EPFL

Thèse n° 9912

A Theory of Finite-Width Neural Networks: Generalization, Scaling Laws, and the Loss Landscape

Présentée le 4 juillet 2023

Faculté des sciences de base
Chaire de théorie des champs statistiques
Programme doctoral en informatique et communications

pour l'obtention du grade de Docteur ès Sciences

par

Berfin SIMSEK

Acceptée sur proposition du jury

Prof. M. Jaggi, président du jury
Prof. C. Hongler, Prof. W. Gerstner, directeurs de thèse
Prof. A. Steger, rapporteuse
Prof. A. Saxe, rapporteur
Prof. E. Abbé, rapporteur

École
polytechnique
fédérale
de Lausanne

2023

I'm older than you, and must know better.
— Lewis Carroll, Alice in Wonderland

To Okan, my brother

Acknowledgements

I am lucky to have had the opportunity to pursue my PhD at EPFL in the beautiful city of Lausanne.
I cannot thank enough my supervisors Clément and Wulfram for supporting me throughout my doctoral journey. I thank Wulfram for his enthusiasm for science, his broad perspective, and insatiable curiosity. Witnessing Wulfram in action during scientific talks has been a constant source of inspiration. Additionally, I am thankful for Wulfram's professionalism which has been essential for progressing forward, especially during challenging times.
I thank Clément for welcoming me to his group and providing continuous guidance during my PhD. I have learned from him that great achievements take time, and I am thankful for his encouragement during the exploration phase that emerged at the end of my doctoral studies. Moreover, I thank Clément for sharing amusing stories and anecdotes that always seemed to arrive at the perfect moment.
Both Wulfram and Clément granted me a great amount of freedom, allowing me to flourish both as a researcher and an individual.
I am also thankful to the members of the jury for my thesis defense: Andrew Saxe, Emmanuel Abbé, Angelika Steger, and Martin Jaggi.
I owe a special debt of gratitude to Johanni for his invaluable support, both during moments of triumph and times of struggle. Without his support in my late PhD, I might have failed to pursue what excited me the most. His commitment to scientific rigor is unmatched, making our collaboration an absolute pleasure. Additionally, I extend my thanks to Franck for his support and for sharing beautiful pieces of mathematics with me.
I am grateful to Levent for supervising my internship at Meta and engaging in enlightening discussions spanning a wide range of AI topics. I also express my appreciation to Florent and Lenka for organizing a fantastic summer school in Les Houches and for their ongoing support of my work.
I want to thank the students that I had the pleasure of supervising: Zhengqing, Hugo, Oğuz, Theodor, and Manu.
To all the fellow PhD students and postdocs I have had the privilege of meeting at LCN and CSFT, I extend my sincere thanks. Arthur, you are a talented and skillful problem solver and on top of that your relaxed attitude made collaborating with you a pleasure. Alireza, from the earliest days of my PhD, I have cherished our conversations and valued your opinions as a close friend, thanks also for reading some sections of this thesis. François, your constant support, encouragement, and reminders of my concrete achievements have been invaluable,

and I appreciate your time in reviewing the introduction of this thesis. Amire, your deep focus and your realistic yet daring approach to problem solving made our collaboration truely enjoyable. I warmly express my gratitude to Chiara, Vasia, and Francesco for the many fun moments that we shared in Lausanne and beyond. Additionally, I extend my heartfelt gratitude to Flavio, Valentin, Guillaume, Samuel, Sophia, Georgios, Bernd, Martin, Florian, Christos, Maxime, Evgenii, Leonard, Vassilis, Ariane, Shuqi, Louis, Marie and Elke for sharing their positive energy and friendship.

To my parents, I am deeply thankful for accompanying me through all the ups and downs of the research process and providing unwavering support. I also extend my gratitude to my brother, Okan, whose visit to Switzerland for two weeks made me happy. To my best friend, Beliz, with whom I have shared every significant life event despite the time difference, I am grateful for your presence and the many delightful vacations we have enjoyed together.

I cannot thank enough my partner Isak for all the support and love he has shown during my PhD, without him this thesis would not be the same. I am grateful for every moment we had together, and cannot wait for many more to come.

Lausanne, June 20, 2023

Abstract

Deep learning has achieved remarkable success in various challenging tasks such as generating images from natural language or engaging in lengthy conversations with humans. The success in practice stems from the ability to successfully train massive neural networks on massive datasets. This thesis studies the theoretical foundations of the simplest architecture, that is, a deep (feedforward) neural network, with a particular emphasis on the role of width.

We first focus on a simple model of finite-width neural networks to study generalization, a central inquiry in machine learning and statistical learning theory. Our study delves into the expected generalization error of a Gaussian random features model in terms of the number of features, number of data points, the kernel that it approximates, and the input distribution. Our formulas closely match numerical experiments.

Next, we explore another simplification of finite-width neural networks to study their training dynamics. We assume a linear activation function, resulting in a linear predictor. However, the training dynamics remain non-trivial. In particular, the loss function is non-convex: the orthogonal symmetry gives rise to manifolds of saddle points at various loss levels. Nevertheless, these saddle points exhibit a unique arrangement, wherein the escape direction of a saddle channels the trajectory towards a subsequent saddle. By gluing the local trajectories between saddles, we describe a so-called saddle-to-saddle dynamics that provably kicks in for very small initializations.

To study finite-width neural networks without devising a simple model, we shift our focus to the structure of network parameterization and permutation symmetry among hidden neurons. We identify a neuron-splitting technique that maps a critical point of a network to a manifold of symmetry-induced critical points of a wider network. By considering all possible neuron partitions and their permutations, we establish the precise scaling law for the number of critical manifolds. The scaling laws behave as $e^{c(\alpha)}m^m$ for large m where m is the width of the wider network and α is shrinkage factor, i.e. is the ratio between the number of distinct neurons to m. Notably, the maximum of $c(\alpha)$ is attained at $\alpha^* = \frac{1}{2\log(2)}$, hence it is the shrinkage factor inducing the most numerous symmetry-induced critical manifolds. We then give an application of this scaling law for overparameterized networks.

The key question is: can we give a rule of thumb for how much overparameterization is needed to ensure reliable convergence to a zero-loss solution? Our approach is based on studying the geometry and topology of the zero-loss solutions in overparameterized neural networks. We prove that *all* zero-loss solution manifolds are identical up to neuron splitting, zero neuron addition, and permutation for input distributions with full support. Additionally,

we give the scaling law of the zero-loss manifolds. The ratio between the two scaling laws yields a measure of the landscape complexity which decays with overparameterization. We observe that the complexity decreases rapidly until reaching an overparameterization factor of approximately $2\log(2)$, beyond which the complexity becomes smaller than one. Overall, we find it recommendable to use at least a factor of 2 to 4 of overparameterization to ensure reliable convergence to a zero-loss solution.

While the scaling laws apply to arbitrary settings, a more detailed analysis is needed to study generalization. We shift our focus to the study of neural networks with few neurons for learning from a standard Gaussian input distribution and a unit-orthonormal teacher network with more neurons. We reformulate the weight-space minimization problem as a constrained optimization problem by factoring out symmetries due to the input distribution. As a non-trivial application, we provide a closed-form expression of the optimal solution and its generalization error for the one-neuron network for ReLU activation. Our reformulation applies to networks with arbitrary width and may be the key to finding the generalization error of underparameterized networks.

Key words: Neural Networks, Machine Learning, Deep Learning, Random Feature Models, Generalization, Loss Landscape, Random Matrix Theory, Combinatorics, Constrained Optimization

Résumé

L'apprentissage profond a connu un succès remarquable dans diverses tâches complexes telles que la génération d'images à partir de textes naturels ou l'engagement dans de longues conversations avec des humains. Ce succès en pratique provient de la capacité à entraîner avec des réseaux immenses sur d'énormes ensembles de données. Cette thèse étudie les fondements théoriques de la plus simple des architectures, à savoir les réseaux neuronaux profonds (feedforward), en mettant l'accent particulier sur le rôle de la largeur.

Nous commençons par nous concentrer sur un modèle simple de réseaux neuronaux de largeur finie afin d'étudier la généralisation, qui a été un sujet central dans l'apprentissage automatique et la théorie de l'apprentissage statistique. Notre étude explore l'erreur de généralisation attendue d'un modèle de caractéristiques aléatoires gaussiennes en fonction du nombre de caractéristiques, du nombre de points de données, du noyau qu'il approche et de la distribution d'entrée. Nos formules correspondent de près aux expériences numériques.

Ensuite, nous explorons une autre simplification des réseaux neuronaux de largeur finie pour étudier leur dynamique d'apprentissage. Nous supposons une fonction d'activation linéaire, ce qui conduit à un prédicteur linéaire. Cependant, la dynamique d'apprentissage reste non triviale. En particulier, la fonction de perte n'est pas convexe : la symétrie orthogonale donne naissance à des points selles à différents niveaux de perte. Néanmoins, ces points selles présentent un agencement unique, où la direction d'échappement d'un point selle canalise la trajectoire vers un prochain point selle. En reliant les trajectoires locales entre les points selles, nous décrivons une dynamique dite de « point selle à point selle » dont nous prouvons qu'elle entre en jeu pour de très petites initialisations.

Pour étudier les réseaux neuronaux de largeur finie sans établir de modèle simple, nous déplaçons notre attention vers la structure de la paramétrisation du réseau et la symétrie de permutation parmi les neurones de la couche cachée. Nous identifions une technique de division des neurones qui envoie un point critique d'un réseau vers une multitude de points critiques induits par la symétrie d'un réseau plus large. En considérant toutes les partitions possibles de neurones et leurs permutations, nous établissons une loi d'échelle précise pour le nombre de variétés critiques. Celui-ci se comportent comme $e^{c(\alpha)}m^m$ pour de grandes valeurs de m, où m représente la largeur du réseau le plus étendu et α est le facteur de réduction, c'est-à-dire le rapport entre le nombre de neurones distincts et m. Notamment, le maximum de $c(\alpha)$ est atteint à $\alpha^* = \frac{1}{2\log(2)}$, ce qui en fait le facteur de réduction induisant le plus grand nombre de variétés critiques induites par la symétrie. Nous donnons ensuite une application de cette loi d'échelle pour les réseaux sur-paramétrisés.

Résumé

La question clé est : pouvons-nous établir une règle empirique sur la sur-paramétrisation nécessaire pour une convergence fiable vers une solution sans perte ? Notre approche repose sur l'étude de la géométrie et de la topologie des solutions sans perte dans les réseaux neuronaux sur-paramétrisés. Nous prouvons que *toutes* les solutions sans perte sont identiques à division des neurones, à ajout de neurones nuls et à permutation près, pour les distributions d'entrée avec un support complet. De plus, nous donnons la loi d'échelle des variétés à perte nulle. Le rapport entre les deux lois d'échelle définit une mesure de la complexité du paysage qui décroît avec la sur-paramétrisation. Nous observons que la complexité diminue rapidement jusqu'à atteindre un facteur de sur-paramétrisation d'environ $2\log(2)$, au-delà duquel la complexité devient inférieure à un. Dans l'ensemble, il s'avère recommandable d'utiliser un facteur de sur-paramétrisation d'au moins 2 à 4 pour garantir une convergence fiable vers une solution sans perte.

Alors que les lois d'échelle s'appliquent à des contextes arbitraires, une analyse plus détaillée est nécessaire pour étudier la généralisation. Nous concentrons notre attention sur l'étude des réseaux neuronaux avec quelques neurones pour l'apprentissage à partir d'une distribution d'entrée gaussienne standard et d'un réseau enseignant unitaire orthonormal avec plus de neurones. Nous reformulons le problème de minimisation de l'espace des poids sous la forme d'un problème d'optimisation contraint en prenant en compte les symétries dues à la distribution d'entrée. Comme application non triviale, nous fournissons une expression analytique de la solution optimale et de son erreur de généralisation pour le réseau à un seul neurone avec une activation ReLU. Notre reformulation s'applique aux réseaux de largeur arbitraire et pourrait être une clé pour trouver l'erreur de généralisation des réseaux sous-paramétrisés.

Contents

List of Figures

1 Introduction

Deep learning has achieved unprecedented success in learning from massive amounts of data (LeCun, Bengio, and G. Hinton, 2015; Devlin et al., 2018; Brown et al., 2020; Dosovitskiy et al., 2020; Brown et al., 2020). The success stems from the combination of increasingly cheap computing power, billion-parameter architectures capable of leveraging parallelization and learning from massive datasets. Deep learning today empowers technology we use in daily life such as image recognition in our smartphones and generating text on demand with chatbots. In particular the recent generative deep learning models such as DALL-E and ChatGPT work unexpectedly well in the difficult tasks of generating realistic images from natural language and engaging in lengthy conversations with humans.

The current practice in deep learning is simple: scale up the dataset and scale up the model to enable learning from the massive dataset (J. Kaplan et al., 2020; Bahri et al., 2021). Scaling up works often well in practice. However, it is very costly, and the resulting models are very complex. The complexity of the models makes it difficult to understand their inner workings and to identify the failure modes.

Explaining the success of deep learning through a theoretical framework seems far away at this early stage. The current general questions and approaches that fall within the scope of this thesis can be listed as follows:

- *Tractable models.* Relevant models can help us understand how neural networks are trained and make predictions on unseen samples. It is possible to completely solve the simplest models. In particular, for simple models such as linear regression, we have the closed-form solution of training since the loss is convex. Then the question of interest is generalization: in particular, how does it scale as a function of parameters and training samples? We study this question in the context of a Gaussian random features model in Chapter 2 and give partial answers. If the loss is not convex, then training is the pressing question which needs to be addressed. Can we find the set of optimal solutions? Can we study training regimes depending on the network initialization? We study these questions in the context of deep linear networks in Chapter 3 and give partial answers.

- *Properties of mid-size and large networks.* Some general properties of neural network families can be studied precisely. An example are novel scaling laws of the loss landscape derived from the permutation symmetries of deep neural networks. This is the topic of Chapter 4 and Chapter 5 of this thesis. This approach is strong as it applies broadly, since permutation-symmetry is an inherent property of neural networks. The implications of such global properties on training dynamics and generalization are yet to be discovered.

- *Toy models.* Neural networks on a tiny scale serve as toy models of their larger versions. We partially address the question of finding the closed-form solution of a non-convex problem in the context of neural networks with few neurons in Chapter 6. This approach is motivated by the scaling studies which show that neural networks improve gradually as they grow larger.

Before we go into details of the particular questions addressed in this thesis, we would like to motivate the study of deep learning theory.

