. Considering the limit where the depth tends to infinity Chen et al. [Chen, 2018] introduced the *Neural Ordinary Differential Equation (NODE)* architecture we presented in Section 1.4.2: with a $1/D$ scaling of residual branches, passing to the limit of infinite depth leads to a model performing the integration of the ODE Eq. (42), with a parametric velocity field.

An important contribution of NODEs is to provide a theoretical framework upon which many other works have been based to study very deep neural network architectures. Chen et al. [Chen, 2018] proposed a method based on *adjoint sensitivity analysis* to compute the gradient of NODEs efficiently without automatic differentiation. Sander et al. [Sander, 2021] proposed a new architecture based on a second-order ODE which can be trained with reduced computational complexity. Inspired by methods from medical imaging and shape analysis, Vialard et al. [Vialard, 2020] proposed a new algorithm for the training of deep ResNets. E, Han, and Li [E, 2019] and E, Ma, and Wu [E, 2021] studied the training and generalization properties of deep ResNets borrowing tools from the mathematical theory of *Optimal Control*.

Notations For a metric space X, $\mathcal{P}(X)$ is the set of Borel probability measures over X. This set is endowed with the *narrow topology*, which is the topology of convergence against the set $\mathcal{C}_b(X)$ of bounded continuous functions. For $x \in X$, we note $\delta_x \in \mathcal{P}(X)$ the Dirac measure at x. For $p \geq 1$, $\mathcal{P}_p(X)$ is the subset of $\mathcal{P}(X)$ of probability measures with finite p-order moment, endowed with the *Wasserstein distance* $\mathcal{W}_p$ defined in Eq. (51) [Villani, 2009; Santambrogio, 2015]. When X is a Hilbert space we define on $\mathcal{P}_p(X)$ the p-Energy $\mathcal{E}_p(\mu) \coloneqq \int_X |x|^p \mathrm{d}\mu(x)$. If $\mu \in \mathcal{P}(X)$ and $f : X \mapsto Y$ is a measurable map between topological spaces we denote by $f_{\#}\mu \in \mathcal{P}(Y)$ the pushforward of μ by f. If $\{f_i : X_i \to Y_i\}_{1 \leq i \leq n}$ is a family of mappings then $(f_1, ..., f_n)$ designates the product map $(f_1, ..., f_n) : (x_1, ..., x_n) \mapsto (f_1(x_1), ..., f_n(x_n))$ and if $X = X^1 \times ... \times X^n$ is a product space we designate by π^i the projection $\pi^i : (x^1, ..., x^n) \in X \mapsto x^i \in X^i$.

I.1.1 Mean-field models of neural networks

We consider in this chapter *Neural ODEs (NODEs)* modeling ResNets whose depth tends to infinity with a proper rescaling of the residuals layers [Chen, 2018]. We also consider residuals of the form Eq. (39). Such "mean-field models" can be thought of as neural networks of arbitrary width and were studied before by Chizat and Bach [Chizat, 2018], Mei, Montanari, and Nguyen [Mei, 2018], Rotskoff and Vanden-Eijnden [Rotskoff, 2018], Wojtowytsch [Wojtowytsch, 2020], and Nguyen and Pham [Nguyen, 2023]. Provided with the *parameter space* $\Theta \subset \mathbb{R}^p$, the input space $\mathbb{R}^d$ and with a Borel map $\psi : \Theta \times \mathbb{R}^d \to \mathbb{R}^d$ (the *basis function*), we consider mappings $F_\mu : \mathbb{R}^d \to \mathbb{R}^d$ parameterized by measures $\mu \in \mathcal{P}(\Theta)$ and defined by:

$$F_\mu : x \in \mathbb{R}^d \mapsto \int_\Theta \psi(\theta, x) \mathrm{d}\mu(\theta) \,. \tag{I.1}$$

Single-hidden-layer perceptrons The above definition encompasses as a particular case standard neural network architectures. For example, for $\Theta = \mathbb{R}^d \times \mathbb{R}^d \times \mathbb{R}$ and $\psi : ((u, w, b), x) \mapsto u\sigma(w^\top x + b)$ with a real-valued function $\sigma : \mathbb{R} \to \mathbb{R}$ (called *activation*), considering the atomic measure $\mu = \frac{1}{M} \sum_{i=1}^M \delta_{(u_i, w_i, b_i)}$ one recovers the classical model of a *single-hidden-layer (SHL) perceptron* of width $M \geq 1$ defined in Eq. (34):

$$F_\mu : x \mapsto \frac{1}{M} \sum_{i=1}^M u_i \sigma(w_i^\top x + b_i) \,. \tag{I.2}$$

We will study this type of architecture in more detail in Chapters II and III.

Convolutional layers Closer to applications, Eq. (I.1) also encompasses the residuals originally used by He et al. [He, 2016a]. These consists of two of the convolutional layers defined in Eq. (31), composed with a nonlinear activation. Consider integers $n, c, k \geq 0$ and $\Theta = \mathbb{R}^{c \times 1 \times k \times k} \times \mathbb{R}^{1 \times c \times k \times k} \times \mathbb{R}^{1 \times n \times n}$, the set of parameters of the form (u, w, b) where u and w are convolutional filters of size $c \times 1 \times k \times k$ and $1 \times c \times k \times k$ respectively and b is a bias term of size $1 \times n \times n$. Then for an image input $x \in \mathbb{R}^{c \times n \times n}$ of size $n \times n$ with c channels, and $(u, w, b) \in \Theta$ consider the basis function $\psi : ((u, w, b), x) \mapsto u \star \sigma(w \star x + b) \in \mathbb{R}^{c \times n \times n}$ where the activation σ is applied component-wise. Then for an empirical measure $\mu = \frac{1}{M} \sum_{i=1}^M \delta_{(u_i, w_i, b_i)}$, Eq. (I.1) gives on input x:

$$F_\mu(x) = \frac{1}{M} \sum_{i=1}^M u_i \star \sigma(w_i \star x + b_i) \,, \tag{I.3}$$

which is the output of a ResNet residual with M intermediary channels in [He, 2016a]. However, the definition in Eq. (I.1) does not model some popular architectures such as normalization or pooling layers which play an important role in the success of ResNets.

Attention layers Finally, Eq. (I.1) also models *attention layers* at the heart of *Transformers* architectures [Vaswani, 2017]. Consider as parameter space $\Theta = \mathbb{R}^{c \times d} \times \mathbb{R}^{c \times d} \times \mathbb{R}^{d \times d}$, the set of triplets (K, Q, V) where $K \in \mathbb{R}^{c \times d}$ is the *key* matrix, $Q \in \mathbb{R}^{c \times d}$ is the *query* matrix and $V \in \mathbb{R}^{d \times d}$ is the *value* matrix. For parameters $(K, Q, V) \in \Theta$ and an input sequence of *tokens* $\mathbf{x} = (x^i)_{1 \leq i \leq N} \in (\mathbb{R}^d)^N$ of length $N \geq 0$, the *attention head* defined in Eq. (35) is:

$$\psi((K, Q, V), \mathbf{x}) = \mathrm{Attention}((K, Q, V), \mathbf{x}) := \left(\sum_{j=1}^N \frac{e^{\langle Kx^i, Qx^j \rangle}}{\sum_{j=1}^N e^{\langle Kx^i, Qx^j \rangle}} V x^j \right)_{1 \leq i \leq N} \in (\mathbb{R}^d)^N \,.$$

Then for an empirical measure $\mu = \frac{1}{M}\sum_{k=1}^{M}\delta_{(K_k,Q_k,V_k)}$, Eq. (I.1) defines a *multi-head attention* layer with M heads:

$$F_\mu(\mathbf{x}) = \frac{1}{M}\sum_{k=1}^{M}\mathrm{Attention}((K_k,Q_k,V_k),\mathbf{x})\,. \tag{I.4}$$

Note that with this definition, we are only able to describe Transformer architectures taking as input sequences of tokens of fixed length N. However, our setting could be adapted to model Transformer architecture taking as inputs sequences of tokens of various finite lengths by considering different basis functions depending on the length of the input sequence.

I.1.2 Mean-field NODEs

We proceed then to the definition of *Neural ODEs (NODEs)* modeling ResNets whose depth tends to infinity with a proper rescaling of the residual layers [Chen, 2018]. Our NODE model is then an ODE whose velocity field (or *residual*) belongs to the class of mappings parameterized by measure defined in Eq. (I.1). Similar models of "mean-field NODEs" or "mean-field limit of ResNets" were studied by Lu et al. [Lu, 2020], Ding et al. [Ding, 2022], and Isobe [Isobe, 2023].

Definition I.1 (Mean-field NODE)**.** *For a family of probability measures $\mu = \{\mu(.|s)\}_{s\in[0,1]} \in \mathcal{P}(\Theta)^{[0,1]}$ and input $x \in \mathbb{R}^d$, we define the NODE model output as* $\mathrm{NODE}_\mu(x) \coloneqq x_\mu(1)$ *where $(x_\mu(s))_{s\in[0,1]}$ satisfies the* Forward ODE*:*

$$\frac{\mathrm{d}}{\mathrm{d}s}x_\mu(s) = F_{\mu(.|s)}(x_\mu(s))\,,\quad x_\mu(0) = x\,. \tag{I.5}$$

When there is no ambiguity, we simply write $x(s)$.

The parameter set $\mathcal{P}_2^{\mathrm{Leb}}([0,1]\times\Theta)$ To justify the well-posedness of Eq. (I.5) it is first necessary to define the adequate set of parameters we will consider. Given a topological space Z, we define $\mathcal{P}_2^{\mathrm{Leb}}([0,1]\times Z)$ as the set of probability measures $\mu \in \mathcal{P}_2([0,1]\times Z)$ whose marginal w.r.t. $[0,1]$ is the Lebesgue measure $\mathrm{Leb}([0,1])$:

$$\mathcal{P}_2^{\mathrm{Leb}}([0,1]\times Z) \coloneqq \left\{\mu \in \mathcal{P}_2([0,1]\times Z)\ :\ \pi^1_\#\mu = \mathrm{Leb}([0,1])\right\}\,.$$

Given $\mu \in \mathcal{P}_2^{\mathrm{Leb}}([0,1]\times\Theta)$, using a disintegration result [Attouch, 2014, Thm.4.2.4], there exists a ds-a.e. uniquely determined family of probability measures $\mu(.|s) \in \mathcal{P}_2(Z)$ such that for every measurable $f : [0,1]\times Z \to \mathbb{R}$ the mapping:

$$s \in [0,1] \mapsto \int_Z f(s,z)\mathrm{d}\mu(z|s)$$

is measurable and

$$\int_{[0,1]\times Z} f(s,z)\mathrm{d}\mu(s,z) = \int_0^1\int_Z f(s,z)\mathrm{d}\mu(z|s)\mathrm{d}s\,.$$

In the following, we will consider as parameters probability measures $\mu \in \mathcal{P}_2^{\mathrm{Leb}}([0,1]\times\Theta)$. Therefore, every parameter $\mu \in \mathcal{P}_2^{\mathrm{Leb}}([0,1]\times\Theta)$ is naturally associated with a (almost everywhere uniquely defined) family of probability measures $\{\mu(.|s)\}_{s\in[0,1]}$. We will provide this set of parameters with a modification of the Wasserstein-2 distance [Villani, 2009;

Santambrogio, 2015] that takes into account the marginal constraint by considering a restriction of Kantorovitch's original optimal coupling problem to the set of couplings that are the identity on the first variable $s \in [0,1]$. The solution of this new optimization problem induces the *Conditional Optimal Transport* (COT) distance on the parameter set [Hosseini, 2025].

Well-posedness of NODEs The following assumption on ψ will be sufficient to show the well-posedness of Eq. (I.5) for any parameter $\mu \in \mathcal{P}_2^{\mathrm{Leb}}([0,1] \times \Theta)$. This is the content of Proposition I.1.1.

Assumption I.1. *Assume $\psi : \Theta \times \mathbb{R}^d \to \mathbb{R}^d$ is measurable and*

(i) (quadratic growth) grows at most quadratically w.r.t. θ and linearly w.r.t. x: there exists a constant C s.t.

$$\forall x \in \mathbb{R}^d\,,\ \forall \theta \in \Theta\,,\quad \|\psi(\theta, x)\| \leq C(1 + \|x\|)(1 + \|\theta\|^2)\,.$$

(ii) (local Lipschitz continuity) is locally Lipschitz w.r.t. x, with a Lipschitz constant that grows at most quadratically with θ: for every $R \geq 0$, there exists a constant $C(R)$ s.t.

$$\forall x, x' \in B(0,R)\,,\ \forall \theta \in \Theta\,,\quad \|\psi(\theta, x) - \psi(\theta, x')\| \leq C(R)(1 + \|\theta\|^2)\|x - x'\|\,.$$

Proposition I.1.1 (Well-posedness of the flow)**.** *Assume $\mu \in \mathcal{P}_2^{\mathrm{Leb}}([0,1] \times \Theta)$ and ψ satisfies Assumption I.1. Then for every $x \in \mathbb{R}^d$ there exists a unique weak solution to Eq. (I.5), that is an absolutely continuous path $(x(s))_{s\in[0,1]}$ such that for every $s \in [0,1]$:*

$$x(s) = x + \int_0^s F_{\mu(.|r)}(x(r))\mathrm{d}r\,. \tag{I.6}$$

Proof. The result follows Caratheodory's theorem for the existence and uniqueness of absolutely continuous solutions [Hale, 2009, Sec.I.5]. Indeed, given $\mu \in \mathcal{P}_2^{\mathrm{Leb}}([0,1] \times \Theta)$ the map $(s,x) \mapsto F_{\mu(.|s)}(x)$ is measurable w.r.t. s and, thanks to Assumption I.1 (local Lipschitz continuity), locally Lipschitz w.r.t. x with a local Lipschitz constant that is integrable w.r.t. s. Moreover, the solutions of Eq. (I.6) are defined up to time $s = 1$ thanks to the growth assumption in Assumption I.1, and if C is the growth constant we get the following bound on the solution:

$$\forall s \in [0,1]\,,\quad \|x(s)\| \leq \exp(C(1 + \mathcal{E}_2(\mu)))(\|x(0)\| + C(1 + \mathcal{E}_2(\mu)))\,. \tag{I.7}$$

□

Supervised learning We consider the supervised learning framework presented in Section 1.1 with input and output space $\mathcal{X} = \mathcal{Y}_{out} = \mathbb{R}^d$ and space of targets $\mathcal{Y}_{targ} = \mathbb{R}^{d'}$ for $d, d' \geq 1$. Given a data distribution $\mathbb{R}^d \times \mathbb{R}^{d'} \ni (x,y) \sim \mathcal{D}$ and loss $\ell : \mathbb{R}^d \times \mathbb{R}^{d'} \to \mathbb{R}_+$, we associate to a parameterization $\mu \in \mathcal{P}_2^{\mathrm{Leb}}([0,1] \times \Theta)$ the training risk:

$$\mathcal{R}(\mu) \coloneqq\ \mathbb{E}_{x,y}\ell(\mathrm{NODE}_\mu(x), y) = \mathbb{E}_{x,y}\ell(x_\mu(1), y)\,. \tag{I.8}$$

In the following, we assume the data distribution $\mathcal{D}$ has compact support and ℓ is a smooth loss. The *risk minimization problem* Eq. (30) for the training of the mean-field NODE model thus reads:

$$\text{Find}\quad \mu^* \in \operatorname*{arg\,min}_{\mu \in \mathcal{P}_2^{\mathrm{Leb}}([0,1]\times\Theta)} \mathcal{R}(\mu)\,.$$

In practice, such an optimization problem is often solved using first order optimization algorithms such a *gradient descent* or *stochastic gradient descent*. In this chapter, we show such training dynamics can be modeled by a gradient flow w.r.t. an appropriate metric structure on the space of parameterizations.

I.1.3 Related works and contributions

Due to the popularity and performance of the ResNet architecture, many works have studied its training dynamics and their convergence properties.

Mean-field models of NODEs Some works have proposed models for ResNets of infinite depth similar to Definition I.1. E, Ma, and Wu [E, 2021] study properties of the functional space induced by considering the flow of functions of the form Eq. (I.1) and define a notion of norm which they use to provide bounds on the Rademacher complexity of this class of function. Chen et al. [Chen, 2023] also provide bounds on this Rademacher complexity which they use to prove an upper bound on the generalization error of trained ResNets.

Closer to our work are the works of Lu et al. [Lu, 2020] and Ding et al. [Ding, 2021; Ding, 2022] studying gradient flow dynamics for the minimization of the risk $\mathcal{R}$ for the ResNet model of Definition I.1. Lu et al. [Lu, 2020] consider gradient flows w.r.t. the true Wasserstein distance on the space of measures. While this point of view motivates a new training strategy, it is not consistent with the way ResNets are trained in practice, that is with a layer-wise-L^2 metric. Ding et al. [Ding, 2021; Ding, 2022] show existence and uniqueness of solutions for gradient flow equation similar to Definition I.3.

As a comparison, a key contribution of our work is to provide the parameter set with the appropriate metric structure allowing us to identify the gradient flow equation, derived formally by *adjoint sensitivity analysis* with a *curve of maximal slope* of the risk. Similarly, Isobe [Isobe, 2023] considers NODEs parameterized on the space of $\mathcal{P}_2(\Theta)$-valued functions equipped with a "L^2-Wasserstein" metric and trained with gradient flow. A notable difference is that [Isobe, 2023] considers adding a regularization term to the risk. This ensures the risk is a coercive function, which is not the case in our setting.

ResNets as a discretization of NODEs While it is not addressed in the present chapter, an interesting question is the one of the consistency of the NODE model with ResNets of finite depth. Marion et al. [Marion, 2023b] shows the convergence of ResNets of finite width towards NODEs, at initialization and during training, when the depth tends to infinity. This convergence is uniform over finite training time intervals but can be made uniform over the whole training dynamic under a convergence condition. For ResNets of arbitrary width, with layers of the form Eq. (I.1), Ding et al. [Ding, 2021; Ding, 2022] give a result of uniform convergence over finite training time intervals. Adding a regularization term, Thorpe and Gennip [Thorpe, 2023] show the Γ-convergence of the risk associated with ResNets to the one associated with NODEs.

Conditional Optimal transport In this chapter, we rely on the properties of the Conditional OT metric (Section I.2) to define a notion of gradient flow for the training of ResNets in the mean-field limit. Similar metrics have been used in recent works for other applications, for example Peszek and Poyato [Peszek, 2023] use gradient flow in the Conditional OT topology to study evolution PDEs with heterogeneities, Hosseini, Hsu, and Taghvaei [Hosseini, 2025] apply Conditional OT to the study of solutions to *Bayesian Inverse Problems*, Chemseddine et al. [Chemseddine, 2024] consider applications to *Bayesian Flow Matching* and Kerrigan, Migliorini, and Smyth [Kerrigan, 2024] consider applications to conditional generative modeling. Important for studying the gradient flow dynamics are the dynamical properties of the Conditional OT metric. Analogously to the Wasserstein case [Ambrosio, 2008b], we show that absolutely continuous curves

are solutions to certain continuity equations (Theorem I.1). Similar results were shown in [Peszek, 2023].

Contributions Our main contribution is to propose a model for ResNets of infinite depth and arbitrary width together with a metric space structure that is consistent with the layer-wise-L^2-metric used in practice when training ResNets with gradient descent and automatic differentiation. Our model thus allows a rigorous analysis of the training of ResNets at infinite depth and arbitrary width.

In detail, the ResNet model of Definition I.1 is parameterized over $\mathcal{P}_2^{\mathrm{Leb}}([0,1]\times\Theta)$ — the set of probability measures on $[0,1]\times\Theta$ whose first marginal is the Lebesgue measure on $[0,1]$ — which we equip in Section I.2 with a L^2-Wasserstein (or *Conditional Optimal Transport*) distance $\mathcal{W}_2^{\mathrm{COT}}$ (Proposition I.2.1). In Section I.3 we leverage results from the theory of gradient flows in metric spaces [Ambrosio, 2008b; Santambrogio, 2017] to define the gradient flow of the risk $\mathcal{R}$. This gradient flow equation corresponds to both notions of *curve of maximal slope* of the risk and the usual gradient flow of ResNets obtained by *adjoint sensitivity analysis* [Chen, 2018]. We conclude this part by showing well-posedness results for the gradient flow equation, that is existence in arbitrary time (Theorem I.3), uniqueness (Theorem I.4) and stability w.r.t. initialization (Theorem I.5). The study of the asymptotic behavior of such gradient flow curves will be the subject of Chapter II.

In addition to this, we study in Section I.2 properties of the space $\mathcal{P}_2^{\mathrm{Leb}}([0,1]\times\Theta)$ equipped with the Conditional OT distance $\mathcal{W}_2^{\mathrm{COT}}$. The literature on this subject being sparse, some of our results might be of their own interest. In particular, we provide in Theorem I.1 a characterization of absolutely continuous curves analogous to the one in the Wasserstein space [Ambrosio, 2008b, Thm.8.3.1].

I.2 Metric structure of the parameter set $\mathcal{P}_2^{\mathrm{Leb}}([0,1]\times\Theta)$

We define here a notion of distance $\mathcal{W}_2^{\mathrm{COT}}$ over the parameter set $\mathcal{P}_2^{\mathrm{Leb}}([0,1]\times\Theta)$ and study its properties. Importantly, the characterization of absolutely continuous curves in the metric space $(\mathcal{P}_2^{\mathrm{Leb}}([0,1]\times\Theta), \mathcal{W}_2^{\mathrm{COT}})$ will be used in Section I.3 to define the notion of gradient flow for the risk $\mathcal{R}$.

In the rest of this chapter as well as in Chapter II, we will assume for simplicity that Θ is the Euclidean space $\mathbb{R}^p$ for some $p \geq 1$. However, the presented results could probably be adapted to the case where Θ is a smooth manifold embedded in $\mathbb{R}^p$ or an (infinite dimensional) separable Hilbert space. In particular, we will extensively use the fact that Θ is a complete, separable metric space. We recall that the Wasserstein-2 distance $\mathcal{W}_2$ on the space $\mathcal{P}_2(\Theta)$ was defined in Eq. (51) as the optimal value of the Kantorovitch's optimal transport problem:

$$\forall \mu, \mu' \in \mathcal{P}_2(\Theta)\,, \quad \mathcal{W}_2(\mu,\mu') \coloneqq \min_{\gamma\in\Gamma(\mu,\mu')} \left(\int_{\Theta\times\Theta} \|\theta-\theta'\|^2 \mathrm{d}\gamma(\theta,\theta')\right)^{1/2}, \tag{I.9}$$

where $\Gamma(\mu,\mu')$ is the set of *couplings* between μ and μ', defined in Eq. (52). We denote by $\Gamma_o(\mu,\mu') \subset \Gamma(\mu,\mu')$ the subset of *optimal couplings* achieving the equality in Eq. (I.9). We refer to the books of Villani [Villani, 2009] and Santambrogio [Santambrogio, 2015] for further properties of the Wasserstein distance.

I.2.1 Conditional Optimal Transport distance

The Conditional Optimal Transport (COT) distance $\mathcal{W}_2^{\mathrm{COT}}$ is a modification of the Wasserstein distance $\mathcal{W}_2$ with the supplementary constraint that the transport plan should pre-

serve the marginal over $[0,1]$. This constraint is introduced to closely model the training dynamic of ResNets where the gradients are computed over the weights of each layer independently. For this purpose, it is natural to define a "layer-wise-L^2" Wasserstein distance, that is a L^2-distance over the set of families of probability measures in $\mathcal{P}_2(\Theta)$, indexed over $s \in [0,1]$.

Proposition I.2.1 (COT distance)**.** *Define for $\mu, \mu' \in \mathcal{P}_2^{\mathrm{Leb}}([0,1]\times\Theta)$:*

$$\mathcal{W}_2^{\mathrm{COT}}(\mu,\mu') \coloneqq \left(\int_0^1 \mathcal{W}_2\left(\mu(.|s),\mu'(.|s)\right)^2 \mathrm{d}s\right)^{1/2}.$$

Then, $\mathcal{W}_2^{\mathrm{COT}}$ defines a metric on $\mathcal{P}_2^{\mathrm{Leb}}([0,1]\times\Theta)$.

Proof. One essentially needs to justify the existence of the integral in the definition of $\mathcal{W}_2^{\mathrm{COT}}$. That $\mathcal{W}_2^{\mathrm{COT}}$ is a metric then follows from the properties of the Wasserstein and L^2 metrics respectively.

For Borel probability measures $\mu,\nu \in \mathcal{P}(\Theta)$ it is known [Villani, 2009, Thm.5.10] that the Monge-Kantorovitch problem admits the dual formulation:

$$\mathcal{W}_2(\mu,\nu)^2 = \sup\left\{\int_\Theta \varphi \mathrm{d}\mu + \int_\Theta \psi \mathrm{d}\nu\right\},$$

where the supremum is taken over all pairs $(\varphi,\psi) \in \mathcal{C}_b(\Theta)\times\mathcal{C}_b(\Theta)$ such that $\varphi(x)+\psi(y) \leq \|x-y\|^2$. We also have the alternative formulation:

$$\mathcal{W}_2(\mu,\nu)^2 = \sup_{\varphi\in\mathcal{C}_b(\Theta)}\left\{\int_\Theta \varphi \mathrm{d}\mu + \int_\Theta \varphi^c \mathrm{d}\nu\right\},$$

where for $\varphi:\Theta\to\mathbb{R}$ the c-transform φ^c of φ is defined as [Santambrogio, 2015, Def.1.10]:

$$\forall\theta\in\Theta, \quad \varphi^c(\theta) \coloneqq \inf_{\theta'\in\Theta} \|\theta'-\theta\|^2 - \varphi(\theta').$$

Consider $(\varphi_n)_{n\geq 0}$ a sequence of functions in $\mathcal{C}_b(\Theta)$ such that for any $\varphi\in\mathcal{C}_b(\Theta)$ we can find a subsequence $m(n)$ with $\varphi_{m(n)}\to\varphi$ for the compact-open topology (uniform convergence on compact subsets) and $\|\varphi_{m(n)}\|_\infty$ is uniformly bounded. Then we also have $\varphi^c_{m(n)}\to\varphi^c$ uniformly on compact subsets with $\|\varphi^c_{m(n)}\|_\infty \leq \|\varphi_{m(n)}\|_\infty$ uniformly bounded, whence:

$$\mathcal{W}_2(\mu,\nu)^2 = \sup_{n\in\mathbb{N}}\left\{\int_\Theta \varphi_n \mathrm{d}\mu + \int_\Theta \varphi_n^c \mathrm{d}\nu\right\}.$$

Thus, for $\mu,\mu'\in\mathcal{P}_2^{\mathrm{Leb}}([0,1]\times\Theta)$, the application $s\mapsto \mathcal{W}_2\left(\mu(.|s),\mu'(.|s)\right)^2$ is measurable as it can be expressed as the supremum of countably many measurable functions. □

Alternatively, the distance $\mathcal{W}_2^{\mathrm{COT}}$ can be viewed as an optimal transport distance with the additional constraint that the transport plans should be the identity on the first marginal. This new formulation is convenient for calculations and, in particular, allows easily estimating the distance $\mathcal{W}_2^{\mathrm{COT}}$ from above. Given $\mu,\mu'\in\mathcal{P}_2^{\mathrm{Leb}}([0,1]\times\Theta)$ we define:

$$\Gamma^{\mathrm{Leb}}(\mu,\mu') \coloneqq \left\{\gamma\in\mathcal{P}_2^{\mathrm{Leb}}([0,1]\times\Theta^2) \ : \ \gamma(.|s)\in\Gamma(\mu(.|s),\mu'(.|s)) \text{ for } \mathrm{d}s\text{-a.e. } s\in[0,1]\right\},$$
$$\Gamma^{\mathrm{diag}}(\mu,\mu') \coloneqq \left\{\gamma\in\Gamma(\mu,\mu') \ : \ \int f(s,s')\mathrm{d}\gamma(s,\theta,s',\theta') = \int_0^1 f(s,s)\mathrm{d}s\,,\ \forall f\in\mathcal{C}([0,1]^2)\right\}.$$

Note that these to sets are closely related as, if $\gamma \in \Gamma^{\mathrm{Leb}}(\mu,\mu')$, then

$$\tilde{\gamma} \coloneqq (\pi^1,\pi^2,\pi^1,\pi^3)_{\#}\gamma \in \Gamma^{\mathrm{diag}}(\mu,\mu')$$

and conversely, if $\tilde{\gamma} \in \Gamma^{\mathrm{diag}}(\mu,\mu')$, then

$$\gamma \coloneqq (\pi^1,\pi^2,\pi^4)_{\#}\tilde{\gamma} \in \Gamma^{\mathrm{Leb}}(\mu,\mu')\,.$$

In both cases we have for any measurable $f:\Theta^2\to\mathbb{R}$:

$$\int_{[0,1]\times\Theta^2} f(\theta,\theta')\mathrm{d}\gamma(s,\theta,\theta') = \int_{([0,1]\times\Theta)^2} f(\theta,\theta')\mathrm{d}\tilde{\gamma}(s,\theta,s',\theta')\,. \tag{I.10}$$

In the same way the Wasserstein distance $\mathcal{W}_2(\mu,\mu')$ can be obtained of the solution of a minimization problem over the set $\Gamma(\mu,\mu')$ (Eq. (I.9)), the COT distance $\mathcal{W}_2^{\mathrm{COT}}$ can be obtained as the solution of minimization problems over the sets $\Gamma^{\mathrm{Leb}}(\mu,\mu')$ and $\Gamma^{\mathrm{diag}}(\mu,\mu')$.

Proposition I.2.2. *Let $\mu,\mu' \in \mathcal{P}_2^{\mathrm{Leb}}([0,1]\times\Theta)$ then:*

$$\begin{aligned}\mathcal{W}_2^{\mathrm{COT}}(\mu,\mu')^2 &= \min_{\gamma\in\Gamma^{\mathrm{Leb}}(\mu,\mu')}\int_{[0,1]\times\Theta^2}\|\theta-\theta'\|^2\mathrm{d}\gamma(s,\theta,\theta')\\ &= \min_{\gamma\in\Gamma^{\mathrm{diag}}(\mu,\mu')}\int_{([0,1]\times\Theta)^2}\|\theta-\theta'\|^2\mathrm{d}\gamma(s,\theta,s',\theta')\,.\end{aligned}$$

We denote respectively by $\Gamma_o^{\mathrm{Leb}}(\mu,\mu')$ and $\Gamma_o^{\mathrm{diag}}(\mu,\mu')$ the set of optimal couplings in both minimization problems. Then for $\gamma\in\Gamma_o^{\mathrm{Leb}}(\mu,\mu')$ we have for $\mathrm{d}s$-a.e. $s\in[0,1]$:

$$\gamma(.|s)\in\Gamma_o(\mu(.|s),\mu'(.|s))\,,\quad \textit{i.e.}\quad \int_{\Theta^2}\|\theta-\theta'\|^2\mathrm{d}\gamma(\theta,\theta'|s) = \mathcal{W}_2(\mu(.|s),\mu'(.|s))^2\,.$$

Proof. Our proof technique is similar to the one of [Hosseini, 2025, Prop.3.3] and relies on the possibility of choosing an optimal transport plan $\gamma(.|s)\in\Gamma_o(\mu(.|s),\mu'(.|s))$ for every $s\in[0,1]$ in a measurable way.

We show equality with the first minimization problem on $\Gamma^{\mathrm{Leb}}(\mu,\mu')$, equality between the two minimization problems then comes from Eq. (I.10). Assume there exists a Borel map $\gamma: s\mapsto\gamma(.|s)\in\mathcal{P}(\Theta^2)$ (where $\mathcal{P}(\Theta^2)$ is equipped with the narrow topology) such that $\gamma(.|s)\in\Gamma_o(\mu(.|s),\mu'(.|s))$ for every $s\in[0,1]$. With such a map, one can define a Borel probability measure on $[0,1]\times\Theta$, that we also denote by γ, which is the measure whose disintegration w.r.t. the Lebesgue measure on $[0,1]$ is $\{\gamma(.|s)\}_{s\in[0,1]}$. In other words, the measure γ is defined by:

$$\int_{[0,1]\times\Theta^2} f(s,\theta,\theta')\mathrm{d}\gamma(s,\theta,\theta') \coloneqq \int_0^1\int_{\Theta^2} f(s,\theta,\theta')\mathrm{d}\gamma(\theta,\theta'|s)\mathrm{d}s\,,\quad \forall f\in\mathcal{C}_b([0,1]\times\Theta^2)\,.$$

Such γ will be a solution to our first optimization problem as we have:

$$\int_{[0,1]\times\Theta^2}\|\theta-\theta'\|^2\mathrm{d}\gamma(s,\theta,\theta') = \mathcal{W}_2^{\mathrm{COT}}(\mu,\mu')^2 \le \inf_{\gamma\in\Gamma^{\mathrm{Leb}}(\mu,\mu')}\int_{[0,1]\times\Theta^2}\|\theta-\theta'\|^2\mathrm{d}\gamma(s,\theta,\theta')\,.$$

To show the existence of such γ we use a measurable selection result, that is considering the set-valued mapping $s\in[0,1]\mapsto\Gamma_o(\mu(.|s),\mu'(.|s))\subset\mathcal{P}(\Theta^2)$ we show it admits a measurable section. Consider the set:

$$\mathcal{G}^* \coloneqq \{(s,\gamma)\ :\ \gamma\in\Gamma_o(\mu(.|s),\mu'(.|s))\}\subset[0,1]\times\mathcal{P}_2(\Theta^2)\,.$$

Using [Bogachev, 2007, Thm.6.9.6], as for every $s \in [0,1]$ the set $\Gamma_o(\mu(.|s), \mu'(.|s))$ is narrowly compact, it is sufficient to show that $\mathcal{G}^* \in \mathcal{B}([0,1] \times \mathcal{P}_2(\Theta^2))$. Let $\{f_n\}_{n\in\mathbb{N}}$ be dense in $\mathcal{C}_b(\Theta)$ for the compact-open topology. Then, for every $n \in \mathbb{N}$ as the mapping $(s,\gamma) \mapsto \int_{\Theta^2} f_n \mathrm{d}\gamma - \int_\Theta f_n \mathrm{d}\mu(.|s)$ is measurable, so are the sets:

$$\mathcal{G}_n \coloneqq \left\{(s,\gamma) \ : \ \int_{\Theta^2} f_n(\theta)\mathrm{d}\gamma(\theta,\theta') = \int_\Theta f_n(\theta)\mathrm{d}\mu(\theta|s)\right\},$$
$$\mathcal{G}'_n \coloneqq \left\{(s,\gamma) \ : \ \int_{\Theta^2} f_n(\theta')\mathrm{d}\gamma(\theta,\theta') = \int_\Theta f_n(\theta')\mathrm{d}\mu'(\theta'|s)\right\}.$$

Also, as the mapping $(s,\gamma) \mapsto \mathcal{W}_2(\mu(.|s), \mu'(.|s))^2 - \int_{\Theta^2} \|\theta-\theta'\|^2 \mathrm{d}\gamma$ is measurable by Proposition I.2.1, so is the set:

$$\mathcal{G}_o \coloneqq \left\{(s,\gamma) \ : \ \int_{\Theta^2} \|\theta-\theta'\|^2 \mathrm{d}\gamma = \mathcal{W}_2(\mu(.|s), \mu'(.|s))^2\right\} \in \mathcal{B}([0,1] \times \mathcal{P}_2(\Theta^2)).$$

Finally we have that $\mathcal{G}^* = \mathcal{G}_o \cap (\bigcap_{n\in\mathbb{N}} \mathcal{G}_n \cap \mathcal{G}'_n)$ is a Borel set, which completes the proof. □

Remark I.2.1 (Comparison of Wasserstein and Conditional-Wasserstein topologies)**.** *Note that, for $\mu, \mu' \in \mathcal{P}_2^{\mathrm{Leb}}([0,1] \times \Theta)$, we have that $\Gamma^{\mathrm{diag}}(\mu,\mu') \subset \Gamma(\mu,\mu')$. Hence from the previous result, it follows:*

$$\mathcal{W}_2(\mu,\mu') \leq \mathcal{W}_2^{\mathrm{COT}}(\mu,\mu')$$

and the topology induced by $\mathcal{W}_2^{\mathrm{COT}}$ on $\mathcal{P}_2^{\mathrm{Leb}}([0,1] \times \Theta)$ is stronger than the Wasserstein topology. It is in fact strictly stronger and, for example, the sequence $\mu_n = \int_0^1 \delta_{(-1)^{\lfloor 2ns \rfloor}} \mathrm{d}s$ and the measure $\mu = \frac{1}{2}\int_0^1 (\delta_1 + \delta_{-1})\mathrm{d}s$ in $\mathcal{P}_2^{\mathrm{Leb}}([0,1] \times \mathbb{R})$ are such that $\mathcal{W}_2(\mu_n,\mu) \to 0$ but $\mathcal{W}_2^{\mathrm{COT}}(\mu_n,\mu) \geq 1$.

The following result states that the metric space $(\mathcal{P}_2^{\mathrm{Leb}}([0,1] \times \Theta), \mathcal{W}_2^{\mathrm{COT}})$ is complete.

Proposition I.2.3 (Completeness)**.** $(\mathcal{P}_2^{\mathrm{Leb}}([0,1] \times \Theta), \mathcal{W}_2^{\mathrm{COT}})$ *is a complete metric space.*

Proof. The proof is analogous to the proof of completeness of the Wasserstein space $\mathcal{P}_2([0,1] \times \Theta)$ (see [Villani, 2009, Thm.6.18]).

Let $(\mu_n)_{n\geq 0}$ be a Cauchy sequence in $\mathcal{P}_2^{\mathrm{Leb}}([0,1] \times \Theta)$. Then, since the $\mathcal{W}_2^{\mathrm{COT}}$-topology is stronger than the (complete) $\mathcal{W}_2$-topology (cf. Remark I.2.1) and since narrow convergence preserves the marginal condition, such a sequence narrowly converges to some $\mu_\infty \in \mathcal{P}_2^{\mathrm{Leb}}([0,1] \times \Theta)$. Then by narrow lower semicontinuity of $\mathcal{W}_2^{\mathrm{COT}}$ (Lemma I.2.1) we have for every $n \geq 0$:

$$\mathcal{W}_2^{\mathrm{COT}}(\mu_\infty,\mu_n) \leq \liminf_{m\to\infty} \mathcal{W}_2^{\mathrm{COT}}(\mu_m,\mu_n),$$

and by taking the lim sup w.r.t. $n \geq 0$:

$$\limsup_{n\to\infty} \mathcal{W}_2^{\mathrm{COT}}(\mu_\infty,\mu_n) \leq \limsup_{\substack{m\to\infty \\ n\to\infty}} \mathcal{W}_2^{\mathrm{COT}}(\mu_m,\mu_n) = 0.$$

Hence (μ_n) $\mathcal{W}_2^{\mathrm{COT}}$-converges to μ_∞. □

Lemma I.2.1 (narrow lower semicontinuity of $\mathcal{W}_2^{\mathrm{COT}}$)**.** *Let $(\mu_n)_{n\geq 0}$ and $(\nu_n)_{n\geq 0}$ be sequences in $\mathcal{P}_2^{\mathrm{Leb}}([0,1] \times \Theta)$ such that $(\mu_n,\nu_n) \xrightarrow[n\to\infty]{} (\mu,\nu)$ narrowly for some $\mu,\nu \in \mathcal{P}_2^{\mathrm{Leb}}([0,1] \times \Theta)$. Then:*

$$\mathcal{W}_2^{\mathrm{COT}}(\mu,\nu) \leq \liminf_{n\to\infty} \mathcal{W}_2^{\mathrm{COT}}(\mu_n,\nu_n).$$

Proof. Up to extraction of a subsequence one can assume:

$$\mathcal{W}_2^{\mathrm{COT}}(\mu_n,\nu_n) \xrightarrow{n\to+\infty} \liminf_{n\to+\infty} \mathcal{W}_2^{\mathrm{COT}}(\mu_n,\nu_n)\,.$$

Then for every $n \geq 0$ consider some $\gamma_n \in \Gamma_o^{\mathrm{diag}}(\mu_n,\nu_n)$. In particular $\gamma_n \in \Gamma(\mu_n,\nu_n)$ and by [Villani, 2009, Lem.4.4] the sequence (γ_n) is tight. Hence it admits a subsequence $(\gamma_{n_k})_{k\geq 0}$ narrowly converging to some γ which is in $\Gamma^{\mathrm{diag}}(\mu,\nu)$ by the properties of narrow convergence. Thus applying [Villani, 2009, Lem.4.3] and using the characterization of $\mathcal{W}_2^{\mathrm{COT}}$ in Proposition I.2.2:

$$\begin{aligned}
\mathcal{W}_2^{\mathrm{COT}}(\mu,\nu)^2 &\leq \int_{([0,1]\times\Theta)^2} \|\theta-\theta'\|^2 \mathrm{d}\gamma \\
&\leq \liminf_{k\to\infty} \int_{([0,1]\times\Theta)^2} \|\theta-\theta'\|^2 \mathrm{d}\gamma_{n_k} \\
&= \liminf_{n\to\infty} \mathcal{W}_2^{\mathrm{COT}}(\mu_n,\nu_n)^2\,,
\end{aligned}$$

from which the result follows. □

I.2.2 Dynamical formulation of Conditional Optimal Transport

We analyze here the properties of absolutely continuous curves in $\mathcal{P}_2^{\mathrm{Leb}}([0,1]\times\Theta)$ when equipped with the COT metric. Similarly to the classical Wasserstein metric, we show that absolutely continuous curves obey a certain continuity equation. This characterization will be crucial to define the gradient flow equation used in the training of our NODE model.

Absolutely continuous curves in the Wasserstein space For $T > 0$, consider $I = (0,T)$ an open interval and $(\mu_t)_{t\in I}$ a family of probability measures over the Euclidean space $\mathbb{R}^p$. Given a Borel velocity field $v : (t,x) \in I\times\mathbb{R}^p \mapsto v_t(x) \in \mathbb{R}^p$ such that $\int_I \int_{\mathbb{R}^p} \|v_t\| \mathrm{d}\mu_t \mathrm{d}t < \infty$, we say that $(\mu_t)_{t\in I}$ satisfies the continuity equation $\partial_t\mu_t + \mathrm{div}(v_t\mu_t)$ in the weak sense if:

$$\int_I \int_{\mathbb{R}^p} \left(\partial_t\varphi(t,x) + \langle \nabla\varphi(t,x), v_t(x)\rangle\right) \mathrm{d}\mu_t(x)\mathrm{d}t = 0\,, \quad \forall \varphi \in \mathcal{C}_c^1(I\times\mathbb{R}^p)\,. \tag{I.11}$$

Equivalently ([Santambrogio, 2015, Prop.4.2]), when the mapping $t \mapsto \mu_t$ is narrowly continuous, this amounts to have that for every $\varphi \in \mathcal{C}_c^1(\mathbb{R}^p)$ the map $t \mapsto \mu_t(\varphi) \coloneqq \int \varphi \mathrm{d}\mu_t$ is absolutely continuous and verifies:

$$\frac{\mathrm{d}}{\mathrm{d}t}\mu_t(\varphi) = \int \langle \nabla\varphi, v_t\rangle \,\mathrm{d}\mu_t, \quad \text{for } \mathrm{d}t\text{-a.e. } t \in I.$$

An important property of the Wasserstein space $\mathcal{P}_2(\mathbb{R}^p)$ endowed with the distance $\mathcal{W}_2$ is the characterization of absolutely continuous curves: a narrowly continuous curve $(\mu_t)_{t\in I}$ is absolutely continuous in $\mathcal{P}_2(\mathbb{R}^p)$ if and only if it is solution to the continuity equation Eq. (I.11) for some velocity field v with $\int_I \|v_t\|_{L^2(\mu_t)}\mathrm{d}t < \infty$ [Ambrosio, 2008b, Thm.8.3.1]. We refer to the book by Ambrosio, Gigli, and Savaré [Ambrosio, 2008b] for a detailed study of absolutely continuous curves in $(\mathcal{P}_2(\mathbb{R}^p), \mathcal{W}_2)$.

Absolutely continuous curves int the Conditional Wasserstein space Similarly to the characterization of absolutely continuous curves in the Wasserstein space $\mathcal{P}_2(\mathbb{R}^p)$, an adaptation of [Ambrosio, 2008b, Thm.8.3.1] provides an analogous characterization of absolutely continuous curves in $\mathcal{P}_2^{\mathrm{Leb}}([0,1]\times\Theta)$, equipped with the Conditional OT

distance $\mathcal{W}_2^{\mathrm{COT}}$. This characterization allows us to (formally) provide the metric space $(\mathcal{P}_2^{\mathrm{Leb}}([0,1]\times\Theta), \mathcal{W}_2^{\mathrm{COT}})$ with a kind of " differential structure" by seeing tangent vectors as velocity fields. This identification will be crucial for defining the gradient flow equation for the training risk $\mathcal{R}$, which will take the form of a continuity equation with an appropriate velocity field (Definition I.3).

Theorem I.1 (adapted from [Ambrosio, 2008b], Thm.8.3.1)**.** *Assume $\Theta = \mathbb{R}^p$. Let $I = (0,T)$ for some $T > 0$ and $(\mu_t)_{t\in I}$ an absolutely continuous curve in $\mathcal{P}_2^{\mathrm{Leb}}([0,1]\times\Theta)$. Then there exists a unique Borel velocity field $v : (t,s,\theta) \in I \times [0,1] \times \Theta \mapsto v_t(s,\theta) \in \Theta$ such that for a.e. $t \in I$:*

$$v_t \in L^2(\mu_t)\,, \quad \|v_t\|_{L^2(\mu_t)} \leq \left|\frac{\mathrm{d}}{\mathrm{d}t}\mu_t\right|\,,$$

and μ is a weak solution of the continuity equation:

$$\partial_t \mu_t + \mathrm{div}((0,v_t)\mu_t) = 0 \quad \text{on } I\times[0,1]\times\Theta. \tag{I.12}$$

We will refer to such v_t as the tangent velocity field *of the curve $(\mu_t)_{t\in I}$. Conversely, if $(\mu_t)_{t\in I}$ is a narrowly continuous curve satisfying Eq. (I.12) for some Borel velocity field v_t with $\|v_t\|_{L^2(\mu_t)} \in L^1(I)$, then $(\mu_t)_{t\in I}$ is absolutely continuous and $|\frac{\mathrm{d}}{\mathrm{d}t}\mu_t| \leq \|v_t\|_{L^2(\mu_t)}$ for a.e. $t \in I$.*

Proof. *Part 1: AC curve $\Rightarrow$ Continuity equation.*

Note that this part is the easiest as $\mathcal{W}_2^{\mathrm{COT}}$-absolute continuity implies $\mathcal{W}_2$-absolute continuity for which the result is well-known, originally proven in [Ambrosio, 2008b, Thm.8.3.1]. Therefore we here only adapt this proof to our specific setting.

Up to reparameterization, one can assume w.l.o.g. that $\left|\frac{\mathrm{d}}{\mathrm{d}t}\mu_t\right| \in L^\infty(I)$. First we show that, for $\varphi \in \mathcal{C}_c^1(I\times[0,1]\times\Theta)$, the map $t \mapsto \mu_t(\varphi) \coloneqq \int_0^1 \int_\Theta \varphi \mathrm{d}\mu_t$ is absolutely continuous. Indeed, for $t,u \in I$, introducing a coupling $\gamma_{t,u} \in \Gamma_o^{\mathrm{Leb}}(\mu_t,\mu_u)$ we have:

$$|\mu_t(\varphi) - \mu_u(\varphi)| \leq \left|\int_0^1 \int_{\Theta^2} (\varphi(s,\theta) - \varphi(s,\theta'))\mathrm{d}\gamma_{t,u}(s,\theta,\theta')\right| \leq \|\nabla_\theta\varphi\|_\infty \mathcal{W}_2^{\mathrm{COT}}(\mu_t,\mu_u)\,,$$

from which absolute continuity follows. Then considering the map:

$$H(s,\theta,\theta') \coloneqq \begin{cases} \|\nabla_\theta\varphi(s,\theta)\| & \text{if } \theta = \theta', \\ \frac{|\varphi(s,\theta)-\varphi(s,\theta')|}{\|\theta-\theta'\|} & \text{else,} \end{cases}$$

we have for every $t,u \in I$:

$$\begin{aligned} \frac{|\mu_t(\varphi)-\mu_u(\varphi)|}{|t-u|} &\leq \frac{1}{|t-u|}\int_0^1\int_{\Theta^2} \|\theta-\theta'\| H(s,\theta,\theta')\mathrm{d}\gamma_{t,u}(s,\theta,\theta') \\ &\leq \frac{\mathcal{W}_2^{\mathrm{COT}}(\mu_t,\mu_u)}{|t-u|}\|H\|_{L^2(\gamma_{t,u})}\,. \end{aligned}$$

As $u \to t$, we have $\mathcal{W}_2^{\mathrm{COT}}(\mu_u,\mu_t) \to 0$ and by the properties of L^2 spaces [Cannarsa, 2015, Prop.3.11] we can take a sequence $u_n \to t$ such that $\mathcal{W}_2(\mu_{u_n}(.|s),\mu_t(.|s)) \to 0$ for ds-a.e. $s \in [0,1]$. This implies for those $s \in [0,1]$ that $\mu_{u_n}(.|s) \to \mu_t(.|s)$ narrowly and that $\gamma_{t,u_n}(.|s) \to \gamma(.|s) \in \Gamma_o(\mu_t(.|s),\mu_t(.|s))$, i.e. the trivial transport plan $\gamma(.|s) =$

$(\mathrm{Id},\mathrm{Id})_\#\mu_t(.|s)$. Thus we have that $\gamma_{t,u_n}\to(\pi^1,\pi^2,\pi^2)_\#\mu_t$ narrowly since, by Lebesgue's theorem, given a bounded continuous function $f\in\mathcal{C}_b([0,1]\times\Theta\times\Theta)$:

$$\int f\mathrm{d}\gamma_{t,u_n}=\int_0^1\left(\int_{\Theta^2}f(s,\theta,\theta')\mathrm{d}\gamma_{t,u_n}(\theta,\theta'|s)\right)\mathrm{d}s\xrightarrow[n\to\infty]{}\int_0^1\left(\int_\Theta f(s,\theta,\theta)\mathrm{d}\mu_t(\theta|s)\right)\mathrm{d}s\,.$$

Hence, at a point where $t\mapsto\mu_t$ is metrically differentiable:

$$\limsup_{u\to t}\frac{|\mu_t(\varphi)-\mu_u(\varphi)|}{|t-u|}\le\left|\frac{\mathrm{d}}{\mathrm{d}t}\mu_t\right|\|\nabla_\theta\varphi\|_{L^2(\mu_t)}\,.$$

Consider $\boldsymbol{\mu}=\int_I\mu_t\mathrm{d}t\in\mathcal{P}(I\times[0,1]\times\Theta)$ the measure whose disintegration w.r.t. Lebesgue's measure on I is $(\mu_t)_{t\in I}$. Then for $\varphi\in\mathcal{C}_c^1(I\times[0,1]\times\Theta)$ we have:

$$\begin{aligned}\int_I\int_{[0,1]\times\Theta}&\partial_t\varphi(t,s,\theta)\mathrm{d}\mu_t(s,\theta)\mathrm{d}t\\&=\lim_{h\to0}\int_I\int_{[0,1]\times\Theta}\frac{\varphi(t,s,\theta)-\varphi(t-h,s,\theta)}{h}\mathrm{d}\mu_t(s,\theta)\mathrm{d}t\\&=\lim_{h\to0}\int_I\frac{1}{h}\left(\int_{[0,1]\times\Theta}\varphi(t,s,\theta)\mathrm{d}\mu_t(s,\theta)-\int_{[0,1]\times\Theta}\varphi(t,s,\theta)\mathrm{d}\mu_{t-h}(s,\theta)\right)\mathrm{d}t\,.\end{aligned}$$

Thus by the previous inequality and Fatou's lemma:

$$\left|\int_I\int_{[0,1]\times\Theta}\partial_t\varphi(t,s,\theta)\mathrm{d}\mu_t(s,\theta)\mathrm{d}t\right|\le\left(\int_I\left|\frac{\mathrm{d}}{\mathrm{d}t}\mu_t\right|^2\mathrm{d}t\right)^{1/2}\left(\int_{I\times[0,1]\times\Theta}\|\nabla_\theta\varphi(t,s,\theta)\|^2\mathrm{d}\boldsymbol{\mu}(t,s,\theta)\right)^{1/2}.$$

Consider the subspace $V\coloneqq\{\nabla_\theta\varphi\ :\ \varphi\in\mathcal{C}_c^1(I\times[0,1]\times\Theta)\}$ and let $\mathcal{V}$ be its closure in $L^2(I\times[0,1]\times\Theta,\boldsymbol{\mu})$. Then by the previous inequality the linear functional $\mathcal{A}:V\to\mathbb{R}$ defined by:

$$\mathcal{A}(\nabla_\theta\varphi)\coloneqq-\int_{I\times[0,1]\times\Theta}\partial_t\varphi(t,s,\theta)\mathrm{d}\boldsymbol{\mu}(t,s,\theta)$$

is continuous on V and thus, by Hahn-Banach's theorem, can be extended to a unique continuous linear functional on $\mathcal{V}$. Therefore, by Lax-Milgram's theorem, the minimization problem

$$\min\left\{\frac{1}{2}\int_{I\times\mathbb{R}^{p+1}}\|w(t,s,\theta)\|^2\mathrm{d}\boldsymbol{\mu}(t,s,\theta)-\mathcal{A}(w)\ :\ w\in\mathcal{V}\right\}$$

admits a unique solution $v\in\mathcal{V}$ which is characterized by the property that:

$$\int_{I\times\mathbb{R}^{p+1}}\langle v(t,s,\theta),\nabla_\theta\varphi(t,s,\theta)\rangle\,\mathrm{d}\boldsymbol{\mu}(t,s,\theta)=\mathcal{A}(\nabla_\theta\varphi)\,,\quad\forall\varphi\in\mathcal{C}_c^1(I\times[0,1]\times\Theta)\,.$$

This is the desired continuity equation by definition of $\mathcal{A}$.

Finally, let $(\nabla_\theta\varphi_n)\subset V$ be a sequence converging to $v\in L^2(\boldsymbol{\mu})$. Considering an interval $J\subset I$ and some $\eta\in\mathcal{C}_c^1(J)$ with $0\le\eta\le1$ we have by the previous arguments:

$$\begin{aligned}\int_{I\times[0,1]\times\Theta}\eta(t)\|v(t,s,\theta)\|^2\mathrm{d}\boldsymbol{\mu}(t,s,\theta)&=\lim_{n\to\infty}\int_{I\times[0,1]\times\Theta}\eta\,\langle v,\nabla_\theta\varphi_n\rangle\,\mathrm{d}\boldsymbol{\mu}\\&=\lim_{n\to\infty}\mathcal{A}(\nabla_\theta(\eta\varphi_n))\\&\le\left(\int_J\left|\frac{\mathrm{d}}{\mathrm{d}t}\mu_t\right|^2\mathrm{d}t\right)^{1/2}\lim_{n\to\infty}\left(\int_{J\times[0,1]\times\Theta}\|\nabla_\theta\varphi_n\|^2\mathrm{d}\boldsymbol{\mu}\right)^{1/2}\\&=\left(\int_J\left|\frac{\mathrm{d}}{\mathrm{d}t}\mu_t\right|^2\mathrm{d}t\right)^{1/2}\left(\int_{J\times[0,1]\times\Theta}\|v\|^2\mathrm{d}\boldsymbol{\mu}\right)^{1/2}.\end{aligned}$$

Hence approximating the characteristic function of J with such an η we get:

$$\int_J \int_{[0,1]\times\Theta} \|v_t\|^2 \mathrm{d}\mu_t \mathrm{d}t \leq \int_J \left|\frac{\mathrm{d}}{\mathrm{d}t}\mu_t\right|^2 \mathrm{d}t\,,$$

implying $\|v_t\|_{L^2(\mu_t)} \leq \left|\frac{\mathrm{d}}{\mathrm{d}t}\mu_t\right|$ for a.e. $t \in I$.

Part 2: Continuity equation ⇒ AC curve.

This part of the proof is new as, according to [Ambrosio, 2008b, Thm.8.3.1], the continuity equation only ensures $\mathcal{W}_2$-absolute continuity, which is strictly weaker that $\mathcal{W}_2^{\mathrm{COT}}$-continuity as explained in Remark I.2.1. We show here that the specific form of the velocity field ensures $\mathcal{W}_2^{\mathrm{COT}}$-absolute continuity.

For $(t,s) \in I \times [0,1]$, we denote by $v_{t,s}$ the Borel vector field $v_{t,s} : \theta \in \Theta \mapsto v_t(s,\theta)$. Note that by Jensen's inequality:

$$\int_I \int_0^1 \|v_{t,s}\|_{L^2(\mu_t(.|s))} \mathrm{d}s\mathrm{d}t \leq \int_I \|v_t\|_{L^2(\mu_t)} \mathrm{d}t < +\infty\,,$$

and we have that for ds-a.e. $s \in [0,1]$, $t \mapsto \|v_{t,s}\|_{L^2(\mu_t(.|s))} \in L^1(I)$. Also if $\varphi \in \mathcal{C}_c^1(I\times\Theta)$ and $\chi \in \mathcal{C}_c^1([0,1])$ then by definition of the continuity equation:

$$\int_I \int_0^1 \int_\Theta \left(\partial_t \varphi + \langle \nabla_\theta \varphi, v_{t,s}\rangle\right) \chi(s) \mathrm{d}\mu_t(.|s) \mathrm{d}s\mathrm{d}t = 0\,.$$

Hence if $J \subset [0,1]$ is an interval, approaching the characteristic function of J with χ we get:

$$\int_I \int_J \int_\Theta \left(\partial_t \varphi + \langle \nabla_\theta \varphi, v_{t,s}\rangle\right) \mathrm{d}\mu_t(.|s) \mathrm{d}s\mathrm{d}t = 0\,,$$

and hence for ds-a.e. $s \in [0,1]$:

$$\int_I \int_\Theta \left(\partial_t \varphi + \langle \nabla_\theta \varphi, v_{t,s}\rangle\right) \mathrm{d}\mu_t(.|s) \mathrm{d}t = 0\,.$$

Now if we consider (φ_n) a countable dense sequence in $\mathcal{C}_c^1(I\times\Theta)$ endowed with the usual topology then we can find a set $\Lambda \subset [0,1]$ of full Lebesgue's measure such that for every $s \in \Lambda$ the above equation holds for every test function $\varphi \in \mathcal{C}_c^1(I\times\Theta)$. In other words we have shown that, for ds-a.e. $s \in [0,1]$, $\mu_t(.|s)$ solves the continuity equation:

$$\partial_t \mu_t(.|s) + \mathrm{div}(v_{t,s}\mu_t(.|s)) = 0 \quad \text{on } I\times\Theta.$$

Note that, without loss of generality, we can consider the curve $(\mu_t(.|s))_{t\in I}$ to be narrowly continuous. Indeed, as it is a solution of the continuity equation we know that the curve $(\mu_t(.|s))_{t\in I}$ admits a narrowly continuous representative $\tilde{\mu}_t(.|s)$ [Ambrosio, 2008b, Lem.8.1.2] and that this representative is characterized by that for every $\varphi \in \mathcal{C}_c^1(\Theta)$ and every $t \in I$:

$$\tilde{\mu}_t(.|s)(\varphi) = \int_0^t \int_\Theta \left(\chi'(u)\varphi + \chi(u)\left\langle \nabla_\theta \varphi, v_{u,s}\right\rangle\right) \mathrm{d}\mu_u(.|s) \mathrm{d}u\,,$$

where $\chi \in \mathcal{C}^1(I)$ is any function such that $\chi = 0$ on a neighbourhood of 0 and $\chi(u) = 1$ for $u \geq t$ (the definition does not depend on χ by definition of the continuity equation). Then it follows that for any $t \in I$ and any $f \in \mathcal{C}_b([0,1]\times\Theta)$ the map $s \mapsto \tilde{\mu}_t(.|s)(f)$ is measurable

and integrating w.r.t. s we get that $\tilde{\mu}_t := \int_0^1 \tilde{\mu}_t(.|s)\mathrm{d}s$ defines a probability measure over $[0,1]\times\Theta$ whose disintegration is $\{\tilde{\mu}_t(.|s)\}_{s\in[0,1]}$. Moreover, for any $\varphi \in \mathcal{C}_c^1([0,1]\times\Theta)$ we have:

$$\tilde{\mu}_t(\varphi) = \int_0^t \int_0^1 \int_\Theta \left(\chi'(u)\varphi + \chi(u)\left\langle \nabla_\theta \varphi, v_{u,s}\right\rangle\right) \mathrm{d}\mu_u(.|s)\mathrm{d}s\mathrm{d}u = \mu_t(\varphi)$$

and hence in fact the equality $\tilde{\mu}_t = \mu_t$.

Then, using [Ambrosio, 2008b, Thm.8.3.1] with the assumption that $t \mapsto \mu_t(.|s)$ is narrowly continuous, we have that for $\mathrm{d}s$-a.e. $s \in [0,1]$ the curve $t \in I \mapsto \mu_t(.|s)$ is absolutely continuous and

$$\mathcal{W}_2(\mu_{t_1}(.|s), \mu_{t_2}(.|s))^2 \leq (t_2 - t_1)\int_{t_1}^{t_2}\int_\Theta \|v_{t,s}\|^2 \mathrm{d}\mu_t(.|s)\mathrm{d}t\,, \quad \forall t_1 < t_2 \in I\,.$$

Hence integrating w.r.t. $s \in [0,1]$ gives:

$$\mathcal{W}_2^{\mathrm{COT}}(\mu_{t_1}, \mu_{t_2})^2 \leq (t_2 - t_1)\int_{t_1}^{t_2}\int_{[0,1]\times\Theta} \|v_t\|^2 \mathrm{d}\mu_t\mathrm{d}t\,, \quad \forall t_1 < t_2 \in I\,,$$

showing that $(\mu_t)_{t\in I}$ is $\mathcal{W}_2^{\mathrm{COT}}$-absolutely continuous and $\left|\frac{\mathrm{d}}{\mathrm{d}t}\mu_t\right| \leq \|v_t\|_{L^2(\mu_t)}$ for a.e. $t \in I$. □

Remark I.2.2. *Note that to study absolutely continuous curves, we introduce the supplementary time variable $t \geq 0$. This time variable will model the optimization time in the Definition I.3 of the gradient flow equation. It is not to be interverted with the NODE flow time $s \in [0,1]$.*

As a consequence of Theorem I.1 we recover two useful results about absolutely continuous curves in $\mathcal{P}_2^{\mathrm{Leb}}([0,1]\times\Theta)$. Those are stated in the following Lemmas I.2.2 and I.2.3. The first result is a result of approximation along absolutely continuous curves. It states that the tangent velocity field $(v_t)_{t\in I}$ defined in Theorem I.1 indeed furnishes a first-order approximation of the curve $(\mu_t)_{t\in I}$ at every time $t \in I$. It will be particularly useful to differentiate quantities related to μ_t (Corollaries I.3.2 and I.3.3). The second result is an application and gives the differential of the square-distance $\mathcal{W}_2^{\mathrm{COT}}(\mu_t, \mu')^2$ along an absolutely continuous curve $(\mu_t)_{t\in I}$.

Lemma I.2.2 (Adapted from [Ambrosio, 2008b, Prop.8.4.6])**.** *Let $(\mu_t)_{t\in I}$ be an absolutely continuous curve in $(\mathcal{P}_2^{\mathrm{Leb}}([0,1]\times\Theta), \mathcal{W}_2^{\mathrm{COT}})$ and let $v : I\times\mathbb{R}^{p+1} \to \mathbb{R}^p$ be the unique velocity field satisfying the conclusions of Theorem I.1. Then for $\mathrm{d}t$-a.e. $t \in I$ it holds that for any choice of $\gamma_t^h \in \Gamma_o^{\mathrm{Leb}}(\mu_{t+h}, \mu_t)$:*

$$\lim_{h\to 0}\left(\pi^1, \pi^2, \frac{1}{h}(\pi^3 - \pi^2)\right)_{\#}\gamma_t^h = \left(\pi^1, \pi^2, v_t\right)_{\#}\mu_t \quad \text{in } \mathcal{W}_2([0,1]\times\Theta\times\Theta)$$

and

$$\lim_{h\to 0}\frac{\mathcal{W}_2^{\mathrm{COT}}(\mu_{t+h}, (\mathrm{Id} + h(0, v_t))_{\#}\mu_t)}{|h|} = 0\,.$$

Proof. The proof only needs to be slightly adapted from the one of [Ambrosio, 2008b, Lem.8.4.6] but we rewrite it here for completeness.

Let $(\varphi_n)_{n\geq 0}$ be a countable dense sequence in $\mathcal{C}^1_c([0,1]\times\Theta)$. Then for $\mathrm{d}t$-a.e. $t\in I$ we have $\lim_{h\to 0}\frac{1}{|h|}\mathcal{W}_2^{\mathrm{COT}}(\mu_{t+h},\mu_t)=\left|\frac{\mathrm{d}}{\mathrm{d}t}\mu_t\right|$ and for every $n\geq 0$:

$$\lim_{h\to 0}\frac{\mu_{t+h}(\varphi_n)-\mu_t(\varphi_n)}{h}=\int\langle\nabla_\theta\varphi_n,v_t\rangle\,\mathrm{d}\mu_t.$$

Introducing some $\gamma_t^h\in\Gamma_o^{\mathrm{Leb}}(\mu_{t+h},\mu_t)$ we consider:

$$\nu^h\coloneqq\left(\pi^1,\pi^2,\frac{1}{h}(\pi^3-\pi^2)\right)_{\#}\gamma_t^h\,.$$

Then for any sequence (h_n) converging to 0 the sequence (ν^{h_n}) is tight in $\mathcal{P}([0,1]\times\Theta\times\Theta)$ and we can consider a narrow limit point ν^0. The marginal of ν^h, and hence of ν^0, on $[0,1]\times\Theta$ is μ_t which allows to write by disintegration $\nu^0=\int\nu^0_{s,\theta}\mathrm{d}\mu_t(s,\theta)$. Then we have for every $n\geq 0$:

$$\begin{aligned}\frac{\mu_{t+h}(\varphi_n)-\mu_t(\varphi_n)}{h}&=\frac{1}{h}\int\left(\varphi_n(s,\theta')-\varphi_n(s,\theta)\right)\mathrm{d}\gamma_t^h(s,\theta,\theta')\\&=\frac{1}{h}\int\left(\varphi_n(s,\theta+hz)-\varphi_n(s,\theta)\right)\mathrm{d}\nu^h(s,\theta,z)\,,\end{aligned}$$

and taking the limit $h\to 0$ gives by Lebesgue's theorem:

$$\int\langle\nabla_\theta\varphi_n,v_t\rangle\,\mathrm{d}\mu_t=\int_{[0,1]\times\Theta}\int_\Theta\langle z,\nabla_\theta\varphi_n(s,\theta)\rangle\,\mathrm{d}\nu^0_{s,\theta}(z)\mathrm{d}\mu_t(s,\theta)\,.$$

For $(s,\theta)\in[0,1]\times\Theta$, let us denote by $\tilde{v}_t(s,\theta)\coloneqq\int_\Theta z\mathrm{d}\nu^0_{s,\theta}(z)$ the first moment of $\nu^0_{s,\theta}$. Then from the last equality and by a density argument it follows:

$$\mathrm{div}((0,\tilde{v}_t-v_t)\mu_t)=0\,,$$

and in particular the continuity equation Eq. (I.12) is satisfied with the vector field $(0,\tilde{v}_t)$. Let us now show:

$$\int_{[0,1]\times\Theta}\int_\Theta\|z\|^2\mathrm{d}\nu^0_{s,\theta}(z)\mathrm{d}\mu_t(s,\theta)\leq\left|\frac{\mathrm{d}}{\mathrm{d}t}\mu_t\right|^2\,.$$

Indeed we have:

$$\begin{aligned}\int_{[0,1]\times\Theta}\int_\Theta\|z\|^2\mathrm{d}\nu^0_{s,\theta}(z)\mathrm{d}\mu_t(s,\theta)&\leq\liminf_{h\to 0}\int_{[0,1]\times\Theta\times\Theta}\|z\|^2\mathrm{d}\nu^h(s,\theta,z)\\&=\liminf_{h\to 0}\int_{[0,1]\times\Theta\times\Theta}\frac{1}{h^2}\|\theta'-\theta\|^2\mathrm{d}\gamma_t^h(s,\theta,\theta')\\&=\liminf_{h\to 0}\frac{\mathcal{W}_2^{\mathrm{COT}}(\mu_{t+h},\mu_t)^2}{h^2}=\left|\frac{\mathrm{d}}{\mathrm{d}t}\mu_t\right|^2\,.\end{aligned}$$

Whence by definition of $\tilde{v}_t$ and Jensen's inequality:

$$\|\tilde{v}_t\|_{L^2(\mu_t)}\leq\left|\frac{\mathrm{d}}{\mathrm{d}t}\mu_t\right|=\|v_t\|_{L^2(\mu_t)}$$

from which it follows that $\tilde{v}_t=v_t$ in $L^2(\mu_t)$ because of the minimality of $\|v_t\|_{L^2(\mu_t)}$ and the strict convexity of the L^2-norm. Moreover the above inequality is strict whenever $\nu^0_{s,\theta}$ is not a Dirac mass in a set of μ_t positive measure. This implies that $\nu^0_{s,\theta}$ is a Dirac

mass for $\mathrm{d}\mu_t$-a.e. $(s,\theta) \in [0,1]\times\Theta$ and that $\nu^0 = (\pi^1,\pi^2,v_t)_\#\mu_t$. This proves the narrow convergence of ν^h towards the desired measure and together with the convergence of the second moments we have $\mathcal{W}_2$ convergence.

Let us now estimate the distance between μ_{t+h} and $(\pi^1,\pi^2+h(0,v_t))_\#\mu_t$ with the coupling $\gamma := (\pi^1,\pi^2+h(0,v_t),\pi^3)_\#\gamma_t^h \in \Gamma^{\mathrm{Leb}}((\pi^1,\pi^2+h(0,v_t))_\#\mu_t,\mu_{t+h})$. By the preceding result:

$$\begin{aligned}\frac{\mathcal{W}_2^{\mathrm{COT}}((\pi^1,\pi^2+hv_t)_\#\mu_t,\mu_{t+h})^2}{h^2} &\leq \int_{[0,1]\times\Theta\times\Theta}\frac{1}{h^2}\|\theta+hv_t(s,\theta)-\theta'\|^2\mathrm{d}\gamma_t^h(s,\theta,\theta')\\ &= \int_{[0,1]\times\Theta\times\Theta}\|v_t(s,\theta)-z\|^2\mathrm{d}\nu^h(s,\theta,z)\xrightarrow[h\to 0]{}0\,.\end{aligned}$$

□

Lemma I.2.3 (Adapted from [Ambrosio, 2008b, Thm.8.4.7])**.** *Let $(\mu_t)_{t\in I}$ be an absolutely continuous curve in $\mathcal{P}_2^{\mathrm{Leb}}([0,1]\times\Theta)$, let $v : I\times[0,1]\times\theta\to\Theta$ be its* tangent vector field *and let $\mu'\in\mathcal{P}_2^{\mathrm{Leb}}([0,1]\times\Theta)$. Then for $\mathrm{d}t$-a.e. $t\in I$:*

$$\frac{\mathrm{d}}{\mathrm{d}t}\mathcal{W}_2^{\mathrm{COT}}(\mu_t,\mu_t')^2 = 2\int_0^1\int_\Theta\langle\theta-\theta',v_t(s,\theta)\rangle\,\mathrm{d}\gamma(s,\theta,\theta')\,,\quad\forall\gamma\in\Gamma_o^{\mathrm{Leb}}(\mu_t,\mu_t')\,.$$

Proof. Having shown Lemma I.2.2, the proof is the same as the one of [Ambrosio, 2008b, Thm.8.4.7]. □

I.3 Gradient flow dynamics

To train the NODE model of Definition I.1 we consider performing *Gradient Flow* on the parameter μ for the risk $\mathcal{R}$ and for the COT metric described in the previous section. However the parameter set $\mathcal{P}_2^{\mathrm{Leb}}([0,1]\times\Theta)$ equipped with the distance $\mathcal{W}_2^{\mathrm{COT}}$ lacks a proper differential structure. We will thus in this section give a sense to the notion of gradient flow of $\mathcal{R}$. First, motivated by formal computations we will introduce a definition of gradient flow that is consistent with the one proposed by Chen et al. [Chen, 2018] for the training of NODEs of finite width. Then, we will show this definition to be equivalent to the notion of *curve of maximal slope* from the theory of gradient flow in metric spaces [Ambrosio, 2008b; Santambrogio, 2017]. Finally, this equivalence will allow us to show well-posedness results for the gradient flow equation.

I.3.1 Backward equation and adjoint variables

The computation of the gradient will make use of a new ODE linked to Eq. (I.6). This ODE should be understood as running backward over the time variable $s\in[0,1]$ with the initial condition at $s=1$. In the same way Eq. (I.6) models the processing of the data by a ResNet of infinite depth, the *adjoint variables* p solutions to this *backward ODE* should be considered as modeling the quantities calculated when performing back-propagation over a deep ResNet.

Definition I.2 (Adjoint variable)**.** *Let $\mu\in\mathcal{P}_2^{\mathrm{Leb}}([0,1]\times\Theta)$ and $(x,y)\in\mathbb{R}^{d+d'}$. Let $(x_\mu(s))_{s\in[0,1]}$ be the solution to Eq. (I.5) with parameter μ and $x_\mu(0)=x$. Then we call* adjoint variable associated to μ, x and y *the solution $(p_{\mu,x,y}(s))_{s\in[0,1]}$ to the* backward ODE*:*

$$\forall s\in[0,1]\,,\quad p_{\mu,x,y}(s) = \nabla_x\ell(x_\mu(1),y) + \int_s^1 \mathrm{D}_x F_{\mu(.|r)}(x_\mu(r))^\top p_{\mu,x,y}(r)\mathrm{d}r\,. \tag{I.13}$$

When no ambiguity, the dependence w.r.t. μ, x and y is omitted and we simply write $p(s)$.

The following proposition states the well-posedness of the backward equation under suitable assumptions on the basis function ψ and gives a useful representation of the adjoint variables.

Proposition I.3.1. *Let $\mu \in \mathcal{P}_2^{\mathrm{Leb}}([0,1] \times \Theta)$ and $(x, y) \in \mathbb{R}^{d+d'}$. Assume ψ satisfies Assumptions I.1 to I.3. Then there exists a unique solution to Eq. (I.13) which is given by:*

$$\forall s \in [0,1]\,, \quad p_{\mu,x,y}(s) = \Phi_{\mu,x}(s)^{-\top} \Phi_{\mu,x}(1)^{\top} \nabla_x \ell(x_\mu(1), y)\,. \tag{I.14}$$

where we define $\Phi_{\mu,x}$ to be the (matrix) solution of the linear ODE:

$$\forall s \in [0,1]\,, \quad \Phi_{\mu,x}(s) = \mathrm{Id} + \int_0^s \mathrm{D}_x F_{\mu(.|r)}(x_\mu(r)) \Phi_{\mu,x}(r) \mathrm{d}r\,. \tag{I.15}$$

When no ambiguity we simply denote by $\Phi_\mu(s)$ or even $\Phi(s)$.

Proof. Note that Eq. (I.13) is non-autonomous linear ODE w.r.t. the variable p. Thus, the representation Eq. (I.14) follows from the existence and uniqueness of Φ and to prove the result it suffices to show the map $s \mapsto \mathrm{D}_x F_{\mu(.|s)}(x(s))$ is integrable.

First, as ψ is continuously differentiable w.r.t. x with integrable differential for almost every fixed $s \in [0,1]$ the map $x \mapsto F_{\mu(.|s)}(x)$ is continuously differentiable with differential given by:

$$\mathrm{D}_x F_{\mu(.|s)}(x) = \int_\Theta \mathrm{D}_x \psi_\theta(x) \mathrm{d}\mu(\theta|s)\,.$$

Moreover, by continuity of $s \mapsto x(s)$ the integrand $\mathrm{D}_x \psi_\theta(x(t))$ is measurable and so is the map $s \mapsto \mathrm{D}_x F_{\mu(.|s)}(x(s))$. Finally integrability follows as $\mathrm{D}_x \psi_\theta(x(s))$ has 2-growth w.r.t. θ and $\int_\Theta \|\theta\|^2 \mathrm{d}\mu(\theta|s)$ is integrable on $[0,1]$. □

The following result gives an alternate point of view on the adjoint variable p. Geometrically, it follows from Eq. (I.14) that p lives in the co-tangent space of the flow x. In the case of a general (not necessarily with finite support) data distribution $\mathcal{D} \in \mathcal{P}(\mathbb{R}^d \times \mathbb{R}^{d'})$ it is convenient to see p as the gradient of a potential ψ over the variables $(x, y) \in \mathbb{R}^d \times \mathbb{R}^{d'}$.