1.1 Why a Theory of Neural Networks?

To understand the role of theory in deep learning, it is instructive to make a comparison with the traditional scientific fields. Generally speaking, physics seeks to explain the laws of nature, and biology looks for answers to underpin the mechanisms running alive beings. Theory has been central for the development of science in understanding the nature of things. Analogous to the role of theory in physics and biology, a theory of deep learning might bring important insights, but still has to be developed.

Additionally, theory can play an important role in developing deep learning models in practice and making them efficient. For instance, theory in deep learning can help make important design choices such as data (sub)selection and hyperparameter selection that would relieve the computational burden of grid search in a high-dimensional space. In this sense, it is an exciting time for theory in deep learning due to the potential impact it may have in the close future.

For a theoretician with an eye towards experimentation, deep learning can also be a delightful playground compared to the classic scientific fields. Biological experiments often take very long time, and may be contaminated and flawed, while doing large scale studies is hard due to ethical concerns. Simulations of a physical process, for example, in cosmology and particle physics, take a huge amount of time and compute power. In comparison, in deep learning, it is easy to run a simple experiment on a computer, ideally on a GPU, which is accessible much more broadly. Scaling studies are important to answer the following question: does the simple model trained on a personal computer represent deep learning employed in practice to some extent?

There are many challenging theoretical questions in deep learning requiring a large variety

of tools, some already successfully addressed, and many more begging for an answer. In fact, thanks to phenomena found in the context of deep learning, our understanding of the more classic models have significantly improved. For instance, the classic problem of linear regression demonstrates the so-called double-descent curve which can be captured with a random matrix theory analysis (Hastie et al., 2022). More generally, the random features model of Rahimi and Recht, 2008a offers a family of interesting problems that can be solved via high-dimensional probability or statistical physics tools. This is relevant because a neural network which has frozen weights except for the last layer corresponds to a random features model. Moreover, neural networks in a particular training regime converge to a so-called Neural Tangent Kernel predictor in the infinite-width limit (Jacot, Gabriel, and Hongler, 2018b). This correspondence between neural networks and kernel methods has sparked a renewed interest in the study of kernel methods (Belkin, Ma, and Mandal, 2018; Jacot, Şimşek, et al., 2020d). At the other end of the spectrum, there are low-dimensional problems of neural networks that may require precise, and rather problem-specific analysis using tools from geometry, topology, and dynamical systems.

1.2 Thesis Focus & General Questions

We study the questions related to generalization, loss landscape, and training dynamics of neural networks and their simpler models covering three loosely related topics:

- *Generalization.* We assume that the input data samples $x_i \in \mathbb{R}^d$ are drawn independently from an input distribution $\mathscr{D}$ and that there is a true function f^* generating targets, i.e. $y_i = f^*(x_i)$ (say, without noise). Given N samples and problem parameters such as the number of random features P, or a ridge parameter λ, we are interested in giving an approximation of the generalization error (i.e., the mismatch between the predictor and the true function with respect to the input distribution). It is possible to obtain such formulas for linear regression (Hastie et al., 2022), for kernel ridge regression (KRR), and for random features (RF) regression as the optimization problem is convex. Moreover, for square loss, there is a closed-form expression for the optimal parameter vector and the corresponding predictor. It is then possible to get approximation for the quantities of interest such as the generalization error, using random matrix theory.

- *Loss Landscape.* Let us consider a shallow neural network $f : \mathbb{R}^d \to \mathbb{R}^{d_{\text{out}}}$, that is $f(x) = \sum_{j=1}^{m} a_j \sigma(w_j \cdot x)$. The loss can be written as $\frac{1}{N}\sum_{i=1}^{N} c(f(x_i), f^*(x_i))$ where $c : \mathbb{R}^{d_{\text{out}}} \times \mathbb{R}^{d_{\text{out}}} \to \mathbb{R}$ is a single-sample cost. The loss function L measures the quality of the parameter[1] $\theta = (w_1, a_1) \oplus ... \oplus (w_m, a_m) \in \mathbb{R}^{(d+d_{\text{out}})m}$ on the training data (x_i, y_i) for $i = 1, ..., N$. This is a non-convex optimization problem for $m \geq 2$ since there is always a permutation-symmetric solution. Finding the critical points (i.e. $\nabla L(\theta) = 0$) of the loss landscape is an important question as well as describing whether they are strict

[1] $\oplus$ denotes the concatenation of two vectors.

saddles, non-strict saddles, local minima, or global minima. Due to the symmetries of network parameterization, the loss function of neural networks exhibits the so-called symmetry-induced critical points (Fukumizu and Amari, 2000; Şimşek, Ged, et al., 2021). Another question of interest is the scaling of the number of critical points to understand the difficulty of the non-convex high-dimensional optimization problem as done in complex systems (Auffinger, Arous, and Černy, 2013).

- *Training.* The scaling analysis of the loss landscape is particularly insightful for finite-width neural networks where, in general, we do not have convergence guarantees. There are three main exceptions:

    1. *Infinite-Width Neural Networks.* Although the loss function is non-convex, an equivalent optimization problem can be formulated either in function space or in measure space, where the loss is then convex in its domain (assuming c is convex). These equivalent formulations allow studying training dynamics either for large initialization through the so-called Neural Tangent Kernel (Jacot, Gabriel, and Hongler, 2018b; S. S. Du, Zhai, et al., 2018) or for small initialization through mean-field dynamics (Chizat and Bach, 2018c; Mei, Montanari, and P.-M. Nguyen, 2018a). Both approaches have received substantial success in explaining training dynamics and establishing convergence guarantees in their respective regimes.
    2. *One-Neuron Network.* In this case, there is neither permutation symmetry between the hidden neurons, nor a non-strict saddle on the line of symmetry-induced critical points. Hence the loss function does not violate the so-called Morse property (Mei, Y. Bai, and Montanari, 2018) and convergence to a global minimum is guaranteed (Tian, 2017; Yehudai and Ohad, 2020).
    3. *Deep Linear Networks.* The loss function in this case is non-convex. Nevertheless, the global loss landscape can be characterized exactly (the network matrix of *all* critical points of the loss function is given by Baldi and Hornik, 1989) and it is simple in the sense that there are only symmetry-induced saddles and the global minima. Relatedly, it is possible to analyze training dynamics even for finite-width networks in this case, as well as the effect of depth, and the low-rank bias (Saxe, McClelland, and Ganguli, 2014; Arora, Cohen, W. Hu, et al., 2019).

1.3 State of the Art

Theory of neural networks can be traced back to the study of perceptron learning (Gardner and Derrida, 1989; Seung, Sompolinsky, and Tishby, 1992) and learning with a few neurons in the hidden layer (Rumelhart, G. E. Hinton, and Williams, 1986). Some classic works have studied the problem of learning with a few neurons with the objective of approximating a sum of neurons (Saad and Solla, 1995) whereas some others studied the organization of critical points of the loss landscape (Fukumizu and Amari, 2000). On another front, approximation theory has been developed for neural networks (Cybenko, 1989; Funahashi, 1989; Hornik,

Stinchcombe, and White, 1989) also in relation to kernels (Poggio and Girosi, 1990; Girosi, Jones, and Poggio, 1995), and the properties at initialization have been studied (R. M. Neal, 1996). It is a delicate task to do justice to the great early works that have contributed to the study of neural networks, which is beyond our scope here. In recent years, the theory of neural networks has taken off and reemerged as an exciting field (J. Lee, Bahri, et al., 2017a; Jacot, Gabriel, and Hongler, 2018b; S. S. Du, Zhai, et al., 2018; Chizat and Bach, 2018c; Mei, Montanari, and P.-M. Nguyen, 2018a; G. M. Rotskoff and Vanden-Eijnden, 2018a; Goldt et al., 2019; J. Lee, Xiao, et al., 2019; Arora, S. Du, et al., 2019; Fan and Z. Wang, 2020; Bordelon, Canatar, and Pehlevan, 2020b; Jacot, Şimşek, et al., 2020b; Şimşek, Ged, et al., 2021; Abbé, Boix-Adserà, Brennan, et al., 2021; Abbé, Adserà, and Misiakiewicz, 2022; Veiga et al., 2022; Arous, Gheissari, and Jagannath, 2022).

Despite much exciting progress, the current approaches to studying finite-width behavior are limited. Systematic empirical studies focus on measuring deviations of the network function from the infinite-width limit (Geiger, Jacot, Spigler, Gabriel, Sagun, d'Ascoli, et al., 2020; J. Lee, Schoenholz, et al., 2020; Vyas, Bansal, and Nakkiran, 2022). For finite-width networks, convergence results to a global minimum are only established when the gradient flow is initialized close to a global minimum (Oymak and Soltanolkotabi, 2020; I. M. Safran, Yehudai, and Shamir, 2021) which requires a priori knowledge of the global minimum as opposed to the random initialization near the origin as done in practice. More precise results are obtained under the assumption that the activation function is quadratic (S. Du and J. Lee, 2018; Sarao Mannelli, Vanden-Eijnden, and Zdeborová, 2020). However, the network function is then limited to a polynomial of the input. In general, it is known that studying models with a finite size might depend on specific problem parameters in comparison to their infinite limits. This thesis focuses on the rich and broad question of learning in finite-width neural networks.

1.3.1 Random Features Model

The conventional wisdom suggests that to ensure good generalization performance, one should choose a model class that is complex enough to learn the signal from the training data, yet simple enough to avoid fitting spurious patterns therein (Bishop, 2006). This view has been questioned by recent developments in machine learning. First, C. Zhang et al., 2016 observed that modern neural network models can perfectly fit randomly labeled training data, while still generalizing well. Second, the test error as a function of parameters exhibits a so-called 'double-descent' curve for many models including neural networks, random forests, and random features models (Advani, Saxe, and Sompolinsky, 2020; Spigler et al., 2018; Belkin, Hsu, Ma, et al., 2018; Mei and Montanari, 2019; Belkin, Hsu, and Xu, 2019; Nakkiran et al., 2019). In general, the risk (i.e. test error) is a random variable with two sources of randomness: the usual one due to the sampling of the training set, and the second one due to the randomness of the model itself.

Kernel Ridge Regression. Despite decades of intense mathematical progress, the rigorous

analysis of the generalization of kernel methods remains a very active and challenging area of research. In recent years, many new kernels have been introduced for both regression and classification tasks; notably, a large number of kernels have been discovered in the context of deep learning, in particular through the so-called Scattering Transform (Mallat, 2012), and in close connection with deep neural networks (Cho and L. K. Saul, 2009; Jacot, Gabriel, and Hongler, 2018b), yielding ever-improving performance for various practical tasks (Arora, S. S. Du, et al., 2019; S. S. Du, Zhai, et al., 2019; Z. Li, R. Wang, et al., 2019; Shankar et al., 2020). Currently, theoretical tools to select the relevant kernel for a given task, i.e. to minimize the generalization error, are however lacking. While a number of bounds for the risk of Linear Ridge Regression (LRR) or KRR (Caponnetto and De Vito, 2007; Gerfo et al., 2008; Sridharan, Shalev-Shwartz, and Srebro, 2009; Marteau-Ferey et al., 2019) exist, most focus on the rate of convergence of the risk: these estimates typically involve constant factors which are difficult to control in practice. Recently, a number of more precise estimates have been given (Louart, Liao, and Couillet, 2017; Dobriban and Wager, 2018; Mei and Montanari, 2019; Liu and Dobriban, 2020; Bordelon, Canatar, and Pehlevan, 2020a).

1.3.2 Deep Linear Networks

DLNs have a non-convex loss landscape and the behavior of training dynamics can be subtle. For shallow networks, the convergence of gradient descent is guaranteed by the fact that the saddles are strict and that all minima are global (Baldi and Hornik, 1989; Kawaguchi, 2016; J. D. Lee, Simchowitz, et al., 2016; J. D. Lee, Panageas, et al., 2019a). In contrast, the deep case features non-strict saddles (Kawaguchi, 2016) and no general proof of convergence exists at the moment, though convergence to a global minimum can be guaranteed in some cases (Arora, Cohen, Golowich, et al., 2019; Eftekhari, 2020).

A recent line of work focuses on the implicit bias of DLNs, and consistently reveals some form of incremental learning and implicit sparsity as in Gissin, Shalev-Shwartz, and Daniely, 2020. Diagonal networks are known to learn minimal L_1 solutions (Moroshko et al., 2020; Woodworth et al., 2020). With a specific initialization and the MSE loss, DLNs learn the singular components of the signal one by one (Saxe, McClelland, and Ganguli, 2014; Advani and Saxe, 2017; Saxe, McClelland, and Ganguli, 2019; Gidel, Bach, and Lacoste-Julien, 2019; Arora, Cohen, W. Hu, et al., 2019). Recently, it has been shown that with losses such as the cross-entropy and the exponential loss, the parameters diverge towards infinity, but end up following the direction of the max-margin classifier w.r.t. the L_p-Schatten (quasi-)norm (Gunasekar, J. Lee, et al., 2018; Gunasekar, J. D. Lee, et al., 2018; Soudry et al., 2018; Ji and Telgarsky, 2018; Ji and Telgarsky, 2020; Chizat and Bach, 2020; Lyu and J. Li, 2020; Moroshko et al., 2020; Yun, Krishnan, and Mobahi, 2021).

In parallel, recent works have shown the existence of two regimes in large-width DNNs: a kernel regime (also called NTK or lazy regime) where learning is described by the so-called Neural Tangent Kernel (NTK) guaranteeing linear convergence (Jacot, Gabriel, and Hongler,

2018b; S. S. Du, Zhai, et al., 2018; Chizat and Bach, 2018a; Arora, S. S. Du, et al., 2019; J. Lee, Xiao, et al., 2019; Huang and Yau, 2020) and an active regime where the dynamics is nonlinear (Chizat and Bach, 2018b; G. Rotskoff and Vanden-Eijnden, 2018; Mei, Montanari, and P.-M. Nguyen, 2018b; Mei, Misiakiewicz, and Montanari, 2019; Chizat and Bach, 2020). For DLNs, both regimes can be observed as well, with evidence that while the linear regime exhibits no sparsity, the active regime favors solutions with some kind of sparsity (Woodworth et al., 2020; Moroshko et al., 2020).