Lemma I.3.1. *Let $\mu \in \mathcal{P}_2^{\mathrm{Leb}}([0,1] \times \Theta)$. Then for every $(x, y) \in \mathbb{R}^{d+d'}$ the associated adjoint variable p can be expressed for every $s \in [0,1]$ as:*

$$p_{\mu,x,y}(s) = \nabla_x \psi_\mu(s, x_\mu(s), y)\,, \tag{I.16}$$

where ψ_μ is the unique solution to the transport equation:

$$\partial_s \psi_\mu + \left\langle \nabla_x \psi_\mu, F_{\mu(.|s)} \right\rangle = 0\,, \quad \psi_\mu(1, x, y) = \ell(x, y)\,, \quad \forall (x, y) \in \mathbb{R}^d \times \mathbb{R}^{d'}\,. \tag{I.17}$$

Proof. The solution to the transport equation can be given in the characteristic form:

$$\forall s \in [0,1]\,, x \in \mathbb{R}^d\,, \quad \psi_\mu(s, x_\mu(s), y) = \psi_\mu(1, x_\mu(1), y) = \ell(x_\mu(1), y)\,.$$

One can then check that the r.h.s. of Eq. (I.16) is indeed a solution of Eq. (I.13). □

I.3.2 The gradient flow equation

We motivate here by formal computations a definition of a gradient flow equation for the risk $\mathcal{R}$. This *adjoint sensitivity analysis* consists in using a Lagrangian form of the risk minimization problem to obtain an expression of the gradient w.r.t. the parameter μ.

One can consider for a parameter $\mu \in \mathcal{P}_2^{\mathrm{Leb}}([0,1] \times \Theta)$ and every time $s \in [0,1]$ the distribution $\rho_\mu(.|s) \coloneqq (x \mapsto x_\mu(s), \mathrm{Id})_\# \mathcal{D}$ of the data at time s. Then, as the inputs are processed by our model through the ODE Eq. (I.5), $\{\rho_\mu(.|s)\}_{s\in[0,1]}$ is a narrowly continuous solution to the continuity equation:

$$\partial_s \rho_\mu^*(.|s) + \mathrm{div}_x(F_{\mu(.|s)} \rho_\mu^*(.|s)) = 0\,, \tag{I.18}$$

and the risk associated to μ is

$$\mathcal{R}(\mu) = \int_{\mathbb{R}^d \times \mathbb{R}^{d'}} \ell(x,y) \mathrm{d}\rho_\mu(x,y|1) = \rho_\mu(.|1)(\ell)\,.$$

We introduce a Lagrange multiplier ψ to penalize the above continuity equation. For a parameter $\mu \in \mathcal{P}_2^{\mathrm{Leb}}([0,1] \times \Theta)$, a measurable family $\rho = \{\rho(.|s)\}_{s\in[0,1]}$ of probability measures over $\mathbb{R}^d \times \mathbb{R}^{d'}$ and a smooth test function $\psi : [0,1] \times \mathbb{R}^d \times \mathbb{R}^{d'} \to \mathbb{R}$, consider the lagrangian $\mathcal{L}$ defined as:

$$\begin{aligned}\mathcal{L}(\mu,\rho,\psi) \coloneqq \rho(.|1)(\ell) - \rho(.|1)(\psi(1)) - \rho(.|0)(\psi(0)) \\ + \int_0^1 \int_{\mathbb{R}^{d+d'}} \left(\partial_s \psi + \left\langle \nabla_x \psi, F_{\mu(.|s)} \right\rangle\right) \mathrm{d}\rho(.|s)\mathrm{d}s\,.\end{aligned} \tag{I.19}$$

Using the definition of F and inverting integrals, the variation of $\mathcal{L}$ w.r.t. μ is given for every ρ and ψ by:

$$\frac{\delta \mathcal{L}}{\delta \mu}(\mu,\rho,\psi) : (s,\theta) \mapsto \int_{\mathbb{R}^{d+d'}} \langle \nabla_x \psi(s,x,y), \psi(\theta,x) \rangle \, \mathrm{d}\rho(x,y|s)\,.$$

Also, if $\rho = \rho_\mu$ is the solution of Eq. (I.18) for the parameter μ, we have the relation $\mathcal{L}(\mu,\rho_\mu,\psi) = \mathcal{R}(\mu)$ for any test function ψ. Hence the variation of $\mathcal{R}$ w.r.t. μ is:

$$\frac{\delta \mathcal{R}}{\delta \mu}(\mu) = \frac{\delta \mathcal{L}}{\delta \mu}(\mu,\rho_\mu,\psi) + \frac{\delta \mathcal{L}}{\delta \rho}(\mu,\rho_\mu,\psi) \frac{\delta \rho_\mu}{\delta \mu}(\mu)\,,$$

where the *Lagrange multiplier* ψ can be chosen arbitrarily. Also the variation of $\mathcal{L}$ w.r.t. the family of probability measures ρ, seen as the probability measure whose disintegration on $[0,1]$ is $\{\rho_s\}_{s\in[0,1]}$ (with the fixed initial condition $\rho(.|0) = \mathcal{D}$), can be formally given by:

$$\frac{\delta \mathcal{L}}{\delta \rho}(\mu,\rho,\psi) = (\ell - \psi(1))\,\delta_{s=1} + \partial_s \psi + \left\langle \nabla_x \psi\,, F_{\mu(.|s)} \right\rangle\,.$$

We see that taking $\psi = \psi_\mu$ to be a solution of Eq. (I.17) cancels $\frac{\delta \mathcal{L}}{\delta \rho}$ for every ρ and hence:

$$\frac{\delta \mathcal{R}}{\delta \mu}(\mu) = \frac{\delta \mathcal{L}}{\delta \mu}(\mu,\rho_\mu,\psi_\mu)\,.$$

By Theorem I.1 we know that, for every absolutely continuous curve (μ_t) passing through μ, its variation at μ is given by $\partial_t \mu_t = -\mathrm{div}((0,v)\mu)$ for some $v \in L^2(\mu)$. A notion of

gradient of $\mathcal{R}$ (for the "differential" structure of $\mathcal{P}_2^{\mathrm{Leb}}([0,1]\times\Theta)$) at μ could thus be defined as the unique solution to the variational problem:

$$\nabla\mathcal{R}(\mu) \in \operatorname*{arg\,min}_{v\in L^2(\mu)} \frac{1}{2}\|v\|^2_{L^2(\mu)} - \left\langle \nabla_\theta \frac{\delta\mathcal{R}}{\delta\mu}(\mu), v \right\rangle_{L^2(\mu)} .$$

This problem admits a unique solution $v^* \in L^2(\mu)$, provided that $\nabla_\theta \frac{\delta\mathcal{R}}{\delta\mu}(\mu) \in L^2(\mu)$, and using the relation Eq. (I.16) between the adjoint variable p_μ and the potential ψ_μ we have:

$$v^* = \nabla_\theta \frac{\delta\mathcal{R}}{\delta\mu}(\mu) = \nabla_\theta \frac{\delta\mathcal{L}}{\delta\mu}(\mu, \rho_\mu, \psi_\mu) ,$$

that is

$$v^* : (s,\theta) \mapsto \int_{\mathbb{R}^{d+d'}} \mathrm{D}_\theta\psi(\theta,x)^\top \nabla_x\psi_\mu(x,y)\mathrm{d}\rho_\mu(x,y|s) = \mathbb{E}_{x,y}\mathrm{D}_\theta\psi(\theta,x_\mu(s))^\top p_{\mu,x,y}(s) .$$

If the above calculations are purely formal they motivate the following definition of gradient flow for $\mathcal{R}$. In particular, this definition will be shown in the next section to be equivalent to the appropriate notion of gradient flow in metric spaces.

Definition I.3 (Gradient flow equation)**.** *Let $I \subset \mathbb{R}$ be an interval. For $\mu \in \mathcal{P}_2^{\mathrm{Leb}}([0,1]\times\Theta)$ let us define:*

$$\nabla\mathcal{R}[\mu] : (s,\theta) \mapsto \mathbb{E}_{x,y}\mathrm{D}_\theta\psi(\theta,x_\mu(s))^\top p_{\mu,x,y}(s) . \tag{I.20}$$

We say a locally absolutely continuous curve $t \in I \mapsto \mu_t \in \mathcal{P}_2^{\mathrm{Leb}}([0,1]\times\Theta)$ is a gradient flow *for $\mathcal{R}$ if it is a weak solution to the continuity equation:*

$$\partial_t\mu_t - \operatorname{div}\left((0, \nabla\mathcal{R}[\mu_t])\,\mu_t\right) = 0 \quad \text{on } I\times[0,1]\times\Theta. \tag{I.21}$$

The following result is a useful representation formula for the gradient flow curves defined by Definition I.3: for every $t \geq 0$ the gradient flow μ_t at time t is the pushforward of the initialization μ_0 by a flow-map. The proof relies on classical results from transport equation theory [Ambrosio, 2008a].

Proposition I.3.2. *Assume ψ is twice continuously differentiable and satisfies Assumptions I.1 to I.3. Let $(\mu_t)_{t\geq 0}$ be a gradient flow for the risk $\mathcal{R}$ and consider for every $t \geq 0$ the vector field:*

$$V_t : (s,\theta) \mapsto (0, \nabla\mathcal{R}[\mu_t](s,\theta)) = \left(0, \mathbb{E}_{x,y}\mathrm{D}_\theta\psi(\theta,x_{\mu_t}(s))^\top p_{\mu_t,x,y}(s)\right) \in \mathbb{R}\times\Theta .$$

Then for every $t \geq 0$ we have $\mu_t = (X_t)_\#\mu_0$ where X_t is the flow-map solution of the ODE:

$$\frac{\mathrm{d}}{\mathrm{d}t}X_t(s,\theta) = V_t(X_t(s,\theta)) , \quad X_0 = \mathrm{Id} . \tag{I.22}$$

Proof. The existence and uniqueness of the flow-map X_t for every $t \geq 0$ follows from the assumptions on ψ (in particular linear growth and local Lipschitz continuity of $\mathrm{D}_\theta\psi$ w.r.t. θ) and classical theory of ODEs. The flow-map representation of μ_t then follows from [Ambrosio, 2008a, Thm.3.2] as for any initial value $(s,\theta) \in [0,1]\times\Theta$ the set of curves solutions to the ODE is the singleton $\{(X_t(s,\theta))_{t\geq 0}\}$. □

Consistency with the adjoint gradient flow A case of particular interest for numerical applications is when the measure μ is discretized and approximated at every time $s \in [0,1]$ by an empirical distribution. Given $M \geq 1$ and $\theta = (\theta^j)_{1 \leq j \leq M} \in L^2([0,1], \Theta)^M$ we define the associated empirical distribution $\mu_\theta \in \mathcal{P}_2^{\mathrm{Leb}}([0,1] \times \Theta)$ by:

$$\text{for } \mathrm{d}s\text{-a.e. } s \in [0,1], \quad \mu_\theta(.|s) \coloneqq \frac{1}{M} \sum_{j=1}^{M} \delta_{\theta^j(s)} \mathrm{d}s \,. \tag{I.23}$$

i.e. μ_θ is the measure whose disintegration at any time $s \in [0,1]$ is the empirical measure $\frac{1}{M}\sum_{j=1}^{M} \delta_{\theta^j(s)}$. Then we denote by $\mathcal{R}(\theta) \coloneqq \mathcal{R}(\mu_\theta)$ the risk associated to θ. In the original work of Chen et al. [Chen, 2018], the authors propose to train the Neural ODE parameterized by θ and minimize $\mathcal{R}(\theta)$ by performing gradient descent for the *adjoint gradient* defined as:

$$\nabla_{\theta^j} \mathcal{R}(\theta) \coloneqq \mathbb{E}_{x,y} \mathrm{D}_\theta \psi(\theta^j, x(s))^\top p(s) \,,$$

where x and p are respectively the solutions of Eqs. (I.6) and (I.13) for the parameter μ_θ. One can observe that the adjoint gradient is the one calculated by Eq. (I.20) when $\mu = \mu_\theta$. Given sufficient regularity assumptions on the basis function ψ, we have by Proposition I.3.2 that $(\mu_t)_{t \geq 0}$ is a gradient flow in the sense of Definition I.3 with $\mu_0 = \mu_{\theta_0}$ if and only if $\mu_t = \mu_{\theta_t}$ for every $t \geq 0$ and $(\theta_t)_{t \geq 0}$ is a gradient flow for the above adjoint gradient.

I.3.3 Gradient flows as curves of maximal slope

There exists a large body of mathematical works devoted to the generalization of the classical theory of gradient flows to functionals over metric spaces. Ambrosio, Gigli, and Savaré [Ambrosio, 2008b] give an in-depth presentation of this theory. Complementary and more synthetic presentations are given by Ambrosio et al. [Ambrosio, 2013] and Santambrogio [Santambrogio, 2017]. Based on those works, we introduce here another definition of gradient flows for the risk $\mathcal{R}$ which is the one of *curves of maximal slope* and show it coincides with the definition from the previous section.

I.3.3.1 Curves of maximal slope in metric spaces

When Z is a Euclidean space, and f is a smooth function the gradient flow of f is defined as the solution of the ODE $\frac{\mathrm{d}}{\mathrm{d}t} z_t = -\nabla f(z_t)$. Then such a gradient flow satisfies:

$$\frac{\mathrm{d}}{\mathrm{d}t} f(z_t) = -\|\nabla f(z_t)\|^2 = -\frac{1}{2}\left(\|\frac{\mathrm{d}}{\mathrm{d}t} z_t\|^2 + \|\nabla f(z_t)\|^2 \right) ,$$

whereas for any other smooth curve (y_t) we have by Young's inequality:

$$\frac{\mathrm{d}}{\mathrm{d}t} f(y_t) = -\left\langle \nabla f(y_t), \frac{\mathrm{d}}{\mathrm{d}t} y_t \right\rangle \geq -\frac{1}{2}\left(\|\frac{\mathrm{d}}{\mathrm{d}t} y_t\|^2 + \|\nabla f(y_t)\|^2 \right) , \tag{I.24}$$

with equality if and only if $\frac{\mathrm{d}}{\mathrm{d}t} y_t = -\nabla f(y_t)$. Hence, we see that imposing equality in the above inequality gives a characterization of gradient flow curves in the Euclidean case. The definition of *curves of maximal slope* is based on the generalization of this characterization to metric spaces. For example, a generalization of the speed's norm $\|\frac{\mathrm{d}}{\mathrm{d}t} z_t\|$ is given by the metric derivative $\left|\frac{\mathrm{d}}{\mathrm{d}t} z_t\right|$. To give a sense to the gradient's norm $\|\nabla f(z)\|$ we need to introduce the notion of *upper gradient*.

Definition I.4 (Upper gradient [Ambrosio, 2008b, Def.1.2.1])**.** *Let (Z, d) be a complete metric space and $f : Z \to \mathbb{R}$ be a function. The map $g : Z \to [0, +\infty]$ is an* upper gradient *for f if for every absolutely continuous curve $(z_t)_{t\in I}$ we have that $t \mapsto g(z_t)$ is measurable and:*

$$|f(z_{t_1}) - f(z_{t_2})| \leq \int_{t_1}^{t_2} g(z_t) \left|\frac{\mathrm{d}}{\mathrm{d}t} z_t\right| \mathrm{d}t\,, \quad \forall t_1 \leq t_2 \in I\,.$$

When no ambiguity an upper gradient of f will simply be denoted by $|\nabla f|$.

Given an upper gradient for f, the definition of curves of maximal slope consists in imposing equality in Eq. (I.24).

Definition I.5 (Curve of maximal slope [Ambrosio, 2008b, Def.1.3.2])**.** *Let (Z, d) be a complete metric space, $I \in \mathbb{R}$ be an interval and $f : Z \to \mathbb{R}$ a function with $|\nabla f|$ an upper gradient for f. We say that $(z_t)_{t\in I}$ is a* curve a maximal slope *for f (w.r.t. $|\nabla f|$) if it satisfies:*

(i) $(z_t)_{t\in I}$ is locally absolutely continuous,

(ii) the map $t \mapsto f(z_t)$ is non-increasing,

(iii) for $\mathrm{d}t$-a.e. $t \in I$ it holds $\frac{\mathrm{d}}{\mathrm{d}t} f(z_t) \leq -\frac{1}{2}\left(\left|\frac{\mathrm{d}}{\mathrm{d}t} z_t\right|^2 + |\nabla f|^2(z_t)\right)$.

If $\lim_{t\to \inf I} z_t = z$ exists then we say $(z_t)_{t\in I}$ is a curve of maximal slope starting at z.

Remark I.3.1 (About the various definitions of *curves of maximal slope*)**.** *There exists various definitions of the notion of curve of maximal slope in metric spaces, see for example [Ambrosio, 2013, Sec.4] for a discussion about the various definitions and their relations. Our definition is the same as the one used in [Hauer, 2019, Def.2.12]. In particular, it implies the following* Energy Dissipation Inequality *[Hauer, 2019, Prop.2.14]:*

$$f(z_{t_1}) - f(z_{t_2}) \geq \frac{1}{2}\int_{t_1}^{t_2} \left(\left|\frac{\mathrm{d}}{\mathrm{d}t} z_t\right|^2 + |\nabla f|^2(z_t)\right) \mathrm{d}t\,, \quad \forall t_1, t_2 \in I\,. \tag{EDI}$$

This definition differs from the one exposed in [Dello Schiavo, 2024, Def.2.2] (see also [Muratori, 2020, Def.4.4]) as the map $t \mapsto f(z_t)$ need not be locally absolutely continuous. The difference between these two definitions is discussed in [Dello Schiavo, 2024, Rem.2.6] but observe that for our purpose (i) *the loss $\mathcal{R}$ will be shown to be locally Lipschitz in Corollary I.3.1, hence implying that $\mathcal{R}(\mu_t)$ is locally absolutely continuous,* (ii) *the gradient norm $\|\nabla \mathcal{R}[\mu]\|_{L^2(\mu)}$ will be shown to be an upper gradient in Proposition I.3.4. For these reasons, we need not here make the distinction between these two definitions and prefer weaker assumptions.*

Note that there is *a priori* no reasons for Definitions I.3 and I.5 to define the same notion of gradient flow for the risk $\mathcal{R}$. In particular, the first definition uses the existence of the adjoint variable p and thus some regularity on ψ. In contrast, the second definition requires an upper gradient which is yet unspecified for $\mathcal{R}$. Taking appropriate assumptions on the basis function ψ, we show in the rest of this section that the risk $\mathcal{R}$ is sufficiently regular for the two definitions to coincide. This will be the content of Theorem I.2.

I.3.3.2 Curve of maximal slope for the risk $\mathcal{R}$

Provided with Definition I.5, we seek to characterize the curves of maximal slope for the risk $\mathcal{R}$ in the metric space $\mathcal{P}_2^{\mathrm{Leb}}([0,1]\times\Theta)$. We first show the NODE's output is a locally Lipschitz function of the parameter $\mu\in\mathcal{P}_2^{\mathrm{Leb}}([0,1]\times\Theta)$.

Assumption I.2 (local Lipschitz continuity w.r.t. θ)**.** *Assume that $\psi:\Theta\times\mathbb{R}^d\to\mathbb{R}^d$ is locally Lipschitz w.r.t. θ with a Lipschitz constant that grows at most linearly w.r.t. Θ: for every $R\geq 0$ there exists a constant $C(R)$ s.t.:*

$$\forall x\in B(0,R)\,,\ \forall\theta,\theta'\in\Theta\,,\quad \|\psi(\theta,x)-\psi(\theta',x)\|\leq C(R)(1+\max(\|\theta\|,\|\theta'\|))\|\theta-\theta'\|\,.$$

Lemma I.3.2 (local Lipschitz continuity of the flow)**.** *Assume ψ satisfies Assumptions I.1 and I.2 and consider some input $x\in\mathbb{R}^d$. Then the map*

$$\mu\in\mathcal{P}_2^{\mathrm{Leb}}([0,1]\times\Theta)\mapsto(x_\mu(s))_{s\in[0,1]}\in\mathcal{C}([0,1],\mathbb{R}^d)$$

is locally Lipschitz. More precisely, for every $\mathcal{E}\geq 0$ there exists a constant $C=C(\mathcal{E})$ s.t.:

$$\sup_{s\in[0,1]}\|x_\mu(s)-x_{\mu'}(s)\|\leq C\mathcal{W}_2^{\mathrm{COT}}(\mu,\mu')\,,$$

for every $\mu,\mu'\in\mathcal{P}_2^{\mathrm{Leb}}([0,1]\times\Theta)$ with $\mathcal{E}_2(\mu),\mathcal{E}_2(\mu')\leq\mathcal{E}$. Moreover, the constant C can be chosen uniformly over x in a compact set.

Proof. Consider $x\in\mathbb{R}^d$, $\mathcal{E}\geq 0$ and μ,μ' such as in the statement. We denote by $(x(s))_{s\in[0,1]}$ and $(x'(s))_{s\in[0,1]}$ the flow associated to x and to the parameters μ and μ' respectively. Let $R\geq 0$ be such that $\|x\|\leq R$. Then by Proposition I.1.1 the trajectories x,x' are uniformly bounded by some $R'=R'(R,\mathcal{E})$. Then using Eq. (I.5) we have for every $s\in[0,1]$:

$$\begin{aligned}\|x(s)-x'(s)\|\leq&\|x(0)-x'(0)\|+\|\int_0^s\int_\Theta\psi(\theta,x(r))\mathrm{d}\mu(\theta|r)\mathrm{d}r-\int_0^s\int_\Theta\psi(\theta,x'(r))\mathrm{d}\mu'(\theta|r)\mathrm{d}r\|\\ \leq&\|x(0)-x'(0)\|+\int_0^s\int_\Theta\|\psi(\theta,x(r))-\psi(\theta,x'(r))\|\mathrm{d}\mu(\theta|r)\mathrm{d}r\\ &+\int_0^s\|\int_\Theta\psi(\theta,x'(r))\mathrm{d}(\mu-\mu')(\theta|r)\|\mathrm{d}r\,.\end{aligned}$$

For the first integral note that using the local Lipschitz continuity of ψ w.r.t. x in Assumption I.1 we have for every $r\in[0,1]$:

$$\int_\Theta\|\psi(\theta,x(r))-\psi(\theta,x'(r))\|\mathrm{d}\mu(\theta|r)\leq C_1\int_\Theta(1+\|\theta\|^2)\mathrm{d}\mu(\theta|r)\|x(r)-x'(r)\|\,,$$

where $C_1=C_1(R,\mathcal{E})$. For the second integral note that at fixed $r\in[0,1]$, if $\gamma\in\Gamma_o(\mu(.|r),\mu'(.|r))$ is a (optimal) coupling between $\mu(.|r)$ and $\mu'(.|r)$, then:

$$\int_\Theta\psi(\theta,x'(r))\mathrm{d}(\mu-\mu')(\theta|r)=\int_{\Theta^2}(\psi(\theta,x'(r))-\psi(\theta',x'(r)))\mathrm{d}\gamma(\theta,\theta')\,.$$

Using the local Lipschitz continuity of ψ w.r.t. θ in Assumption I.2 and the optimality of γ:

$$\begin{aligned}\|\int_\Theta\psi(\theta,x'(r))\mathrm{d}(\mu-\mu')(\theta|r)\|\leq&\int_{\Theta^2}\|\psi(\theta,x'(r))-\psi(\theta',x'(r))\|\mathrm{d}\gamma(\theta,\theta')\\ \leq&\int_{\Theta^2}C_2(1+\max(\|\theta\|,\|\theta'\|))\|\theta-\theta'\|\mathrm{d}\gamma(\theta,\theta')\\ \leq&\sqrt{3}C_2\left(1+\mathcal{E}_2(\mu(.|r))+\mathcal{E}_2(\mu'(.|r))\right)^{1/2}\mathcal{W}_2(\mu(.|r),\mu'(.|r))\,,\end{aligned}$$

where $C_2 = C_2(R, \mathcal{E})$. Integrating those inequalities gives by Grönwall's lemma that for $s \in [0,1]$:

$$\|x(s) - x'(s)\| \leq \sqrt{3} e^{C_1(1+\mathcal{E}_2(\mu))} C_2 \int_0^s \left(1 + \mathcal{E}_2(\mu(.|r)) + \mathcal{E}_2(\mu'(.|r))\right)^{1/2} \mathcal{W}_2(\mu(.|r), \mu'(.|r)) \mathrm{d}r \\ \leq C_3 \mathcal{W}_2^{\mathrm{COT}}(\mu, \mu') \,,$$

where $C_3 = C_3(R, \mathcal{E})$. □

As an immediate corollary of the above proposition we get that, provided the loss function ℓ is itself locally Lipschitz then the risk $\mathcal{R}$ is also a locally Lipschitz function of μ.

Corollary I.3.1 (local Lipschitz continuity of the risk)**.** *Assume that ψ satisfies Assumptions I.1 and I.2 and ℓ is locally Lipschitz w.r.t. x. Then the risk map $L : \mu \in \mathcal{P}_2^{\mathrm{Leb}}([0,1] \times \Theta) \mapsto \mathcal{R}(\mu)$ is locally Lipschitz.*

Assuming more regularity on the map ψ, one can express the first variation δx of the flow map with respect to a variation of the parameter transported by a velocity field $v \in L^2(\mu)$.

Assumption I.3 (Differentiability of ψ)**.** *Assume that ψ is continuously differentiable and s.t.*

(i) $\mathrm{D}_x \psi$ *grows at most quadratically with θ: for every $R \geq 0$ there exists a constant $C(R)$ such that*

$$\forall x \in B(0,R) \,, \ \forall \theta \in \Theta \,, \quad \|\mathrm{D}_x \psi(\theta, x)\| \leq C(R)(1 + \|\theta\|^2) \,.$$

(ii) $\mathrm{D}_\theta \psi$ *grows at most linearly with θ: for every $R \geq 0$ there exists a constant $C = C(R)$ such that*

$$\forall x \in B(0,R) \,, \ \forall \theta \in \Theta \,, \quad \|\mathrm{D}_\theta \psi(\theta, x)\| \leq C(R)(1 + \|\theta\|) \,.$$

Proposition I.3.3. *Assume ψ satisfies Assumptions I.1 to I.3. Consider $\mu \in \mathcal{P}_2^{\mathrm{Leb}}([0,1] \times \Theta)$ and a velocity field $v : [0,1] \times \Theta \to \Theta$ in $L^2(\mu)$. For $t \in \mathbb{R}$, define $\mu_t \coloneqq (\mathrm{Id} + t(0,v))_{\#}\mu$. Then, for $x \in \mathbb{R}^d$, $(x_{\mu_t})_{t \in \mathbb{R}}$ is differentiable in $\mathcal{C}([0,1], \mathbb{R}^d)$ at $t = 0$ and $\delta x \coloneqq \frac{\mathrm{d}}{\mathrm{d}t} x_{\mu_t}|_{t=0}$ is the solution to:*

$$\forall s \in [0,1] \,, \ \delta x(s) = \int_0^s \mathrm{D} F_{\mu(.|r)}(x_\mu(r)) \delta x(r) \mathrm{d}r + \int_0^s \int_\Theta \mathrm{D}_\theta \psi(\theta, x_\mu(r)) v(r, \theta) \mathrm{d}\mu(\theta|r) \mathrm{d}r \,. \tag{I.25}$$

Proof. First, thanks to Assumption I.3, for $\nu \in \mathcal{P}_2(\Theta)$ the map $F_\nu : \mathbb{R}^d \to \mathbb{R}^d$ is differentiable with $\mathrm{D}F_\nu : x \mapsto \int_\Theta \mathrm{D}_x \psi(\theta, x) \mathrm{d}\nu(\theta)$. Also, δx is well-defined as the unique solution of Eq. (I.25) and:

$$\forall s \in [0,1] \,, \quad \delta x(s) = \int_0^s \int_\Theta \Phi_{\mu,x}(s) \Phi_{\mu,x}(r)^{-1} \mathrm{D}_\theta \psi(\theta, x(r)) v(r, \theta) \mathrm{d}\mu(r, \theta) \,.$$

For simplicity, in the rest of the proof we will write $x_t \coloneqq x_{\mu_t}$ for any $t \in \mathbb{R}$. Let us then show that δx is the derivative of x_t at $t = 0$. For $t \neq 0$, consider the normalized increment:

$$z_t \coloneqq \frac{1}{t}(x_t - x_0) \in \mathcal{C}([0,1], \mathbb{R}^d) \,.$$

Then we have by definition of x_t and x_0 that for every $s \in [0,1]$:

$$\begin{aligned} z_t(s) &= \frac{1}{t}\int_0^s \int_\Theta \psi(\theta, x_t(r))\mathrm{d}\mu_t(r,\theta) - \frac{1}{t}\int_0^s \int_\Theta \psi(\theta, x_0(r))\mathrm{d}\mu(r,\theta) \\ &= \frac{1}{t}\int_0^s \int_\Theta \left(\psi(\theta + tv(r,\theta), x_t(r)) - \psi(\theta, x_0(r))\right)\mathrm{d}\mu(r,\theta) \\ &= \int_0^s \int_\Theta \left(\int_0^1 \mathrm{D}_x\psi(\theta, x_0(r) + utz_t(r))\mathrm{d}u\right)\cdot z_t(r)\mathrm{d}\mu(r,\theta) \\ &\qquad + \int_0^s \int_\Theta \left(\int_0^1 \mathrm{D}_\theta\psi(\theta + utv(r,\theta), x_t(r))\mathrm{d}u\right)\cdot v(r,\theta)\mathrm{d}\mu(r,\theta)\,. \end{aligned}$$

Hence z_t is solution of the linear ODE $z_t(s) = \int_0^s \left(A_t(r)\cdot z_t(r) + b_t(r)\right)\mathrm{d}r$ where we defined for $\mathrm{d}r$-a.e. $r \in [0,1]$:

$$\begin{aligned} A_t(r) &\coloneqq \int_\Theta \int_0^1 \mathrm{D}_x\psi(\theta, x_0(r) + utz_t(r))\mathrm{d}u\mathrm{d}\mu(\theta|r)\,, \\ b_t(r) &\coloneqq \int_\Theta \int_0^1 \mathrm{D}_\theta\psi(\theta + utv(r,\theta), x_t(r))\cdot v(r,\theta)\mathrm{d}u\mathrm{d}\mu(\theta|r)\,, \end{aligned}$$

and in order to prove that $z_t \to \delta x$ in $\mathcal{C}([0,1],\mathbb{R}^d)$ as $t \to 0$ it suffices to show that A_t and b_t converge respectively in $L^1([0,1])$ to:

$$A(r) \coloneqq \int_\Theta \mathrm{D}_x\psi(\theta, x_0(r))\mathrm{d}\mu(\theta|r)\,, \quad \text{and} \quad b(r) \coloneqq \int_\Theta \mathrm{D}_\theta\psi(\theta, x_0(r))\cdot v(r)\mathrm{d}\mu(\theta|r)\,.$$

Indeed, note that $\mathcal{W}_2^{\mathrm{COT}}(\mu_t,\mu) \le t\|v\|_{L^2(\mu)}$ and the family $(z_t)_{t\in[-1,1]}$ is bounded $\mathcal{C}([0,1],\mathbb{R}^d)$ by Lemma I.3.2. Thus for $t \in \mathbb{R}$:

$$\int_0^1 |A_t(r) - A(r)|\,\mathrm{d}r \le \int_0^1 \int_\Theta \left|\int_0^1 \mathrm{D}_x\psi(\theta, x_0(r) + utz_t(r))\mathrm{d}u - \mathrm{D}_x\psi(\theta, x_0(r))\right|\mathrm{d}\mu(r,\theta) \xrightarrow[t\to 0]{} 0$$

where Assumption I.3 allows to bound the integrand by an integrable function and to apply Lebesgue's theorem, showing that $A_t \to A$ as $t \to 0$ in $L^1([0,1])$. Similarly for b_t:

$$\int_0^1 |b_t(r) - b(r)|\,\mathrm{d}r \le \int_0^1 \int_\Theta \left|\int_0^1 \mathrm{D}_\theta\psi(\theta + utv(r,\theta), x_t(r))\mathrm{d}u - \mathrm{D}_\theta\psi(\theta, x_0(r))\right|\|v(r,\theta)\|\mathrm{d}\mu(r,\theta) \xrightarrow[t\to 0]{} 0.$$

□

A direct consequence of the previous result is the differentiability of the flow map and consequently of the risk along absolutely continuous curves.

Corollary I.3.2 (Differentiability of the flow)**.** *Assume ψ satisfies Assumptions I.1 to I.3. Let $I \subset \mathbb{R}$ be an interval and consider $(\mu_t)_{t\in I}$ an absolutely continuous curve in $\mathcal{P}_2^{\mathrm{Leb}}([0,1]\times\Theta)$ satisfying the continuity equation:*

$$\partial_t\mu_t + \mathrm{div}((0,v_t)\mu_t) = 0 \quad \text{on } I\times[0,1]\times\Theta.$$

Consider some $x \in \mathbb{R}^d$. Then $(x_{\mu_t})_{t\in I}$ is an absolutely continuous curve in $\mathcal{C}([0,1],\mathbb{R}^d)$ and is differentiable in $\mathcal{C}([0,1],\mathbb{R}^d)$ for $\mathrm{d}t$-a.e. $t \in I$ with $\delta x_t \coloneqq \frac{\mathrm{d}}{\mathrm{d}t}x_{\mu_t}$ the solution to:

$$\forall s \in [0,1]\,,\ \delta x_t(s) = \int_0^s \mathrm{D}F_{\mu_t(.|r)}(x_{\mu_t}(r))\delta x_t(r)\mathrm{d}r + \int_0^s \int_\Theta \mathrm{D}_\theta\psi(\theta, x_{\mu_t}(r))v_t(r,\theta)\mathrm{d}\mu_t(\theta|r)\mathrm{d}r\,. \tag{I.26}$$

Proof. For $t \in I$ we use the shortcut notation $x_t \coloneqq x_{\mu_t}$. The fact that $(x_t)_{t\in I}$ is absolutely continuous follows from Lemma I.3.2 stating that the flow map is locally Lipschitz. To prove the result it hence suffices to show that δx_t is the derivative of $t \mapsto x_t$ in $\mathcal{C}([0,1], \mathbb{R}^d)$.

Note that, without loss of generality we can consider v_t to be the (uniquely defined) tangent velocity field of the curve $(\mu_t)_{t\in I}$. Indeed if $\tilde{v}_t$ is the tangent velocity field then we have by Theorem I.1 that in the sense of distributions:

$$\operatorname{div}((0, v_t - \tilde{v}_t)\mu_t) = 0\,.$$

Hence for every $x \in \mathbb{R}^d$ and every $s \in [0,1]$:

$$\int_0^s \int_\Theta \mathrm{D}_\theta \psi(\theta, x_t(r)) v_t(r,\theta) \mathrm{d}\mu_t(r,\theta) = \int_0^s \int_\Theta \mathrm{D}_\theta \psi(\theta, x_t(r)) \tilde{v}_t(r,\theta) \mathrm{d}\mu_t(r,\theta)$$

and the definition of δx_t stays unchanged. Then, assuming v_t is the tangent velocity field to the curve μ_t, we can consider a subset $\Lambda \subset I$ of full Lebesgue measure such that the conclusions of Lemma I.2.2 hold. For every $t \in \Lambda$ and every $h \neq 0$ consider $\tilde{\mu}_t^h \coloneqq (\mathrm{Id} + h(0, v_t))_\# \mu_t$ and $\tilde{x}_t^h$ the associated flow. Then by Proposition I.3.3:

$$\|\frac{x_{t+h} - x_t}{h} - \delta x_t\|_{\mathcal{C}([0,1])} \leq \|\frac{\tilde{x}_t^h - x_t}{h} - \delta x_t\|_{\mathcal{C}([0,1])} + \|\frac{x_{t+h} - \tilde{x}_t^h}{h}\|_{\mathcal{C}([0,1])} \xrightarrow[h\to 0]{} 0$$

where the first term goes to 0 by application of Proposition I.3.3.The second term also goes to 0 by the fact that the flow map is locally Lipschitz, thus

$$\|x_{t+h} - \tilde{x}_t^h\|_{\mathcal{C}([0,1])} \leq C \mathcal{W}_2^{\mathrm{COT}}(\mu_{t+h}, \tilde{\mu}_t^h)$$

for some constant C and $\frac{1}{h}\|x_{t+h} - \tilde{x}_t^h\|_{\mathcal{C}([0,1])} \to 0$ by Lemma I.2.2. Note that, as $\Lambda \subset I$ is independent of x, it follows that the curve $t \mapsto x_t$ is differentiable at every $t \in \Lambda$ for every $x \in \mathbb{R}^d$. □

Corollary I.3.3 (Differentiability of the loss)**.** *Assume ψ satisfies Assumptions I.1 to I.3 and ℓ is continuously differentiable. Let $I \subset \mathbb{R}$ be an interval and $(\mu_t)_{t\in I}$ be as in Corollary I.3.2. Then $(\mathcal{R}(\mu_t))_{t\in I}$ is absolutely continuous and for almost every $t \in I$:*

$$\frac{\mathrm{d}}{\mathrm{d}t}\mathcal{R}(\mu_t) = \int_{[0,1]\times\Theta} \langle \nabla \mathcal{R}[\mu_t](s,\theta), v_t(s,\theta)\rangle \,\mathrm{d}\mu_t(s,\theta)\,.$$

Proof. First, the fact that $t \mapsto \mathcal{R}(\mu_t)$ is absolutely continuous follows from the fact that $\mu \mapsto \mathcal{R}(\mu)$ is locally Lipschitz, as shown in Corollary I.3.1. It remains to show the formula for its derivative.

For $t \in I$ and $(x,y) \in \mathbb{R}^d \times \mathbb{R}^{d'}$ use the shortcut notations $x_t \coloneqq x_{\mu_t}$, $p_t \coloneqq p_{\mu_t,x,y}$ and $\Phi_t \coloneqq \Phi_{\mu_t,x}$. By the proof of Corollary I.3.2 we know that there exists a subset $\Lambda \subset I$ of full Lebesgue measure such that for every $t \in \Lambda$, the map $t \mapsto x_t$ is differentiable at t for every $x \in \mathbb{R}^d$. By Lebesgue theorem, the map $t \mapsto \mathcal{R}(\mu_t)$ is differentiable at every $t \in \Lambda$ with:

$$\frac{\mathrm{d}}{\mathrm{d}t}\mathcal{R}(\mu_t) = \mathbb{E}_{x,y}\left\langle \nabla_x \ell(x_t(1), y), \delta x_t(1)\right\rangle\,,$$

where, at fixed $x \in \mathbb{R}^d$, δx_t verifies Eq. (I.26) and is given by

$$\delta x_t(1) = \int_{[0,1]\times\Theta} \Phi_t(1)\Phi_t(s)^{-1} \mathrm{D}_\theta \psi(\theta, x_t(s)) v_t(s,\theta) \mathrm{d}\mu_t(s,\theta)\,.$$

Also, the adjoint variable p_t is given by $p_t(s) = \Phi_t(s)^{-\top}\Phi_t(1)^\top \nabla_x \ell(x_t(1), y)$. Hence by inverting the integration order, we see that for $t \in \Lambda$:

$$\frac{\mathrm{d}}{\mathrm{d}t}\mathcal{R}(\mu_t) = \mathbb{E}_{x,y}\left\langle \nabla_x \ell(x_t(1), y), \delta x_t(1)\right\rangle = \int_{[0,1]\times\Theta} \left\langle \mathbb{E}_{x,y} \mathrm{D}_\theta \psi(\theta, x_t(s))^\top p_t(s), v_t(s,\theta)\right\rangle \mathrm{d}\mu_t(s,\theta)\,.$$

By the expression of $\nabla\mathcal{R}[\mu_t]$ in Eq. (III.10), this is the desired result. □

Thanks to Corollary I.3.2 (which can be seen as a chain-rule formula) one can show that the gradient norm $\|\nabla\mathcal{R}[\mu]\|_{L^2(\mu)}$ gives an upper gradient for the risk $\mathcal{R}$ in the sense of Definition I.4. Moreover the following Proposition I.3.4 shows it corresponds to the notion of *local slope* ([Ambrosio, 2008b, Def.1.2.4]). The result relies on the following lemma.

Lemma I.3.3 (Continuity of the adjoint variable p)**.** *Assume ψ satisfies Assumptions I.1 to I.3. Then, for fixed $(x, y) \in \mathbb{R}^{d+d'}$, the map $\mu \mapsto p_{\mu,x,y} \in \mathcal{C}([0,1]\times\mathbb{R}^d)$ is $\mathcal{W}_2^{\mathrm{COT}}$-continuous on $\mathcal{P}_2^{\mathrm{Leb}}([0,1]\times\Theta)$.*

Proof. Let $\mu \in \mathcal{P}_2^{\mathrm{Leb}}([0,1]\times\Theta)$ and consider a sequence $(\mu_n)_{n\geq 0}$ in $\mathcal{P}_2^{\mathrm{Leb}}([0,1]\times\Theta)$ s.t. $\mathcal{W}_2^{\mathrm{COT}}(\mu_n, \mu) \to 0$. Fix a pair $(x, y) \in \mathbb{R}^{d+d'}$ and use the shortcuts $x_n := x_{\mu_n}$ (resp. $x := x_\mu$) and $p_n := p_{\mu_n,x,y}$ (resp. $p := p_{\mu,x,y}$). By Lemma I.3.2 we already have $x_n \to x$ in $\mathcal{C}([0,1])$ and we show now that $p_n \to p$ in $\mathcal{C}([0,1])$ using Ascoli's theorem.

Remark that by the assumptions on ψ, all the trajectories x, x_n, p and p_n stay in a bounded set $B(0,R)$ for some $R \geq 0$. Also, as $\mu_n \to \mu$, we have that the sequence (μ_n) has uniformly integrable second moment and for every $\varepsilon > 0$ we can find a $k \geq 0$ s.t.

$$\int_{\|\theta\|\geq k} (1 + \|\theta\|^2)\mathrm{d}\mu \leq \varepsilon\,, \quad \text{and} \quad \sup_{n\geq 0} \int_{\|\theta\|\geq k} (1 + \|\theta\|^2)\mathrm{d}\mu_n \leq \varepsilon\,.$$

Then for $n \geq 0$ and $s_1 \leq s_2 \in [0,1]$ we have by Eq. (I.13) and the assumptions on ψ:

$$\begin{aligned}
\|p_n(s_2) - p_n(s_1)\| &\leq \int_{s_1}^{s_2} \int_\Theta \|\mathrm{D}_x\psi(\theta, x_n(r))\|\|p_n(r)\|\mathrm{d}\mu_n(r,\theta) \\
&\leq C\int_{s_1}^{s_2} \int_\Theta (1 + \|\theta\|^2)\mathrm{d}\mu_n(r,\theta) \\
&\leq C(\varepsilon + (1+k^2)|s_2 - s_1|)\,,
\end{aligned}$$

where $C = C(R)$. Hence the sequence $(p_n)_{n\geq 0}$ is equicontinuous and, up to a subsequence, we have $p_n \to \bar{p} \in \mathcal{C}([0,1])$. Let us then show $\bar{p} = p$. Indeed using the initial condition we have for $n \geq 0$ and $s \in [0,1]$:

$$p_n(s) = \nabla_x \ell(x_n(1), y) + \int_s^1 \int_\Theta \mathrm{D}_x\psi(\theta, x_n(r))^\top p_n(r)\mathrm{d}\mu_n(r,\theta)\,.$$

First we have $\nabla_x \ell(x_n(1), y) \xrightarrow[n\to\infty]{} \nabla_x \ell(x(1), y)$. Also, note that by the assumptions on ψ we have $\mathrm{D}_x\psi(\theta, x_n(r))^\top p_n(r) \leq C(1+\|\theta\|^2)$ and $\mathrm{D}_x\psi(\theta, x_n(r))^\top p_n(r) \to \mathrm{D}_x\psi(\theta, x(r))^\top \bar{p}(r)$ locally uniformly over $[0,1]\times\Theta$. Hence by the properties of $\mathcal{W}_2$-convergence, we can take the limit in the above equation to obtain:

$$\bar{p}(s) = \nabla_x \ell(x(1), y) + \int_s^1 \int_\Theta \mathrm{D}_x\psi(\theta, x(r))^\top \bar{p}(r)\mathrm{d}\mu(r,\theta)\,,$$

i.e. $\bar{p} = p$ by uniqueness of the solutions to Eq. (I.13). □

Lemma I.3.4 (Continuity of $\|\nabla\mathcal{R}(\mu)\|_{L^2(\mu)}$)**.** *Assume ψ satisfies Assumptions I.1 to I.3. Then the map $\mu \mapsto \|\nabla\mathcal{R}[\mu]\|_{L^2(\mu)}$ is $\mathcal{W}_2^{\mathrm{COT}}$-continuous on $\mathcal{P}_2^{\mathrm{Leb}}([0,1]\times\Theta)$.*

Proof. Let $\mu \in \mathcal{P}_2^{\mathrm{Leb}}([0,1]\times\Theta)$ and consider a sequence $(\mu_n)_{n\geq 0}$ in $\mathcal{P}_2^{\mathrm{Leb}}([0,1]\times\Theta)$ s.t. $\mathcal{W}_2^{\mathrm{COT}}(\mu_n,\mu)\to 0$. For an input $x\in\mathbb{R}^d$, denote by x_n (resp. x) the flow associated to μ_n (resp. μ) and starting from x. Similarly introduce the adjoint variables $(p_n)_{n\geq 0}$ and p. Then by Lemma I.3.2 and Lemma I.3.3 we have that $x_n\to x$ and $p_n\to p$ in $\mathcal{C}([0,1],\mathbb{R}^d)$. As a consequence the sequence of continuous maps

$$f_n : (r,\theta)\mapsto \mathbb{E}_{x,y}\mathrm{D}_\theta\psi(\theta,x_n(r))^\top p_n(r)$$

converges locally uniformly towards the map $f:(r,\theta)\mapsto \mathbb{E}_{x,y}\mathrm{D}_\theta\psi(\theta,x(r))^\top p(r)$ and is uniformly bounded by a function of linear growth. As $\mathcal{W}_2^{\mathrm{COT}}$-convergence implies $\mathcal{W}_2$-convergence and by the properties of $\mathcal{W}_2$-convergence ([Villani, 2009, Thm.6.9]) this implies:

$$\|\nabla\mathcal{R}[\mu_n]\|^2_{L^2(\mu_n)} = \int_{[0,1]\times\Theta}\|f_n\|^2\mathrm{d}\mu_n \xrightarrow[n\to\infty]{} \int_{[0,1]\times\Theta}\|f\|^2\mathrm{d}\mu = \|\nabla\mathcal{R}[\mu]\|^2_{L^2(\mu)}\,.$$

□

Proposition I.3.4 ($\|\nabla\mathcal{R}(\mu)\|_{L^2(\mu)}$ is an upper-gradient)**.** *Assume ψ satisfies Assumptions I.1 to I.3. Let $\mu\in\mathcal{P}_2^{\mathrm{Leb}}([0,1]\times\Theta)$, then $\|\nabla\mathcal{R}[\mu]\|_{L^2}(\mu)$ is the* local slope *of the risk $\mathcal{R}$ at μ, that is:*

$$\|\nabla\mathcal{R}[\mu]\|_{L^2(\mu)} = \limsup_{\nu\to\mu}\frac{(\mathcal{R}(\mu)-\mathcal{R}(\nu))^+}{\mathcal{W}_2^{\mathrm{COT}}(\mu,\nu)}\,. \tag{I.27}$$

Moreover, it is an upper-gradient in the sense of Definition I.4.

Proof. The last part of the result follows from Theorem I.1, since if $(\mu_t)_{t\in I}$ is an absolutely continuous curve then it satisfies the continuity equation with a vector field v such that $\|v_t\|_{L^2(\mu_t)}\leq\left|\frac{\mathrm{d}}{\mathrm{d}t}\mu_t\right|$ for a.e. $t\in I$. Hence by Corollary I.3.3 and Cauchy-Schwarz we have:

$$\forall t_1\leq t_2\in I\,,\quad |\mathcal{R}(\mu_{t_1})-\mathcal{R}(\mu_{t_2})|\leq\int_{t_1}^{t_2}\|\nabla\mathcal{R}[\mu_t]\|_{L^2(\mu_t)}\left|\frac{\mathrm{d}}{\mathrm{d}t}\mu_t\right|\mathrm{d}t\,.$$

Let us then show Eq. (I.27). Consider some parameter $\mu\in\mathcal{P}_2^{\mathrm{Leb}}([0,1]\times\Theta)$ and denote by $|\nabla\mathcal{R}|(\mu)$ the r.h.s. of Eq. (I.27). Then for $\varepsilon>0$, by continuity of $\|\nabla\mathcal{R}[\mu]\|_{L^2(\mu)}$ (Lemma I.3.4) and by definition of $|\nabla\mathcal{R}|(\mu)$ one can find a $\nu\in\mathcal{P}_2^{\mathrm{Leb}}([0,1]\times\Theta)$ s.t.:

$$\frac{(\mathcal{R}(\mu)-\mathcal{R}(\nu))^+}{\mathcal{W}_2^{\mathrm{COT}}(\mu,\nu)}\geq|\nabla\mathcal{R}|(\mu)-\varepsilon$$

and

$$\left|\|\nabla\mathcal{R}[\mu]\|_{L^2(\mu)}-\|\nabla\mathcal{R}[\nu']\|_{L^2(\nu')}\right|\leq\varepsilon$$

if $\mathcal{W}_2^{\mathrm{COT}}(\mu,\nu')\leq\mathcal{W}_2^{\mathrm{COT}}(\mu,\nu)$. Consider $(\mu_t)_{t\in[0,1]}$ a constant speed geodesics with endpoints $\mu_0=\mu$ and $\mu_1=\nu$ (such a geodesic can easily be constructed by similarity

with classical Wasserstein geodesics, see [Ambrosio, 2008b, Thm.7.2.2]). Then by Theorem I.1 the tangent velocity field v of the curve (μ_t) satisfies for $\mathrm{d}t$-a.e. $t \in [0,1]$, $\|v_t\|_{L^2(\mu_t)} \leq \left|\frac{\mathrm{d}}{\mathrm{d}t}\mu_t\right| = \mathcal{W}_2^{\mathrm{COT}}(\mu,\nu)$ and using Corollary I.3.3:

$$\begin{aligned}\mathcal{R}(\nu) &= \mathcal{R}(\mu) + \int_0^1 \langle \nabla\mathcal{R}[\mu_t], v_t\rangle_{L^2(\mu_t)}\,\mathrm{d}t \\ &\leq \mathcal{R}(\mu) + \mathcal{W}_2^{\mathrm{COT}}(\mu,\nu)\int_0^1 \|\nabla\mathcal{R}[\mu_t]\|_{L^2(\mu_t)}\mathrm{d}t \\ &\leq \mathcal{R}(\mu) + \mathcal{W}_2^{\mathrm{COT}}(\mu,\nu)\|\nabla\mathcal{R}[\mu]\|_{L^2(\mu)} + \varepsilon\,.\end{aligned}$$

Similarly we have

$$\mathcal{R}(\mu) \leq \mathcal{R}(\nu) + \mathcal{W}_2^{\mathrm{COT}}(\mu,\nu)\|\nabla\mathcal{R}[\mu]\|_{L^2(\mu)} + \varepsilon$$

and hence $|\nabla\mathcal{R}|(\mu) \leq \|\nabla\mathcal{R}[\mu]\|_{L^2(\mu)} + \varepsilon$.

For the converse inequality consider for $t \in \mathbb{R}$ the parameter $\mu_t = (\mathrm{Id} + t(0, \nabla\mathcal{R}[\mu]))_{\#}\mu$. Then, by Proposition I.3.3 with $v = \nabla\mathcal{R}[\mu]$, the map $t \mapsto \mathcal{R}(\mu_t)$ is differentiable at $t = 0$ and applying the same calculations as in Corollary I.3.3:

$$\frac{\mathrm{d}}{\mathrm{d}t}\mathcal{R}(\mu_t)\Big|_{t=0} = \langle \nabla\mathcal{R}[\mu], v\rangle_{L^2(\mu)} = \|\nabla\mathcal{R}[\mu]\|^2_{L^2(\mu)}.$$

Hence observing that $\mathcal{W}_2^{\mathrm{COT}}(\mu_t,\mu) \leq t\|v\|_{L^2(\mu_t)}$ we have

$$\liminf_{t\to 0^+} \frac{(\mathcal{R}(\mu_t) - \mathcal{R}(\mu))^+}{\mathcal{W}_2^{\mathrm{COT}}(\mu_t,\mu)} \geq \|\nabla\mathcal{R}[\mu]\|_{L^2(\mu)}\,.$$

□

As a consequence of the previous result, we will from now on only consider as upper gradient of $\mathcal{R}$ the one given for every $\mu \in \mathcal{P}_2^{\mathrm{Leb}}([0,1]\times\Theta)$ by:

$$|\nabla\mathcal{R}|(\mu) \coloneqq \|\nabla\mathcal{R}[\mu]\|_{L^2(\mu)} = \left(\int_{[0,1]\times\Theta} \|\mathbb{E}_{x,y}\mathrm{D}_\theta\psi(\theta,x(s))^\top p(s)\|^2 \mathrm{d}\mu(s,\theta)\right)^{1/2}. \tag{I.28}$$

Note that the vector field $\nabla\mathcal{R}[\mu]$ was used in Definition I.3 to define the notion of *gradient flow* whereas the upper-gradient $|\nabla\mathcal{R}|(\mu)$ is used in the Definition I.5 of curves of maximal slope. The following theorem is the main result of this section and shows these two notions coincide.

Theorem I.2. *Assume ψ satisfies Assumptions I.1 to I.3 and ℓ is smooth. Let $I \subset \mathbb{R}$ be an open interval. Then a curve $(\mu_t)_{t\in I}$ is a* gradient flow *in the sense of Definition I.3 if and only if it is a* curve of maximal slope *for $\mathcal{R}$ in the sense of Definition I.5.*

Proof. *Part 1: Gradient flows are curves of maximal slope.*

Let $(\mu_t)_{t\in I}$ be a gradient flow for $\mathcal{R}$ in the sense of Definition I.3. Then (μ_t) is a locally absolutely continuous curve satisfying the continuity equation $\partial_t\mu_t + \mathrm{div}(v_t\mu_t) = 0$ with $v_t = -\nabla\mathcal{R}[\mu_t]$. Hence by Theorem I.1 we have for a.e. $t \in I$:

$$\left|\frac{\mathrm{d}}{\mathrm{d}t}\mu_t\right| \leq \|v_t\|_{L^2(\mu_t)} = \|\nabla\mathcal{R}[\mu_t]\|_{L^2(\mu_t)}\,.$$

Also, by Corollary I.3.2, $(\mathcal{R}(\mu_t))_{t\in I}$ is absolutely continuous with for a.e. $t \in I$:

$$-\frac{\mathrm{d}}{\mathrm{d}t}\mathcal{R}(\mu_t) = \langle v_t, \nabla\mathcal{R}[\mu_t]\rangle_{L^2(\mu_t)} = \|\nabla\mathcal{R}[\mu_t]\|^2_{L^2(\mu_t)}\,.$$

Thus recalling that $|\nabla\mathcal{R}|(\mu) = \|\nabla\mathcal{R}[\mu]\|_{L^2(\mu)}$ we get Definition I.5 by putting together the two previous equations.

Part 2: Curves of maximal slope are gradient flows.

Let $(\mu_t)_{t\in I}$ be a curve of maximal slope for $\mathcal{R}$ in the sense of Definition I.5. Then in particular $(\mu_t)_{t\in I}$ is locally absolutely continuous in $(\mathcal{P}_2^{\mathrm{Leb}}([0,1]\times\Theta)\,,\,\mathcal{W}_2^{\mathrm{COT}})$ and by Theorem I.1 there exists a Borel velocity field $v : I\times[0,1]\times\Theta \to \Theta$ such that μ satisfies the continuity equation:

$$\partial_t\mu_t + \mathrm{div}((0,v_t)\mu_t) = 0, \quad \text{on } I\times[0,1]\times\Theta,$$

and such that the metric derivative satisfies $|\frac{\mathrm{d}}{\mathrm{d}t}\mu_t| \geq \|v_t\|_{L^2(\mu_t)}$ for a.e. $t\in I$. Hence it follows from Corollary I.3.3 that $(\mathcal{R}(\mu_t))_{t\in I}$ is absolutely continuous and for a.e. $t\in I$:

$$-\frac{\mathrm{d}}{\mathrm{d}t}\mathcal{R}(\mu_t) = -\langle v_t, \nabla\mathcal{R}[\mu_t]\rangle\,.$$

Using the EDE condition we thus have:

$$-\langle v_t, \nabla\mathcal{R}[\mu_t]\rangle \geq \frac{1}{2}(|\frac{\mathrm{d}}{\mathrm{d}t}\mu_t|^2 + |\nabla\mathcal{R}|^2(\mu_t)) \geq \frac{1}{2}(\|v_t\|^2_{L^2(\mu_t)} + \|\nabla\mathcal{R}[\mu_t]\|^2_{L^2(\mu_t)})$$

from which it follows by Young's inequality that $v_t = -\nabla\mathcal{R}[\mu_t]$ in $L^2(\mu_t)$ for a.e. $t\in I$. □

Note that, although it does not appear in Definition I.3, the above equivalence shows that if $(\mu_t)_{t\in I}$ is a gradient flow for $\mathcal{R}$ then $\left|\frac{\mathrm{d}}{\mathrm{d}t}\mu_t\right| = \|\nabla\mathcal{R}[\mu_t]\|_{L^2(\mu_t)}$ i.e. $\nabla\mathcal{R}[\mu_t]$ is in fact the (uniquely defined) tangent velocity field of the curve $(\mu_t)_{t\in I}$.

I.3.4 Existence, uniqueness, and stability of gradient flow curves

We show here the well-posedness result for the gradient flow equation of the risk $\mathcal{R}$, namely we show the existence, uniqueness, and stability of gradient flow curves starting from any initialization $\mu_0 \in \mathcal{P}_2^{\mathrm{Leb}}([0,1]\times\Theta)$. For the "existence" part we will rely on classical results from the theory of gradient flows in metric spaces showing the convergence of proximal sequences towards a curve known as *(Generalized) Minimising Movements* [De Giorgi, 1993]. For the "uniqueness" part we will show that gradient flow trajectories are stable, that is if two initializations μ_0, μ_0' are close (in the sense of the metric $\mathcal{W}_2^{\mathrm{COT}}$), then the emanating gradient flow curves $(\mu_t)_{t\geq 0}$, $(\mu_t')_{t\geq 0}$ stay close in finite time.

I.3.4.1 Existence

We proceed to show the existence of gradient flow curves as defined in Definition I.3. For that purpose, we need a strengthening of Assumption I.1. Notably, Assumption I.A allows to show the flow map $\mu \mapsto x_\mu$ is continuous for the topology of narrow convergence over $\mathcal{P}_2^{\mathrm{Leb}}([0,1]\times\Theta)$.

Assumption I.A. *For some $\alpha \in [1,2)$ we assume that:*

(i) The basis function ψ has α-growth w.r.t. θ, locally w.r.t. x. That is for every compact $K \subset \mathbb{R}^d$ there exists a constant $C = C(K)$ s.t:

$$\forall x \in K\,,\ \forall \theta \in \Theta\,, \quad \|\psi(\theta, x)\| \leq C(1 + \|\theta\|^\alpha)\,.$$

(ii) The basis function ψ is continuously differentiable and its differential $\mathrm{D}_x\psi$ w.r.t. x has α-growth w.r.t. θ, locally w.r.t. x. That is for every compact $K \subset \mathbb{R}^d$ there exists a constant $C = C(K)$ s.t.:

$$\forall x \in K\,, \forall \theta \in \Theta\,, \quad \|\mathrm{D}_x\psi(\theta, x)\| \leq C(1 + \|\theta\|^\alpha)\,.$$

Theorem I.3 (Existence of curves of maximal slope)**.** *Assume ψ satisfies Assumptions I.1 to I.3 and Assumption I.A. Let $\mu_0 \in \mathcal{P}_2^{\mathrm{Leb}}([0,1]\times\Theta)$. Then there exists a curve of maximal slope $(\mu_t)_{t\in[0,+\infty)}$ starting from μ_0 and $\left(\left|\frac{\mathrm{d}}{\mathrm{d}t}\mu_t\right|\right)_{t\geq 0} \in L^2_{loc}([0,+\infty))$.*

Proof. The result follows from the successive application of [Ambrosio, 2008b, Thm.2.2.3 and Thm.2.3.3], the first result ensuring the existence of *Generalized Minimizing Movements* and the second result stating that these curves are curves of maximal slope for the local slope. The proof proceeds by verifying the assumptions of these theorems. We consider here $\mathcal{P}_2^{\mathrm{Leb}}([0,1] \times \Theta)$ equipped with the topology induced by the distance $\mathcal{W}_2^{\mathrm{COT}}$ and with the (weaker) topology of narrow convergence, denoted by τ. Note that $(\mathcal{P}_2^{\mathrm{Leb}}([0,1] \times \Theta)\,, \mathcal{W}_2^{\mathrm{COT}})$ is a complete metric space (Proposition I.2.3) and that the distance $\mathcal{W}_2^{\mathrm{COT}}$ is τ-lower-semicontinuous (Lemma I.2.1).

Part 1: $\mathcal{W}_2^{\mathrm{COT}}$-bounded sets are τ-relatively compact.

This property is verified as $\mathcal{W}_2^{\mathrm{COT}}$-bounded sets are tight and hence τ-relatively compact by Prokhorov's theorem.

Part 2: $\mathcal{R}$ is τ-continuous on $\mathcal{W}_2^{\mathrm{COT}}$-bounded sets.

Let (μ_n) be a $\mathcal{W}_2^{\mathrm{COT}}$-bounded sequence in $\mathcal{P}_2^{\mathrm{Leb}}([0,1] \times \Theta)$ such that $\mu_n \xrightarrow{\tau} \mu$ for some $\mu \in \mathcal{P}_2^{\mathrm{Leb}}([0,1] \times \Theta)$ and let us show that $\mathcal{R}(\mu_n) \to \mathcal{R}(\mu)$. Take $x \in \mathbb{R}^d$ and denote by $x_n := x_{\mu_n}$ the flow trajectory starting from x and associated to μ_n. By Lebesgue's theorem, it suffices to show that $x_n(1) \to x_\mu(1)$. Using Ascoli's theorem, we will proceed by showing that $x_n \to x_\mu$ in $\mathcal{C}([0,1], \mathbb{R}^d)$.

From the $\mathcal{W}_2^{\mathrm{COT}}$-boundedness and the proof of Proposition I.1.1, it follows that the trajectories x_n stay in a compact set. Moreover given $s_1 < s_2 \in [0,1]$ we have using the α-growth assumption:

$$\|x_n(s_2) - x_n(s_1)\| \leq \int_{[s_1,s_2]\times\Theta} \|\psi(\theta, x_n(r))\| \mathrm{d}\mu_n(r,\theta) \leq \int_{[s_1,s_2]\times\Theta} C(1 + \|\theta\|^\alpha) \mathrm{d}\mu_n(r,\theta)\,.$$

Also as the sequence (μ_n) is $\mathcal{W}_2^{\mathrm{COT}}$-bounded it has uniformly integrable α-moments. Given $\varepsilon > 0$ we can thus find a $k \geq 0$ such that, for every $n \geq 0$, $\int_{\|\theta\|\geq k} C(1 + \|\theta\|^\alpha)\mathrm{d}\mu_n \leq \varepsilon$. Using this in the previous inequality and the fact that the marginal of μ_n on $[0,1]$ is the Lebesgue measure gives:

$$\forall s_1 < s_2 \in [0,1]\,, \quad \|x_n(s_2) - x_n(s_1)\| \leq \varepsilon + C(1 + k^\alpha)(s_2 - s_1)\,.$$

Thus the trajectories (x_n) are equicontinuous and, by Arzela-Ascoli's theorem, we have (up to a subsequence) that $x_n \to \bar{x}$ in $\mathcal{C}([0,1], \mathbb{R}^d)$.

Let us show that $\bar{x} = x_\mu$ is the flow generated by μ. This will conclude this part of the proof as it will imply $x_n(1) \to x_\mu(1)$ and then $\mathcal{R}(\mu_n) \to \mathcal{R}(\mu)$ by Lebesgue's convergence theorem. Considering $s \in [0,1]$ we have:

$$\begin{aligned} x_n(s) &= x + \int_{[0,1]\times\Theta} \mathbb{1}_{r\le s}\psi(\theta, x_n(r))\mathrm{d}\mu_n(r,\theta) \\ &= x + \int_{[0,1]\times\Theta} \mathbb{1}_{r\le s}\psi(\theta, \bar{x}(r))\mathrm{d}\mu(r,\theta) \\ &\quad + \int_{[0,1]\times\Theta} \mathbb{1}_{r\le s}\left(\psi(\theta, x_n(r)) - \psi(\theta, \bar{x}(s))\right)\mathrm{d}\mu_n(r,\theta) \qquad \text{(L1)} \\ &\quad + \int_{[0,1]\times\Theta} \mathbb{1}_{r\le s}\psi(\theta, \bar{x}(r))\mathrm{d}(\mu_n - \mu)(r,\theta)\,. \qquad \text{(L2)} \end{aligned}$$

In this last equality, we need to show that L1 and L2 vanish as $n \to \infty$. In L1 the integrand has α-growth. Hence, given $\varepsilon > 0$, we have using the uniform integrability of $\|\theta\|^\alpha$ on the sequence (μ_n) that for every $n \ge 0$:

$$\|\int \mathbb{1}_{r\le s}\left(\psi(\theta, x_n(r)) - \psi(\theta, \bar{x}(r))\right)\mathrm{d}\mu_n(r,\theta)\| \le \varepsilon + \int \mathbb{1}_{r\le s, \|\theta\|\le k}\|\psi(\theta, x_n(r)) - \psi(\theta, \bar{x}(r))\|\mathrm{d}\mu_n(r,\theta)\,.$$

Then, as ψ is locally-Lipschitz w.r.t. x and $x_n \to \bar{x}$, we have that the integrand on the r.h.s. converges uniformly to 0 and hence:

$$\limsup_{n\to\infty} \|\int_{[0,1]\times\Theta} \mathbb{1}_{r\le s}\left(\psi(\theta, x_n(r)) - \psi(\theta, \bar{x}(s))\right)\mathrm{d}\mu_n(r,\theta)\| \le \varepsilon\,.$$

In L2 the integrand is not continuous so we can't simply apply the definition of narrow convergence and need to leverage the fact that (μ_n) is a $\mathcal{W}_2^{\mathrm{COT}}$-bounded sequence in $\mathcal{P}_2^{\mathrm{Leb}}([0,1]\times\Theta)$. Note that for every $(r,\theta) \in [0,1]\times\Theta$, $\|\psi(\theta, \bar{x}(r))\| \le C(1 + \|\theta\|^\alpha)$. Given $\varepsilon > 0$ and using the uniform integrability of $\|\theta\|^\alpha$ we can thus have a $k \ge 0$ such that:

$$\sup_{n\ge 0} \int_{\|\theta\|\ge k} C(1 + \|\theta\|^\alpha)\mathrm{d}\mu_n,\ \int_{\|\theta\|\ge k} C(1 + \|\theta\|^\alpha)\mathrm{d}\mu \ \le \varepsilon\,.$$

Then, whenever $\delta > 0$, we can find a continuous function $\varphi : [0,1]\times\Theta \to \mathbb{R}^d$ such that $\|\varphi(r,\theta)\| \le C(1 + \|\theta\|^\alpha)$, $\varphi(r,\theta) = \psi(\theta, \bar{x}(r))$ for every $r \le s$ and $\varphi(r,\theta) = 0$ whenever $r \ge s + \delta$. Considering such a function φ we have for L2:

$$\begin{aligned} \|\int \mathbb{1}_{r\le s}\psi(\theta, \bar{x}(r))\mathrm{d}(\mu_n - \mu)(r,\theta)\| &\le \|\int \varphi\mathrm{d}(\mu_n - \mu)\| + \int \|\mathbb{1}_{r\le s}\psi(\theta, \bar{x}(r)) - \varphi(r,\theta)\|\mathrm{d}(\mu_n + \mu)(r,\theta) \\ &\le \|\int \varphi\mathrm{d}(\mu_n - \mu)\| + 4\varepsilon + C(1 + k^\alpha)\delta \end{aligned}$$

where we used the fact that, in the second term, the integrand has α-growth and is only non-zero for $r \in [s, s+\delta]$. Hence having chosen δ sufficiently small gives:

$$\limsup_{n\to\infty} \|\int \mathbb{1}_{r\le s}\psi(\theta, \bar{x}(r))\mathrm{d}(\mu_n - \mu)(r,\theta)\| \le \limsup_{n\to\infty} \|\int \varphi\mathrm{d}(\mu_n - \mu)\| + 5\varepsilon \le 5\varepsilon\,,$$

where the first lim sup is 0 by definition of narrow convergence. We have thus shown that for every $s \in [0,1]$, taking the limit as $n \to \infty$:

$$\bar{x}(s) = x + \int_0^s \psi(\theta, \bar{x}(r))\mathrm{d}\mu(\theta|r)\mathrm{d}r\,,$$

i.e. $\bar{x} = x_\mu$ is the flow trajectory associated to μ and starting from x.

Part 3: $\|\nabla\mathcal{R}(\mu)\|_{L^2(\mu)}$ *is* τ*-lower semicontinuous on* $\mathcal{W}_2^{\mathrm{COT}}$*-bounded subsets.*

We previously considered the upper-gradient $|\nabla\mathcal{R}|$ defined in Eq. (I.28) as $|\nabla\mathcal{R}|(\mu) \coloneqq \|\nabla\mathcal{R}[\mu]\|_{L^2(\mu)}$. However, Ambrosio, Gigli, and Savaré [Ambrosio, 2008b, Thm.2.3.3] state that the obtained curve of maximal slope is a curve of maximal slope for another definition of gradient, referred to as *relaxed slope*. To show these two notions coincide here we show that the map $\mu \mapsto \|\nabla\mathcal{R}[\mu]\|_{L^2(\mu)}$ is τ-lower-semicontinuous on $\mathcal{W}_2^{\mathrm{COT}}$-bounded subsets (*c.f.* [Ambrosio, 2008b, Rm.2.3.4]).

As before consider a $\mathcal{W}_2^{\mathrm{COT}}$-bounded sequence (μ_n) in $\mathcal{P}_2^{\mathrm{Leb}}([0,1]\times\Theta)$ which narrowly converges towards some $\mu \in \mathcal{P}_2^{\mathrm{Leb}}([0,1]\times\Theta)$. Then we previously have shown that for every $x \in \mathbb{R}^d$ we have $x_{\mu_n} \to x_\mu$ in $\mathcal{C}([0,1],\mathbb{R}^d)$. Proceeding with similar arguments one could show the same for the adjoint variable p, that is $p_n \coloneqq p_{\mu_n,x,y} \to p_{\mu,x,y}$ in $\mathcal{C}([0,1],\mathbb{R}^d)$. Then using a generalization of Fatou's lemma with varying measures (*e.g.* [Feinberg, 2020, Thm.2.4]) we have:

$$\begin{aligned}
\liminf_{n\to\infty} \|\nabla\mathcal{R}[\mu_n]\|^2_{L^2(\mu_n)} &= \liminf_{n\to\infty} \int_{[0,1]\times\Theta} \|\mathbb{E}_{x,y}\mathrm{D}_\theta\psi(\theta, x_n(r))^\top p_n(r)\|^2 \mathrm{d}\mu_n(r,\theta) \\
&\geq \int_{[0,1]\times\Theta} \liminf_{\substack{n\to\infty \\ (r',\theta')\to(r,\theta)}} \|\mathbb{E}_{x,y}\mathrm{D}_\theta\psi(\theta', x_n(r'))^\top p_n(r')\|^2 \mathrm{d}\mu(r,\theta) \\
&= \|\nabla\mathcal{R}[\mu]\|^2_{L^2(\mu)}\,,
\end{aligned}$$

which is the desired property. □

I.3.4.2 Uniqueness

We present here a uniqueness result for solutions of the gradient flow equation, which is the content of the following Theorem I.4. The proof is standard and relies on the lipschitzness of the gradient vector field $\nabla\mathcal{R}[\mu]$ w.r.t. the measure μ. It uses the following Assumption I.B on the basis function ψ to ensure local lipschitzness of the adjoint variable p (Lemma I.3.5).

Assumption I.B. *Assume that* ψ *is twice continuously differentiable with* $\mathrm{D}^2_{\theta,\theta}\psi$ *uniformly bounded,* $\mathrm{D}^2_{\theta,x}\psi$ *having linear growth and* $\mathrm{D}^2_{x,x}\psi$ *having quadratic growth w.r.t.* θ*. Namely, for every* $R \geq 0$ *there exists a constant* $C = C(R)$ *s.t. for every* $x, x' \in B(0,R)$ *and every* $\theta, \theta' \in \Theta$ *it holds:*

$$\|\mathrm{D}^2_{\theta,\theta}\psi(\theta,x)\| \leq C(R)\,, \quad \|\mathrm{D}^2_{\theta,x}\psi(\theta,x)\| \leq C(R)(1+\|\theta\|)\,, \quad \|\mathrm{D}^2_{x,x}\psi(\theta,x)\| \leq C(R)(1+\|\theta\|^2)\,.$$

Theorem I.4 (Uniqueness of curves of maximal slope)**.** *Assume* ψ *satisfies Assumptions I.1 to I.3 and Assumption I.B and that* $\nabla_x\ell$ *is locally Lipschitz w.r.t.* x*. Let* $\mu_0 \in \mathcal{P}_2^{\mathrm{Leb}}([0,1]\times\Theta)$*. Then the gradient flow for the risk* $\mathcal{R}$ *starting from* μ_0*, if it exists, is unique.*

Proof. Let $(\mu_t)_{t\geq 0}$ and $(\mu'_t)_{t\geq 0}$ be two gradient flow curves for the risk $\mathcal{R}$ starting from μ_0. We will proceed to show that $\mathcal{W}_2^{\mathrm{COT}}(\mu_t,\mu'_t) = 0$ for every $t \geq 0$.

We use the shorter notations $v_t \coloneqq \nabla\mathcal{R}[\mu_t]$, $v'_t \coloneqq \nabla\mathcal{R}[\mu'_t]$ to refer to the *tangent vector fields* of μ and μ' respectively. Observe that the map $t \mapsto \mathcal{W}_2^{\mathrm{COT}}(\mu_t,\mu'_t)^2$ is locally absolutely continuous and by Lemma I.2.3 its differential is given at almost every $t \geq 0$ by:

$$\frac{\mathrm{d}}{\mathrm{d}t}\mathcal{W}_2^{\mathrm{COT}}(\mu_t,\mu'_t)^2 = 2\int_0^1 \int_\Theta \langle \theta' - \theta, v'_t(s,\theta') - v_t(s,\theta)\rangle \,\mathrm{d}\gamma_t(s,\theta,\theta'),$$

where $\gamma_t \in \Gamma_o^{\mathrm{Leb}}(\mu_t, \mu'_t)$ can be any optimal coupling.

Let $T \geq 0$ and define $\mathcal{E} \coloneqq \sup_{t\in[0,T]} \max(\mathcal{E}_2(\mu_t), \mathcal{E}_2(\mu'_t)) < \infty$. Fix some $t \in [0,T]$ and consider (x,y) in the support of the data ditsribution $\mathcal{D}$ with the shortcuts $x_t \coloneqq x_{\mu_t}$, $p_t \coloneqq p_{\mu_t,x,y}$ and similarly x'_t, p'_t for μ'_t. As the data distribution has compact support, we have that there exists some $R = R(\mathcal{E})$ such that $\|x_t(s)\|$, $\|x'_t(s)\|$, $\|p_t(s)\|$, $\|p'_t(s)\| \leq R$. Then using Lemmas I.3.2 and I.3.5 as well as the assumptions on ψ we have for every $(s,\theta,\theta') \in [0,1]\times\Theta^2$:

$$\begin{aligned}
&\|\mathrm{D}_\theta\psi(\theta, x_t(s))^\top p_t(s) - \mathrm{D}_\theta\psi(\theta', x'_t(s))^\top p'_t(s)\| \\
&\qquad \leq \|\mathrm{D}_\theta\psi(\theta, x_t(s))^\top p_t(s) - \mathrm{D}_\theta\psi(\theta', x_t(s))^\top p_t(s)\| \\
&\qquad\qquad + \|\mathrm{D}_\theta\psi(\theta', x_t(s))^\top p_t(s) - \mathrm{D}_\theta\psi(\theta', x'_t(s))^\top p'_t(s)\| \\
&\qquad \leq \|\mathrm{D}_\theta\psi(\theta, x_t(s)) - \mathrm{D}_\theta\psi(\theta', x_t(s))\|\|p_t(s)\| \\
&\qquad\qquad + \|\mathrm{D}_\theta\psi(\theta', x_t(s)) - \mathrm{D}_\theta\psi(\theta', x'_t(s))\|\|p_t(s)\| \\
&\qquad\qquad + \|\mathrm{D}_\theta\psi(\theta', x'_t(s))\|\|p_t(s) - p'_t(s)\| \\
&\leq C_1\|\theta - \theta'\| + C_1(1 + \|\theta'\|)\mathcal{W}_2^{\mathrm{COT}}(\mu_t, \mu'_t) + C_1(1 + \|\theta'\|)\mathcal{W}_2^{\mathrm{COT}}(\mu_t, \mu'_t)\,, \qquad \text{(I.29)}
\end{aligned}$$

with $C_1 = C_1(\mathcal{E})$ some constant. Fixing some $\gamma_t \in \Gamma_o^{\mathrm{Leb}}(\mu_t, \mu'_t)$, using that $2\langle a, b\rangle \leq \|a\|^2 + \|b\|^2$ and integrating the previous inequality over (x,y) and (s,θ,θ') we get:

$$\frac{\mathrm{d}}{\mathrm{d}t}\mathcal{W}_2^{\mathrm{COT}}(\mu_t, \mu'_t)^2 \leq \int_{[0,1]\times\Theta^2} \left(\|\theta - \theta'\|^2 + \|v_t(s,\theta) - v'_t(s,\theta')\|^2\right) \mathrm{d}\gamma_t(s,\theta,\theta') \leq C_2\mathcal{W}_2^{\mathrm{COT}}(\mu_t, \mu'_t)^2\,,$$

for some constant $C_2 = C_2(\mathcal{E})$. We can then conclude using Grönwall's inequality to:

$$\forall t \in [0,T]\,, \quad \mathcal{W}_2^{\mathrm{COT}}(\mu_t, \mu'_t)^2 \leq e^{C_2 t}\mathcal{W}_2^{\mathrm{COT}}(\mu_0, \mu_0)^2 = 0\,.$$

□

The above proof relied on the following lemma, showing that the adjoint variable map $\mu \mapsto p_{\mu,x,y}$ is locally Lipschitz under Assumption I.B.

Lemma I.3.5. *Assume ψ satisfies Assumptions I.1 to I.3 and Assumption I.B and that $\nabla_x\ell$ is locally Lipschitz w.r.t. x. Then for fixed $(x,y) \in \mathbb{R}^d \times \mathbb{R}^{d'}$ the adjoint variable map $\mu \in \mathcal{P}_2^{\mathrm{Leb}}([0,1]\times\Theta) \mapsto p_{\mu,x,y} \in \mathcal{C}([0,1], \mathbb{R}^d)$ is locally Lipschitz. Namely, for every $\mathcal{E} \geq 0$ there exists a constant $C = C(\mathcal{E})$ such that:*

$$\sup_{s\in[0,1]} \|p_{\mu,x,y}(s) - p_{\mu',x,y}(s)\| \leq C\mathcal{W}_2^{\mathrm{COT}}(\mu, \mu')$$

for every parameterization μ, μ' such that $\mathcal{E}_2(\mu)$, $\mathcal{E}_2(\mu') \leq \mathcal{E}$. Moreover, the constant C can be chosen uniformly over (x,y) in a compact subset.

Proof. Consider $(x,y) \in \mathbb{R}^d \times \mathbb{R}^{d'}$, $\mathcal{E} \geq 0$ and parameterizations μ, μ' as in the proposition. We denote by $(x(s))$ and $(x'(s))$ the forward flows and $(p(s))$ and $(p'(s))$ the backward flows associated to x, y and to μ and μ' respectively. Let $R \geq 0$ be such that $\|x\| + \|y\| \leq R$. Using Proposition I.1.1 and Eq. (I.13) we can assume that the trajectories x, x', p and p' are uniformly bounded by some $R' = R'(R, \mathcal{E})$. Then we get from Eq. (I.13) that at every $s \in [0,1]$:

$$\begin{aligned}
\|p(s) - p'(s)\| &\leq \|p(1) - p'(1)\| \\
&\quad + \int_s^1 \| \int_\Theta \mathrm{D}_x\psi(\theta, x(r))^\top p(r)\mathrm{d}\mu(\theta|r) - \int_\Theta \mathrm{D}_x\psi(\theta, x'(r))^\top p'(r)\mathrm{d}\mu'(\theta|r)\|\mathrm{d}r.
\end{aligned}$$

Fixing $r \in [s,1]$, the integrand on the r.h.s. can be decomposed as:

$$\begin{aligned}\| \int_\Theta \mathrm{D}_x\psi(\theta, x(r))^\top p(r)\mathrm{d}\mu(\theta|r) - \int_\Theta \mathrm{D}_x\psi(\theta, x'(r))^\top p'(r)\mathrm{d}\mu'(\theta|r)\| \\ \leq \int_\Theta \|\mathrm{D}_x\psi(\theta, x(r)) - \mathrm{D}_x\psi(\theta, x'(r))\|\|p(r)\|\mathrm{d}\mu(\theta|r) \\ + \int_\Theta \|\mathrm{D}_x\psi(\theta, x'(r))\|\|p(r) - p'(r)\|\mathrm{d}\mu(\theta|r) \\ + \| \int_\Theta \mathrm{D}_x\psi(\theta, x'(r))p'(r)\mathrm{d}(\mu' - \mu)(\theta|r)\| \\ =: I_1(r) + I_2(r) + I_3(r)\,.\end{aligned}$$

Then using the assumptions on ψ and Lemma I.3.2:

$$\begin{aligned}I_1(r) &\leq C_1 \mathcal{W}_2^{\mathrm{COT}}(\mu, \mu') \int_\Theta (1 + \|\theta\|^2)\mathrm{d}\mu(\theta|r)\,, \\ I_2(r) &\leq C_2\|p(r) - p'(r)\| \int_\Theta (1 + \|\theta\|^2)\mathrm{d}\mu(\theta|r)\,,\end{aligned}$$

with $C_1 = C_1(R, \mathcal{E})$ and $C_2 = C_2(R, \mathcal{E})$. For I_3, considering $\gamma \in \Gamma_o(\mu(.|r), \mu'(.|r))$ gives:

$$\begin{aligned}I_3(r) &\leq \int_{\Theta^2} \|\mathrm{D}_x\psi(\theta, x'(r)) - \mathrm{D}_x\psi(\theta', x'(r))\|\|p'(r)\|\mathrm{d}\gamma(\theta, \theta') \\ &\leq C_3(1 + \mathcal{E}_2(\mu(.|r)) + \mathcal{E}_2(\mu'(.|r)))^{1/2}\mathcal{W}_2(\mu(.|r), \mu'(.|r))\,,\end{aligned}$$

with $C_3 = C_3(R, \mathcal{E})$. Assembling all the previous inequalities we get by Grönwall's lemma:

$$\|p(s) - p'(s)\| \leq e^{C_2(1+\mathcal{E}_2(\mu))} \begin{pmatrix} \|p(1) - p'(1)\| \\ + \mathcal{W}_2^{\mathrm{COT}}(\mu, \mu') \left(C_1(1 + \mathcal{E}_2(\mu)) + C_3(1 + \mathcal{E}_2(\mu) + \mathcal{E}_2(\mu'))^{1/2}\right) \end{pmatrix}.$$

To conclude it suffices to note that by definition $p(1) = \nabla_x\ell(x(1), y)$ and $p'(1) = \nabla_x\ell(x'(1), y)$ and using Lemma I.3.2 with the assumptions on ℓ:

$$\|p(1) - p'(1)\| \leq C_4 \mathcal{W}_2^{\mathrm{COT}}(\mu, \mu')\,, \quad \text{where } C_4 = C_4(R, \mathcal{E}).$$

□

I.3.4.3 Stability

We now turn to a stability result on the gradient flow equation. The following Theorem I.5 is stronger than the above Theorem I.4. It implies that if a sequence of initializations $(\mu_0^n)_{n\geq 0}$ $\mathcal{W}_2^{\mathrm{COT}}$-converges to some initialization μ_0 then the associated gradient flows $(\mu_t^n)_{n\geq 0}$ $\mathcal{W}_2^{\mathrm{COT}}$-converge to μ_t, uniformly over finite time intervals. For simplicity we consider here that ℓ is the square loss $\ell : (x, y) \mapsto \frac{1}{2}\|x - y\|^2$, but the result could be extended to any other loss satisfying $\|\nabla_x\ell\| \leq \varphi(\ell)$ for a concave increasing function φ. We also consider the following supplementary assumptions allowing to control the growth of $\mathcal{E}_2(\mu_t)$ along the gradient flow (Lemma I.3.6).

Assumption I.C. *Assume that ψ is continuously differentiable and such that $\mathrm{D}_x\psi$ is uniformly bounded and $\mathrm{D}_\theta\psi$ is of linear growth w.r.t. θ. Namely, there exists an absolute constant C s.t.:*

$$\forall x \in \mathbb{R}^d\,,\ \forall \theta \in \Theta\,, \quad \|\mathrm{D}_x\psi(\theta, x)\| \leq C\,, \quad \|\mathrm{D}_\theta\psi(\theta, x)\| \leq C(1 + \|\theta\|)\,.$$

Theorem I.5 (Stability of curves of maximal slope)**.** *Assume ψ satisfies Assumptions I.1 to I.3 and Assumptions I.B and I.C and assume ℓ is the square loss. Let $(\mu_t)_{t\geq 0}$, $(\mu'_t)_{t\in t\geq 0}$ be gradient flow curves for the risk $\mathcal{R}$ starting from $\mu_0, \mu'_0 \in \mathcal{P}_2^{\mathrm{Leb}}([0,1]\times\Theta)$ respectively and let $\mathcal{E}_0$ be such that $\mathcal{E}_2(\mu_0), \mathcal{E}_2(\mu'_0) \leq \mathcal{E}_0$. Then for every $T \geq 0$ there exists a constant $C = C(\mathcal{E}_0, T)$ such that:*

$$\forall t \in [0,T]\,, \quad \mathcal{W}_2^{\mathrm{COT}}(\mu_t, \mu'_t) \leq e^{Ct}\mathcal{W}_2^{\mathrm{COT}}(\mu_0, \mu'_0)\,.$$

Proof. Let $T \geq 0$. By the energy bound of Lemma I.3.6 below, we know that we can find a $\mathcal{E} = \mathcal{E}(\mathcal{E}_0, T)$ such that for every $t \in [0,T]$:

$$\mathcal{E}_2(\mu_t)\,,\, \mathcal{E}_2(\mu'_t) \leq \mathcal{E}.$$

Then using Assumption I.B and proceeding as in the proof of the above Theorem I.4 (c.f. Eq. (I.29)) we get a constant $C = C(\mathcal{E})$ such that for $\mathrm{d}t$-a.e. $t \in [0,T]$:

$$\frac{\mathrm{d}}{\mathrm{d}t}\mathcal{W}_2^{\mathrm{COT}}(\mu_t, \mu'_t)^2 \leq C\mathcal{W}_2^{\mathrm{COT}}(\mu_t, \mu'_t)^2\,,$$

which gives the result using Grönwall's inequality. □

In the above proof, we used the following technical result giving an upper bound on the energy $\mathcal{E}_2(\mu_t)$ along a gradient flow curve $(\mu_t)_{t\geq 0}$.

Lemma I.3.6. *Assume ψ satisfies Assumptions I.1 to I.3 and Assumptions I.B and I.C and ℓ is the square loss. Let $(\mu_t)_{t\geq 0}$ be a gradient flow for the risk $\mathcal{R}$ and let $\mathcal{E} \geq 0$ be s.t. $\mathcal{E}_2(\mu_0) \leq \mathcal{E}$. Then there exists a constant $C = C(\mathcal{E})$ such that $\mathcal{E}_2(\mu_t) \leq e^{Ct}(\mathcal{E}_2(\mu_0) + Ct)$ for every $t \geq 0$.*

Proof. For $(x,y) \in \mathbb{R}^d \times \mathbb{R}^{d'}$ use the shortcuts $x_t := x_{\mu_t}$, $p_t := p_{\mu_t,x,y}$. Note that the map $t \mapsto \mathcal{E}_2(\mu_t) = \mathcal{W}_2^{\mathrm{COT}}(\mu_t, \mathrm{Leb}([0,1]) \otimes \delta_0)^2$ is locally absolutely continuous and that by Lemma I.2.3 its derivative is given at $\mathrm{d}t$-a.e. $t \geq 0$ by:

$$\frac{\mathrm{d}}{\mathrm{d}t}\mathcal{E}_2(\mu_t) = 2\int_{[0,1]\times\Theta} \Big\langle \theta, \mathbb{E}_{x,y}\mathrm{D}_\theta\psi(\theta, x_t(s))^\top p_t(s)\Big\rangle \,\mathrm{d}s\,.$$

By Assumption I.C there exists an absolute constant C_1 such that $\|\mathrm{D}_x F_{\mu_t(.|s)}\| \leq C_1$ and hence $\|p_t(s)\| \leq e^{C_1}\|p_t(1)\|$ for every $s \in [0,1]$. Using the initial condition on $p_t(1)$ and the fact that ℓ is the quadratic loss:

$$\mathbb{E}_{x,y}\|p_t(s)\| \leq e^{C_1}\mathbb{E}_{x,y}\|x_t(1) - y\| \leq C_2\sqrt{\mathcal{R}(\mu_t)} \leq C_2\sqrt{\mathcal{R}(\mu_0)}\,,$$

for some universal constant C_2. Then using that by Assumption I.C we have $\|\mathrm{D}_\theta\psi(\theta, x)\| \leq C_1(1 + \|\theta\|)$ and with the previous inequality we get:

$$\frac{\mathrm{d}}{\mathrm{d}t}\mathcal{E}_2(\mu_t) \leq 2C_1C_2\left(\mathcal{E}_2(\mu_t) + \sqrt{\mathcal{E}_2(\mu_t)}\right)\sqrt{\mathcal{R}(\mu_0)}\,.$$

Noting that $\mathcal{R}(\mu_0) \leq C_3$ for some constant $C_3 = C_3(\mathcal{E})$, the result follows by Grönwall's inequality □

Appendices

I.A Well-posedness of the gradient flow equation for SHL residuals

While Theorems I.3 and I.4 show existence and uniqueness to gradient flow equation of the training risk $\mathcal{R}$ under mild assumptions on the basis function ψ, those assumptions are however not met for residuals which are 2-perceptrons as defined in Eq. (34). Indeed, in this case ψ is of the form:

$$\psi((u,w,b),x) = u\sigma(w^\top x + b)\,, \tag{I.30}$$

where $x \in \mathbb{R}^d$, (u,w,b) are parameters in $\Theta = \mathbb{R}^d \times \mathbb{R}^d \times \mathbb{R}$ and $\sigma : \mathbb{R} \to \mathbb{R}$ is a non-linear activation. In particular, we will be considering such residuals in Section II.5. We thus justify here why the existence and uniqueness results of Theorems I.3 and I.4 still apply in the case of the SHL architecture where ψ is given by Eq. (I.30), even if Assumptions I.A and I.B are not satisfied. The idea is to restrict ourselves to compactly supported parameterizations $\mu \in \mathcal{P}_2^{\mathrm{Leb}}([0,1]\times\Theta)$ where both assumptions are satisfied $\mathrm{d}\mu$ almost everywhere.

In the rest of this section, we consider the parameter space $\Theta = \mathbb{R}^d \times \mathbb{R}^d \times \mathbb{R}$ and the basis function is supposed to be given by Eq. (I.30) with some activation σ satisfying Assumption II.3. Note in particular that Assumptions I.1 to I.3 are satisfied and that the representation result of Proposition I.3.2 holds. The following preliminary result states that if the initialization μ_0 is compactly supported, so is a solution μ_t of the gradient flow at every time $t \geq 0$.

Lemma I.A.1. *Assume ψ if of the form Eq. (I.30) with some activation σ satisfying Assumption II.3. Let $\mu_0 \in \mathcal{P}_2^{\mathrm{Leb}}([0,1]\times\Theta)$ be some compactly supported initialization with $\mathrm{Supp}(\mu_0) \subset B(0,R_0)$ for some $R_0 \geq 0$. If $(\mu_t)_{t\geq 0}$ is a gradient flow of the risk $\mathcal{R}$ then for every $T \geq 0$ there exists $R_T \geq 0$ such that:*

$$\forall t \in [0,T]\,, \quad \mathrm{Supp}(\mu_t) \subset B(0,R_T)\,. \tag{I.31}$$

Proof. Let $(\mu_t)_{t\geq 0}$ be such as in the statement. In view of Proposition I.3.2 such a gradient flow is given for every $t \geq 0$ by $\mu_t = (X_t)_{\#}\mu_0$, where X is a solution of Eq. (I.22). Let us then consider some $T \geq 0$. The energy $\mathcal{E}_2(\mu_t)$ is a continuous function along the gradient flow time t so that $\mathcal{E} \coloneqq \sup_{t\in[0,T]} \mathcal{E}_2(\mu_t) < \infty$. Then using Assumption II.3 we have a constant $C = C(\mathcal{E})$ such that for every $t \in [0,T]$ and every $(s,\theta) \in [0,1]\times\Theta$:

$$\|\frac{\mathrm{d}}{\mathrm{d}t} X_t(s,\theta)\| \leq C(1 + \|X_t(s,\theta)\|)\,.$$

Hence by Grönwall's lemma:

$$\|X_t(s,\theta)\| \leq e^{Ct}(\|X_0(s,\theta)\| + Ct)\,,$$

from which the result follows by taking $R_T = e^{CT}(R_0 + CT)$. □

Note that ψ may not satisfy Assumption I.B when considering $\theta \in \Theta$ but it does when considering θ in bounded regions $B(0,R) \subset \Theta$. Hence using the above Lemma I.A.1 and restricting ourselves to finite time intervals, one can show uniqueness of the gradient flow curves whenever the initialization is compactly supported.

Proposition I.A.1. *Assume ψ if of the form Eq. (I.30) with some activation σ satisfying Assumption II.3. Let $\mu_0 \in \mathcal{P}_2^{\mathrm{Leb}}([0,1]\times\Theta)$ be some compactly supported initialization. Then the gradient flow $(\mu_t)_{t\geq 0}$ of the risk $\mathcal{R}$ starting from μ_0, if it exists, is unique.*

Proof. Let $(\mu_t)_{t\geq 0}, (\mu'_t)_{t\geq 0}$ be two gradient flow curves starting from μ_0. We will proceed to show that $\mathcal{W}_2^{\mathrm{COT}}(\mu_t,\mu'_t)=0$ for every $t\geq 0$.

Fix some $T\geq 0$. By the above Lemma I.A.1, we can find some $R\geq 0$ such that for every $t\in[0,T]$ we have $\mathrm{Supp}(\mu_t),\mathrm{Supp}(\mu'_t)\subset B(0,R)$. Then note that ψ satisfies Assumption I.B when restricted to $\theta\in B(0,R)$ in the sense that for every compact set $K\subset\mathbb{R}^d$ there exists a constant $C=C(K,R)$ s.t. for every $x,x'\in K$ and $\theta,\theta'\in B(0,R)$:

$$\|\mathrm{D}^2_{\theta,\theta}\psi(\theta,x)\|\leq C,\quad \|\mathrm{D}^2_{\theta,x}\psi(\theta,x)\|\leq C(1+\|\theta\|)\,,\quad \|\mathrm{D}^2_{x,x}\psi(\theta,x)\|\leq C(1+\|\theta\|^2)\,.$$

Also note that, for every $t\in[0,T]$ and every optimal Conditional OT coupling $\gamma_t\in\Gamma_o^{\mathrm{diag}}(\mu_t,\mu'_t)$, we have that γ_t is compactly supported with $\mathrm{Supp}(\gamma_t)\subset B(0,R)\times B(0,R)$. Hence proceeding as in the proof of Theorem I.4 (c.f. Eq. (I.29)) we find a constant $C=C(R)$ s.t. for $\mathrm{d}t$-a.e. $t\in[0,T]$:

$$\frac{\mathrm{d}}{\mathrm{d}t}\mathcal{W}_2^{\mathrm{COT}}(\mu_t,\mu'_t)^2\leq C\mathcal{W}_2^{\mathrm{COT}}(\mu_t,\mu'_t)^2\,,$$

from which it follows by Grönwall's inequality that

$$\mathcal{W}_2^{\mathrm{COT}}(\mu_t,\mu'_t)^2\leq e^{Ct}\mathcal{W}_2^{\mathrm{COT}}(\mu_0,\mu_0)^2=0\,.$$

□

Finally, the following result states the existence of a gradient flow curve of the risk $\mathcal{R}$ for the SHL architecture when the initialization is compactly supported.

Proposition I.A.2. *Assume ψ if of the form Eq. (I.30) with some activation σ satisfying Assumption II.3. Let $\mu_0\in\mathcal{P}_2^{\mathrm{Leb}}([0,1]\times\Theta)$ be some compactly supported initialization. Then there exists a gradient flow $(\mu_t)_{t\geq 0}$, defined for every $t\geq 0$, for the risk $\mathcal{R}$ starting at μ_0.*

Proof. Let μ_0 be such as in the statement and consider some $T_0\geq 0$. Then by the previous Proposition I.A.1, the gradient flow $(\mu_t)_{t\geq 0}$ starting from μ_0, if it exists, is unique and by Lemma I.A.1 there exists a $R_{T_0}>0$ such $\mathrm{Supp}(\mu_t)\subset B(0,R_{T_0})$ for every $t\in[0,T_0]$.

It is then easy to modify ψ into some $\tilde{\psi}$ such that $\psi(\theta,x)=\tilde{\psi}(\theta,x)$ whenever $\theta\in B(0,2R_{T_0})$, $\tilde{\psi}$ satisfies Assumptions I.1 to I.3 but also Assumption I.A. For example, consider for every $x\in\mathbb{R}^d$ and $(u,w,b)\in\Theta$:

$$\tilde{\psi}((u,w,b),x):=\pi(u)\sigma(w^\top x+b)\,,$$

for some smooth map $\pi:\mathbb{R}^d\to\mathbb{R}^d$ (depending on R_{T_0}) such that $\|\pi\|$ and $\|\mathrm{D}\pi\|$ are uniformly bounded and $\pi(u)=u$ if $\|u\|<2R_{T_0}$. Denote by $\tilde{L}$ the modified risk associated to the modified basis function $\tilde{\psi}$. Then Theorem I.3 applies and there exists a gradient flow $(\tilde{\mu}_t)_{t\geq 0}$ for the modified risk $\tilde{L}$ starting from μ_0. Consider the the time T defined by:

$$T^*=\sup\left\{T\geq 0\ :\ \mathrm{Supp}(\tilde{\mu}_t)\subset B(0,2R_{T_0})\,,\ \forall t\in[0,T]\right\}\,.$$

Note that by the definition of T^* and $\tilde{\psi}$, if $T<T^*$ then for every $t\in[0,T]$, $\nabla\tilde{L}[\tilde{\mu}_t]=\nabla\mathcal{R}[\tilde{\mu}_t]$ in $L^2(\tilde{\mu}_t)$ and hence $(\tilde{\mu}_t)_{t\in[0,T]}$ is a gradient flow for the original risk $\mathcal{R}$, starting

from μ_0. We show by contradiction that $T^* > T_0$, implying that there exists a gradient flow for $\mathcal{R}$ starting from μ_0 and defined up to time T_0.

Assume $T^* \leq T_0$. Consider $\mathcal{E} \coloneqq \sup_{t\in[0,T_0+1]} \mathcal{E}_2(\tilde{\mu}_t)$. For any $T < T^*$, we have that $(\tilde{\mu}_t)_{t\in[0,T]}$ is a gradient flow for $\mathcal{R}$ starting from μ_0 and in particular $\mathrm{Supp}(\tilde{\mu}_T) \subset B(0, R_{T_0})$. But then, reasoning as in the proof of Lemma I.A.1, there exists a constant $C = C(\mathcal{E})$ (independent of T) such that for every $\varepsilon \in [0, 1]$:

$$\forall t \in [T, T+\varepsilon]\,, \quad \mathrm{Supp}(\tilde{\mu}_t) \subset B(0, e^{C\varepsilon}(R_{T_0} + C\varepsilon))\,,$$

which is included in $B(0, 2R_{T_0})$ for ε sufficiently small. Hence chosing T sufficiently close from T^* we get a $T + \varepsilon > T^*$ such that $\mathrm{Supp}(\tilde{\mu}_t) \subset B(0, 2R_{T_0})$ for every $t \in [0, T+\varepsilon]$. This is in contradiction with the definition of T^*. □

Chapter II

Convergence in the training of residual architectures: Polyak-Łojasiewicz inequalities and local convergence guarantees

Contents

II.1 Introduction **79**
II.1.1 Related works and contributions 81
II.2 Polyak-Łojasiewicz property and convergence of gradient flow **83**
II.2.1 The Polyak-Łojasiewicz property in Hilbert spaces 83
II.2.2 The Polyak-Łojasiewicz property in metric spaces 86
II.3 Convergence for general architectures **89**
II.3.1 Conditioning of the tangent kernel implies the P-Ł property 89
II.3.2 Expressivity and functional properties of the set of residuals 91
II.4 Linear parameterization of the residuals **92**
II.4.1 Gradient flow equation in the case of RKHS residuals 96
II.4.2 Convergence of RKHS-NODE 99
II.4.3 Convergence with finite width 101
II.5 The case of SHL residuals **104**
II.5.1 Comparison with the case of a linear parameterization 106
II.5.2 Convergence of NODEs with SHL residuals 107
II.5.3 Examples of activations and quantitative convergence results 108
II.6 Ensuring convergence with lifting and scaling **111**
II.7 Numerical results **112**
II.7.1 Experiments on MNIST 113
II.7.2 Experiments on CIFAR10 116
II.8 Conclusion **118**

II.1 Introduction

A central question in modern machine learning is to understand why neural networks perform so well in practice, despite the apparent complexity of their training dynamics. At the heart of this process lies a challenging non-convex optimization problem, typically

approached using first order optimization methods such as *(stochastic) gradient descent*. However, while there has been an important amount of work on the subject [Hardt, 2016a; Bartlett, 2018; Zou, 2019; Li, 2017; Li, 2018; Du, 2018; Du, 2019; Allen-Zhu, 2019; Lee, 2019; Zou, 2020; Liu, 2020; Chen, 2020; Nguyen, 2021; Marion, 2023b], convergence properties of those algorithms still lacks theoretical understanding. Due to the exponential increase in the size of state-of-the-art models, a particular focus was placed on overparameterized architectures, whose number of parameters is very high w.r.t. the number of data. Following Eq. (36), such architectures take the form of mappings $F_{(\theta_i)_{1\leq i\leq M}} : \mathbb{R}^d \to \mathbb{R}^d$ defined by:

$$F_{(\theta_i)_{1\leq i\leq M}} : x \in \mathbb{R}^d \mapsto \frac{1}{M}\sum_{i=1}^{M} \psi(\theta_i, x)\,, \tag{II.1}$$

where Θ is some space of parameters, $\psi : \Theta \times \mathbb{R}^d \to \mathbb{R}^d$ is some *basis function* and $(\theta_i)_{1\leq i\leq M} \in \Theta^M$ is some family of parameters whose size M, the *width* of the model, is usually large. Depending on the choice of Θ and ψ, Eq. (II.1) can model various types of neural network architectures ranging from simple *single-hidden-layer perceptrons* (Eq. (I.2)) to more complex *convolutional layers* (Eq. (I.3)), used in the original ResNet architecture for image classification [He, 2016a], or even *multi-head attention layers* (Eq. (I.4)), used in Transformer architectures [Vaswani, 2017]. Several authors have then proposed to model architectures with an arbitrary (finite or infinite) number of parameters by using parameterization in the space of measures. The obtained *mean-field* models, described in Eq. (39), are mappings $F_\mu : \mathbb{R}^d \to \mathbb{R}^d$ defined by:

$$F_\mu : x \in \mathbb{R}^d \mapsto \int_\Theta \psi(\theta, x)\mathrm{d}\mu(\theta)\,, \tag{II.2}$$

where $\mu \in \mathcal{P}(\Theta)$ is a distribution of parameters. This setting first provides a convenient framework for studying the training of overparameterized neural network architectures with dedicated mathematical tools such as *Wasserstein gradient flows*. Moreover, this setting also allows for favorable training properties such as the absence of spurious critical points in the loss-landscape, permitting the development of a theory of convergence for the training of shallow architectures at large depth [Chizat, 2018; Mei, 2018; Javanmard, 2020; Wojtowytsch, 2020].

Following on Chapter I, we focus here on studying the training dynamics of deep neural networks and more precisely of deep Residual Neural Networks (ResNets) [He, 2016a], which we presented in Section 1.4.1. The defining characteristic of ResNets is the use of *skip connections*, a mechanism enabling the efficient training of extremely deep models, marking the beginning of the modern era of machine learning. *Neural Ordinary Differential Equations (NODEs)*, proposed by Chen et al. [Chen, 2018] and which we described in Section 1.4.2, correspond to the limit of ResNets whose number of layers tends to infinity and treat deep networks as ODE solvers with trainable parametric vector fields. We consider here models of ResNets whose both depth and width are very large. Such mean-field models of NODEs are ODEs whose velocity field (called *residual*) are parameterized by a distribution of parameters. Let us recall Definition I.1:

Definition I.1 (Mean-field NODE)**.** *For a family of probability measures $\mu = \{\mu(.|s)\}_{s\in[0,1]} \in \mathcal{P}(\Theta)^{[0,1]}$ and input $x \in \mathbb{R}^d$, we define the NODE model output as* $\mathrm{NODE}_\mu(x) \coloneqq x_\mu(1)$ *where $(x_\mu(s))_{s\in[0,1]}$ satisfies the* Forward ODE*:*

$$\frac{\mathrm{d}}{\mathrm{d}s}x_\mu(s) = F_{\mu(.|s)}(x_\mu(s))\,, \quad x_\mu(0) = x\,. \tag{I.5}$$

When there is no ambiguity, we simply write $x(s)$.

We proposed in Chapter I to parameterize such models by measures in $\mathcal{P}_2^{\mathrm{Leb}}([0,1]\times\Omega)$, the set of probability measures over $[0,1]\times\Theta$ with uniform marginal on $[0,1]$. We then showed in Proposition I.1.1 that the above definition is well-posed provided mild regularity assumptions on the basis function ψ are satisfied (cf. Assumption I.1).

Supervised learning As in Chapter I we consider a supervised learning framework where we are given a *training data distribution* $\mathcal{D}$ consisting of pairs of input data $x\in\mathbb{R}^d$ and *target* or *labels* $y\in\mathbb{R}^{d'}$. Then for a parameterization $\mu\in\mathcal{P}_2^{\mathrm{Leb}}([0,1]\times\Theta)$, the training risk is defined as:

$$\mathcal{R}(\mu) \coloneqq \mathbb{E}_{(x,y)\sim y}\ell(\mathrm{NODE}_\mu(x),y) = \mathbb{E}_{(x,y)\sim y}\ell(x_\mu(1),y)\,, \tag{II.3}$$

where $\ell:\mathbb{R}^d\times\mathbb{R}^{d'}\to\mathbb{R}_{\geq 0}$ is some *loss function*. Of particular interest in this section will be the case of a finite number of data, that is of an empirical data distribution $\mathcal{D}=\frac{1}{N}\sum_{i=1}^N\delta_{(x^i,y^i)}$, where $N\geq 1$ is the number of samples. In this case, $\mathcal{R}$ is the *empirical risk* given by:

$$\mathcal{R}(\mu)=\frac{1}{N}\sum_{i=1}^N\ell(x_\mu^i(1),y^i)\,.$$

We showed in Chapter I, that the training of deep ResNets for the minimization of this risk is modeled by a gradient flow w.r.t. to a *Conditional Optimal Transport* metric structure on the space of parameterizations $\mathcal{P}_2^{\mathrm{Leb}}([0,1]\times\Theta)$. Concretely, this gradient flow takes the form Eq. (I.21), a nonlinear advection PDE solved by the parameter distribution. We in particular showed in Section I.3.4 that such a PDE is well-posed. We ask here the question of the *convergence* of this dynamic:

Given an initial parameterization μ_0*, does the gradient flow* $(\mu_t)_{t\geq 0}$ *converge to an optimal parameterization* $\mu^*\in\arg\min\mathcal{R}$*?*

II.1.1 Related works and contributions

Recently, several works have addressed the problem of proving convergence of gradient descent algorithms in the training of neural networks. If convergence properties of gradient descent are well understood for models that are linear w.r.t. their input [Hardt, 2016a; Bartlett, 2018; Zou, 2019; Achour, 2024], it is not the case for non-linear neural network architectures.

Finitely deep architectures In [Li, 2017; Li, 2018; Du, 2018], the authors focus on the training of "shallow" two layer fully connected neural networks and establish convergence of GD in an overparameterized setting where width of the intermediary layer scales polynomially with the size N of the dataset. The works of [Du, 2019; Allen-Zhu, 2019; Zou, 2019; Lee, 2019; Zou, 2020; Liu, 2020; Chen, 2020; Nguyen, 2021] extend those results to arbitrary deep neural networks in the overparameterized setting. Specifically, the results in [Du, 2019; Allen-Zhu, 2019; Liu, 2020] apply to deep ResNets. A common feature for the above cited works is to rely on the fact that, for a sufficiently high number of parameters, the model can be well approximated by a linear model corresponding to its first order expansion around the initialization. In [Chizat, 2019] this phenomenon, called "lazy regime", is attributed to an inappropriate scaling of the parameters. On the other hand, [Liu, 2020] refer to this phenomenon as "linear" or "kernel regime" and relate it to the constancy of the *Neural Tangent Kernel (NTK)* introduced in [Jacot, 2018].

Infinitely deep ResNets Allen-Zhu, Li, and Song [Allen-Zhu, 2019], Du et al. [Du, 2019], and Liu, Zhu, and Belkin [Liu, 2020] give convergence results for the training of deep neural networks with gradient descent and their results can be applied to ResNets. However, in those works the width of intermediary layers has to depend on the depth of the network. Therefore, these results do not apply to the training of the model in Definition I.1, corresponding to the limit $D \to +\infty$. Marion et al. [Marion, 2023b] give local convergence results for the training of infinitely deep ResNets / NODEs based on a local Polyak-Łojasiewicz condition. They assume a model of finite width and their result therefore does not hold in the mean-field limit where residuals are of the form Eq. (II.2). They also consider parameter initializations that are Lipschitz w.r.t. depth which is not consistent with applications where the parameters are initialized at random, independently at each layer. As a comparison Eq. (II.2) models residuals of both finite and infinite width and we only assume the family $\{\mu(.|s)\}_{s\in[0,1]}$ is measurable w.r.t. $s \in [0,1]$.

Mean-field models of NODEs The result presented here should be compared with several other works [Lu, 2020; Ding, 2021; Ding, 2022; Isobe, 2023] that have also studied the convergence of gradient descent for the training of infinitely deep and arbitrarily wide ResNet models similar to Definition I.1. Ding et al. [Ding, 2021; Ding, 2022] — and also Lu et al. [Lu, 2020], but with a different training dynamic — give a result of optimality at convergence: if the parameter distribution converges then its limit is a global minimizer of the risk. They do not however provide proofs of convergence and this convergence assumption seems hard to justify a priori as the loss-landscape of ResNets can have non-compact subsets. In fact, cases where the gradient flow fails to converge have been identified for simple architectures [Bartlett, 2018]. In contrast, our results ensure the convergence of the parameter distribution provided the risk at initialization is sufficiently low.

Borrowing tools from the study of asymptotic behavior of evolution PDEs, Isobe [Isobe, 2023] studies the asymptotic behavior of gradient flow curves associated to the training of ResNets. Precisely, he shows the risk $\mathcal{R}$ satisfy functional inequalities similar to the Polyak-Łojasiewicz inequality in the neighborhood of critical points and shows convergence of the gradient flow to a critical point. Aside from technical differences, our work differs in at least two fundamental aspects. First, [Isobe, 2023] considers adding a regularization term to the risk. Such a regularization ensures gradient flow curves stay in strongly compact sets [Isobe, 2023, Prop.5.4] and admit convergent sub-sequences. Also, the obtained functional inequality does not rule out the presence of non-optimal critical points and the obtained limit is thus not necessarily a minimizer of the risk. In contrast, we consider an unregularized risk whose level sets may be non-compact and show convergence of the gradient flow towards a global minimum for well-chosen initializations.

Contributions We study in this chapter the asymptotic behavior of gradient flow curves for the minimization of the risk $\mathcal{R}$ associated to the training of deep ResNets or NODEs. Specifically, we show that, for standard examples of residual architectures, the risk $\mathcal{R}$ satisfies a *Polyak-Łojasiewicz (P-Ł) property* around well-chosen initializations. The risk has thus no saddles in these regions and decreases at a constant rate along the gradient flow. We study the case of residuals that are *random feature models* [Rahimi, 2007] in Section II.4 and the case of residuals that are 2-layer perceptrons in Section II.5. Based on previous works on the convergence of curves of maximal slope under the P-Ł assumption [Hauer, 2019; Dello Schiavo, 2024], we can then formulate a convergence result: for initializations with a sufficiently large but finite number of features and sufficiently low risk the gradient flow converges towards a global minimizer (Theorems II.6 and II.7). The dependence of

these convergence conditions w.r.t. the data distribution can be numerically quantified and checked for common architectures and initializations. Our results are to be compared with the ones of Lu et al. [Lu, 2020] and Ding et al. [Ding, 2022]. Both works give a result of optimality under a convergence assumption but do not give conditions guaranteeing convergence of the gradient flow. Moreover, their results hold under the assumption of an infinite number of features whereas our convergence conditions can be obtained with a finite number of features.

Finally, we implemented and trained ResNets for solving image classification problems on the MNIST [LeCun, 2010] and CIFAR10 [Krizhevsky, 2009] datasets. In addition to support our theoretical results, the numerical results presented in Section II.7 also show that reduction of the training risk go along with an increase of the classification accuracy on test data.

II.2 Polyak-Łojasiewicz property and convergence of gradient flow

Our approach to show convergence of the gradient flow is to show that the risk satisfies a local Polyak-Łojasiewicz (P-Ł) property around well-chosen parameterizations. The P-Ł inequality provides a lower bound on the ratio between the square gradient's risk $|\nabla\mathcal{R}|^2$ and the risk $\mathcal{R}$. It thus prevents the existence of spurious critical points and guarantees that the risk decreases at a constant rate along gradient flow.

II.2.1 The Polyak-Łojasiewicz property in Hilbert spaces

We consider here the problem of minimizing a function $f : \mathcal{H} \to \mathbb{R}$ defined on some separable (possibly infinite dimensional) Hilbert space $\mathcal{H}$. In the context of training machine learning models, the function f corresponds to some training objective such as the training risk and the task of minimizing f is solved by a *gradient descent algorithm*. For an initialization $z_0 \in \mathcal{H}$, the gradient descent with stepsize $\tau > 0$ is the iterative scheme:

$$\forall k \geq 0\,, \quad z_{k+1} = z_k - \tau\nabla f(z_k)\,.$$

For theoretical purposes, it is also convenient to consider the *gradient flow dynamic*, corresponding to the limit of gradient descent when the stepsize τ tends to 0. For a function $f : \mathcal{H} \to \mathbb{R}$ and an initialization $z_0 \in \mathcal{H}$ the gradient flow for f starting from z_0 is defined as the solution $(z_t)_{t\geq 0}$ to the Cauchy problem:

$$\forall t \geq 0\,, \quad \frac{\mathrm{d}}{\mathrm{d}t} z_t = -\nabla f(z_t)\,. \tag{II.4}$$

Theory for the existence of solutions to such gradient systems under mild regularity assumptions on f was originally developed in [Brezis, 1973]. We are here concerned with analyzing the *convergence* of such optimization methods. A first question is the one of the effective minimization of f, that is whether $f(z_k)$ (resp. $f(z_t)$) tends to $f^* := \inf f$ when $k \to +\infty$ (resp. $t \to +\infty$). A second question is the one of the ability for these methods to find a minimizer, that is whether z_k (resp. z_t) tends to some $z^* \in \arg\min f$ when $k \to +\infty$ (resp. $t \to +\infty$).

Along gradient flow curves, the decrease of f is given by $\frac{\mathrm{d}}{\mathrm{d}t} f(z_t) = -\|\nabla f(z_t)\|^2$. Thus, to obtain a constant decay rate for f, it is natural to ask that the square norm of the gradient is lower-bounded by f itself, namely a condition of the form:

$$\forall z \in \mathcal{H}\,, \quad \|\nabla f(z)\|^2 \geq m\,(f(z) - f^*)\,, \tag{II.5}$$

for some constant $m > 0$. Such *Polyak-Łojasiewicz inequality* was originally used by Polyak [Polyak, 1963] to establish the convergence of the gradient descent algorithm. It also bears the name of Łojasiewicz who showed at the same time that (a generalization of) this inequality is a generic property of analytic functions near their critical points [Łojasiewicz, 1963], a property later generalized by Kurdyka [Kurdyka, 1998]. A generalization to infinite dimensional spaces was also introduced by Simon [Simon, 1983], with application to the study of the asymptotic behavior of evolution PDEs [Chill, 2003]. More recently, the Polyak-Łojasiewicz inequality has proven convenient in non-convex optimization for studying the convergence of various other first order optimization methods [Karimi, 2016], including applications to the training of neural networks [Oymak, 2019; Chatterjee, 2022]. An advantage of Eq. (II.5) is that it is a local condition which can be checked pointwise, only requiring the knowledge of f and of its gradient. This stands in contrast with other assumptions used to obtain convergence of first order methods which generally requires the convexity of f and/or the existence of a minimizer z^*. Moreover, unlike convexity, this condition is robust to small perturbations or reparameterizations of the space $\mathcal{H}$.

Still, Eq. (II.5) is usually too strong to be satisfied in many applications. For example, when studying the training of neural networks, it is known that the loss-landscape has saddle points and Eq. (II.5) can hence not be satisfied on the whole parameter space. For this reason, we consider here a local variant of the Polyak-Łojasiewicz inequality which was proposed in [Oymak, 2019; Chatterjee, 2022].

Definition II.1 (Local P-Ł inequality)**.** *Let $f : \mathcal{H} \to \mathbb{R}_{\geq 0}$ be some non-negative continuously differentiable function and consider $z_0 \in \mathcal{H}$. For constants $R, m > 0$, we say f satisfies the (R, m)-Polyak-Łojasiewicz inequality around z_0 if for every $z \in B(z_0, R)$ it holds:*

$$\|\nabla f(z)\|^2 \geq m f(z)\,. \tag{II.6}$$

Remark II.2.1. *Different formulations of the (R, m)-P-Ł property have been proposed in the literature. For example [Chatterjee, 2022] introduced the local ratio:*

$$\alpha(z_0, R) \coloneqq \inf_{\substack{z \in B(z_0, R) \\ f(z) > 0}} \frac{\|\nabla f(z)\|^2}{f(z)}\,.$$

A direct consequence of the (R, m)-P-Ł property in Definition II.1 is that f admits no spurious critical points (saddles, local maxima or local minima) around z_0 and that all critical points are global minimizers. On its own, such a local property is however insufficient to conclude to convergence of gradient flow to a global minimizer $z^* \in \arg\min f$. Indeed, Eq. (II.6) only controls the decrease rate of f inside a ball and if the gradient flow dynamic escape this ball, it might get stuck at a spurious critical point. Nonetheless, using a confinement argument it is possible to conclude to a local convergence result: if f is sufficiently small at initialization then the gradient flow of f converges with a linear convergence rate.

Theorem II.1 (Convergence of gradient flow)**.** *Let $f : \mathcal{H} \to \mathbb{R}_{\geq 0}$ be some non-negative continuously differentiable function with locally Lipschitz gradient and consider $z_0 \in \mathcal{H}$. Assume that f satisfies a (R, m)-P-Ł inequality around z_0 for some $R, m > 0$ and that $f(z_0)$ satisfies:*

$$f(z_0) < \frac{R^2 m}{4}\,.$$

Let $(z_t)_{t\geq[0,T)}$ be the gradient flow curve starting from z_0, defined until some time $T > 0$. Then the following statements holds:

(i) confinement: $T = +\infty$ and $z_t \in B(z_0, R)$ for every $t \in [0, +\infty)$,

(ii) convergence: $z_\infty \coloneqq \lim_{+\infty} z_t$ exists, is in $B(z_0, R)$ and s.t. $f(z_\infty) = 0$,

(iii) convergence rate: for every $t \in [0, +\infty)$ it holds:

$$f(z_t) \leq f(z_0)e^{-mt} \quad \text{and} \quad \|z_t - z_\infty\| \leq Re^{-mt/2}\,.$$

Proof. The above result is the content of [Chatterjee, 2022, Thm. 1.1] which generalizes to infinite dimensional Hilbert spaces. However, we give here a proof for the sake of completeness.

First, by application of the Cauchy-Lipschitz theorem, there exists a maximal time $T > 0$ s.t. the solution z_t to the gradient flow equation is uniquely defined for $t \in [0, T)$. Note that, if there exists $t_0 \in [0, T)$ s.t. $f(z_{t_0}) = 0$, then a stationary point of the gradient flow has been reached in finite time and all the above claim follow. Thus we can assume w.l.o.g. that $f(z_t) > 0$ for every $t \in [0, T)$. Define for $t \in [0, T)$:

$$\mathcal{E}(t) \coloneqq \sqrt{\frac{4f(z_t)}{m}} + \int_0^t \|\nabla f(z_{t'})\|\,\mathrm{d}t'\,.$$

and consider $T_R \coloneqq \inf\{t \in [0, T) \,:\, \|z_t - z_0\| \geq R\}$ and $T^* \coloneqq T \wedge T_R$. Then the map $t \mapsto \mathcal{E}(t)$ is locally absolutely continuous on $[0, T^*)$ and for a.e. $t \in [0, T^*)$ we have:

$$\frac{\mathrm{d}}{\mathrm{d}t}\mathcal{E}(t) = -\frac{\|\nabla f(z_t)\|^2}{\sqrt{mf(z_t)}} + \|\nabla f(z_t)\| \leq 0\,.$$

Thus $\mathcal{E}$ is decreasing with t and for every $t \in [0, T^*)$ it holds:

$$\int_0^t \|\nabla f(z_{t'})\mathrm{d}t'\| \leq \sqrt{\frac{4f(z_0)}{m}} < R\,.$$

This shows that the curve $(z_t)_{t\in[0,T^*)}$ has finite length, hence that $\lim_{t\to T^*} z_t =: z^*$ exists and that it is in $B(z_0, R)$. Moreover, this also shows that

$$\|z_0 - z_t\| \leq \sqrt{\frac{4f(z_0)}{m}} < R$$

for every $t \in [0, T^*)$ and as a consequence $T_R > T^*$, i.e. $T^* = T$. But then, since the curve (z_t) admits a limit when $t \to T$, this means $T = +\infty$ since otherwise the gradient flow could be extended to a strictly larger time interval, leading to a contradiction with the definition of T. Thus we have shown that $T = +\infty$, that $z_t \in B(z_0, R)$ for every $t \in [0, +\infty)$ and that $\lim_{+\infty} z_t =: z_\infty$ exists and is equal to $z^* \in B(z_0, R)$.

For the convergence rates, observe that, following from the fact that $z_t \in B(z_0, R)$, we have for a.e. $t \geq 0$:

$$\frac{\mathrm{d}}{\mathrm{d}t}f(z_t) = -\|\nabla f(z_t)\|^2 \leq -mf(z_t)\,,$$

leading to $f(z_t) \leq f(z_0)e^{-mt}$ for every $t \geq 0$. Also, for $t \geq 0$ it holds:

$$\begin{aligned}\|z_t - z_\infty\| &\leq \int_t^\infty \|\nabla f(z_{t'})\| \mathrm{d}t' \\ &= \sqrt{\frac{4f(z_t)}{m}} + \left(\lim_{t' \to +\infty} \mathcal{E}(t') - \mathcal{E}(t)\right) \\ &\leq \sqrt{\frac{4f(z_t)}{m}} \\ &\leq Re^{-mt/2}\,,\end{aligned}$$

which finishes the proof. □

It is important to stress that, in the above theorem, there is a result of convergence of gradient flow curves as well as a result of existence of minimizers. Indeed, we only assume that $f \geq 0$ and the existence of a global minimizer $z^* \in Z$ s.t. $f(z^*) = 0$ is part of the conclusion. Finally, the same result hold for gradient descent.

Theorem II.2 (Convergence of gradient descent)**.** *Let $f : \mathcal{H} \to \mathbb{R}_{\geq 0}$ be some non-negative continuously differentiable function with locally Lipschitz gradient and consider $z_0 \in \mathcal{H}$. Assume that f satisfies a (R, m)-P-Ł inequality around z_0 for some $R, m > 0$ and that $f(z_0)$ satisfies:*

$$f(z_0) < \frac{R^2 m}{4}\,.$$

Then, for any $\alpha \in \left(\sqrt{\frac{4f(z_0)}{mR^2}}, 1\right)$, there exists $\tau > 0$ sufficiently small such that the iterates $(z_k)_{k \geq 0}$ of gradient descent with step-size τ satisfy:

(i) *confinement:* *$z_k \in B(z_0, R)$ for every $k \geq 0$,*

(ii) *convergence:* *$z_\infty \coloneqq \lim_{+\infty} z_k$ exists, is in $\bar{B}(z_0, R)$ and s.t. $f(z_\infty) = 0$,*

(iii) *convergence rate:* *for every $k \geq 0$ it holds:*

$$f(z_k) \leq (1 - \alpha m\tau)^k f(z_0) \quad \text{and} \quad \|z_t - z_\infty\| \leq (1 - \alpha m\tau)^{k/2} \|z_0 - z_\infty\|\,.$$

Proof. Proof of this result can be found in [Chatterjee, 2022, Thm. 1.2]. □

II.2.2 The Polyak-Łojasiewicz property in metric spaces

We now consider the case where we want to minimize a function $f : Z \to \mathbb{R}$ defined on a complete metric space (Z, d). In this setting, there is no notion of gradient which could be used as a direction of descent in an iterative algorithm. Instead, for an initialization $z_0 \in Z$ and a step-size $\tau > 0$, one can consider the *proximal descent scheme* which produces iteratively:

$$\forall k \geq 0\,, \quad z_{k+1} \in \operatorname*{arg\,min}_{z \in Z} f(z) + \frac{1}{2\tau} d(z, z_k)^2\,.$$

In turn, this proximal sequence defines a limiting dynamic when the stepsize τ tends to 0, thereby generalizing the notion of gradient flow to the setting of metric spaces. In Hilbert spaces, gradient flow curves can be characterized as solutions to variational inequalities

involving the gradient norm. For example, the curve $(z_t)_{t\in\mathbb{R}}$ is the solution to the gradient flow equation Eq. (II.4) if and only if it is an absolutely continuous curve satisfying the *Energy Dissipation Inequality (EDI)*:

$$\forall t \in \mathbb{R}\,, \quad \frac{\mathrm{d}}{\mathrm{d}t} f(z_t) \leq -\frac{1}{2}\left(\left\|\frac{\mathrm{d}}{\mathrm{d}t} z_t\right\|^2 + \|\nabla f(z_t)\|^2\right)\,. \tag{II.7}$$

The above characterization then generalizes to the setting of metric spaces by replacing the objects with their metric counterparts. The norm of the velocity $\left\|\frac{\mathrm{d}}{\mathrm{d}t} z_t\right\|$ is replaced by the metric derivative $\left|\frac{\mathrm{d}}{\mathrm{d}t} z_t\right|$ ([Ambrosio, 2008b, Def. 1.1.2]) and the norm of the gradient $\|\nabla f(z_t)\|$ is replaced by the notion of *upper gradient*. Recalling Definition I.4, a function $g : Z \to [0, +\infty]$ is an upper gradient for f if for every absolutely continuous curve $(z_t)_{t\in I}$ on an interval $I \subset \mathbb{R}$ it holds:

$$|f(z_{t_1}) - f(z_{t_2})| \leq \int_{t_1}^{t_2} g(z_t) \left|\frac{\mathrm{d}}{\mathrm{d}t} z_t\right| \mathrm{d}t\,, \quad \forall t_1 \leq t_2 \in I\,.$$

In Section I.3, we for example showed that $\|\nabla \mathcal{R}\|_{L^2(\mu)}$ is an upper gradient for the training risk $\mathcal{R}$ defined on the space of parameter distributions $\mathcal{P}_2^{\mathrm{Leb}}([0,1] \times \Theta)$ equipped with the metric $\mathcal{W}_2^{\mathrm{COT}}$ (Proposition I.3.4). *Curves of maximal slopes* of f can then be defined as absolutely continuous curves in Z for which Eq. (II.7) holds. We recall here Definition I.5.

Definition I.5 (Curve of maximal slope [Ambrosio, 2008b, Def.1.3.2])**.** *Let (Z, d) be a complete metric space, $I \in \mathbb{R}$ be an interval and $f : Z \to \mathbb{R}$ a function with $|\nabla f|$ an upper gradient for f. We say that $(z_t)_{t\in I}$ is a* curve a maximal slope *for f (w.r.t. $|\nabla f|$) if it satisfies:*

(i) $(z_t)_{t\in I}$ is locally absolutely continuous,

(ii) the map $t \mapsto f(z_t)$ is non-increasing,

(iii) for $\mathrm{d}t$-a.e. $t \in I$ it holds $\frac{\mathrm{d}}{\mathrm{d}t} f(z_t) \leq -\frac{1}{2}\left(\left|\frac{\mathrm{d}}{\mathrm{d}t} z_t\right|^2 + |\nabla f|^2(z_t)\right)$.

If $\lim_{t\to \inf I} z_t = z$ exists then we say $(z_t)_{t\in I}$ is a curve of maximal slope starting at z.

There is an important amount of literature devoted to the study of gradient flow dynamics in metric spaces [Ambrosio, 2008b; Ambrosio, 2013; Santambrogio, 2017]. Of particular interest is the case of the space $\mathcal{P}(X)$ of probability measures over some metric space X, equipped with the Wasserstein distance $\mathcal{W}_p$ for some $p \geq 1$. In this case, the seminal work of Jordan, Kinderlehrer, and Otto [Jordan, 1998] has shown that some evolution PDEs such as Fokker-Planck equations can be interpreted as gradient flows of functionals defined on the space of probability measures w.r.t. the Wasserstein metric. More recently, this formalism has attracted growing interest for studying the training dynamics of neural networks, modeled by Wasserstein gradient flows on the distribution of their parameters [Chizat, 2018; Mei, 2018]. Similarly, we showed in Section I.3 that the training of our mean-field NODE model can be modeled with a gradient flow w.r.t. the conditional OT metric $\mathcal{W}_2^{\mathrm{COT}}$, corresponding to solutions of some advection PDE on the space of parameters (Definition I.3).

We are here interested in analyzing the *convergence* of curves of maximal slopes for a function $f : Z \to \mathbb{R}$. The strategy is the same as in the case of Hilbert spaces, observing

that the decrease of f along such curves is formally given by $\frac{\mathrm{d}}{\mathrm{d}t} f(z_t) = -|\nabla f|^2(z_t)$. Thus, the natural generalization of Eq. (II.5) is:

$$\forall z \in Z\,, \quad |\nabla f|^2(z) \geq m(f(z) - f^*)\,, \tag{II.8}$$

for some constant $m > 0$ and where $f^* \coloneqq \inf f$. The above Polyak-Łojasiewicz inequality in a metric space setting has encountered applications in the analysis of non-convex and non-smooth optimization problems [Bolte, 2010]. Following from the interpretation of PDEs as gradient flows, Eq. (II.8) takes the form of functional inequalities used in the classical entropy method to establish quantitative contraction properties of solutions [Blanchet, 2018; Hauer, 2019]. A classical example is the heat equation, where the logarithmic-Sobolev inequality can be interpreted as a (metric) P-Ł inequality for the Boltzmann entropy. As before, the general form in Eq. (II.8) will however be too strong to be satisfied and we will consider instead a local variant of Eq. (II.8) which was proposed in [Dello Schiavo, 2024].

Definition II.2 (Local P-Ł property in metric spaces)**.** *Let $f : Z \to \mathbb{R}_{\geq 0}$ be a non-negative function with upper gradient $|\nabla f|$ and consider $z_0 \in Z$. For constants $R, m > 0$, we say that f satisfies a (R, m)-Polyak-Łojasiewicz inequality (w.r.t. $|\nabla f|$) around a z_0 if for every $z \in B(z_0, R)$ it holds:*

$$|\nabla f|^2(z) \geq m f(z)\,. \tag{II.9}$$

As for the case of Hilbert spaces, while such a local P-Ł property is insufficient to conclude to unconditional convergence of curves of maximal slope, it allows obtaining convergence when f is already sufficiently small at initialization. The following result can be found in [Dello Schiavo, 2024].

Theorem II.3 ([Dello Schiavo, 2024, Cor. 1.5])**.** *Let $f : Z \to R_{\geq 0}$ be lower semicontinuous and non-negative, let $|\nabla f|$ be an upper-gradient for f and consider $z_0 \in Z$. Assume that f satisfies a (R, m)-P-Ł inequality around z_0 and that $f(z_0)$ satisfies:*

$$f(z_0) < \frac{mR^2}{4}\,. \tag{II.10}$$

For $T > 0$, let $(z_t)_{t \in [0,T)}$ be a curve of maximal slope for f (w.r.t. upper gradient $|\nabla f|$) starting from z_0. Then the following statements hold:

(i) *confinement:* *$z_t \in B(z_0, R)$ for every $t \in [0, T)$,*

(ii) *convergence:* *$z_T \coloneqq \lim_{t \to T} z_t$ exists and is in $B(z_0, R)$,*

(iii) *convergence rate:* *for every $t \in [0, T]$ it holds:*

$$f(z_t) \leq f(z_0) e^{-mt} \quad \text{and} \quad d(z_t, z_T) \leq R e^{-mt/2}\,,$$

with the convention that $e^{-\infty} = 0$.

Finally, we conclude this section by noticing that above convergence result is open in the sense that if its assumptions are satisfied for some initialization z_0 then it is also the case for any initialization z_0' sufficiently close to z_0.

Proposition II.2.1. *Let $f : Z \to \mathbb{R}_{\geq 0}$ be continuous and non-negative and let the assumptions of Theorem II.3 be satisfied at some $z_0 \in Z$. Then there exists a neighborhood $\mathcal{U}$ of z_0 such that the assumptions of Theorem II.3 are also satisfied at any initialization $z_0' \in \mathcal{U}$.*

Proof. By definition of the (R, m)-P-Ł property, if it is satisfied around z_0 then, for any $\delta \in (0, R)$ a $(R-\delta, m)$-P-Ł property is satisfied around z_0' for any $z_0' \in B(z_0, \delta)$. Moreover it follows from the continuity of f that the condition Eq. (II.10) is open: if it is satisfied at z_0 with R and m then it is satisfied at any $z_0' \in B(z_0, \delta)$ with $R - \delta$ and m provided δ is sufficiently small. □

II.3 Convergence for general architectures

We explain here how one can prove convergence of the gradient flow towards a minimizer of the risk $\mathcal{R}$ when training the NODE model defined in Definition I.1. Precisely, we show that the risk satisfies a Polyak-Łojasiewicz inequality of the form Eq. (II.8) where the P-Ł constant depends both on the dataset and on the parameterization. We then explain how this constant is related to functional properties of the space of residuals and give conditions to ensure its positivity. Those conditions are described here at a general level but will be specified in Sections II.4 and II.5 for practical examples of architectures and initializations.

Finite number of data As in Chapter I, we consider training the NODE model for the minimization of the training risk associated to a distribution of labeled training data $\mathbb{R}^d \times \mathbb{R}^{d'} \ni (x, y) \sim \mathcal{D}$. Specifically, to obtain convergence results, we focus on the case where the data distribution is the empirical distribution $\mathcal{D} = \frac{1}{N}\sum_{i=1}^N \delta_{(x^i, y^i)}$ for a dataset $\{(x^i, y^i)\}_{1 \leq i \leq N} \in (\mathbb{R}^d \times \mathbb{R}^d)^N$. In this case, the risk $\mathcal{R}$ for a parameterization μ is given by the *empirical risk*:

$$\mathcal{R}(\mu) = \frac{1}{2N}\sum_{i=1}^N \ell(x_\mu^i(1), y^i)\,, \tag{II.11}$$

where x_μ^i is the flow of Eq. (I.5) starting at x^i and with parameterization μ, which we will simply denote by x^i when no ambiguity. Similarly we denote by $p_\mu^i \coloneqq p_{\mu, x^i, y^i}$, or simply p^i when no ambiguity, the associated adjoint variables solution to Eq. (I.13). We also suppose that ψ satisfies Assumptions I.1 to I.3.

Assumption on the loss To show the convergence of gradient methods for the training of our NODE architecture, we will show the empirical risk $\mathcal{R}$ satisfies a local Polyak-Łojasiewicz property. In this purpose, a minimal working assumption is that the loss function ℓ itself satisfies the P-Ł property. This assumption is in particular satisfied in regression problems by the quadratic loss $\ell(x, y) = \frac{1}{2}\|x-y\|^2$ or locally by the cross entropy loss $\ell(x, y) = -\log\left(\frac{\sum_j y[j]\exp(x[j])}{\sum_j \exp(x[j])}\right)$ in classification. For the sake of simplicity, we assume the P-Ł constant is 2 but all the results of course still hold with other constants.

Assumption II.1. *The loss function $\ell : \mathbb{R}^d \times \mathbb{R}^{d'}$ is smooth and satisfies the P-Ł property w.r.t. $x \in \mathbb{R}^d$, uniformly w.r.t $(x, y) \in \mathbb{R}^d \times \mathbb{R}^{d'}$, that is:*

$$\forall (x, y) \in \mathbb{R}^d \times \mathbb{R}^{d'}\,, \quad \|\nabla_x \ell(x, y)\|^2 \geq 2\ell(x, y)\,.$$

II.3.1 Conditioning of the tangent kernel implies the P-Ł property

We showed in Proposition I.3.4 that, for the mean-field NODE model defined in Definition I.1, an upper gradient of the training risk $\mathcal{R}$ is given by the norm the the gradient

field $\nabla\mathcal{R}[\mu]$ obtained in Eq. (I.20). In the setting of a finite number of data samples, we thus have for any $\mu \in \mathcal{P}_2^{\mathrm{Leb}}([0,1]\times\Theta)$:

$$|\nabla\mathcal{R}|^2(\mu) = \int_0^1 \int_\Theta \left\| \frac{1}{N}\sum_{i=1}^N \mathrm{D}_\theta\psi(\theta, x^i(s))^\top p^i(s) \right\|^2 \mathrm{d}\mu(s,\theta)$$
$$= \frac{1}{N^2}\int_0^1 \left(\sum_{1\le i,j\le N} p^i(s)^\top K[\mu(.|s)](x^i(s), x^j(s)) p^j(s) \right) \mathrm{d}s\,,$$

where for a parameterization $\nu \in \mathcal{P}_2(\Theta)$ we define the kernel $K[\nu] : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}^{d\times d}$ as:

$$\forall x, x' \in \mathbb{R}^d\,, \quad K[\nu](x,x') \coloneqq \int_\theta \mathrm{D}_\theta\psi(\theta,x)\mathrm{D}_\theta\psi(\theta,x')^\top \mathrm{d}\nu(\theta)\,. \tag{II.12}$$

For a point cloud $\mathbf{z} = (z^i)_{1\le i\le N} \in (\mathbb{R}^d)^N$ we will denote by $\mathbb{K}[\nu,\mathbf{z}] \in \mathbb{R}^{dN\times dN}$ the *kernel matrix* associated to $K[\nu]$ and defined as the block matrix:

$$\mathbb{K}[\nu,\mathbf{z}] \coloneqq \Big(K[\nu](z^i,z^j)\Big)_{1\le i,j\le N}\,. \tag{II.13}$$

In particular, we see that the conditioning $\lambda_{\min}(\mathbb{K}[\nu,\mathbf{z}])$ of the kernel matrix will play an important role in proving a local P-Ł property for the risk $\mathcal{R}$. Indeed, in terms of the kernel matrix $\mathbb{K}$, the square gradient can then be written :

$$|\nabla\mathcal{R}|^2(\mu) = \frac{1}{N^2}\int_0^1 \langle \mathbf{p}_\mu(s), \mathbb{K}[\mu(.|s), \mathbf{x}_\mu(s)]\mathbf{p}_\mu(s)\rangle\,\mathrm{d}s\,, \tag{II.14}$$

where for every $s\in[0,1]$ we defined the point cloud $\mathbf{x}_\mu(s) \coloneqq (x^i_\mu(s)) \in (\mathbb{R}^d)^N$ and where we concatenated the adjoint variables into $\mathbf{p}_\mu(s) \coloneqq (p^i_\mu(s))_{1\le i\le N} \in \mathbb{R}^{dN}$. In the following, when no ambiguity, we will write $\mathbf{x} = \mathbf{x}_\mu \in \mathcal{C}([0,1],(\mathbb{R}^d)^N)$ and $\mathbf{p} = \mathbf{p}_\mu \in \mathcal{C}([0,1],\mathbb{R}^{dN})$.

Lemma II.3.1. *Assume ψ satisfies Assumptions I.1 to I.3 and ℓ satisfies Assumption II.1. Consider $\mu \in \mathcal{P}_2^{\mathrm{Leb}}([0,1]\times\Theta)$. Then there exists a constant $C = C(\mathcal{E}_2(\mu))$ s.t.:*

$$|\nabla\mathcal{R}|^2(\mu) \ge \frac{2e^{-C}}{N}\left(\int_0^1 \lambda_{\min}(\mathbb{K}[\mu(.|s),\mathbf{x}_\mu(s)])\mathrm{d}s\right)\mathcal{R}(\mu)\,, \tag{II.15}$$

Proof. Thanks to Assumption I.3 and to the definition of p^i_μ, there exists a constant $C = C(\mathcal{E}_2(\mu))$ such that for every $1\le i\le N$ we have the estimate:

$$\|p^i_\mu(s)\|^2 \ge e^{-C}\|p^i_\mu(1)\|^2\,, \quad \forall s\in[0,1]\,.$$

Using that $p^i_\mu(1) = \nabla_x\ell(x^i_\mu(1), y^i)$ and with the previous Assumption II.1 we have

$$\|\mathbf{p}_\mu(s)\|^2 \ge e^{-C}\|\mathbf{p}_\mu(1)\|^2 \ge 2Ne^{-C}\mathcal{R}(\mu)\,.$$

Putting this lower bound in Eq. (II.14) then gives the result. □

The above Lemma II.3.1 shows that the conditioning $\lambda_{\min}(\mathbb{K})$ of the kernel matrix provides a lower bound on the ratio between the square gradient and the risk: assuming $\lambda_{\min}(\mathbb{K}) > 0$ — which will for example always be true in Section II.4.2 — implies the P-Ł inequality Eq. (II.9) for the risk. It in particular implies that the risk has no spurious critical points — every critical point is a global minimizer. This remarkable property arises from the combination of skip connections and the infinite-depth limit, which together

enable NODEs to implement an invertible warping of the input space. This stands in stark contrast to finite-width feedforward architectures, which typically exhibit numerous saddle points [Achour, 2024].

The P-Ł constant in Eq. (II.15) could for example be computed numerically during training. However, at this point, it is not clear how one can be sure, before training, that the P-Ł inequality will hold along the gradient flow. We investigate this problem in the next sections for special kinds of architectures. Nonetheless, a direct corollary of Lemma II.3.1 is that gradient flow converges if it stays bounded and if we assume the kernel matrix stays well-conditioned.

Corollary II.3.1. *Assume ψ satisfies Assumptions I.1 to I.3 and ℓ satisfies Assumption II.1. For an initialization $\mu_0 \in \mathcal{P}_2^{\mathrm{Leb}}([0,1]\times\Theta)$, let $(\mu_t)_{t\geq 0}$ be a gradient flow of $\mathcal{R}$ starting from μ_0. If there exists a constant $C>0$ s.t., for every $t\geq 0$, $\mathcal{E}_2(\mu_t)\leq C$ and $\int_0^1 \lambda_{\min}(\mathbb{K}[\mu_t(.|s),\mathbf{x}_{\mu_t}(s)])\mathrm{d}s \geq C^{-1}$, then the gradient flow converges in the sense that $\mu_t \xrightarrow{t\to+\infty} \mu_\infty \in \mathcal{P}_2^{\mathrm{Leb}}([0,1]\times\Theta)$ and there exists a constant $C'>0$ s.t.:*

$$\mathcal{R}(\mu_t) \leq e^{-C't}\mathcal{R}(\mu_0), \quad \forall t \geq 0\,.$$

II.3.2 Expressivity and functional properties of the set of residuals

The kernel $K[\mu]$ defined in Eq. (II.12) corresponds to the *Neural Tangent Kernel (NTK)* associated to the architecture in Eq. (II.2) [Jacot, 2018]. Properties of the NTK, and especially conditioning of the associated kernel matrix, have been identified by several works has a key ingredient to show convergence of gradient methods for the training of neural networks [Allen-Zhu, 2019; Du, 2019; Lee, 2019; Zou, 2020; Liu, 2020]. In turn, the positivity of the kernel matrix $\mathbb{K}$ here readily plays a role in Corollary II.3.1 to establish convergence of gradient flow for the training of NODEs. We explain here how this conditioning is related to functional properties of the set of residuals and more precisely to their expressivity. Later-on we will give examples of architectures and parameterizations for which sufficient expressivity of the residuals can be ensured to show convergence of gradient methods for the training of deep ResNets.

Positive kernels and RKHS By construction, the kernel $K[\mu]$ defined in Eq. (II.12) is a (vector valued) *positive kernel* over $\mathbb{R}^d$. Indeed, for $\mu\in\mathcal{P}_2(\Theta)$, we have that for every $(x^i)_{1\leq i\leq N}$ and every $(p^i)_{1\leq i\leq N}\in(\mathbb{R}^d)^N$:

$$\sum_{1\leq i,j\leq N}\left\langle p^i, K[\mu](x^i,x^i)p^j\right\rangle = \int_\Theta \left\|\sum_{i=1}^N \mathrm{D}_\theta\psi(\theta,x^i(s))^\top p^i(s)\right\|^2 \mathrm{d}\mu(\theta) \geq 0\,.$$

It is a classical result that every such kernel defines a unique structure of *Reproducing Kernel Hilbert Space (RKHS)* over $\mathbb{R}^d$, a Hilbert space of functions for which the evaluation function is continuous [Steinwart, 2008; Carmeli, 2010]. The kernel $K[\mu]$ is here directly given by a *feature representation*, that is a representation of the form $K[\mu](x,y)=\chi(x)^\top\chi(y)$ with a map $\chi:\mathbb{R}^d\to\mathcal{L}(\mathbb{R}^d,\mathcal{H})$ for some Hilbert space $\mathcal{H}$. If such a representation always defines a positive kernel, one can conversely show that such a representation always exists whenever K is a positive kernel [Carmeli, 2010]. This representation can however not be expected to be unique and corresponds to a certain square root of $K[\mu]$ viewed as an integral operator [Bach, 2017b]. Here, one can for example consider $\mathcal{H}=L^2(\mu)$ and χ is given by:

$$\forall x,p\in\mathbb{R}^d, \quad \chi(x)\cdot p = \mathrm{D}_\theta\psi(.,x)^\top p\,.$$

The associated RKHS is defined by:

$$\mathcal{F}_\mu \coloneqq \left\{ F : x \mapsto \chi(x)^\top \cdot \delta\theta = \int_\Theta \mathrm{D}_\theta \psi(\theta, x) \delta\theta(\theta) \mathrm{d}\mu(\theta) \ : \ \delta\theta \in L^2(\mu) \right\} \tag{II.16}$$

and

$$\forall F \in \mathcal{F}_\mu \,, \quad \|F\|_{\mathcal{F}_\mu} \coloneqq \inf \left\{ \|\delta\theta\|_{L^2(\mu)} \ : \ F(x) = \chi(x)^\top \cdot \delta\theta \,, \ \forall x \in \mathbb{R}^d \right\} . \tag{II.17}$$

The RKHS $\mathcal{F}_\mu$ can be seen as the linearization of the space of residuals at some parameterization $\mu \in \mathcal{P}_2(\Theta)$. Its functional properties depend on the choice of architecture, materialized by the basis function ψ, which is fixed, but also on the choice of the parameterization, which will vary during training.

Universality of residuals A lower bound on the minimum eigenvalue of the kernel matrix $\mathbb{K}$ is assumed in Corollary II.3.1 to ensure convergence of gradient flow. For a kernel K, the property of having its associated kernel matrix being (strictly) positive on every separated point cloud is a property we refer to as *strict positivity*. We make the following definition:

Definition II.3 (Strict positivity)**.** *We say a positive kernel K is strictly positive if for every family $\mathbf{z} = (z^i)_{1 \leq i \leq N} \in (\mathbb{R}^d)^N$ of mutually disjoint points the associated kernel matrix $\mathbb{K}[\mathbf{z}]$ is positive definite.*

The notion of strict positivity is related to the stronger notion of *universality* which is the property for a RKHS to be dense in the space of continuous functions [Micchelli, 2006; Sriperumbudur, 2011] (the two notions are for example equivalent for radial kernels [Sriperumbudur, 2011]). In particular, this condition is satisfied by a large class of common kernels such as Gaussian or Matérn kernels. More generally, being strictly positive in the sense of Definition II.3 requires for the associated RKHS $\mathcal{F}$ to be at least of dimension $M \geq dN$, since it implies that, for any family of N mutually-disjoint points $\mathbf{z} = (z^i)_{1 \leq i \leq N} \in (\mathbb{R}^d)^N$ and any family of vectors $(F^i)_{1 \leq i \leq N} \in (\mathbb{R}^d)^N$ there exists some $F \in \mathcal{F}$ s.t. $F(x^i) = F^i$ for every index $i \in \{1, ..., N\}$. However, when considering a fixed family $\mathbf{z} = (z^i)_{1 \leq i \leq N} \in (\mathbb{R}^d)^N$, the strict positivity assumption can be satisfied for finite dimensional RKHSs of dimension $M \leq N^d$, for example by considering a polynomial kernel, or by RKHSs of dimension $M \geq poly(N)$ with high probability over the sampling of random features.

For a RKHS $\mathcal{F}_\mu$ as in Eq. (II.16), the expressivity of $\mathcal{F}_\mu$ depends on ψ and on the parameterization μ. An example we develop further in Sections II.4 and II.5 is the case of 2-layer perceptrons of Eq. (34) where trained parameters are weight matrices $U, W \in \mathbb{R}^{d \times M}$ and a bias vector $b \in \mathbb{R}^M$ and $\psi((U, W, b), x) = U\sigma(W^\top x + b)$, with σ an activation function applied component-wise. In this case, when considering the ReLU activation or the trigonometric activation cos, the strict positivity of the NTK is ensured provided the width $M \geq 1$ is sufficiently large.

II.4 Linear parameterization of the residuals

Most often in the literature studying the training properties of ResNets, the considered residual transformations are *Multi-Layer Perceptrons (MLP)* [Du, 2019; Allen-Zhu, 2019; Hardt, 2016a]. These consist in the composition of several trained linear layers alternatively composed with a non-linear activation function. In contrast, we first consider here

a simplified architecture where the residual term is linear w.r.t. the parameters while still being nonlinear w.r.t the inputs. While retaining the expressivity properties of MLPs, such a parameterization has the advantage of simplifying the learning dynamic.

Precisely, we consider residuals that are of the form Eq. (II.2) where the parameter space is $\Theta = \mathcal{H}^d$, for some a Hilbert space $\mathcal{H}$, and $\psi : \Theta \times \mathbb{R}^d \to \mathbb{R}^d$ is defined as:

$$\forall \theta = (\theta_i)_{1 \leq i \leq d} \in \Theta\,, \quad \forall x \in \mathbb{R}^d, \quad \psi(\theta, x) = \theta \cdot \phi(x) = \begin{pmatrix} \langle \theta_1, \phi(x) \rangle_{\mathcal{H}} \\ \vdots \\ \langle \theta_d, \phi(x) \rangle_{\mathcal{H}} \end{pmatrix} \in \mathbb{R}^d\,, \tag{II.18}$$

where $\phi : \mathbb{R}^d \to \mathcal{H}$ is some measurable map which we call *feature map*.

Examples of ResNets with linear parameterization Depending on the choice of a feature map ϕ and on the choice of a Hilbert space $\mathcal{H}$, the above defined class of residuals encompasses several interesting examples of architectures.

- Linear residuals: This corresponds to the case where $\mathcal{H} = \mathbb{R}^d$ and $\phi = \mathrm{Id}$. In this case the set of parameters identifies to $\Theta = \mathbb{R}^{d \times d}$, the set of matrices of size $d \times d$, and the residuals simply consist in the left matrix-vector multiplication.
- Perceptrons with fixed hidden layer: Recalling Eq. (34), the single-hidden-layer perceptron model is defined by:

$$F_{(U,W,b)} : x \in \mathbb{R}^d \mapsto U\sigma(W^\top x + b)\,, \tag{II.19}$$

  where $U, W \in \mathbb{R}^{d \times M}$ are trainable weight matrices and $b \in \mathbb{R}^M$ is a trainable bias vector. In comparison, Eq. (II.20) encompasses the case of *random feature models* [Rahimi, 2007] where the inner weight matrix W and the bias vector are fixed. This corresponds to a feature space $\mathcal{H} = \mathbb{R}^M$ and a feature map $\phi : x \mapsto \sigma(W^\top x + b)$. We will show convergence for ResNets with this type of residuals in Section II.4.3, provided the width M is sufficiently large.