1.3.3 The Loss Landscape of Non-Linear Neural Networks

Neural network landscapes are highly non-convex landscapes, where non-optimal critical points may harm gradient-descent by slowing it down (due to saddles) or making it stop at local minima. Earlier works have argued in favor of a proliferation of saddles in high-dimensional neural network landscapes through an analogy with random error functions (Dauphin et al., 2014). One of the earliest attempts to build a theory of *modern* neural networks was made by drawing a connection between neural networks and spherical spin glasses (Choromanska et al., 2015). However, later numerical work showed that this analogy could lead to incorrect conclusions by revealing fundamental phenomenological differences (Baity-Jesi et al., 2018). The Kac-Rice formula from probability theory is commonly used for giving an estimate for the landscape complexity in high dimensions such as spherical spin glasses and spiked matrix tensor models (Auffinger, Arous, and Černy, 2013; Arous, Mei, et al., 2019). However, this methodology cannot be directly applied to neural networks beyond the one-neuron case (Maillard, Arous, and Biroli, 2020) because of the presence of degenerate critical points emerging due to symmetries in network parameterization (Fukumizu and Amari, 2000).

1.3.4 Overparameterized Neural Networks

Neural network landscapes in practice are found to exhibit surprising properties, such as the connectivity of global minima (Draxler et al., 2018; Garipov et al., 2018) and the convergence to a global minimum in the so-called overparameterized regime (Jacot, Gabriel, and Hongler, 2018b), thereby ruling out proliferating saddles as a problem in this regime. Yet, in mildly overparameterized networks, gradient descent may find a global minimum only for a small fraction of random initializations (Sagun, Guney, et al., 2014; Chizat and Bach, 2018c; Frankle and Carbin, 2018). Neural networks that have more parameters than needed to interpolate the dataset are shown to reach a zero-loss solution more easily (Neyshabur, Bhojanapalli, et al., 2017; Neyshabur, Z. Li, et al., 2018; Geiger, Jacot, Spigler, Gabriel, Sagun, d'Ascoli, et al., 2020).

1.3.5 Neural Networks with Few Neurons

Our theoretical understanding of the neural network training can be improved by studying different, tractable limits, like infinitely-wide networks (Jacot, Gabriel, and Hongler, 2018b;

Chizat and Bach, 2018c; S. S. Du, Zhai, et al., 2018; Mei, Montanari, and P.-M. Nguyen, 2018a; G. M. Rotskoff and Vanden-Eijnden, 2018b; Arora, S. Du, et al., 2019; Sirignano and Spiliopoulos, 2020). However, a theoretical understanding of under-parametrized neural networks is still lacking. We are interested in the following fundamental algorithm-free question: *Can we characterize the optimal solution of a neural network with few neurons?*

In this chapter, we focus on the opposite end: the one-neuron network. There is a large body of work on the one-neuron case with notable examples of Tian, 2017; Mei, Y. Bai, and Montanari, 2018; Yehudai and Ohad, 2020. Our paper develops it further by giving a characterization (closed-form formula) of the optimal solution for unit-orthonormal teacher networks and for standard Gaussian input. Our approach is likely to be generalizable to two and more neurons as we neither assume that the loss function has the Morse property as in Mei, Y. Bai, and Montanari, 2018 (requires that all critical points are isolated; two-neuron network breaks it) nor track the L^2-distance between student and teacher parameters as in Tian, 2017; Yehudai and Ohad, 2020 (fails due to permutation symmetry).

1.4 Main Thesis Contributions

In the first part of this thesis, we will present results from tractable and simplified models of finite-width neural networks.

In Chapter 2, we study the Gaussian random features model. This is a simplification of neural networks where every layer except for the last layer is frozen at initialization hence training happens only in the last layer. Note that when the number of hidden neurons goes to infinity and when the weights are initialized with large values, the neural networks in fact operate in this regime. The closed-form expression of the resulting training dynamics is then captured by the so-called Neural Tangent Kernel (Jacot, Gabriel, and Hongler, 2018b). Importantly, we do not argue for any direct correspondence between the random features model and neural networks beyond this limit. Indeed, Ba et al., 2022 showed that even one step of gradient descent moves the first hidden layer parameters by a non-negligible amount from their random initialization when the numbers of neurons and data points approach infinity at a constant rate.

In Chapter 3, we study deep linear networks, i.e. neural networks with linear activation function. While it is true that the network function expressed by a deep linear network is simply linear, the training dynamics is non-trivial (Saxe, McClelland, and Ganguli, 2014; Arora, Cohen, Golowich, et al., 2019). Moreover, the saddles of linear networks are arranged in a special way: the typical escape direction of a rank ℓ saddle (that is, the network matrix has rank ℓ) falls within the stable manifold of a next saddle of rank $\ell+1$ and visits its neighborhood. In particular, we study the very small initialization regime, so that the parameter vector falls in the proximity of the known saddle at the origin. In this regime, the training dynamics traverses from one saddle to the next following the so-called saddle-to-saddle training regime. Numerically, for finite-width networks that are initialized in the NTK scaling, i.e. parameters

have variance $m^{-\gamma}$ with $\gamma < 1$ (Jacot, Ged, Şimşek, et al., 2021), the trajectories are drawn to the first few saddles which is manifested as learning plateaus in the loss curves.

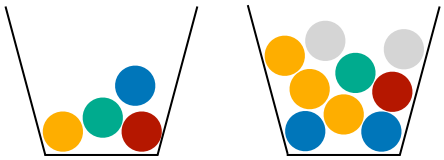
In the second part of this thesis, we study finite-width neural networks (neural networks of fixed width, and with a non-linear activation function).

In Chapter 4, we focus on permutation symmetry and the so-called symmetry-induced saddles of the loss landscape. Splitting one of the neurons into two with a mixing ratio $\mu \in \mathbb{R}$

$$(w_j, a_j) \to (w_j, \mu a_j) \oplus (w_j, (1-\mu) a_j) \tag{1.1}$$

does not only preserve the network function but also preserves criticality (Fukumizu and Amari, 2000). By splitting multiple neurons into many neurons, a critical point θ of a network induces symmetry-induced critical points in any wider network (Şimşek, Ged, et al., 2021). The symmetry-induced critical points form manifolds (i.e. affine subspaces in the base case without the scaling symmetry of ReLU) as the mixing ratios are arbitrary. Hence, they are relevant objects to count to measure the complexity of the loss landscape. Our main result of the chapter is a scaling law of critical manifolds of finite-width neural networks derived from counting all partitions and permutation due to neuron splittings from an initial set of neurons. Due to permutation symmetry, the scaling law in neural networks behaves as a factorial in the number of neurons m but the parameter space also grows linearly in m. This is faster than the exponential growth of the number of critical points of other high-dimensional complex loss landscapes (Auffinger, Arous, and Černy, 2013; Ros et al., 2019; Arous, Mei, et al., 2019).

In Chapter 5, we give an analysis of overparameterization independent of the optimization algorithm, from the landscape complexity point of view. On the one end, this applies to overparameterized networks of any width; on the other end, it does not give convergence guarantees. We assume that the true function is a sum of k neurons, that is $f^*(x) = \sum_{\ell=1}^{k} b_\ell \sigma(v_\ell \cdot x)$. Said differently, we assume that the dataset can be 'solved' by a finite-width network, i.e. $f^*(x_i) = y_i$. The target function is also called a teacher network (Saad and Solla, 1995) or a multi-index model (Mousavi-Hosseini et al., 2022). We focus on overparameterized neural networks with width $m \geq k$ (including zero-overparameterization) and give the scaling law of the zero-loss manifolds that comes from neuron splitting and zero-neuron addition (see Figure 1.1). Importantly, the scaling law is *exact* for the global minima manifold of the population loss: we show that all zero-loss solutions are identical up to neuron splitting, zero neuron addition, and permutation. We then compare the scaling law of the symmetry-induced saddles with the scaling law of the zero-loss manifolds. The resulting measure of landscape complexity gradually decreases and drops to



(a) initial neurons *(b)* equivalent neurons

Figure 1.1 – The two symmetry operations generate an equivalent set of neurons in a wider network: neuron splitting 1.1 (orange and blue neurons) and zero-neuron addition 1.5 (gray neurons cancel out each other).

zero for infinitely wide networks.

In Chapter 6, we study the problem of learning with a few neurons. We assume that the true function is a unit-orthonormal teacher network with k neurons and the input data is standard Gaussian. We study the optimal loss given $n < k$ neurons, assuming we have access to the input distribution. The classic problem of learning with a few neurons is challenging, and only the one-neuron case is studied in detail in the literature (Tian, 2017; Yehudai and Ohad, 2020; Mei, Montanari, and P.-M. Nguyen, 2018a). We also study the one-neuron case in detail and give a closed-form expression for the optimal solution for ReLU activation function in this thesis. Moreover, for odd activation functions such as erf, the extension to the whole underparameterized regime and also $n = k$ has been developed in Şimşek, Bendjeddou, et al., 2023. Our approach includes a reformulation of the problem in the weight space in terms of angular variables and study of the problem in terms of the so-called interaction functions for which in general we do not have analytical formula. Our formulation is applicable to neural networks with arbitarily many neurons and may be the key to characterizing the optimal solutions of underparameterized neural networks.

This thesis includes a variety of methods to study finite-width neural networks to understand their generalization and the loss landscape. Note that not all of the original papers are written with the objective of studying finite-width neural networks, nevertheless, the results are closely linked to the study of finite-width neural networks. In this thesis, we will reinterpret the results accordingly.

1.4.1 Gaussian Random Features Model

Implicit Regularization of Random Features

We consider the Random Feature (RF) model (Rahimi and Recht, 2008b) with features sampled from a Gaussian Process (GP) and study the RF predictor $\hat{f}$ minimizing the regularized least squares error, isolating the randomness of the model by considering fixed training data points. RF models have been the subject of intense research activity: they are (randomized) approximations of Kernel Methods aimed at easing the computational challenges of Kernel Methods while being asymptotically equivalent to them (Rahimi and Recht, 2008b; T. Yang et al., 2012; Sriperumbudur and Szabó, 2015; Yu et al., 2016). Unlike the asymptotic behavior, which is well studied, RF models with a finite number of features are much less understood.

Random Feature (RF) models are used as efficient parametric approximations of kernel methods. We investigate, by means of random matrix theory, the connection between Gaussian RF models and Kernel Ridge Regression (KRR). For a Gaussian RF model with P features, N data points, and a ridge λ, we show that the average (i.e. expected) RF predictor is close to a KRR predictor with an *effective ridge* $\tilde{\lambda}$. We show that $\tilde{\lambda} > \lambda$ and $\tilde{\lambda} \searrow \lambda$ monotonically as P grows, thus revealing the *implicit regularization effect* of finite RF sampling. We then compare the risk (i.e. test error) of the $\tilde{\lambda}$-KRR predictor with the average risk of the λ-RF predictor and obtain a

precise and explicit bound on their difference. Finally, we empirically find an extremely good agreement between the test errors of the average λ-RF predictor and $\tilde{\lambda}$-KRR predictor.

Generalization of KRR

KRR is a widely-used statistical method to learn a function from its values on a training set (Schölkopf, Smola, and Müller, 1998a; Shawe-Taylor and Cristianini, 2004). It is a non-parametric generalization of linear regression to infinite-dimensional feature spaces. Given a positive-definite kernel function K and (noisy) observations y^ϵ of a true function f^* at a list of points $X = \{x_1, \dots, x_N\}$, the λ-KRR estimator $\hat{f}_\lambda^\epsilon$ of f^* is defined by

$$\hat{f}_\lambda^\epsilon(x) = \frac{1}{N} K(x, X) \left(\frac{1}{N} K(X, X) + \lambda I_N \right)^{-1} y^\epsilon,$$

where $K(x, X) = (K(x, x_i))_{i=1,..,N} \in \mathbb{R}^N$ and $K(X, X) = (K(x_i, x_j))_{i,j=1,..,N} \in \mathbb{R}^{N \times N}$.

We study the generalization error of KRR for a kernel K with ridge $\lambda > 0$ and i.i.d. observations. For this, we introduce a so-called Signal Capture Threshold (SCT), which is a function of the data distribution: it can be used to identify the components of the data that the KRR predictor captures, and to approximate the (expected) KRR risk.

1.4.2 Deep Linear Networks

The dynamics of Deep Linear Networks (DLNs) is dramatically affected by the variance σ^2 of the parameters at initialization θ_0. For DLNs of width m, we show a transition w.r.t. the scaling γ of the variance $\sigma^2 = m^{-\gamma}$ as $m \to \infty$: for large variance ($\gamma < 1$), θ_0 is very close to a global minimum but far from any saddle point, and for small variance ($\gamma > 1$), θ_0 is close to a saddle point and far from any global minimum. While the first case corresponds to the well-studied NTK regime, the second case is less understood. This motivates the study of the case $\gamma \to +\infty$, where we conjecture a so-called saddle-to-saddle dynamics: throughout training, gradient flow visits the neighborhoods of a sequence of saddles, each corresponding to linear maps of increasing rank, until reaching a sparse global minimum (Z. Li, Y. Luo, and Lyu, 2020; Jacot, Ged, Şimşek, et al., 2021). We support this conjecture with a theorem for the dynamics between the first two saddles, as well as some numerical experiments. Saddle-to-saddle dynamics are also observed and studied in (non-linear) neural networks (Boursier, Pillaud-Vivien, and Flammarion, 2022; Abbé, Boix-Adserà, and Misiakiewicz, 2023).

1.4.3 The Loss Landscape of (Non-Linear) Neural Networks

It is crucial to understand the loss landscapes of neural networks to study various training regimes (Jacot, Gabriel, and Hongler, 2018b; Chizat and Bach, 2018c; Jacot, Ged, Şimşek, et al., 2021), to find the optimal solutions with closed-form expressions, and to characterize the

possible failure modes (I. Safran and Shamir, 2018).

We give a detailed second-order analysis of the so-called symmetry-induced critical points originating from the optimal solution of a narrower neural network building upon Fukumizu and Amari, 2000; Şimşek, Ged, et al., 2021. In particular, we specify the conditions under which the splitting of neurons leads to local minima, strict saddles, or non-strict saddles. We analyze the second-order derivatives of a line of critical points in the network with $n+1$ neurons induced by neuron splitting of a minimum of the network with n neurons. The line always contains two non-strict saddles, a continuum of strict saddles, and potentially also local minima on line segments forming a *plateau saddle* since there is an escape direction via a non-strict saddle at its boundary. Overall, the loss functions of neural networks with more than one neuron are qualitatively different from the complex loss landscapes with isolated critical points and non-convex loss functions satisfying the strict saddle property.

We also derive a new scaling law for the number of critical manifolds for finite-width neural networks (Şimşek, Ged, et al., 2021). The number of splittings of n neurons onto m neurons including the permutation symmetry between the latter is given by the expansion factor

$$G(n,m) = \sum_{i=1}^{n} \binom{n}{i} (-1)^{n-i} i^m = \left\{ {m \atop n} \right\} n! \tag{1.2}$$

where the curly brackets denote Stirling numbers of the second kind. Numerically, we observe that

$$\lim_{m\to\infty} \frac{1}{m \log m} G(\alpha m, m) \to c(\alpha) \tag{1.3}$$

for $\alpha \in [0,1]$ where $c(\alpha)$ is a unimodal curve with peak at $\frac{1}{2log2}$ (clearly, the exact constant is not possible to determine numerically; it is taken from the exciting mathoverflow post). Importantly, studying the scaling law and the hiererchical organization of the saddles gives a lens to see the stucture of the loss landscape of finite-width neural networks.