RKHS parameterization of residuals Linear parameterization of ψ greatly simplifies the parameterization of the residuals since, leveraging the linearity w.r.t. $\theta \in \Theta$, a parameter distribution is equivalently described by its mean. Recalling Eq. (II.2), the output of a residual F_μ parameterized by $\mu \in \mathcal{P}_2(\Theta)$ on an input $x \in \mathbb{R}^d$ is:

$$F_\mu(x) = \int_\Theta \theta \cdot \phi(x) \mathrm{d}\mu(\theta) = \mathbb{E}_\mu[\theta] \cdot \phi(x)\,.$$

As a consequence, the space of residuals can be described by a single parameter $\theta \in \Theta$:

$$\mathcal{F} \coloneqq \{F : x \mapsto \theta \cdot \phi(x) \;:\; \theta \in \Theta\}\,. \tag{II.20}$$

In particular, this space of residuals is a vector space which is independent of the parameter distribution. Concretely, for every parameter distribution $\mu \in \mathcal{P}_2(\Theta)$, we have $\mathcal{F}_\mu = \mathcal{F}$ where $\mathcal{F}_\mu$ is the linearization of the space of residuals defined in Eq. (II.16). Moreover, this space is in fact isometric to the space of parameters. The following result is a direct application of [Carmeli, 2010, Prop. 1].

Proposition II.4.1. *The vector space $\mathcal{F}$ is a Reproducing Kernel Hilbert Space of vector fields over $\mathbb{R}^d$ whose kernel is given by:*

$$\forall x, x' \in \mathbb{R}^d\,, \quad K(x,x') \coloneqq \langle \phi(x), \phi(x') \rangle_{\mathcal{H}} \operatorname{Id} \in \mathbb{R}^{d\times d}\,. \tag{II.21}$$

Moreover the mapping $\mathrm{T} : \theta \in \Theta \mapsto F = \theta \cdot \phi(.) \in \mathcal{F}$ is a partial isometry and it holds:

$$\forall F \in \mathcal{F}\,, \quad \|F\|_{\mathcal{F}}^2 = \inf \left\{ \|\theta\|_{\Theta}^2 = \sum_{i=1}^d \|\theta_i\|_{\mathcal{H}}^2 \;:\; F(x) = \theta \cdot \phi(x)\,,\ \forall x \in \mathbb{R}^d \right\}\,.$$

Upon restricting Θ to $\operatorname{Ker}(\mathrm{T})^\top$, we will assume in the following that T is injective and hence an isometry, which is equivalent to assume that $\operatorname{Span}(\{\phi(x),\, x \in \mathbb{R}^d\})$ is dense in $\mathcal{H}$. Moreover, we abuse notations and extend the operator T to $L^2([0,1],\Theta)$ by defining for $\theta \in L^2([0,1],\Theta)$:

$$\mathrm{T}(\theta)(s) \coloneqq \mathrm{T}(\theta(s))\,, \quad \text{for a.e. } s \in [0,1].$$

Then $\mathrm{T} : L^2([0,1],\Theta) \to L^2([0,1],\mathcal{F})$ is an isometry and its inverse T^{-1} is defined similarly. We also denote by $\pi : \mathcal{P}_2^{\mathrm{Leb}}([0,1],\Theta) \to L^2([0,1],\mathcal{F})$ the (surjective) mapping associating to a parameter distribution its corresponding residuals and defined for $\mu \in \mathcal{P}_2^{\mathrm{Leb}}([0,1],\Theta)$ by:

$$\text{for a.e. } s \in [0,1], \quad \pi(\mu)(s) \coloneqq \mathrm{T}\left(\mathbb{E}_{\mu(.|s)}[\theta]\right) = \mathbb{E}_{\mu(.|s)}[\theta] \cdot \phi(.) \in \mathcal{F}\,. \tag{II.22}$$

It admits a natural right-inverse which we denote by $\pi^{-1} : L^2([0,1],\mathcal{F}) \to \mathcal{P}_2^{\mathrm{Leb}}([0,1],\Theta)$ and which consists in considering parameter distributions that are single dirac masses at each layer. Namely, if $F \in L^2([0,1],\mathcal{F})$, then $\theta = \mathrm{T}^{-1}(F) \in L^2([0,1],\Theta)$ and we define:

$$\pi^{-1}(F) \coloneqq \int_0^1 \delta_{\theta(s)} \mathrm{d}s\,,$$

i.e. the measure $\mu \in \mathcal{P}_2^{\mathrm{Leb}}([0,1],\Theta)$ whose disintegration is $\{\delta_{\theta(s)}\}_{s\in[0,1]}$.

The RKHS-NODE model In this section, we are interested in understanding the convergence properties of first order methods such as Gradient Descent (GD) on infinitely deep ResNet models for which the residual layers are encoded in a vector-valued RKHS. Instantiating the NODE model in Definition I.1 to the case of residuals in a RKHS give the following definition of RKHS-NODEs:

Definition II.4 (RKHS-NODE)**.** *Let $\mathcal{F}$ be a RKHS of vector-fields over $\mathbb{R}^d$. Then for $F \in L^2([0,1],\mathcal{F})$ and a data input $x \in \mathbb{R}^d$, the RKHS-NODE model is given by:*

$$\mathrm{NODE}_F(x) = x_F(1)$$

where x_F is the solution to the forward problem*:*

$$\forall s \in [0,1]\,, \quad x_F(s) = x + \int_0^s F(r, x_F(r))\mathrm{d}r\,. \tag{II.23}$$

Note that there is a slight abuse of notation as we denote by NODE the model parameterized either by parameter distributions $\mu \in \mathcal{P}_2^{\mathrm{Leb}}([0,1],\Theta)$ (Definition I.1) or by residuals $F \in L^2([0,1],\mathcal{F})$ (Definition II.4). Indeed, for every parameter distribution $\mu \in \mathcal{P}_2^{\mathrm{Leb}}([0,1],\Theta)$ and every input $x \in \mathbb{R}^d$, we have $x_\mu = x_{\pi(\mu)}$ and hence:

$$\mathrm{NODE}_\mu(.) = \mathrm{NODE}_{\pi(\mu)}(.)\,,$$

where $\pi : \mathcal{P}_2^{\mathrm{Leb}}([0,1],\Theta) \to L^2([0,1],\mathcal{F})$ is the surjection defined in Eq. (II.22).

Relevance of the RKHS-NODE model. The main difference between the model of Definition II.4 and standard ResNets is linearity in the parameters of the residual blocks. As a comparison, a 2-layer MLP is nonlinear w.r.t. the parameters of the hidden layers. However, this linearity assumption does not impact the expressivity of the model, but only its training dynamic. **(i)** Indeed, considering $\mathcal{F}$ to be a random feature approximation (c.f. Eq. (II.33)) of some universal RKHS, the residual blocks are as expressive as a 2-layer MLP since both are dense in the space of continuous functions. **(ii)** Up to the cost of adding a supplementary variable, the dynamical system parameterized by a 2-layer MLP can be expressed as a model which is linear w.r.t. its parameters [Vialard, 2020, Section 3.2]. Only the training dynamic between these two architectures differs. Also, this assumption of linearity in the parameters also prevents the use of normalization layers. In this direction, Zhang, Dauphin, and Ma [Zhang, 2018] have shown that ResNets without normalization but proper initialization of the weights can lead to robust training and similar generalization on the test set than standard ResNets. Finally, the model of Definition II.4 still retains the effect of depth and the nonlinearity w.r.t. the input. Due to composition of these residual blocks the model's output is still highly non-linear w.r.t. parameters. For these reasons, we consider this model as an important step towards the study of the general case.

In turn, this linearity in parameters naturally leads to an RKHS parameterization which has two important benefits on the theoretical side: **(i)** Flows of vector-fields as implemented by our model in Eq. (II.23) have already been studied theoretically and for applications in image registration problems [Younes, 2010; Beg, 2005; Niethammer, 2011]. Under some regularity assumptions on the considered RKHS $\mathcal{F}$, one can show that the model's output corresponds to the invertible action of a diffeomorphism by composition on the input [Trouvé, 1998]. This property was already used in [Salman, 2018] to implement models of *Normalizing Flows* [Kobyzev, 2020] with applications in generative modeling. **(ii)** There is an important literature in Machine Learning about Kernel methods [Schölkopf, 2002]. In practice, various sub-sampling methods exist in order to approximate infinite-dimensional RKHSs with finite-dimensional spaces generated by *random features* [Rahimi, 2007; Rahimi, 2008].

To further support the practical applicability and the relevance of this model in comparison with standard architectures, we report in Section II.7 numerical experiments on MNIST and CIFAR10 datasets. They show that — as predicted by our theory — the model can be trained in these cases to almost zero loss. But more importantly, they show that the model is able to generalize well on the test dataset with performances that are similar to those of classical ResNets.

Supervised learning We consider a supervised learning problem where the RKHS-NODE model of Definition II.4 is trained for the minimization of the training risk associated to the data distribution $\mathbb{R}^d \times \mathbb{R}^{d'} \ni (x, y) \sim \mathcal{D}$ and, as in Section II.3, we consider a finite training dataset $\mathcal{D} = \{(x^i, y^i)\}_{1 \leq i \leq N} \in (\mathbb{R}^d \times \mathbb{R}^d)^N$. Then, instantiating the risk defined in Eq. (I.8) to the case of a linear parameterization of residuals, gives here:

$$\mathcal{R}(F) \coloneqq \frac{1}{N} \sum_{i=1}^{N} \ell(\mathrm{NODE}_F(x^i), y^i) = \frac{1}{N} \sum_{i=1}^{N} \ell(x_F^i(1), y^i)\,, \tag{II.24}$$

where $F \in L^2([0,1], \mathcal{F})$ is the family of residuals and x_F^i is the flow of Eq. (II.23) starting at x^i with parameterization F. Note that there is again a slight abuse of notation with the risk $\mathcal{R}$ defined in Eq. (I.8). As before, this is justified since for every parameter distribution

$\mu \in \mathcal{P}_2^{\mathrm{Leb}}([0,1],\Theta)$ we have:

$$\mathcal{R}(\mu) = \mathcal{R}(\pi(\mu))\,,$$

where $\pi : \mathcal{P}_2^{\mathrm{Leb}}([0,1],\Theta) \to L^2([0,1],\mathcal{F})$ is the surjection defined in Eq. (II.22).

II.4.1 Gradient flow equation in the case of RKHS residuals

Gradient flows for the mean-field models of NODEs were defined in Definition I.3 as solutions to some non-linear advection PDE. Before turning to the convergence analysis of such dynamics, we discuss the interpretation of this PDE in the case where the parameterization of residuals is linear w.r.t. parameters. Precisely, in this case, leveraging the isometry T between the parameter space Θ and the residual space $\mathcal{F}$ (cf. Proposition II.4.1), we show this dynamic corresponds to an actual gradient flow on the space of parameters.

To place ourselves in the setting of Chapter I we will consider in this section that the feature space $\mathcal{H}$ is finite-dimensional, though most results could probably apply to the case where $\mathcal{H}$ is an arbitrary separable Hilbert space. More importantly, we make the following regularity assumption on the feature map ϕ.

Assumption II.2 (Admissibility)**.**
We say that the RKHS $\mathcal{F}$ is admissible *if the feature map $\phi : \mathbb{R}^d \to \mathcal{H}$ is in $\mathcal{C}^2(\mathbb{R}^d,\mathcal{H})$. This in particular implies that $\mathcal{F}$ is continuously embedded in $\mathcal{C}^2(\mathbb{R}^d,\mathbb{R}^d)$ and for every $F \in \mathcal{F}$ it holds:*

$$\|F\|_{\mathcal{C}^2(\mathbb{R}^d,\mathbb{R}^d)} \leq \|\phi\|_{\mathcal{C}^2(\mathbb{R}^d,\mathcal{H})}\|F\|_{\mathcal{F}}$$

Note that the above Assumption II.2 implies that the basis function ψ defined in Eq. (II.18) satisfies all the assumptions considered in Chapter I, namely Assumptions I.1 to I.3 and Assumptions I.A to I.C. In particular, for residuals $F \in L^2([0,1],\mathcal{F})$ and data $(x,y) \in \mathbb{R}^d \times \mathbb{R}^{d'}$ the adjoint variable is defined by $p_{F,x,y} \coloneqq p_{\mu,x,y}$ where one can consider any parameterization $\mu \in \mathcal{P}_2^{\mathrm{Leb}}([0,1]\times\Theta)$ s.t. $\pi(\mu) = F$. As in Eq. (I.13), the *backward ODE* reads here:

$$\forall s \in [0,1]\,, \quad p_{F,x,y}(s) = \nabla_x \ell(x_F(1),y) + \int_s^1 \mathrm{D}_x F(r,x_F(r))^\top p_{F,x,y}(s)\,. \tag{II.25}$$

In this section, for every index $i \in \{1,...,N\}$, we will denote by p_F^i the adjoint variable associated to the data point (x^i,y^i).

In Chapter I, the adjoint variables were used in Definition I.3 to define a notion of gradient velocity field. For a parameter distribution $\mu \in \mathcal{P}_2^{\mathrm{Leb}}([0,1]\times\Theta)$, the velocity field $\nabla\mathcal{R}[\mu] \in L^2(\mu)$ reads here using the definition of ψ:

$$\forall (s,\theta) \in [0,1]\times\Theta\,, \quad \nabla\mathcal{R}[\mu](s,\theta) = \frac{1}{N}\sum_{i=1}^N \mathrm{D}_\theta\psi(\theta,x_\mu^i(s))^\top p_\mu^i(s) = \frac{1}{N}\sum_{i=1}^N p_\mu^i(s)\otimes\phi(x_\mu^i(s))\,.$$

Notably, due to the linearity of ψ w.r.t. the parameters, this velocity field is here independent of $\theta \in \Theta$. Thus, leveraging the isometry between the space of parameters Θ and the space of residuals $\mathcal{F}$, this dual vector can be used to define a dual vector on the space of residuals. Namely, for $\mu \in \mathcal{P}_2^{\mathrm{Leb}}([0,1]\times\Theta)$ and $F = \pi(\mu) \in L^2([0,1],\mathcal{F})$ we define for a.e. $s \in [0,1]$:

$$\nabla\mathcal{R}(F)(s) \coloneqq \mathrm{T}\left(\frac{1}{N}\sum_{i=1}^N p_\mu^i(s)\otimes\phi(x_\mu^i(s))\right) = \frac{1}{N}\sum_{i=1}^N K(.,x_F^i(s))p_F^i(s) \;\in \mathcal{F}\,, \tag{II.26}$$

where the second equality comes from the reproducing property of the kernel K. This definition is unambiguous since the flows x^i_μ and the adjoint variables p^i_μ only depend on the residuals $F = \pi(\mu) \in L^2([0,1],\mathcal{F})$. We show here that it corresponds to a true notion of gradient for the risk $\mathcal{R}$ defined on the space of residuals.

Proposition II.4.2. *Assume $\mathcal{F}$ is admissible according to Assumption II.2. Then $\mathcal{R}$ is continuously differentiable and $\nabla\mathcal{R}$ defined in Eq. (II.26) is its gradient.*

Proof. Let $I \subset \mathbb{R}$ be an open interval s.t. $0 \in I$ and let $(F_t)_{t\in I}$ be a smooth (at least continuously differentiable) curve in $L^2([0,1],\mathcal{F})$ s.t. $F_0 = F$. Define, for every $t \in I$, $\Theta_t = \mathrm{T}^{-1}(F_t) \in L^2([0,1],\Theta)$ and define $\mu_t = \pi^{-1}(F_t) = \int_0^1 \delta_{\theta_t(s)}\mathrm{d}s \in \mathcal{P}_2^{\mathrm{Leb}}([0,1],\Theta)$. Then by construction, we have $\mathcal{R}(F_t) = \mathcal{R}(\mu_t)$ for every $t \in I$. Also, the curve $t \in I \mapsto \mu_t$ is absolutely continuous in $\mathcal{P}_2^{\mathrm{Leb}}([0,1],\Theta)$ and satisfy the continuity equation:

$$\partial_t\mu_t + \mathrm{div}(\mu_t(0,\theta_0')) = 0 \quad \text{over } I\times[0,1]\times\Theta,$$

where $\theta_0' = \mathrm{T}^{-1}(F_0')$. Thus, applying Corollary I.3.3, the risk is differentiable at $t = 0$ and writing without ambiguity $x^i_0 = x^i_{\mu_0} = x^i_{F_0}$ and $p^i_0 = p^i_{\mu_0} = p^i_{F_0}$ we have:

$$\frac{\mathrm{d}}{\mathrm{d}t}\mathcal{R}(F_t)\Big|_{t=0} = \int_{[0,1]\times\Theta} \langle \nabla\mathcal{R}[\mu_0](s,\theta), \theta_0'(s)\rangle_\Theta \,\mathrm{d}\mu_0(s,\theta)\,.$$

Using that $\theta_0' = \mathrm{T}^{-1}(F_0')$ and the definition of $\nabla\mathcal{R}$ in Eq. (II.26) this equation reads:

$$\frac{\mathrm{d}}{\mathrm{d}t}\mathcal{R}(F_t)\Big|_{t=0} = \int_0^1 \left\langle \frac{1}{N}\sum_{i=1}^N K(.,x^i_0(s))p^i_0(s), F_0'(s)\right\rangle_{\mathcal{F}} \mathrm{d}s = \langle \nabla\mathcal{R}(F_0), F_0'\rangle_{L^2([0,1],\mathcal{F})}\,.$$

This hence shows that $\nabla\mathcal{R}$, as defined by Eq. (II.26), is the directional (or Gâteaux) derivative of $\mathcal{R}$. Since the applications $F \mapsto x^i_F$ and $F \mapsto p^i_F$ are continuous (cf. Lemmas I.3.2 and I.3.5) it follows that the map $F \mapsto \nabla\mathcal{R}(F)$ is also continuous. By classical results, this imply $\mathcal{R}$ is continuously differentiable and $\nabla\mathcal{R}$ is its gradient (see e.g. [Younes, 2010, Prop. C.1]). □

Note that, for $\mathcal{F}$ satisfying Assumption II.2, the forward flow map $F \mapsto x_F$ and the adjoint flow map $F \mapsto p_{F,x,y}$ are locally Lipschitz by Lemmas I.3.2 and I.3.5. As a consequence, the gradient map $F \in L^2([0,1],\mathcal{F}) \mapsto \mathcal{R}(F)$ is locally Lipschitz and the gradient flow equation is well-posed. That is, for any $F_0 \in L^2([0,1],\mathcal{F})$, there exists a unique solution $F \in \mathcal{C}^1_{loc}([0,+\infty), L^2([0,1],\mathcal{F}))$ of the Cauchy problem:

$$\forall t \geq 0\,, \quad \frac{\mathrm{d}}{\mathrm{d}t}F_t = -\nabla\mathcal{R}(F_t)\,.$$

We show now that the notion of gradient flow of the risk w.r.t. the parameter distribution μ as defined in the previous chapter (Definition I.3) here corresponds to the classical notion of gradient flow w.r.t. the Hilbert metric structure on the space of residuals $\mathcal{F}$.

Proposition II.4.3. *Let $\mu_0 \in \mathcal{P}_2^{\mathrm{Leb}}([0,1]\times\Theta)$ and $(\mu_t)_{t\in[0,+\infty)}$ be a gradient flow for the risk $\mathcal{R}$ starting from μ_0 given by Theorem I.3. For $t \in [0,+\infty)$, define $F_t := \pi(\mu_t) \in L^2([0,1],\mathcal{F})$. Then $(F_t)_{t\in[0,+\infty)}$ is the solution of the gradient flow equation:*

$$\forall t \geq 0\,, \quad \frac{\mathrm{d}}{\mathrm{d}t}F_t = -\nabla\mathcal{R}(F)\,.$$

Conversely, consider $(F_t)_{t\in[0,+\infty)}$ the solution of the above gradient flow equation for some $F_0 \in L^2([0,1],\mathcal{F})$. Then, defining $\mu_t = \pi^{-1}(F_t)$ for every $t \in [0,+\infty)$, the curve $(\mu_t)_{t\in[0,+\infty)}$ is the gradient flow of the risk starting from μ_0.

Proof. Note that by Jensen's inequality the mapping $\pi : \mathcal{P}_2^{\mathrm{Leb}}([0,1]\times\Theta) \to L^2([0,1],\mathcal{F})$ defined in Eq. (II.22) is a contraction. Indeed, for $\mu,\mu' \in \mathcal{P}_2^{\mathrm{Leb}}([0,1]\times\Theta)$:

$$\begin{aligned}\left\|\pi(\mu)-\pi(\mu')\right\|^2_{L^2([0,1],\mathcal{F})} &= \int_0^1 \left\|\mathbb{E}_{\mu(.|s)}[\theta]-\mathbb{E}_{\mu'(.|s)}[\theta]\right\|^2_\Theta \mathrm{d}s \\ &\leq \int_0^1 \mathcal{W}_2(\mu(.|s),\mu'(.|s))^2 \mathrm{d}s \\ &\leq \mathcal{W}_2^{\mathrm{COT}}(\mu,\mu')^2\,.\end{aligned}$$

Thus, it directly follows from the local absolute continuity of $(\mu_t)_{t\in[0,+\infty)}$ that the curve $(F_t)_{t\in[0,+\infty)}$ is locally absolutely continuous in $L^2([0,1],\mathcal{F})$. As a consequence (see e.g. [Ambrosio, 2008b, Rem. 1.1.3]), it is almost everywhere differentiable with a differential $F'_t \in L^1_{loc}([0,+\infty), L^2([0,1],\mathcal{F}))$ and for every $t_1 < t_2 \in [0,+\infty)$ it holds:

$$F_{t_2} = F_{t_1} + \int_{t_1}^{t_2} F'_t \mathrm{d}t\,.$$

Hence, to conclude it suffices to show that $F'_t = -\nabla\mathcal{R}(F_t)$ for a.e. $t\in[0,+\infty)$. For this, let us consider some $G \in L^2([0,1],\mathcal{F})$. Using the density of $\mathcal{C}^\infty([0,1],\mathcal{F})$ in $L^2([0,1],\mathcal{F})$ we can consider w.l.o.g. that G is smooth and using the isometry between Θ and $\mathcal{F}$ we write $G(s) = \mathrm{T}(\omega(s))$ for some $\omega \in \mathcal{C}^\infty([0,1],\Theta)$. Then by construction for every $t \geq 0$:

$$\langle G, F_t\rangle_{L^2([0,1],\mathcal{F})} = \int_0^1 \left\langle \omega(s), \mathbb{E}_{\mu_t(.|s)}[\theta]\right\rangle_\Theta \mathrm{d}s = \int_0^1\int_\Theta \langle\omega(s),\theta\rangle_\Theta \,\mathrm{d}\mu_t(s,\theta)\,.$$

Since, $\langle G(s),\theta\rangle_\Theta$ is smooth and bounded by a function of linear growth it follows from definition of gradient flow curves (Definition I.3) that for a.e. $t\in[0,+\infty)$:

$$\frac{\mathrm{d}}{\mathrm{d}t}\langle G, F_t\rangle_{L^2([0,1],\mathcal{F})} = -\int_{[0,1]\times\Theta}\langle\omega(s),\nabla\mathcal{R}[\mu_t](s,\theta)\rangle_\Theta \,\mathrm{d}\mu_t(s,\theta) = -\langle G, \nabla\mathcal{R}(F_t)\rangle_{L^2([0,1],\mathcal{F})}\,,$$

where we used Eq. (II.26). Hence, for every $t_1 < t_2 \in [0,+\infty)$ it holds:

$$\langle G, F_{t_2}-F_{t_1}\rangle_{L^2([0,1],\mathcal{F})} = -\int_{t_1}^{t_2}\langle G,\nabla\mathcal{R}(F_t)\rangle_{L^2([0,1],\mathcal{F})}\,\mathrm{d}t = \left\langle G, -\int_{t_1}^{t_2}\nabla\mathcal{R}(F_t)\mathrm{d}t\right\rangle_{L^2([0,1],\mathcal{F})}\,,$$

which shows that $F_{t_2}-F_{t_1} = -\int_{t_1}^{t_2}\nabla\mathcal{R}(F_t)\mathrm{d}t$ and implies the desired result.

For the converse result, note that, for every $t\geq 0$, $\mu_t = \int_0^1 \delta_{\theta_t(s)}\mathrm{d}s$ where $\theta_t = \mathrm{T}^{-1}(F_t)$. Then, for a test function $\varphi \in \mathcal{C}^\infty_c([0,1]\times\Theta)$ and for $t\geq 0$:

$$\mu_t(\varphi) = \int_{[0,1]\times\Theta}\varphi(s,\theta)\mathrm{d}\mu_t(s,\theta) = \int_0^1 \varphi(s,\theta_t(s))\,.$$

Hence differentiating w.r.t. t:

$$\frac{\mathrm{d}}{\mathrm{d}t}\mu_t(\varphi) = \int_0^1 \langle\nabla\varphi(s,\theta_t(s)),\theta'_t(s)\rangle_\Theta\,\mathrm{d}s = -\int_{[0,1]\times\Theta}\langle\nabla\varphi(s,\theta),\nabla\mathcal{R}[\mu_t](s,\theta)\rangle_\Theta\,\mathrm{d}\mu_t(s,\theta)\,,$$

where we used Eq. (II.26) and that $\mathrm{T}(\theta'_t) = F'_t = -\nabla\mathcal{R}(F_t)$. □

II.4.2 Convergence of RKHS-NODE

Following the line of proof sketched in Section II.3, we show how to derive P-Ł inequalities of the form Eq. (II.5) for the empirical risk associated with the RKHS-NODE model. For this purpose we will rely on expressivity properties of the set of residuals $\mathcal{F}$ and more precisely on the strict positivity of the kernel matrix, as defined by Definition II.3.

Similarly as the space of residuals, the associated kernel is independent of the parameter distribution. Indeed, instantiating Eq. (II.12) with ψ of the form Eq. (II.18) gives that for every $\nu \in \mathcal{P}_2(\Theta)$ and every $x, x' \in \mathbb{R}^d$:

$$K[\mu](x, x') = K(x, x') = \langle \phi(x), \phi(x') \rangle_{\mathcal{H}} \, \mathrm{Id} \,,$$

i.e. $K[\nu]$ is the kernel K associated to the RKHS $\mathcal{F}$. As before, we denote by $\mathbb{K}$ the kernel matrix defined for $\mathbf{z} = (z^i)_{1 \leq i \leq N} \in (\mathbb{R}^d)^N$ as the block matrix

$$\mathbb{K}[\mathbf{z}] \coloneqq (K(z^i, z^j))_{1 \leq i,j \leq N} \in \mathbb{R}^{dN \times dN} \,. \tag{II.27}$$

The square norm of the gradient in Eq. (II.14) thus reads here for $F \in L^2([0,1], \mathcal{F})$:

$$\|\nabla \mathcal{R}(F)\|^2_{L^2([0,1],\mathcal{F})} = \frac{1}{N^2} \int_0^1 \langle \mathbf{p}_F(s), \mathbb{K}[\mathbf{x}_F(s)] \mathbf{p}_F(s) \rangle \, \mathrm{d}s \,, \tag{II.28}$$

where $\mathbf{x}_F(s) = (x^i(s))_{1 \leq i \leq N} \in (\mathbb{R}^d)^N$ and $\mathbf{p}_F(s) \coloneqq (p^i(s))_{1 \leq i \leq N} \in \mathbb{R}^{dN}$ for every $s \in [0,1]$. Then, as in Lemma II.3.1, one can show that the risk $\mathcal{R}$ satisfies a P-Ł property whenever the kernel K is assumed to be strictly positive.

Proposition II.4.4 (RKHS-NODE satisfy P-Ł)**.** *Assume $\mathcal{F}$ satisfies Assumption II.2, its associated kernel K is strictly positive in the sense of Definition II.3 and the input data has separation $\delta \coloneqq \min_{i \neq j} \|x^i - x^j\| > 0$. Then, for every $R \geq 0$, the empirical risk $\mathcal{R}$ satisfies the (R, m)-P-Ł property of Definition II.1 with m given by:*

$$m = \frac{1}{N} \lambda_K \left(\delta e^{-\kappa R} \right) e^{-2\kappa R} \,. \tag{II.29}$$

where $\kappa = \kappa(\phi)$ and $\lambda_K : \mathbb{R}_{>0} \to \mathbb{R}_{>0}$ is the positive increasing function (possibly depending on N) defined by:

$$\lambda_K(\delta) \coloneqq \inf_{\substack{\mathbf{z} = (z^i) \in (\mathbb{R}^d)^N \\ \min_{i \neq j} \|z^i - z^j\| \geq \delta}} \lambda_{\min}(\mathbb{K}[\mathbf{z}]) \tag{II.30}$$

Proof. Let $R > 0$ and consider $F \in L^2([0,1], \mathcal{F})$ s.t. $\|F\|_{L^2([0,1],\mathcal{F})} \leq R$. First, note that it follows from Assumption II.2 and from the forward and backward ODEs in Eqs. (II.23) and (II.25) that we have for every index $i, j \in \{1, ..., N\}$ and every $s \in [0,1]$:

$$\left\| x^i_F(s) - x^j_F(s) \right\| \geq e^{-\kappa R} \delta$$

and

$$\|p^i_F(s)\| \geq e^{-\kappa R} \|\nabla_x \ell(x^i_F(1), y^i)\| \,,$$

where $\kappa = \|\phi\|_{\mathcal{C}^2}$. Then, plugging this into Eq. (II.28) and using the definition of λ_K gives:

$$\begin{aligned}
\|\nabla\mathcal{R}(F)\|^2_{L^2([0,1],\mathcal{F})} &= \frac{1}{N^2}\int_0^1 \langle \mathbf{p}_F(s), \mathbb{K}[\mathbf{x}_F(s)]\mathbf{p}_F(s)\rangle \,\mathrm{d}s \\
&\geq \frac{1}{N^2}\int_0^1 \lambda_{\min}(\mathbb{K}[\mathbf{x}_F(s)])\|\mathbf{p}_F(s)\|^2 \mathrm{d}s \\
&\geq \frac{1}{N^2}\lambda_K(e^{-\kappa R}\delta)e^{-2\kappa R}\sum_{i=1}^N \|\nabla_x \ell(x_F^i(1), y^i)\|^2 \\
&\geq \frac{1}{N^2}\lambda_K(e^{-\kappa R}\delta)e^{-2\kappa R}\sum_{i=1}^N \ell(x_F^i(1), y^i) \\
&= \frac{1}{N}\lambda_K(e^{-\kappa R}\delta)e^{-2\kappa R}\mathcal{R}(F)\,,
\end{aligned}$$

where we used Assumption II.1 in the penultimate line. □

Since the empirical risk $\mathcal{R}$ satisfies a local P-Ł property, the analysis of Section II.2 gives the convergence of gradient flow curves, provided the risk at initialization is already sufficiently low. This condition depend on the kernel K through its conditioning λ_K defined in Eq. (II.30). While we keep here an abstract condition for the sake of generality, the quantitative dependence of λ_K w.r.t. the data separation $\delta > 0$ will be made explicit for a large class of kernels in Section II.4.3.

Theorem II.4. *Let the assumptions of Proposition II.4.4 be satisfied and consider the associated constants δ, κ and the function λ_K defined in Eq. (II.30). Let $F_0 \in L^2([0,1],\mathcal{F})$ be some initialization and write $\|F_0\|_{L^2} = R_0$. Assume there exists $R \geq 0$ s.t.:*

$$4N\mathcal{R}(F_0) < R^2\lambda_K(\delta e^{-\kappa(R+R_0)})e^{-2\kappa(R+R_0)}\,. \tag{II.31}$$

Then, the gradient flow $(F_t)_{t\geq 0}$ of $\mathcal{R}$ with initialization F_0 converges to some $F_\infty \in L^2([0,1],\mathcal{F})$ and for every $t \geq 0$ it holds:

$$\mathcal{R}(F_t) \leq e^{-mt}\mathcal{R}(F^0)\,, \quad \text{and} \quad \|F_t - F_\infty\|_{L^2([0,1],\mathcal{F})} \leq e^{-mt/2}R\,,$$

where $m = \frac{1}{N}\lambda_K(\delta e^{-\kappa(R+R_0)})e^{-2\kappa(R+R_0)}$.

Proof. It follows from the assumptions and from Proposition II.4.4 that $\mathcal{R} : L^2([0,1],\mathcal{F}) \to \mathbb{R}$ satisfies the (R,m)-P-Ł property of Definition II.1 around F_0 with

$$m = \frac{1}{N}\lambda_K(\delta e^{-\kappa(R+R_0)})e^{-2\kappa(R+R_0)}\,.$$

The result then follows from an application of Theorem II.1. □

Moreover, a similar conclusion holds for gradient descent on the risk $\mathcal{R}$. The following result is an application of Theorem II.2.

Theorem II.5. *Let the assumptions of Proposition II.4.4 be satisfied and consider the associated constants δ, κ and the function λ_K defined in Eq. (II.30). Let $F_0 \in L^2([0,1],\mathcal{F})$ be some initialization and write $\|F_0\|_{L^2} = R_0$. Assume there exists $R \geq 0$ s.t.:*

$$4N\mathcal{R}(F_0) < R^2\lambda_K(\delta e^{-\kappa(R+R_0)})e^{-2\kappa(R+R_0)}\,.$$

Define $m = \frac{1}{N}\lambda_K(\delta e^{-\kappa(R+R_0)})e^{-2\kappa(R+R_0)}$. *Then, for any* $\alpha \in \left(\sqrt{\frac{2\mathcal{R}(F_0)}{mR^2}}, 1\right)$, *there exists a sufficiently small step-size* $\tau > 0$, *s.t. the iterates* $(F_k)_{k\geq 0}$ *of gradient descent on* $\mathcal{R}$ *with initialization* F_0 *and step-size* τ *satisfy for every* $k \geq 0$*:*

$$\mathcal{R}(F_k) \leq (1-\alpha m\tau)^k \mathcal{R}(F^0)\,, \quad \textit{and} \quad \|F_k - F_\infty\|_{L^2([0,1],\mathcal{F})} \leq (1-\alpha m\tau)^{k/2} R\,,$$

where $F_\infty \in \bar{B}(F_0, R)$ *is s.t.* $\mathcal{R}(F_\infty) = 0$.

Proof. It follows from the assumptions and from Proposition II.4.4 that $\mathcal{R} : L^2([0,1],\mathcal{F}) \to \mathbb{R}$ satisfies the (R,m)-P-Ł property of Definition II.1 around F_0 with

$$m = \frac{1}{N}\lambda_K(\delta e^{-\kappa(R+R_0)})e^{-2\kappa(R+R_0)}\,.$$

The result then follows from an application of Theorem II.2. □

Note that there are two important parameters determining the P-Ł constant m in Proposition II.4.4 and thus the convergence rate of gradient flow in Theorem II.4 and of gradient descent in Theorem II.5. The first one is the data-separation $\delta = \min_{i\neq j}\|x^i - x^j\|$ which is a priori imposed by the dataset but which could be increased by an appropriate preprocessing of the data such as normalization, rescaling or embedding in high dimension. The other parameter is the function λ_K which depends on the choice of the kernel K and thus on the choice of a functional space $\mathcal{F}$ for the residuals. In Section II.4.3, we use results on condition number for radial basis function interpolation problems [Schaback, 1995] to provide a lower bound on λ_K in the case of radial kernels (e.g. Gaussian or Matérn kernels). However, if in theory, prior knowledge of the data might allow to optimize the choice of kernel, we expect the selection of an optimal kernel to be an intractable problem in practice. Instead, cross-validation techniques can be used to select a suitable kernel.

Finally, the degeneracy of the P-Ł constant m as $R \to +\infty$ readily appears in Proposition II.4.4. Thus one should not expect these bounds to imply global convergence of gradient descent without making any further assumption. Indeed, cases where gradient descent fails to converge towards a global optimizer of the loss are observed in [Bartlett, 2018]. Instead, Theorems II.4 and II.5 are local convergence results in which the condition in Eq. (II.31) expresses a threshold between two kinds of behaviours: **(i)** if $\mathcal{R}(F^0)$ is sufficiently small, the training dynamic converges towards a global minimizer. The limiting behaviour is when the l.h.s. of Eq. (II.31) tends to 0. Because of a regularizing effect of gradient descent (i.e. that $\|F^k - F^0\|_{L^2} \leq R$), the parameter stays in a ball of arbitrary small radius R all along the training dynamic. In this limit, we recover a "linear" or "kernel" regime where the model is well approximated by its linearization at F^0 [Chizat, 2018; Liu, 2020; Jacot, 2018]. **(ii)** If $\mathcal{R}(F^0)$ is too large, the result says nothing about the convergence of gradient descent. However, it is still observed in practice that the training dynamic often converges towards a global minimizer of the loss [Zhang, 2021]. Explaining this phenomenon in a general setting remains a challenging open question, even for simple linear models.

II.4.3 Convergence with finite width

While Theorems II.4 and II.5 describe local convergence results for the training of deep ResNets with gradient descent and gradient flow, the strict positivity assumption requires the space of residuals to be of very high dimension. Actually, typical examples of RKHSs satisfying Assumption II.2 and having a strictly positive kernel in the sense of Definition II.3 would be Sobolev spaces H^ν of regularity $\nu > d/2 + 2$. Those are described

by considering as feature map the Fourier coefficients $\phi(x) = (e^{\imath x^\top w})_{w\in\mathbb{R}^d}$ and as feature space $\mathcal{H} = L^2(\rho_\nu)$, where $\rho_\nu \in \mathcal{P}(\mathbb{R}^d)$ is the probability distribution with density:

$$\forall w \in \mathbb{R}^d\,, \quad \rho_\nu(w) \propto \left(1 + \frac{\|w\|^2}{2\nu}\right)^\nu .$$

The associated kernel K_ν is the translation-invariant kernel whose Fourier transform is ρ_ν:

$$\forall x, x' \in \mathbb{R}^d\,, \quad K_\nu(x,x') = k_\nu(x,x')\mathrm{Id} \quad \text{with} \quad k_\nu(x,x') = \int_{\mathbb{R}^d} e^{\imath(x-x')^\top w}\mathrm{d}\rho_\nu(w)\,. \tag{II.32}$$

Note in particular that, taking the limit $\nu \to +\infty$, one recovers a Gaussian kernel, which we denote by $K_\infty = k_\infty \mathrm{Id}$. In any case, for every $\nu \in (d/2+2, +\infty]$, the RKHS H^ν has infinite dimension.

In contrast, in standard ResNets, the space of residuals usually consists of a parametric function space of large but finite dimension. In the case of a linear parameterization, a typical example would be a *random feature model* [Rahimi, 2007] consisting in a 2-layer perceptron (Eq. (34)) whose hidden layer weights are fixed. For a width $M \geq 1$ and parameters $\theta \in \mathbb{R}^{d\times M}$, the residuals are of the form:

$$\forall x \in \mathbb{R}^d\,, \quad F_\theta(x) = \sqrt{\frac{2}{M}}\,\theta \cdot \sigma(W^\top x + b)\,, \tag{II.33}$$

where σ is some activation, $W = (w_1|...|w_M) \in \mathbb{R}^{d\times M}$ is a fixed weight matrix, whose column are called *features*, and $b = (b_i)_{1\leq i\leq M} \in \mathbb{R}^M$ is some bias vector. In particular, if considering the trigonometric activation $\sigma = \cos$, i.i.d. features $w_i \sim \rho_\nu$ for $\nu > d/2+2$ and i.i.d. biases $b_i \sim \mathcal{U}([0,\pi])$, it follows from Proposition II.4.1 that the space of residual maps $\mathcal{F}$ described in Eq. (II.20) is the RKHS associated to the kernel:

$$\forall x, x' \in \mathbb{R}^d\,, \quad \hat{K}_\nu(x,x') = \hat{k}_\nu(x,x')\mathrm{Id}\,, \tag{II.34}$$

with

$$\hat{k}_\nu(x,x') = \frac{2}{M}\sum_{i=1}^M \cos(w_i^\top x + b_i)\cos(w_i^\top x' + b_i)\,.$$

Using the law of large numbers, we see after some calculations that:

$$\hat{k}_\nu(x,x') \xrightarrow{M\to+\infty} 2\int_{\mathbb{R}^d\times[0,\pi]} \cos(w^\top x + b)\cos(w^\top x' + b)\mathrm{d}\rho_\nu(w)\mathrm{d}b = k_\nu(x,x')\,.$$

Thus, this space of residuals, which we will from now-on denote by $\hat{H}^\nu$, is a finite-dimensional approximation of H^ν. In the rest of this section, we show that the convergence results in Theorems II.4 and II.5 hold for residuals in $\hat{H}^\nu$ provided the width M is sufficiently large w.r.t. the number of samples.

We start by showing that the admissibility and strict-positivity assumptions are both satisfied by $\hat{H}^\nu$, respectively almost surely and with great probability over the sampling of random features. We will then conclude to convergence in Theorem II.6.

Lemma II.4.1. *Let $\nu \in (d/2+2, +\infty]$ and, for $M \geq 1$, let $\mathcal{F} = \hat{H}^\nu$ be the RKHS with kernel $\hat{K}_\nu$ defined in Eq. (II.34). Then, almost surely, $\mathcal{F}$ satisfies Assumption II.2 with a a constant $\hat{\kappa}$ independent of $M \geq 1$.*

Proof. Note that the feature map associated to the RKHS $\mathcal{F} = \hat{H}^\nu$ is defined by:

$$\forall x \in \mathbb{R}^d\,, \quad \hat{\phi}_\nu(x) = \sqrt{\frac{2}{M}}\sigma\left(w^\top x + b\right) \in \mathbb{R}^M\,.$$

In particular, since here $\sigma = \cos$, we have $\sup_x \|\hat{\phi}_\nu(x)\| \leq \sqrt{2}$. Moreover, for $x \in \mathbb{R}^d$:

$$\mathrm{D}\hat{\phi}_\nu(x)^\top \mathrm{D}\hat{\phi}_\nu(x) = \frac{2}{M}\sum_{i=1}^M \cos(w_i^\top x + b_i)^2 w_i w_i^\top \preceq \frac{2}{M}\sum_{i=1}^M \|w_i\|^2 \mathrm{Id}\,.$$

Since $\nu > d/2+1$, ρ_ν has finite second-order moments and this sequence converges almost surely by the law of large numbers. In particular, $\sup_x \|\mathrm{D}\hat{\phi}_\nu(x)\|$ is almost surely bounded w.r.t. $M \geq 1$. Using that $\nu > d/2+2$, a similar argument shows that $\sup_x \|\mathrm{D}^2\hat{\phi}_\nu(x)\|$ is almost surely bounded w.r.t. $M \geq 1$. This shows the result. □

Lemma II.4.2. *Let $\nu \in (d/2+2, +\infty]$ and, for $M \geq 1$, let $\mathcal{F} = \hat{H}^\nu$ be the RKHS with kernel $\hat{K}_\nu$ defined in Eq. (II.34). Consider a number of data sample $N \geq 1$ and let $\hat{\mathbb{K}}_\nu$ and $\mathbb{K}_\nu$ be the kernel matrices associated to the kernels $\hat{K}_\nu$ and K_ν respectively. Consider $\varepsilon, \gamma > 0$ and $R \geq 1$. There exists a constant $C > 0$ s.t. if $M \geq C\varepsilon^{-2}N^2(1+\log(R)+\gamma)$, then with probability greater than $1-e^{-\gamma}$, for any $F \in L^2([0,1],\mathcal{F})$ s.t. $\|F\|_{L^2([0,1],\mathcal{F})} \leq R$ it holds:*

$$\forall s \in [0,1]\,, \quad \lambda_{\min}\left(\hat{\mathbb{K}}_\nu(\mathbf{x}(s))\right) \geq \lambda_{\min}\left(\mathbb{K}_\nu(\mathbf{x}(s))\right) - \varepsilon\,,$$

where $\mathbf{x}(s) = (x^i(s))_{1\leq i\leq N}$ are the solutions to the forward ODE Eq. (II.23).

Proof. Consider $\hat{\kappa}$, independent of $M \geq 1$, such that $\mathcal{F} = \hat{H}^\nu$ satisfies Assumption II.2 with constant $\hat{\kappa}$. Then, for $\|F\|_{L^2([0,1],\mathcal{F})} \leq R$, it holds for every index $i \in \{1,...,N\}$ that $\|x^i(s)\| \leq \|x^i(0)\| + \hat{\kappa}R$ for every $s \in [0,1]$. Then using [Sriperumbudur, 2015, Thm.1], there exists a constant $C = C(d,\nu)$ s.t.:

$$\mathbb{P}\left(\sup_{\|F\|_{L^2}\leq R}\sup_{1\leq i,j\leq N}\sup_{s\in[0,1]}\left\|\hat{k}_\nu(x^i(s),x^j(s)) - k_\nu(x^i(s),x^j(s))\right\| \geq \frac{C(1+\log(R))+\sqrt{2\gamma}}{\sqrt{M}}\right) \leq e^{-\gamma}\,.$$

This gives the desired result by considering that $\lambda_{\min}$ is N-Lipschitz continuous on the set $N \times N$ symmetric matrices. □

As a consequence of the two above lemma, we recover convergence of gradient flow and gradient descent for the training of the RKHS-NODE model defined in Definition II.4 with residuals of the form Eq. (II.33), provided the width M is sufficiently large. Note that, in the following theorem, one can distinguish between two kind of assumptions: the assumption that the risk at initialization is sufficiently small, allowing the application of the local convergence results in Theorems II.1 and II.2, and the assumption of a sufficiently large number of random features, allowing for the RKHS $\hat{H}^\nu$ to recover the expressivity property of H^ν with great probability. In particular, taking the limit $M \to \infty$, one recovers that convergence hold with probability 1 when considering residuals in the infinite dimensional space H^ν.

Theorem II.6. *Let $\nu \in (d/2+2, +\infty]$ and, for $M \geq 1$, let $\mathcal{F} = \hat{H}^\nu$ be the RKHS with kernel $\hat{K}_\nu$ defined in Eq. (II.34). Consider $N \geq 1$ input data samples $(x^i)_{1\leq i\leq N} \in (\mathbb{R}^d)^N$ with data separation $\delta := \min_{i\neq j}\|x^i - x^j\| > 0$. Consider the initialization $F_0 = 0 \in L^2([0,1],\mathcal{F})$. Then there exists a constant $C > 0$, s.t. for every $\gamma > 0$ the conclusions of Theorems II.4 and II.5 apply with probability greater than $1-e^{-\gamma}$ if:*

- *Sobolev / Matérn kernel ($\nu < +\infty$):*

$$\mathcal{R}(0) < C^{-1}N^{-1}\delta^{2\nu-d}, \quad \textit{and} \quad M \geq CN^2\delta^{2d-4\nu}(1+\gamma)\,,$$

- *Gaussian kernel ($\nu = +\infty$):*

$$\mathcal{R}(0) < C^{-1}N^{-1}\delta^{-d}e^{-C\delta^{-2}}, \quad \textit{and} \quad M \geq CN^2\delta^{2d}e^{+2C\delta^{-2}}(1+\gamma)\,.$$

Proof. The result follows from lower bounds on the conditioning of translation-invariant kernels of the form K_ν in Eq. (II.32) [Schaback, 1995]. We make the proof in the case of Matérn kernels, that is for $\nu \in (d/2+2, +\infty)$, but the arguments straightforwardly adapt to the case of Gaussian kernels, i.e. $\nu = +\infty$.

For $\nu \in (d/2+2, +\infty)$, Schaback [Schaback, 1995] gives that, for a data separation $\delta' > 0$, a lower bound on the conditioning of the kernel K_ν is:

$$\lambda_{K_\nu}(\delta') \geq C_1^{-1}(\delta')^{2\nu-d}\,,$$

where $C_1 = C_1(d,\nu)$. Let $\hat{\kappa}$ be the constant provided by Lemma II.4.1. Consider $R = 1$ and $\varepsilon = \frac{1}{2}C^{-1}(\delta e^{-\hat{\kappa}})^{2\nu-d}$ in Lemma II.4.2. Then there exists a constant $C_2 = C_2(d,\nu)$ such that, for every $\gamma > 0$, if $M \geq C_2\delta^{2d-4\nu}N^2(1+\gamma)$, with probability greater than $1-e^{-\gamma}$, for every $F \in L^2([0,1],\mathcal{F})$ s.t. $\|F\|_{L^2([0,1],\mathcal{F})} \leq 1$ it holds:

$$\lambda_{\min}\left(\hat{\mathbb{K}}_\nu(\mathbf{x}(s))\right) \geq \lambda_{\min}\left(\mathbb{K}_\nu(\mathbf{x}(s))\right) - \varepsilon \geq \frac{1}{2}C_1^{-1}e^{-(2\nu-d)\hat{\kappa}}\delta^{2\nu-d}\,,$$

where $\mathbf{x}(s) = (x^i(s))_{1\leq i\leq N}$ are the solutions to the forward ODE Eq. (II.23). As a consequence, the risk $\mathcal{R}$ satisfies the (R,m)-P-Ł property of Definition II.1 around the initialization $F = 0$ with $R = 1$ and $m = C_3^{-1}N^{-1}\delta^{2\nu-d}$ for some constant $C_3 = C_3(d,\nu)$. The condition on $\mathcal{R}(0)$ then allows applying Theorem II.1 for the convergence of gradient flow or Theorem II.1 for the convergence of gradient descent. □

II.5 The case of SHL residuals

In the above Section II.4, we derived convergence results for the training of deep ResNets or NODEs whose residuals are linearly parameterized. We study here the case where residuals are *single-hidden-layer (SHL)* perceptrons of the form in Eq. (34). For $M \geq 1$, weight matrices $U, W \in \mathbb{R}^{d\times M}$ and bias vector $b \in \mathbb{R}^M$, a SHL perceptron of width M is described by:

$$\forall x \in \mathbb{R}^d\,, \quad F_{(U,W,b)}(x) = \frac{1}{M}U\sigma(W^\top x + b)\,,$$

where $\sigma : \mathbb{R} \to \mathbb{R}$ is an activation function applied component-wise. In particular, it is similar to the random feature model (Eq. (II.33)) previously considered in Section II.4 with the notable difference that both outer weights in U and inner weights in W, b are learned parameters.

Note that we consider the mean-field scaling factor $1/M$. With this choice of scaling, the SHL architecture is an instance of Eq. (II.2) which we define by setting the parameter space $\Theta = \mathbb{R}^d \times \mathbb{R}^d \times \mathbb{R}$ and the map:

$$\psi : ((u,w,b),x) \in \Theta \times \mathbb{R}^d \mapsto u\sigma(w^\top x + b)\,. \tag{II.35}$$

Indeed, if $(u_i)_{1\leq i\leq M}$ and $(w_i)_{1\leq i\leq M}$ are the columns of U and W respectively then for every $x \in \mathbb{R}^d$:

$$F_{(U,W,b)}(x) = \frac{1}{M}\sum_{i=1}^{M} u_i \sigma(w_i^\top x + b_i)\,.$$

We will generically make the following assumption on the activation σ to ensure that the results of Chapter I on existence and uniqueness of the gradient flow dynamic still hold when considering the basis function ψ.

Assumption II.3. *The activation $\sigma : \mathbb{R} \to \mathbb{R}$ is a twice continuously differentiable function with a uniformly bounded derivative. Defining $C = C(\sigma) := |\sigma(0)| + \|\sigma'\|_\infty$, we then have for $(x, \theta) \in \mathbb{R}^d \times \Theta$:*

$$\begin{aligned} &\|\psi(\theta, x)\| \leq C(1 + \|x\|)(1 + \|\theta\|^2)\,, \\ &\|\mathrm{D}_\theta\psi(\theta, x)\| \leq C(1 + \|x\|)(1 + \|\theta\|)\,, \\ &\|\mathrm{D}_x\psi(\theta, x)\| \leq C\|\theta\|^2\,. \end{aligned} \tag{II.36}$$

Thus, this assumption ensures that Assumptions I.1 to I.3 are satisfied. It does not however imply Assumptions I.A and I.B. Still we are able to show that Theorems I.3 and I.4 both hold for SHL architectures (c.f. Propositions I.A.1 and I.A.2 in Section I.A).

Remark II.5.1. *Assumption II.3 is in particular satisfied for the popular choices that are $\sigma =$ tanh or any smooth approximation of* ReLU *such as* GeLU *or* Swish*, but considering the* ReLU *activation itself is expected to create two kinds of issues. First, the non-differentiability of* ReLU *at* 0 *could create singularities in the continuity equation. As a consequence, while existence of solutions to the gradient flow equation (Definition I.3) might still hold, one should not expect those solutions to be unique (Theorem I.4). Then, and perhaps most importantly, those solutions might not coincide with curves of maximal slope. Indeed, a cornerstone of our analysis is Theorem I.2, identifying gradient flow curves (Definition I.3) with curves of maximal slopes for the risk (Definition I.5). This result requires minimal regularity on ψ and allows showing existence and uniqueness of gradient flow curves in Section I.3.4.*

Following the lines of Section II.3, our proof strategy to show convergence of gradient flow for the training of deep ResNets with SHL residuals will be to study the conditioning of the associated Neural Tangent Kernel (NTK) during training. Recall the definition Eq. (II.12) of the tangent kernel K associated to the architecture defined by ψ and to some parameterization $\mu \in \mathcal{P}_2(\Theta)$:

$$\forall x, x' \in \mathbb{R}^d\,, \quad K[\mu](x, x') := \int_\Theta \mathrm{D}_\theta\psi(\theta, x)\mathrm{D}_\theta\psi(\theta, x')^\top \mathrm{d}\mu(\theta)\,. \tag{II.37}$$

In the case of the SHL architecture defined by Eq. (II.35), the associated kernel can be decomposed into two parts. For $\mu \in \mathcal{P}_2(\Theta)$ we have $K[\mu] = k^1[\mu]\mathrm{Id} + K^2[\mu]$ where we define for every $x, y \in \mathbb{R}^d$:

$$\begin{aligned} k^1[\mu](x, y) &:= \int_{\mathbb{R}^d\times\mathbb{R}^d\times\mathbb{R}} \sigma(w^\top x + b)\sigma(w^\top y + b)\mathrm{d}\mu(u, w, b)\,, \\ K^2[\mu](x, y) &:= \int_{\mathbb{R}^d\times\mathbb{R}^d\times\mathbb{R}} \sigma'(w^\top x + b)\sigma'(w^\top y + b)(x^\top y + 1)(u \otimes u)\mathrm{d}\mu(u, w, b)\,. \end{aligned} \tag{II.38}$$

Observing that both $k^1[\mu]$ and $K^2[\mu]$ define positive kernels over $\mathbb{R}^d$ we have that $K[\mu] \geq k^1[\mu]\mathrm{Id}$ in the sense of positive kernels. Therefore $\lambda_{\min}(\mathbb{K}^1[\mu(.|s), \mathbf{x}_\mu(s)])$ provides a natural lower bound for $\lambda_{\min}(\mathbb{K}[\mu(.|s), \mathbf{x}_\mu(s)])$ where, similarly to Eq. (II.13) the kernel matrix $\mathbb{K}^1[\mu, \mathbf{z}] \in \mathbb{R}^{N\times N}$ is defined for a point cloud $\mathbf{z} = (z^i)_{1\leq i\leq N}$ and for $\mu \in \mathcal{P}_2(\Theta)$ as:

$$\mathbb{K}^1[\mu, \mathbf{z}] \coloneqq \left(k^1[\mu](z^i, z^j)\right)_{1\leq i,j\leq N} \in \mathbb{R}^{N\times N}.$$

In Theorem II.7, we will rely on the conditioning of the kernel k^1 during training to ensure convergence of gradient flow.

II.5.1 Comparison with the case of a linear parameterization

An important improvement of this section w.r.t. the analysis performed in Section II.4 is that we consider a more realistic setting where residuals are 2-layer neural networks whose hidden layer weights are learned.

Leveraging the linearity of ψ w.r.t. the outer layer weights, one can replace u with its conditional expectation w.r.t. the inner layer weights (w, b). For a parameterization $\mu \in \mathcal{P}_2(\Theta)$, the residual is then equivalently represented by the marginal μ^2 of μ w.r.t. (w, b) — the *feature distribution* — and by the conditional expectation $u(w, b) = \mathbb{E}_\mu[u|w, b] \in L^2(\mu^2)$:

$$\forall x \in \mathbb{R}^d, \quad F_\mu(x) = \int_\Theta u\sigma(w^\top x + b)\mathrm{d}\mu(u, w, b) = \int_{\mathbb{R}^d\times\mathbb{R}} u(w, b)\sigma(w^\top x + b)\mathrm{d}\mu^2(w, b).$$

Such a residual belongs to the RKHS associated with the feature space $\mathcal{H}_\mu = L^2(\mu^2)$ and the feature map $\phi : x \mapsto \sigma(w^\top x+b)$. The associated kernel is $k^1[\mu]$, which as in Eq. (II.21) reads:

$$\forall x, x' \in \mathbb{R}^d, \quad k^1[\mu](x, x') = \langle\phi(x), \phi(x')\rangle_{L^2(\mu^2)} = \int_{\mathbb{R}^d\times\mathbb{R}} \sigma(w^\top x + b)\sigma(w^\top x' + b)\mathrm{d}\mu^2(w, b).$$

Thus, fixing the inner weight distribution μ^2 one would recover the setting of Section II.4, with residuals in a RKHS independent of the parameterization. In contrast, in Theorem II.7, the distribution of the inner layer weights evolves during training. Tracking evolution of the feature distribution in deep neural networks is however a difficult theoretical problem. In Theorem II.7 we will overcome this issue by assuming the risk at initialization is sufficiently small for the training dynamic to stay close from some "nice" feature distribution. We will quantify this condition w.r.t. the number of data sample N in Corollary II.5.1 but, as a consequence of not being able to track the learning of the feature distribution, we will ask for the risk at initialization to scale as N^{-3} in contrast to N^{-1} in Theorem II.6. This gap motivates a detailed analysis of the evolution of the feature distribution in shallow architectures which will be the content of Chapter III.

Mathematically, training of inner weights also materializes as a change of metric on the space of residual mappings which is no longer isometric to its parameter space. Here the NTK in fact decomposes as a sum of two terms: k^1 corresponds to gradients w.r.t. linear parameters while K^2 corresponds to gradients w.r.t. nonlinear parameters. The space of residuals is described by the so-called "Barron space"

$$\mathcal{B} \coloneqq \left\{F : x \mapsto \int u\sigma(w^\top x + b)\mathrm{d}\mu(u, w, b) \ : \ \mu \in \mathcal{P}_2(\mathbb{R}\times\mathbb{R}^d\times\mathbb{R})\right\}.$$

In the case $\sigma = \mathrm{ReLU}$, E and Wojtowytsch [E, 2022] show that $\mathcal{B}$ can be endowed with a Banach norm:

$$\forall F \in \mathcal{B}, \quad \|F\|_{\mathcal{B}} \coloneqq \inf\left\{\int |u|(\|w\| + |b|)\mathrm{d}\mu \ : \ \mu \in \mathcal{P}_2(\mathbb{R}\times\mathbb{R}^d\times\mathbb{R}), \ F = F_\mu\right\}.$$

However, if it satisfies enjoyable approximation properties such as density in the space of continuous functions [Cybenko, 1989], this Banach space is generically not separable nor reflexive.

II.5.2 Convergence of NODEs with SHL residuals

We show here a local convergence result for the training of NODEs with gradient flow in the case of SHL residuals. Theorem II.7 here assumes the risk at initialization is already sufficiently small and we will show in Section II.5.3 that this assumption can be quantified explicitly when specifying the activation and the initial parameterization.

First, we show the conditioning of the kernel k^1 defined in Eq. (II.38) is well behaved w.r.t. to the metric $\mathcal{W}_2^{\mathrm{COT}}$ on the parameter set $\mathcal{P}_2^{\mathrm{Leb}}([0,1]\times\Theta)$.

Lemma II.5.1. *Assume σ satisfies Assumption II.3. Then the map*

$$\mu \mapsto \int_0^1 \lambda_{\min}(\mathbb{K}^1[\mu(.|s), \mathbf{x}_\mu(s)])\mathrm{d}s$$

is locally-Lipschitz continuous on $(\mathcal{P}_2^{\mathrm{Leb}}([0,1]\times\Theta), \mathcal{W}_2^{\mathrm{COT}})$. Moreover there exists some constant C such that if μ, μ' are such that $\mathcal{E}_2(\mu)$, $\mathcal{E}_2(\mu') \leq \mathcal{E}$ then :

$$\left|\int_0^1 \lambda_{\min}(\mathbb{K}^1[\mu, \mathbf{x}_\mu]) - \int_0^1 \lambda_{\min}(\mathbb{K}^1[\mu', \mathbf{x}_{\mu'}])\right| \leq NCe^{C\mathcal{E}_2(\mu_0)}\mathcal{W}_2^{\mathrm{COT}}(\mu, \mu')\,.$$

Proof. Let $C = C(\sigma)$ be the constant appearing in Eq. (II.36) and let $R \geq 0$ be such that $\mathrm{Supp}(\mathcal{D}) \subset B(0,R)$. We have by Proposition I.1.1 that for $\mu \in \mathcal{P}_2^{\mathrm{Leb}}([0,1]\times\Theta)$ and for $x \in \mathrm{Supp}(\mathcal{D}_x)$ the flow verifies:

$$\forall s \in [0,1]\,, \quad \|x_\mu(s)\| \leq e^{C(1+\mathcal{E}_2(\mu))}(R + C(1+\mathcal{E}_2(\mu))) \leq C_1 e^{C_1\mathcal{E}_2(\mu)}\,,$$

where $C_1 = C_1(R,\sigma)$. Using the previous bound on the trajectories as well as the bounds in Eq. (II.36) we see following the proof of Lemma I.3.2 that if $\mathcal{E}_2(\mu)$, $\mathcal{E}_2(\mu') \leq \mathcal{E}$ then for every $s \in [0,1]$:

$$\|x_\mu(s) - x_{\mu'}(s)\| \leq e^{C\mathcal{E}}(1 + C_1 e^{C_1\mathcal{E}})\sqrt{2+4\mathcal{E}}\mathcal{W}_2^{\mathrm{COT}}(\mu,\mu') \leq C_2 e^{C_2\mathcal{E}}\mathcal{W}_2^{\mathrm{COT}}(\mu,\mu')\,,$$

where $C_2 = C_2(R,\sigma)$. Also, it follows from the assumptions on σ that, for fixed $\mu \in \mathcal{P}_2(\Theta)$, the map $(x,y) \in \mathbb{R}^{2d} \mapsto k^1[\mu](x,y)$ is locally Lipschitz and, for any x, x', y, $y' \in \mathbb{R}^d$,

$$\left|k^1[\mu](x,y) - k^1[\mu](x',y')\right| \leq C^2\mathcal{E}_2(\mu)(1 + \|x'\| + \|y\|)(\|x - x'\| + \|y - y'\|)\,.$$

For fixed $x, y \in \mathbb{R}^d$, the map $\mu \in \mathcal{P}_2(\Theta) \mapsto k^1[\mu](x,y)$ is also locally lipschitz and using Assumption II.3 we have that if $\mathcal{E}_2(\mu)$, $\mathcal{E}_2(\mu') \leq \mathcal{E}$ then for some constant C_4:

$$\forall x, y \in \mathbb{R}^d\,, \quad \left|k^1[\mu](x,y) - k^1[\mu'](x,y)\right| \leq C_4(1 + \|x\| + \|y\|)(1 + \sqrt{\mathcal{E}})\mathcal{W}_2(\mu,\mu')\,.$$

Compiling the previous inequalities we have that if $\mathcal{E}$ is such that $\mathcal{E}_2(\mu)$, $\mathcal{E}_2(\mu') \leq \mathcal{E}$ then:

$$\|\mathbb{K}^1[\mu, \mathbf{x}_\mu] - \mathbb{K}^1[\mu', \mathbf{x}_{\mu'}]\|_\infty \leq C_5 e^{C_5\mathcal{E}}\mathcal{W}_2^{\mathrm{COT}}(\mu,\mu')$$

where $C_5 = C_5(R,\sigma)$ and $\|.\|_\infty$ is the supremum norm on matrices. Finally, the result follows from the N-Lipschitz continuity of the map $S \mapsto \lambda_{\min}(S)$ on the space of $N \times N$ symmetric matrices provided with $\|.\|_\infty$. □

The following result gives sufficient conditions for the convergence of the gradient flow towards a global minimizer of the risk in the case of a NODE with SHL residuals.

Theorem II.7. *Assume ψ is of the form Eq. (II.35) with an activation σ satisfying Assumption II.3 and that ℓ satisfies Assumption II.1. Then for any $\mu_0 \in \mathcal{P}_2^{\mathrm{Leb}}([0,1] \times \Theta)$ there exists a positive constant $C = C(\mathcal{E}_2(\mu_0))$ s.t. if*

$$\lambda_0 := \int_0^1 \lambda_{\min}(\mathbb{K}^1[\mu_0(.|s), \mathbf{x}_{\mu_0}(s)])\mathrm{d}s > 0 \quad \textit{and} \quad \mathcal{R}(\mu_0) < CN^{-3}\lambda_0^3\,, \tag{II.39}$$

then any gradient flow $(\mu_t)_{t\geq 0}$ starting from μ_0 satisfies:

$$\mathcal{R}(\mu_t) \leq \mathcal{R}(\mu_0)\exp(-\frac{C\lambda_0}{N}t)\,, \quad \textit{and} \quad \mu_t \xrightarrow{t\to\infty} \mu_\infty \in \mathcal{P}_2^{\mathrm{Leb}}([0,1]\times\Theta)\,.$$

Proof. Let C_1 be the universal constant appearing in Lemma II.5.1 and consider the radius $R = \min\left\{1, \frac{1}{2NC_1}\lambda_0 e^{-C_1(\sqrt{\mathcal{E}_2(\mu_0)}+1)^2}\right\}$. Then we have that for every $\mu \in B(\mu_0, R)$, $\mathcal{E}_2(\mu) \leq (\sqrt{\mathcal{E}_2(\mu_0)}+1)^2$ and hence by the local Lipschitz property of:

$$\int_0^1 \lambda_{\min}(\mathbb{K}^1[\mu, \mathbf{x}_\mu]) \geq \frac{\lambda_0}{2}\,.$$

Then, as a consequence of Eq. (II.15), we obtain that $\mathcal{R}$ satisfies the (R,m)-P-Ł property of Definition II.2 around μ_0 with $m = N^{-1}e^{-C_2}\lambda_0$ and $C_2 = C_2(\mathcal{E}_2(\mu_0))$ is a constant depending on μ_0. Combined with Theorem II.3 we obtain that the condition in Eq. (II.39) is sufficient for the gradient flow initialized at μ_0 to converge towards a global minimizer of the risk.

Note that by Lemma II.5.1, Eq. (II.15), and Assumption II.3 we can take the constant C in Eq. (II.39) to be of the form $C = C_3 e^{-C_3\mathcal{E}_2(\mu_0)}$ for some constant C_3. □

II.5.3 Examples of activations and quantitative convergence results

As one can see in the previous Theorem II.7, the better the conditioning of the kernel matrix, the better the constants in the local P-Ł property, and hence the easier it is to satisfy the condition for convergence. This conditioning depends on the choice of activation and initialization and it is important to keep in mind that the P-Ł property is not expected to hold around any initialization. For example, there is a saddle at every initialization μ_0 with feature distribution $\mu_0^2 = \delta_{(w,b)=0}$, whenever $\sigma(0) = \sigma'(0) = 0$. However, in general, the feature distribution μ^2 having dense support is a sufficient condition to ensure strict positivity. The following proposition is a direct consequence of [Sun, 2019, Thm.III.4] and [Carmeli, 2010, Cor.4.3].

Proposition II.5.1. *Assume σ has linear growth and is not a polynomial. Then if the feature distribution $\mu^2 \in \mathcal{P}_2(\mathbb{R}^d \times \mathbb{R})$ has dense support in $\mathbb{R}^d \times \mathbb{R}$, the kernel $k^1[\mu^2]$ is strictly positive.*

In the following, we provide examples of activations σ and initializations μ_0 for which the kernel matrix is well-conditioned. Moreover, in the case of the trigonometric activation function $\sigma = \cos$, quantitative lower bounds on the conditioning of the kernel matrix allow us to give quantitative conditions for convergence of the gradient flow.

Identity (or *FixUp*) initialization It will be particularly convenient to consider initial parameterization of the form $\mu_0 = \mathrm{Leb}([0,1]) \otimes \delta_0 \otimes \mu_0^2$ for some $\mu_0^2 \in \mathcal{P}_2(\mathbb{R}^d \times \mathbb{R})$, i.e. parameterization whose disintegration $\mu_0(.|s) = \delta_0 \otimes \mu_0^2$ is independent of $s \in [0,1]$ and has support in $\{0\} \times \mathbb{R}^d \times \mathbb{R}$. Such an initialization has been proposed for ResNets in [Zhang, 2018] and is shown to be associated with robust training and good generalization performances. Moreover, note that such an initialization is particularly natural for NODEs: in this case F_{μ_0} is identically 0 and the associated NODE flow is the identity. As a consequence the kernel matrix $\mathbb{K}^1[\mu_0]$ is independent of s and can be expressed as the block matrix:

$$\mathbb{K}^1[\mu_0] = \left(k^1[\mu_0](x^i, x^j)\right)_{1 \leq i,j \leq N} ,$$

only depending on the feature distribution μ_0^2 and on the input data distribution.

Positively homogeneous activation with uniform distribution of the features on the sphere The kernel $k^1[\mu]$ has been particularly studied in the case of a positively homogeneous activation σ [Cho, 2009; Bach, 2017b]. Motivated by applications in machine learning, a popular choice for such activation is the *Rectified Linear Unit* (ReLU):

$$\mathrm{ReLU} : x \mapsto \max\{x, 0\}$$

However, for $\sigma = \mathrm{ReLU}$, the associated basis function ψ would only satisfy Assumptions I.1 and I.2 and the only choice of positively homogeneous σ satisfying Assumption II.3 would be the trivial choice $\sigma = \mathrm{Id}$.

Nonetheless, whatever the choice of activation σ, Eq. (II.38) still defines a positive kernel k^1_μ over $\mathbb{R}^d$. Properties of this kernel in the case where σ is a positively homogeneous activation have been extensively investigated in the literature. In the case of $\sigma = \mathrm{ReLU}$ the previous Proposition II.5.1 can be improved thanks to the homogeneity of the activation:

Proposition II.5.2. *Assume $\sigma = \mathrm{ReLU}$. Then if the feature distribution $\mu^2 \in \mathcal{P}_2(\mathbb{R}^d \times \mathbb{R})$ has dense support in the sphere $\mathbb{S}^d$, the associated kernel $k^1[\mu^2]$ is strictly positive.*

Proof. The result is a direct application of [Sun, 2019, Prop.III.5] and [Carmeli, 2010, Cor.4.3]. □

Remark II.5.2. *In the case $\sigma = \mathrm{ReLU}^\alpha$ with some non-negative integer α, [Cho, 2009] provides an explicit computation of k^1 as a so-called* arc-cosine kernel *in the case $\mu^2 = \mathcal{U}(\mathbb{S}^d)$ is the uniform distribution on the sphere. Properties of these kernels and of the corresponding RKHSs have been studied in [Bach, 2017a]. It is for example shown that the induced RKHS is the Sobolev H^s of order $s = d/2 + \alpha + 1$.*

Trigonometric activation with strictly positive feature distribution An important case is also the choice of the trigonometric activation $\sigma = \cos$ for which, considering $\mu \in \mathcal{P}_2(\Theta)$, Eq. (II.2) gives:

$$\forall x \in \mathbb{R}^d , \quad F_\mu(x) = \int_{\mathbb{R}^d \times \mathbb{R}^d \times \mathbb{R}} u \cos(w^\top x + b) \mathrm{d}\mu(u, w, b) ,$$

and the definition $k^1[\mu]$ in Eq. (II.38) gives:

$$\forall x, y \in \mathbb{R}^d , \quad k^1[\mu](x, y) = \int_{\mathbb{R}^d \times \mathbb{R}} \cos(w^\top x + b) \cos(w^\top y + b) \mathrm{d}\mu^2(w, b) .$$

In the case where $\mu^2 = \mu^w \otimes \mathcal{U}([0,\pi])$ for some probability measure $\mu^w \in \mathcal{P}_2(\mathbb{R}^d)$ this last expression can be simplified into:

$$k^1[\mu](x,y) = \frac{1}{2}\int_{\mathbb{R}^d} \cos(w^\top(x-y))\mathrm{d}\mu^w(w)\,. \tag{II.40}$$

That is $k^1[\mu]$ is a positive translation-invariant kernel over $\mathbb{R}^d$ whose Fourier Transform is μ^w. It is a well-known theorem of Bochner (see [Wendland, 2004, Thm.6.6]) that having a non-negative Fourier Transform is a necessary and sufficient condition for a continuous function to define a positive translation-invariant kernel. Moreover, for some initial feature distributions, lower bounds on the conditioning of the kernel matrix as a function of the data separation are given in [Schaback, 1995].

Corollary II.5.1. *Let ψ be of the form Eq. (II.35) with activation $\sigma = \cos$. Assume the input data points $\{x^i\}_{1\leq i\leq N}$ are located in the ball $B(0,R)$ of radius $R>0$ and have separation $\delta \coloneqq \min_{i\neq j}\|x^i - x^j\| > 0$. Consider the initialization $\mu_0 = \mathrm{Leb}([0,1])\otimes\mu$ for some weight distribution $\mu \in \mathcal{P}_2(\Theta)$. Then the assumptions of Theorem II.7 are satisfied if:*

- *Sobolev / Matérn kernel $\mu = \delta_0 \otimes \mu^w \otimes \mathcal{U}([0,\pi])$ with $\mu^w(w) \propto (1+\|w\|^2)^{-\nu}$ for some $\nu > d/2+2$ and $\mathcal{R}(\mu_0) < CN^{-3}\delta^{6(\nu-d/2)}$, for some constant $C = C(R,\nu,d)$.*
- *Gaussian kernel $\mu = \delta_0 \otimes \mu^w \otimes \mathcal{U}([0,\pi])$ with $\mu^w(w) \propto \exp(-\frac{\|w\|^2}{2\rho^2})$ for some $\rho > 0$ and $\mathcal{R}(\mu_0) < CN^{-3}\delta^{-3d}e^{-C\delta^{-2}}$, for some constant $C = C(R,\rho,d)$.*
- *Random features: Finally assume $\mu_0 = \mathrm{Leb}([0,1])\otimes\hat{\mu}$ where $\hat{\mu} = M^{-1}\sum_{i=1}^M \delta_{(u_i,w_i,b_i)}$ and (u_i,w_i,b_i) are sampled i.i.d. from a distribution $\mu \in \mathcal{P}_2(\Theta)$ s.t. $\mathrm{Leb}([0,1])\otimes\mu$ satisfies the assumptions of Theorem II.7. Then for every $\varepsilon > 0$ there exists $M_\varepsilon \geq 0$ s.t. the assumptions of Theorem II.7 are satisfied with probability greater than $1-\varepsilon$ (over the sampling of $\{(u_i,w_i,b_i)\}_{1\leq i\leq M}$) whenever $M \geq M_\varepsilon$.*

Proof. This is a consequence of results on the conditioning of translation-invariant kernel of the form Eq. (II.40).

- Sobolev / Matérn kernel: using Eq. (II.40) the RKHS associated to $k^1[\mu]$ corresponds to the Sobolev space $H^\nu(\mathbb{R}^d)$ and [Schaback, 1995] gives that there exists a constant $C = C(\varepsilon,d)$ s.t.: $\lambda_{\min}(\mathbb{K}^1[\mu,\mathbf{x}]) \geq C^{-1}\delta^{2\nu-d}$.
- Gaussian kernel: using Eq. (II.40) the kernel $k^1[\mu]$ is the gaussian kernel given by $k^1[\mu](x,y) = \exp(-\frac{1}{2}\rho^2\|x-y\|^2)$ and [Schaback, 1995] gives that there exists a constant $C = C(\rho,d)$ s.t. $\lambda_{\min}(\mathbb{K}^1[\mu,\mathbf{x}]) \geq C^{-1}\delta^{-d}e^{-C\delta^{-2}}$.
- Random features: the assumptions of Theorem II.7 are satisfied with high probability when M tends to infinity as all the involved quantities in Eq. (II.39) are continuous w.r.t. the weight distribution $\mu \in \mathcal{P}_2(\Theta)$.