1.4.4 Overparameterized Networks

We propose a notion of landscape complexity that measures the competition between the scaling law of the saddle manifolds and that of the global minima. In particular, assuming a teacher network[2]generates the targets, we call any network with a larger width *overparameterized*. For networks with a few neurons more than the teacher, so-called mild overparameterization, we proved that the landscape complexity approaches to infinity in the case of complex targets (i.e. infinite teacher width). With further overparameterization, it decreases and drops to zero for infinite-width networks. Our average-case analysis of finite-width networks provides fresh insights into understanding overparameterization that is impossible to obtain with worst-case

analysis.

Our approach to study landscape complexity of neural networks is based on permutation symmetry (Brea, Şimşek, et al., 2019). Consider a two-layered neural network function $f : \mathbb{R}^d \to \mathbb{R}^{d_{\text{out}}}$ with n neurons

$$f(x) = \sum_{j=1}^{n} a_j \sigma(w_j \cdot x) \tag{1.4}$$

where w_j and a_j are incoming and outgoing vectors to the hidden neurons respectively, and σ is the activation function. For the critical points at zero loss, i.e. global minima, we need to include the zero-neuron addition symmetry

$$(w', 0), \;\; (w', a) \oplus (w', -a), \;\; \ldots \tag{1.5}$$

which preserves the network function, but breaks criticality when applied to a critical point that has greater than zero loss. Under the assumption that a finite-width network, say with k neurons, achieves a zero-loss solution, the number of zero-loss manifolds in an overparameterized network with $m > k$ neurons is given by the scaling law denoted by $T(k, m)$ (Şimşek, Ged, et al., 2021). We can view these scaling laws G and T as a generalization of the usual factorial: in the case $m = n = k$, there is neither room to split neurons nor to add zero neurons, hence both expansion factors reduce to $k!$.

According to our numerics, the critical factor $2\log(2)$ is relevant for the trainability of neural networks for difficult teachers (Martinelli et al., 2023). An exciting future direction is a characterization of difficult teachers/datasets, as it could potentially be used in practice for choosing the optimal network width for reliable convergence to a zero-loss solution. In the case of deep networks, the landscape complexity grows exponentially with the number of hidden layers since permutation symmetry applies to every one of them. We therefore expect a sharper crossover at the critical factor of overparameterization.

Ensembling

In practice, the best performance is typically obtained by an ensemble of the same deep network where the variability comes from the randomness in initialization and the ordering of data samples. In our idealized setting of the population loss limit, if the training algorithm converges to zero loss, ensembling randomly initialized networks would not improve any accuracy since we proved that the network function at convergence is unique. Indeed in practical settings, the solutions reached by stochastic gradient descent can be mapped to the same linear region of the landscape up to a permutation of hidden neurons as shown by (S. P. Singh and Jaggi, 2020; Entezari et al., 2021; Ainsworth, Hayase, and Srinivasa, 2022; K. Jordan et al., 2022; Benzing et al., 2022) through large-scale experiments. These recent results

[2] This is a natural assumption as any target function can be approximated arbitrarily well with a neural network thanks to the universal approximation theorem.

point to an exciting research direction towards bridging the gap between weight averaging and ensembling with substantial implications in distributed training and federated learning which might be studied theoretically by generalizing our symmetry analysis to finite training dataset scenarios.

1.4.5 Neural Networks with Few Neurons

In this chapter, we consider a student network of n neurons that learns from data generated by a teacher network with $k > n$ neurons. There is little work focused on the study of under-parameterized networks, i.e. $n < k$, with the exception of a recent empirical study by Elhage et al., 2022. In particular, we prove the closed-form formula of the optimal solution of the one-neuron network for ReLU and erf activations when learning from a unit-orthonormal teacher network with multiple neurons, going beyond the realizable case of the one-neuron teacher (i.e. single-index model). More generally, our work offers a novel approach to studying the classic teacher-student model as a concrete step toward understanding finite-width neural networks exhibiting rich and intriguing phenomena.

Our approach relies on reparameterizing the loss in terms of the so-called interactions that can be expressed as a function of the standard deviations and correlation of two Gaussian random variables. The interactions in general do not have an explicit formula except for the activation functions such as erf, ReLU, and linear (Saad and Solla, 1995; Goldt et al., 2019). We show that the fixed point equation corresponding to the zero-derivative constraints at a critical point should satisfy some bounds on the norm of the incoming vector and outgoing weight for softplus activation for which the interaction does not admit an analytical formula: the incoming vector computes a damped average of the teacher incoming vectors, and the outgoing weight compensates for the missing neurons.

Part I Tractable Models

2 Gaussian Random Features Model

In this chapter, we study the generalization of the random features predictor learning from a finite data set of size N, finite number of features P, and with ridge λ. We present the main results in Section 2.1 and related works in Section 2.2. We first introduce the λ-RF predictor considering a fixed dataset in Section 2.3 following our paper Jacot, Şimşek, et al., 2020b. In Subsection 2.4.1, preliminary results on the distribution of the λ-RF model are provided. In Subsection 2.4.2, the first main theorem is stated (Theorem 2.4.3): the average (expected) λ-RF predictor is close to the $\tilde{\lambda}$-KRR predictor for an explicit $\tilde{\lambda} > \lambda$. As a consequence (Corollary 2.4.5), the test errors of these two predictors are close; and numerical experiments show that the test errors are in fact virtually identical (Figure 2.2). In Subsection 2.4.3, the second main theorem is stated (Theorem 2.4.6): a bound on the variance of the λ-RF predictor is given, which shows that it concentrates around its average. The ridgeless $\lambda \searrow 0$ case is then investigated: a lower bound on the variance of the λ-RF predictor is given, suggesting an explanation for the double-descent curve in the ridgeless case.

To study the risk of the expected RF predictor, we need to study the generalization error of the Kernel Ridge Regression (KRR) predictor with the effective ridge. In Section 2.5, we introduce the KRR predictor, then introduce the relevant operators (Section 2.5) to study its train error and risk. The rest of the chapter is then devoted to obtaining approximations for the KRR risk. In Section 2.6.1, the Signal Capture Threshold (SCT) is introduced and used to study the mean and variance of the KRR predictor in Subsection 2.6.1. In Section 2.6.3, the expected risk is approximated in terms of the SCT and its derivative w.r.t. the ridge λ. This second part is based on our paper Jacot, Şimşek, et al., 2020c.

2.1 Main Results

We consider a model of Random Features (RF) approximating a kernel method with kernel K. This model consists of P Gaussian features, sampled i.i.d. from a (centered) Gaussian process with covariance kernel K. For a given training set of size N, we study the distribution of the RF predictor $\hat{f}^{(RF)}$ with ridge parameter $\lambda > 0$ (L^2 penalty on the parameters) and denote it by

λ-RF. We show the following in Jacot, Şimşek, et al., 2020a:

- The distribution of $\hat{f}^{(RF)}$ is that of a mixture of Gaussian processes.
- The expected RF predictor is close to the $\tilde{\lambda}$-KRR (Kernel Ridge Regression) predictor for an effective ridge parameter $\tilde{\lambda} > \lambda$.
- The effective ridge $\tilde{\lambda}$ is determined by the number of features P, the ridge λ and the Gram matrix of K on the dataset; $\tilde{\lambda}$ decreases monotonically to λ as P grows, revealing the implicit regularization effect of finite RF sampling. Conversely, when using random features to approximate a kernel method with a specific ridge λ^*, one should choose a smaller ridge $\lambda < \lambda^*$ to ensure $\tilde{\lambda}(\lambda) = \lambda^*$.
- The test errors of the expected λ-RF predictor and of the $\tilde{\lambda}$-KRR predictor are numerically found to be extremely close, even for small P and N.
- The RF predictor's concentration around its expectation can be explicitly controlled in terms of P and of the data; this yields in particular $\mathbb{E}[R(\hat{f}_\lambda^{(RF)})] = R(\hat{f}_{\tilde{\lambda}}^{(K)}) + \mathcal{O}(P^{-1})$ as $N, P \to \infty$ with a fixed ratio $\gamma = P/N$ where R is the MSE risk.

Since we compare the behavior of λ-RF and $\tilde{\lambda}$-KRR predictors on the same fixed training set, our result does not rely on any probabilistic assumption on the training data (in particular, we do not assume that our training data is sampled i.i.d.) in Jacot, Şimşek, et al., 2020a.

To study the generelization of the expected RF predictor, we need to study the generalization error of the Kernel Ridge Regression (KRR) predictor with the effective ridge parameter. We consider the KRR predictor $\hat{f}^{(K)}$: one tries to reconstruct a true function f^* from noisy observations $y^\epsilon = \left(f^*(x_1) + \epsilon e_1, ..., f^*(x_N) + \epsilon e_N\right)$, where the observations x_i are data points sampled from a distribution $\mathcal{D}$, ϵ is the level of noise, and the $e_1, ..., e_N$ are centered of unit variance. We work under the universality assumption that, for large N, only the first two moments of $\phi(x)$ and $f^*(x)$ determine the behavior of the first two moments of $\hat{f}^{(K)}$ where $\phi : \mathbb{R}^d \to \mathcal{H}$ is the feature map to the corresponding RKHS $\mathcal{H}$, i.e. $K(x, x') = \langle \phi(x), \phi(x') \rangle_{\mathcal{H}}$. We obtain the following results in Jacot, Şimşek, et al., 2020d:

- We introduce the Signal Capture Threshold (SCT) ϑ, which is determined by the ridge λ, the size of the training set N, the kernel K, and the data distribution $\mathcal{D}$ (more precisely, the dependence on $\mathcal{D}$ is only through its first two moments). We give approximations for the expectation and variance of the KRR predictor in terms of the SCT.
- Decomposing f^* along the kernel principal components of the data distribution, we observe that in expectation, the predictor $\hat{f}^{(K)}$ captures only the signal along the principal components with eigenvalues larger than the SCT. If N increases or λ decreases, the SCT ϑ shrinks, allowing the predictor to capture more signal. At the same time, the variance of $\hat{f}^{(K)}$ scales with the derivative $\partial_\lambda \vartheta$, which grows as $\lambda \to 0$, supporting the classical bias-variance tradeoff picture (Geman, Bienenstock, and Doursat, 1992a).

- We give an explicit approximation for the expected MSE risk $R^{\epsilon}(\hat{f}^{(K)})$ and expected empirical MSE risk $\hat{R}^{\epsilon}(\hat{f}^{(K)})$. We find that, surprisingly, the expected risk and expected empirical risk are approximately related by

$$\mathbb{E}[R^{\epsilon}(\hat{f}_{\lambda}^{(K)})] \approx \frac{\vartheta^2}{\lambda^2}\mathbb{E}[\hat{R}^{\epsilon}(\hat{f}_{\lambda}^{(K)})].$$

Our proofs rely on a generalized and refined version of the finite-size analysis of Jacot, Şimşek, et al., 2020a of generalized Wishart matrices, obtaining sharper bounds and generalizing the results to operators. Our analysis relies in particular on the complex Stieltjes transform $m_G(z)$, evaluated at $z = -\lambda$, and on fixed-point arguments.

2.2 Related works

Generalization of Random Features. The generalization behavior of Random Feature models has seen intense study in the Statistical Learning Theory framework. Rahimi and Recht, 2009 find that $\mathcal{O}(N)$ features are sufficient to ensure the $\mathcal{O}(1/\sqrt{N})$ decay of the generalization error of Kernel Ridge Regression (KRR). Rudi and Rosasco, 2017 improve on their result and show that $\mathcal{O}(\sqrt{N}\log N)$ features is actually enough to obtain the $\mathcal{O}(1/\sqrt{N})$ decay of the KRR error.

Hastie et al., 2022 use random matrix theory tools to compute the asymptotic risk when both $P, N \to \infty$ with $P/N \to \gamma > 0$. When the training data is sampled i.i.d. from a Gaussian distribution, the variance is shown to explode at $\gamma = 1$. In the same linear regression setup, Bartlett et al. (2019) establish general upper and lower bounds on the excess risk. Mei and Montanari (2019) prove that the double-descent (DD) curve also arises for random ReLU features, and adding a ridge suppresses the explosion around $\gamma = 1$.

Double-descent and the effect of regularization. For the cross-entropy loss, Neyshabur, Tomioka, and Srebro (2014) observed that for two-layer neural networks the test error exhibits the double-descent (DD) curve as the network width increases (without regularizers, without early stopping). For MSE and hinge losses, the DD curve was observed also in multilayer networks on the MNIST dataset (Spigler et al., 2018; Advani, Saxe, and Sompolinsky, 2020). B. Neal et al. (2018) study the variance due to stochastic training in neural networks and find that it increases until a certain width, but then decreases down to 0. Nakkiran et al. (2019) establish the DD phenomenon across various models including convolutional and recurrent networks on more complex datasets (e.g. CIFAR-10, CIFAR-100).

Belkin, Hsu, Ma, et al., 2018; Belkin, Hsu, and Xu, 2019 find that the DD curve is not peculiar to neural networks and observe the same for random Fourier features and decision trees. In Geiger, Jacot, Spigler, Gabriel, Sagun, d'Ascoli, et al., 2019, the DD curve for neural networks is related to the variance associated with the random initialization of the Neural Tangent Kernel (Jacot, Gabriel, and Hongler, 2018a); as a result, ensembling is shown to suppress the DD phenomenon in this case, and the test error stays constant in the overparameterized regime.

Recent theoretical work (d'Ascoli et al., 2020) study the same setting and derive formulas for the asymptotic error, relying on the so-called replica method.

Generalization of Kernel Ridge Regression. The theoretical analysis of the risk of KRR has seen tremendous developments in the recent years. In particular, a number of upper and lower bounds for kernel risk have been obtained in various settings (Caponnetto and De Vito, 2007; Sridharan, Shalev-Shwartz, and Srebro, 2009; Marteau-Ferey et al., 2019): notably, convergence rates (i.e. without control of the constant factors) are obtained in general settings. This allows one to abstract away a number of details about the kernels (e.g. the lengthscale), which don't influence the asymptotic rates. However, this does not give access to the risk at finite data size (crucial to pick e.g. the correct lengthscale or the NTK depth).