□

Note that, in order to obtain convergence in the above Corollary II.5.1, we assume the risk at initialization scales like $N^{-3}\delta^{3(2\nu-d)}$, which is the cube of the scaling required in Theorem II.6 when training only linear parameters. This bad scaling is a consequence of Lemma II.5.1, giving a worst case estimate of the conditioning of the tangent kernels during training if the feature distribution became degenerate. In contrast, one would expect training with gradient flow to lead to the learning of meaningful features, thus improving on the conditioning of the tangent kernels.

Remark II.5.3. *In this section we have leveraged the conditioning of the kernel matrix $\mathbb{K}^1$, that is the square norm of the gradient w.r.t. the outer weights u, to show a Polyak-Łojasiewicz inequality holds along the gradient flow. One might ask to what extent the kernel K^2 which takes into account the norm of the gradient w.r.t. the weights (w, b) might help improve on our convergence result. In fact, this kernel plays a negligible role in our analysis for the following reasons:*

We consider a "Fixup" initialization where the outer weights u are initialized to 0 at every layer. Initially proposed in [Zhang, 2018], this kind of initialization is shown to have favorable properties when training ResNets without normalization layer. Observing that K^2 is quadratic w.r.t. u, we have in this case that $K^2 = 0$ and $\partial_t K^2 = 0$ at $t = 0$. Thus, the kernel K^2 can only significantly improve the convergence result for large times in the gradient flow and cannot provide us with a good condition number at the beginning of the flow.

In addition, following the lines of Proposition 4.2, one could show the kernel matrix $\mathbb{K}^2$ (defined analogously as the kernel matrices $\mathbb{K}$ and $\mathbb{K}^1$) is locally Lipschitz w.r.t. μ with a Lipschitz constant scaling linearly with N, under additional mild hypotheses on the measure μ. Moreover, Theorem 4.2 ensures that during gradient flow the weight distribution will stay in a ball of radius $R \simeq \lambda_0/N$ around the weight distribution at initialization. Thus $\lambda_{\min}(\mathbb{K}^2[\mu])$ will be at most of order λ_0, which is the same order as $\lambda_{\min}(\mathbb{K}^1[\mu])$.

As a consequence of these two arguments, the local convergence result cannot be explained by the kernel K^2.

II.6 Ensuring convergence with lifting and scaling

The conditions derived in Sections II.4 and II.5 for convergence of the gradient flow notably asks for the loss at initialization to be sufficiently low, a condition which is difficult to check in practice. We conclude the present chapter by showing how this condition can always be enforced, that is how, for a given training dataset, one can modify the ResNet architecture in such a way that the convergence conditions are satisfied. The modification we propose is inspired by the work of Chizat, Oyallon, and Bach [Chizat, 2019] and consists in embedding the data in a higher dimensional space and performing a rescaling.

As before, we consider an empirical data distribution $\mathcal{D} = \frac{1}{N}\sum_{i=1}^N \delta_{x^i,y^i}$, with data $(x^i, y^i) \in \mathbb{R}^d \times \mathbb{R}^{d'}$. Consider also respectively the *embedding* and *projection* matrices:

$$A \coloneqq (\mathrm{Id}_d, 0_{d,d'})^\top \in \mathbb{R}^{(d+d')\times d}, \quad B \coloneqq (0_{d',d}, \mathrm{Id}_{d'}) \in \mathbb{R}^{d'\times(d+d')}.$$

Using the matrix A we embed the input variables $x^i \in \mathbb{R}^d$ in the space $\mathbb{R}^{d+d'}$ by defining $z^i \coloneqq Ax^i$. We then consider the NODE model of Definition I.1 with either:

- the linear parameterization of the residuals described in Section II.4, that is ψ of the form Eq. (II.18),
- residuals that are SHL perceptrons as described in Section II.5, that is ψ of the form Eq. (II.35).

For an input $z^i = Ax^i$ and a parameterization $\mu \in \mathcal{P}_2^{\mathrm{Leb}}([0,1] \times \Theta)$ we denote by z_μ^i the associated flow defined by Eq. (I.6). Also, for a scaling factor $\alpha > 0$, we consider the modified loss function ℓ^α defined by:

$$\forall (z, y) \in \mathbb{R}^{d+d'} \times \mathbb{R}^{d'}, \quad \ell^\alpha(z, y) \coloneqq \frac{1}{2}\|\alpha Bz - y\|^2.$$

We consider training the parameter $\mu \in \mathcal{P}_2^{\mathrm{Leb}}([0,1] \times \Theta)$ by performing gradient flow for the risk $\mathcal{R}^\alpha$ defined as:

$$\mathcal{R}^\alpha(\mu) \coloneqq \frac{1}{N} \sum_{i=1}^{N} \ell^\alpha(z_\mu^i(1), y^i) \,.$$

Note that, by construction, ℓ^α satisfies the P-Ł inequality $\|\nabla_x \ell^\alpha(x,y)\|^2 \geq 2\alpha^2 \ell(x,y)$. Thus, analogously to Eq. (II.15), we obtain the following P-Ł inequality for $\mathcal{R}^\alpha$:

$$\left|\nabla \mathcal{R}^\alpha\right|^2(\mu) \geq \frac{2\alpha^2 e^{-C}}{N} \left(\int_0^1 \lambda_{\min}(\mathbb{K}[\mu(.|s), \mathbf{z}_\mu(s)]) \mathrm{d}s \right) \mathcal{R}^\alpha(\mu) \,, \tag{II.41}$$

where $\mathbf{z}_\mu$ is the point cloud $(z_\mu^i)_{1 \leq i \leq N}$, the kernel matrix $\mathbb{K}$ is defined by Eq. (II.13) and $C = C(\mathcal{E}_2(\mu))$ is a constant depending on μ. Together with Theorem II.7, the above inequality implies that gradient flow converges towards a minimizer of the risk whenever α is sufficiently big.

Proposition II.6.1. *Assume one of the following condition if satisfied:*

- *In the case of a linear parameterization of the residuals, assume that the associated RKHS has a strictly positive kernel in the sense of Definition II.3. Moreover consider the initialization* $\mu_0 = \mathrm{Leb}([0,1]) \times \delta_0 \in \mathcal{P}_2^{\mathrm{Leb}}([0,1] \times \Theta)$.
- *In the case of SHL residuals, consider the initialization* $\mu_0 = \mathrm{Leb}([0,1]) \otimes \delta_0 \otimes \mu_0^2$ *for some* $\mu_0^2 \in \mathcal{P}_2(\mathbb{R}^{d+d'+1})$ *s.t.* $\lambda_0 \coloneqq \lambda_{\min}(\mathbb{K}^1[\mu_0, \mathbf{z}]) > 0$, *where* $\mathbb{K}^1$ *is defined in Eq. (II.38).*

Then there exists $\alpha_0 > 0$ *s.t. if* $\alpha > \alpha_0$ *then the gradient flow initialized at* μ_0 *converges towards a global minimizer of* $\mathcal{R}^\alpha$.

Proof. Using Lemma II.5.1 in the case of SHL residuals or Proposition II.4.4 in the case of RKHS residuals, we know a local P-Ł inequality is satisfied around μ_0. Then note that, as at initialization $\mathcal{R}^\alpha(\mu_0) = N^{-1} \sum_{i=1}^N \|y^i\|^2$ is independent of α and as increasing α increases the P-Ł in Eq. (II.41), the convergence condition in Eq. (II.10) is necessarily satisfied for α sufficiently large. □

II.7 Numerical results

We derived in this chapter theoretical results showing that deep ResNets or NODEs trained with gradient descent are able to interpolate the training dataset. The goal of this section is to verify those predictions numerically and to quantify how much our NODE models with RKHS and SHL residuals are able to generalize on unseen data. This is also useful to compare the performances of our models with those of standard ResNet architectures (which for example integrate batch normalization). We implemented our model in Pytorch [Paszke, 2017] and trained it on two image classification datasets, MNIST [LeCun, 2010] and CIFAR10 [Krizhevsky, 2009]. Source code is available at `https://github.com/rbarboni/FlowResNets`.

Classification task In the context of classification problem with K classes, the output dimension of the model is $d' = K$ and targets $y \in \mathbb{R}^K$ are one-hot vectors encoding the target classes. In both MNIST and CIFAR10, the number of classes is $K = 10$. We

consider evaluating the model using the *Cross Entropy* loss ℓ defined in Eq. (28). For a prediction z and a target one-hot vector label $y \in \mathbb{R}^K$ we have:

$$\ell(z, y) = \texttt{CrossEntropy}(z, y) \coloneqq -\log\left(\frac{\sum_{j=1}^K y_j e^{z_j}}{\sum_{j=1}^K e^{z_j}}\right).$$

Note that ℓ does not satisfy Assumption II.1, however, it does satisfy locally a modification of the Polyak-Łojasiewicz inequality. Indeed, assuming without loss of generality that $y = e_1$ is the indicator of class 1, then $\nabla_{z_1}\ell(z, y) = 1 - e^{-\ell(z,y)}$, leading to

$$\|\nabla_z \ell(z, y)\|^2 \geq (1 - e^{-\ell(z,y)})^2 \geq \left(\frac{1 - e^{-\ell_0}}{\ell_0}\right)^2 \ell(z, y)^2,$$

when $\ell(z, y) \leq \ell_0$.

Training As in Section II.3, our training dataset are constituted of a finite (large) number N of data samples. Then for a predictor $F : \mathbb{R}^d \to \mathbb{R}^C$ with parameters $\theta \in \Theta$ the empirical risk reads:

$$\mathcal{R}(\theta) \coloneqq \frac{1}{N}\sum_{i=1}^N \ell(F_\theta(x^i), y^i).$$

We consider training NODE models with *Stochastic Gradient Descent (SGD)* for the minimization of this training risk. Note that while the convergence results in Sections II.4 and II.5 only apply for full batch gradient descent, several similar convergence results under the P-Ł assumption hold for stochastic optimization methods [Karimi, 2016]. Finally, the performance of the models are assessed by the *Top-1 error* on a set of test data.

II.7.1 Experiments on MNIST

We implemented the NODE model in Definition I.1 with RKHS and SHL residuals on Pytorch using the `torchdiffeq` package [Chen, 2018] and performed experiments on the MNIST dataset [LeCun, 2010].

Implementation We implement the NODE model in Definition I.1 with residuals that are 2-layer convolutional neural networks. This corresponds to a modification of the residuals originally considered in [He, 2016a] where the final nonlinearity and batch normalizations are removed.

Given a depth $D \geq 1$ the trained parameters consist of convolution matrices $W_d \in \mathbb{R}^{C \times C_{int} \times 3 \times 3}$ and $U_d \in \mathbb{R}^{C_{int} \times C \times 3 \times 3}$ for $d \in \{0, ..., D\}$, with C the number of channels of the input image and C_{int} some number of channels for the hidden layers. The residuals are then defined at discrete time steps $\{d/D\}_{0 \leq d \leq D}$ by:

$$F_{d/D}(x) \coloneqq W_d \star \mathrm{ReLU}(U_d \star x),$$

where $x \in \mathbb{R}^{C \times n_w \times n_h}$ is the input signal and $\star$ is the discrete convolution operator defined in Eq. (32). When the inner convolutional filters U_k are fixed, this corresponds to the RKHS residuals considered in Section II.4. On the opposite, when the U_k are learned, this is similar to the SHL residuals considered in Section II.5. Then, for any $s \in [0, 1]$, the residual at time s is defined by affine interpolation. For an input signal $x \in \mathbb{R}^{C \times n \times n}$:

$$F_s(x) \coloneqq F_{d/D}(x) + (tD - d)\left(F_{(d+1)/D}(x) - F_{d/D}(x)\right),$$

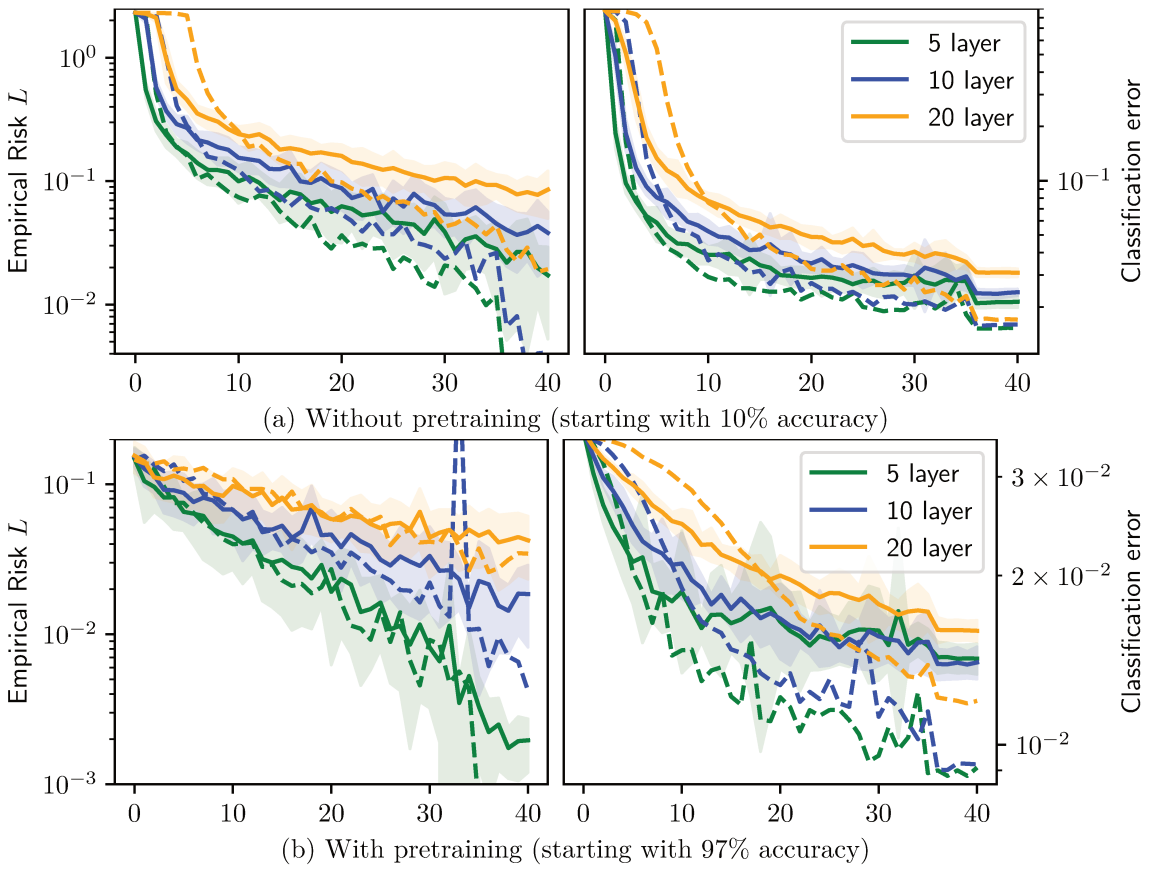
with $k = \lfloor sD \rfloor$. The forward pass through the network consists in integrating the ODE in Eq. (I.5) with the residuals $F = (F_s)_{s\in[0,1]}$ using the `torchdiffeq.odeint` method from Chen et al. [Chen, 2018]. For an input signal $x \in \mathbb{R}^{C\times n\times n}$ the output of the NODE is given by:

$$\mathrm{NODE}_{(W_d,U_d)_{1\leq d\leq D}} \coloneqq \texttt{torchdiffeq.odeint}(F, x, [0,1])\,.$$

Hyperparameter tuning. Several hyperparameters can affect the training.

- The convolution matrices U_d: as detailed in Section II.4, the way the weights U_d are sampled determines the RKHS of residuals and has thus a significant impact on training. For the sake of simplicity we choose to sample the coefficients of U_d as i.i.d. Gaussians.
- The initialization of (W_k): the weights of the convolution matrices W_k are initialized to 0. This is a standard choice when considering NODEs without normalization layers [Zhang, 2018].
- The integration method: `torchdiffeq.odeint` allows the user to choose an integration method. We observed an *explicit midpoint* method to offer a good trade-off between performance and numerical stability w.r.t. other fixed-steps methods such as *explicit Euler* or *RK4*.
- The number of layers D: we tested our model for $D \in \{5, 10, 20\}$. This parameter controls the total number of parameters of the model.
- Pre- and postprocessing: We consider pre- and postprocessing the signal with small neural networks `A` and `B` respectively. While MNIST is composed of gray-scale images of size 28×28 with 1 channel, the purpose of this is to downsample the image while adjusting the number of channels $C \geq 1$. As explained in Section II.6, rising the number of channels is expected to ease the training problem. To isolate the effect of training the NODE, both `A` and `B` are fixed during training but we consider different level of pretraining of the concatenation $\texttt{B} \circ \texttt{A}$, corresponding to the NODE when initialiazed with $W_d = 0$. In any case, we see that training the NODE improves on the performance of the simple concatenation $\texttt{B} \circ \texttt{A}$.

Results. Fig. II.1 shows the evolution of the performances of the NODEs with both RKHS and SHL residuals while trained on the MNIST dataset. One can observe in both cases that the training risk converges to 0 at a linear rate, supporting the results of Section II.4 and Section II.5. Decay of the risk is also directly related to the decay of the classification error showing the NODE models exhibit generalization abilities. Without pretraining of `A` and `B` (Fig. II.1a), the models start with random guesses (10% accuracy) and achieve up to 98% accuracy on the test set for RKHS residuals and up-to 99.5% accuracy for SHL residuals. When `A` and `B` are pretrained (Fig. II.1b), the NODE still improves on the starting accuracy: in this setting more than 99% accuracy is reached for both RKHS and SHL residuals. While there is a difference between the performance of NODEs with RKHS and SHL residuals one can thus observe here that it is not significant when inner layers of the residuals are sampled appropriately. One can see the effect of varying the depth D of the model and observe that deeper model seem to have poorer performances. We explain this by the fact that deeper models are harder to train.



(a) Without pretraining (starting with 10% accuracy)

(b) With pretraining (starting with 97% accuracy)

Figure II.1: Performances of NODE with 32 channels while trained on MNIST with SGD. Left column reports evolution of the empirical risk and right column reports evolution of classification error, both for ResNets with RKHS residuals (plain) and SHL residuals (dashed). The x-axis is the number of pass through the dataset. Experiments are performed with different levels of pretraining of `A` and `B`, corresponding to different starting accuracy ((a)-(b)), and with different number of layers. Learning rate and batch size are fixed, learning rate is divided by 10 after 35 iterations. Plots are average over 20 runs, lines are means and, for RKHS residuals, colored areas are mean $\pm$ one standard deviation.

II.7.2 Experiments on CIFAR10

Implementation. For experiments on the CIFAR10 dataset, we did not rely on numerical integration of the NODE using `torchdiffeq` but instead use a ResNet architecture inspired from ResNet18 [He, 2016a].

As before, residual blocks are simplified by removing the final non-linearity and the batch-normalization. For an input image $x \in \mathbb{R}^{C\times n\times n}$ the output of a residual is:

$$F_d(x) = W_d \star \mathrm{ReLU}(U_d \star x),$$

where $U_d \in \mathbb{R}^{C_{int}\times C\times 3\times 3}$, $W_d \in \mathbb{R}^{C\times C_{int}\times 3\times 3}$ are convolution matrices, C is the number of channels of the input image and C_{int} is the number of channels of the hidden layer. Also, while ResNet18 consists of 4 blocks each containing 2 residual layers, we keep 2 of our residuals in the first, second and fourth block but stack an arbitrary number D of residual layers in the third block. Thereby, we refer to this third block as the NODE block.

Initialization The weights of the convolutional filters W_d are initialized at 0, corrresponding to the initialization proposed in [Zhang, 2018]. Also, the weights of the convolutional filters U_d are initialized as i.i.d. Gaussians and rescaled by a $C_{int}^{-1/2}$ factor.

Results. Fig. II.2 reports the training of our ResNet model on the CIFAR10 dataset. Fig. II.2a reports evolution of the training risk and classification error when inner weights U_d are fixed (RKHS residuals) and is to be compared with Fig. II.2b, showing the same quantities when hidden weights are learned (SHL residuals). In particular, one can observe that the training risk is reduced to nearly 0 at the end of training with SGD, as predicted in Sections II.4 and II.5 for gradient descent. This reduction of the training risk goes with an augmentation of the accuracy. Our experiments show that similar performances can be achieved with RKHS or SHL residuals: both ResNets achieve up to 88% accuracy on the test dataset. As a comparison, the ResNet18 original architecture can be trained to achieve up to 94% accuracy in a similar setting.

Finally, Fig. II.2 also compares the performances of the model depending on the number of layers inside the NODE block. One observes significantly different behavior when there is no NODE (1 layer) and one there is (10 and 20 layers): more layers are related to better performances both on the train dataset and on the test dataset and both when hidden layers are trained or not. However, one sees that the improvement related to adding more layers is limited: performances with 10 and 20 layers are very similar and a NODE block with 1 layer already achieves 82% accuracy with RKHS residuals and 84% accuracy with SHL residuals. This hints towards the fact that our discrete ResNet model indeed converges towards a NODE when the depth increases.

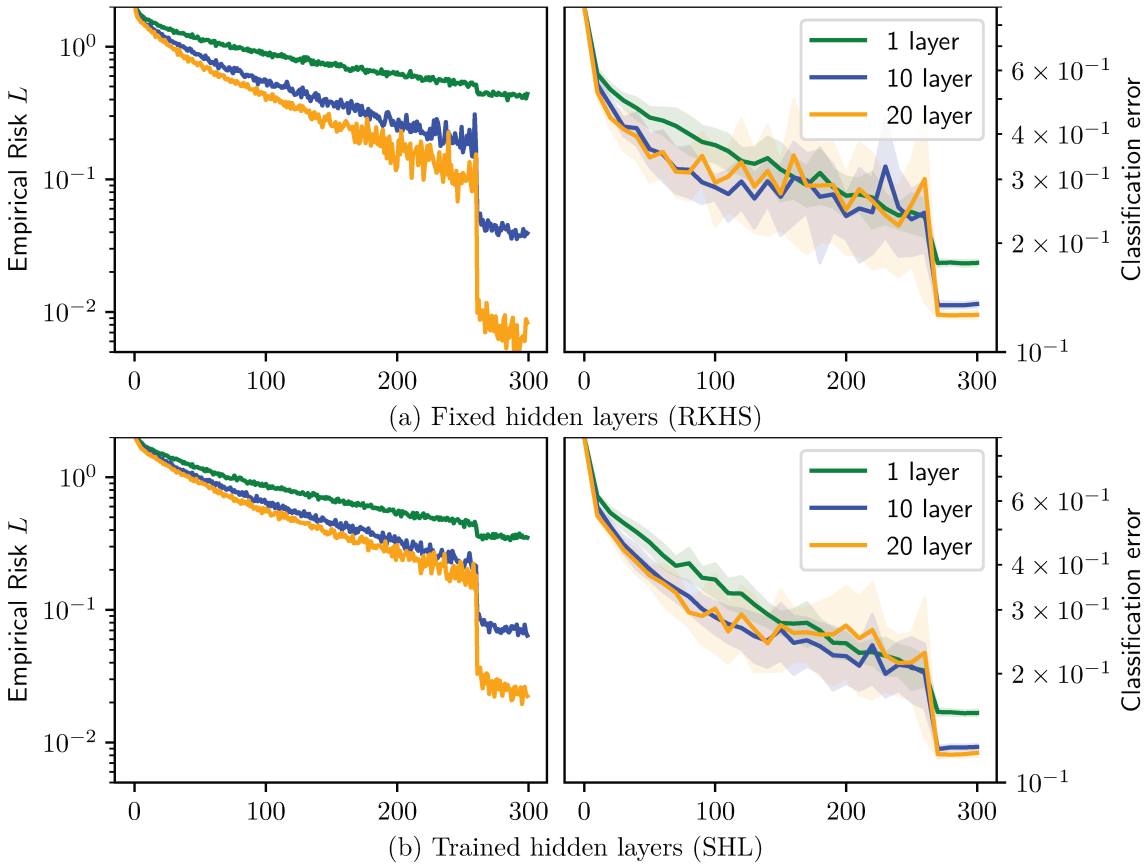


Figure II.2: Performances of ResNets while trained on CIFAR10 with SGD (256 images per batch) and trained (a) or fixed (b) hidden layers. Left column reports evolution of the empirical risk on the train set and right column reports the classification error on the test set. The x-axis is the number of pass through the dataset. Learning rate and batch size are fixed, learning rate is divided by 10 after 260 iterations. Plots are average over 20 runs, lines are means and colored areas are mean $\pm$ one standard deviation.

II.8 Conclusion

Relying on the mathematical framework previously developed in Chapter I, we showed in this chapter that convergence of mean-field models of ResNets can be proved by using a Polyak-Łojasiewicz inequality. This inequality is satisfied locally around well-chosen initializations for which the residuals have sufficiently (but possibly finitely) many features, ensuring their expressivity. As a consequence, assuming the risk is already sufficiently small at those initializations, the gradient flow provably converges towards a global minimizer of the risk with a linear convergence rate. For practical examples of architectures — such as random feature models [Rahimi, 2007] or SHL perceptrons — and parameter initializations, we also quantified explicitly the convergence condition as a function of the number of data points.

This is the first convergence result of this type for mean-field models of ResNets with unregularized risk as previous works only showed results of optimality under the assumption of convergence. Moreover, we showed through numerical experiments that deep ResNets or NODEs trained with gradient descent are indeed amenable to zero training risk while still being able to generalize on test data.

We point out some limitations and possible extensions of these results:

- We make regularity assumptions on the basis function ψ that might be improved on. In particular, Assumption I.3 assumes ψ to be at least continuously differentiable which does not allow us to consider SHL residuals with ReLU activations. This assumption might be weakened, for example by using the recently introduced notion of *conservative gradient* [Bolte, 2021].

- We only considered in our convergence analysis the case of an empirical data distribution $\mathcal{D} = \frac{1}{N}\sum_{i=1}^{N} \delta_{(x^i,y^i)}$. This assumption is crucial as the P-Ł constant in Eq. (II.15) scales as N^{-1} and become degenerate for large N. It would therefore be interesting to extend our analysis to the case of a data distribution with density.

- An important aspect of our convergence analysis is to only leverage information about gradients w.r.t. the outer weights of the residuals (denoted by u) to obtain the Polyak-Łojasiewicz inequality. In doing so, we are unable to provide information about the behavior of the feature distribution during training and unable to ensure that gradient flow will escape the "kernel regime".

  Quantifying in what extent feature learning helps the training of neural networks is indeed an active area of research. In this direction, we will perform in Chapter III a detailed analysis of the evolution of the feature distribution during the training of simpler shallow architectures.

Chapter III

Feature learning in shallow architectures: a study of two-timescale learning algorithms

Contents

III.1 Introduction **119**
III.1.1 Mean-field neural networks and two-timescale learning 120
III.1.2 Contributions and related works 125
III.2 Reduced risk associated to the VarPro algorithm **127**
III.2.1 Primal formulation of the reduced risk 128
III.2.2 Partial minimization on the space of measures 129
III.2.3 Dual formulation of the reduced risk 130
III.2.4 Kernel learning in the case of quadratic regularization 132
III.3 Properties of minimizers of the reduced risk **132**
III.3.1 Existence and uniqueness of minimizers 132
III.3.2 Convergence of minimizers 134
III.4 Training with gradient flow **135**
III.4.1 Wasserstein gradient flows in the case $\lambda > 0$ 136
III.4.2 Wasserstein gradient flows in the case $\lambda = 0$ and ultra-fast diffusions 140
III.5 Convergence of gradient flow **143**
III.5.1 Algebraic convergence rate 143
III.5.2 Convergence to ultra-fast diffusion. 145
III.6 Numerics **148**
III.6.1 Single-hidden-layer neural networks with 1-dimensional feature space 148
III.6.2 VarPro for image classification on CIFAR10 155
III.7 Conclusion **159**
Appendices **160**
III.A Positive definite kernels and RKHS **160**
III.B Radial basis function neural network on the 2-dimensional torus **162**

III.1 Introduction

Machine learning methods based on *artificial neural networks* have recently experienced a significant increase in popularity due to their efficiency in solving numerous supervised or unsupervised learning tasks. This success owes to their capacity to perform *feature*

learning, that is to extract meaningful representations from the data during the training process [Goodfellow, 2016, Chap. 15], standing in contrast with *kernel methods* for which feature representations are designed by hand and fixed during training [Hofmann, 2008]. Indeed, we have observed in Sections II.4 and II.5 that the learning of appropriate feature representations at each layer plays a fundamental role in the training of deep architectures. Moreover, feature learning is also believed to play an important role in the generalization performance of neural networks. For example, adaptivity to low-dimensional representations of the data can prevent the *curse of dimensionality* [Bach, 2017a; Ghorbani, 2020].

However, the process through which features are learned remains largely misunderstood. Indeed, adaptivity of the representations comes in neural networks at the price of a nonlinear parameterization, making the training dynamic more difficult to analyze. Specifically, for *overparameterized* neural network architectures where the dimension of the parameter space greatly exceeds the number of training samples, recent works have put forward the crucial role played by the choice of scaling w.r.t. the number of parameters in the training dynamic [Chizat, 2019; Liu, 2020; Yang, 2021]. For single-hidden-layer neural networks, the "kernel regime", corresponding to a scaling of $1/\sqrt{M}$ where M is the width, has been identified as a scaling for which the model is well-approximated by its linearization around initialization, therefore reducing to a kernel method [Jacot, 2018]. Relying on the good conditioning of the "Neural Tangent Kernel (NTK)" (Eq. (38)), this regime provides convergence of gradient descent towards a global minimizer of the risk at a linear rate [Allen-Zhu, 2019; Du, 2019; Lee, 2019; Zou, 2020]. However, this regime has also been shown to suffer from a "lazy training" behavior preventing significant modification of the feature distribution and associated to poor generalization guarantees [Chizat, 2019].

In contrast, another line of work has been focused on the "mean-field" regime (Eq. (39)) corresponding to a scaling of $1/M$ for which the neural network is parameterized by a probability distribution over the space of weights [Chizat, 2018; Mei, 2019; Rotskoff, 2019; Sirignano, 2020]. While such a choice of scaling has been shown to enable nonlinear feature learning behaviors [Yang, 2021], existing convergence results are primarily qualitative, lacking explicit convergence rates. To bridge this gap, we are interested in this chapter in the dynamic of the feature distribution in the training of mean-field models of shallow neural network architectures. We study more particularly a *variable projection* or *two-timescale* learning strategy which allows reducing the learning problem to the training of the feature distribution.

III.1.1 Mean-field neural networks and two-timescale learning

We consider in this chapter shallow neural networks with a parameter space that decomposes as $\Theta = \mathbb{R} \times \Omega$ where $\mathbb{R}$ is the space of linear parameters and Ω is the space of nonlinear parameters of the model. In the following, we will assume Ω to be either the n-dimensional torus $\mathbb{T}^n = \mathbb{R}^n/\mathbb{Z}^n$, or a closed, bounded and convex domain of $\mathbb{R}^n$. Following Eq. (36), such shallow neural networks can be expressed as a sum of basis functions of the form:

$$\forall (u, \omega) \in \mathbb{R} \times \Omega, \ \forall x \in \mathbb{R}^d, \quad \psi((u, \omega), x) = u\phi(\omega, x)\,,$$

where $\phi : \Omega \times \mathbb{R}^d \to \mathbb{R}$, is some *feature map*. For an integer $M \geq 1$, the obtained *single-hidden-layer (SHL)* neural network of width M with *inner weights* $\{\omega_i\}_{1 \leq i \leq M} \in \Omega^M$ and *outer weights* $\{u_i\}_{1 \leq i \leq M} \in \mathbb{R}^M$ is the map:

$$F_{\{(\omega_i, u_i)\}} : x \in \mathbb{R}^d \mapsto \frac{1}{M} \sum_{i=1}^{M} u_i \phi(\omega_i, x) \in \mathbb{R}\,, \tag{III.1}$$

taking inputs in the *input space* $\mathbb{R}^d$ and returning values in the *output space* $\mathbb{R}$. Following Eq. (39), using the interchangeability of the indices and the normalisation factor $1/M$, the above model can then be reparameterized in terms of the empirical distribution of the inner weights $\{\omega_i\}_{1\leq i\leq M}$. Given an arbitrary probability distribution $\mu \in \mathcal{P}(\Omega)$ on the space of inner weights and a measurable map $u \in L^1(\mu)$ we define:

$$F_{\mu,u} : x \in \mathbb{R}^d \mapsto \int_\Omega u(\omega)\phi(\omega, x)\mathrm{d}\mu(\omega) \in \mathbb{R}\,. \tag{III.2}$$

In particular, for the empirical distribution $\hat{\mu} = \frac{1}{M}\sum_{i=1}^M \delta_{\omega_i}$ and outer weights $\hat{u}(\omega_i) = u_i$ we recover the finite width SHL $F_{\hat{\mu},\hat{u}} = F_{\{(\omega_i,u_i)\}}$. Such a "mean-field" model of neural network has been proposed by several authors to study the training of neural networks at arbitrary large width [Chizat, 2018; Mei, 2019; Rotskoff, 2019; Sirignano, 2020].

Supervised learning As in Chapters I and II we consider a supervised learning framework where training a neural network consists in minimizing a *training risk* associated to the evaluation of the model on some *training data*. Precisely, we consider in this chapter a univariate regression setting where the neural network weights are trained for minimizing the mean square error with a *target signal* $Y \in L^2(\rho)$ evaluated on training data with distribution $\rho \in \mathcal{P}(\mathbb{R}^d)$. However, in contrast with Chapters I and II we add here a supplementary regularization term on the linear parameters of the model.

For a regularization strength $\lambda > 0$ and $\mu \in \mathcal{P}(\Omega)$, $u \in L^1(\mu)$ we define the training risk as:

$$\mathcal{R}^\lambda(\mu, u) \coloneqq \frac{1}{2}\|F_{\mu,u} - Y\|^2_{L^2(\rho)} + \lambda\|u\|^2_{L^2(\mu)}\,, \tag{III.3}$$

where we assume $\mathcal{R}^\lambda(\mu, u) = +\infty$ if $u \notin L^2(\mu)$. Training the neural network then amounts to finding parameters $(\mu, u) \in \arg\min \mathcal{R}^\lambda$.

Example of applications Note that the mean-field neural network model of Eq. (III.2) can be seen as a linear model acting on (signed) measures. Indeed, for $\mu \in \mathcal{P}(\Omega)$ and $u \in L^1(\mu)$, we have $F_{\mu,u} = \Phi \star (u\mu)$ where for every finite Borel measure $\nu \in \mathcal{M}(\Omega)$ we define:

$$\Phi \star \nu \coloneqq \int_\Omega \phi(\omega, .)\mathrm{d}\nu(\omega)\,. \tag{III.4}$$

This structural property is in strong contrast with the ODE based models considered in Chapters I and II and will be crucial to our analysis in this chapter. Also, minimization of functionals of the form in Eq. (III.3) with linear models acting on the space of measures have numerous applications depending on the choice of the feature map ϕ.

- Two-layer perceptron: The perceptron model defined in Eq. (34) is arguably the prototypical example of a neural network. It consists here in considering a parameter space $\Omega \subset \mathbb{R}^{d+1}$ and a feature map $\phi : (\omega, x) \mapsto \sigma(\omega^\top \bar{x})$ where $\bar{x} = (x, 1) \in \mathbb{R}^{d+1}$ and $\sigma : \mathbb{R} \to \mathbb{R}$ is some nonlinear activation function such as the *Rectified Linear Unit (ReLU)* or hyperbolic tangent. Owing to their great expressivity [Cybenko, 1989], this class of models is ubiquitous in applications where an unknown signal is to be recovered from data observations.

- Radial Basis Function (RBF) neural networks and signal deconvolution: RBF neural networks [Pereyra, 2006; Karamichailidou, 2024] is an example of a simple architecture in which the feature map consists of a translation invariant kernel k i.e.

$\Omega \subset \mathbb{R}^d$ and $\phi : (\omega, x) \mapsto k(\omega - x)$. The network $F_{\mu,u}$ then implements a convolution with the kernel k and minimization of the risk $\mathcal{R}^\lambda$ amounts to solve a form of deconvolution problem. This has important applications in signal processing where one wants to recover an unknown signal given noisy or filtered observations [De Castro, 2012; Duval, 2015].

Training with gradient descent and two-timescale learning In supervised learning, minimization of the training risk is usually performed using first order optimization methods such as gradient descent or stochastic variants on the neural network's weights [Bottou, 2018].

We consider here the two-timescale version of gradient descent described in Eq. (47). For a SHL of finite width $M \geq 1$ with weights $\{(\omega_i, u_i)\}_{1 \leq i \leq M} \in (\Omega \times \mathbb{R})^M$ the associated risk is $\hat{\mathcal{R}}^\lambda(\{(\omega_i, u_i)\}_{1 \leq i \leq M}) \coloneqq \mathcal{R}^\lambda(\hat{\mu}, \hat{u})$, where $\hat{\mu} = \frac{1}{M}\sum_{i=1}^M \delta_{\omega_i}$ and $\hat{u}(\omega_i) = u_i$. For an initialization $\{(\omega_i^0, u_i^0)\}_{1 \leq i \leq M}$, a step-size $\tau > 0$ and a timescale parameter $\eta > 0$, the *two-timescale gradient descent* dynamic reads:

$$\forall k \geq 0,\ \forall i \in \{1, ..., M\}, \quad \left\{ \begin{array}{rcl} \omega_i^{k+1} & = & \omega_i^k - M\tau \nabla_{\omega_i} \hat{\mathcal{R}}^\lambda(\{(\omega_i^k, u_i^k)\}_{1 \leq i \leq M}) \,, \\ u_i^{k+1} & = & u_i^k - \eta M \tau \nabla_{u_i} \hat{\mathcal{R}}^\lambda(\{(\omega_i^k, u_i^k)\}_{1 \leq i \leq M}) \,. \end{array} \right. \tag{III.5}$$

For the purpose of theoretical analysis we study here the limit of the gradient descent algorithm when the step-size τ tends to 0. For an initialization $\{(\omega_i(0), u_i(0))\}_{1 \leq i \leq M}$, this *gradient flow* dynamic reads:

$$\forall i \in \{1, ..., M\}, \quad \left\{ \begin{array}{rcl} \frac{\mathrm{d}}{\mathrm{d}t}\omega_i(t) & = & -M \nabla_{\omega_i} \hat{\mathcal{R}}^\lambda(\{(\omega_i(t), u_i(t))\}_{1 \leq i \leq M}) \,, \\ \frac{\mathrm{d}}{\mathrm{d}t}u_i(t) & = & -\eta M \nabla_{u_i} \hat{\mathcal{R}}^\lambda(\{(\omega_i(t), u_i(t))\}_{1 \leq i \leq M}) \,. \end{array} \right. \tag{III.6}$$

Note the role of the *timescale parameter* $\eta > 0$ controlling the ratio of learning timescales between inner and outer weights. When $\eta < 1$ the outer-weights u_i are learned more "slowly" than the inner-weights ω_i and conversely, when $\eta > 1$ the outer-weights u_i are learned more "quickly" than the inner-weights ω_i. In particular, the limiting training dynamics when $\eta \to +\infty$ correspond (formally) to the case where the outer weights are learned "instantaneously", that is, at each time $t \geq 0$, we have

$$\{u_i(t)\}_{1 \leq i \leq M} \in \operatorname*{arg\,min}_{u \in \mathbb{R}^M} \hat{\mathcal{R}}^\lambda(\{(\omega_i(t), u_i)\}_{1 \leq i \leq M}) \,.$$

Such limiting dynamics correspond to the variable projection algorithm described in Eq. (48).

Variable Projection The *Variable Projection (VarPro)* algorithm performs elimination of the linear variable u and enables here reducing the training of a neural network to the sole problem of learning the feature distribution. Introduced in [Golub, 1973] for the minimization of separable nonlinear least squares problems, such a strategy has proven to be efficient in various applications [Golub, 2003; Osborne, 2007] including the training of neural networks [Sjoberg, 1997; Pereyra, 2006; Newman, 2021; Karamichailidou, 2024]. A reason for this popularity is that partial optimization over one variable can lead to a better conditioning of the Hessian [Sjoberg, 1997; Vialard, 2019].

Exploiting here the linearity w.r.t. the outer weights in the definition of F, it is convenient to read a neural network's output $F_{\{(\omega_i, u_i)\}}(x) = \frac{1}{M}\sum u_i \phi(\omega_i, x)$ as a linear combination of the *features* $\{\phi(\omega_i, x)\}_{i=1}^M$. From this point of view, neural networks should be compared to *kernel methods* for which the features are built in advance and fixed during

training, whereas only the weights of the linear combination are learned [Hofmann, 2008]. In contrast, both inner weights $\{\omega_i\}_{i=1}^M$ and outer weights $\{u_i\}_{i=1}^M$ of a neural network are usually trained. In the following, we refer to the parameters $\omega \in \Omega$ as the neural network's *features* and to $\mu \in \mathcal{P}(\Omega)$ as the *feature distribution*. More generally in the mean-field limit, for $\mu \in \mathcal{P}(\Omega)$ and $u \in L^1(\mu)$, we have:

$$F_{\mu,u} = \int_\Omega \phi(\omega, .)u(\omega)\mathrm{d}\mu(\omega) = \Phi_\mu \cdot u\,, \tag{III.7}$$

where we introduced the *feature operator* $\Phi_\mu : u \in L^1(\mu) \mapsto \int_\Omega u(\omega)\phi(\omega, .)\mathrm{d}\mu(\omega) \in L^2(\rho)$. One can thus notice that the problem of minimizing the risk $\mathcal{R}^\lambda$ belongs to the class of *separable nonlinear least squares problems* as, by definition, for a fixed inner weights distribution $\mu \in \mathcal{P}(\Omega)$:

$$\mathcal{R}^\lambda(\mu, u) = \frac{1}{2}\|\Phi_\mu \cdot u - Y\|^2_{L^2(\rho)} + \lambda\|u\|^2_{L^2(\mu)}\,.$$

Thus the problem of minimizing $\mathcal{R}^\lambda$ w.r.t. u is a *ridge regression problem* which can be efficiently numerically solved by inverting a linear system. For $\lambda > 0$, there exists a unique solution $u^\lambda[\mu] \in \arg\min_{u\in L^2(\mu)} \mathcal{R}^\lambda(\mu, u)$ given by $u^\lambda[\mu] \coloneqq (\Phi_\mu^\top\Phi_\mu + 2\lambda)^{-1}\Phi_\mu^\top Y$. Plugging this in $\mathcal{R}^\lambda$ gives rise to a *reduced risk* which we define for any $\mu \in \mathcal{P}(\Omega)$ by:

$$\mathcal{L}^\lambda(\mu) \coloneqq \frac{1}{\lambda}\mathcal{R}^\lambda(\mu, u^\lambda[\mu]) = \min_{u\in L^2(\mu)} \frac{1}{2\lambda}\|\Phi_\mu \cdot u - Y\|^2_{L^2(\rho)} + \|u\|^2_{L^2(\mu)}\,. \tag{III.8}$$

This definition also extends to the limiting case $\lambda \to 0^+$ by considering:

$$\mathcal{L}^0(\mu) \coloneqq \min_{\Phi_\mu\cdot u=Y} \|u\|^2_{L^2(\mu)}\,. \tag{III.9}$$

where the infimum is taken to be $+\infty$ whenever the signal Y is not in the range of Φ_μ. In the case where $Y \in \mathrm{Range}(\Phi_\mu)$, this minimization problem admits a unique solution and $\mathcal{L}^0(\mu) = \|u^\dagger[\mu]\|^2_{L^2(\mu)}$, where $u^\dagger[\mu] = \Phi_\mu^\dagger \cdot Y$ and $\Phi_\mu^\dagger$ is the generalized pseudo-inverse of Φ_μ restricted to $L^2(\mu)$.

The *VarPro algorithm* consists here in performing *gradient descent* over the reduced risk $\mathcal{L}^\lambda$. For a neural network of finite width $M \geq 1$ with features $\{\omega_i\}_{1\leq i\leq M} \in \Omega^M$, the associated reduced risk is $\hat{\mathcal{L}}^\lambda(\{\omega_i\}_{1\leq i\leq M}) \coloneqq \mathcal{L}^\lambda(\hat{\mu})$, where $\hat{\mu}$ is the empirical distribution $\hat{\mu} = \frac{1}{M}\sum_{i=1}^M \delta_{\omega_i}$. For an initialization $\{\omega_i^0\}_{1\leq i\leq M} \in \Omega^M$ and a step-size $\tau > 0$, the VarPro dynamic reads:

$$\forall k \geq 0,\ \forall i \in \{1, ..., M\}, \quad \omega_i^{k+1} = \omega_i^k - M\tau\nabla_{\omega_i}\hat{\mathcal{L}}^\lambda(\{\omega_i^k\}_{1\leq i\leq M})\,.$$

As before, the *gradient flow* of $\hat{\mathcal{L}}_f^\lambda$ is the continuous counterpart of gradient descent when the step-size τ tends to 0. For an initialization $\{\omega_i(0)\}_{1\leq i\leq M} \in \Omega^M$, it is defined for every time $t \geq 0$ as the solution $\{\omega_i(t)\}_{1\leq i\leq M} \in \Omega^M$ to the ODE:

$$\forall i \in \{1, ..., M\}, \quad \frac{\mathrm{d}}{\mathrm{d}t}\omega_i(t) = -M\nabla_{\omega_i}\hat{\mathcal{L}}^\lambda(\{\omega_i(t)\}_{1\leq i\leq M})\,. \tag{III.10}$$

Note that the above gradient can be efficiently calculated numerically once optimization on the outer weights u_i has been performed, for example by means of standard automatic differentiation libraries. Indeed, if $\{u_i(t)\}_{1\leq i\leq M} \in \arg\min_{u\in\mathbb{R}^M} \hat{\mathcal{R}}^\lambda(\{(\omega_i(t), u_i)\}_{1\leq i\leq M})$, then by the envelope theorem $\nabla_{\omega_i}\hat{\mathcal{R}}^\lambda(\{(\omega_i(t), u_i(t))\}_{1\leq i\leq M}) = \lambda\nabla_{\omega_i}\hat{\mathcal{L}}^\lambda(\{\omega_i(t)\}_{1\leq i\leq M})$. For the same reason, the above dynamic can be seen, at least formally, as the limit of the gradient flow dynamic Eq. (III.6) over the (unreduced) risk $\hat{\mathcal{R}}^\lambda$ when the timescale parameter η tends to $+\infty$. Thus, we equivalently refer to Eq. (III.10) as the *VarPro gradient flow* or as the *two-timescale regime of gradient flow*.

Wasserstein gradient flows and ultra-fast diffusions Relying on the mathematical framework provided by theory of gradient flows in the space of probability measures [Ambrosio, 2008b; Santambrogio, 2017], we show in Section III.4 that the dynamic of the feature distribution when trained with gradient flow for the minimization of the reduced risk $\mathcal{L}^\lambda$ is solution to an advection PDE of the form:

$$\partial \mu_t - \operatorname{div}(\mu_t \nabla \mathcal{L}^\lambda[\mu_t]) = 0\,,$$

for some nonlinear velocity field $\nabla \mathcal{L}_f^\lambda[\mu_t]$. We study in Section III.5 the asymptotics of this equation when the training time t tends to $+\infty$ and the regularization strength λ tends to 0^+. We are more particularly interested in the case where the signal Y itself can be exactly represented by a neural network. We consider the following assumption:

Assumption III.1 (Teacher student setup)**.**
Let $\Phi\star$ be defined by Eq. (III.4). We assume that,

(i) there exists a finite measure $\bar\nu \in \mathcal{M}(\Omega)$ s.t. $Y = \Phi \star \bar\nu$,

(ii) the operator $\Phi\star : \mathcal{M}(\Omega) \to L^2(\rho)$ is injective.

In this case, we refer to $\bar\nu \in \mathcal{M}(\Omega)$ as the teacher measure *and to $\bar\mu \coloneqq |\bar\nu|/\|\bar\nu\|_{\mathrm{TV}} \in \mathcal{P}(\Omega)$ as the* teacher (feature) distribution*.*

In such a "teacher-student" framework, we are interested in determining to what extent the teacher feature distribution can be learned by the student neural network. Observe that, under Assumption III.1, $\mathcal{L}^0$ can be simply expressed in terms of the χ^2-divergence between the teacher feature distribution $\bar\mu$ and μ. By definition $\chi^2(\bar\mu|\mu) = \int_\Omega \left|\frac{\mathrm{d}\bar\mu}{\mathrm{d}\mu} - 1\right|^2 \mathrm{d}\mu$ and it follows from Eq. (III.17) that:

$$\mathcal{L}^0(\mu) = \int_\Omega \left|\frac{\mathrm{d}\bar\nu}{\mathrm{d}\mu}\right|^2 \mathrm{d}\mu = \|\bar\nu\|_{\mathrm{TV}}^2 \left(\int_\Omega \left|\frac{\mathrm{d}\bar\mu}{\mathrm{d}\mu} - 1\right|^2 \mathrm{d}\mu + 1\right) = \|\bar\nu\|_{\mathrm{TV}}^2 (\chi^2(\bar\mu|\mu) + 1)\,.$$

Then, following Eq. (50), the Wasserstein gradient flow of $\mathcal{L}^0$ corresponds to a nonlinear diffusion equation of the form:

$$\partial_t \mu = \operatorname{div}\left(\bar\mu \nabla \left(\frac{\mu}{\bar\mu}\right)^m\right)\,, \tag{III.11}$$

with $m < 0$ and $\bar\mu \in \mathcal{P}(\Omega)$, referred to as *ultra-fast diffusion* equation [Iacobelli, 2019b]. Note that this class of nonlinear diffusion equations stands out from the class of *linear diffusion* and *porous medium* equations (corresponding to the case $m \geq 1$ [Vázquez, 2006; Vázquez, 2007]) by the fact that the exponent m is negative and the diffusivity μ^{m-1} is singular at 0. In [Iacobelli, 2019a; Caglioti, 2018; Iacobelli, 2019b], the study of solutions to Eq. (III.11) is motivated by the convergence analysis of algorithms for the quantization of measures. In particular, Iacobelli, Patacchini, and Santambrogio [Iacobelli, 2019b] show the well-posedness of Eq. (III.11) on the d-dimensional torus or on bounded convex domains with Neumann boundary conditions and prove convergence of solutions towards the stationary state $\bar\mu$ in L^2. We prove in Theorem III.5 that Wasserstein gradient flows of our reduced risk $\mathcal{L}^\lambda$ converge towards solutions of the ultra-fast diffusion equation when the regularization strength λ vanishes.

Remark III.1.1. *Some remarks about Assumption III.1:*

- *At fixed $\lambda > 0$, the teacher-student assumption that $Y = \Phi \star \bar{\nu}$ is not restrictive since one can always replace Y by its orthogonal projection on the set $\{\Phi \star \nu,\, \nu \in \mathcal{M}(\Omega)\}$, thereby only modifying $\mathcal{L}^\lambda$ by subtracting a constant term. However, this assumption becomes crucial in the limit $\lambda \to 0^+$ to ensure the feasibility of the optimization problem in Eq. (III.14).*

- *The injectivity assumption on $\Phi\star$ ensures uniqueness of the reference measure $\bar{\nu}$. In the limit where $\lambda \to 0^+$, this allows rewriting $\mathcal{L}^0$ only in terms of a divergence between $\bar{\nu}$ and μ (Eq. (III.17)). In the case $\lambda > 0$, $\mathcal{L}^\lambda$ is an infimal convolution between this divergence and a* kernel discrepancy *(Eq. (III.18)) and the injectivity assumption ensures this discrepancy is a distance on the space of measures (Lemma III.A.1). It will be useful in Section III.5 to prove convergence of Wasserstein gradient flows of $\mathcal{L}^\lambda$ to solutions of the ultra-fast diffusion equation. In the case of a two-layer perceptron, the feature map is of the form $\phi((w,b),x) = \sigma(w^\top x + b)$ and the injectivity assumption is satisfied as soon as σ is not a polynomial and the data distribution has full support on $\mathbb{R}^d$ ([Sun, 2019, Thm. III.4]).*

III.1.2 Contributions and related works

Contributions This chapter studies the convergence of the VarPro algorithm — or two-timescale regime of gradient descent — for the training of mean-field models of neural networks. Precisely, we study the dynamic of the feature distribution $\mu \in \mathcal{P}(\Omega)$ when trained with gradient flow for the minimization of the reduced risk $\mathcal{L}^\lambda$, for $\lambda \geq 0$. In the teacher-student scenario defined by Assumption III.1, we establish guarantees for the convergence of μ towards the teacher feature distribution $\bar{\mu}$:

- In the case $\lambda = 0$, we show in Section III.4 that the training dynamic corresponds to an *ultra-fast diffusion* equation. Relying on the work of Iacobelli, Patacchini, and Santambrogio [Iacobelli, 2019b], this allows stating convergence towards the teacher feature distribution $\bar{\mu}$ (Theorem III.3), with a linear convergence rate.

- At fixed $\lambda > 0$, we establish in Theorem III.4 convergence of μ towards the teacher feature distribution $\bar{\mu}$ with an algebraic rate.

- In the limit $\lambda \to 0^+$, we show that, under regularity assumptions, the dynamic of the feature distribution μ converges locally uniformly in time to the solution of the *ultra-fast diffusion* equation with weights $\bar{\mu}$ (Theorem III.5).

- Finally, we show in Section III.6 that numerical results on low-dimensional learning problems with synthetic data are well-aligned with our theory. Overall, these experiments indicate that, when the regularization is sufficiently low, the VarPro dynamic indeed enters an "ultra-fast diffusion regime" where the student feature distribution converges to the teacher's at a linear rate. We also show with experiments on CIFAR10 that the VarPro algorithm can be adapted to the training of more complex architectures such as ResNets and achieves generalization on supervised learning problems with large datasets.

Convergence analysis for the training mean-field neural networks Several works have studied the convergence of gradient based methods for the training of neural network models similar to Eq. (III.1) with the mean-field scaling $\frac{1}{M}$. Chizat and Bach [Chizat, 2018] show that, for two layer neural networks with a homogeneous activation, if gradient

flow on the weights distribution converges then it converges towards a global minimizer of the risk. Rotskoff et al. [Rotskoff, 2019] show a similar result for a modification of the gradient flow dynamic where a supplementary "birth-death" term is added.

Several works have also analyzed the convergence of noisy gradient descent, or *Langevin dynamic*, for the training of mean-field models of two layer neural networks [Chizat, 2022; Mei, 2019; Nitanda, 2022; Hu, 2021; Suzuki, 2023]. Thanks to the addition of an entropic regularization term, these works provide a convergence rate for the sampling of an invariant weight distribution.

Two-timescale learning While two-timescale learning strategies have a broad range of applications in the fields of stochastic approximation and optimization [Borkar, 1997; Borkar, 2008], there has been a recent interest in these methods for the training of neural networks [Marion, 2023a; Berthier, 2024; Wang, 2024; Bietti, 2023; Takakura, 2024]. Specifically, Berthier, Montanari, and Zhou [Berthier, 2024] study the training of two-layer neural networks and exhibit a separation of timescales and different learning phases whose respective sizes depend on the timescale parameter η. Marion and Berthier [Marion, 2023a] study two-timescale gradient descent for a simple model of 1-dimensional neural network and show that the teacher network is recovered as soon as both the number of neurons of the student and the timescale parameter are sufficiently large. Bietti, Bruna, and Pillaud-Vivien [Bietti, 2023] consider a multi-index regression problem. Relying on the assumption of high dimensional Gaussian data, they consider a linear layer composed with a nonparameteric model whose projection can be computed in the Hermite basis. They show this instance of the VarPro algorithm results in a saddle-to-saddle dynamic on the linear layer and establish guarantees for the recovery of the teacher model.

Finally, Takakura and Suzuki [Takakura, 2024] and Wang, Mousavi-Hosseini, and Chizat [Wang, 2024] study the training of mean-field models of neural networks in the two-timescale limit with noisy gradient descent. In contrast with these works, we do not consider here additional entropic or L^2-regularization on the feature weights.

Wasserstein gradient flows of statistical distances Under our Assumption III.1, Eq. (III.18) shows $\mathcal{L}^\lambda$ is an infimal convolution of statistical divergences between the feature distribution μ and the teacher $\bar{\nu}$, interpolating between the χ^2-divergence $\chi^2(\bar{\nu}|\mu)$ — or more generally a f-divergence $\mathrm{D}_f(\bar{\nu}|\mu)$ — when $\lambda \to 0^+$ and a (squared) kernel discrepancy $\mathrm{MMD}(\bar{\nu},\mu)^2$ when $\lambda \to \infty$. In the case $\lambda \to \infty$, gradient flows of MMD-discrepancies and applications to sampling were studied in several works [Arbel, 2019; Sejdinovic, 2013; Hertrich, 2023a; Hertrich, 2023b; Hertrich, 2024; Boufadène, 2023]. Those flows are known to get trapped in local minima but discrepancies associated to non-smooth kernels have been observed to behave better in terms of convergence [Hertrich, 2023b; Hertrich, 2024]. In the case of the coulomb kernel, Boufadène and Vialard [Boufadène, 2023] prove that the discrepancy loss admits no spurious local minima and that the discrepancy flow converges towards the target measure under regularity assumptions.

In the intermediate regime $\lambda \in (0,\infty)$, several other works have also proposed regularization of f-divergences based on the infimal convolution with a kernel distance. For Glaser, Arbel, and Gretton [Glaser, 2021], the *KL Approximate Lower bound Estimator (KALE)* kernelizes the variational formulation of the KL-divergence and for Chen et al. [Chen, 2024] the *(De)-regularized Maximum Mean Discrepancy (DrMMD)* kernelizes the χ^2-distance. More generally, the work of Neumayer, Stein, and Steidl [Neumayer, 2024] studied kernelized variational formulations — or "Moreau envelopes in a RKHS" — of f-divergences. Similar to our Lemma III.3.3, they showed Γ-convergence of these func-

tionals towards the generating f-divergence when the regularization parameter λ tends to 0. They also studied numerically the convergence of the associated Wasserstein gradient flow towards the target distribution. The most notable difference between these regularized distances and the functional $\mathcal{L}^\lambda$ appearing in this chapter is that (w.r.t. [Neumayer, 2024, eq. (14)]) the role of the target $\bar{\nu}$ and parameter μ, over which optimization is performed, are interchanged. In other words, we consider optimizing over a statistical discrepancy which is the "reverse" of the one considered by Neumayer, Stein, and Steidl [Neumayer, 2024] and for this reason, though the mathematical tools to analyze it might be similar, the gradient flow dynamics will a priori have different behaviors.

Mathematical preliminaries and notations In the following, Ω will either be the n-dimensional torus or a closed bounded convex domain of $\mathbb{R}^n$, for some $n \geq 1$. We denote by $\mathcal{M}(\Omega)$ the set of finite Borel measures over Ω and by $\mathcal{P}(\Omega)$ the subset of $\mathcal{M}(\Omega)$ consisting of probability measures. We will denote by $\pi \in \mathcal{P}(\Omega)$ the uniform distribution over Ω. For a measure $\nu \in \mathcal{M}(\Omega)$, $|\nu|$ is its total variation measure and $\|\nu\|_{\mathrm{TV}}$ is the total variation of ν. For $p \in [1, +\infty)$, we denote by $\mathcal{W}_p$ the Wasserstein-p distance defined in Eq. (51). For probability measures $\mu, \mu' \in \mathcal{P}(\Omega)$,

$$\mathcal{W}_p(\mu, \mu') \coloneqq \min_{\gamma \in \Gamma(\mu,\mu')} \left(\int_{\Omega\times\Omega} \|\omega - \omega'\|^2 \mathrm{d}\gamma(\omega, \omega') \right)^{1/p},$$

where $\Gamma(\mu, \mu') \subset \mathcal{P}(\Omega \times \Omega)$ is the set of couplings between μ and μ' defined in Eq. (52). Standard references on the properties of the Wasserstein distance are the textbooks of Villani [Villani, 2009] and Santambrogio [Santambrogio, 2015]. If not otherwise specified, $\mathcal{M}(\Omega)$ and $\mathcal{P}(\Omega)$ are endowed with the topology of *narrow* convergence, that is the weak-* topology of $\mathcal{M}(\Omega)$ in duality with continuous functions. Importantly, because Ω is compact, this topology on $\mathcal{P}(\Omega)$ is equivalent to the $\mathcal{W}_p$-topology for any $p \in [1, +\infty)$ and $\mathcal{P}(\Omega)$ is compact.

For an integer $k \geq 0$ and for $s \in (0, 1]$, we denote by $\mathcal{C}^{k,s}(\Omega)$ (or just $\mathcal{C}^{k,s}$) the Hölder space of k-times continuously differentiable real-valued functions over Ω with s-Hölder k^{th}-derivative. We denote by $\|.\|_{\mathcal{C}^{k,s}}$ the Hölder norm on $\mathcal{C}^{k,s}(\Omega)$. For a probability measure $\rho \in \mathcal{P}(\mathbb{R}^d)$ and $p \in [1, +\infty]$, we denote by $L^p(\rho, \mathcal{C}^{k,s})$ the space of measurable functions $\phi : \Omega \times \mathbb{R}^d \to \mathbb{R}$ s.t. $\phi(., x) \in \mathcal{C}^{k,s}(\Omega)$ for $\mathrm{d}\rho$-a.e. $x \in \Omega$ and

$$\|\phi\|_{L^p(\rho,\mathcal{C}^{k,s})} \coloneqq \left(\int_{\mathbb{R}^d} \|\phi(., x)\|^p_{\mathcal{C}^{k,s}} \mathrm{d}\rho(x) \right)^{1/p} < +\infty\,.$$

We will often use that, if $\phi \in L^2(\rho, \mathcal{C}^{k,s})$ and $\alpha \in L^2(\rho)$, then the Bochner integral $\int_{\mathbb{R}^d} \phi(., x)\alpha(x)\mathrm{d}\rho(x)$ is in $\mathcal{C}^{k,s}$ with:

$$\left\| \int_{\mathbb{R}^d} \phi(., x)\alpha(x)\mathrm{d}\rho(x) \right\|_{\mathcal{C}^{k,s}} \leq \|\phi\|_{L^2(\rho,\mathcal{C}^{k,s})} \|\alpha\|_{L^2(\rho)}\,.$$

III.2 Reduced risk associated to the VarPro algorithm

We study in this chapter a VarPro algorithm or two-timescale regime of gradient descent for the training of neural networks. This strategy amounts to performing gradient descent on the *reduced risk* defined as the result of a partial minimization on a regularized version of the risk.

III.2.1 Primal formulation of the reduced risk

Whereas regularizing the risk with the Euclidean square norm of the weights is a popular practice, the variable projection procedure can be used with other kinds of regularization. Generally, for a convex function $f : \mathbb{R} \to \mathbb{R}$ and a regularization strength $\lambda > 0$ we consider for $\mu \in \mathcal{P}(\Omega)$ and $u \in L^1(\mu)$:

$$\mathcal{R}_f^\lambda(\mu, u) \coloneqq \frac{1}{2} \|F_{\mu,u} - Y\|_{L^2(\rho)}^2 + \lambda \int_\Omega f(u) \mathrm{d}\mu = \frac{1}{2} \|\Phi_\mu \cdot u - Y\|_{L^2(\rho)}^2 + \lambda \int_\Omega f(u) \mathrm{d}\mu \,, \tag{III.12}$$

where we assume $\mathcal{R}_f^\lambda(\mu, u) = +\infty$ if $f(u)$ is not integrable w.r.t. μ. As before we consider the *reduced risk* obtained by minimizing $\mathcal{R}_f^\lambda$ w.r.t. the outer weights u. For every $\mu \in \mathcal{P}(\Omega)$ we define:

$$\mathcal{L}_f^\lambda(\mu) \coloneqq \min_{u \in L^1(\mu)} \frac{1}{\lambda} \mathcal{R}_f^\lambda(\mu, u) = \min_{u \in L^1(\mu)} \frac{1}{2\lambda} \|\Phi_\mu \cdot u - Y\|_{L^2(\rho)}^2 + \int_\Omega f(u) \mathrm{d}\mu \,, \tag{III.13}$$

and this definition extends to the limiting case $\lambda \to 0^+$ by considering:

$$\mathcal{L}_f^0(\mu) \coloneqq \min_{\Phi_\mu \cdot u = Y} \int_\Omega f(u) \mathrm{d}\mu \,. \tag{III.14}$$

In the following, we always assume that $\phi \in L^2(\rho, \mathcal{C}^0(\Omega))$ (Assumption III.3). This in particular implies that, for any $\mu \in \mathcal{P}(\Omega)$, the map $\Phi_\mu : L^1(\mu) \to L^2(\rho)$ is weakly continuous. We also consider the following assumption on the regularization function:

Assumption III.2. *The function $f : \mathbb{R} \to \mathbb{R} \cup +\infty$ is nonnegative, strictly convex and superlinear i.e. such that $\lim_{\pm\infty} \frac{f(t)}{|t|} = +\infty$.*

By Lemma III.2.1, this is sufficient to ensure the existence of a unique minimizer

$$u_f^\lambda[\mu] \in \arg\min \mathcal{R}_f^\lambda(\mu, u) \,,$$

when $\lambda > 0$, and

$$u_f^0[\mu] \in \arg\min_{\Phi_\mu \cdot u = Y} \int_\Omega f(u) \mathrm{d}\mu \,,$$

when $\lambda = 0$. Of particular interest in this chapter and more precisely in Section III.4 is the case where $f(t) = |t|^r/(r-1)$ for some $r > 1$. In this case we denote the corresponding reduced risk by $\mathcal{L}_r^\lambda$. In particular, for $r = 2$ we recover the "L^2-regularized" reduced risk defined in Eq. (III.8) and Eq. (III.9).

Lemma III.2.1. *Assume Assumption III.2 holds. Then, for every $\mu \in \mathcal{P}(\Omega)$, the functional*

$$\mathcal{I}_f : u \in L^1(\mu) \mapsto \int_\Omega f(u) \mathrm{d}\mu$$

is strictly convex, weakly lower semicontinuous and has weakly compact sublevel sets. In particular, Eq. (III.13) (and Eq. (III.14) if feasible) admits a unique minimizer $u_f^\lambda[\mu]$.

Proof. Clearly $\mathcal{I}_f$ is strictly convex. Weak lower semicontinuity is a classical consequence of the fact that $\mathcal{I}_f$ is convex and strongly lower semicontinuous (using Fatou's lemma), hence its epigraph is convex and strongly closed and hence also weakly closed. For weak compacity of sublevel sets, if $(u_n)_{n \geq 0}$ is a sequence s.t. $\int_\Omega f(u_n) \mathrm{d}\mu \leq C$ for every $n \geq 0$,

then using that f has super-linear growth, for every $\varepsilon > 0$ there exists a $T \geq 0$ s.t. $|t| \leq \varepsilon f(t)$ for every $|t| \geq T$ and for every $n \geq 0$:

$$\int_{|u_n| \geq T} |u_n| \mathrm{d}\mu \leq \varepsilon \int f(u_n) \mathrm{d}\mu \leq \varepsilon C \,.$$

Thus the sequence $(u_n)_{n \geq 0}$ is uniformly integrable and admits a weakly converging subsequence by Dunford-Pettis theorem. □

III.2.2 Partial minimization on the space of measures

The reduced risk can also be obtained as the result of partial minimization of a convex functional over the space of measures. Whereas we have previously separated the role of the outer weights u and of the feature distribution μ in Eq. (III.2), our neural network model can equivalently be seen as a linear operator acting on the space $\mathcal{M}(\Omega)$ of finite measures on Ω.

For $\mu \in \mathcal{P}(\Omega)$ and $u \in L^1(\mu)$, we have by definition of $\Phi\star$ in Eq. (III.4) and of Φ_μ in Eq. (III.7) that $\phi_\mu \cdot u = \Phi \star \nu$ where $\nu \in \mathcal{M}(\Omega)$ is s.t. $\mathrm{d}\nu = u \mathrm{d}\mu$. Also

$$\int_\Omega f(u) \mathrm{d}\mu = \int_\Omega f(\frac{\mathrm{d}\nu}{\mathrm{d}\mu}) \mathrm{d}\mu = \mathrm{D}_f(\nu|\mu) \,,$$

where, for f satisfying Assumption III.2, D_f is the divergence defined by:

$$\forall (\nu, \mu) \in \mathcal{M}(\Omega) \times \mathcal{P}(\Omega), \quad \mathrm{D}_f(\nu|\mu) \coloneqq \begin{cases} \int_\Omega f(\frac{\mathrm{d}\nu}{\mathrm{d}\mu}) \mathrm{d}\mu & \text{if } \nu \ll \mu, \\ +\infty & \text{otherwise.} \end{cases} \tag{III.15}$$

In particular, in the case where f is an *entropy function* and $\nu \in \mathcal{P}(\Omega)$ is a probability measure, $\mathrm{D}_f(\nu|\mu)$ is the standard Csiszàr f-divergence [Liero, 2018]. Performing a change of variable, one can thus define the functional $\mathcal{L}_f^\lambda$ as the value resulting from a minimization problem over the space of measures. For $\mu \in \mathcal{P}(\Omega)$, minimizing over $\nu \in \mathcal{M}(\Omega)$ instead of $u \in L^1(\mu)$, we get:

$$\mathcal{L}_f^\lambda(\mu) = \begin{cases} \min\limits_{\nu \in \mathcal{M}(\Omega)} \frac{1}{2\lambda} \|\Phi \star \nu - Y\|^2 + \mathrm{D}_f(\nu|\mu) & \text{if } \lambda > 0, \\ \min\limits_{\nu \in \mathcal{M}(\Omega)} \iota_{\Phi\star\nu = Y} + \mathrm{D}_f(\nu|\mu) & \text{if } \lambda = 0. \end{cases} \tag{III.16}$$

As presented in Assumption III.1, of particular interest is the case where the signal Y itself can be exactly represented by a neural network, that is $Y = \Phi \star \bar{\nu}$, for some $\bar{\nu} \in \mathcal{M}(\Omega)$. Then in the case $\lambda = 0$, using the injectivity of $\Phi\star$, $\bar{\nu}$ is the only feasible solution in Eq. (III.16) and we obtain:

$$\mathcal{L}_f^0(\mu) = \int_\Omega f(\frac{\mathrm{d}\bar{\nu}}{\mathrm{d}\mu}) \mathrm{d}\mu = \mathrm{D}_f(\bar{\nu}|\mu) \,. \tag{III.17}$$

In the case $\lambda > 0$, $\mathcal{L}_f^\lambda$ can be interpreted as the infimal convolution between a *Maximum Mean Discrepancy (MMD)* and the divergence D_f. Indeed, naturally associated to the data distribution $\rho \in \mathcal{P}(\mathbb{R}^d)$ and to the feature map ϕ is a structure of Reproducing Kernel Hilbert Space (RKHS) of functions on Ω. We refer to Section III.A for results on the theory of RKHSs we use in this chapter. The RKHS $\mathcal{H}$ is defined in Eq. (III.47), and corresponds to the kernel $\kappa : \Omega \times \Omega \to \mathbb{R}$ defined by:

$$\forall \omega, \omega' \in \Omega, \quad \kappa(\omega, \omega') \coloneqq \int_\Omega \phi(\omega, x) \phi(\omega', x) \mathrm{d}\rho(x) \,.$$

It then follows from the definition of κ and $\mathcal{H}$ that, under Assumption III.1, the data attachment term in Eq. (III.16) can be interpreted as a kernel distance between ν and $\bar{\nu}$. By Eq. (III.49) we have $\|\Phi \star (\nu - \bar{\nu})\|_{L^2(\rho)} = \mathrm{MMD}_\kappa(\nu, \bar{\nu})$ where MMD_κ is the *Maximum Mean Discrepancy (MMD)* with kernel κ [Muandet, 2017; Gretton, 2012]. For $\lambda > 0$, the functional $\mathcal{L}_f^\lambda$ can then be expressed for every $\mu \in \mathcal{P}(\Omega)$ as:

$$\mathcal{L}_f^\lambda(\mu) = \min_{\nu \in \mathcal{M}(\Omega)} \frac{1}{2\lambda} \mathrm{MMD}_\kappa^2(\nu, \bar{\nu}) + \mathrm{D}_f(\nu|\mu)\,. \tag{III.18}$$

This last formulation of the functional $\mathcal{L}_f^\lambda$ resembles the notion of *Moreau envelope in a RKHS* of the divergence D_f introduced by Neumayer, Stein, and Steidl [Neumayer, 2024]. This notion encompasses the particular cases of *De-regularized MMD* studied in [Chen, 2024] and *KL Approximate Lower bound Estimator* studied in [Glaser, 2021]. Nonetheless, w.r.t. [Neumayer, 2024, eq. (14)], the role of the target measure $\bar{\nu}$ and of the optimized measure μ are here interchanged, which is expected to play an important role in the gradient flow dynamic.

III.2.3 Dual formulation of the reduced risk

In Eqs. (III.13) and (III.14), the objectives $\mathcal{L}_f^\lambda$ and $\mathcal{L}_f^0$ are expressed as the value of a minimization problem over the outer weights u. Taking the dual of those minimization problems, $\mathcal{L}_f^\lambda$ and $\mathcal{L}_f^0$ can be expressed as the value of a maximization problem over the dual variable $\alpha \in L^2(\rho)$. In contrast with the primal formulation Eq. (III.13), the dual formulation of Proposition III.2.1 has the advantage of conveniently expressing $\mathcal{L}_f^\lambda$ for both $\lambda > 0$ and $\lambda = 0$ as the value of an optimization problem over the space $L^2(\rho)$ which is independent of μ.

Proposition III.2.1 (Dual representation)**.** *Let Assumption III.2 hold and consider $\mu \in \mathcal{P}(\Omega)$. Then we have for $\lambda > 0$:*

$$\mathcal{L}_f^\lambda(\mu) = \max_{\alpha \in L^2(\rho)} -\int_\Omega f^*(\Phi^\top \alpha) \mathrm{d}\mu + \langle \alpha, Y \rangle_{L^2(\rho)} - \frac{\lambda}{2} \|\alpha\|_{L^2(\rho)}^2\,, \tag{III.19}$$

where f^* is the Legendre transform of f and $\Phi^\top : L^2(\rho) \to \mathcal{C}^0(\Omega)$ is defined by:

$$\forall \alpha \in L^2(\rho), \quad \Phi^\top \alpha \coloneqq \int_{\mathbb{R}^d} \phi(., x) \alpha(x) \mathrm{d}\rho(x)\,.$$

The supremum in Eq. (III.19) is attained at some $\alpha_f^\lambda[\mu] \in L^2(\rho)$ and for $u_f^\lambda[\mu] \in L^1(\mu)$ the optimizer in Eq. (III.13) it holds:

$$\lambda \alpha_f^\lambda[\mu] = \Phi_\mu \cdot u_f^\lambda[\mu] - Y \quad \textit{and} \quad f(u_f^\lambda[\mu]) + f^*(\Phi^\top \alpha_f^\lambda[\mu]) = u_f^\lambda[\mu](\Phi^\top \alpha_f^\lambda[\mu])\,. \tag{III.20}$$

Moreover, Eq. (III.19) also holds in the case $\lambda = 0$ under Assumption III.1.

When $\lambda > 0$, this result yields a convenient reformulation of the functional $\mathcal{L}_f^\lambda$. For $\mu \in \mathcal{P}(\Omega)$, $\alpha_f^\lambda[\mu] \in L^2(\rho)$ being the maximizer in Eq. (III.19) and $u_f^\lambda[\mu] \in L^1(\mu)$ the minimizer in Eq. (III.13), we have:

$$\mathcal{L}_f^\lambda(\mu) = \frac{\lambda}{2} \left\| \alpha_f^\lambda[\mu] \right\|_{L^2(\rho)}^2 + \int_\Omega f(u_f^\lambda[\mu]) \mathrm{d}\mu\,. \tag{III.21}$$

Proof. Consider $\mu \in \mathcal{P}(\Omega)$ and $\lambda > 0$. First, by definition of $\Phi^\top$ we have for every $\alpha \in L^2(\rho)$ and every $u \in L^1(\mu)$ that $\int_\Omega (\Phi^\top \alpha) u \mathrm{d}\mu = \langle \alpha, \Phi_\mu \cdot u \rangle_{L^2(\rho)}$ i.e. $\Phi^\top$ is the adjoint of $\Phi_\mu : L^1(\mu) \to L^2(\mu)$. Also, it follows from the assumption on f that the map $\mathcal{I}_f : u \in L^1(\mu) \mapsto \int_\Omega f(u)\mathrm{d}\mu$ is a convex, weakly lower semicontinuous functional whose Legendre transform is given for $h \in L^\infty(\mu)$ by:

$$\mathcal{I}_f^*(h) = \sup_{u \in L^1(\mu)} \int_\Omega hu\mathrm{d}\mu - \int_\Omega f(u)\mathrm{d}\mu = \int_\Omega f^*(h)\mathrm{d}\mu\,,$$

with f^* the Legendre transform of f and where the supremum is attained for $u \in L^1(\mu)$ satisfying the duality relation $f(u) + f^*(h) = uh$ [Rockafellar, 1968, Thm. 2]. Similarly, for $u \in L^1(\mu)$ we have:

$$\sup_{\alpha \in L^2(\rho)} -\langle \alpha, \Phi_\mu \cdot u - Y \rangle_{L^2(\rho)} - \frac{\lambda}{2}\|\alpha\|^2_{L^2(\rho)} = \frac{1}{2\lambda}\|\Phi_\mu \cdot u - Y\|^2_{L^2(\rho)}\,.$$

where the supremum is reached at $\alpha = \lambda(\Phi_\mu \cdot u - Y)$ when $\lambda > 0$. Moreover, the functional $\alpha \mapsto \frac{1}{2\lambda}\|\alpha\|^2_{L^2(\rho)}$ being continuous, we can apply [Rockafellar, 1967, Thm. 3] and Eq. (III.19) holds by strong duality. The optimums are attained in both Eq. (III.13) and Eq. (III.19) and thus Eq. (III.20) expresses the optimality conditions.

Finally, for the case $\lambda = 0$, when Assumption III.1 holds we have by Eq. (III.16) that $\mathcal{L}_f^0(\mu) = \mathrm{D}_f(\bar{\nu}|\mu)$. Also, the assumptions on f ensures $\mathrm{dom}(f^*) = \mathbb{R}$ and using [Rockafellar, 1971, Thm. 4] we obtain:

$$\mathcal{L}_f^\lambda(\mu) = \mathrm{D}_f(\bar{\nu}|\mu) = \sup_{h \in \mathcal{C}^0(\Omega)} \int_\Omega h\mathrm{d}\bar{\nu} - \int_\Omega f^*(h)\mathrm{d}\mu\,.$$

The result follows as the injectivity of $\Phi\star$ ensures $\mathrm{Range}(\Phi^\top)$ is dense in $\mathcal{C}^0(\Omega)$ (Lemma III.A.1). □

Observing that $\Phi^\top$ defines a partial isometry from $L^2(\rho)$ to the RKHS $\mathcal{H}$ (Eq. (III.47)), a similar dual formulation of $\mathcal{L}_f^\lambda$ also holds in duality with $\mathcal{H}$.

Proposition III.2.2. *Let Assumption III.2 and Assumption III.1 hold and consider $\mu \in \mathcal{P}(\Omega)$. Then we have for $\lambda \geq 0$:*

$$\mathcal{L}_f^\lambda(\mu) = \sup_{h \in \mathcal{H}} -\int_\Omega f^*(h)\mathrm{d}\mu + \int_\Omega h\mathrm{d}\bar{\nu} - \frac{\lambda}{2}\|h\|^2_{\mathcal{H}}\,, \tag{III.22}$$

where f^* is the Legendre transform of f. For $\lambda > 0$, the supremum in Eq. (III.22) is attained at some $h_f^\lambda[\mu] \in \mathcal{H}$ and for $\nu_f^\lambda[\mu] \in L^1(\mu)$ the optimizer in Eq. (III.18) it holds:

$$\lambda h_f^\lambda[\mu] = \Phi^\top \Phi \star (\nu_f^\lambda[\mu] - \bar{\nu}) \quad \textit{and} \quad f\Big(\frac{\mathrm{d}\nu_f^\lambda[\mu]}{\mathrm{d}\mu}\Big) + f^*(h_f^\lambda[\mu]) = h_f^\lambda[\mu]\frac{\mathrm{d}\nu_f^\lambda[\mu]}{\mathrm{d}\mu}\,. \tag{III.23}$$

Proof. The formula Eq. (III.22) is directly deduced from Eq. (III.19) and the characterization of the RKHS $\mathcal{H}$ in Eq. (III.47). Also Eq. (III.23) is a rewritting of Eq. (III.20) since $\nu_f^\lambda[\mu] \in \mathcal{M}(\Omega)$ and $h_f^\lambda[\mu] \in \mathcal{H}$ are related to $u_f^\lambda[\mu] \in L^1(\mu)$ and $\alpha_f^\lambda[\mu] \in L^2(\rho)$ by

$$\mathrm{d}\nu_f^\lambda[\mu] = u_f^\lambda[\mu]\mathrm{d}\mu \quad \text{and} \quad h_f^\lambda[\mu] = \Phi^\top \alpha_f^\lambda[\mu]\,.$$

□

III.2.4 Kernel learning in the case of quadratic regularization

The case of a quadratic regularization is of particular interest since the partial optimization problem over u admits a closed-form solution which can be efficiently obtained numerically by solving a linear system. In this case, the task of minimizing the reduced risk is equivalent to solving a *Multiple Kernel Learning* problem [Bach, 2004].

For the L^2-regularization $f(t) = |t|^2$, the reduced risk $\mathcal{L}_2^\lambda[\mu]$ is the value of the *ridge regression problem* in Eq. (III.8) and for $\lambda > 0$, the optimizer is given by

$$u_2^\lambda[\mu] = (\Phi_\mu^\top \Phi_\mu + 2\lambda)^{-1} \Phi_\mu^\top Y \,,$$

where $\Phi_\mu^\top : L^2(\rho) \to L^2(\mu)$ is the adjoint of the operator Φ_μ restricted to $L^2(\mu)$. Also, the dual problem in Eq. (III.19) here reads:

$$\mathcal{L}_2^\lambda(\mu) = \sup_{\alpha \in L^2(\rho)} -\frac{1}{2} \langle \alpha, (K_\mu + 2\lambda)\alpha \rangle_{L^2(\rho)} + \langle \alpha, Y \rangle_{L^2(\rho)} \,,$$

where $K_\mu : L^2(\rho) \to L^2(\rho)$ is the self-adjoint operator defined by $K_\mu = \Phi_\mu \Phi_\mu^\top$. The supremum is attained at $\alpha_2^\lambda[\mu] = (K_\mu + 2\lambda)^{-1} Y$ and by Eq. (III.21) we obtain for every $\mu \in \mathcal{P}(\Omega)$:

$$\mathcal{L}_2^\lambda(\mu) = \frac{1}{2} \left\langle Y, (K_\mu + 2\lambda)^{-1} Y \right\rangle_{L^2(\rho)} \,. \tag{III.24}$$

This is the optimal value of the kernel ridge regression problem with kernel K_μ, where K_μ is parameterized by the feature distribution μ. Moreover this parameterization is linear w.r.t. $\mu \in \mathcal{P}(\Omega)$ since, considering for $\omega \in \Omega$ the rank the rank-one self-adjoint operator $k(\omega) \coloneqq \phi(\omega, .) \otimes \phi(\omega, .)$, we have:

$$K_\mu = \int_\Omega k(\omega) \mathrm{d}\mu(\omega) \,.$$

Therefore, minimizing the reduced risk $\mathcal{L}_2^\lambda$ over the feature distribution μ amounts to finding the best kernel for solving the ridge regression problem in Eq. (III.8) among convex combinations of "simple" basis kernels $(k(\omega))_{\omega \in \Omega}$ i.e. a *Multiple Kernel Learning* task. Other convex optimization strategies for solving such task have been studied in [Lanckriet, 2004; Bach, 2004].

III.3 Properties of minimizers of the reduced risk

Before turning to the analysis of gradient methods for the minimization of the reduced risk $\mathcal{L}_f^\lambda$ in Sections III.4 and III.5, we study here variational properties of $\mathcal{L}_f^\lambda$.

III.3.1 Existence and uniqueness of minimizers

We first investigate existence and uniqueness of minimizers of $\mathcal{L}_f^\lambda$. Importantly, we use here that $\mathcal{L}_f^\lambda$ is obtained as the result of a partial minimization. Namely, for $\lambda \geq 0$ and $\mu \in \mathcal{P}(\Omega)$, we have from Eq. (III.16) that $\mathcal{L}_f^\lambda(\mu) = \min_{\nu \in \mathcal{M}(\Omega)} \mathcal{E}_f^\lambda(\nu, \mu)$, where $\mathcal{E}_f^\lambda$ is defined for $\nu \in \mathcal{M}(\Omega)$ and $\mu \in \mathcal{P}(\Omega)$ by:

$$\mathcal{E}_f^\lambda(\nu, \mu) \coloneqq \begin{cases} \mathrm{D}_f(\nu|\mu) + \frac{1}{2\lambda} \|\Phi \star \nu - Y\|^2 & \text{if } \lambda > 0, \\ \mathrm{D}_f(\nu|\mu) + \iota_{\Phi \star \nu = Y} & \text{if } \lambda = 0. \end{cases} \tag{III.25}$$

In particular, it follows from variational formulations of f-divergences that D_f is (jointly) convex and lower semicontinuous w.r.t. its arguments $(\nu, \mu) \in \mathcal{M}(\Omega) \times \mathcal{P}(\Omega)$ [Rockafellar, 1971, Thm. 4]. The following Lemma III.3.1 uses this fact to establish convexity and lower semicontinuity of $\mathcal{L}_f^\lambda$, implying the existence of minimizers. We then discuss cases in which $\mathcal{L}_f^\lambda$ has in fact a unique minimizer.

Lemma III.3.1. *Assume f satisfies Assumption III.2. Then, for $\lambda \geq 0$, $\mathcal{L}_f^\lambda : \mathcal{P}(\Omega) \to \mathbb{R}$ is a convex, lower semicontinuous function (w.r.t. the narrow convergence on $\mathcal{M}(\Omega)$).*

Proof. By the definition of the divergence D_f in Eq. (III.15) and by [Rockafellar, 1971, Thm. 4], we have for every $(\nu, \mu) \in \mathcal{M}(\Omega) \times \mathcal{P}(\Omega)$:

$$\mathrm{D}_f(\nu|\mu) = \sup_{h \in \mathcal{C}^0(\Omega)} \int_\Omega h \mathrm{d}\nu - \int_\Omega f^*(h) \mathrm{d}\mu \,.$$

Thus D_f is a (jointly) convex and lower semicontinuous function as a supremum of (jointly) convex and lower semicontinuous functions. As a consequence, for $\lambda \geq 0$, $\mathcal{E}_f^\lambda$ is also (jointly) convex and lower semicontinuous. The convexity of $\mathcal{L}_f^\lambda = \min_\nu \mathcal{E}_f^\lambda(\nu, .)$ follows as partial minimization preserves convexity. Also, if $(\mu_n)_{n \geq 0}$ is a sequence in $\mathcal{P}(\Omega)$ converging narrowly to some $\mu \in \mathcal{P}(\Omega)$, then we have $\mathcal{L}_f^\lambda(\mu_n) = \mathcal{E}_f^\lambda(\nu_n, \mu_n)$ for some $\nu_n \in \mathcal{M}(\Omega)$. Without loss of generality one can assume $\mathcal{L}_f^\lambda(\mu_n)$ is bounded, thus $\mathrm{D}_f(\nu_n|\mu_n)$ and then $\|\nu_n\|_{\mathrm{TV}}$ are bounded as f is superlinear. Then, up to extraction of a subsequence, (ν_n) converges narrowly to $\nu \in \mathcal{M}(\Omega)$ and we get by lower semicontinuity of $\mathcal{E}_f^\lambda$:

$$\liminf_{n \to \infty} \mathcal{L}_f^\lambda(\mu_n) = \liminf_{n \to \infty} \mathcal{E}_f^\lambda(\nu_n, \mu_n) \geq \mathcal{E}_f^\lambda(\nu, \mu) \geq \mathcal{L}_f^\lambda(\mu) \,,$$

which shows that $\mathcal{L}_f^\lambda$ is lower semicontinuous. □

The above result implies the existence of minimizers of the reduced risk $\mathcal{L}_f^\lambda$ for every $\lambda \geq 0$ but it does not establish uniqueness and $\mathcal{L}_f^\lambda$ may, a priori, have several minimizers. However, there are cases in which uniqueness can be ensured. We give two examples:

- In the teacher-student setup where Assumption III.1 holds, if $\bar{\nu}$ is a positive measure with $\bar{m} = \bar{\nu}(\Omega) > 0$ and if f is nonnegative, strictly convex and s.t. $f(\bar{m}) = 0$ then the teacher feature distribution $\bar{\mu} = \bar{\nu}/\bar{m}$ is the unique minimizer of $\mathcal{L}_f^\lambda$, whatever $\lambda \geq 0$. Indeed, from Eq. (III.18) we have that $\mathcal{L}_f^\lambda(\bar{\mu}) = 0$ and $\mathcal{L}_f^\lambda(\mu) > 0$ for every $\mu \neq \bar{\mu}$. With these assumptions, we prove in Theorem III.4 that the gradient flow of $\mathcal{L}_f^\lambda$ converges towards the teacher feature distribution $\bar{\mu}$ with an algebraic convergence rate.

- For general data Y, relying on a variational characterization of the total variation, the following Lemma III.3.2 establishes uniqueness of a minimizer to $\mathcal{L}_r^\lambda$ in the case the regularization is of the form $f(t) = |t|^r$ for some $r > 1$.

Lemma III.3.2. *Let $\lambda \geq 0$ and assume $f(t) = |t|^r$ for some $r > 1$. Then $\mathcal{L}_r^\lambda$ admits a unique minimizer $\bar{\mu}_r^\lambda \in \mathcal{P}(\Omega)$.*

Proof. We use arguments similar to the one of [Wang, 2024, Prop. 3.3]. By duality $\inf_{\mu \in \mathcal{P}(\Omega)} \mathcal{L}_r^\lambda(\mu) = \inf_{\nu \in \mathcal{M}(\Omega)} \mathcal{G}_r^\lambda(\nu)$ where $\mathcal{G}_r^\lambda$ is defined for $\nu \in \mathcal{M}(\Omega)$ by:

$$\mathcal{G}^\lambda(\nu) \coloneqq \begin{cases} \|\nu\|_{\mathrm{TV}}^r + \frac{1}{2\lambda} \|\Phi \star \nu - Y\|^2 & \text{if } \lambda > 0, \\ \|\nu\|_{\mathrm{TV}}^r + \iota_{\Phi \star \nu = Y} & \text{if } \lambda = 0, \end{cases}$$

where we used the variational representation $\|\nu\|_{\mathrm{TV}}^r = \inf_{\mu\in\mathcal{P}(\Omega)} \int_\Omega \left|\frac{\mathrm{d}\nu}{\mathrm{d}\mu}\right|^r \mathrm{d}\mu$ (the case $r=2$ is used in [Wang, 2024; Lanckriet, 2004]). Indeed, for measures $\nu \in \mathcal{M}(\Omega)$ and $\mu \in \mathcal{P}(\Omega)$ s.t. $\nu \ll \mu$ we have:

$$\int_\Omega \left|\frac{\mathrm{d}\nu}{\mathrm{d}\mu}\right|^r \mathrm{d}\mu = \int_\Omega \left(\frac{\mathrm{d}\mu}{\mathrm{d}|\nu|}\right)^{1-r} \mathrm{d}|\nu|.$$

The Lagrangian of the convex problem $\inf_\mu \int_\Omega \left(\frac{\mathrm{d}\mu}{\mathrm{d}|\nu|}\right)^{1-r} \mathrm{d}|\nu|$ is given by:

$$\mathcal{J}(\mu,\gamma) = \int_\Omega \left(\frac{\mathrm{d}\mu}{\mathrm{d}|\nu|}\right)^{1-r} \mathrm{d}|\nu| + \gamma\left(\int_\Omega \mathrm{d}\mu - 1\right).$$

The optimality condition gives that $\frac{\mathrm{d}\mu}{\mathrm{d}|\nu|}$ is constant and the minimum is attained for $\mu = |\nu|/\|\nu\|_{\mathrm{TV}}$, giving $\int_\Omega \left|\frac{\mathrm{d}\nu}{\mathrm{d}\mu}\right|^r \mathrm{d}\mu = \|\nu\|_{\mathrm{TV}}^r$. Finally, the map $\nu \mapsto \|\nu\|_{\mathrm{TV}}^r$ is strictly convex so that $\mathcal{G}_r^\lambda$ admits a unique minimizer $\bar\nu_r^\lambda \in \mathcal{M}(\Omega)$, thus $\mathcal{L}_r^\lambda$ has also a unique minimizer $\bar\mu_r^\lambda \in \mathcal{P}(\Omega)$ and we have the duality relation $\bar\mu_r^\lambda = |\bar\nu_r^\lambda|/\|\bar\nu_r^\lambda\|_{\mathrm{TV}}$. Notably, in the case where Assumption III.1 holds and $\lambda = 0$, we have $\bar\nu_r^0 = \bar\nu$ and $\bar\mu_r^0 = \bar\mu$ for every $r > 1$. □

III.3.2 Convergence of minimizers

Of particular interest to us is the case $\lambda = 0$ for which minimizers of $\mathcal{L}_f^0$ are related to the teacher measure $\bar\nu$ by Eq. (III.17). However, in practice, minimization of $\mathcal{L}_f^\lambda$ is easier in the presence of a regularization parameter $\lambda > 0$. For this reason, we are interested in the asymptotic behavior of minimizers of $\mathcal{L}_f^\lambda$ when $\lambda \to 0^+$.

We show here, when $\lambda \to 0^+$, that any converging sequence of minimizers to $\mathcal{L}_f^\lambda$ converges to some minimizer of $\mathcal{L}_f^0$. In particular, if $\mathcal{L}_f^0$ has a unique minimizer $\bar\mu^0$ then any sequence of minimizers to $\mathcal{L}_f^\lambda$ converges to $\bar\mu^0$. This result is a consequence of the following Lemma III.3.3 which states the Γ-convergence of the functionals $\mathcal{L}_f^\lambda$ to $\mathcal{L}_f^0$. We refer to [Santambrogio, 2023, Chap. 7] for an introduction to Γ-convergence. This is in particular stronger than pointwise convergence and is the appropriate notion of convergence for studying the behavior of minimizers. In the case where Assumption III.1 holds and $\mathcal{L}_f^\lambda$ admits the representation Eq. (III.18), a similar result was established by Neumayer, Stein, and Steidl [Neumayer, 2024], with a notable difference being here that we prove Γ-convergence w.r.t. the variable μ instead of $\bar\nu$.

Lemma III.3.3 (Γ-convergence)**.** *Assume Assumptions III.1 and III.2 hold, $\phi \in L^2(\rho, \mathcal{C}^{0,1})$ and $f^* \in \mathcal{C}^{0,1}_{loc}(\mathbb{R})$. Then the family of functionals $(\mathcal{L}_f^\lambda)_{\lambda>0}$ Γ-converges towards $\mathcal{L}_f^0$ as $\lambda \to 0^+$ in the sense that for every sequence $(\lambda_n)_{n\geq 0}$ converging to 0^+ and every $\mu \in \mathcal{P}(\Omega)$ it holds:*

(i) for every sequence $(\mu_n)_{n\geq 0}$ converging narrowly to μ,

$$\liminf_{n\to+\infty} \mathcal{L}_f^{\lambda_n}(\mu_n) \geq \mathcal{L}_f^0(\mu)_f\,,$$

(ii) there exists a sequence $(\mu_n)_{n\geq 0}$ converging narrowly to μ s.t.

$$\limsup_{n\to+\infty} \mathcal{L}_f^{\lambda_n}(\mu_n) \leq \mathcal{L}_f^0(\mu)\,.$$

Proof. For the second part of the result it suffices to consider the constant sequence $\mu_n = \mu$ for every $n \geq 0$. Indeed, it then directly follows from the definition of $\mathcal{L}_f^\lambda$ and $\mathcal{L}_f^0$ in Eq. (III.13) and Eq. (III.14) that $\mathcal{L}_f^\lambda(\mu) \leq \mathcal{L}_f^0(\mu)$ for every $\lambda > 0$.

To prove the first part of the definition, consider a sequence $(\mu_n)_{n\geq 0}$ converging narrowly to μ in $\mathcal{P}(\Omega)$. By the dual formulation of $\mathcal{L}_f^0$ in Eq. (III.19), we can also consider a sequence $(\alpha_k)_{k\geq 0}$ in $L^2(\rho)$ such that:

$$-\int_\Omega f^*(\Phi^\top \alpha_k)\mathrm{d}\mu + \langle \alpha_k, Y\rangle \xrightarrow{k\to\infty} \mathcal{L}_f^0(\mu)\,.$$

Then, by the dual formulation of $\mathcal{L}_f^\lambda$ in Eq. (III.19), for every $n, k \geq 0$:

$$\begin{aligned}
\mathcal{L}^{\lambda_n}(\mu_n) &\geq -\int_\Omega f^*(\Phi^\top \alpha_k)\mathrm{d}\mu_n + \langle \alpha_k, Y\rangle - \frac{\lambda_n}{2}\|\alpha_k\|_{L^2\rho}^2 \\
&= -\int_\Omega f^*(\Phi^\top \alpha_k)\mathrm{d}(\mu_n - \mu) - \frac{\lambda_n}{2}\|\alpha_k\|_{L^2\rho}^2 - \int_\Omega f^*(\Phi^\top \alpha_k)\mathrm{d}\mu + \langle \alpha_k, Y\rangle \\
&\geq -\left\|f^*(\Phi^\top \alpha_k)\right\|_{\mathcal{C}^{0,1}} \mathcal{W}_1(\mu_n, \mu) - \frac{\lambda_n}{2}\|\alpha_k\|_{L^2\rho}^2 - \int_\Omega f^*(\Phi^\top \alpha_k)\mathrm{d}\mu + \langle \alpha_k, Y\rangle\,.
\end{aligned}$$

But then, since $\mathcal{W}_1(\mu_n, \mu) \to 0$ and $\lambda_n \to 0$, one can find an increasing sequence $(k_n)_{n\geq 0}$ s.t.:

$$\lambda_n \|\alpha_{k_n}\|_{L^2(\rho)}^2 \xrightarrow{n\to+\infty} 0 \quad \text{and} \quad \left\|f^*(\Phi^\top \alpha_{k_n})\right\|_{\mathcal{C}^{0,1}} \mathcal{W}_1(\mu_n, \mu) \xrightarrow{n\to+\infty} 0\,.$$

Thus $\mathcal{L}_f^{\lambda_n}(\mu_n) \geq \mathcal{L}_f^0(\mu) + o(1)$ for every $n \geq 0$ and the result follows. □

It is a direct consequence of the above Γ-convergence result that the limit when $\lambda \to 0^+$ of a sequence of minimizers of $\mathcal{L}_f^\lambda$ is a minimizer of $\mathcal{L}_f^0$ [Santambrogio, 2023, Prop. 7.5]. In the case where $\mathcal{L}_f^\lambda$ has a unique minimizer $\bar{\mu}_f^0$, this implies every sequence of minimizers of $\mathcal{L}_f^\lambda$ converges to $\bar{\mu}_f^0$ when $\lambda \to 0^+$.

Proposition III.3.1 (Convergence of minimizers)**.** *Assume the result of Lemma III.3.3 holds. For every $\lambda > 0$, consider $\bar{\mu}_f^\lambda \in \arg\min \mathcal{L}_f^\lambda$ and assume $\bar{\mu}_f^\lambda \xrightarrow{\lambda\to 0^+} \mu$. Then $\mu \in \arg\min \mathcal{L}_f^0$. Notably, if $\mathcal{L}_f^0$ has a unique minimizer $\bar{\mu}_f^0$, then $\bar{\mu}_f^\lambda \xrightarrow{\lambda\to 0^+} \bar{\mu}_f^0$.*

III.4 Training with gradient flow

In the rest of this chapter we consider the optimization over the feature distribution $\mu \in \mathcal{P}(\Omega)$ for the minimization of the reduced risk $\mathcal{L}_f^\lambda$, for $\lambda \geq 0$. Specifically, we consider a *gradient flow* algorithm. In the case of a finite number of features $\{\omega_i\}_{1\leq i\leq M} \in \Omega^M$ such gradient flow is defined as the solution of the equation:

$$\forall i \in \{1, ..., M\}\,, \quad \frac{\mathrm{d}}{\mathrm{d}t}\omega_i(t) = -M\nabla_{\omega_i}\hat{\mathcal{L}}_f^\lambda(\{\omega_i(t)\}_{1\leq i\leq M}) \tag{III.26}$$

where $\hat{\mathcal{L}}_f^\lambda(\{\omega_i\}_{1\leq i\leq M}) \coloneqq \mathcal{L}_f^\lambda(\hat{\mu})$ and $\hat{\mu}$ is the empirical distribution $\hat{\mu} = \frac{1}{M}\sum_{i=1}^M \delta_{\omega_i}$. More generally, in terms of the feature distribution $\mu \in \mathcal{P}(\Omega)$, the above equation corresponds to a *Wasserstein gradient flow* over the functional $\mathcal{L}_f^\lambda$, namely:

$$\partial_t \mu_t - \mathrm{div}\left(\mu_t \nabla \frac{\delta \mathcal{L}_f^\lambda}{\delta \mu}[\mu_t]\right) = 0, \quad \text{on } (0,\infty)\times\Omega, \tag{III.27}$$

where for $\mu \in \mathcal{P}(\Omega)$, $\frac{\delta \mathcal{L}_f^\lambda}{\delta \mu}[\mu]$ is the *Fréchet differential* of $\mathcal{L}_f^\lambda$ at μ [Santambrogio, 2015, Def. 7.12]. Importantly, Jordan, Kinderlehrer, and Otto [Jordan, 1998] have shown that Wasserstein gradient flows can be obtained as limits of proximal update schemes when the discretization step tends to 0. Here, the curve $(\mu_t)_{t\geq 0}$ is the limit of the piecewise-constant curve with values $(\mu_k^\tau)_{k\geq 0}$ where, given a time-step $\tau > 0$ and an initialization $\mu_0^\tau \in \mathcal{P}(\Omega)$, the sequence $(\mu_k^\tau)_{k\geq 0}$ is defined recursively by:

$$\forall k \geq 0, \quad \mu_{k+1}^\tau \in \operatorname*{arg\,min}_{\mu \in \mathcal{P}(\Omega)} \mathcal{L}_f^\lambda(\mu) + \frac{1}{2\tau} \mathcal{W}_2(\mu, \mu_k^\tau)^2 \,. \tag{III.28}$$

We study in this section the well-posedness of the above Wasserstein gradient flow equation by distinguishing the case where $\lambda > 0$ and the case $\lambda = 0$. In the latter case, we show the Wasserstein gradient flow corresponds to a *weighted ultra-fast diffusion equation* [Iacobelli, 2019b].

III.4.1 Wasserstein gradient flows in the case $\lambda > 0$

In the case where $\lambda > 0$, the presence of the regularization induces sufficient regularity on the objective to study the training dynamic through the lens of classical results from the theory of gradient flows in the Wasserstein space [Ambrosio, 2008b; Santambrogio, 2017]. In particular, one can derive the gradient flow equation leveraging the dual representation of $\mathcal{L}_f^\lambda$. Indeed, Eq. (III.19) expresses $\mathcal{L}_f^\lambda$ as a maximum over linear functionals, and thus by the envelope theorem one can formally differentiate $\mathcal{L}_f^\lambda$ w.r.t. μ and obtain the Fréchet differential:

$$\frac{\delta \mathcal{L}_f^\lambda}{\delta \mu}[\mu](\omega) = -f^*(\Phi^\top \alpha_f^\lambda[\mu])(\omega) \,,$$

with $\alpha_f^\lambda[\mu] \in L^2(\rho)$ the maximizer in Eq. (III.19). We show that the gradient field of this potential indeed defines a notion of "gradient" for the functional $\mathcal{L}_f^\lambda$ w.r.t. the Wasserstein topology on $\mathcal{P}(\Omega)$.

Locally absolutely continuous curves $(\mu_t)_{t\in[0,1]}$ in the space $\mathcal{P}(\Omega)$, equipped with the Wasserstein distance $\mathcal{W}_2$, are characterised as solutions to a continuity equation:

$$\partial_t \mu_t + \operatorname{div}(\mu_t v_t) = 0 \quad \text{on } (0, +\infty) \times \Omega \tag{III.29}$$

for some velocity field v such that $\|v_t\|_{L^2(\mu_t)} \in L^1_{loc}((0, +\infty))$ [Santambrogio, 2015, Thm. 5.14]. This equation has to be understood in the sense of distributions, that is in duality with the set $\mathcal{C}_c^\infty((0, +\infty) \times \Omega)$ of smooth compactly supported test functions, i.e.:

$$\int_0^1 \int_\Omega (\partial_t \varphi + \langle \nabla \varphi, v_t \rangle) \, \mathrm{d}\mu_t \mathrm{d}t = 0, \quad \forall \varphi \in \mathcal{C}_c^\infty((0, +\infty) \times \Omega) \,. \tag{III.30}$$

The following result shows that the functional $\mathcal{L}_f^\lambda(\mu_t)$ is differentiable along those curves and expresses its derivative in terms of the gradient field $\nabla \frac{\delta \mathcal{L}_f^\lambda}{\delta \mu}$.

Lemma III.4.1 (Wasserstein chain rule for $\mathcal{L}_f^\lambda$)**.** *Assume $\phi \in L^2(\rho, \mathcal{C}^1)$, f satisfies Assumption III.2 with $f^* \in \mathcal{C}^1_{loc}(\mathbb{R})$ and consider $\lambda > 0$. Let $(\mu_t)_{t\in(0,+\infty)}$ be a locally absolutely continuous curve in $\mathcal{P}(\Omega)$ solution of the continuity equation Eq. (III.30) for some velocity field v such that $\|v_t\|_{L^2(\mu_t)} \in L^1_{loc}((0, +\infty))$. Then $(\mathcal{L}_f^\lambda(\mu_t))_{t\in(0,+\infty)}$ is locally absolutely continuous and for a.e. $t', t \in (0, +\infty)$:*

$$\mathcal{L}_f^\lambda(\mu_{t'}) - \mathcal{L}_f^\lambda(\mu_t) = \int_t^{t'} \left\langle \nabla \mathcal{L}_f^\lambda[\mu_s], v_s \right\rangle_{L^2(\mu_s)} \mathrm{d}s \,,$$

where for $\mu \in \mathcal{P}_2(\Omega)$ the velocity field $\nabla \mathcal{L}_f^\lambda[\mu] \in L^2(\mu)$ is defined by:

$$\forall \omega \in \Omega, \quad \nabla \mathcal{L}_f^\lambda[\mu](\omega) \coloneqq -\nabla \left(f^*(\Phi^\top \alpha_f^\lambda[\mu]) \right)(\omega) \,,$$

with $\alpha_f^\lambda[\mu]$ the maximizer in Eq. (III.19).

Proof. Consider the dual formulation of $\mathcal{L}_f^\lambda$ in Eq. (III.19). For every $\mu \in \mathcal{P}(\Omega)$ we have:

$$\mathcal{L}_f^\lambda(\mu) = \sup_{\alpha \in L^2(\rho)} \mathcal{V}_\alpha(\mu) - \frac{\lambda}{2}\|\alpha\|^2 + \langle \alpha, Y \rangle \ ,$$

where for $\alpha \in L^2(\rho)$ we defined:

$$\mathcal{V}_\alpha(\mu) \coloneqq -\int_\Omega f^*(\Phi^\top \alpha)(\omega) \mathrm{d}\mu(\omega) \,.$$

In particular, at fixed $\alpha \in L^2(\rho)$, it follows from the assumptions on ϕ and f^* that the potential $f^*(\Phi^\top \alpha)$ is in $\mathcal{C}^1(\Omega)$ with $\|f^*(\Phi^\top \alpha)\|_{\mathcal{C}^1} \leq C(\|\alpha\|_{L^2(\rho)})$ for some continuous function C. Thus, by properties of the continuity equation, $\mathcal{V}_\alpha(\mu_t)$ is absolutely continuous and its derivative is given for a.e. $t \in (0, +\infty)$ by:

$$\frac{\mathrm{d}}{\mathrm{d}t}\mathcal{V}_\alpha(\mu_t) = -\int_\Omega \left\langle \nabla f^*(\Phi^\top \alpha), v_t \right\rangle \mathrm{d}\mu_t \,.$$

Moreover, for $\mu \in \mathcal{P}(\Omega)$, using Eq. (III.21) and the fact that $\mathcal{L}_f^\lambda \leq \frac{1}{2\lambda}\|Y\|^2_{L^2(\rho)} + f(0)$ (by taking $u = 0$ in Eq. (III.13)) we have at the optimum in Eq. (III.19) that

$$\|\alpha_f^\lambda[\mu]\|_{L^2(\rho)} \leq \lambda^{-1} \left(\|Y\|^2_{L^2(\rho)} + \lambda f(0) \right)^{1/2} =: R_\lambda \,.$$

Thus, $\mathcal{L}_f^\lambda$ is equivalently defined by restricting the supremum to $\alpha \in L^2(\rho)$ such that $\|\alpha\|_{L^2(\rho)} \leq R_\lambda$. For such α we have $\left|\frac{\mathrm{d}}{\mathrm{d}t}\mathcal{V}_\alpha(\mu_t)\right| \leq C'$ for some constant $C' = C'(f^*, \lambda)$ independent of α. Thus we can apply the envelope theorem in [Milgrom, 2002, Thm. 2], which shows the desired result. □

The preceding result has defined a notion of gradient field for the functional $\mathcal{L}_f^\lambda$. One can thus define gradient flows of $\mathcal{L}_f^\lambda$ for the $\mathcal{W}_2$ metric as the curves solution to the continuity equation:

$$\partial_t \mu_t - \mathrm{div}(\mu_t \nabla \mathcal{L}_f^\lambda[\mu_t]) = 0 \quad \text{on } (0, \infty) \times \Omega. \tag{III.31}$$

We make the following definition:

Definition III.1 (Gradient flow of $\mathcal{L}_f^\lambda$)**.** *Let $\mu_0 \in \mathcal{P}_2(\Omega)$. We say $(\mu_t)_{t \geq 0}$ is a gradient flow for $\mathcal{L}_f^\lambda$ starting at μ_0 if it is a locally absolutely continuous curve on $(0, +\infty)$ s.t. $\lim_{t \to 0^+} \mu_t = \mu_0$ and if it satisfies the continuity equation Eq. (III.31) in the sense of distribution, i.e.:*

$$\int_0^\infty \int_\Omega \left(\partial_t \varphi - \nabla \varphi \cdot \nabla \mathcal{L}_f^\lambda[\mu_t] \right) \mathrm{d}\mu_t \mathrm{d}t = 0, \quad \forall \varphi \in \mathcal{C}_c^\infty((0, +\infty) \times \Omega) \,. \tag{III.32}$$

Remark III.4.1 (Boundary conditions)**.** *Note that, in the case where Ω is a closed, bounded, smooth and convex domain, our definition Eq. (III.30) of solutions to the continuity equation enforces no-flux conditions on the boundary $\partial\Omega$. Indeed we consider test function $\varphi \in \mathcal{C}_c^\infty((0,1)\times\Omega)$ that can be supported on the whole domain Ω (which is always assumed closed). Thus, Eq. (III.30) enforces $\langle \mu_t v_t, \vec{n}\rangle = 0$ in the sense of distribution, where $\vec{n}$ is the outer normal vector to the boundary $\partial\Omega$.*

In case of the gradient flow equation Eq. (III.32), this boundary condition is for example satisfied if one assumes $\langle \nabla_\omega \phi(\omega, x), \vec{n}\rangle = 0$ for every $x \in \mathbb{R}^d$ and every $\omega \in \partial\Omega$. Another way of ensuring the no-flux condition is to remove the outer part of the gradient field $\nabla \mathcal{L}_f^\lambda$ on the boundary $\partial\Omega$, which can be performed by clipping the features.

Well-posedness of the gradient flow equation To show the well-posedness of gradient flows, we rely on convexity properties of the functional $\mathcal{L}_f^\lambda$. Indeed, by the dual formulation in Eq. (III.19), we can express $\mathcal{L}_f^\lambda$ as a supremum over semiconvex functionals. As a consequence, the Lemma III.4.2 below shows that, for $\lambda > 0$, $\mathcal{L}_f^\lambda$ is semiconvex along (generalized) geodesics of the Wasserstein space (see [Ambrosio, 2008b, Def. 9.2.4] for the definition of *generalized geodesics*). However, note that such an argument can not be extended to the case $\lambda = 0$ since the semiconvexity constant blows-up when $\lambda \to 0^+$. For example, in the case $f(t) = |t|^2$ this constant scales as λ^{-2}.

Lemma III.4.2 (Geodesic semiconvexity)**.** *Assume $\phi \in L^2(\rho, \mathcal{C}^{1,1})$, f satisfies Assumption III.2 with $f^* \in \mathcal{C}^{1,1}_{loc}(\mathbb{R})$ and let $\lambda > 0$. Then $\mathcal{L}_f^\lambda$ is C-semiconvex along (generalized) geodesics for some constant $C = C(f^*, \lambda)$.*

Proof. Consider the dual formulation of $\mathcal{L}_f^\lambda$ in Eq. (III.19). For every $\mu \in \mathcal{P}(\Omega)$ we have:

$$\mathcal{L}_f^\lambda(\mu) = \sup_{\alpha \in L^2(\rho)} -\int_\Omega f^*(\Phi^\top \alpha)(\omega) \mathrm{d}\mu(\omega) - \frac{\lambda}{2}\|\alpha\|^2 + \langle \alpha, Y\rangle \ .$$

Then, at fixed $\alpha \in L^2(\rho)$, it follows from the assumptions on ϕ that $\Phi^\top \alpha \in \mathcal{C}^{1,1}(\Omega)$ with

$$\|\Phi^\top \alpha\|_{\mathcal{C}^{1,1}} \leq \|\alpha\|_{L^2(\rho)} \|\phi\|_{L^2(\rho, \mathcal{C}^{1,1})} \ .$$

Then, from the assumptions on f^*, the composition $f^*(\Phi^\top \alpha)$ is also in $\mathcal{C}^{1,1}(\Omega)$ and by [Ambrosio, 2008b, Prop.9.3.2] the functional $\mu \mapsto \int_\Omega f^*(\Phi^\top \alpha)\mathrm{d}\mu$ is C-semiconvex along generalized geodesics for some constant $C = C(f^*, \|\alpha\|_{L^2(\rho)}\|\phi\|_{L^2(\rho,\mathcal{C}^{1,1})})$. Moreover, similarly as in the proof of Lemma III.4.1, one can restrict the definition of $\mathcal{L}_f^\lambda$ to the supremum over $\alpha \in L^2(\rho)$ with $\|\alpha\|_{L^2(\rho)} \leq R_\lambda$. The result then follows by taking a supremum over (uniformly) semiconvex functionals. □

The semiconvexity of $\mathcal{L}_f^\lambda$ along generalized geodesics ensures the existence and uniqueness of gradient flows in the sense of Definition III.1.

Theorem III.1 (Well-posedness of the gradient flow equation for $\lambda > 0$)**.** *Assume the assumptions of Lemma III.4.2 hold. Then for any $\lambda > 0$ and any initialization $\mu_0 \in \mathcal{P}_2(\Omega)$ there exists a unique gradient flow for $\mathcal{L}_f^\lambda$ starting from μ_0 in the sense of Definition III.1. Moreover, if $(\mu_t)_{t\geq 0}, (\mu'_t)_{t\geq 0}$ are gradient flows for $\mathcal{L}_f^\lambda$ with respective initializations $\mu_0, \mu'_0 \in \mathcal{P}(\Omega)$ then for every $t \geq 0$:*

$$\mathcal{W}_2(\mu_t, \mu'_t) \leq e^{Ct} \mathcal{W}_2(\mu_0, \mu'_0) \, ,$$

for some constant $C = C(f^*, \lambda)$.

Proof. The chain rule formula established in Lemma III.4.1 shows that for every $\mu \in \mathcal{P}(\Omega)$ the vector field $\mathcal{L}_f^\lambda[\mu]$ is a strong subdifferential of $\mathcal{L}_f^\lambda$ in the sense of [Ambrosio, 2008b, Def. 10.3.1 and eq. (10.3.12)]. Existence, uniqueness and contractivity properties of the gradient flow then follow from the geodesic semiconvexity of $\mathcal{L}_f^\lambda$ established in Lemma III.4.2 and the application of [Ambrosio, 2008b, Def. 11.2.1] □

Finally, it is a classical property of weak solutions to continuity equations that gradient flows of $\mathcal{L}_f^\lambda$ can be represented in terms of push-forward by a flow map.

Proposition III.4.1. *Let the assumptions of Lemma III.4.2 hold. Then, for any $\lambda > 0$ and any initialization $\mu_0 \in \mathcal{P}_2(\Omega)$, the gradient flow $(\mu_t)_{t\geq 0}$ of $\mathcal{L}_f^\lambda$ starting from μ_0 satisfies $\mu_t = (X_t)_\# \mu_0$ for every $t \geq 0$, where $(X_t)_{t\geq 0}$ is the flow-map solution of the ODE:*

$$\forall t \geq 0, \quad \frac{\mathrm{d}}{\mathrm{d}t} X_t = -\nabla \mathcal{L}_f^\lambda[\mu_t] \circ X_t, \quad \textit{with} \quad X_0 = \mathrm{Id}_\Omega \,.$$

In particular, if $(\omega_i(t))_{t\geq 0}$ for $i \in \{1, ..., M\}$ are solutions to Eq. (III.26) then the empirical distribution $\hat{\mu}_t := \frac{1}{M}\sum_{i=1}^M \delta_{\omega_i(t)}$ is a gradient flow for $\mathcal{L}_f^\lambda$ in the sense of Definition III.1 and thus $\omega_i(t) = X_t(\omega_i(0))$ for $i \in \{1, ..., M\}$ and $t \geq 0$.

Proof. For every $t \geq 0$, similarly as in the proof of Lemma III.4.1, we have that the dual variable is bounded by $\|\alpha_f^\lambda[\mu_t]\|_{L^2(\rho)} \leq R_\lambda$ and from the assumption on the regularity of ϕ it follows that:

$$\|f^*(\Phi^\top \alpha_f^\lambda[\mu_t])\|_{\mathcal{C}^{1,1}} \leq C \,.$$

for some constant $C = C(f^*, \lambda)$. Then by definition $\nabla \mathcal{L}^\lambda[\mu_t] = -\nabla f^*(\Phi^\top \alpha_f^\lambda[\mu_t]) \in \mathcal{C}^{0,1}$ and the first part of the result follows from classical results of ODE theory [Hale, 2009] and on representation of solutions to continuity equations [Ambrosio, 2008b, Thm. 8.1.8]. For the second part of the result, it suffices to remark that, by the definition of $\hat{\mathcal{L}}_f^\lambda$ and $\nabla \mathcal{L}_f^\lambda$, for $\{\omega_i\}_{1\leq i\leq M} \in \Omega^M$ and $j \in \{1, ..., M\}$:

$$M \nabla_{\omega_j} \hat{\mathcal{L}}^\lambda(\{\omega_i\}_{1\leq i\leq M}) = \nabla \mathcal{L}_f^\lambda[\hat{\mu}](\omega_j) \,, \tag{III.33}$$

where $\hat{\mu} = \frac{1}{M}\sum_{i=1}^M \delta_{\omega_i}$. Therefore, by Eq. (III.26) we have that for any test function $\varphi \in \mathcal{C}_c^\infty((0,\infty) \times \Omega)$:

$$0 = \frac{1}{M}\sum_{i=1}^M \int_0^\infty \frac{\mathrm{d}}{\mathrm{d}t}\varphi(t, \omega_i(t))\mathrm{d}t = \int_0^\infty \int_\Omega \left(\partial_t \varphi - \nabla\varphi \cdot \nabla \mathcal{L}_f^\lambda[\hat{\mu}_t]\right) \mathrm{d}\hat{\mu}_t \mathrm{d}t \,,$$

meaning $(\hat{\mu}_t)_{t\geq 0}$ is a gradient flow for $\mathcal{L}_f^\lambda$ according to Definition III.1. □

Particle approximation In the case where $\lambda > 0$, associating the contraction rate of the gradient flow obtained in Theorem III.1 with classical results on the approximation of measures by empirical distributions we obtain an approximation result for the minimization of $\mathcal{L}_f^\lambda$ with a finite number of features. For conciseness, we only state the result in the case $d \geq 3$, but similar results hold for $d \in \{1, 2\}$.

Corollary III.4.1 (Particle approximation)**.** *Let the assumptions of Lemma III.3.1 hold and let $d \geq 3$. Consider some initialization $\mu_0 \in \mathcal{P}(\Omega)$ and, for some $N \geq 0$, denote by $\hat{\mu}_0 := N^{-1}\sum_{i=1}^N \delta_{\omega_i}$ the empirical measure where $\{\omega_i\}_{1\leq i\leq N}$ are i.i.d. samples of μ_0. For $\lambda > 0$, let $(\mu_t^\lambda)_{t\geq 0}$ and $(\hat{\mu}_t^\lambda)_{t\geq 0}$ be the gradient flow of $\mathcal{L}_f^\lambda$ starting from μ_0 and $\hat{\mu}_0$*

respectively. Then there exists a constant $A = A(d, \Omega)$ s.t. for every $t \geq 0$ and every $\varepsilon > 0$:

$$\mathbb{P}\left(\mathcal{W}_1(\hat{\mu}_t^\lambda, \mu_t^\lambda) \geq \varepsilon\right) \leq \frac{A}{\varepsilon} N^{-1/d} e^{Ct},$$

where $C = C(f^*, \lambda)$ is the constant in Theorem III.1.

Proof. Using [Fournier, 2015, Thm. 1] we obtain at initialization $t = 0$:

$$\mathbb{E}\left[\mathcal{W}_1(\hat{\mu}_0^\lambda, \mu_0^\lambda)\right] \leq A N^{-1/d}$$

for some constant $A = A(d, \Omega) > 0$ depending on the dimension and on the domain Ω. Then using the contraction rate in Theorem III.1 we have a constant $C = C(f^*, \lambda) > 0$ such that for every $t \geq 0$:

$$\mathbb{E}\left[\mathcal{W}_1(\hat{\mu}_t^\lambda, \mu_t^\lambda)\right] \leq A N^{-1/d} e^{Ct}.$$

The result then follows by applying Markov's inequality. □

III.4.2 Wasserstein gradient flows in the case $\lambda = 0$ and ultra-fast diffusions

We now consider the limit of the proximal scheme Eq. (III.28) when the step size τ tends to 0 and λ is set to 0. We focus on the case where Assumption III.1 holds and the regularization is of the form $f(t) = |t|^r/(r-1)$ for some $r > 1$ and recall that we use the shortcut $\mathcal{L}_r^0 \coloneqq \mathcal{L}_f^0$. Then, following Eq. (III.17), we have for $\mu \in \mathcal{P}(\Omega)$:

$$\mathcal{L}_r^0(\mu) = \frac{1}{r-1} \mathrm{D}_r(\bar{\nu}|\mu) = \frac{1}{r-1} \int_\Omega \left|\frac{\mathrm{d}\bar{\nu}}{\mathrm{d}\mu}\right|^r \mathrm{d}\mu = \frac{\|\bar{\nu}\|_{\mathrm{TV}}^r}{r-1} \int_\Omega \left|\frac{\mathrm{d}\bar{\mu}}{\mathrm{d}\mu}\right|^r \mathrm{d}\mu. \tag{III.34}$$

The first variation of $\mathcal{L}_r^0$ w.r.t. μ is formally given by $\frac{\delta \mathcal{L}_r^0}{\delta \mu}[\mu](\omega) = -\|\bar{\nu}\|_{\mathrm{TV}}^r \left(\frac{\bar{\mu}}{\mu}\right)^r$ and thus, following Eq. (III.27), the Wasserstein gradient flow of $\mathcal{L}_r^0$ is formally defined as the solution to the continuity equation:

$$\partial_t \mu_t = -\|\bar{\nu}\|_{\mathrm{TV}}^r \mathrm{div}\left(\mu_t \nabla \left(\frac{\bar{\mu}}{\mu_t}\right)^r\right). \tag{III.35}$$

Moreover, calculating formally, $\nabla \left(\frac{\bar{\mu}}{\mu}\right)^r = r \frac{\bar{\mu}}{\mu} \nabla \left(\frac{\bar{\mu}}{\mu}\right)^{r-1}$ and Eq. (III.35) can be written equivalently:

$$\partial_t \mu_t = -r\|\bar{\nu}\|_{\mathrm{TV}}^r \mathrm{div}\left(\bar{\mu} \nabla \left(\frac{\bar{\mu}}{\mu_t}\right)^{r-1}\right).$$

When the target distribution is uniform, i.e. with density $\bar{\mu} = 1$, this corresponds to a nonlinear diffusion equation of the form Eq. (III.11) with the coefficient $m = 1 - r < 0$, that is an *ultra-fast diffusion*. Such an equation, with general inhomogeneous weights $\bar{\mu}$ was studied in [Iacobelli, 2019a; Caglioti, 2018; Iacobelli, 2019b] in the context of particle algorithms for finding an optimal quantization of the measure $\bar{\mu}$. We rely particularly here on the work of Iacobelli, Patacchini, and Santambrogio [Iacobelli, 2019b] which establishes the well-posedness of Eq. (III.35) as well as the convergence of the solution μ_t towards the target measure $\bar{\mu}$. We consider the following definition of solutions for Eq. (III.35):

Definition III.2 (Gradient flow of $\mathcal{L}_r^0$ (Def. 1.1 in [Iacobelli, 2019b])). *Let $\mu_0 \in \mathcal{P}(\Omega)$ admit a density $\mu_0 \in L^{r+2}(\Omega)$. We say $(\mu_t)_{t\geq 0}$ is a weak solution of Eq. (III.35) or a gradient flow for $\mathcal{L}_r^0$ starting from μ_0 if it is a narrowly continuous curve in $\mathcal{P}(\Omega)$ with $\lim_{t\to 0^+} \mu_t = \mu_0$, s.t.*

$$\int_0^\infty \int_\Omega \left(\partial_t \varphi - \|\bar{\nu}\|_{\mathrm{TV}}^r \nabla\varphi \cdot \nabla \left(\frac{\bar{\mu}}{\mu_t} \right)^r \right) \mathrm{d}\mu_t \mathrm{d}t = 0, \quad \forall \varphi \in \mathcal{C}_c^\infty((0,\infty)\times\Omega)\,. \tag{III.36}$$

and satisfying:

$$\left(\frac{\mu_t}{\bar{\mu}}\right)^{r-1} \in L^2_{loc}((0,\infty), H^1(\Omega)), \quad \frac{\bar{\mu}}{\mu_t} \in L^2_{loc}((0,\infty), H^1(\Omega))\,.$$

Existence and uniqueness of solutions In [Iacobelli, 2019b], the authors establish the existence and uniqueness of gradient flows for the functional $\mathcal{L}_r^0$. More precisely, they show that, under appropriate assumptions on the initialization μ_0 and on the target $\bar{\mu}$, the iterates of the proximal scheme in Eq. (III.28) converge towards a curve $(\mu_t)_{t\geq 0}$ that is a gradient flow of the functional $\mathcal{L}_r^0$ in the sense of Definition III.2.

Theorem III.2 ([Iacobelli, 2019b, Thm. 1.2]). *Assume μ_0 and $\bar{\mu}$ are absolutely continuous and have bounded log-densities. Then there exists a unique weak solution of Eq. (III.35) starting from μ_0 in the sense of Definition III.2.*

Convergence towards the target distribution In the case $\lambda = 0$, Iacobelli, Patacchini, and Santambrogio [Iacobelli, 2019b] establish a linear convergence rate of the *weighted ultra-fast diffusion* Eq. (III.35) towards the target distribution $\bar{\mu}$. Precisely, they show convergence in the L^2-sense of the density μ_t towards the target density $\bar{\mu}$. We state their result in the following theorem.

Theorem III.3 ([Iacobelli, 2019b, Thm. 1.4]). *Assume μ_0 and $\bar{\mu}$ are absolutely continuous and have bounded log-densities. For $\mu_0 \in \mathcal{P}(\Omega)$, let $(\mu_t)_{t\geq 0}$ be a weak solution of Eq. (III.35) starting from μ_0 in the sense of Definition III.2. Then the log-density of μ_t is bounded, uniformly over $t \geq 0$, and there exists a constant $C = C(\Omega, \bar{\mu}, \mu_0) > 0$ s.t. for every $t \geq 0$ it holds:*

$$\|\bar{\mu} - \mu_t\|_{L^2(\Omega)} \leq Ce^{-Ct}\,.$$

For completeness, we give here some of the key arguments of the proof of the above Theorem III.3 in the case where $r = 2$. In this case, we have for every $\mu \in \mathcal{P}(\Omega)$:

$$\mathcal{L}_2^0(\mu) = \|\bar{\nu}\|_{\mathrm{TV}}^2 \int_\Omega \left| \frac{\mathrm{d}\bar{\mu}}{\mathrm{d}\mu} \right|^2 \mathrm{d}\mu = \|\bar{\nu}\|_{\mathrm{TV}}^2 \left(\chi^2(\bar{\mu}|\mu) + 1 \right)\,, \tag{III.37}$$

where χ^2 is the chi-square divergence. The following Lemma III.4.3 establishes the desired linear convergence rate for the proximal scheme defined in Eq. (III.28) with the loss $\mathcal{L}_2^0$. The result in continuous time then follows from the lower semicontinuity of the χ^2-divergence as the curve $(\mu_t)_{t\geq 0}$ is obtained by taking the limit of the discrete process $(\mu_k^\tau)_{k\geq 0}$ when the discretization time τ tends to zero.

From a technical perspective, the proof of Lemma III.4.3 relies on a Poincaré inequality satisfied by μ_t. It is indeed well-known that such inequality controls the convergence rate of Fokker-Planck equations towards their stationary distribution in χ^2-distance [Pavliotis, 2014, Thm. 4.4]. This can for example be used to prove the convergence of sampling

algorithms such as *Langevin Monte Carlo* [Chewi, 2024; Chewi, 2020]. In our case, the ultra-fast diffusion Eq. (III.35) is to be interpreted as a Wasserstein gradient flow for $\mathcal{L}_2^0$, which, by the above Eq. (III.37), is the *reverse* χ^2-divergence between μ_t and $\bar{\mu}$ and the convergence rate is controlled by the Poincaré constant of μ_t. This rate may thus a priori evolve and vanish during training but, crucially, [Iacobelli, 2019b, Lem. 2.4] shows that it is here a property of solutions to the ultra-fast diffusion equation that the log-density ratio $\| \log \left(\frac{\bar{\mu}}{\mu_t} \right) \|_\infty$ decreases with time. As a consequence, it is sufficient to assume that the log-density is bounded at initialization to obtain a control over the Poincaré constant of μ_t, for $t \geq 0$, by a classical perturbation argument [Ané, 2000, Thm. 3.4.1].

Lemma III.4.3. *Assume μ_0 and $\bar{\mu}$ are absolutely continuous with bounded log-densities. Let $\tau > 0$ and let $(\mu_k^\tau)_{k \geq 0}$ be the sequence defined by Eq. (III.28) with $\lambda = 0$, $f(t) = |t|^2$ ($r = 2$) and initialization $\mu_0^\tau = \mu_0 \in \mathcal{P}(\Omega)$. Then there exists a constant $C > 0$ s.t.:*

$$\forall k \geq 0, \quad \chi^2(\bar{\mu} | \mu_k^\tau) \leq (1 + C\tau)^{-k} \chi^2(\bar{\mu} | \mu_0) \, .$$

Proof. From [Iacobelli, 2019b, Thm.2.1 and Lem.2.4] we know the sequence $(\mu_k^\tau)_{k \geq 0}$ is uniquely defined. Moreover μ_k^τ is absolutely continuous w.r.t. Lebesgue measure and their exists a constant $C = C(\bar{\mu}, \mu_0) > 0$ s.t. the log-densities $\log(\mu_k^\tau)$ satisfy:

$$\forall k \geq 0, \quad \| \log(\mu_k^\tau) \|_\infty \leq C \, .$$

Then, at step $k \geq 0$, we get from the expression of $\mathcal{L}_2^0$ in Eq. (III.37) and from the optimality condition in Eq. (III.28) (see e.g. [Santambrogio, 2015, Prop.7.20]) that:

$$- \| \bar{\nu} \|_{\mathrm{TV}}^2 \left(\frac{\bar{\mu}}{\mu_{k+1}^\tau} \right)^2 + \frac{\varphi}{\tau} = cte, \quad \text{almost everywhere on } \Omega,$$

where φ is the Kantorovitch potential from μ_{k+1}^τ to μ_k^τ. Also this potential is necessarily Lipschitz, hence a.e. differentiable and so is $\bar{\nu}/\mu_{k+1}^\tau$. Then from the definition of $\mu_{k+1}^{0,\tau}$ we have:

$$\begin{aligned}
\mathcal{L}_2^0(\mu_k^\tau) - \mathcal{L}_2^0(\mu_{k+1}^\tau) &\geq \frac{1}{2\tau} \mathcal{W}_2(\mu_{k+1}^\tau, \mu_k^\tau)^2 \\
&= \frac{1}{2\tau} \int_\Omega |\nabla \varphi|^2 \, \mathrm{d}\mu_{k+1}^\tau \\
&= \frac{\tau \| \bar{\nu} \|_{\mathrm{TV}}^2}{2} \int_\Omega \left\| \nabla \left(\frac{\bar{\mu}}{\mu_{k+1}^\tau} \right)^2 \right\|^2 \mathrm{d}\mu_{k+1}^\tau \, ,
\end{aligned}$$

where we used the definition of the potential φ and the optimality condition. Using that μ_{k+1}^τ has log-density bounded by $C = C(\bar{\mu}, \mu_0)$ and that the domain Ω satisfies a Poincaré inequality with constant $C_P = C_P(\Omega)$, it follows from a classical perturbation argument that μ_{k+1}^τ satisfies a Poincaré inequality with constant $e^{2C} C_P(\Omega)$ [Ané, 2000, Thm. 3.4.1]. As a consequence:

$$\begin{aligned}
\int_\Omega \left\| \nabla \left(\frac{\bar{\mu}}{\mu_{k+1}^\tau} \right)^2 \right\|^2 \mathrm{d}\mu_{k+1}^\tau &\geq 4 e^{-4C} \int_\Omega \left\| \nabla \left(\frac{\bar{\mu}}{\mu_{k+1}^\tau} \right) \right\|^2 \mathrm{d}\mu_{k+1}^\tau \\
&\geq 4 C_P^{-1} e^{-6C} \left(\int_\Omega \left(\frac{\bar{\mu}}{\mu_{k+1}^\tau} \right)^2 \mathrm{d}\mu_{k+1}^\tau - 1 \right) \\
&= 4 C_P^{-1} e^{-6C} \| \bar{\nu} \|_{\mathrm{TV}}^{-2} \left(\mathcal{L}_2^0(\mu_{k+1}^\tau) - \| \bar{\nu} \|_{\mathrm{TV}}^2 \right) \, ,
\end{aligned}$$

where $\|\bar{\nu}\|_{\mathrm{TV}}^2 = \inf \mathcal{L}_2^0$. Combining this with the previous inequality finally gives:

$$\left(1 + 2\tau C_P^{-1} e^{-6C}\right)\left(\mathcal{L}_2^0(\mu_{k+1}^\tau) - \inf \mathcal{L}_2^0\right) \leq \mathcal{L}_2^0(\mu_k^\tau) - \inf \mathcal{L}_2^0\,,$$

and inductively:

$$\forall k \geq 0, \quad \mathcal{L}_2^0(\mu_k^\tau) - \inf \mathcal{L}_2^0 \leq \left(1 + 2\tau C_P^{-1} e^{-6C}\right)^{-k}\left(\mathcal{L}_2^0(\mu_0) - \inf \mathcal{L}_2^0\right)\,.$$

By the definition of $\mathcal{L}_2^0$ in Eq. (III.37), this is the desired result. □

Remark III.4.2 (Dependence of the convergence rate w.r.t. the dimension)**.** *It follows from the proof that the convergence rate C in Lemma III.4.3 scales linearly with $C_P(\Omega)^{-1}$ where $C_P(\Omega)$ is the Poincaré constant of the domain Ω. For bounded, Lipschitz and convex domains of $\mathbb{R}^n$ or for the flat torus $\mathbb{T}^n$, this constant is in particular independent of the dimension n [Payne, 1960]. Therefore, the correspondence between the training of neural networks in the two-timescale regime and solutions to ultra-fast diffusions points towards the fact that gradient methods, with suitable hyperparameter scaling, are amenable to efficient feature learning in the training of neural networks, without suffering from the curse of dimensionality [Donoho, 2000]. Note however that the convergence rate C in Lemma III.4.3 is exponentially bad in the log-density ratio $\|\log(\bar{\mu}/\mu_0)\|_\infty$. In particular the convergence rate does not hold in case the teacher feature distribution is supported on a finite number of atoms.*

III.5 Convergence of gradient flow

The main purpose of this chapter is to study in what extent the gradient flow dynamics defined Definitions III.1 and III.2 allow recovering the teacher feature distribution $\bar{\mu}$ associated to the observed signal Y in Assumption III.1. Whereas Theorem III.3 shows convergence of the gradient flow of $\mathcal{L}_f^0$, that is solutions to the *ultra-fast diffusion* Eq. (III.35), such dynamics are hardly numerically tractable in practice due to the absence of the regularization parameter λ. For this reason we are interested here in the asymptotic behavior of the gradient flow of $\mathcal{L}_f^\lambda$ in the case where $\lambda > 0$. A difficulty is that, in the case $\lambda = 0$, the proof of Theorem III.3 relies on the implicit behavior of Eq. (III.35) which preserves the density of solutions. Such a behavior is a priori not expected to hold when $\lambda > 0$. As a consequence, the results in this section hold under supplementary regularity assumptions on the solutions to Eq. (III.31).

III.5.1 Algebraic convergence rate

At fixed $\lambda > 0$, we are able to obtain convergence towards the minimizer $\bar{\mu}_f^\lambda$ of $\mathcal{L}_f^\lambda$ under mild regularity assumptions on solutions to the gradient flow Eq. (III.31). Specifically, for a probability measure $\mu \in \mathcal{P}(\Omega)$ and a function $h \in \mathcal{C}^1$ we define the weighted Sobolev seminorm of h as:

$$\|h\|_{\dot{H}^1(\mu)} \coloneqq \left(\int_\Omega \|\nabla h\|^2 \mathrm{d}\mu\right)^{1/2}\,.$$

Then, for a measure $\nu \in \mathcal{M}(\Omega)$ s.t. $\int_\Omega \mathrm{d}\nu = 0$, the negative weighted Sobolev seminorm $\|\nu\|_{\dot{H}^{-1}(\mu)}$ is defined by duality with $\dot{H}^1(\mu)$:

$$\|\nu\|_{\dot{H}^{-1}(\mu)} \coloneqq \sup_{\|h\|_{\dot{H}^1(\mu)} \leq 1} \int_\Omega h \mathrm{d}\nu\,.$$

The following Theorem III.4 states convergence of $\mathcal{L}_f^\lambda$ towards 0 with an algebraic convergence rate provided $\|\mu_t - \frac{\bar\nu}{\bar m}\|_{\dot H^{-1}(\mu_t)}$ stays bounded along the gradient flow. As discussed below, since the domain Ω is compact, this assumption is satisfied for example when both distribution have bounded log-densities. The arguments are similar to the one presented in [Glaser, 2021], where the authors consider an infimal convolution between a kernel discrepancy and the KL-divergence. Importantly, the obtained convergence rate depends on the bound on $\|\mu_t - \frac{\bar\nu}{\bar m}\|_{\dot H^{-1}(\mu_t)}$ but is independent of $\lambda > 0$.

Theorem III.4. *Let Assumption III.1 hold. Consider $\lambda > 0$ and some initialization $\mu_0 \in \mathcal{P}(\Omega)$. Let $(\mu_t)_{t\geq 0}$ be the gradient flow of $\mathcal{L}_f^\lambda$ starting from μ_0 in the sense of Eq. (III.31). Assume that:*

- *$\bar\nu$ is a positive measure and f is s.t. $\min f = f(\bar m) = 0$ where $\bar m := \bar\nu(\Omega) > 0$.*
- *the gradient flow $(\mu_t)_{t\geq 0}$ is s.t. $\|\frac{\bar\nu}{\bar m} - \mu_t\|_{\dot H^{-1}(\mu_t)}$ is bounded, uniformly over $t \geq 0$.*

Then there exists a constant $C > 0$ s.t. for any $t \geq 0$:

$$\mathcal{L}_f^\lambda(\mu_t) \leq \left(\mathcal{L}_f^\lambda(\mu_0)^{-1} + Ct\right)^{-1} .$$

In particular, μ_t converges to $\bar\mu = \bar\nu/\bar m$ when t tends to $+\infty$.

Proof. Note that it follows from the assumptions on f and $\bar\nu$ that $\inf \mathcal{L}_f^\lambda = 0$ and that this infimum is attained only for $\mu = \bar\nu/\bar m$. Thus, the last statement on the convergence of μ_t follows from the convergence of $\mathcal{L}_f^\lambda(\mu_t)$ to 0 and from the lower semicontinuity of $\mathcal{L}_f^\lambda$ (see Section III.3).

To obtain the convergence rate, consider $\mu \in \mathcal{P}(\Omega)$ and note that by Eq. (III.19) we have $\mathcal{L}_f^\lambda(\mu) = \max_\alpha \mathcal{K}(\alpha, \mu)$, where for every $\alpha \in L^2(\rho)$:

$$\mathcal{K}(\alpha, \mu) := \int_\Omega (\Phi^\top \alpha) \mathrm{d}\bar\nu - \int_\Omega f^*(\Phi^\top \alpha) \mathrm{d}\mu - \frac{\lambda}{2}\|\alpha\|^2_{L^2(\rho)} ,$$

where f^* is the Legendre transform of f. Let us denote by $\alpha^\lambda = \alpha_f^\lambda[\mu]$ the maximizer of $\mathcal{K}(\alpha, \mu)$. Then, using the convexity of f^*, we have for every $\omega \in \Omega$:

$$f^*(0) + \partial f^*(0)(\Phi^\top \alpha^\lambda)(\omega) \leq f^*((\Phi^\top \alpha^\lambda)(\omega)) .$$

Also by assumption $\partial f(\bar m) = 0$ and hence by properties of the Legendre transform $\partial f^*(0) = \bar m$. Also $f^*(0) = -f(\bar m) = 0$ and after integrating w.r.t. $\bar\nu$:

$$\int_\Omega (\Phi^\top \alpha^\lambda) \mathrm{d}\bar\nu \leq \int_\Omega f^*(\Phi^\top \alpha^\lambda) \frac{\mathrm{d}\bar\nu}{\bar m} .$$

Then, replacing α by α^λ in $\mathcal{K}$ and using the previous inequality:

$$\mathcal{L}_f^\lambda(\mu) = \mathcal{K}(\alpha^\lambda, \mu) \leq \int_\Omega f^*(\Phi^\top \alpha^\lambda) \mathrm{d}(\frac{\bar\nu}{\bar m} - \mu) \leq \left\|f^*(\Phi^\top \alpha^\lambda)\right\|_{\dot H^1(\mu)} \|\frac{\bar\nu}{\bar m} - \mu\|_{\dot H^{-1}(\mu)} .$$

Also, by the gradient flow equation Eq. (III.31), the dissipation of $\mathcal{L}_f^\lambda$ along the gradient flow curve $(\mu_t)_{t\geq 0}$ is given for every $t \geq 0$ by:

$$\frac{\mathrm{d}}{\mathrm{d}t}\mathcal{L}_f^\lambda(\mu_t) = -\int_\Omega \left\|\nabla(f^*(\Phi^\top \alpha_t^\lambda))\right\|^2 \mathrm{d}\mu_t = -\left\|f^*(\Phi^\top \alpha_t^\lambda)\right\|^2_{\dot H^1(\mu_t)} ,$$

where $\alpha_t^\lambda = \alpha_f^\lambda[\mu_t]$ maximizes $\mathcal{K}(\alpha, \mu_t)$. Thus using the previous inequality on $\mathcal{L}_f^\lambda(\mu_t)$ and that $\|\frac{\bar\nu}{\bar m} - \mu_t\|_{\dot H^{-1}(\mu_t)}$ is bounded, uniformly over $t \geq 0$, we get for every $t \geq 0$:

$$\frac{\mathrm{d}}{\mathrm{d}t}\mathcal{L}_f^\lambda(\mu_t) \leq -C^{-1}\mathcal{L}_f^\lambda(\mu_t)^2\,,$$

for some constant $C > 0$. The desired convergence rate follows from this inequality by applying a Grönwall lemma. □

Let us comment on the assumptions of Theorem III.4. The second assumption specifically, is automatically satisfied in case $\bar\nu$ has bounded density and μ_t has bounded log-density, uniformly over $t \geq 0$. Indeed, for $\mu \in \mathcal{P}(\Omega)$ having a lower-bounded log-density, we have that the *weighted* Sobolev seminorm $\|.\|_{\dot H^1(\mu)}$ is lower-bounded by the classical Sobolev seminorm $\|.\|_{\dot H^1(\pi)}$, where we recall that π is the (normalized) Lebesgue measure over Ω. Precisely, if $\pi \ll \mu$ and $\mathrm{d}\pi/\mathrm{d}\mu \leq C_1$ then for every $f \in \mathcal{C}^1$:

$$\|f\|_{\dot H^1(\pi)} \leq C_1\|f\|_{\dot H^1(\mu)}\,.$$

In this case, the weighted *negative* Sobolev seminorm $\|.\|_{\dot H^{-1}(\mu)}$ is upper-bounded by the seminorm $\|.\|_{\dot H^{-1}(\pi)}$ and for every $\nu \in \mathcal{M}(\Omega)$ with $\int_\Omega \mathrm{d}\nu = 0$ we have:

$$\|\nu\|_{\dot H^{-1}(\mu)} \leq C_1\|\nu\|_{\dot H^{-1}(\pi)}\,.$$

Moreover, this last quantity can be estimated by the Wasserstein distance. Indeed, for probability measures having bounded log-densities, the Wasserstein distance $\mathcal{W}_2$ is equivalent to the negative Sobolev seminorm $\|.\|_{\dot H^{-1}(\pi)}$. If $\mu, \nu \in \mathcal{P}(\Omega)$ are such that $\frac{\mathrm{d}\mu}{\mathrm{d}\pi}, \frac{\mathrm{d}\nu}{\mathrm{d}\pi} \leq C_2$ for some constant $C_2 > 0$ we have [Santambrogio, 2015, Lem. 5.33 and Thm. 5.34]:

$$\|\mu - \nu\|_{\dot H^{-1}(\pi)} \leq C_2^{1/2}\mathcal{W}_2(\mu, \nu)\,.$$

Finally, the Wasserstein distance $\mathcal{W}_2(\mu, \nu)$ is always bounded by $\mathrm{diam}(\Omega)$ which is finite, hence ensuring the second assumption of Theorem III.4 is satisfied.

III.5.2 Convergence to ultra-fast diffusion.

The algebraic convergence rate stated in the above Theorem III.4 in the case $\lambda > 0$ stands in contrast with the faster linear convergence stated in Theorem III.3 in the case $\lambda = 0$. For this reason, we are interested in comparing the gradient flow dynamics with and without regularization.

Below we assume $f(t) = |t|^r/(r-1)$ for some $r > 1$ and Theorem III.5 shows local uniform in time convergence of gradient flows of $\mathcal{L}_r^\lambda$ to gradient flows of $\mathcal{L}_r^0$, i.e. solutions to the ultra-fast diffusion equation Eq. (III.34), when the regularization strength λ vanishes. To obtain such a result we assume regularity on the density ratio $\frac{\mathrm{d}\bar\nu}{\mathrm{d}\mu_t^\lambda}$. Namely, we assume that the Legendre-conjugate $\partial f(\frac{\mathrm{d}\bar\nu}{\mathrm{d}\mu_t^\lambda})$ stays bounded in the RKHS $\mathcal{H}$, defined as the image of the convolution operator $\Phi^\top : L^2(\rho) \to \mathcal{C}^0(\Omega)$ (Eq. (III.47)). Using classical results from the theory of inverse problems, such a *source condition* ensures the dual variable $\alpha \in L^2(\rho)$ stays uniformly bounded for $\lambda > 0$ (Lemma III.5.1). Therefore, provided $\mathcal{H}$ is sufficiently regular, such a regularity assumption ensures compactness of the Wasserstein gradient $\nabla\mathcal{L}_r^\lambda[\mu_t] = \nabla f^*(\Phi^\top\alpha_t^\lambda)$ in $\mathcal{C}^1$ and allows passing to the limit in Eq. (III.32) to obtain Eq. (III.36).

Theorem III.5. *Assume Assumption III.1 hold with $\bar{\nu}$ a positive measure with bounded log-density, $f(t) = |t|^r/(r-1)$ for some $r > 1$ and the assumptions of Theorem III.1 and Theorem III.2 are satisfied. Consider some initialization $\mu_0 \in \mathcal{P}(\Omega)$ s.t. μ_0 has bounded log-density. For $\lambda \geq 0$, let $(\mu_t^\lambda)_{t\geq 0}$ be the gradient flow of $\mathcal{L}_r^\lambda$, starting from μ_0 in the sense of Definition III.1 (when $\lambda > 0$) and Definition III.2 (when $\lambda = 0$). Moreover, for $\mathcal{H}$ defined by Eq. (III.47), assume $\mathcal{H}$ is compactly embedded in $\mathcal{C}^1(\Omega)$ and $\partial f(\frac{\mathrm{d}\bar{\nu}}{\mathrm{d}\mu_t^\lambda})$ is bounded in $\mathcal{H}$, locally uniformly over $t \geq 0$ and uniformly over $\lambda > 0$. Then for any $T \geq 0$:*

$$\lim_{\lambda \to 0^+} \sup_{t\in[0,T]} \mathcal{W}_2(\mu_t^0, \mu_t^\lambda) = 0\,.$$

Proof. For $\lambda > 0$, the curves $(\mu_t)^\lambda$ are gradient flows for the functionals $\mathcal{L}_r^\lambda$ and classical computations show that for every $t, s \geq 0$:

$$\mathcal{W}_2(\mu_t^\lambda, \mu_s^\lambda)^2 \leq |t-s| \left| \mathcal{L}_r^\lambda(\mu_t^\lambda) - \mathcal{L}_r^\lambda(\mu_s^\lambda) \right| \leq |t-s| \mathcal{L}_r^0(\mu_0)\,,$$

where we used that the functionals $\mathcal{L}_r^\lambda$ converge pointwise from below to $\mathcal{L}_r^0$. Thus, for $T \geq 0$, the sequence $(\mu_t^\lambda)_{t\in[0,T]}$ is uniformly equicontinuous with value in the compact space $\mathcal{P}(\Omega)$ and Arzela-Ascoli's theorem ensures the existence of a subsequence $\lambda_n \to 0^+$ s.t.:

$$(\mu_t^\lambda)_{t\in[0,T]} \xrightarrow{n\to\infty} (\mu_t)_{t\in[0,T]} \in \mathcal{C}^0([0,T], \mathcal{P}(\Omega))\,.$$

To prove the result one needs to identify μ_t with μ_t^0 and the supplementary regularity assumptions on μ_t^λ are sufficient for this purpose. Let us fix some $t \in [0,T]$ and denote by $u_t^\lambda = u_f^\lambda[\mu_t^\lambda]$ the minimizer in Eq. (III.13), $\nu_t^\lambda \in \mathcal{M}(\Omega)$ the minimizer in Eq. (III.18) s.t. $\frac{\mathrm{d}\nu_t^\lambda}{\mathrm{d}\mu_t^\lambda} = u_t^\lambda$ and $\alpha_t^\lambda \in L^2(\rho)$ the maximizer in Eq. (III.19).

Then for every $\lambda > 0$, since μ_0 has bounded log-density we have by the flow-map representation in Proposition III.4.1 that μ_t^λ has bounded log-density. Also, since $\bar{\nu}$ is positive with bounded log-density and $\Phi\star$ is injective, we have that $u_t^\dagger := \frac{\mathrm{d}\bar{\nu}}{\mathrm{d}\mu_t^\lambda}$ is the unique solution to Eq. (III.14). But then, by the characterization of the RKHS $\mathcal{H}$ in Theorem III.6, we have that $\Phi^\top : L^2(\rho) \to \mathcal{H}$ is a partial isometry and the assumption that $\partial f(\frac{\mathrm{d}\bar{\nu}}{\mathrm{d}\mu_t^\lambda}) \in \mathcal{H}$ is equivalent to a source condition of the form Eq. (III.38). Hence by Lemma III.5.1, the dual variable α_t^λ is bounded in $L^2(\rho)$, uniformly over $\lambda > 0$, which implies that, up to extraction of a subsequence, $\Phi^\top \alpha_t^\lambda$ converges to some h_t in $\mathcal{C}^1(\Omega)$.