We introduce the Signal Capture Threshgold (SCT) to study the risk achieved when learning from finite data that is related to a number of objects from previous works, such as the effective dimension of T. Zhang, 2003; Caponnetto and De Vito, 2007, the companion Stieltjes transform of Dobriban and Wager, 2018; Liu and Dobriban, 2020, and particularly the effective ridge of Jacot, Şimşek, et al., 2020a. The SCT can actually be viewed as a direct translation to the KRR risk setting of Jacot, Şimşek, et al., 2020a.

General Wishart Matrices. A number of recent results have given precise descriptions of the risk for ridge regression (Dobriban and Wager, 2018; Liu and Dobriban, 2020), for random features (Mei and Montanari, 2019), and in relation to neural networks (Louart, Liao, and Couillet, 2017; Bordelon, Canatar, and Pehlevan, 2020a). These results rely on the analysis of the asymptotic spectrum of general Wishart random matrices, in particular through the Stieltjes transform Silverstein, 1995; Z. Bai and Z. Wang, 2008. The limiting Stieltjes transform can be recovered from the formula for the product of freely independent matrices (Gabriel, 2015; Speicher, 2017). To extend these asymptotic results to finite-size settings, we generalize and adapt the results of (Jacot, Şimşek, et al., 2020a).

While these techniques have given simple formulae for the KRR predictor expectation, approximating its variance has remained more challenging. For this reason the description of the expected risk in Louart, Liao, and Couillet, 2017 is stated as a conjecture. In Liu and Dobriban, 2020 only the bias component of the risk is approximated. In Dobriban and Wager, 2018 the expected risk is given only for random true functions (in a Bayesian setting) with a specific covariance. In Bordelon, Canatar, and Pehlevan, 2020a, the expected risk follows from a heuristic spectral analysis combining a PDE approximation and replica tricks. In this paper, we approximate the variance of the predictor along the principal components, giving an approximation of the risk for arbitrary true functions.

Our analysis relies on the study of the spectrum of the general Wishart matrices of the form $W\Sigma W^T$ (for a fixed square matrix Σ and a rectangular matrix W with i.i.d. standard Gaussian entries) and in particular their Stieltjes transform $m_P(z) = \frac{1}{P}\mathrm{Tr}\left(W\Sigma W^T - zI_P\right)^{-1}$. In this paper, we provide non-asymptotic variants of these results for an arbitrary matrix Σ (which in our setting is the kernel Gram matrix or the kernel integral operator); the proofs in our setting are

detailed in Appendix A.2.

2.3 Setup

Linear regression is a parametric model consisting of linear combinations

$$f_\theta = \frac{1}{\sqrt{P}}\left(\theta_1\phi^{(1)} + \cdots + \theta_P\phi^{(P)}\right)$$

of (deterministic) features $\phi^{(1)},\ldots,\phi^{(P)} : \mathbb{R}^d \to \mathbb{R}$. We consider an arbitrary training dataset (X, y) with $X = [x_1, ..., x_N] \in \mathbb{R}^{d\times N}$ and $y = [y_1, \ldots, y_N] \in \mathbb{R}^N$, where the labels could be noisy observations. For a ridge parameter $\lambda > 0$, the linear estimator corresponds to the parameters $\hat{\theta} = [\hat{\theta}_1, \ldots, \hat{\theta}_P] \in \mathbb{R}^P$ that minimize the (regularized) Mean Square Error (MSE) functional $\hat{R}_\lambda$ defined by

$$\hat{R}_\lambda(f_\theta) = \frac{1}{N}\sum_{i=1}^{N}\left(f_\theta(x_i) - y_i\right)^2 + \frac{\lambda}{N}\|\theta\|^2. \tag{2.1}$$

The *data matrix* F is defined as the $N \times P$ matrix with entries $F_{ij} = \frac{1}{\sqrt{P}}\phi^{(j)}(x_i)$. The minimization of (2.1) can be rewritten in terms of F as

$$\hat{\theta} = \operatorname{argmin}_\theta \|F\theta - y\|^2 + \lambda\|\theta\|^2. \tag{2.2}$$

The optimal solution $\hat{\theta}$ is then given by

$$\hat{\theta} = F^T\left(FF^T + \lambda I_N\right)^{-1} y \tag{2.3}$$

and the optimal predictor $\hat{f} = f_{\hat{\theta}}$ by

$$\hat{f}(x) = \frac{1}{\sqrt{P}}\sum_{j=1}^{P}\phi^{(j)}(x)F^T_{:,j}\left(FF^T + \lambda I_N\right)^{-1} y. \tag{2.4}$$

In this paper, we consider linear models of Gaussian random features associated with a kernel $K : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$. We take $\phi^{(j)} = f^{(j)}$, where $f^{(1)},\ldots,f^{(P)}$ are sampled i.i.d. from a Gaussian Process of zero mean (i.e. $\mathbb{E}[f^{(j)}(x)] = 0$ for all $x \in \mathbb{R}^d$) and with covariance K (i.e. $\mathbb{E}[f^{(j)}(x)f^{(j)}(x')] = K(x, x')$ for all $x, x' \in \mathbb{R}^d$). In our setup, the optimal parameter $\hat{\theta}$ still satisfies (2.3) where F is now a random matrix. The associated predictor, called λ-RF predictor, is then given by

Definition 2.3.1 (Random Feature Predictor)**.** *Consider a kernel $K : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$, a ridge $\lambda > 0$, and random features $f^{(1)},\ldots,f^{(P)}$ sampled i.i.d. from a centered Gaussian Process of covariance K. Let $\hat{\theta}$ be the optimal solution to (2.1) taking $\phi^{(j)} = f^{(j)}$. The Random Feature predictor with ridge λ is the random function $\hat{f}^{(RF)}_\lambda : \mathbb{R}^d \to \mathbb{R}$ defined by*

$$\hat{f}^{(RF)}_\lambda(x) = \frac{1}{\sqrt{P}}\sum_{j=1}^{P}\hat{\theta}_j f^{(j)}(x). \tag{2.5}$$

The λ-RF can be viewed as an approximation of kernel ridge predictors: observing from (2.4) that $\hat{f}_\lambda^{(RF)}$ only depends on the scalar product $K_P(x,x') = \frac{1}{P}\sum_{j=1}^P f^{(j)}(x)f^{(j)}(x')$ between datapoints, we see that as $P \to \infty$, $K_P \to K$ and hence $\hat{f}_\lambda^{(RF)}$ converges (Rahimi and Recht, 2008b) to a kernel predictor with ridge λ (Schölkopf, Smola, and Müller, 1998b), which we call λ-KRR predictor.

Definition 2.3.2 (Kernel Predictor)**.** *Consider a kernel function $K : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ and a ridge $\lambda > 0$. The Kernel Predictor is the function $\hat{f}_\lambda^{(K)} : \mathbb{R}^d \to \mathbb{R}$*

$$\hat{f}_\lambda^{(K)}(x) = K(x,X)(K(X,X) + \lambda \mathrm{I}_N)^{-1} y \tag{2.6}$$

where $K(X,X)$ is the $N \times N$ matrix of entries $K(X,X)_{ij} = K(x_i,x_j)$ and $K(\,\cdot\,,X) : \mathbb{R}^d \to \mathbb{R}^N$ is the map $K(x,X)_i = K(x,x_i)$.

Bias-Variance Decomposition of the Risk

Let us assume that there exists a true regression function $f^* : \mathbb{R}^d \to \mathbb{R}$ and an input data generating distribution $\mathscr{D}$ on $\mathbb{R}^d$. The risk of a predictor $f : \mathbb{R}^d \to \mathbb{R}$ is measured by the MSE defined as (in the noiseless setting)

$$R(f) = \mathbb{E}_{\mathscr{D}}\left[(f(x) - f^*(x))^2\right].$$

Let π denote the joint distribution of the i.i.d. sample $f^{(1)}, ..., f^{(P)}$ from the centered Gaussian process with covariance kernel K. The risk of $\hat{f}^{(RF)}$ can be decomposed into a bias-variance form as

$$\mathbb{E}_\pi\left[R(\hat{f}^{(RF)})\right] = R\left(\mathbb{E}_\pi[\hat{f}^{(RF)}]\right) + \mathbb{E}_{\mathscr{D}}\left[\mathrm{Var}_\pi(\hat{f}^{(RF)}(x))\right].$$

This decomposition into the risk of the *average* RF predictor and of the $\mathscr{D}$-expectation of its variance will play a crucial role in the next sections. This is in contrast with the classical bias-variance decomposition in Geman, Bienenstock, and Doursat (1992b)

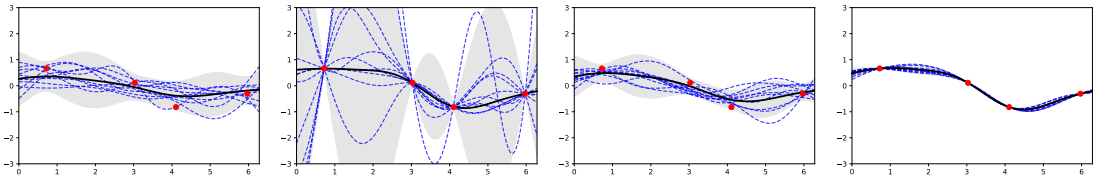
$$\mathbb{E}_{\mathscr{D}^{\otimes N}}[R(f)] = R(\mathbb{E}_{\mathscr{D}^{\otimes N}}[f]) + \mathbb{E}_{\mathscr{D}}[\mathrm{Var}_{\mathscr{D}^{\otimes N}}[f(x)]]$$

where $\mathscr{D}^{\otimes N}$ denotes the joint distribution on $x_1, ..., x_N$, sampled i.i.d. from $\mathscr{D}$. Note that in our decomposition no probabilistic assumption is made on the data, which is fixed.

Additional Notation

In this paper, we consider a fixed dataset (X, y) with distinct data points and a kernel K (i.e. a positive definite symmetric function $\mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$).

Let UDU^T be the spectral decomposition of the kernel matrix $K(X,X)$, with $D = \mathrm{diag}(d_1, \ldots, d_N)$. Let $D^{\frac{1}{2}} = \mathrm{diag}(\sqrt{d_1}, \ldots, \sqrt{d_N})$ and set $K^{\frac{1}{2}} = UD^{\frac{1}{2}}U^T$. The law of the (random) data matrix F is now that of $\frac{1}{\sqrt{P}}K^{\frac{1}{2}}W^T$ where W is a $P \times N$ matrix of i.i.d. standard Gaussian entries, so that



(a) $P = 2, \lambda = 10^{-4}$ (b) $P = 4, \lambda = 10^{-4}$ (c) $P = 4, \lambda = 0.1$ (d) $P = 100, \lambda = 10^{-4}$

Figure 2.1 – *Distribution of the RF Predictor.* Red dots represent a sinusoidal dataset $y_i = \sin(x_i)$ for $N = 4$ points x_i in $[0, 2\pi)$. For selected P and λ, we sample ten RF predictors (blue dashed lines) and compute empirically the average RF predictor (black lines) with ± 2 standard deviations intervals (shaded regions).

$\mathbb{E}[FF^T] = K(X, X)$.

We will denote by $\gamma = P/N$ the parameter-to-datapoint ratio: the *underparameterized regime* corresponds to $\gamma < 1$, while the *overparameterized regime* corresponds to $\gamma \geq 1$. In order to stress the dependence on the ratio parameter γ, we write $\hat{f}^{(RF)}_{\lambda,\gamma}$ instead of $\hat{f}^{(RF)}_{\lambda}$.

2.4 Implicit Regularization of Random Features

2.4.1 First Observations

The distribution of the RF predictor features a variety of behaviors depending on γ and λ, as displayed in fig. 2.1. In the underparameterized regime $P < N$, sample RF predictors induce some *implicit regularization* and do not interpolate the dataset (2.1a); at the interpolation threshold $P = N$, RF predictors interpolate the dataset but the variance explodes when there is no ridge (2.1b), however adding some ridge suppresses variance explosion (2.1c); in the overparameterized regime $P \geq N$ with large P, the variance vanishes thus the RF predictor converges to its average (2.1d). We will investigate the average RF predictor (solid lines) in detail in Section 2.4.2 and study its variance in Section 2.4.3.

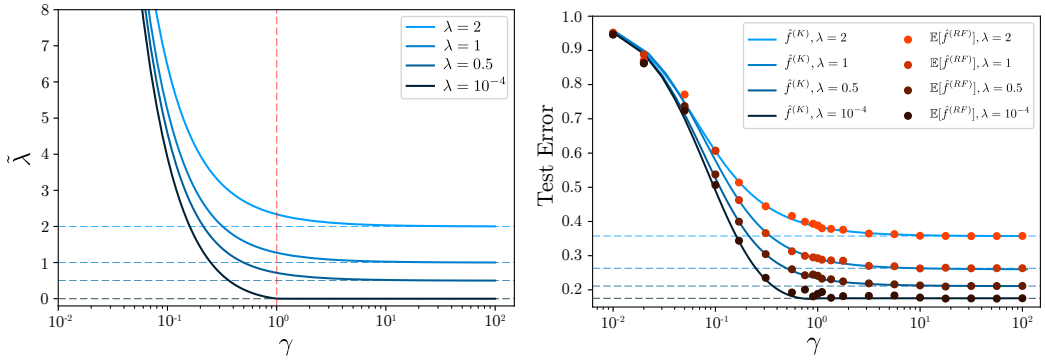
We start by characterizing the distribution of the RF predictor as a Gaussian mixture:

Proposition 2.4.1. *Let $\hat{f}^{(RF)}_{\lambda,\gamma}(x)$ be the random features predictor as in (2.5) and let $\hat{y} = F\hat{\theta}$ be the prediction vector on training data, i.e. $\hat{y}_i = \hat{f}^{(RF)}_{\lambda,\gamma}(x_i)$. The process $\hat{f}^{(RF)}_{\lambda,\gamma}$ is a mixture of Gaussians: conditioned on F, we have that $\hat{f}^{(RF)}_{\lambda,\gamma}$ is a Gaussian process. The mean and covariance of $\hat{f}^{(RF)}_{\lambda,\gamma}$ conditioned on F are given by*

$$\mathbb{E}[\hat{f}^{(RF)}_{\lambda,\gamma}(x)|F] = K(x, X)K(X, X)^{-1}\hat{y}, \tag{2.7}$$

$$\mathrm{Cov}[\hat{f}^{(RF)}_{\lambda,\gamma}(x), \hat{f}^{(RF)}_{\lambda,\gamma}(x')|F] = \frac{\|\hat{\theta}\|^2}{P}\tilde{K}(x, x'), \tag{2.8}$$

with $\tilde{K}(x, x') = K(x, x') - K(x, X)K(X, X)^{-1}K(X, x')$ denoting the posterior covariance kernel.



(a) Evolution of $\tilde{\lambda}$

(b) Average λ-RF predictor vs. $\tilde{\lambda}$-KRR

Figure 2.2 – *Comparison of the test errors of the average λ-RF predictor and the $\tilde{\lambda}$-KRR predictor.* We train the RF predictors on $N = 100$ MNIST data points where K is the RBF kernel, i.e. $K(x,x') = \exp\left(-\|x-x'\|^2/\ell\right)$. We approximate the average λ-RF on 100 random test points for various ridges λ. In *(a)*, given γ and λ, the effective ridge $\tilde{\lambda}$ is computed numerically using (2.10). In *(b)*, the test errors of the $\tilde{\lambda}$-KRR predictor (blue lines) and the empirical average of the λ-RF predictor (red dots) agree perfectly.

The proof of Proposition 2.4.1 relies on the fact that $f^{(j)}$ conditioned on $\left(f^{(j)}(x_i)\right)_{i=1,\dots,N}$ is a Gaussian Process. Note that (2.7) and (2.8) depend on λ and P through $\hat{y}$ and $\|\hat{\theta}\|^2$; in fact, as the proof shows, these identities extend to the ridgeless case $\lambda \searrow 0$. For the ridgeless case, when one is in the overparameterized regime ($P \geq N$), one can (with probability one) fit the labels y and hence $\hat{y} = y$:

Corollary 2.4.2. *When $P \geq N$, the average ridgeless RF predictor is equivalent to the ridgeless KRR predictor*

$$\mathbb{E}\left[\hat{f}^{(RF)}_{\lambda\searrow 0,\gamma}(x)\right] = K(x,X)K(X,X)^{-1}y = \hat{f}^{(K)}_{\lambda\searrow 0}(x).$$

This corollary shows that in the overparameterized case, the ridgeless RF predictor is an unbiased estimator of the ridgeless kernel predictor. The difference between the expected loss of ridgeless RF predictor and that of the ridgeless KRR predictor is hence equal to the variance of the RF predictor. As will be demonstrated in this article, outside of this specific regime, a systematic bias appears, which reveals an implicit regularizing effect of random features.

2.4.2 Average Predictor

In this section, we study the average RF predictor $\mathbb{E}[\hat{f}^{(RF)}_{\lambda,\gamma}]$. As shown by Corollary 2.4.2 above, in the ridgeless overparmeterized regime, the RF predictor is an unbiased estimator of the ridgeless kernel predictor. However, in the presence of a non-zero ridge, we see the following *implicit regularization effect*: the average λ-RF predictor is close to the $\tilde{\lambda}$-KRR predictor for an effective ridge $\tilde{\lambda} > \lambda$ (in other words, sampling a finite number P of features amounts to taking a greater kernel ridge $\tilde{\lambda}$).

Theorem 2.4.3. *For $N, P > 0$ and $\lambda > 0$, we have*

$$\left|\mathbb{E}[\hat{f}^{(RF)}_{\lambda,\gamma}(x)] - \hat{f}^{(K)}_{\lambda}(x)\right| \leq \frac{c\sqrt{K(x,x)}\, y^T K(X,X)^{-1} y}{P} \tag{2.9}$$

where the effective ridge $\tilde{\lambda}(\lambda,\gamma) > \lambda$ is the unique positive number satisfying

$$\tilde{\lambda} = \lambda + \frac{\tilde{\lambda}}{\gamma}\frac{1}{N}\sum_{i=1}^{N}\frac{d_i}{\tilde{\lambda}+d_i}, \tag{2.10}$$

and where $c > 0$ depends on λ, γ, and $\frac{1}{N}\mathrm{Tr}K(X,X)$ only.

Proof. (Sketch; see Supp. Mat. for details) Set $A_\lambda = F(F^TF + \lambda I_P)^{-1}F^T$. The vector of the predictions on the training set is given by $\hat{y} = A_\lambda y$ and the expected predictor is given by

$$\mathbb{E}\left[\hat{f}^{(RF)}_{\lambda,\gamma}(x)\right] = K(x,X)K(X,X)^{-1}\mathbb{E}[A_\lambda]\, y.$$

By a change of basis, we may assume the kernel Gram matrix to be diagonal, i.e. $K(X,X) = \mathrm{diag}(d_1,\ldots,d_N)$. In this basis $\mathbb{E}[A_\lambda]$ turns out to be diagonal too. For each $i = 1,\ldots,N$ we can isolate the contribution of the i-th row of F: by the Sherman-Morrison formula, we have $(A_\lambda)_{ii} = \frac{d_i g_i}{1+d_i g_i}$, where

$$g_i = \frac{1}{P} W_i^T (F_{(i)}^T F_{(i)} + \lambda \mathrm{I}_P)^{-1} W_i,$$

with W_i denoting the i-th column of $W = \sqrt{P}F^TK^{-\frac{1}{2}}$ and $F_{(i)}$ being obtained by removing the i-th row of F. The g_i's are all within $\mathcal{O}(1/\sqrt{P})$ distance to the Stieltjes transform

$$m_P(-\lambda) = \frac{1}{P}\mathrm{Tr}\left(F^TF + \lambda \mathrm{I}_P\right)^{-1}.$$

By a fixed point argument, the Stieltjes transform $m_P(-\lambda)$ is itself within $\mathcal{O}(1/\sqrt{P})$ distance to the deterministic value $\tilde{m}(-\lambda)$, where it is is the unique positive solution to

$$\gamma = \frac{1}{N}\sum_{i=1}^{N}\frac{d_i\tilde{m}(-\lambda)}{1+d_i\tilde{m}(-\lambda)} + \gamma\lambda\tilde{m}(-\lambda).$$

(The detailed proof in the Supp. Mat. uses non-asymptotic variants of arguments found in (Z. Bai and Z. Wang, 2008); the constants in the $\mathcal{O}$ bounds are in particular made explicit).

As a consequence, from the above results, we obtain

$$\mathbb{E}[(A_\lambda)_{ii}] = \mathbb{E}\left[\frac{d_i g_i}{1+d_i g_i}\right] \approx \frac{d_i\tilde{m}}{1+d_i\tilde{m}} = \frac{d_i}{\tilde{\lambda}+d_i},$$

revealing the effective ridge $\tilde{\lambda} = 1/\tilde{m}(-\lambda)$. This implies that $\mathbb{E}[A_\lambda] \approx K(X,X)(K(X,X)+\tilde{\lambda}\mathrm{I}_N)^{-1}$

and

$$\mathbb{E}\left[\hat{f}^{(RF)}_{\lambda,\gamma}(x)\right] \approx K(x,X)(K(X,X)+\tilde{\lambda}\mathrm{I}_N)^{-1}y = \hat{f}^{(K)}_{\lambda}(x),$$

yielding the desired result. □

Note that asymptotic forms of equations similar to the ones in the above proof appear in different settings (Dobriban and Wager, 2018; Mei and Montanari, 2019; Liu and Dobriban, 2020), related to the study of the Stieltjes transform of the product of asymptotically free random matrices.

While the above theorem does not make assumptions on P, N, and K, the case of interest is when the right hand side $\frac{c}{P}\sqrt{K(x,x)}y^T K(X,X)^{-1}y$ is small. The constant $c>0$ is uniformly bounded whenever γ and λ are bounded away from 0 and $\frac{1}{N}\mathrm{Tr}K(X,X)$ is bounded from above. As a result, to bound the right hand side of (2.9), the two quantities we need to bound are $T=\frac{1}{N}\mathrm{Tr}K(X,X)$ and $y^T K(X,X)^{-1}y$.

- The boundedness of T is guaranteed for kernels that are translation-invariant, i.e. of the form $K(x,y)=k(\|x-y\|)$: in this case, one has $T=k(0)$.
- If we assume $\mathbb{E}_{\mathscr{D}}[K(x,x)]<\infty$ as is commonly done in the literature (Rudi and Rosasco, 2017), T converges to $\mathbb{E}_{\mathscr{D}}[K(x,x)]$ as $N\to\infty$ (assuming i.i.d. data points).
- For $y^T K(X,X)^{-1}y$, under the assumption that the labels are of the form $y_i=f^*(x_i)$ for a true regression function f^* lying in Reproducing Kernel Hilbert Space (RKHS) $\mathscr{H}$ of the kernel K (Schölkopf, Smola, and Müller, 1998b), we have $y^T K(X,X)^{-1}y \le \|f^*\|_{\mathscr{H}}$.

Our numerical experiments in Figure (2.2b) show excellent agreement between the test error of the expected λ-RF predictor and the one of the $\tilde{\lambda}$-KRR predictor suggesting that the two functions are indeed very close, even for small N, P.

Thanks to the implicit definition of the effective ridge $\tilde{\lambda}$ (which depends on λ,γ,N and on the eigenvalues d_i of $K(X,X)$) we obtain the following:

Proposition 2.4.4. *The effective ridge $\tilde{\lambda}$ satisfies the following properties:*

1. *for any $\gamma>0$, we have $\lambda<\tilde{\lambda}(\lambda,\gamma)\le\lambda+\frac{1}{\gamma}T$;*
2. *the function $\gamma\mapsto\tilde{\lambda}(\lambda,\gamma)$ is decreasing;*
3. *for $\gamma>1$, we have $\tilde{\lambda}\le\frac{\gamma}{\gamma-1}\lambda$;*
4. *for $\gamma<1$, we have $\tilde{\lambda}\ge\frac{1-\sqrt{\gamma}}{\sqrt{\gamma}}\min_i d_i$.*

The above proposition shows the implicit regularization effect of the RF model: sampling fewer features (i.e. decreasing γ) increases the effective ridge $\tilde{\lambda}$.

Furthermore, as $\lambda \to 0$ (ridgeless case), the effective ridge $\tilde{\lambda}$ behave as follows:

- in the overparameterized regime ($\gamma > 1$), $\tilde{\lambda}$ goes to 0;
- in the underparameterized regime ($\gamma < 1$), $\tilde{\lambda}$ goes to a limit $\tilde{\lambda}_0 > 0$.

These observations match the profile of $\tilde{\lambda}$ in Figure (2.2a).

Remark. When $\lambda \searrow 0$, the constant c in our bound (2.9) explodes (see Supp. Mat.). As a result, this bound is not directly useful when $\lambda = 0$. However, we know from Corollary 2.4.2 that in the ridgeless overparametrized case ($\gamma > 1$), the average RF predictor is equal to the ridgeless KRR predictor. In the underparametrized case ($\gamma < 1$), our numerical experiments suggest that the ridgeless RF predictor is an excellent approximation of the $\tilde{\lambda}_0$-KRR predictor.

Effective Dimension

The effective ridge $\tilde{\lambda}$ is closely related to the so-called effective dimension appearing in statistical learning theory. For a linear (or kernel) model with ridge λ, the *effective dimension* $\mathcal{N}(\lambda) \leq N$ is defined as $\sum_{i=1}^{N} \frac{d_i}{\lambda + d_i}$ (T. Zhang, 2003; Caponnetto and De Vito, 2007). It allows one to measure the effective complexity of the Hilbert space in the presence of a ridge.

For a given $\lambda > 0$, the effective ridge $\tilde{\lambda}$ introduced in Theorem 2.4.3 is related to the effective dimension $\mathcal{N}(\tilde{\lambda})$ by

$$\mathcal{N}(\tilde{\lambda}) = P\left(1 - \frac{\lambda}{\tilde{\lambda}}\right).$$

In particular, we have that $\mathcal{N}(\tilde{\lambda}) \leq \min(N, P)$: this shows that the choice of a finite number of features corresponds to an automatic lowering of the effective dimension of the related kernel method.

Note that in the ridgeless underparameterized case ($\lambda \searrow 0$ and $\gamma < 1$), the effective dimension $\mathcal{N}(\tilde{\lambda})$ equals precisely the number of features P.

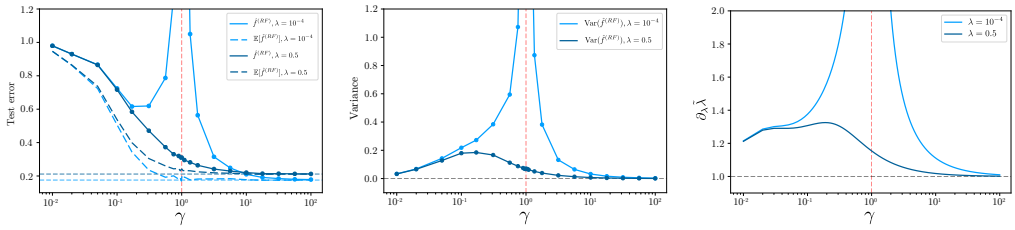
Risk of the Average Predictor

A corollary of Theorem 2.4.3 is that the loss of the expected RF predictor is close to the loss of the KRR predictor with ridge $\tilde{\lambda}$:

Corollary 2.4.5. *If $\mathbb{E}_{\mathscr{D}}[K(x,x)] < \infty$, we have that the difference of errors $\delta_E = \left| L(\mathbb{E}[\hat{f}^{(RF)}_{\lambda,\gamma}]) - L(\hat{f}^{(K)}_{\tilde{\lambda}}) \right|$ is bounded from above by*

$$\delta_E \leq \frac{C y^T K(X,X)^{-1} y}{P} \left(2\sqrt{L\left(\hat{f}^{(K)}_{\tilde{\lambda}}\right)} + \frac{C y^T K(X,X)^{-1} y}{P} \right),$$

where C is given by $c\sqrt{\mathbb{E}_{\mathscr{D}}[K(x,x)]}$, with c the constant appearing in (2.9) above.



(a) Ridgeless vs. Ridge (b) Variance of λ-RF (c) Evolution of $\partial_\lambda \tilde{\lambda}$

Figure 2.3 – *Average test error of the ridgeless vs. ridge λ-RF predictors.* In *(a)*, the average test errors of the ridgeless and the ridge RF predictors (solid lines) and the effect of ensembling (dashed lines) for $N = 100$ MNIST data points. In *(b)*, the variance of the RF predictors and in *(c)*, the evolution of $\partial_\lambda \tilde{\lambda}$ in the ridgeless and ridge cases. The experimental setup is the same as in Figure 2.2.

As a result, δ_E can be bounded in terms of $\lambda, \gamma, T, y^T K(X,X)^{-1} y$, which are discussed above, and of the kernel generalization error $L(f_{\tilde{\lambda}}^{(K)})$. Such a generalization error can be controlled in a number of settings as N grows (Caponnetto and De Vito, 2007; Marteau-Ferey et al., 2019). For instance, the loss is shown to vanish as $N \to \infty$. Figure (2.2b) shows that the two test losses are indeed very close.