Also for every $\lambda > 0$, by the duality relations in Eq. (III.20), we have $\frac{\mathrm{d}\nu_t^\lambda}{\mathrm{d}\mu_t^\lambda} = \partial f^*(\Phi^\top \alpha_t^\lambda)$ and hence — recalling that $f(t) = |t|^r/(r-1)$ for some $r > 1$ — $\frac{\mathrm{d}\nu_t^\lambda}{\mathrm{d}\mu_t^\lambda} \to \partial f^*(h_t)$ in $\mathcal{C}^0(\Omega)$. Since $\mathcal{L}_r^\lambda(\mu_t^\lambda) \leq \mathcal{L}_r^0(\mu_0)$ is bounded we have by Eq. (III.18) that $\nu_t^\lambda \to \bar{\nu}$ narrowly and then for every $\varphi \in \mathcal{C}^0(\Omega)$:

$$\int_\Omega \varphi \mathrm{d}\nu_t^\lambda = \int_\Omega \varphi \frac{\mathrm{d}\nu_t^\lambda}{\mathrm{d}\mu_t^\lambda} \mathrm{d}\mu_t^\lambda \xrightarrow{\lambda\to 0^+} \int_\Omega \varphi \mathrm{d}\bar{\nu} = \int_\Omega \varphi \partial f^*(h_t) \mathrm{d}\mu_t\,.$$

This shows that $\bar{\nu}$ is absolutely continuous w.r.t. μ_t and that $\frac{\mathrm{d}\bar{\nu}}{\mathrm{d}\mu_t} = \partial f^*(h_t)$. By duality, this is equivalent to $h_t = \partial f(\frac{\mathrm{d}\bar{\nu}}{\mathrm{d}\mu_t})$, which shows that $\Phi^\top \alpha_t^\lambda$ converges to h_t in $\mathcal{C}^1(\Omega)$.

Finally, using the gradient flow equation in Eq. (III.32), the previously described convergence of $\Phi^\top \alpha_t^\lambda$ is sufficient to have for every test function $\varphi \in \mathcal{C}_c^\infty((0,T)\times\Omega)$:

$$\begin{aligned} &\int_0^T \int_\Omega \left(\partial_t \varphi_t - \frac{1}{2} \nabla \varphi_t \cdot \nabla f^*(\Phi^\top \alpha_t^\lambda) \right) \mathrm{d}\mu_t^\lambda \mathrm{d}t = 0 \\ \xrightarrow{\lambda\to 0^+} &\int_0^T \int_\Omega \left(\partial_t \varphi_t - \frac{1}{2} \nabla \varphi_t \cdot \nabla f^*(\partial f(\frac{\mathrm{d}\bar{\nu}}{\mathrm{d}\mu_t})) \right) \mathrm{d}\mu_t \mathrm{d}t = 0\,. \end{aligned}$$

Since $f(t) = |t|^r/(r-1)$ for some $r > 1$ the above equation is equivalent to Eq. (III.36) which shows μ_t is the weak solution starting from μ_0 of the ultra-fast diffusion equation Eq. (III.34) according to Definition III.2, that is $\mu_t = \mu_t^0$. □

The proof of the above Theorem III.5 relies on the following result on solutions to inverse problems with nonlinear regularization [Benning, 2018]. The following Lemma III.5.1 is similar to [Iglesias, 2018, Prop. 3].

Lemma III.5.1. *Assume f satisfies Assumption III.2 and Assumption III.1 holds. For $\mu \in \mathcal{P}(\Omega)$, let $u^\dagger \in L^1(\mu)$ be a solution of Eq. (III.14). We say $u^\dagger$ satisfies a* source condition *if there exists $\alpha \in L^2(\rho)$ s.t.*

$$\Phi^\top \alpha \in \partial f(u^\dagger) \quad \text{in } L^1(\mu). \tag{III.38}$$

Then in this case, noting $\alpha^\dagger \in L^2(\rho)$ the α of minimal norm satisfying the above source condition, we have for every $\lambda > 0$:

$$\|\alpha_f^\lambda[\mu]\|_{L^2(\rho)} \leq \|\alpha^\dagger\|_{L^2(\rho)} \quad \text{and} \quad \alpha_f^\lambda[\mu] \xrightarrow{\lambda \to 0^+} \alpha^\dagger\,,$$

where $\alpha_f^\lambda[\mu]$ is the solution to Eq. (III.19).

Proof. Let $u^\dagger \in L^1(\mu)$ and $\alpha^\dagger \in L^2(\rho)$ be as in the statement. By the source condition Eq. (III.38) we have in $L^1(\mu)$:

$$-f^*(\Phi^\top \alpha^\dagger) + (\Phi^\top \alpha^\dagger) u^\dagger \geq f(u^\dagger)$$

and integrating w.r.t. μ and using that $\int_\Omega (\Phi^\top \alpha^\dagger) u^\dagger \mathrm{d}\mu = \left\langle \alpha^\dagger, Y \right\rangle_{L^2(\rho)}$ we obtain:

$$-\int_\Omega f^*(\Phi^\top \alpha^\dagger)\mathrm{d}\mu + \left\langle \alpha^\dagger, Y \right\rangle_{L^2(\rho)} \geq \int_\Omega f(u^\dagger)\mathrm{d}\mu = \inf_{\Phi_\mu \cdot u = Y} \int_\Omega f(u)\mathrm{d}\mu\,.$$

Thus, $\alpha^\dagger$ achieves the supremum in Eq. (III.19) with $\lambda = 0$ and we have for any $\alpha \in L^2(\rho)$:

$$-\int_\Omega f^*(\Phi^\top \alpha^\dagger)\mathrm{d}\mu + \left\langle \alpha^\dagger, Y \right\rangle_{L^2(\rho)} \geq -\int_\Omega f^*(\Phi^\top \alpha)\mathrm{d}\mu + \langle \alpha, Y \rangle_{L^2(\rho)}\,.$$

Moreover, for $\lambda > 0$, noting $\alpha^\lambda \coloneqq \alpha_f^\lambda[\mu]$, we have by definition:

$$-\int_\Omega f^*(\Phi^\top \alpha^\lambda)\mathrm{d}\mu + \left\langle \alpha^\lambda, Y \right\rangle_{L^2(\rho)} - \frac{\lambda}{2}\|\alpha^\lambda\|_{L^2(\rho)}^2 \geq -\int_\Omega f^*(\Phi^\top \alpha^\dagger)\mathrm{d}\mu + \left\langle \alpha^\dagger, Y \right\rangle_{L^2(\rho)} - \frac{\lambda}{2}\|\alpha^\dagger\|_{L^2(\rho)}^2\,.$$

Substracting the two previous inequalities and simplifying gives:

$$\|\alpha^\lambda\|_{L^2(\rho)} \leq \|\alpha^\dagger\|_{L^2(\rho)}\,.$$

Thus α^λ is bounded, uniformly over $\lambda > 0$. For the convergence part, note that since it is bounded in $L^2(\rho)$ it converges weakly to some $\alpha^0 \in L^2(\rho)$. Also, taking the optimality condition for α^λ, we obtain for every $\alpha \in L^2(\rho)$:

$$-\int_\Omega f^*(\Phi^\top \alpha^\lambda)\mathrm{d}\mu + \left\langle \alpha^\lambda, Y \right\rangle_{L^2(\rho)} - \frac{\lambda}{2}\|\alpha^\lambda\|_{L^2(\rho)}^2 \geq -\int_\Omega f^*(\Phi^\top \alpha)\mathrm{d}\mu + \langle \alpha, Y \rangle_{L^2(\rho)} - \frac{\lambda}{2}\|\alpha\|_{L^2(\rho)}^2\,,$$

and taking the limit when $\lambda \to 0^+$ leads to:

$$-\int_\Omega f^*(\Phi^\top \alpha^0)\mathrm{d}\mu + \left\langle \alpha^0, Y \right\rangle_{L^2(\rho)} \geq -\int_\Omega f^*(\Phi^\top \alpha)\mathrm{d}\mu + \langle \alpha, Y\rangle_{L^2(\rho)} \ ,$$

which shows α^0 is also a maximizer of the dual problem Eq. (III.19) when $\lambda = 0$ and, as a consequence, also satisfies the source condition Eq. (III.38). But, by minimality of the norm of $\alpha^\dagger$ and by weak lower semicontinuity of the norm we have:

$$\|\alpha^\dagger\|_{L^2(\rho)} \leq \|\alpha^0\|_{L^2(\rho)} \leq \liminf_{\lambda \to 0^+} \|\alpha^\lambda\|_{L^2(\rho)} \leq \|\alpha^\dagger\|_{L^2(\rho)} \,,$$

which shows that in fact $\alpha^\lambda \to \alpha^\dagger$ strongly in $L^2(\rho)$. □

III.6 Numerics

We report in this section numerical results. First, to assess the validity of our theory, we tested the VarPro algorithm on simple low-dimensional examples with synthetic data: experiments with a 1-dimensional feature space are detailed in Section III.6.1 and supplementary experiments in 2-d are detailed in Section III.B. Those experiments indicate that, when the regularization is sufficiently low, the VarPro dynamic indeed enters an ultra-fast diffusion regime where the student feature distribution converges to the teacher's at a linear rate. Moreover, if the stepsize is sufficiently small, the VarPro dynamic can also be efficiently approximated by a two-timescale learning strategy.

Finally, to investigate the large-scale applicability and generalization capabilities of the VarPro algorithm, we tested it on an image classification problem with the CIFAR10 dataset [Krizhevsky, 2009] and compare its performances with other standard stochastic optimization methods. Those results are detailed in Section III.6.2.

The code for reproducing the results is available at: `https://github.com/rbarboni/VarPro`.

III.6.1 Single-hidden-layer neural networks with 1-dimensional feature space

We tested the VarPro algorithm for the training of a simple SHL with features on the 1-dimensional sphere $\mathbb{S}^1$. The feature space is here $\Omega = \mathbb{S}^1$, the data dimension is $d = 2$, and the feature map is given by $\phi : (\omega, x) \in \mathbb{S}^1 \times \mathbb{R}^2 \mapsto \mathrm{ReLU}(\omega^\top x)$ where ReLU is the *Rectified Linear Unit* activation. Recalling Eq. (III.1), we thus consider a SHL of width M defined for inner weights $\{\omega_i\}_{i=1}^M \in (\mathbb{S}^1)^M$ and outer weights $\{u_i\}_{i=1} \in \mathbb{R}^M$ by:

$$F_{\{(\omega_i,u_i)\}} : x \in \mathbb{R}^2 \mapsto \frac{1}{M}\sum_{i=1}^M u_i \mathrm{ReLU}(\omega_i^\top x) \,. \tag{III.39}$$

We consider a target signal Y that is given by a teacher network of width $\bar{M}$:

$$\forall x \in \mathbb{R}^2, \quad Y(x) = \frac{1}{\bar{M}}\sum_{i=1}^{\bar{M}} \mathrm{ReLU}(\bar{\omega}_i^\top x) \,.$$

The teacher feature distribution is hence $\bar{\mu}_\gamma = \frac{1}{\bar{M}}\sum_{i=1}^{\bar{M}} \delta_{\bar{\omega}_i}$ with i.i.d. features $\bar{\omega}_i \sim \mu_\gamma$ where, for $\gamma > 0$, we consider $\mu_\gamma := \left(\frac{2}{3}\delta_{\omega_1^*} + \frac{1}{3}\delta_{\omega_2^*}\right) \star \pi_\gamma$. The target feature modes are

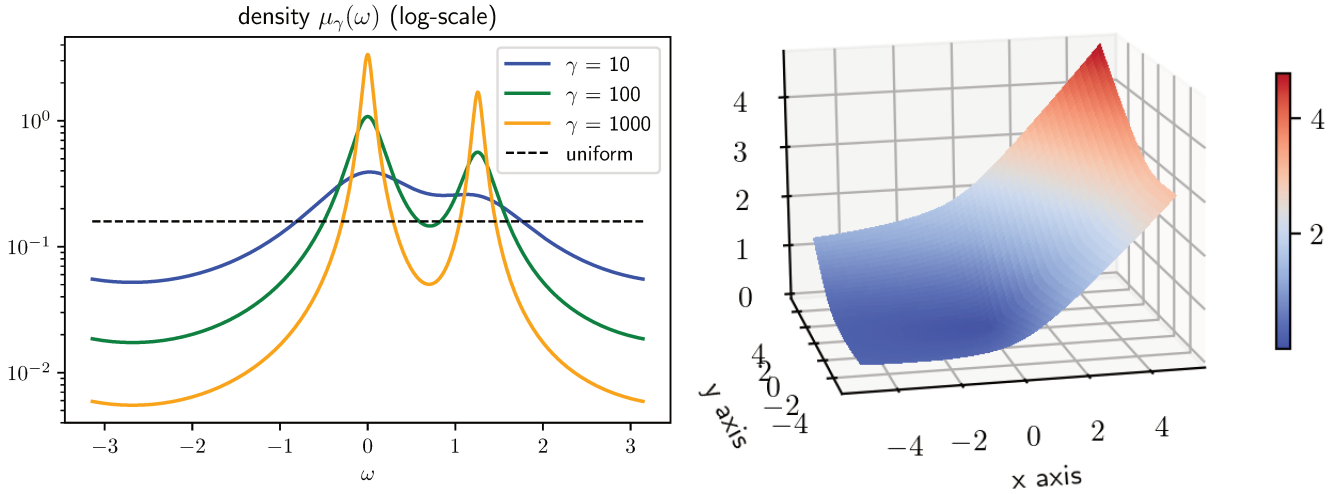


Figure III.1: Left: density of the teacher distributions μ_γ for $\gamma \in \{10, 100, 1000\}$. Right: corresponding teacher signal for $\gamma = 100$.

here fixed to $\omega_1^* = 0$ and $\omega_2^* = 0.4\pi$ and $\pi_\gamma \in \mathcal{P}(\mathbb{S}^1)$ is the distribution with density:

$$\pi_\gamma(\omega) \propto \frac{1}{1 + \gamma \sin^2(\omega/2)}, \quad \forall \omega \in \mathbb{S}^1 \,, \tag{III.40}$$

where by abuse of notation we identify $\omega \in \mathbb{S}^1$ with the corresponding angle in $\mathbb{R}/2\pi\mathbb{Z}$. In particular, the parameter $\gamma \geq 0$ controls the shape of the distribution μ_γ and the concentration around its modes: when $\gamma = 0$, μ_γ is the uniform distribution and, when $\gamma \to \infty$, we have $\mu_\gamma \to \mu_\infty := \frac{2}{3}\delta_{\omega_1^*} + \frac{1}{3}\delta_{\omega_2^*}$. Plots of the density μ_γ and of the corresponding teacher signal are shown in Fig. III.1. Finally, we consider the input data x to be distributed according to an empirical distribution $\hat{\rho} = \frac{1}{N}\sum_{i=1}^N \delta_{x_i}$ with i.i.d. standard Gaussian samples $x_i \sim \mathcal{N}(0, \mathrm{Id})$.

Recall that, for a regularization function $f : \mathbb{R} \to \mathbb{R}$ and a regularization strength $\lambda > 0$, the reduced risk defined by Eq. (III.13) associated to the features $\{\omega_i\}_{i=1}^M \in (\mathbb{S}^1)^M$ reads:

$$\hat{\mathcal{L}}_f^\lambda(\{\omega_i\}_{i=1}^M) = \min_{u \in \mathbb{R}^M} \frac{1}{2\lambda N} \sum_{j=1}^N \left| F_{\{(\omega_i, u_i)\}}(x_j) - Y(x_j) \right|^2 + \frac{1}{M}\sum_{i=1}^M f(u_i) \,. \tag{III.41}$$

In this setting, the *VarPro algorithm* is the time discretization of the particle evolution Eq. (III.10) and consists in performing gradient descent over the reduced risk $\hat{\mathcal{L}}_f^\lambda$:

$$\forall i \in \{1, ..., M\}, \forall k \geq 0, \quad \omega_i^{k+1} = \omega_i^k - M\tau \nabla_{\omega_i} \hat{\mathcal{L}}_f^\lambda(\{\omega_i^k\}_{1 \leq i \leq M}) \,. \tag{III.42}$$

where $\tau > 0$ is some stepsize parameter and $\{\omega_i^0\}_{i=1}^M \in (\mathbb{S}^1)^M$ is some random initialization. We consider here an uniform initialization with i.i.d. $\omega_i^0 \sim \mathcal{U}(\mathbb{S}^1)$.

Experimental setting We test the performance of the VarPro algorithm (Eq. (III.42)) for the training of SHLs (Eq. (III.39)) of varying width $M \in \{32, 128, 512, 1024\}$. We use either the "biased" quadratic regularization $f_b : t \mapsto \frac{1}{2}t^2$, for which the minimizer of the reduced risk differs from $\bar{\mu}_\gamma$, or the "unbiased" quadratic regularization $f_u : t \mapsto \frac{1}{2}|t-1|^2$,

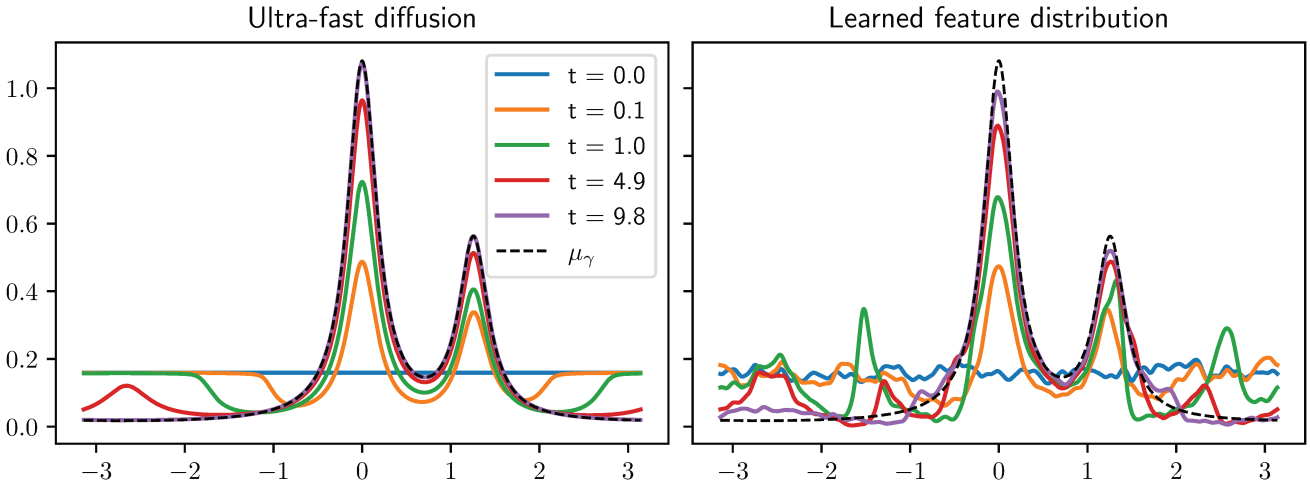


Figure III.2: Left: Solution μ_t to the ultra-fast diffusion Eq. (III.35) equation with exponent $r = 2$ and weights μ_γ, $\gamma = 100$. Right: Evolution of the feature distribution learned by gradient descent on a SHL of width $M = 1024$ for the minimization the reduced risk $\hat{\mathcal{L}}_f^\lambda$ with regularization function $f_b : t \mapsto \frac{1}{2}t^2$ and $\lambda = 10^{-4}$ (c.f. Eqs. (III.41) and (III.42)). The density is obtained by convolving the empirical feature distribution $\hat{\mu}$ with a gaussian kernel of variance $\sigma^2 = (0.03)^2$ and the plots are averages over 6 independent runs.

for which the minimizer of the reduced risk is the teacher distribution $\bar{\mu}_\gamma$ (c.f. Section III.3), and we consider varying regularization strength $\lambda \in \{10^{-1}, 10^{-2}, 10^{-3}, 10^{-4}\}$. We also consider different teacher distributions $\bar{\mu}_\gamma$ by changing the parameter $\gamma \in \{10, 100, 1000\}$. In order to stick with our theoretical results, we consider a number of data samples $N = 4096 \gg M$, such that the injectivity assumption in Assumption III.1 is satisfied, and we consider the teacher has a width $\bar{M} = 4096 \gg M$, such that the approximation $\bar{\mu}_\gamma \simeq \mu_\gamma$ holds. Finally, to closely model the gradient flow equation Eq. (III.26) we consider a stepsize $\tau = 2^{-10}$.

Qualitative comparison with ultra-fast diffusion on $\mathbb{S}^1$ Conveniently, the choice of the 1-dimensional domain $\mathbb{S}^1$ enables the use of standard numerical schemes to solve the weighted ultra-fast diffusion equation Eq. (III.35). This setting thus allows for comparison of the solutions to ultra-fast diffusion computed with high accuracy on a fine grid — we use here the "LSODA" integration method [Hindmarsh, 1983] — and the training dynamics computed with our VarPro method with particles Eq. (III.42). The two dynamics can be compared in Fig. III.2. Qualitatively, one can observe a close resemblance between the two dynamics, especially around the modes of the target distribution μ_γ where the densities progressively concentrates. While the learned feature distribution seems to concentrate less than the exact solution, this is likely due to the convolution with a gaussian kernel which is used to plot the density. However, the dynamics seems to differ more on the sides of the plots. These are indeed regions where the density μ_t becomes very low and thus where approximation of the velocity field $\nabla \left(\frac{\mu_\gamma}{\mu_t}\right)^2$ likely suffers from numerical instabilities.

Neural networks of varying width We investigate the behavior of the gradient descent dynamic for the minimization of the reduced risk (Eq. (III.42)) when varying the width M of the neural network. For this purpose we consider the teacher distribution

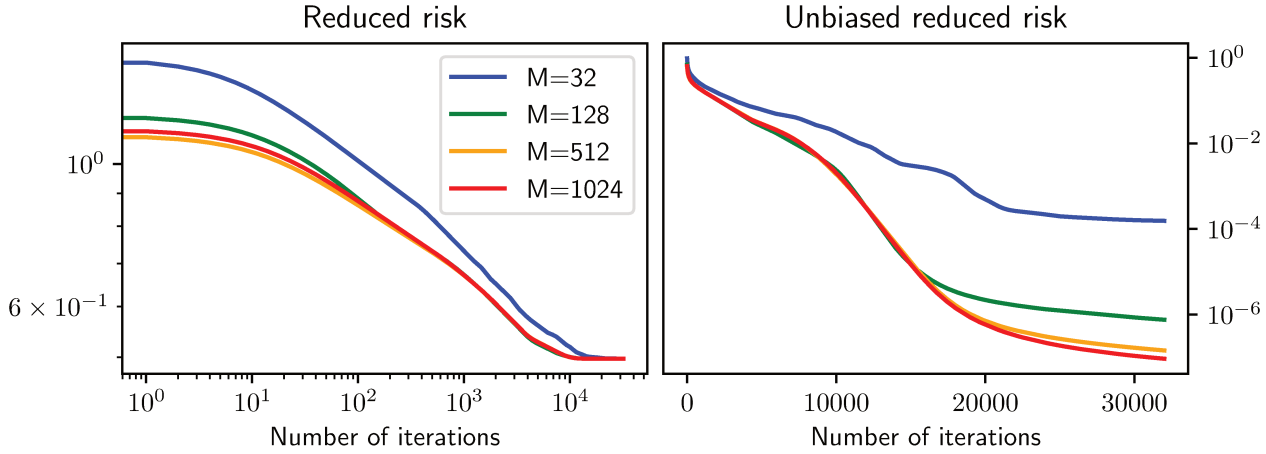


Figure III.3: Evolution of the reduced risk $\hat{\mathcal{L}}_f^\lambda$ (Eq. (III.42)) along iterations of gradient descent for a SHL of width $M \in \{32, 128, 512, 1024\}$. The regularization strength is $\lambda = 10^{-3}$ and the regularization function is either $f_b : t \mapsto \frac{1}{2}t^2$ (left) or $f_u : t \mapsto \frac{1}{2}|t-1|^2$ (right). Plots are averages over 6 independent runs.

$\bar{\mu}_\gamma \simeq \mu_\gamma$ with $\gamma = 100$, fix the regularization strength to $\lambda = 10^{-3}$ and consider SHLs of varying width $M \in \{32, 128, 512, 1024\}$ with regularization either f_b or f_u.

In this setting, Fig. III.3 reports evolution of the reduced risk $\hat{\mathcal{L}}_f^\lambda$ along iterations of gradient descent. In the case of the biased regularization f_b, the reduced risk monotonically decreases to the same (strictly positive) value for every width. This is normal since one should expect the feature distribution to converge to a minimizer $\bar{\mu}_\gamma^\lambda \neq \bar{\mu}_\gamma$ for which the reduced risk is strictly positive. On the contrary, in the case of the unbiased regularization f_u, the reduced risk monotonically decreases to different values depending on the width M. Indeed, in this case the gradient descent is expected to converge to the true teacher distribution $\bar{\mu}_\gamma \simeq \mu_\gamma$ and these different values corresponds to different levels of discretization of μ_γ. Also, in this case, the convergence speed seems to increase with the width.

In Fig. III.4, we report the evolution of a MMD distance between the learned feature distribution and two references which are the teacher distribution $\bar{\mu}_\gamma \simeq \mu_\gamma$ and the exact ultra-fast diffusion dynamic. We used the MMD distance Eq. (III.49) associated to the energy-distance kernel $\kappa(\omega, \omega') = -\|\omega - \omega'\|$. In coherence with what was observed before, in the case of the unbiased regularization f_u, the distance to the teacher distribution decreases monotonically to some value which is lower when the width increases. Illustrating our Theorem III.4, this shows gradient descent converges to a feature distribution discretizing the teacher distribution. On the contrary, when considering the biased regularization f_b, the positive regularization strength introduces a bias. In turn, plots of the distance to the diffusion dynamic show this distance decreases with the width, which is normal since a higher number of features corresponds to a better discretization. These plots also show that gradient descent stays close from the diffusion limit, as predicted by Theorem III.3.

Role of the regularization strength λ We now investigate the role played in the gradient descent dynamic by the regularization strength $\lambda > 0$. For this purpose, we consider a neural network of fixed width $M = 1024$ and train it with gradient descent for the minimization of the reduced risk $\hat{\mathcal{L}}_f^\lambda$ for varying values $\lambda \in \{10^{-1}, 10^{-2}, 10^{-3}, 10^{-4}\}$ of

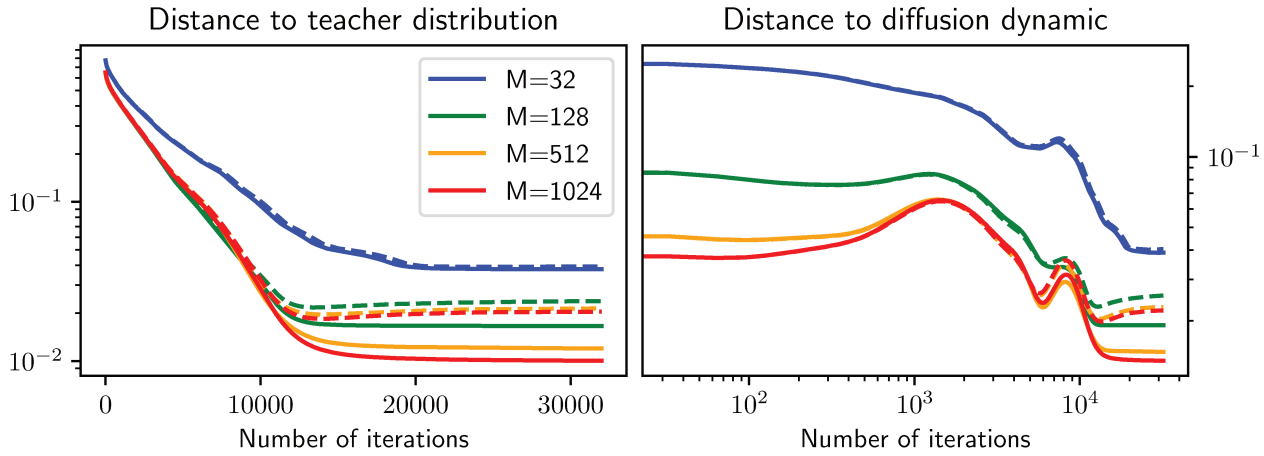


Figure III.4: Evolution of the MMD distance to the teacher distribution and to the diffusion dynamic along iterations of gradient descent over the reduced risk $\hat{\mathcal{L}}_f^\lambda$ (Eq. (III.42)) for a SHL of width $M \in \{32, 128, 512, 1024\}$. Left: distance to the teacher distribution $\bar{\mu}_\gamma \simeq \mu_\gamma$ ($\gamma = 100$). Right: distance to the diffusion dynamic. The regularization strength is $\lambda = 10^{-3}$ and the regularization function is either $f_b : t \mapsto \frac{1}{2}t^2$ (dashed) or $f_u : t \mapsto \frac{1}{2}|t-1|^2$ (plain). Plots are averages over 6 independent runs.

the regularization strength.

Evolution of the MMD distance between the learned feature distribution and respectively the teacher feature distribution and the diffusion dynamic are shown in Fig. III.5. On the plots of distance to the teacher distribution, one can first observe that the bias introduced in the case of the regularization f_b decreases with the regularization strength λ. This illustrates well our Proposition III.3.1, showing convergence of minimizers of the reduced risk towards the true teacher distribution when the regularization strength vanishes. In the case of the unbiased regularization f_u, one can observe a difference of behavior between low regularization regimes $\lambda \in \{10^{-2}, 10^{-3}, 10^{-4}\}$ and large regularization $\lambda = 10^{-1}$. While in the former case convergence seems to operate at a linear rate, which is the convergence rate of the diffusion limit (Theorem III.3), in the latter the convergence rate is significantly slower which could indicate an algebraic rate as predicted by Theorem III.4. Indeed, $\lambda = 10^{-1}$ is the order of magnitude of the most significant eigenvalues of the tangent kernel K_μ (numerically, the spectrum of K_μ is, in descending order, $\mathrm{Sp}(K_\mu) \simeq (0.2, 0.1, 0.1, 0.02, ...)$). Recalling that the risk can be expressed in terms of $(K_\mu + \lambda)^{-1}$ (Eq. (III.24)), an explanation is thus that, the unregularized reduced risk is well approximated only when $\lambda \ll K_\mu$. In contrast, in the high regularization regime ($\lambda \gtrsim K_\mu$), the reduced risk receive more influence from the MMD distance term than from the f-divergence term in Eq. (III.18) and gradient flows of MMD distances are known to be associated with slower convergence rates.

Finally, plots of the distance between the gradient flow and ultra-fast diffusion dynamics show this distance is lower and stays also lower for longer time when the regularization strength decreases. This supports the "local uniform in time convergence" behavior predicted by Theorem III.5. Note however that this result says nothing about the long time behavior of the dynamic, which is why the number of iterations is displayed in log-scale.

Role of the shape of the teacher distribution We investigate the role played by the shape of the distribution μ_γ, controlled by the parameter γ. We consider teacher

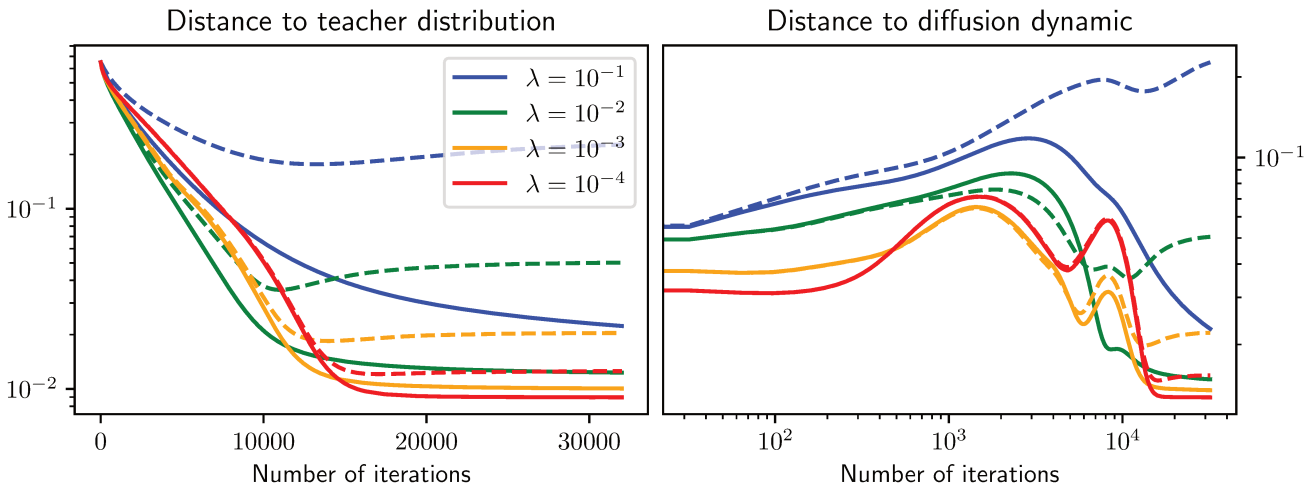


Figure III.5: Evolution of the MMD distance to the teacher distribution and to the diffusion dynamic along iterations of gradient descent over the reduced risk $\hat{\mathcal{L}}_f^\lambda$ (Eq. (III.42)) for a SHL of width $M = 1024$ with regularization $\lambda \in \{10^{-1}, 10^{-2}, 10^{-3}, 10^{-4}\}$. Left: distance to the teacher distribution $\bar{\mu}_\gamma \simeq \mu_\gamma$ ($\gamma = 100$). Right: distance to the diffusion dynamic. The regularization function is either $f_b : t \mapsto \frac{1}{2}t^2$ (dashed) or $f_u : t \mapsto \frac{1}{2}|t-1|^2$ (plain). Plots are averages over 6 independent runs.

distributions $\bar{\mu}_\gamma \simeq \mu_\gamma$ for $\gamma \in \{10, 100, 1000\}$ and train a neural network of fixed width $M = 1024$ with gradient descent over the reduced risk (Eq. (III.42)) with the unbiased regularization f_u and $\lambda = 10^{-4}$. Plot of the log-densities μ_γ are shown in Fig. III.1. In particular the distribution μ_γ approximates the atomic distribution $\mu_\infty = \frac{2}{3}\delta_{\omega_1^*} + \frac{1}{3}\delta_{\omega_2^*}$ in the limit $\gamma \to \infty$.

Plots of the evolution of the reduced risk, of the distance to the teacher distribution and of the distance to the ultra-fast diffusion dynamic are shown in Fig. III.6. One can clearly observe that the convergence speed of gradient descent is affected by the parameter γ. In particular, looking at the distance to the teacher distribution, every curve exhibits a linear convergence rate but this convergence rate deteriorates when γ increases. This supports the conclusions of Theorem III.3 in which the convergence rate of ultra-fast diffusion towards the target distribution is exponentially bad in the log-density ratio $\log(\mu_\gamma/\mu_0)$ (see also Remark III.4.2). Finally, on can observe in the last plot that gradient descent deviates more quickly from the diffusion dynamic when γ increases. When γ is large, there are indeed regions where the density μ_t will become very low, hence leading to numerical instabilities when estimating the velocity field $\nabla\left(\frac{\mu_\gamma}{\mu_t}\right)^2$.

Comparison with two-timescale gradient descent Since performing exact projection of the outer layer at every gradient step might have a prohibitive algorithmic cost, it is interesting to compare the VarPro algorithm with the two-timescale gradient descent which consist in affecting a different learning rate to the inner and outer weights of the neural network. For a regularization function $f : \mathbb{R} \to \mathbb{R}$ and a regularization strength $\lambda > 0$ we recall that the risk defined by Eq. (III.12) associated to the parameters

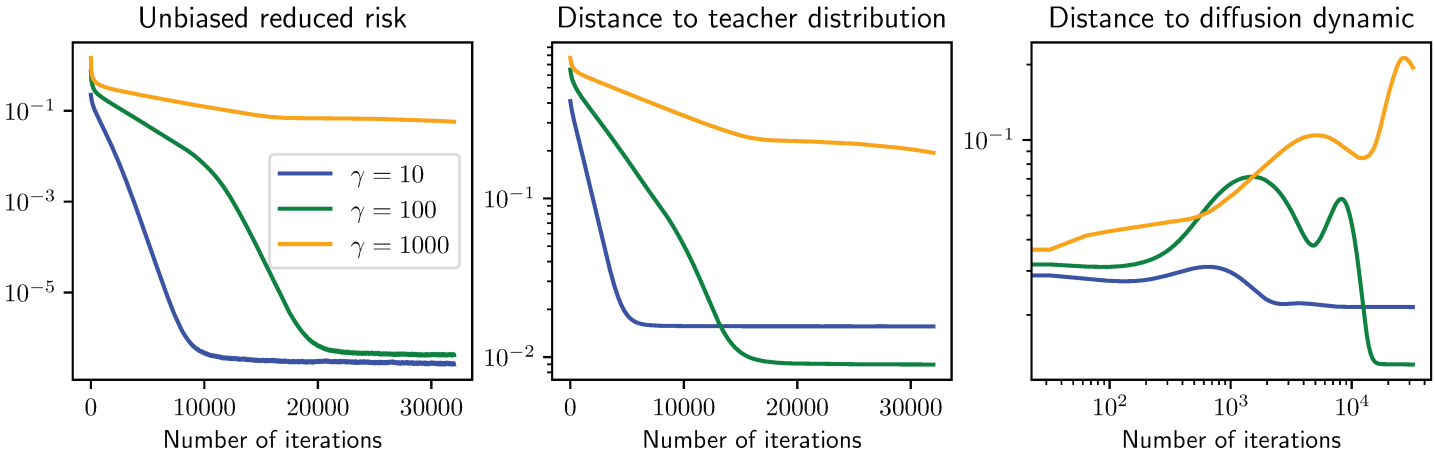


Figure III.6: Gradient descent over the reduced risk (Eq. (III.42)) for a SHL of width $M = 1024$ with unbiased regularization $f_u : t \mapsto \frac{1}{2}|t-1|^2$, $\lambda = 10^{-4}$ and teacher distribution $\bar{\mu}_\gamma \simeq \mu_\gamma$ for $\gamma \in \{10, 100, 1000\}$. Left: Evolution of the reduced risk. Middle: Evolution of the MMD distance to the teacher distribution $\bar{\mu}_\gamma \simeq \mu_\gamma$. Right: Distance to the ultra-fast diffusion dynamic. Plots are averages over 6 independent runs.

$\{(\omega_i, u_i)\}_{i=1}^M \in (\mathbb{S}^1 \times \mathbb{R})^M$ reads:

$$\frac{1}{\lambda}\hat{\mathcal{R}}_f^\lambda(\{(\omega_i,u_i)\}_{i=1}^M) = \frac{1}{2\lambda N}\sum_{j=1}^N \left|F_{\{(\omega_i,u_i)\}}(x_j) - Y(x_j)\right|^2 + \frac{1}{M}\sum_{i=1}^M f(u_i). \tag{III.43}$$

Then, for a timescale parameter $\eta > 0$, we implement the two-timescale gradient descent algorithm defined by :

$$\forall i \in \{1,...,M\},\ \forall k \geq 0, \quad \begin{cases} \omega_i^{k+1} &= \omega_i^k - \frac{M\tau}{\lambda}\nabla_{\omega_i}\hat{\mathcal{R}}_f^\lambda(\{(\omega_i^k,u_i^k)\}_{1\leq i\leq M})\,, \\ u_i^{k+1} &= u_i^k - \frac{\eta}{\lambda}\nabla_{u_i}\hat{\mathcal{R}}_f^\lambda(\{(\omega_i^k,u_i^k)\}_{1\leq i\leq M})\,. \end{cases} \tag{III.44}$$

As for the VarPro algorithm (Eq. (III.42)), we take the stepsize parameter $\tau = 2^{-10}$ and $\{\omega_i^0\}_{i=1}^M \in (\mathbb{S}^1)^M$ is some random initialization with i.i.d. $\omega_i^0 \sim \mathcal{U}(\mathbb{S}^1)$. For a fair comparison with VarPro, we first perform one projection step before training such that the outer weights initialization verifies:

$$u^0 \in \operatorname*{arg\,min}_{u\in\mathbb{R}^M} \hat{\mathcal{R}}_f^\lambda(\{(\omega_i^0,u_i)\}_{1\leq i\leq M})\,.$$

Concerning the timescale parameter η, we find it empirically efficient to set it to $\eta = \lambda M$. Lower values of η leads to slower training and higher values to numerical instabilities. An explanation for this is that, in the case of a quadratic regularization, by Eq. (III.43) the risk as a function of the outer weights $u \in \mathbb{R}^M$ reads:

$$\frac{1}{\lambda}\hat{\mathcal{R}}_f^\lambda(\{(\omega_i,u_i)\}_{1\leq i\leq M}) = \frac{1}{2\lambda N}\sum_{j=1}^N \left|\frac{1}{M}(\hat{\Phi}\cdot u)_j - Y(x_j)\right|^2 + \frac{1}{2M}\sum_{i=1}^M u_i^2\,,$$

where $\hat{\Phi} \in \mathbb{R}^{N\times M}$ is some feature matrix depending on the features $\{\omega_i\}_{1\leq i\leq M}$. Numerically, one observes $\frac{1}{NM}\lambda_{\max}(\hat{\Phi}^\top\hat{\Phi}) \simeq 1$, such that $\eta^{-1} = \frac{1}{\lambda M} \simeq \lambda_{\max}(\lambda^{-1}\nabla_{u,u}^2\hat{\mathcal{R}}_f^\lambda)$ indeed corresponds to the smoothness constant of the ridge regression problem w.r.t. u.

Remark III.6.1. *Note that, as explain above, for numerical stability, one can not consider an arbitrarily large time-scale parameter η and we fix here $\eta = \lambda M$. In this setting, the ratio between the lerning rates of inner and outer weights is given by $\frac{\eta}{M\tau} = \frac{\lambda}{\tau}$. Therefore, we can only expect to be in the two-timescale regime, i.e. when the two-timescale gradient descent is a good approximation of VarPro, if the stepsize τ is chosen s.t. $\tau \ll \lambda$.*

We stress that, for low-regularization regimes, this can be numerically prohibitive and VarPro, i.e. exact optimization of the outer weights at each step, can provide an efficient alternative to gradient descent. Interestingly, we in fact observe in our case that, as soon as $\tau \gg \lambda$ and thus $\eta \gg M\tau$ (which for examples happens here for $\lambda = 10^{-4}$), the VarPro algorithm (Eq. (III.42)) efficiently learns the teacher feature distribution (see e.g. Fig. III.5), whereas two-timescale gradient descent (Eq. (III.44)) does not converge.

In this setting we train SHLs of varying width using either the VarPro algorithm (Eq. (III.42)) or the two-timescale gradient descent algorithm (Eq. (III.44)) and report results in Fig. III.7. As predicted, one can observe the two dynamics are very close in the case case of a sufficiently high regularization, here $\lambda \geq 10^{-2}$, for which we have $\eta \gg M\tau$. This supports the fact that, in this regime, the VarPro dynamic can be obtained as the two-timescale limit of gradient descent. On the other hand, the two dynamics significantly differ in the low regularization regime $\lambda = 10^{-3}$ for which we have $\eta = \lambda M \simeq M\tau$. In this case, independently of the width M, the VarPro algorithm converges at a linear rate, while two-timescale gradient descent is slower and even seems to introduce a bias in the learned feature distribution. An explanation is that, in this regime, the two-timescale gradient descent quickly deviates from the ultra-fast diffusion dynamic, which one can observe in the last column of Fig. III.7. Overall, the most favorable setting seems to be when $\lambda = 10^{-2}$. Indeed, in this case $\eta = \lambda M \gg \tau M$ s.t. two-timescale gradient descent efficiently emulates the VarPro dynamic, while $\lambda \ll \|K_\mu\|_{op} \simeq 0.5$, the spectral norm of tangent kernel, s.t. both dynamics benefit from the linear convergence rate of ultra-fast diffusion (see also Fig. III.5).

III.6.2 VarPro for image classification on CIFAR10

We conclude this section by performing experiments on an image classification task with the CIFAR10 dataset [Krizhevsky, 2009]. We thereby aim at testing the large scale applicability of the VarPro algorithm. Note that applications of variable projection strategies to the training of deep neural network architectures were also studied in [Newman, 2021]. However, such setting goes outside of the scope of the theory developed in this chapter as the neural network can no longer be represented as a linear operator acting on measures.

We consider a *residual neural network (ResNet)* architecture with 20 layers and 0.27M parameters, whose precise description can be found in [He, 2016a, Sec. 4.2]. This model has a Euclidean parameter space Ω and for parameters $\omega \in \Omega$ and images $x \in \mathbb{R}^{3\times 32\times 32}$ it produces features which we denote by $\mathrm{ResNet}(\omega, x) \in \mathbb{R}^M$, with $M = 64$. We consider the last fully connected layer separately as a weight matrix $U \in \mathbb{R}^{c\times M}$ with here $c = 10$ the number of classes. Overall, for parameters $(\omega, U) \in \Omega \times \mathbb{R}^{c\times M}$ and an input image $x \in \mathbb{R}^{3\times 32\times 32}$, the output of the model is given by:

$$F_{(\theta,U)}(x) := \frac{1}{M} U \cdot \mathrm{ResNet}(\omega, x) \in \mathbb{R}^c .$$

To apply the VarPro algorithm we need to have an efficient way of computing the exact projection of the linear parameters U. For this purpose and instead of a cross-entropy loss, we consider here simply the square error between the outputs of our model and the

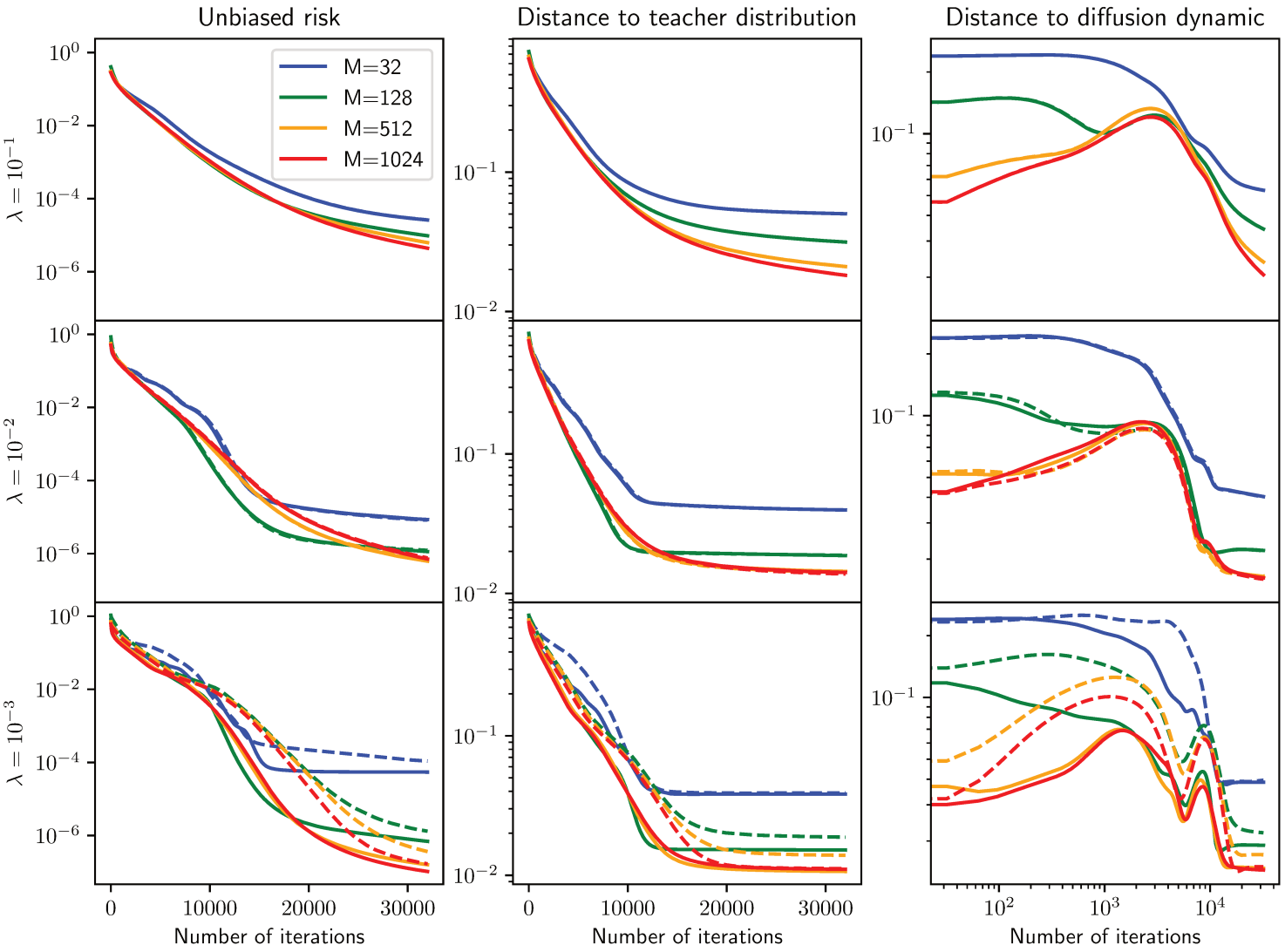


Figure III.7: VarPro (Eq. (III.42), plain lines) and two-timescale gradient descent (Eq. (III.44), dashed lines) over the risk for SHLs of varying width $M \in \{32, 128, 512, 1024\}$ with unbiased regularization function $f_u : t \mapsto \frac{1}{2}|t-1|^2$ and regularization strength $\lambda = 10^{-1}$ (top), $\lambda = 10^{-2}$ (middle) or $\lambda = 10^{-3}$ (bottom). The teacher distribution is $\bar{\mu}_\gamma \simeq \mu_\gamma$ with $\gamma = 100$. Left: Evolution of the risk. Middle: Evolution of the MMD distance to the teacher distribution. Right: Distance to the ultra-fast diffusion dynamic. Plots are averages over 6 independent runs.

true labels converted to one-hot vectors $y \in \{0,1\}^c$. In this manner, the training risk for a batch of data $\mathcal{D}$ and parameters $(\omega, U) \in \Omega \times \mathbb{R}^{c\times M}$ reads:

$$\hat{\mathcal{R}}^{\lambda}_{\mathcal{D}}(\omega, U) \coloneqq \frac{1}{2\#\mathcal{D}|} \sum_{(x,y)\in\mathcal{D}} \|F_{(\omega,U)}(x) - y\|^2 + \frac{\lambda}{2M}\|U\|^2 \,. \tag{III.45}$$

We then consider training our model with an adaptation of the SGD algorithm with momentum described in Eqs. (45) and (46). For an initialization $(\omega^0, U^0) \in \Omega \times \mathbb{R}^{c\times M}$, a timestep $\tau > 0$ and a momentum parameter $\mathrm{m} > 0$ the training dynamic reads:

$$\forall k \geq 0, \quad \begin{cases} U^{k+1} &= \mathrm{m}U^k + (1-\mathrm{m})\bar{U}^k \,, \\ \omega^{k+1} &= \omega^k - \frac{M\tau}{\lambda}\nabla_\omega \hat{\mathcal{R}}^{\lambda}_{\mathcal{D}_k}(\omega^k, U^{k+1}) \,. \end{cases} \tag{III.46}$$

where $\mathcal{D}_k$ is the mini-batch at step k and $\bar{U}_k$ is the corresponding projection of the outer weights i.e. $\bar{U}^k \in \arg\min_{U\in\mathbb{R}^{c\times M}} \hat{\mathcal{R}}^{\lambda}_{\mathcal{D}_k}(\omega^k, U)$. The introduction of the momentum parameter $\mathrm{m} > 0$ is here to compensate the variability of the projection $\bar{U}_k$ w.r.t. the sampling of mini-batches at each step. Indeed, intuitively, for evaluation on test-data, rather than having a classifier computed only on the last mini-batch, it is preferable to have an average of the last computed classifiers.

Experimental setting In practice, we find it effective to consider a regularization strength $\lambda = 10^{-3}$, a momentum $\mathrm{m} = 0.9$ and a stepsize $\tau = 10^3$. We consider different values of the batch size $\#\mathcal{D} \in \{64, 128, 256, 512, 1024\}$. We train our model by performing 110 passes over the training set, evaluating the model accuracy on the test set in-between each pass. The stepsize is divided by 2 for the last 10 passes on the training set. Note that this setting allows for a fair comparison of performances with the results presented in [He, 2016a, Sec. 4.2] for the training of ResNets on the CIFAR10 dataset. We also follow the same data-augmentation procedure.

Comparison with other stochastic optimization methods We compare the above described VarPro algorithm (Eq. (III.46)) with other stochastic optimization methods for the minimization of the training risk in Eq. (III.45). We compare with standard *Stochastic Gradient Descent (SGD)* on the full parameterization $(\omega, U) \in \Omega \times \mathbb{R}^{c\times M}$ with momentum $\mathrm{m} = 0.9$ and stepsize $\tau = 10^{-3}$. We also compare with the *Shampoo* algorithm [Gupta, 2018] which is a preconditioned gradient method[1] and set the learning rate to $\tau = 10^{-2}$.

Fig. III.8 reports the evolution of the training risk (Eq. (III.45)) along training. One can observe that, in terms of minimization of the training risk, performances of VarPro at convergence are similar to the one of SGD and better than Shampoo. Compared with these last two methods, the convergence speed of VarPro however seems to be slower during the first stages of training. Behavior of the methods w.r.t. the batch size is also different. While the batch size has no or little influence on the convergence speed of SGD or Shampoo, one can observe that the VarPro algorithm tends to converge more slowly when the batch size increases. Since this method is based on the exact resolution of a quadratic minimization problem on each mini-batch at each step, an explanation is thus that this subproblem becomes less well-conditioned when the size of the mini-batches increases.

Fig. III.8 reports the evolution of the top-1 accuracy of the model on the test set. All optimization methods seems to achieve the same generalization performances on the test set, that is more than 90% accuracy, which is in par with the 91.25% reported for the

[1] We used the implementation from https://github.com/moskomule/shampoo.pytorch

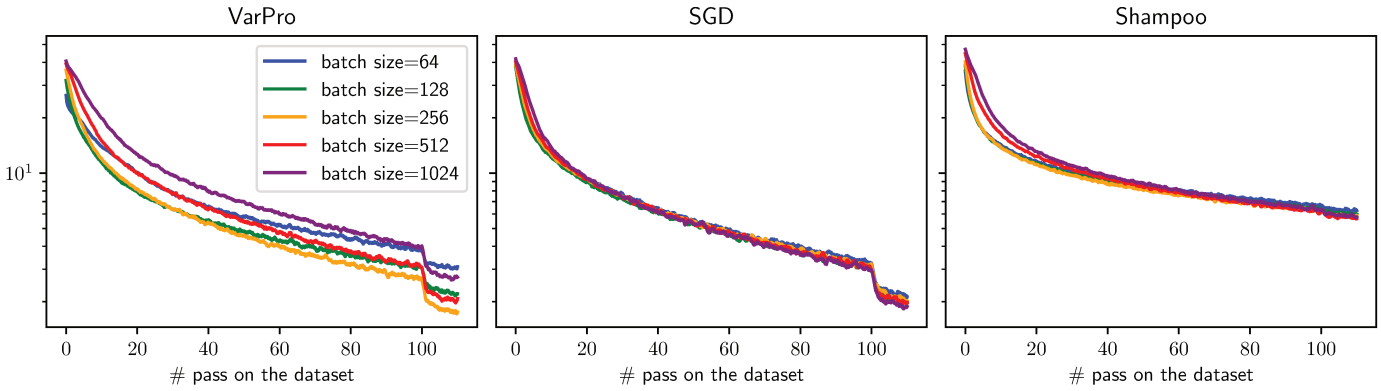


Figure III.8: Evolution of the training risk $\frac{1}{\lambda}\hat{\mathcal{R}}_{\mathcal{D}}^{\lambda}$ (Eq. (III.45)) along training for different batch sizes and different optimization methods. Plots are averages of the risk associated to each mini-batch encountered during one pass. VarPro corresponds to Eq. (III.46).

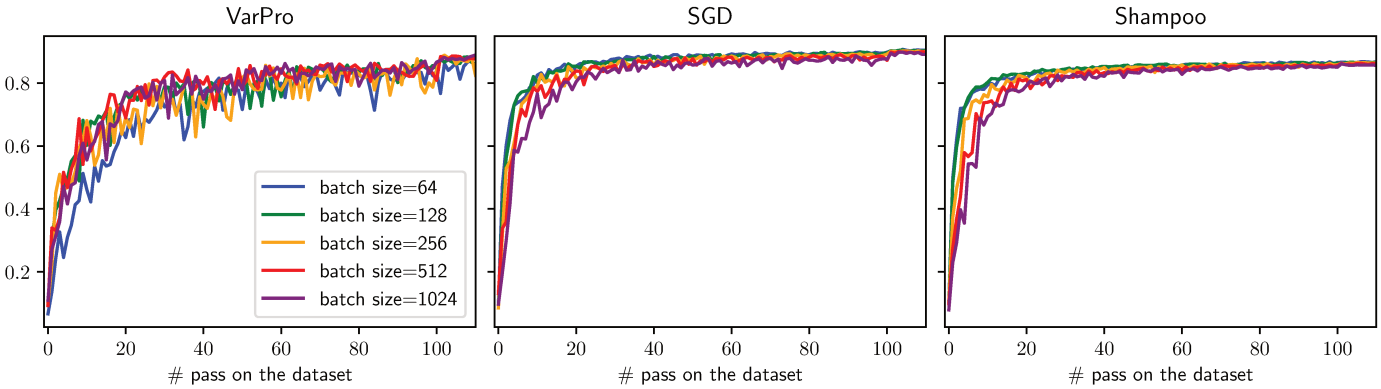


Figure III.9: Evolution of the top-1 accuracy along training for different batch sizes and different optimization methods. VarPro corresponds to Eq. (III.46).

same model in [He, 2016a]. As before, one can observe the Varpro algorithm (Eq. (III.46)) seems to take more time to achieve the same accuracy. Also, whereas SGD and Shampoo generalize better when the batch size is smaller, the converse happens for VarPro and one can see the generalization performance of our ResNet model trained with the VarPro algorithm deteriorates for smaller batch sizes.

III.7 Conclusion

In this chapter we have investigated the convergence of gradient based methods for the training of mean-field models of shallow neural networks. To this end, we have adopted a *Variable Projection (VarPro)* strategy, which eliminates the linear parameters and reduces the training problem to the learning of the nonlinear features. Using tools from the theory of Wasserstein gradient flows, we have shown theoretically that, when the regularization strength λ vanishes, the training dynamic converges, under regularity assumptions, to solutions of *weighted ultra-fast diffusion* PDEs (Theorem III.5). In such a low regularization regime, this allows establishing convergence of the learned feature distribution to the teacher's at a linear rate (Theorem III.3). Moreover, in presence of regularization, we also obtain a quantitative convergence result but with a slower algebraic rate (Theorem III.4).

Our theoretical predictions are supported by numerical results on simple experiments with synthetic data. One can observe that, when the regularization strength λ is negligible compared to the tangent kernel, the VarPro and ultra-fast diffusion dynamics are similar and converge to the teacher feature distribution at a linear rate (Fig. III.5). Moreover, if the time step is sufficiently small, this dynamic is also recovered with a simple two-timescale gradient descent algorithm (Fig. III.7). Finally, experiments with a ResNet architecture on the CIFAR10 dataset show that a VarPro strategy can be easily adapted to the training of complex architectures on large datasets.

We conclude by mentioning possible future research directions:

- On a theoretical perspective, our convergence results in Section III.5 hold under regularity assumptions on the training dynamic. It would be interesting to see if one can relax or ensure these assumptions, possibly by strengthening Assumption III.1.
- We have considered here simple 2-layer neural network architectures but, as pointed out in Chapter II, the learning of good nonlinear representations of the data also plays an important role in the training of deep architectures such as ResNets or Transformers [Gao, 2024]. It might thus be interesting to see in what extent the mathematical framework developed in Chapter I could be extended to model two-timescale approaches for the training of deeper architectures. Ultimately, an objective could to relax the convergence conditions obtained in Section II.5 by ensuring the learning of good feature representations during training. A difficulty is that, in deep architectures, separability of the regression problem w.r.t. linear and nonlinear variables of each layer is lost due to composition.

Appendices

III.A Positive definite kernels and RKHS

We recall in this section basic properties on the theory of Reproducing Kernel Hilbert Spaces and refer to classical textbooks for a complete presentation of the topic [Steinwart, 2008; Schölkopf, 2002]. In this chapter we consider a mapping $\phi : \Omega \times \mathbb{R}^d \to \mathbb{R}$ as well as a probability measure $\rho \in \mathcal{P}(\mathbb{R}^d)$. This choice of ϕ and ρ determines a *symmetric, positive definite kernel* [Steinwart, 2008, Def. 4.15] $\kappa : \Omega \times \Omega \to \mathbb{R}$ defined by:

$$\forall \omega, \omega' \in \Omega, \quad \kappa(\omega, \omega') \coloneqq \int_X \phi(\omega, x)\phi(\omega', x)\mathrm{d}\rho(x) = \left\langle \phi(\omega, .), \phi(\omega', .) \right\rangle_{L^2(\rho)} .$$

Thus, associated to κ is a (uniquely defined) structure of *Reproducing Kernel Hilbert Space (RKHS)* $\mathcal{H}$ with scalar product $\langle ., . \rangle_{\mathcal{H}}$, that is a Hilbert space of functions on Ω such that [Steinwart, 2008, Def. 4.18]: (i) $\kappa(\omega, .) \in \mathcal{H}$ for every $\omega \in \Omega$ and (ii) the following *reproducing property* holds:

$$\forall h \in \mathcal{H},\ \omega \in \Omega, \quad h(\omega) = \langle h, \kappa(\omega, .) \rangle_{\mathcal{H}} .$$

Following the definition of κ, $L^2(\rho)$ is a *feature space* for $\mathcal{H}$ and ϕ a *feature map* [Steinwart, 2008, Def. 4.1]. Also, $\mathcal{H}$ can be isometrically identified as a subspace of $L^2(\rho)$ and the convolution with ϕ is a partial isometry [Steinwart, 2008, Thm. 4.21]. Precisely, we have

$$\mathcal{H} = \left\{ h : \Omega \to \mathbb{R} \ : \ \exists \alpha \in L^2(\rho) \text{ s.t. } h = \int_{\mathbb{R}^d} \phi(., x)\alpha(x)\mathrm{d}\rho(x) \right\}$$

and the RKHS norm on $\mathcal{H}$ satisfies:

$$\forall h \in \mathcal{H}, \quad \|h\|_{\mathcal{H}} = \inf \left\{ \|\alpha\|_{L^2(\rho)} \ : \ h = \int_{\mathbb{R}^d} \phi(., x)\alpha(x)\mathrm{d}\rho(x) \right\} . \tag{III.47}$$

In this chapter, we always work with the following minimal assumption on the feature map ϕ:

Assumption III.3 (Assumption on ϕ)**.**
The feature map ϕ is in $L^2(\rho, \mathcal{C}^0)$. In particular, it implies the kernel κ is continuous.

Kernel embeddings and kernel discrepancy between measures The above assumption is sufficient to ensure $\mathcal{H}$ is included in $\mathcal{C}(\Omega)$ and guarantees the existence of kernel embeddings for finite Borel measures [Muandet, 2017; Gretton, 2012]. For a measure $\nu \in \mathcal{M}(\Omega)$ its *kernel embedding* $M_\kappa(\nu)$ is defined as the unique element of $\mathcal{H}$ satisfying:

$$\forall h \in \mathcal{H}, \quad \int_\Omega h\mathrm{d}\nu = \langle h, M_\kappa(\nu) \rangle_{\mathcal{H}} . \tag{III.48}$$

Equivalently, the kernel embedding is given by the Bochner integral $M_\kappa(\nu) = \int_\Omega \kappa(., \omega)\mathrm{d}\nu(\omega) \in \mathcal{H}$. This embedding defines a discrepancy between measures by seeing them as element of the Hilbert space $\mathcal{H}$. For two measures $\nu, \nu' \in \mathcal{M}(\Omega)$ the *Maximum Mean Discrepancy (MMD)* between ν and ν' is defined as [Muandet, 2017; Gretton, 2012]:

$$\mathrm{MMD}_\kappa(\nu, \nu') \coloneqq \left\| M_\kappa(\nu) - M_\kappa(\nu') \right\|_{\mathcal{H}} .$$

Alternatively, Assumption III.3 is sufficient to ensure the convolution $\Phi\star : \mathcal{M}(\Omega) \to L^2(\rho)$ defined in Eq. (III.4) is a bounded operator and by construction we have:

$$\mathrm{MMD}_\kappa(\nu,\nu') = \left(\iint_{\Omega\times\Omega} \kappa(\omega,\omega')\mathrm{d}(\nu-\nu')(\omega)\mathrm{d}(\nu-\nu')(\omega')\right)^{1/2} = \left\|\Phi\star(\nu-\nu')\right\|_{L^2(\rho)} . \tag{III.49}$$

The discrepancy MMD_κ is in particular a distance between measures whenever the kernel κ is *universal*, that is when the associated RKHS $\mathcal{H}$ is dense in the space of continuous functions on Ω [Micchelli, 2006; Sriperumbudur, 2011]. One can show this condition is equivalent to an injectivity assumption on $\Phi\star$.

Lemma III.A.1 (see also [Micchelli, 2006, Prop. 1])**.** *Let ϕ satisfy Assumption III.3. Then $\Phi\star : \mathcal{M}(\Omega) \to L^2(\rho)$ is injective if and only if $\mathcal{H}$ is dense in the space $\mathcal{C}^0(\Omega)$ of continuous functions over Ω. In this case, MMD_κ is a distance on $\mathcal{M}(\Omega)$.*

Proof. The fact the MMD is a distance on $\mathcal{M}(\Omega)$ when $\Phi\star$ is injective directly follows from Eq. (III.49). For the direct implication, assume $\Phi\star$ is injective and consider some measure $\nu \in \mathcal{H}^\perp$ i.e. such that for every $h \in \mathcal{H}$ we have $\int h\mathrm{d}\nu = 0$. Then by the characterisation in Eq. (III.47) we have for every $\alpha \in L^2(\rho)$:

$$0 = \int_\Omega \left(\int_{\mathbb{R}^d} \phi(\omega,x)\alpha(x)\mathrm{d}\rho(x)\right)\mathrm{d}\nu(\omega) = \langle\alpha, \Phi\star\nu\rangle_{L^2(\rho)} .$$

Hence $\Phi\star\nu = 0$, implying $\nu = 0$ and thus that $\mathcal{H}^\perp = \{0\}$ i.e. $\mathcal{H}$ is dense in $\mathcal{C}^0(\Omega)$ by Hahn-Banach theorem. For the converse implication, assume that $\mathcal{H}$ is dense in $\mathcal{C}^0(\Omega)$ and consider some $\nu \in \mathcal{M}(\Omega)$ s.t. $\Phi\star\nu = 0$. Then for $\alpha \in L^2(\rho)$ we have $\langle\alpha, \Phi\star\nu\rangle_{L^2(\rho)} = 0$ and by similar calculations this implies $\nu \in \mathcal{H}^\perp$ i.e. $\nu = 0$. □

Kernel and integral operator In this chapter we have used properties of the RKHS $\mathcal{H}$ seen as a subspace of the Hilbert space $L^2(\mu)$ for probability measures $\mu \in \mathcal{P}(\Omega)$. For such a probability measure $\mu \in \mathcal{P}(\Omega)$, it indeed follows from Assumption III.3 that $\mathcal{H}$ is compactly embedded in $L^2(\mu)$ [Steinwart, 2012, Lem. 2.3]. Also, the kernel defines an integral operator $J_\mu : L^2(\mu) \to L^2(\mu)$ given by:

$$\forall f \in L^2(\mu)\,, \quad J_\mu \cdot f = \int_\Omega k(.,\omega)f(\omega)\mathrm{d}\mu(\omega)\,.$$

Then J_μ is a compact, self-adjoint and positive operator and, by the spectral theorem, it can be diagonalized in an orthonormal basis $(e_i)_{i\geq 0}$ of $L^2(\mu)$ with associated eigenvalues $(\lambda_i)_{i\geq 0}$ s.t. $\lambda_1 \geq \lambda_2 \geq ... \geq 0$. In particular, $J_\mu = \Phi_\mu^\top\Phi_\mu$ with $\Phi_\mu : L^2(\mu) \to L^2(\rho)$ the *feature operator* defined in Eq. (III.7), thus $(\sqrt{\lambda_i})_{i\geq 0}$ are the (right) singular values of Φ_μ (which is a compact operator) and, if $\Phi\star$ is injective, then $\lambda_i > 0$ for every $i \geq 0$. Mercer's theorem gives a representation of the kernel κ and of the associated RKHS $\mathcal{H}$ in terms of this eigenvalue decomposition [Steinwart, 2008, Thm. 4.51].

Theorem III.6 (Mercer representation of RKHSs)**.** *Assume $\Phi\star$ is injective and let $\mu \in \mathcal{P}(\Omega)$ be a probability measure with full support on Ω. Consider $(\lambda_i)_{i\geq 0}$ and $(e_i)_{i\geq 0}$ the eigenvalue decomposition of the operator J_μ. Then we have:*

$$\forall \omega,\omega' \in \Omega\times\Omega\,, \quad \kappa(\omega,\omega') = \sum_{i\geq 0} \lambda_i e_i(\omega)e_i(\omega')\,,$$

where the convergence is absolute and uniform over $\Omega \times \Omega$. *Moreover*

$$\mathcal{H} = \left\{ \sum_{i \geq 0} a_i \sqrt{\lambda_i} e_i \ : \ (a_i)_{i \geq 0} \in \ell^2(\mathbb{N}) \right\}$$

is the RKHS associated to the kernel κ *and the scalar product* $\langle ., . \rangle_{\mathcal{H}}$ *is given for every* $f = \sum_{i \geq 0} a_i \sqrt{\lambda_i}$ *and* $g = \sum_{i \geq 0} b_i \sqrt{\lambda_i}$ *by* $\langle f, g \rangle_{\mathcal{H}} = \sum_{i \geq 0} a_i b_i$.

III.B Radial basis function neural network on the 2-dimensional torus

We performed numerical experiments to test the performance of the VarPro algorithm for the training of *Radial Basis Function (RBF)* neural networks. Notably, due to the particular structure of the architecture, the learning problem corresponds to performing a deconvolution, which has important applications in signal processing [De Castro, 2012; Duval, 2015].

The feature space is here the 2-dimensional torus $\Omega = \mathbb{R}^2/4\mathbb{Z}^2 \subset \mathbb{R}^2$ and the data dimension is $d = 2$. The RBF neural network architecture performs the convolution with a kernel $k : \mathbb{R}^2 \to \mathbb{R}$ and corresponds to considering the feature map $\phi : (\omega, x) \mapsto k(\omega - x)$. We will use here the Laplace kernel $k : x \in \mathbb{R}^2 \mapsto 8 \exp(-\frac{1}{2}\|[x]\|)$, where $[x]$ represents the projection of x in $\Omega = \mathbb{R}^2/4\mathbb{Z}^2$. For features $\{\omega_i\}_{i=1}^M \in (\Omega)^M$ and outer weights $\{u_i\}_{i=1} \in \mathbb{R}^M$ the RBF neural network model reads:

$$F_{\{(\omega_i, u_i)\}} : x \in \mathbb{R}^2 \mapsto \frac{1}{M} \sum_{i=1}^{M} u_i k(\omega_i - x) = (k \star \hat{\nu})(x) \,, \tag{III.50}$$

where $\hat{\nu} = \frac{1}{M} \sum_{i=1}^M u_i \delta_{\omega_i} \in \mathcal{M}(\Omega)$ and $\star$ is the convolution operator. We consider a teacher feature distribution $\bar{\mu}_\gamma = \frac{1}{\bar{M}} \sum_{i=1}^{\bar{M}} \delta_{\bar{\omega}_i}$ for features $\{\bar{\omega}_i\}_{1 \leq i \leq \bar{M}} \in \Omega^{\bar{M}}$ and the target signal Y is thus:

$$\forall x \in \mathbb{R}^2, \quad Y(x) = \frac{1}{\bar{M}} \sum_{i=1}^{\bar{M}} k(\omega_i - x) = (k \star \bar{\mu}_\gamma) \,.$$

The teacher features are i.i.d. with $\bar{\omega}_i \sim \mu_\gamma = (\frac{1}{2}\delta_{\omega_1^*} + \frac{1}{2}\delta_{\omega_2^*}) \star \pi_\gamma$, where $\omega_1^* = (-1, 0)$, $\omega_2^* = (1, 1)$ are two target modes and π_γ is the product measure with density:

$$\forall (z_1, z_2) \in \mathbb{R}^2/4\mathbb{Z}^2, \quad \pi_\gamma(z_1, z_2) \propto \frac{1}{1 + \gamma \sin^2(z_1 \pi/4)} \times \frac{1}{1 + \gamma \sin^2(z_2 \pi/4)} \,.$$

The parameter γ controls the shape of the distribution μ_γ such that in the large γ limit one recovers $\mu_\gamma \simeq \mu_\infty := \frac{1}{2}\delta_{\omega_1^*} + \frac{1}{2}\delta_{\omega_2^*}$. A scatter plot of the teacher measure $\bar{\mu}_\gamma$ and of the resulting teacher signal for $\gamma = 100$ are shown in Fig. III.10. Finally, we consider the input data x to be distributed according to an empirical distribution $\hat{\rho} = \frac{1}{N} \sum_{i=1}^N \delta_{x_i}$ with i.i.d. standard Gaussian samples $x_i \sim \mathcal{N}(0, \mathrm{Id})$. In this setting we consider training the model by performing a VarPro algorithm i.e. gradient descent over the reduced risk, as in Eq. (III.42).

Experimental setting We test the performances of the VarPro algorithm (Eq. (III.42)) for the training of a RBF neural network (Eq. (III.50)) of varying width $M \in \{32, 128, 512, 1024\}$. We use either the "biased" quadratic regularization $f_b : t \mapsto \frac{1}{2}t^2$ or the "unbiased"

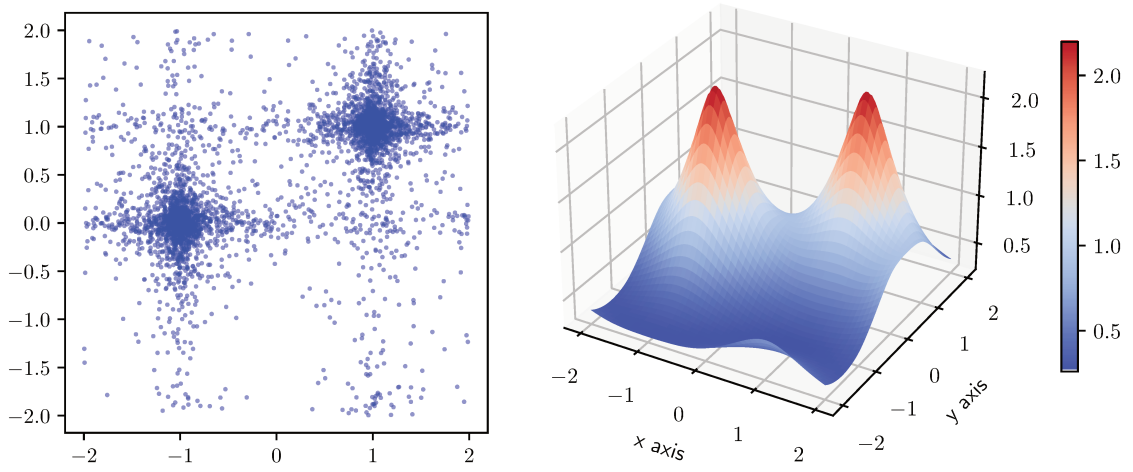


Figure III.10: Left: scatter plot the empirical teacher distribution $\bar{\mu}_\gamma$ for $\gamma = 100$. Right: corresponding target signal.

quadratic regularization $f_u : t \mapsto \frac{1}{2}|t-1|^2$ and we consider varying regularization strength $\lambda \in \{10^{-1}, 10^{-2}, 10^{-3}\}$. We consider different teacher distributions $\bar{\mu}_\gamma$ by changing the parameter $\gamma \in \{100, +\infty\}$. As in Section III.6.1, we consider a number of data samples $N = 4096 \gg M$, a teacher of width $\bar{M} = 4096 \gg M$ — s.t. the approximation $\bar{\mu}_\gamma \simeq \mu_\gamma$ holds — and a stepsize $\tau = 2^{-10}$.

Student of varying width As before, we first investigate the role played by the width M of the student in the training dynamic. For this purpose, we fix the regularization strength to $\lambda = 10^{-3}$ and consider training RBF neural networks (Eq. (III.50)) of varying width $M \in \{32, 128, 512, 1024\}$ with the teacher distribution $\bar{\mu}_\gamma$, $\gamma = 100$.

Fig. III.11 reports evolution of the biased and unbiased reduce risk during training and Fig. III.12 reports evolution of the distance to the teacher distribution $\bar{\mu}_\gamma$. As for our 1-dimensional experiments, one can observe that the VarPro algorithm converges to lower values of the unbiased reduced risk when the width of the student increases. In turn, at convergence, this corresponds to learned feature distributions that approximate the teacher distribution with different levels of discretization. On the contrary, using the biased regularization $f_b : t \mapsto \frac{1}{2}t^2$ introduces a bias in the learned distribution.