2.4.3 Bounding the Variance of the Predictor

In the previous sections, we analyzed the loss of the expected predictor $\mathbb{E}[\hat{f}_{\lambda,\gamma}^{(RF)}]$. In order to analyze the expected loss of the RF predictor $\hat{f}_{\lambda,\gamma}^{(RF)}$, it remains to control the variance of the RF predictor: this follows from the bias-variance decomposition

$$\mathbb{E}\left[L(\hat{f}_{\lambda,\gamma}^{(RF)})\right] = L\left(\mathbb{E}[\hat{f}_{\lambda,\gamma}^{(RF)}]\right) + \mathbb{E}_{\mathscr{D}}\left[\mathrm{Var}(\hat{f}_{\lambda,\gamma}^{(RF)}(x))\right],$$

introduced in Section 2.3. The variance $\mathrm{Var}\left(\hat{f}_{\lambda,\gamma}^{(RF)}(x)\right)$ of the RF predictor can itself be written as the sum

$$\mathrm{Var}\left(\mathbb{E}\left[\hat{f}_{\lambda,\gamma}^{(RF)}(x) \mid F\right]\right) + \mathbb{E}\left[\mathrm{Var}\left(\hat{f}_{\lambda,\gamma}^{(RF)}(x) \mid F\right)\right].$$

By Proposition 2.4.1, we have

$$\mathbb{E}\left[\hat{f}_{\lambda,\gamma}^{(RF)}(x) \mid F\right] = K(x,X)K(X,X)^{-1}\hat{y}, \quad \mathrm{Var}\left(\hat{f}_{\lambda,\gamma}^{(RF)}(x) \mid F\right) = \frac{\|\hat{\theta}\|^2}{P}\tilde{K}(x,x).$$

RF Predictor Concentration

The following theorem allows us to bound both terms:

Theorem 2.4.6. *There are constants $c_1, c_2 > 0$ depending on λ, γ, T only such that*

$$\mathrm{Var}\left(K(x,X)K(X,X)^{-1}\hat{y}\right) \leq \frac{c_1 K(x,x)(y^T K(X,X)^{-1}y)^2}{P},$$
$$\left|\mathbb{E}[\|\hat{\theta}\|^2] - \partial_\lambda \tilde{\lambda}\, y^T M_{\tilde{\lambda}} y\right| \leq \frac{c_2 (y^T K(X,X)^{-1}y)^2}{P},$$

where $\partial_\lambda \tilde{\lambda}$ is the derivative of $\tilde{\lambda}$ with respect to λ and for $M_{\tilde{\lambda}} = K(X,X)(K(X,X)+\tilde{\lambda} I_N)^{-2}$. As a result

$$\mathrm{Var}\left(\hat{f}^{(RF)}_{\lambda,\gamma}(x)\right) \leq \frac{c_3 K(x,x)(y^T K(X,X)^{-1}y)^2}{P},$$

where $c_3 > 0$ depends on λ, γ, T.

Putting the pieces together, we obtain the following bound on the difference $\Delta_E = |\mathbb{E}[L(\hat{f}^{(RF)}_{\lambda,\gamma})] - L(\hat{f}^{(K)}_{\lambda})|$ between the expected RF loss and the KRR loss:

Corollary 2.4.7. *If $\mathbb{E}_D[K(x,x)] < \infty$, we have*

$$\Delta_E \leq \frac{C_1 y^T K(X,X)^{-1} y}{P}\left(\sqrt{L(\hat{f}^{(K)}_{\lambda})} + C_2 y^T K(X,X)^{-1} y\right).$$

where C_1 and C_2 depend on λ, γ, T and $\mathbb{E}_{\mathscr{D}}[K(x,x)]$ only.

Double Descent Curve

We now investigate the neighborhood of the frontier $\gamma = 1$ between the under- and overparameterized regimes, known empirically to exhibit a double descent curve, where the test error explodes at $\gamma = 1$ (i.e. when $P \approx N$) as exhibited in Figure 2.3.

Thanks to Theorem 2.4.6, we get a lower bound on the variance of $\hat{f}^{(RF)}_{\lambda,\gamma}$:

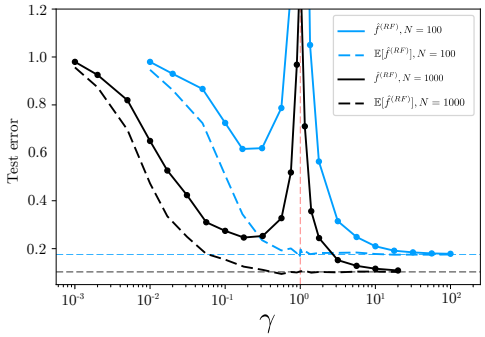
Corollary 2.4.8. *There exists $c_2 > 0$ depending only on λ, γ, T (same as in Theorem 2.4.6) such that* $\mathrm{Var}(\hat{f}^{(RF)}_{\lambda,\gamma}(x))$ *is bounded from below by*

$$\partial_\lambda \tilde{\lambda} \frac{y^T M_{\tilde{\lambda}} y}{P} \tilde{K}(x,x) - \frac{c_2 (y^T K(X,X)^{-1}y)^2}{P^2} \tilde{K}(x,x).$$

If we assume the second term of Corollary A.6.2 to be negligible, then the only term which depends on P is the first term. The derivative $\partial_\lambda \tilde{\lambda}$ has an interesting behavior as a function of λ and γ:

Proposition 2.4.9. *For $\gamma > 1$, as $\lambda \to 0$, the derivative $\partial_\lambda \tilde{\lambda}$ converges to $\frac{\gamma}{\gamma-1}$. As $\lambda\gamma \to \infty$, we have $\partial_\lambda \tilde{\lambda}(\lambda,\gamma) \to 1$.*

The explosion of $\partial_\lambda \tilde{\lambda}$ in $(\gamma = 1, \lambda = 0)$ is displayed in Figure (2.3c). Corollary A.6.2 can be used to explain the double-descent curve numerically observed for small $\lambda > 0$. It is natural to



(a) $N = 100$ vs. $N = 1000$

Figure 2.4 – *Average test error of the λ-RF predictor for two values of N and $\lambda = 10^{-4}$.* For $N = 1000$, the test error is naturally lower and the cusp at $\gamma = 1$ is narrower than for $N = 100$. The experimental setup is the same as in Figure 2.2.

assume that in this case $\partial_\lambda \tilde{\lambda} \gg 1$ around $\gamma = 1$, dominating the lower bound in Corollary A.6.2. In turn, by Proposition A.4.2 this implies that the variance of $\hat{f}^{(RF)}$ gets large. Finally, by the bias-variance decomposition, we obtain a sharp increase of the test error around $\gamma = 1$, which is in line with the results of (Hastie et al., 2022; Mei and Montanari, 2019).

2.5 Additional Setup

Given a compact $\Omega \subset \mathbb{R}^d$, let $\mathscr{C}$ denote the space of continuous $f : \Omega \to \mathbb{R}$, endowed with the supremum norm $\|f\|_\infty = \sup_{x\in\Omega} |f(x)|$. In the classical regression setting, we want to reconstruct a true function $f^* \in \mathscr{C}$ from its values on a training set $x_1, \dots, x_N$, i.e. from the noisy labels $y^\epsilon = \left(f^*(x_1) + \epsilon e_1, \dots, f^*(x_N) + \epsilon e_N\right)$ for some i.i.d. centered noise $e_1, \dots, e_N$ of unit variance and noise level $\epsilon \geq 0$.

In the paper Jacot, Şimşek, et al., 2020d, the observed values (without noise) of the true function f^* consist in observations $o_1, \dots, o_N \in \mathscr{C}^*$, where $\mathscr{C}^*$ is the dual space, i.e. the space of bounded linear functionals $\mathscr{C} \to \mathbb{R}$. We thus represent the training set of N observations $o_1, \dots, o_N$ by the *sampling operator* $\mathcal{O} : \mathscr{C} \to \mathbb{R}^N$ which maps a function $f \in \mathscr{C}$ to the vector of observations $\mathcal{O}(f) = (o_1(f), \dots, o_N(f))^T$. We refer the reader to the paper Jacot, Şimşek, et al., 2020d for the statement of the results and proofs in the general setting.

The classical setting corresponds to the case where the observations are evaluations of f^* at points $x_1, \dots, x_N \in \Omega$, i.e. $o_i\left(f^*\right) = f^*(x_i)$ for $i = 1, \dots, N$. We will restate the operators and the results from Jacot, Şimşek, et al., 2020d that we are interested in this chapter in the classical setting.

The regression problem in the noisy setting is now stated as follows: given noisy observations $y_i^\epsilon = f^*(x_i) + \epsilon e_i$ with i.i.d. centered noises $e_1, \dots, e_N$ of unit variance, how can one reconstruct

f^*? The Kernel Ridge Regression (KRR) predictor with ridge λ is the function $\hat{f}^\epsilon_\lambda : \Omega \to \mathbb{R}$

$$\hat{f}^\epsilon_\lambda = \frac{1}{N} K(x,X)\left(\frac{1}{N}K(X,X) + \lambda I_N\right)^{-1} y^\epsilon$$

where we introduced the rescaling factor of $\frac{1}{N}$ in comparison to the KRR predictor in (2.6). We call the $N \times N$ matrix $G = K(X,X)$ the *Gram matrix*.

Remark. *The KRR predictor arises naturally in the following setup: assuming a (centered) Gaussian Bayesian prior on the true function with covariance operator K and noise amplitude ϵ, the expected posterior, for observed labels y^ϵ is given by $\hat{f}^\epsilon_\lambda$ for $\lambda = \epsilon^2$.*

Useful Operators

We consider the least-squares error (MSE loss) of the KRR predictor, taking into account randomness of: (1) the test point $(x, f^*(x) + \epsilon e)$ which is added a noise ϵe (2) the training data, made of N points $(x_i, f^*(x_i) + \epsilon e_i)$ where $x, x_1, \dots, x_n \sim \mathscr{D}$ and $e, e_1, \dots, e_N \sim \nu$ are i.i.d. (ν has zero mean and unit variance). The expected risk of the KRR predictor is thus taken w.r.t. the test and training observations and their noises. Unless otherwise specified, the expectations are taken w.r.t. all these sources of randomness.

For fixed $x_1, \dots, x_N$, the *empirical risk* or *training error* of the KRR predictor $\hat{f}^\epsilon_\lambda$ is

$$\hat{R}^\epsilon(\hat{f}^\epsilon_\lambda) = \frac{1}{N}\sum_{i=1}^N (\hat{f}^\epsilon_\lambda(x_i) - y^\epsilon_i)^2 = \frac{1}{N}\left\|\hat{f}^\epsilon_\lambda(X) - y^\epsilon\right\|^2.$$

For a test point x sampled from $\mathscr{D}$ and a noise ϵe (where $e \sim \nu$ is centered of unit variance as before), the *risk* $R^\epsilon(\hat{f}^\epsilon_\lambda)$ of the KRR predictor $\hat{f}^\epsilon_\lambda$ is defined by

$$R^\epsilon(\hat{f}^\epsilon_\lambda) = \mathbb{E}_{x\sim\mathscr{D}, e\sim\nu}\left[(\hat{f}^\epsilon_\lambda(x) - f^*(x) - \epsilon e)^2\right].$$

The risk can be rewritten as $R^\epsilon(\hat{f}^\epsilon_\lambda) = \int (\hat{f}^\epsilon_\lambda(x) - f^*(x))^2 \mathscr{D}(dx) + \epsilon^2$. The following three operators enable expressing the risk and empirical risk hence are central to our analysis:

Definition 2.5.1. *The KRR Integral Operator $T_K : \mathscr{C} \to \mathscr{C}$, its empirical version $T_K^N : \mathscr{C} \to \mathscr{C}$, the KRR reconstruction operator $A_\lambda : \mathscr{C} \to \mathscr{C}$ are defined by*

$$[T_K f](x) = \mathbb{E}_{x'\sim\mathscr{D}}\left[f(x')K(x,x')\right], \quad [T_K^N f](x) = \frac{1}{N}\sum_{i=1}^N f(x_i)K(x,x_i),$$
$$A_\lambda = T_K^N (T_K^N + \lambda I_{\mathscr{C}})^{-1}.$$

As $N \to \infty$, we have that $T_K^N \to T_K$, and it follows that

$$A_\lambda \to \tilde{A}_\lambda := T_K(T_K + \lambda I_{\mathscr{C}})^{-1}. \tag{2.11}$$

Eigendecomposition of the Kernel

We will assume that the kernel K can be diagonalized by a countable family of eigenfunctions $(f^{(k)})_{k\in\mathbb{N}}$ in $\mathscr{C}$ with eigenvalues $(d_k)_{k\in\mathbb{N}}$, orthonormal with respect to the scalar product $\int f^{(k)}(x)f^{(\ell)}(x)\mathscr{D}(dx) = \delta_{k\ell}$, such that we have (with uniform convergence):

$$K(x,x') = \sum_{k=1}^{\infty} d_k f^{(k)}(x) f^{(k)}(x').$$

It will be handy to introduce the scalar product notation

$$\langle f, g\rangle_{\mathscr{D}} = \int f(x)g(x)\mathscr{D}(dx). \tag{2.12}$$

The functions $f^{(k)}$ are also eigenfunctions of T_K: we have $T_K f^{(k)} = d_k f^{(k)}$. We will also assume that $\mathrm{Tr}\,[T_K] = \sum_{k=1}^{\infty} d_k$ is finite. Note that in the classical setting K can be diagonalized as above by Mercer's theorem in the domain Ω, and $\mathrm{Tr}\,[T_K] = \mathbb{E}_{x\sim\mathscr{D}}\,[K(x,x)]$ is finite if $\mathscr{D}$ has compact support.

Computing the eigendecomposition of T_K is difficult for general kernels and data distributions, but explicit formulas exist for special cases, such as for the RBF kernel and isotropic Gaussian inputs as described in Section 1.5 of the Appendix of Jacot, Şimşek, et al., 2020c.

2.6 Generalization of Kernel Ridge Regression

2.6.1 Predictor Moments and Signal Capture Threshold (SCT)

A central tool in our analysis of the KRR predictor $\hat{f}^{\epsilon}_{\lambda}$ is the Signal Capture Threshold (SCT):

Definition 2.6.1. *For $\lambda > 0$, the* Signal Capture Threshold ϑ *is the unique positive solution (see Section 2.2 in the Appendix) to the equation:*

$$\vartheta = \lambda + \frac{\vartheta}{N}\mathrm{Tr}\left[T_K\left(T_K + \vartheta I_{\mathscr{C}}\right)^{-1}\right].$$

In this section, we use ϑ and the derivative $\partial_\lambda\vartheta$ for the estimation of the mean and variance of the KRR predictor $\hat{f}^{\epsilon}_{\lambda}$.