Role of the regularization strength λ We now investigate the role of the regularization strength $\lambda > 0$. We thus consider training RBF neural networks (Eq. (III.50)) of fixed width $M = 1024$ with the teacher distribution $\bar{\mu}_\gamma$, $\gamma = 100$, and we perform gradient descent over the reduced risk (Eq. (III.42)) with varying regularization $\lambda \in \{10^{-1}, 10^{-2}, 10^{-3}\}$. Note that here, compared to our 1-dimensional, the case of regularization lower than $\lambda = 10^{-3}$ is numerically impracticable, at least with our choice of stepsize $\tau = 2^{-10}$.

Evolution of the distance to the teacher distribution $\bar{\mu}_\gamma$ along training is reported in Fig. III.13. As in our 1-dimensional experiments, one can observe that the convergence speed gets slower when the regularization strength increases. There is also a significant change of behavior between $\lambda = 10^{-1}$ and $\lambda \in \{10^{-2}, 10^{-3}\}$. In the former case the convergence seems to exhibit an algebraic rate, supporting the conclusions of Theorem III.4,

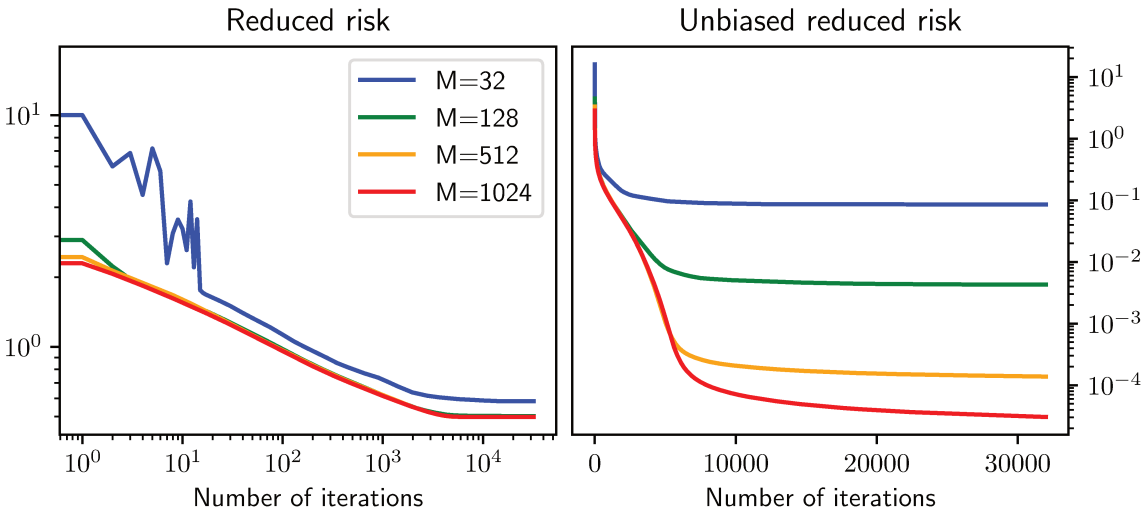


Figure III.11: Evolution of the reduced risk along iterations of gradient descent for a RBF neural network (Eq. (III.50)) of width $M \in \{32, 128, 512, 1024\}$. The regularization strength is $\lambda = 10^{-3}$ and the regularization function is either $f_b : t \mapsto \frac{1}{2}t^2$ (left) or $f_u : t \mapsto \frac{1}{2}|t-1|^2$ (right). Plots are averages over 6 independent runs.

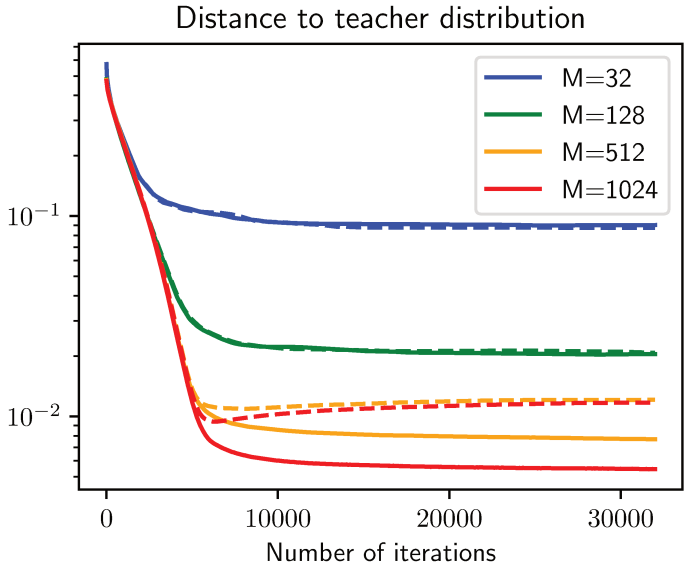


Figure III.12: Evolution of the MMD distance to the teacher distribution $\bar{\mu}_\gamma$ along gradient descent over the reduced risk for a RBF neural network (Eq. (III.50)) of width $M \in \{32, 128, 512, 1024\}$. The regularization strength is $\lambda = 10^{-3}$ and the regularization function is either $f_b : t \mapsto \frac{1}{2}t^2$ (dashed) or $f_u : t \mapsto \frac{1}{2}|t-1|^2$ (plain). Plots are averages over 6 independent runs.

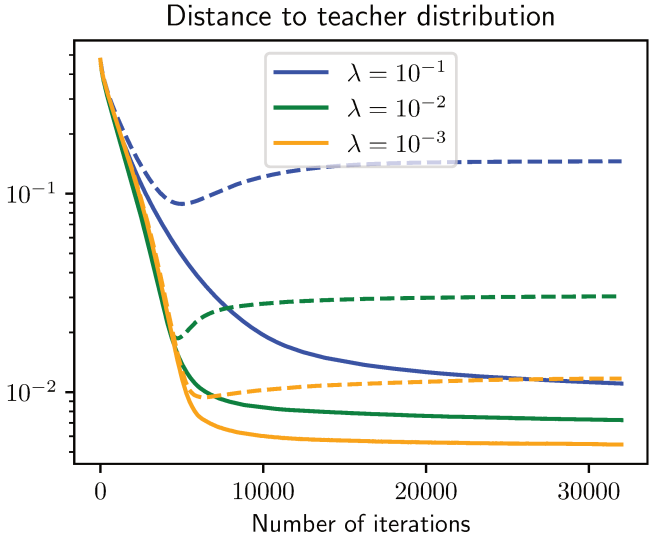


Figure III.13: Evolution of the MMD distance to the teacher distribution $\bar{\mu}_\gamma$ along gradient descent over the reduced risk (Eq. (III.42)) for a RBF neural network (Eq. (III.50)) of width $M = 1024$ with regularization $\lambda \in \{10^{-1}, 10^{-2}, 10^{-3}, 10^{-4}\}$. The regularization function is either $f_b : t \mapsto \frac{1}{2}t^2$ (dashed) or $f_u : t \mapsto \frac{1}{2}|t-1|^2$ (plain). Plots are averages over 6 independent runs.

while in the latter the convergence rate is linear, indicating a behavior closer to the ultra-fast diffusion limit (Theorem III.3). As for the 1-dimensional case, this can be explained by the fact that $\lambda = 10^{-1}$ is the order of magnitude of the most significant eigenvalues of the tangent kernel K_μ in Eq. (III.24). Thus, for higher values of λ, one enters in a high regularization regime where the reduced risk receive more influence from the MMD distance term than from the f-divergence term in Eq. (III.18). One also observes that the bias introduced in the case of the regularization $f_b : t \mapsto \frac{1}{2}t^2$ vanishes with the regularization strength λ, supporting the conclusions of our Proposition III.3.1.

Role of the shape of the teacher distribution Finally, we investigate the impact of the shape of the teacher distribution on the VarPro dynamic. We are particularly interested in the limit $\gamma = +\infty$ in which the teacher distribution is $\bar{\mu}_\gamma = \frac{1}{2}\delta_{\omega_1^*} + \frac{1}{2}\delta_{\omega_2^*}$. While such setting is not covered by our theory (in particular the ultra-fast diffusion equation is not necessarily well-posed), it is of interest to see if the VarPro algorithm is able to recover sparse feature representations.

In this context, we consider teacher distributions $\bar{\mu}_\gamma$ for $\gamma \in \{100, +\infty\}$ and train RBF neural networks (Eq. (III.50)) of fixed width $M = 1024$ with gradient descent over the reduced risk (Eq. (III.42)) with the unbiased regularization $f_u : t \mapsto \frac{1}{2}|t-1|^2$ and $\lambda = 10^{-3}$. Plots of the evolution of the reduced risk and of the MMD distance to the teacher distribution are reported in Fig. III.14. As in our 1-dimensional experiments, one can see that the convergence speed of VarPro deteriorates when γ increases, both in terms of convergence of the risk and in terms of convergence of the learned feature distribution to the teacher's. In case of a sparse teacher distribution ($\gamma = +\infty$), convergence towards the teacher seems to not necessarily be governed by a linear rate. Indeed, as the teacher distribution is not absolutely continuous, one could expect the comparison with ultra-fast diffusion dynamics to no longer hold.

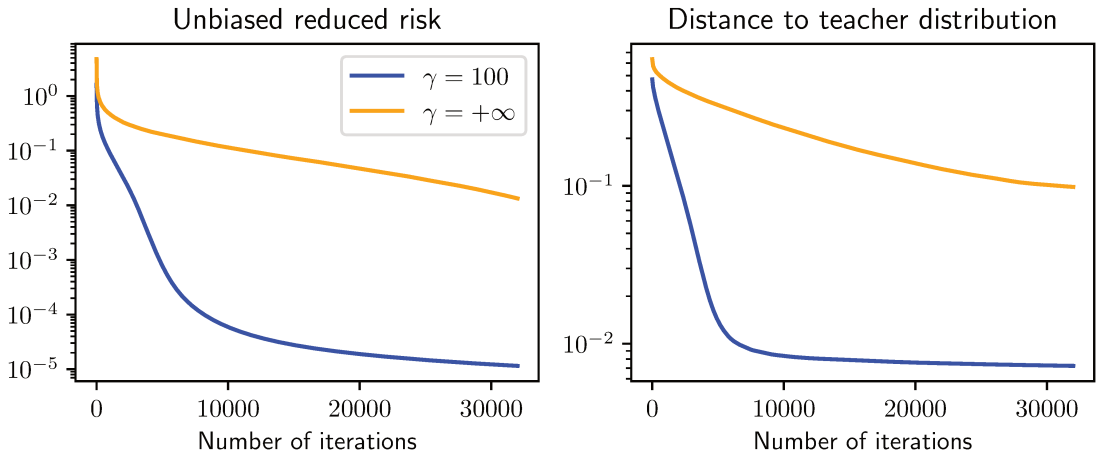


Figure III.14: Gradient descent over the reduced risk (Eq. (III.42)) for a RBF neural network (Eq. (III.50)) of width $M = 1024$ with unbiased regularization $f_u : t \mapsto \frac{1}{2}|t-1|^2$, $\lambda = 10^{-3}$ and teacher distributions $\bar{\mu}_\gamma$ for $\gamma \in \{100, +\infty\}$. Left: Evolution of the unbiased reduced risk. Right: Evolution of the MMD distance to the teacher distribution $\bar{\mu}_\gamma$. Plots are averages over 6 independent runs.

Conclusion

From a theoretical standpoint, the recent successes of learning algorithms have challenged our current understanding and highlighted the need for developing new theoretical frameworks. This line of research is driven by two objectives. First, although these models are often treated as black-boxes, gaining a deeper understanding of how learning algorithms work is essential for improving the interpretability of their predictions. Second, it can guide the design of more efficient architectures and training methods, ultimately leading to improved performance or reduced computational cost.

In this manuscript, we focused specifically on the case of overparameterized architectures, where the number of model parameters is large relative to the amount of training data. Residual architectures form an important class of such models — encompassing popular designs like ResNets and Transformers — thanks to the presence of skip connections, which enable the effective training of very deep networks. In this overparameterized regime, we saw that the choice of a scaling scheme to normalize parameters as the size of the model grows plays a critical role in the success of the learning process. Our study concentrated in particular on the mean-field limit with respect to network width and on the Neural ODE limit with respect to network depth in residual architectures.

These scaling choices are motivated not only by practical considerations, but also by theoretical ones. From a practical perspective, mean-field neural networks tend to exhibit stronger abilities to extract nonlinear representations from data, often resulting in better generalization. Neural ODEs, on the other hand, allow for more efficient training procedures. From a theoretical standpoint, Neural ODEs implement a smooth deformation of the input space, which contributes to a simpler optimization landscape and helps avoid spurious critical points. Also, for both shallow and deep architectures under the mean-field scaling, relaxation of the training objective in a space of measure yields favorable optimization properties. The training dynamics then correspond to interacting particle systems, which are "metric" gradient flows that can be analyzed through the lens of partial differential equations.

Mean-field limits of NODEs In Chapter I, we investigated the training dynamics of mean-field models of Neural ODEs (NODEs), which correspond to residual networks of infinite depth and arbitrarily large width. To analyze the training process in this dynamic, we developed a mathematical framework in which the model is parameterized by probability measures over a product space of layers and parameters, subject to a uniform marginal constraint on the space of layers. This parameter space is endowed with a Conditional Optimal Transport (COT) metric, designed to reflect the Euclidean geometry of the ResNet parameter space.

In Section I.3, we showed that the gradient flow dynamics of deep ResNets can be interpreted as a metric gradient flow with respect to the COT metric. As in the case

of shallow architectures, this dynamic is governed by an advection partial differential equation, whose well-posedness we established in Section I.3.4.

Convergence of training dynamics for NODEs In Chapter II, building on the framework developed in Chapter I, we studied the convergence properties of the gradient flow dynamics arising in the training of deep ResNets. A key finding is that, for these models, the training risk satisfies a local Polyak–Łojasiewicz (P–Ł) inequality. This structure allows us to establish linear convergence of the gradient flow toward an optimal parameterization, under the assumption that the learning problem is sufficiently "easy" to solve. In Theorems II.6 and II.7, we quantified this assumption in terms of the number of training samples. We also verified our theoretical results with numerical experiments on large-scale image classification tasks.

This analysis underscores the pivotal role played by the structure of residual blocks during training. Specifically, a lower bound on the P–Ł constant is governed by the conditioning of the tangent kernel associated with the residuals. We focused in particular on the case of linearly parameterized residuals, such as random feature models, as well as on the case of single-hidden-layer perceptrons. In both settings, we identified the importance of having a well-chosen distribution of features to ensure favorable convergence properties for gradient flow.

Feature learning in shallow architectures Finally, in Chapter III, we investigated feature learning behavior in the training of shallow neural network architectures. To this end, we considered the *Variable Projection (VarPro)* algorithm, which can be interpreted as a two-timescale version of gradient descent—where the linear parameters are updated "faster" than the nonlinear ones. In a teacher–student setting, we showed that this training dynamic yields linear convergence in approximating the teacher distribution. In contrast, existing convergence results for mean-field training of shallow neural networks typically only describe qualitative convergence, without providing explicit rates.

Once again, our results highlight the power of relaxing the learning problem to the space of parameter distributions, in which the training dynamics take the form of partial differential equations. More precisely, we showed that, in a regime of small regularization, the VarPro dynamics converge to the solution of a *weighted ultra-fast diffusion equation* — a nonlinear PDE whose long-time behavior was established in the literature.

We also validated our theory through numerical experiments. On simple learning tasks that satisfy our assumptions, we observed convergence towards ultra-fast diffusion. Moreover, the VarPro algorithm can be adapted to achieve state-of-the-art performance when training ResNets for image classification tasks.

Future research direction

Transformer architectures Transformer architectures now represent the prevailing approach for solving image classification and language modeling tasks. Architecturally, Transformers are also residual networks and can, like ResNets, be modeled by continuous-time dynamics. However, the key difference lies in the nature of their residual blocks, which are based on the attention mechanism described in Eq. (35). Unlike standard feedforward neural networks, attention-based models define sequence-to-sequence mappings that are inherently permutation-equivariant. From a mathematical standpoint, this corresponds to replacing the forward ODE in Eq. (42) with an interacting particle system governed by a nonlocal advection PDE [Sander, 2022a; Geshkovski, 2025].

Adapting our convergence results from Chapter II to the training of Transformer models thus represents an appealing research direction. Since our analysis relies on ensuring the expressivity of the residual maps at each layer, this would require a deeper understanding of the approximation capabilities of attention mechanisms within the space of permutation-equivariant sequence-to-sequence functions. Several recent works have begun to explore this direction [Geshkovski, 2024; Furuya, 2024].

Neural SDE scaling The Neural ODE scaling — in which residual branches are scaled by $1/D$, with D denoting the network depth — in association with smooth initialization of the weights can lead to more memory-efficient training of ResNets. However, it is observed in practice that a scaling of $1/\sqrt{D}$ combined with random initialization of the weights is more effective for feature learning and generalization [Yang, 2023]. Under this alternative scaling and initialization, the infinite-depth limit of ResNets is no longer described by a deterministic ODE, as in Eq. (42), but rather by a stochastic differential equation (SDE) [Marion, 2025].

Adapting our theoretical framework from Chapters I and II to the training of Neural SDE models thus presents a compelling research direction. A key challenge would be dealing with the inherently stochastic nature of the training dynamics. Some recent approaches have begun to address this issue using the formalism of *rough paths* [Gassiat, 2024].

Feature learning in deeper architectures As in the case of shallow architectures, we showed in Chapter II that learning meaningful feature representations from data is crucial to the success of the training process. However, in contrast with Chapter III, we were not able to quantify the extent to which such feature learning occurs during the training of deep ResNets. A natural extension of our work would therefore be to design and analyze algorithms that explicitly promote feature learning in deep neural networks.

In Chapter III, we studied the VarPro algorithm, which leverages the closed-form solution of the regression problem with respect to the linear parameters to eliminate them via partial optimization. A challenge in extending this strategy to deep architectures is that such closed-form elimination is no longer possible due to the compositional structure of multiple layers. However, we also interpreted VarPro as a two-timescale limit of gradient descent. This suggests a promising direction: studying two-timescale variants of gradient descent for the training of deep networks, and analyzing the limiting training dynamics as the timescale separation parameter tends to infinity.

Stochastic optimization In this manuscript, we have primarily focused on the analysis of deterministic training strategies, whereas in practice, stochastic optimization algorithms are the workhorse of modern deep learning. These methods enable the training of large-scale models on massive datasets and often improve generalization performance on unseen data. A natural and exciting extension of our work would be to adapt the convergence analyses from Chapters II and III to stochastic training frameworks.

From a mathematical perspective, the introduction of noise fundamentally alters the training dynamics and raise important questions about the modeling of the noise process. Several works have studied the training of shallow and deep networks under the assumption of isotropic noise [Mei, 2018; Jabir, 2019; Chizat, 2022; Nitanda, 2022], leading in the mean-field limit to Langevin dynamics where a linear diffusion term is added in Eq. (III.31). Although this assumption may be overly simplistic, Langevin dynamics reveal a compelling connection with the theory of sampling algorithms and lead to strong

convergence guarantees [Chewi, 2024]. A possible direction would be to try and extend those results to more realistic noise models.

On the numerical side, our experiments in Section III.6 showed that adapting the VarPro algorithm to a stochastic optimization setting requires specific modifications, such as the introduction of a momentum term. It would be interesting to investigate under what modeling assumptions such stochastic variants of VarPro lead to consistent mean-field dynamics, and whether one can still establish convergence rates for the learning of the feature distribution.

List of publications

Journal publications

R. B., Gabriel Peyré, François-Xavier Vialard. "Understanding the training of infinitely deep and wide resnets with conditional optimal transport". *Communications on Pure and Applied Mathematics*. (2025).

Conference proceedings

R. B., Gabriel Peyré, François-Xavier Vialard. "On global convergence of ResNets: From finite to infinite width using linear parameterization". *Advances in Neural Information Processing Systems*. (2022).

Preprints

R. B., Gabriel Peyré, François-Xavier Vialard. "Ultra-fast feature learning for the training of two-layer neural networks in the two-timescale regime". *arXiv preprint arXiv:2504.18208*. (2025).

References

[Achour, 2024] El Mehdi Achour, François Malgouyres, and Sébastien Gerchinovitz. "The loss landscape of deep linear neural networks: a second-order analysis". *Journal of Machine Learning Research* 25.242 (2024), pp. 1–76 (cit. on pp. 81, 91).

[Allen-Zhu, 2019] Zeyuan Allen-Zhu, Yuanzhi Li, and Zhao Song. "A convergence theory for deep learning via over-parameterization". *International Conference on Machine Learning*. PMLR. 2019, pp. 242–252 (cit. on pp. 8, 23, 36, 39, 80–82, 91, 92, 120).

[Ambrosio, 2008a] Luigi Ambrosio. "Transport equation and Cauchy problem for non-smooth vector fields". *Calculus of variations and nonlinear partial differential equations*. Springer, 2008, pp. 1–41 (cit. on p. 58).

[Ambrosio, 2013] Luigi Ambrosio, Alberto Bressan, Dirk Helbing, Axel Klar, Enrique Zuazua, Luigi Ambrosio, et al. "A user's guide to optimal transport". *Modelling and Optimisation of Flows on Networks: Cetraro, Italy 2009, Editors: Benedetto Piccoli, Michel Rascle* (2013), pp. 1–155 (cit. on pp. 59, 60, 87).

[Ambrosio, 2008b] Luigi Ambrosio, Nicola Gigli, and Giuseppe Savaré. "Gradient flows: in metric spaces and in the space of probability measures". *Lectures in mathematics ETH Zürich* (2008) (cit. on pp. 9, 15, 24, 30, 32, 33, 44, 45, 49, 50, 52, 53, 55, 59, 60, 65, 67, 69, 71, 87, 98, 124, 136, 138, 139).

[Ané, 2000] Cécile Ané, Sébastien Blachère, Djalil Chafaï, Pierre Fougères, Ivan Gentil, Florent Malrieu, et al. *Sur les inégalités de Sobolev logarithmiques*. Vol. 10. Société mathématique de France Paris, 2000 (cit. on p. 142).

[Arbel, 2019] Michael Arbel, Anna Korba, Adil Salim, and Arthur Gretton. "Maximum mean discrepancy gradient flow". *Advances in Neural Information Processing Systems* 32 (2019) (cit. on p. 126).

[Attouch, 2014] Hedy Attouch, Giuseppe Buttazzo, and Gérard Michaille. *Variational analysis in Sobolev and BV spaces: applications to PDEs and optimization*. SIAM, 2014 (cit. on p. 42).

[Ba, 2016] Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E. Hinton. "Layer normalization". *arXiv preprint arXiv:1607.06450* (2016) (cit. on pp. 7, 22).

[Bach, 2017a] Francis Bach. "Breaking the curse of dimensionality with convex neural networks". *The Journal of Machine Learning Research* 18.1 (2017), pp. 629–681 (cit. on pp. 8, 23, 24, 36, 109, 120).

[Bach, 2017b] Francis Bach. "On the equivalence between kernel quadrature rules and random feature expansions". *The Journal of Machine Learning Research* 18.1 (2017), pp. 714–751 (cit. on pp. 91, 109).

[Bach, 2004] Francis Bach, Gert R. G. Lanckriet, and Michael I. Jordan. "Multiple kernel learning, conic duality, and the SMO algorithm". *Proceedings of the twenty-first international conference on Machine learning*. 2004, p. 6 (cit. on p. 132).

[Bah, 2022] Bubacarr Bah, Holger Rauhut, Ulrich Terstiege, and Michael Westdickenberg. "Learning deep linear neural networks: Riemannian gradient flows and convergence to global minimizers". *Information and Inference: A Journal of the IMA* 11.1 (2022), pp. 307–353 (cit. on p. 39).

[Bahdanau, 2014] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. "Neural machine translation by jointly learning to align and translate". *arXiv preprint arXiv:1409.0473* (2014) (cit. on pp. 6, 21).

[Barron, 1993] Andrew R. Barron. "Universal approximation bounds for superpositions of a sigmoidal function". *IEEE Transactions on Information Theory* 39.3 (1993), pp. 930–945 (cit. on pp. 9, 24).

[Bartlett, 2018] Peter Bartlett, Dave Helmbold, and Philip Long. "Gradient descent with identity initialization efficiently learns positive definite linear transformations by deep residual networks". *International Conference on Machine Learning*. PMLR. 2018, pp. 521–530 (cit. on pp. 39, 80–82, 101).

[Baydin, 2018] Atilim Gunes Baydin, Barak A. Pearlmutter, Alexey Andreyevich Radul, and Jeffrey Mark Siskind. "Automatic differentiation in machine learning: a survey". *Journal of machine learning research* 18 (2018) (cit. on pp. 12, 27).

[Beg, 2005] Mirza Faisal Beg, Michael I. Miller, Alain Trouvé, and Laurent Younes. "Computing large deformation metric mappings via geodesic flows of diffeomorphisms". *International journal of computer vision* 61 (2005), pp. 139–157 (cit. on p. 95).

[Belkin, 2019] Mikhail Belkin, Daniel Hsu, Siyuan Ma, and Soumik Mandal. "Reconciling modern machine-learning practice and the classical bias–variance trade-off". *Proceedings of the National Academy of Sciences* 116.32 (2019), pp. 15849–15854 (cit. on pp. 7, 22).

[Bengio, 1994] Yoshua Bengio, Patrice Simard, and Paolo Frasconi. "Learning long-term dependencies with gradient descent is difficult". *IEEE Transactions on Neural Networks* 5.2 (1994), pp. 157–166 (cit. on pp. 10, 25, 40).

[Benning, 2018] Martin Benning and Martin Burger. "Modern regularization methods for inverse problems". *Acta numerica* 27 (2018), pp. 1–111 (cit. on p. 147).

[Berthier, 2024] Raphaël Berthier, Andrea Montanari, and Kangjie Zhou. "Learning time-scales in two-layers neural networks". *Foundations of Computational Mathematics* (2024), pp. 1–84 (cit. on pp. 14, 29, 126).

[Bietti, 2023] Alberto Bietti, Joan Bruna, and Loucas Pillaud-Vivien. "On learning gaussian multi-index models with gradient flow". *arXiv preprint arXiv:2310.19793* (2023) (cit. on pp. 14, 29, 126).

[Blanchet, 2018] Adrien Blanchet and Jérôme Bolte. "A family of functional inequalities: Łojasiewicz inequalities and displacement convex functions". *Journal of Functional Analysis* 275.7 (2018), pp. 1650–1673 (cit. on p. 88).

[Bogachev, 2007] Vladimir I. Bogachev. *Measure Theory*. Springer Berlin, Heidelberg, 2007 (cit. on p. 48).

[Bolte, 2010] Jérôme Bolte, Aris Daniilidis, Olivier Ley, and Laurent Mazet. "Characterizations of Łojasiewicz inequalities: subgradient flows, talweg, convexity". *Transactions of the American Mathematical Society* 362.6 (2010), pp. 3319–3363 (cit. on p. 88).

[Bolte, 2021] Jérôme Bolte, Tam Le, Edouard Pauwels, and Tony Silveti-Falls. "Nonsmooth implicit differentiation for machine-learning and optimization". *Advances in neural information processing systems* 34 (2021), pp. 13537–13549 (cit. on p. 118).

[Bordelon, 2025] Blake Bordelon and Cengiz Pehlevan. "Deep Linear Network Training Dynamics from Random Initialization: Data, Width, Depth, and Hyperparameter Transfer". *arXiv preprint arXiv:2502.02531* (2025) (cit. on pp. 7, 22, 27).

[Borkar, 1997] Vivek S. Borkar. "Stochastic approximation with two time scales". *Systems & Control Letters* 29.5 (1997), pp. 291–294 (cit. on p. 126).

[Borkar, 2008] Vivek S. Borkar. *Stochastic approximation: a dynamical systems viewpoint*. Vol. 9. Springer, 2008 (cit. on p. 126).

[Bottou, 2018] Léon Bottou, Frank E. Curtis, and Jorge Nocedal. "Optimization methods for large-scale machine learning". *SIAM review* 60.2 (2018), pp. 223–311 (cit. on pp. 13, 28, 122).

[Boufadène, 2023] Siwan Boufadène and François-Xavier Vialard. "On the global convergence of Wasserstein gradient flow of the Coulomb discrepancy" (2023) (cit. on p. 126).

[Brezis, 1973] Haïm Brezis. *Opérateurs maximaux monotones et semi-groupes de contractions dans les espaces de Hilbert*. Vol. 5. Elsevier, 1973 (cit. on p. 83).

[Caglioti, 2018] Emanuele Caglioti, François Golse, and Mikaela Iacobelli. "Quantization of Measures and Gradient Flows: a Perturbative Approach in the 2-Dimensional Case". *Annales de l'Institut Henri Poincaré C, Analyse non linéaire*. Vol. 35. 2018, pp. 1531–1555 (cit. on pp. 124, 140).

[Cannarsa, 2015] Piermarco Cannarsa and Teresa D'Aprile. *Introduction to measure theory and functional analysis*. Vol. 89. Springer, 2015 (cit. on p. 50).

[Carmeli, 2010] Claudio Carmeli, Ernesto De Vito, Alessandro Toigo, and Veronica Umanitá. "Vector valued reproducing kernel Hilbert spaces and universality". *Analysis and Applications* 8.01 (2010), pp. 19–61 (cit. on pp. 91, 93, 108, 109).

[Chatterjee, 2022] Sourav Chatterjee. "Convergence of gradient descent for deep neural networks". *arXiv preprint arXiv:2203.16462* (2022) (cit. on pp. 34, 84–86).

[Chemseddine, 2024] Jannis Chemseddine, Paul Hagemann, Christian Wald, and Gabriele Steidl. "Conditional Wasserstein Distances with Applications in Bayesian OT Flow Matching". *arXiv preprint arXiv:2403.18705* (2024) (cit. on pp. 31, 44).

[Chen, 2018] Ricky T. Q. Chen, Yulia Rubanova, Jesse Bettencourt, and David Duvenaud. "Neural Ordinary Differential Equations". *Advances in Neural Information Processing Systems* (2018) (cit. on pp. 11, 12, 26, 32, 40–42, 45, 55, 59, 80, 113, 114).

[Chen, 2023] Yihang Chen, Fanghui Liu, Yiping Lu, Grigorios Chrysos, and Volkan Cevher. "Generalization Guarantees of Deep ResNets in the Mean-Field Regime". *NeurIPS 2023 Workshop on Mathematics of Modern Machine Learning*. 2023 (cit. on p. 44).

[Chen, 2020] Zixiang Chen, Yuan Cao, Difan Zou, and Quanquan Gu. "How Much Over-parameterization Is Sufficient to Learn Deep ReLU Networks?" *International Conference on Learning Representations*. 2020 (cit. on pp. 39, 80, 81).

[Chen, 2024] Zonghao Chen, Aratrika Mustafi, Pierre Glaser, Anna Korba, Arthur Gretton, and Bharath K. Sriperumbudur. "(De)-regularized Maximum Mean Discrepancy Gradient Flow". *arXiv preprint arXiv:2409.14980* (2024) (cit. on pp. 126, 130).

[Chewi, 2024] Sinho Chewi, Murat A. Erdogdu, Mufan Li, Ruoqi Shen, and Matthew S. Zhang. "Analysis of langevin monte carlo from poincare to log-sobolev". *Foundations of Computational Mathematics* (2024), pp. 1–51 (cit. on pp. 142, 170).

[Chewi, 2020] Sinho Chewi, Thibaut Le Gouic, Chen Lu, Tyler Maunu, and Philippe Rigollet. "SVGD as a kernelized Wasserstein gradient flow of the chi-squared divergence". *Advances in Neural Information Processing Systems* 33 (2020), pp. 2098–2109 (cit. on p. 142).

[Chill, 2003] Ralph Chill. "On the Łojasiewicz–Simon gradient inequality". *Journal of Functional Analysis* 201.2 (2003), pp. 572–601 (cit. on p. 84).

[Chizat, 2022] Lénaïc Chizat. "Mean-Field Langevin Dynamics: Exponential Convergence and Annealing". *Transactions on Machine Learning Research* (2022) (cit. on pp. 9, 15, 24, 30, 126, 169).

[Chizat, 2018] Lénaïc Chizat and Francis Bach. "On the Global Convergence of Gradient Descent for Over-parameterized Models using Optimal Transport". *Advances in Neural Information Processing Systems* 31 (2018), pp. 3036–3046 (cit. on pp. 8, 9, 11, 15, 24, 26, 30, 31, 36, 40, 41, 80, 87, 101, 120, 121, 125).

[Chizat, 2019] Lénaïc Chizat, Edouard Oyallon, and Francis Bach. "On Lazy Training in Differentiable Programming". *NeurIPS 2019-33rd Conference on Neural Information Processing Systems*. 2019, pp. 2937–2947 (cit. on pp. 8, 23, 81, 111, 120).

[Cho, 2009] Youngmin Cho and Lawrence Saul. "Kernel methods for deep learning". *Advances in neural information processing systems* 22 (2009) (cit. on pp. 36, 109).

[Cohen, 2021] Alain-Sam Cohen, Rama Cont, Alain Rossier, and Renyuan Xu. "Scaling properties of deep residual networks". *International Conference on Machine Learning*. PMLR. 2021, pp. 2039–2048 (cit. on p. 27).

[Cybenko, 1989] George Cybenko. "Approximation by superpositions of a sigmoidal function". *Mathematics of control, signals and systems* 2.4 (1989), pp. 303–314 (cit. on pp. 9, 24, 107, 121).

[De Castro, 2012] Yohann De Castro and Fabrice Gamboa. "Exact reconstruction using Beurling minimal extrapolation". *Journal of Mathematical Analysis and applications* 395.1 (2012), pp. 336–354 (cit. on pp. 122, 162).

[De Giorgi, 1993] Ennio De Giorgi. "New problems on minimizing movements". *Boundary value problems for PDE and applications* (1993) (cit. on p. 68).

[Dello Schiavo, 2024] Lorenzo Dello Schiavo, Jan Maas, and Francesco Pedrotti. "Local conditions for global convergence of gradient flows and proximal point sequences in metric spaces". *Transactions of the American Mathematical Society* 377.06 (2024), pp. 3779–3804 (cit. on pp. 34, 60, 82, 88).

[Deng, 2009] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. "Imagenet: A large-scale hierarchical image database". *2009 IEEE conference on computer vision and pattern recognition*. Ieee. 2009, pp. 248–255 (cit. on pp. 11, 26).

[Devlin, 2019] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. "Bert: Pre-training of deep bidirectional transformers for language understanding". *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*. 2019, pp. 4171–4186 (cit. on pp. 6, 21).

[Ding, 2021] Zhiyan Ding, Shi Chen, Qin Li, and Stephen Wright. "On the global convergence of gradient descent for multi-layer resnets in the mean-field regime". *arXiv preprint arXiv:2110.02926* (2021) (cit. on pp. 15, 30, 44, 82).

[Ding, 2022] Zhiyan Ding, Shi Chen, Qin Li, and Stephen J. Wright. "Overparameterization of deep ResNet: zero loss and mean-field analysis". *The Journal of Machine Learning Research* 23.1 (2022), pp. 2282–2346 (cit. on pp. 34, 42, 44, 82, 83).

[Donoho, 2000] David L. Donoho et al. "High-dimensional data analysis: The curses and blessings of dimensionality". *AMS math challenges lecture* 1.2000 (2000), p. 32 (cit. on p. 143).

[Dosovitskiy, 2020] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, et al. "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale". *International Conference on Learning Representations*. 2020 (cit. on pp. 6, 21).

[Du, 2019] Simon S. Du, Jason Lee, Haochuan Li, Liwei Wang, and Xiyu Zhai. "Gradient descent finds global minima of deep neural networks". *International Conference on Machine Learning*. PMLR. 2019, pp. 1675–1685 (cit. on pp. 8, 23, 36, 39, 80–82, 91, 92, 120).

[Du, 2018] Simon S. Du, Xiyu Zhai, Barnabas Poczos, and Aarti Singh. "Gradient Descent Provably Optimizes Over-parameterized Neural Networks". *International Conference on Learning Representations*. 2018 (cit. on pp. 80, 81).

[Duval, 2015] Vincent Duval and Gabriel Peyré. "Exact support recovery for sparse spikes deconvolution". *Foundations of Computational Mathematics* 15.5 (2015), pp. 1315–1355 (cit. on pp. 122, 162).

[E, 2019] Weinan E, Jiequn Han, and Qianxiao Li. "A mean-field optimal control formulation of deep learning". *Research in the Mathematical Sciences* 6.1 (2019), p. 10 (cit. on pp. 12, 27, 40).

[E, 2021] Weinan E, Chao Ma, and Lei Wu. "The Barron Space and the Flow-Induced Function Spaces for Neural Network Models". *Constructive Approximation* (2021) (cit. on pp. 9, 12, 24, 27, 40, 44).

[E, 2022] Weinan E and Stephan Wojtowytsch. "Representation formulas and pointwise properties for Barron functions". *Calculus of Variations and Partial Differential Equations* 61.2 (2022), p. 46 (cit. on pp. 9, 10, 24, 25, 106).

[Feinberg, 2020] Eugene A. Feinberg, Pavlo O. Kasyanov, and Yan Liang. "Fatou's lemma for weakly converging measures under the uniform integrability condition". *Theory of Probability & Its Applications* 64.4 (2020), pp. 615–630 (cit. on p. 71).

[Fournier, 2015] Nicolas Fournier and Arnaud Guillin. "On the rate of convergence in Wasserstein distance of the empirical measure". *Probability theory and related fields* 162.3 (2015), pp. 707–738 (cit. on p. 140).

[Furuya, 2024] Takashi Furuya, Maarten V. de Hoop, and Gabriel Peyré. "Transformers are universal in-context learners". *arXiv preprint arXiv:2408.01367* (2024) (cit. on p. 169).

[Gao, 2024] Cheng Gao, Yuan Cao, Zihao Li, Yihan He, Mengdi Wang, Han Liu, et al. "Global convergence in training large-scale transformers". *Advances in Neural Information Processing Systems* 37 (2024), pp. 29213–29284 (cit. on p. 159).

[Gassiat, 2024] Paul Gassiat and Florin Suciu. "A gradient flow on control space with rough initial condition". *arXiv preprint arXiv:2407.11817* (2024) (cit. on p. 169).

[Geshkovski, 2025] Borjan Geshkovski, Cyril Letrouit, Yury Polyanskiy, and Philippe Rigollet. "A mathematical perspective on transformers". *Bulletin of the American Mathematical Society* 62.3 (2025), pp. 427–479 (cit. on p. 168).

[Geshkovski, 2024] Borjan Geshkovski, Philippe Rigollet, and Domènec Ruiz-Balet. "Measure-to-measure interpolation using Transformers". *arXiv preprint arXiv:2411.04551* (2024) (cit. on p. 169).

[Ghorbani, 2019] Behrooz Ghorbani, Song Mei, Theodor Misiakiewicz, and Andrea Montanari. "Limitations of lazy training of two-layers neural network". *Advances in Neural Information Processing Systems* 32 (2019) (cit. on pp. 8, 23).

[Ghorbani, 2020] Behrooz Ghorbani, Song Mei, Theodor Misiakiewicz, and Andrea Montanari. "When do neural networks outperform kernel methods?" *Advances in Neural Information Processing Systems* 33 (2020), pp. 14820–14830 (cit. on pp. 8, 23, 120).

[Glaser, 2021] Pierre Glaser, Michael Arbel, and Arthur Gretton. "KALE flow: A relaxed KL gradient flow for probabilities with disjoint support". *Advances in Neural Information Processing Systems* 34 (2021), pp. 8018–8031 (cit. on pp. 126, 130, 144).

[Glorot, 2010] Xavier Glorot and Yoshua Bengio. "Understanding the difficulty of training deep feedforward neural networks". *Proceedings of the thirteenth international conference on artificial intelligence and statistics*. JMLR Workshop and Conference Proceedings. 2010, pp. 249–256 (cit. on pp. 10, 25, 40).

[Golub, 1973] Gene H. Golub and Victor Pereyra. "The differentiation of pseudo-inverses and nonlinear least squares problems whose variables separate". *SIAM Journal on numerical analysis* 10.2 (1973), pp. 413–432 (cit. on pp. 14, 29, 122).

[Golub, 2003] Gene H. Golub and Victor Pereyra. "Separable nonlinear least squares: the variable projection method and itsapplications". *Inverse problems* 19.2 (2003), R1 (cit. on p. 122).

[Goodfellow, 2016] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep Learning*. MIT Press, 2016 (cit. on p. 120).

[Gretton, 2012] Arthur Gretton, Karsten M Borgwardt, Malte J. Rasch, Bernhard Schölkopf, and Alexander Smola. "A kernel two-sample test". *The Journal of Machine Learning Research* 13.1 (2012), pp. 723–773 (cit. on pp. 130, 160).

[Gupta, 2018] Vineet Gupta, Tomer Koren, and Yoram Singer. "Shampoo: Preconditioned stochastic tensor optimization". *International Conference on Machine Learning*. PMLR. 2018, pp. 1842–1850 (cit. on p. 157).

[Hale, 2009] Jack K. Hale. *Ordinary differential equations*. Courier Corporation, 2009 (cit. on pp. 43, 139).

[Hardt, 2016a] Moritz Hardt and Tengyu Ma. "Identity Matters in Deep Learning". *International Conference on Learning Representations*. 2016 (cit. on pp. 39, 80, 81, 92).

[Hardt, 2016b] Moritz Hardt, Ben Recht, and Yoram Singer. "Train faster, generalize better: Stability of stochastic gradient descent". *International conference on machine learning*. PMLR. 2016, pp. 1225–1234 (cit. on pp. 13, 28).

[Hauer, 2019] Daniel Hauer and José Mazón. "Kurdyka–Łojasiewicz–Simon inequality for gradient flows in metric spaces". *Transactions of the American Mathematical Society* 372.7 (2019), pp. 4917–4976 (cit. on pp. 60, 82, 88).

[He, 2016a] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. "Deep residual learning for image recognition". *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016, pp. 770–778 (cit. on pp. 10, 11, 25, 26, 31, 40, 41, 80, 113, 116, 155, 157, 158).

[He, 2016b] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. "Identity mappings in deep residual networks". *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part IV 14*. Springer. 2016, pp. 630–645 (cit. on pp. 10, 11, 25, 26, 40).

[Hertrich, 2023a] Johannes Hertrich, Robert Beinert, Manuel Gräf, and Gabriele Steidl. "Wasserstein gradient flows of the discrepancy with distance kernel on the line". *International Conference on Scale Space and Variational Methods in Computer Vision*. Springer. 2023, pp. 431–443 (cit. on p. 126).

[Hertrich, 2024] Johannes Hertrich, Manuel Gräf, Robert Beinert, and Gabriele Steidl. "Wasserstein steepest descent flows of discrepancies with Riesz kernels". *Journal of Mathematical Analysis and Applications* 531.1 (2024), p. 127829 (cit. on p. 126).

[Hertrich, 2023b] Johannes Hertrich, Christian Wald, Fabian Altekrüger, and Paul Hagemann. "Generative sliced MMD flows with Riesz kernels". *arXiv preprint arXiv:2305.11463* (2023) (cit. on p. 126).

[Hindmarsh, 1983] Alan C. Hindmarsh. "ODEPACK, a systemized collection of ODE solvers". *Scientific computing* (1983) (cit. on p. 150).

[Hinton, 2012] Geoffrey E. Hinton. "Lecture 6.5-rmsprop: Divide the gradient by a running average of its recent magnitude". *COURSERA: Neural networks for machine learning* 4.2 (2012), p. 26 (cit. on pp. 13, 28).

[Hofmann, 2008] Thomas Hofmann, Bernhard Schölkopf, and Alexander J. Smola. "Kernel methods in machine learning" (2008) (cit. on pp. 120, 123).

[Hosseini, 2025] Bamdad Hosseini, Alexander W. Hsu, and Amirhossein Taghvaei. "Conditional optimal transport on function spaces". *SIAM/ASA Journal on Uncertainty Quantification* 13.1 (2025), pp. 304–338 (cit. on pp. 31, 43, 44, 47).

[Hu, 2021] Kaitong Hu, Zhenjie Ren, David Šiška, and Łukasz Szpruch. "Mean-field Langevin dynamics and energy landscape of neural networks". *Annales de l'Institut Henri Poincare (B) Probabilites et statistiques*. Vol. 57. 4. Institut Henri Poincaré. 2021, pp. 2043–2065 (cit. on p. 126).

[Iacobelli, 2019a] Mikaela Iacobelli. "Asymptotic analysis for a very fast diffusion equation arising from the 1D quantization problem". *Discrete and Continuous Dynamical Systems* 39.9 (2019), pp. 4929–4943 (cit. on pp. 124, 140).

[Iacobelli, 2019b] Mikaela Iacobelli, Francesco S. Patacchini, and Filippo Santambrogio. "Weighted ultrafast diffusion equations: from well-posedness to long-time behaviour". *Archive for Rational Mechanics and Analysis* 232 (2019), pp. 1165–1206 (cit. on pp. 36, 38, 124, 125, 136, 140–142).

[Iglesias, 2018] José A. Iglesias, Gwenael Mercier, and Otmar Scherzer. "A note on convergence of solutions of total variation regularized linear inverse problems". *Inverse Problems* 34.5 (2018), p. 055011 (cit. on p. 147).

[Ioffe, 2015] Sergey Ioffe and Christian Szegedy. "Batch normalization: Accelerating deep network training by reducing internal covariate shift". *International conference on machine learning*. pmlr. 2015, pp. 448–456 (cit. on pp. 7, 22).

[Isobe, 2023] Noboru Isobe. "A Convergence result of a continuous model of deep learning via Łojasiewicz–Simon inequality". *arXiv preprint arXiv:2311.15365* (2023) (cit. on pp. 15, 30, 34, 42, 44, 82).

[Jabir, 2019] Jean-François Jabir, David Šiška, and Łukasz Szpruch. "Mean-field neural odes via relaxed optimal control". *arXiv preprint arXiv:1912.05475* (2019) (cit. on p. 169).

[Jacot, 2018] Arthur Jacot, Franck Gabriel, and Clément Hongler. "Neural tangent kernel: Convergence and generalization in neural networks". *Advances in Neural Information Processing Systems* 31 (2018) (cit. on pp. 8, 23, 31, 81, 91, 101, 120).

[Javanmard, 2020] Adel Javanmard, Marco Mondelli, and Andrea Montanari. "Analysis of a two-layer neural network via displacement convexity". *The Annals of Statistics* 48.6 (2020) (cit. on pp. 40, 80).

[Jordan, 1998] Richard Jordan, David Kinderlehrer, and Felix Otto. "The variational formulation of the Fokker–Planck equation". *SIAM journal on mathematical analysis* 29.1 (1998), pp. 1–17 (cit. on pp. 15, 30, 87, 136).

[Karamichailidou, 2024] Despina Karamichailidou, Georgios Gerolymatos, Panagiotis Patrinos, Haralambos Sarimveis, and Alex Alexandridis. "Radial basis function neural network training using variable projection and fuzzy means". *Neural Computing and Applications* 36.33 (2024), pp. 21137–21151 (cit. on pp. 121, 122).

[Karimi, 2016] Hamed Karimi, Julie Nutini, and Mark Schmidt. "Linear convergence of gradient and proximal-gradient methods under the Polyak-Łojasiewicz condition". *Joint European conference on machine learning and knowledge discovery in databases*. Springer. 2016, pp. 795–811 (cit. on pp. 84, 113).

[Kerrigan, 2024] Gavin Kerrigan, Giosue Migliorini, and Padhraic Smyth. "Dynamic conditional optimal transport through simulation-free flows". *Advances in Neural Information Processing Systems* 37 (2024), pp. 93602–93642 (cit. on p. 44).

[Kingma, 2014] Diederik P. Kingma and Jimmy Ba. "Adam: A method for stochastic optimization". *arXiv preprint arXiv:1412.6980* (2014) (cit. on pp. 13, 28).

[Kobyzev, 2020] Ivan Kobyzev, Simon Prince, and Marcus Brubaker. "Normalizing Flows: An Introduction and Review of Current Methods". *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2020), pp. 1–1 (cit. on pp. 12, 26, 95).

[Krizhevsky, 2009] Alex Krizhevsky, Geoffrey Hinton, et al. "Learning multiple layers of features from tiny images" (2009) (cit. on pp. 11, 26, 36, 83, 112, 148, 155).

[Kurdyka, 1998] Krzysztof Kurdyka. "On gradients of functions definable in o-minimal structures". *Annales de l'institut Fourier*. Vol. 48. 3. 1998, pp. 769–783 (cit. on p. 84).

[Lanckriet, 2004] Gert R. G. Lanckriet, Nello Cristianini, Peter Bartlett, Laurent El Ghaoui, and Michael I. Jordan. "Learning the kernel matrix with semidefinite programming". *Journal of Machine learning research* 5.Jan (2004), pp. 27–72 (cit. on pp. 132, 134).

[LeCun, 2015] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. "Deep learning". *Nature* 521.7553 (2015), pp. 436–444 (cit. on pp. 5, 20).

[LeCun, 1989] Yann LeCun, Bernhard Boser, John Denker, Donnie Henderson, Richard Howard, Wayne Hubbard, et al. "Handwritten digit recognition with a back-propagation network". *Advances in neural information processing systems* 2 (1989) (cit. on pp. 5, 20).

[LeCun, 2010] Yann LeCun, Corinna Cortes, Chris Burges, et al. *MNIST handwritten digit database*. 2010 (cit. on pp. 83, 112, 113).

[Lee, 2019] Jaehoon Lee, Lechao Xiao, Samuel Schoenholz, Yasaman Bahri, Roman Novak, Jascha Sohl-Dickstein, et al. "Wide neural networks of any depth evolve as linear models under gradient descent". *Advances in neural information processing systems* 32 (2019), pp. 8572–8583 (cit. on pp. 8, 23, 36, 39, 80, 81, 91, 120).

[Li, 2018] Yuanzhi Li and Yingyu Liang. "Learning overparameterized neural networks via stochastic gradient descent on structured data". *Advances in neural information processing systems* 31 (2018) (cit. on pp. 80, 81).

[Li, 2017] Yuanzhi Li and Yang Yuan. "Convergence Analysis of Two-layer Neural Networks with ReLU Activation". *Advances in Neural Information Processing Systems* 30 (2017), pp. 597–607 (cit. on pp. 39, 80, 81).

[Liero, 2018] Matthias Liero, Alexander Mielke, and Giuseppe Savaré. "Optimal entropy-transport problems and a new Hellinger–Kantorovich distance between positive measures". *Inventiones mathematicae* 211.3 (2018), pp. 969–1117 (cit. on p. 129).

[Liu, 2020] Chaoyue Liu, Libin Zhu, and Mikhail Belkin. "On the linearity of large non-linear models: when and why the tangent kernel is constant". *Advances in Neural Information Processing Systems* 33 (2020) (cit. on pp. 80–82, 91, 101, 120).

[Łojasiewicz, 1963] Stanislaw Łojasiewicz. "A topological property of real analytic subsets". *Coll. du CNRS, Les équations aux dérivées partielles* 117.87-89 (1963), p. 2 (cit. on p. 84).

[Lu, 2020] Yiping Lu, Chao Ma, Yulong Lu, Jianfeng Lu, and Lexing Ying. "A mean field analysis of deep ResNet and beyond: Towards provably optimization via overparameterization from depth". *International Conference on Machine Learning*. PMLR. 2020, pp. 6426–6436 (cit. on pp. 15, 30, 34, 42, 44, 82, 83).

[Marion, 2023a] Pierre Marion and Raphaël Berthier. "Leveraging the two-timescale regime to demonstrate convergence of neural networks". *Advances in Neural Information Processing Systems* 36 (2023), pp. 64996–65029 (cit. on pp. 14, 29, 126).

[Marion, 2025] Pierre Marion, Adeline Fermanian, Gérard Biau, and Jean-Philippe Vert. "Scaling resnets in the large-depth regime". *Journal of Machine Learning Research* 26.56 (2025), pp. 1–48 (cit. on pp. 27, 169).

[Marion, 2023b] Pierre Marion, Yu-Han Wu, Michael E. Sander, and Gérard Biau. "Implicit regularization of deep residual networks towards neural ODEs" (2023) (cit. on pp. 12, 27, 33, 34, 44, 80, 82).

[Mei, 2019] Song Mei, Theodor Misiakiewicz, and Andrea Montanari. "Mean-field theory of two-layers neural networks: dimension-free bounds and kernel limit". *Conference on Learning Theory*. PMLR. 2019, pp. 2388–2464 (cit. on pp. 8, 9, 15, 24, 30, 120, 121, 126).

[Mei, 2018] Song Mei, Andrea Montanari, and Phan-Minh Nguyen. "A mean field view of the landscape of two-layer neural networks". *Proceedings of the National Academy of Sciences* 115.33 (2018), E7665–E7671 (cit. on pp. 40, 41, 80, 87, 169).

[Micchelli, 2006] Charles A. Micchelli, Yuesheng Xu, and Haizhang Zhang. "Universal Kernels." *Journal of Machine Learning Research* 7.12 (2006) (cit. on pp. 92, 161).

[Milgrom, 2002] Paul Milgrom and Ilya Segal. "Envelope theorems for arbitrary choice sets". *Econometrica* 70.2 (2002), pp. 583–601 (cit. on p. 137).

[Montufar, 2014] Guido F. Montufar, Razvan Pascanu, Kyunghyun Cho, and Yoshua Bengio. "On the number of linear regions of deep neural networks". *Advances in neural information processing systems* 27 (2014) (cit. on pp. 10, 25).

[Muandet, 2017] Krikamol Muandet, Kenji Fukumizu, Bharath Sriperumbudur, Bernhard Schölkopf, et al. "Kernel mean embedding of distributions: A review and beyond". *Foundations and Trends in Machine Learning* 10.1-2 (2017), pp. 1–141 (cit. on pp. 130, 160).

[Muratori, 2020] Matteo Muratori and Giuseppe Savaré. "Gradient flows and evolution variational inequalities in metric spaces. I: Structural properties". *Journal of Functional Analysis* 278.4 (2020), p. 108347 (cit. on p. 60).

[Nesterov, 1983] Yurii Nesterov. "A method of solving a convex programming problem with convergence rate $\mathcal{O}(1/k^2)$". *Doklady Akademii Nauk*. Vol. 269. 3. Russian Academy of Sciences. 1983, pp. 543–547 (cit. on pp. 13, 28).

[Neumayer, 2024] Sebastian Neumayer, Viktor Stein, and Gabriele Steidl. "Wasserstein Gradient Flows for Moreau Envelopes of f-Divergences in Reproducing Kernel Hilbert Spaces". *arXiv preprint arXiv:2402.04613* (2024) (cit. on pp. 126, 127, 130, 134).

[Newman, 2021] Elizabeth Newman, Lars Ruthotto, Joseph Hart, and Bart van Bloemen Waanders. "Train like a (Var) Pro: Efficient training of neural networks with variable projection". *SIAM Journal on Mathematics of Data Science* 3.4 (2021), pp. 1041–1066 (cit. on pp. 122, 155).

[Nguyen, 2023] Phan-Minh Nguyen and Huy Tuan Pham. "A rigorous framework for the mean field limit of multilayer neural networks". *Mathematical Statistics and Learning* 6.3 (2023), pp. 201–357 (cit. on pp. 40, 41).

[Nguyen, 2021] Quynh Nguyen. "On the proof of global convergence of gradient descent for deep relu networks with linear widths". *International Conference on Machine Learning*. PMLR. 2021, pp. 8056–8062 (cit. on pp. 39, 80, 81).

[Niethammer, 2011] Marc Niethammer, Yang Huang, and François-Xavier Vialard. "Geodesic regression for image time-series". *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2011: 14th International Conference, Toronto, Canada, September 18-22, 2011, Proceedings, Part II 14*. Springer. 2011, pp. 655–662 (cit. on p. 95).

[Nitanda, 2022] Atsushi Nitanda, Denny Wu, and Taiji Suzuki. "Convex analysis of the mean field langevin dynamics". *International Conference on Artificial Intelligence and Statistics*. PMLR. 2022, pp. 9741–9757 (cit. on pp. 9, 15, 24, 30, 126, 169).

[Oh, 2025] YongKyung Oh, Seungsu Kam, Jonghun Lee, Dong-Young Lim, Sungil Kim, and Alex Bui. "Comprehensive review of neural differential equations for time series analysis". *arXiv preprint arXiv:2502.09885* (2025) (cit. on pp. 11, 26).

[OpenAI, 2023] OpenAI. "Gpt-4 technical report". *arXiv preprint arXiv:2303.08774* (2023) (cit. on pp. 7, 22, 27).

[Osborne, 2007] Michael R. Osborne. "Separable least squares, variable projection, and the Gauss-Newton algorithm". *Electronic Transactions on Numerical Analysis* 28.2 (2007), pp. 1–15 (cit. on p. 122).

[Oymak, 2019] Samet Oymak and Mahdi Soltanolkotabi. "Overparameterized nonlinear learning: Gradient descent takes the shortest path?" *International Conference on Machine Learning*. PMLR. 2019, pp. 4951–4960 (cit. on pp. 34, 84).

[Paszke, 2017] Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, et al. "Automatic differentiation in pytorch" (2017) (cit. on pp. 13, 28, 112).

[Pavliotis, 2014] Grigorios A. Pavliotis. "Stochastic processes and applications". *Texts in applied mathematics* 60 (2014) (cit. on p. 141).

[Payne, 1960] Lawrence E. Payne and Hans F. Weinberger. "An optimal Poincaré inequality for convex domains". *Archive for Rational Mechanics and Analysis* 5.1 (1960), pp. 286–292 (cit. on p. 143).

[Pereyra, 2006] Víctor Pereyra, Godela Scherer, and F. Wong. "Variable projections neural network training". *Mathematics and Computers in Simulation* 73.1-4 (2006), pp. 231–243 (cit. on pp. 121, 122).

[Peszek, 2023] Jan Peszek and David Poyato. "Heterogeneous gradient flows in the topology of fibered optimal transport". *Calculus of Variations and Partial Differential Equations* 62.9 (2023), p. 258 (cit. on pp. 31, 44, 45).

[Petersen, 2020] Philipp Petersen and Felix Voigtlaender. "Equivalence of approximation by convolutional neural networks and fully-connected networks". *Proceedings of the American Mathematical Society* 148.4 (2020), pp. 1567–1581 (cit. on p. 25).

[Polyak, 1963] Boris T. Polyak. "Gradient methods for the minimisation of functionals". *USSR Computational Mathematics and Mathematical Physics* 3.4 (1963), pp. 864–878 (cit. on pp. 34, 84).

[Polyak, 1964] Boris T. Polyak. "Some methods of speeding up the convergence of iteration methods". *Ussr computational mathematics and mathematical physics* 4.5 (1964), pp. 1–17 (cit. on pp. 13, 28).

[Radford, 2018] Alec Radford, Karthik Narasimhan, Tim Salimans, Ilya Sutskever, et al. "Improving language understanding by generative pre-training" (2018) (cit. on pp. 6, 22).

[Rahimi, 2007] Ali Rahimi and Benjamin Recht. "Random features for large-scale kernel machines". *Proceedings of the 20th International Conference on Neural Information Processing Systems*. 2007, pp. 1177–1184 (cit. on pp. 5, 21, 34, 35, 82, 93, 95, 102, 118).

[Rahimi, 2008] Ali Rahimi and Benjamin Recht. "Weighted Sums of Random Kitchen Sinks: Replacing minimization with randomization in learning". *Advances in Neural Information Processing Systems* 21 (2008), pp. 1313–1320 (cit. on p. 95).

[Raiko, 2012] Tapani Raiko, Harri Valpola, and Yann LeCun. "Deep learning made easier by linear transformations in perceptrons". *Artificial intelligence and statistics*. PMLR. 2012, pp. 924–932 (cit. on pp. 11, 26, 40).

[Rezende, 2015] Danilo Rezende and Shakir Mohamed. "Variational inference with normalizing flows". *International conference on machine learning*. PMLR. 2015, pp. 1530–1538 (cit. on pp. 12, 26).

[Rockafellar, 1967] Ralph Rockafellar. "Duality and stability in extremum problems involving convex functions". *Pacific Journal of Mathematics* 21.1 (1967), pp. 167–187 (cit. on p. 131).

[Rockafellar, 1968] Ralph Rockafellar. "Integrals which are convex functionals". *Pacific journal of mathematics* 24.3 (1968), pp. 525–539 (cit. on p. 131).

[Rockafellar, 1971] Ralph Rockafellar. "Integrals which are convex functionals. II". *Pacific journal of mathematics* 39.2 (1971), pp. 439–469 (cit. on pp. 131, 133).

[Rosenblatt, 1958] Frank Rosenblatt. "The perceptron: a probabilistic model for information storage and organization in the brain." *Psychological review* 65.6 (1958), p. 386 (cit. on pp. 6, 21).

[Rotskoff, 2019] Grant Rotskoff, Samy Jelassi, Joan Bruna, and Eric Vanden-Eijnden. "Global convergence of neuron birth-death dynamics". *International Conference on Machine Learning*. 2019 (cit. on pp. 9, 15, 24, 30, 36, 120, 121, 126).

[Rotskoff, 2018] Grant Rotskoff and Eric Vanden-Eijnden. "Parameters as interacting particles: long time convergence and asymptotic error scaling of neural networks". *Advances in neural information processing systems* 31 (2018) (cit. on pp. 8, 24, 41).

[Salman, 2018] Hadi Salman, Payman Yadollahpour, Tom Fletcher, and Kayhan Batmanghelich. "Deep Diffeomorphic Normalizing Flows". *arXiv e-prints* (2018), arXiv–1810 (cit. on p. 95).

[Sander, 2021] Michael E. Sander, Pierre Ablin, Mathieu Blondel, and Gabriel Peyré. "Momentum residual neural networks". *International Conference on Machine Learning*. PMLR. 2021, pp. 9276–9287 (cit. on pp. 12, 26, 40).

[Sander, 2022a] Michael E. Sander, Pierre Ablin, Mathieu Blondel, and Gabriel Peyré. "Sinkformers: Transformers with doubly stochastic attention". *International Conference on Artificial Intelligence and Statistics*. PMLR. 2022, pp. 3515–3530 (cit. on p. 168).

[Sander, 2022b] Michael E. Sander, Pierre Ablin, and Gabriel Peyré. "Do residual neural networks discretize neural ordinary differential equations?" *Advances in Neural Information Processing Systems* 35 (2022), pp. 36520–36532 (cit. on pp. 12, 27).

[Santambrogio, 2015] Filippo Santambrogio. *Optimal transport for applied mathematicians*. Vol. 87. Springer, 2015 (cit. on pp. 15, 30, 40, 43, 45, 46, 49, 127, 136, 142, 145).

[Santambrogio, 2017] Filippo Santambrogio. "{Euclidean, metric, and Wasserstein} gradient flows: an overview". *Bulletin of Mathematical Sciences* 7 (2017), pp. 87–154 (cit. on pp. 9, 24, 45, 55, 59, 87, 124, 136).

[Santambrogio, 2023] Filippo Santambrogio. *A Course in the Calculus of Variations: Optimization, Regularity, and Modeling*. Springer Nature, 2023 (cit. on pp. 134, 135).

[Schaback, 1995] Robert Schaback. "Error estimates and condition numbers for radial basis function interpolation". *Advances in Computational Mathematics* 3.3 (1995), pp. 251–264 (cit. on pp. 36, 101, 104, 110).

[Schölkopf, 2002] Bernhard Schölkopf, Alexander J. Smola, Francis Bach, et al. *Learning with kernels: support vector machines, regularization, optimization, and beyond*. 2002 (cit. on pp. 5, 8, 21, 23, 95, 160).

[Sejdinovic, 2013] Dino Sejdinovic, Bharath Sriperumbudur, Arthur Gretton, and Kenji Fukumizu. "Equivalence of distance-based and RKHS-based statistics in hypothesis testing". *The annals of statistics* (2013), pp. 2263–2291 (cit. on p. 126).

[Simon, 1983] Leon Simon. "Asymptotics for a class of non-linear evolution equations, with applications to geometric problems". *Annals of Mathematics* 118.3 (1983), pp. 525–571 (cit. on p. 84).

[Sirignano, 2020] Justin Sirignano and Konstantinos Spiliopoulos. "Mean field analysis of neural networks: A central limit theorem". *Stochastic Processes and their Applications* 130.3 (2020), pp. 1820–1852 (cit. on pp. 8, 24, 120, 121).

[Sjoberg, 1997] Jonas Sjoberg and Mats Viberg. "Separable non-linear least-squares minimization-possible improvements for neural net fitting". *Neural networks for signal processing VII. Proceedings of the 1997 IEEE signal processing society workshop*. IEEE. 1997, pp. 345–354 (cit. on p. 122).

[Sriperumbudur, 2015] Bharath Sriperumbudur and Zoltán Szabó. "Optimal rates for random Fourier features". *Advances in neural information processing systems* 28 (2015) (cit. on p. 103).

[Sriperumbudur, 2011] Bharath K. Sriperumbudur, Kenji Fukumizu, and Gert R. G. Lanckriet. "Universality, Characteristic Kernels and RKHS Embedding of Measures." *Journal of Machine Learning Research* 12.7 (2011) (cit. on pp. 92, 161).

[Srivastava, 2015] Rupesh Kumar Srivastava, Klaus Greff, and Jürgen Schmidhuber. "Highway networks". *arXiv preprint arXiv:1505.00387* (2015) (cit. on pp. 10, 25).

[Steinwart, 2008] Ingo Steinwart and Andreas Christmann. *Support vector machines*. Springer Science & Business Media, 2008 (cit. on pp. 5, 8, 21, 23, 91, 160, 161).

[Steinwart, 2012] Ingo Steinwart and Clint Scovel. "Mercer's theorem on general domains: On the interaction between measures, kernels, and RKHSs". *Constructive Approximation* 35 (2012), pp. 363–417 (cit. on p. 161).

[Sun, 2019] Yitong Sun. "Random Features Methods in Supervised Learning". PhD thesis. 2019 (cit. on pp. 108, 109, 125).

[Sutskever, 2013] Ilya Sutskever, James Martens, George Dahl, and Geoffrey E. Hinton. "On the importance of initialization and momentum in deep learning". *International conference on machine learning*. PMLR. 2013, pp. 1139–1147 (cit. on pp. 13, 28).

[Suzuki, 2023] Taiji Suzuki, Denny Wu, Kazusato Oko, and Atsushi Nitanda. "Feature learning via mean-field langevin dynamics: classifying sparse parities and beyond". *Advances in Neural Information Processing Systems* 36 (2023), pp. 34536–34556 (cit. on p. 126).

[Szegedy, 2017] Christian Szegedy, Sergey Ioffe, Vincent Vanhoucke, and Alexander Alemi. "Inception-v4, inception-resnet and the impact of residual connections on learning". *Proceedings of the AAAI conference on artificial intelligence*. Vol. 31. 1. 2017 (cit. on pp. 11, 26, 40).

[Takakura, 2024] Shokichi Takakura and Taiji Suzuki. "Mean-field analysis on two-layer neural networks from a kernel perspective". *Proceedings of the 41st International Conference on Machine Learning*. 2024, pp. 47475–47509 (cit. on pp. 14, 29, 126).

[Thorpe, 2023] Matthew Thorpe and Yves van Gennip. "Deep limits of residual neural networks". *Research in the Mathematical Sciences* 10.1 (2023), p. 6 (cit. on p. 44).

[Trouvé, 1998] Alain Trouvé. "Diffeomorphisms groups and pattern matching in image analysis". *International journal of computer vision* 28.3 (1998), pp. 213–221 (cit. on p. 95).

[Vaswani, 2017] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, et al. "Attention is all you need". *Advances in neural information processing systems* 30 (2017) (cit. on pp. 6, 7, 10, 21, 22, 25, 31, 41, 80).

[Vázquez, 2006] Juan Luis Vázquez. *Smoothing and decay estimates for nonlinear diffusion equations: equations of porous medium type*. Vol. 33. OUP Oxford, 2006 (cit. on p. 124).

[Vázquez, 2007] Juan Luis Vázquez. *The porous medium equation: mathematical theory*. Oxford university press, 2007 (cit. on p. 124).

[Vialard, 2019] François-Xavier Vialard. "Partial optimization and Schur complement" (2019) (cit. on p. 122).

[Vialard, 2020] François-Xavier Vialard, Roland Kwitt, Susan Wei, and Marc Niethammer. "A shooting formulation of deep learning". *Advances in Neural Information Processing Systems* 33 (2020) (cit. on pp. 12, 26, 40, 95).

[Villalobos, 2022] Pablo Villalobos, Jaime Sevilla, Tamay Besiroglu, Lennart Heim, Anson Ho, and Marius Hobbhahn. "Machine learning model sizes and the parameter gap". *arXiv preprint arXiv:2207.02852* (2022) (cit. on pp. 7, 22).

[Villani, 2009] Cédric Villani. *Optimal transport: old and new*. Vol. 338. Springer, 2009 (cit. on pp. 15, 30, 40, 42, 45, 46, 48, 49, 66, 127).

[Wang, 2024] Guillaume Wang, Alireza Mousavi-Hosseini, and Lénaïc Chizat. "Mean-field langevin dynamics for signed measures via a bilevel approach". *Advances in Neural Information Processing Systems* 37 (2024), pp. 35165–35224 (cit. on pp. 126, 133, 134).

[Wendland, 2004] Holger Wendland. *Scattered data approximation*. Vol. 17. Cambridge university press, 2004 (cit. on p. 110).

[Wojtowytsch, 2020] Stephan Wojtowytsch. "On the convergence of gradient descent training for two-layer relu-networks in the mean field regime". *arXiv preprint arXiv:2005.13530* (2020) (cit. on pp. 40, 41, 80).

[Yang, 2021] Greg Yang and Edward J. Hu. "Tensor programs iv: Feature learning in infinite-width neural networks". *International Conference on Machine Learning*. PMLR. 2021, pp. 11727–11737 (cit. on pp. 7, 8, 13, 22, 23, 28, 120).

[Yang, 2023] Greg Yang, Dingli Yu, Chen Zhu, and Soufiane Hayou. "Tensor Programs VI: Feature Learning in Infinite Depth Neural Networks". *The Twelfth International Conference on Learning Representations*. 2023 (cit. on pp. 27, 169).

[Yarotsky, 2022] Dmitry Yarotsky. "Universal approximations of invariant maps by neural networks". *Constructive Approximation* 55.1 (2022), pp. 407–474 (cit. on p. 25).

[Younes, 2010] Laurent Younes. *Shapes and Diffeomorphisms*. Vol. 171. Applied Mathematical Sciences. Berlin, Heidelberg: Springer Berlin Heidelberg, 2010 (cit. on pp. 95, 97).

[Yun, 2020] Chulhee Yun, Srinadh Bhojanapalli, Ankit Singh Rawat, Sashank Reddi, and Sanjiv Kumar. "Are Transformers universal approximators of sequence-to-sequence functions?" *International Conference on Learning Representations*. 2020 (cit. on p. 25).

[Zhang, 2021] Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. "Understanding deep learning (still) requires rethinking generalization". *Communications of the ACM* 64.3 (2021), pp. 107–115 (cit. on pp. 7, 22, 101).

[Zhang, 2018] Hongyi Zhang, Yann N. Dauphin, and Tengyu Ma. "Fixup Initialization: Residual Learning Without Normalization". *International Conference on Learning Representations*. 2018 (cit. on pp. 95, 109, 111, 114, 116).

[Zou, 2020] Difan Zou, Yuan Cao, Dongruo Zhou, and Quanquan Gu. "Gradient descent optimizes over-parameterized deep ReLU networks". *Machine learning* 109 (2020), pp. 467–492 (cit. on pp. 8, 23, 36, 39, 80, 81, 91, 120).

[Zou, 2019] Difan Zou, Philip M Long, and Quanquan Gu. "On the Global Convergence of Training Deep Linear ResNets". *International Conference on Learning Representations*. 2019 (cit. on pp. 39, 80, 81).

RÉSUMÉ

Les progrès récents des modèles d'apprentissage profond dans de nombreuses applications ont mis en lumière la nécessité d'une meilleure compréhension de leurs dynamiques d'entraînement. Dans cette thèse, nous contribuons à l'étude théorique des algorithmes de descente de gradient pour l'entraînement de réseaux de neurones surparamétrés. Des travaux récents ont en effet montré que, pour des architectures peu profondes, il est possible d'obtenir de bonnes garanties de convergence en relaxant le problème d'optimisation dans l'espace des distributions de paramètres.
Nous prolongeons cette approche au cas des architectures profondes en étudiant des limites de champ-moyen de *réseaux de neurones résiduels (ResNets)*. Ces modèles sont paramétrés par des distributions sur le produit de l'espace des couches et d'un espace de paramètres, avec la contrainte d'une marginale uniforme sur l'espace des couches. Dans ce cadre, nous proposons de modéliser l'apprentissage comme un flot de gradient pour une distance de *Transport Optimal Conditionnel (TOC)*, une variante du transport optimal classique incorporant cette contrainte de marginale. En nous appuyant sur la théorie des flots de gradient dans les espaces métriques, nous démontrons l'existence et la cohérence de ce flot avec l'entraînement des ResNets de largeur finie. Ce travail est également l'occasion d'explorer plus en détail les propriétés du TOC et de sa formulation dynamique.
Nous étudions ensuite le comportement asymptotique des flots de gradient en nous appuyant sur des inégalités de type *Polyak-Łojasiewicz locales*. Nous montrons que ces inégalités sont génériquement satisfaites par les ResNets profonds, et établissons des résultats de convergence pour certains exemples d'architectures et d'initialisations : si le nombre de neurones est fini mais suffisamment grand, et si le risque est suffisamment faible à l'initialisation, alors le flot de gradient converge vers un minimiseur global.
Enfin, afin d'étudier l'émergence de *représentations* non-linéaires durant l'apprentissage, nous considérons le cas de réseaux à une seule couche cachée avec une fonction de perte quadratique. Pour ce problème d'optimisation non convexe et de grande dimension, les résultats existants sont souvent qualitatifs, ou fondés sur une analyse par le *neural tangent kernel*, dans laquelle les représentations des données restent figées. Exploitant le fait qu'il s'agit d'un *problème quadratique non-linéaire séparable*, nous analysons un algorithme de *Variable Projection (VarPro)* ou d'*apprentissage à deux vitesses* qui permet d'éliminer les variables linéaires et de réduire le problème d'apprentissage à l'entraînement des paramètres non-linéaires. Dans un cadre "enseignant-élève", nous montrons que, dans la limite d'une régularisation nulle, la dynamique de la distribution des représentations est décrite par une équation de *weighted ultra-fast diffusion*, permettant ainsi d'établir un taux de convergence linéaire pour l'échantillonnage de la distribution enseignante.
Le code pour reproduire les résultats numériques présentés est en open source.

MOTS CLÉS

Théorie de l'apprentissage, Apprentissage profond, Optimisation, EDOs neuronales, Flots de gradient Wasserstein

ABSTRACT

The recent successes of deep learning models across a wide range of applications have underscored the need for a deeper understanding of their training dynamics. This research is ultimately motivated by the design of more efficient architectures and learning algorithms.
In this PhD work, we contribute to the theoretical understanding of the dynamics of gradient-based methods for the training of neural networks by studying the case of overparameterized models. Indeed, a recent line of work has proven that, for shallow architectures, good convergence guarantees can be obtained by relaxing the training problem in the space of parameter distributions.
We extend this analysis to the case of deep architectures by studying mean-field models of *deep Residual Neural Networks (ResNets)*. These are parameterized by distributions over a product set of layers and parameter space, with a uniform marginal condition on the set of layers. We then propose to model training with a gradient flow w.r.t. the *Conditional Optimal Transport distance*: a restriction of the classical Optimal Transport distance which enforces the marginal condition. Relying on the theory of gradient flows in metric spaces, we show the well-posedness of the gradient flow equation and its consistency with the training of ResNets at finite width. In addition, this is an opportunity to study in more detail the Conditional Optimal Transport distance, particularly its dynamic formulation.
We then study the asymptotic behavior of gradient flow curves by relying on *local Polyak-Łojasiewicz inequalities*. We show such inequalities are generically satisfied by deep ResNets and prove convergence for well-chosen examples of architectures and initializations: if the number of neurons is finite but sufficiently large and the risk is sufficiently small at initialization, then gradient flow converges to a global minimizer of the training risk at a linear rate.
Finally, to study the learning of nonlinear *features* during training with gradient descent we consider the case of shallow single-hidden-layer neural networks with square loss. For this high-dimensional and non-convex optimization problem, most known convergence results are either qualitative or rely on a *neural tangent kernel* analysis where hidden representations of the data are fixed. Using that this problem belongs to the class of separable nonlinear least squares problems, we consider a *Variable Projection (VarPro)* or *two-timescale learning* algorithm, thereby eliminating the linear variables and reducing the learning problem to the training of nonlinear features. In a "teacher-student" scenario, we show that, in the limit where the regularization strength vanishes, the training dynamic on the feature distribution corresponds to a *weighted ultra-fast diffusion equation*. This provides a linear convergence rate for the sampling of the teacher distribution.
The code for reproducing the numerical results presented in this thesis is open-sourced.

KEYWORDS

Machine learning theory, Deep learning, Optimization, Neural ODEs, Wasserstein gradient flows