Mean predictor

The expected KRR predictor can be expressed in terms of the expected reconstruction operator A_λ

$$\mathbb{E}[\hat{f}^{\epsilon}_{\lambda}] = \mathbb{E}\,[A_\lambda]\, f^*,$$

where we used the fact that $\mathbb{E}_{e_1,\dots,e_N\sim\nu}[y^{\epsilon}] = f^*(X)$.

Theorem 2.6.1. *The expected reconstruction operator* $\mathbb{E}[A_\lambda]$ *is approximated by the operator* $\tilde{A}_\vartheta = T_K (T_K + \vartheta I_{\mathscr{C}})^{-1}$ *in the sense that for all* $f, g \in \mathscr{C}$,

$$\left|\langle f, \left(\mathbb{E}[A_\lambda] - \tilde{A}_\vartheta\right) g\rangle_{\mathscr{D}}\right| \leq \left(\frac{1}{N} + \boldsymbol{P}_0\left(\frac{\mathrm{Tr}[T_K]}{\lambda N}\right)\right) \left|\langle f, \tilde{A}_\vartheta (I_{\mathscr{C}} - \tilde{A}_\vartheta) g\rangle_{\mathscr{D}}\right|,$$

for a polynomial $\boldsymbol{P}_0$ *with nonnegative coefficients and* $\boldsymbol{P}_0(0) = 0$.

This theorem gives the following motivation for the name SCT: if the true function f^* is an eigenfunction of T_K, i.e. $T_K f^* = \delta f^*$, then we have $\tilde{A}_\vartheta f^* = \frac{\delta}{\vartheta+\delta} f^*$ which implies

- if $\delta \gg \vartheta$, then $\frac{\delta}{\vartheta+\delta} \approx 1$ and $\mathbb{E}[A_\lambda] f^* \approx f^*$, i.e. the function is learned on average,
- if $\delta \ll \vartheta$, then $\frac{\delta}{\vartheta+\delta} \approx 0$ and $\mathbb{E}[A_\lambda] f^* \approx 0$, i.e. the function is not learned on average.

More generally, if we decompose a true function f^* along the principal components (i.e. eigenfunctions) of T_K, the signal along the k-th principal component $f^{(k)}$ is captured whenever the corresponding eigenvalue $d_k \gg \vartheta$ and lost when $d_k \ll \vartheta$.

Variance of the predictor

We now estimate the variance of $\hat{f}_\lambda^\epsilon$ along each principal component in terms of the SCT ϑ and its derivative $\partial_\lambda \vartheta$. Along the eigenfunction $f^{(k)}$, the variance is estimated by V_k, where

$$V_k = \frac{\partial_\lambda \vartheta}{N} \left(\left\| (I_{\mathscr{C}} - \tilde{A}_\vartheta) f^* \right\|_{\mathscr{D}}^2 + \epsilon^2 + \langle f^{(k)}, f^* \rangle_{\mathscr{D}}^2 \frac{\vartheta^2}{(\vartheta + d_k)^2} \right) \frac{d_k^2}{(\vartheta + d_k)^2}.$$

Theorem 2.6.2. *There is a constant* $\boldsymbol{C}_1 > 0$ *and a polynomial* $\boldsymbol{P}_1$ *with nonnegative coefficients and with* $\boldsymbol{P}_1(0) = 0$ *such that*

$$\left| \mathrm{Var}\left(\langle f^{(k)}, \hat{f}_\lambda^\epsilon \rangle_{\mathscr{D}} \right) - V_k \right| \leq \left(\frac{\boldsymbol{C}_1}{N} + \boldsymbol{P}_1\left(\frac{\mathrm{Tr}[T_K]}{\lambda N^{\frac{1}{2}}}\right) \right) V_k.$$

Understanding the variance along the principal components (rather than the covariances between the principal components) is enough to describe the risk.

2.6.2 Behavior of the SCT

The behavior of the SCT can be controlled by the following (agnostic of the spectrum of T_K)

Proposition 2.6.3. *For any* $\lambda > 0$, *we have*

$$\lambda < \vartheta \leq \lambda + \frac{1}{N}\mathrm{Tr}[T_K], \qquad 1 \leq \partial_\lambda \vartheta \leq \frac{1}{\lambda}\vartheta,$$

moreover ϑ *is decreasing as a function of* N.

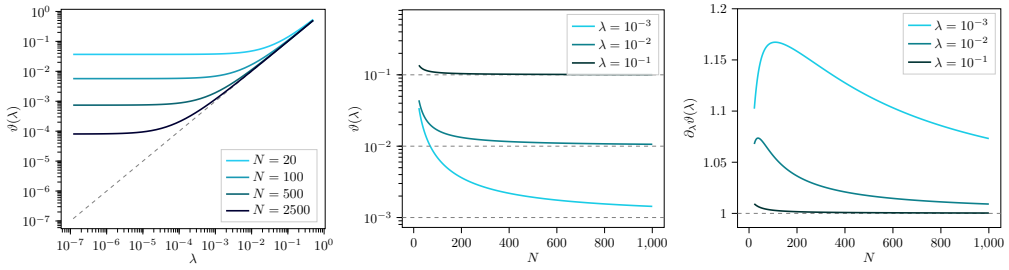


Figure 2.5 – *Signal Capture Threshold and its Derivative.* We consider the RBF Kernel on the standard d-dimensional Gaussian with $\ell = d = 20$. In blue lines, exact formulas for the SCT ϑ and $\partial_\lambda \vartheta$, computed using the explicit formula for the eigenvalues d_k of the integral operator T_K given in Section 1.5 of the Appendix of the paper Jacot, Şimşek, et al., 2020c.

Remark. *As $N \to \infty$, we have ϑ decreases down to λ (see also Figure 2.5), in agreement with the fact that $A_\lambda \to \tilde{A}_\lambda$.*

As $\lambda \to 0$, the above upper bound for $\partial_\lambda \vartheta$ becomes useless. Still, assuming that the spectrum of K has a sufficiently fast power-law decay, we get:

Proposition 2.6.4. *If $d_k = \Theta(k^{-\beta})$ for some $\beta > 1$, there exist $c_0, c_1, c_2 > 0$ such that for any $\lambda > 0$*

$$\lambda + c_0 N^{-\beta} \leq \vartheta \leq c_2 \lambda + c_1 N^{-\beta}, \qquad 1 \leq \partial_\lambda \vartheta \leq c_2.$$

2.6.3 Expected Risk

The expected risk is approximated, in terms of the SCT and the true function f^*, by

$$\tilde{R} = \partial_\lambda \vartheta (\| (I_{\mathscr{C}} - \tilde{A}_\vartheta) f^* \|_{\mathscr{D}}^2 + \epsilon^2),$$

as shown by the following:

Theorem 2.6.5. *There exists a constant $\boldsymbol{C}_2 > 0$ and a polynomial $\boldsymbol{P}_2$ with nonnegative coefficients and with $\boldsymbol{P}_2(0) = 0$, such that we have*

$$\left| \mathbb{E}[R^\epsilon(\hat{f}_\lambda^\epsilon)] - \tilde{R} \right| \leq \left(\frac{\boldsymbol{C}_2}{N} + \boldsymbol{P}_2\left(\frac{\mathrm{Tr}[T_K]}{\lambda N^{\frac{1}{2}}} \right) \right) \tilde{R}.$$

Proof. (Sketch). From the bias-variance decomposition:

$$\mathbb{E}[R(\hat{f}_\lambda^\epsilon)] = R(\mathbb{E}[\hat{f}_\lambda^\epsilon]) + \sum_{k=1}^{\infty} \mathrm{Var}(\langle f^{(k)}, \hat{f}_\lambda^\epsilon \rangle_{\mathscr{D}}).$$

By Theorem 2.6.1 and a small calculation, the bias is approximately $\| (I_{\mathscr{C}} - \tilde{A}_\vartheta) f^* \|_{\mathscr{D}}^2 + \epsilon^2$. By Theorem 2.6.2 and a calculation, the variance is approximately $(\partial_\lambda \vartheta - 1)(\| (I_{\mathscr{C}} - \tilde{A}_\vartheta) f^* \|_{\mathscr{D}}^2 + \epsilon^2)$. □

The approximate expected risk $\tilde{R}$ is increasing in both ϑ and $\partial_\lambda \vartheta$. As λ increases, the bias increases with ϑ, while the variance decreases with $\partial_\lambda \vartheta$: this leads to the bias-variance tradeoff. On the other hand, as a function of N, ϑ is decreasing but $\partial_\lambda \vartheta$ is generally not monotone: this can lead to so-called multiple descent curves in the risk as a function of N (Liang, Rakhlin, and Zhai, 2020).

Note also that if the true function is in RKHS, we can decompose it along the principal components $f^* = \sum_{k=1}^{\infty} b_k f^{(k)}$. The risk is then approximated by

$$\tilde{R}(f^*) = \partial_\lambda \vartheta \left(\sum_{k=1}^{\infty} \frac{\vartheta^2}{(\vartheta + d_k)^2} b_k^2 + \epsilon^2 \right).$$

Remark. *For a decaying ridge $\lambda = cN^{-\gamma}$ for $0 < \gamma < \frac{1}{2}$, as $N \to \infty$, by Proposition 2.6.3, we get $\vartheta \to 0$ and $\partial_\lambda \vartheta \to 1$: this implies that $\mathbb{E}[R(\hat{f}_\lambda^\epsilon)] \to \epsilon^2$ if f^* is in the RKHS associated with K.*

Remark. *In a Bayesian setting, assuming that f^* is random with zero mean and covariance kernel Σ, the optimal choices for the KRR predictor are $K = \Sigma$ and $\lambda = \epsilon^2/N$. When $K = \Sigma$ and $\lambda = \epsilon^2/N$, the formula of $\tilde{R}$ simplifies to*

$$\mathbb{E}\left[R\left(\hat{f}_\lambda^\epsilon\right)\right] \approx N\vartheta.$$

The empirical risk (or train error) $\hat{R}(\hat{f}_\lambda^\epsilon) = \lambda^2 (y^\epsilon)^T (\frac{1}{N} G + \lambda I_N)^{-2} y^\epsilon$ can be analyzed with the same theoretical tools. Its approximation in terms of the SCT is given as follows:

Theorem 2.6.6 (Theorem 17 in the Appendix)**.** *There exists a constant $\boldsymbol{C}_3 > 0$ and a polynomial $\boldsymbol{P}_3$ with nonnegative coefficients and with $\boldsymbol{P}_3(0) = 0$ such that we have*

$$\left| \mathbb{E}[\hat{R}(\hat{f}_\lambda^\epsilon)] - \frac{\lambda^2}{\vartheta^2} \tilde{R} \right| \leq \left(\frac{1}{N} + \boldsymbol{P}_3\left(\frac{\mathrm{Tr}[T_K]}{\lambda N}\right) \right) \tilde{R}.$$

2.7 Conclusion

In the first part, we have identified the implicit regularization arising from the finite sampling of Random Features (RF): using a Gaussian RF model with ridge parameter $\lambda > 0$ (λ-RF) is in expectation equivalent to using a Kernel Ridge Regression with effective ridge $\tilde{\lambda} > \lambda$ ($\tilde{\lambda}$-KRR) which we characterize explicitly (Theorem 2.4.3). The λ-RF predictor concentrates around its expectation when λ is bounded away from zero for large P (Theorem 2.4.6); this implies in particular that the risks of the λ-RF and $\tilde{\lambda}$-KRR predictors are close to each other (Corollary A.6.1). Both theorems are proven using tools from random matrix theory, in particular finite-size results on the concentration of the Stieltjes transform of general Wishart matrix models.

Our numerical verifications on the expected λ-RF predictor and the $\tilde{\lambda}$-KRR predictor have shown that both are in excellent agreement. This shows in particular that in order to use RF predictors to approximate KRR predictors with a given ridge, one should choose both the

number of features and the explicit ridge appropriately.

Finally, we investigate the ridgeless limit case $\lambda \searrow 0$. In this case, we see a sharp transition at $\gamma = 1$: in the overparameterized regime $\gamma > 1$, the effective ridge goes to zero, while in the underparameterized regime $\gamma < 1$, it converges to a positive value. At the interpolation threshold $\gamma = 1$, the variance of the λ-RF explodes, leading to the double descent curve emphasized in (Advani, Saxe, and Sompolinsky, 2020; Spigler et al., 2018; Belkin, Hsu, Ma, et al., 2018; Nakkiran et al., 2019). We investigate this numerically and prove a lower bound yielding a plausible explanation for this phenomenon.

In the second part, we studied the Kernel Ridge Regression (KRR) predictor and its risk. We obtain new precise estimates for the test and train error in terms of a new object, the Signal Capture Threshold (SCT), which identifies the components of a true function that are being learned by the KRR. Our estimates reveal a remarkable relation between the expected risk and expected empirical risk of the KRR predictor.

While our current proofs require the Gaussianity assumption, it seems natural to postulate that the results and the proofs extend to more general setups, along the lines of Louart, Liao, and Couillet, 2017; Benigni and Péché, 2019. To complete the study of the risk of the RF predictor of the gaussian random features model, what remains is the study of the variance of the RF predictor.

3 Deep Linear Networks

In this chapter, we present the preprint Jacot, Ged, Şimşek, et al., 2022, focusing on the parts where the thesis author has more contributions than the other parts, in particular the loss landscape. We present the main results in Section 3.1, related works in Section 3.2, and problem setup in Section 3.3. We present the first main theorem on the scaling of the distance from a typical initialization to the closest saddle point and to the closest global minimum in terms of width in Section 3.4. The proof is presented in Section B. Section 3.5 presents a new training regime called Saddle-to-Saddle which may be relevant to the training of non-linear networks too, due to symmetry-induced saddles (see Chapter 4). The main theorem here shows that the typical gradient flow path initialized very small follows the path followed by the width 1 linear network up to an inclusion and rotation (see the preprint Jacot, Ged, Şimşek, et al., 2022 for the proof).

In Section 3.6, we compare the saddle-to-saddle regime with the commonly studied limiting training dynamics of Mean-Field and the Neural Tangent Kernel. We note here that our main theorem in Section 3.5 applies to finite-width networks and we conjecture that the gradient flow follows the trajectories of the narrower networks for general widths. In this thesis, we argue that this is due to the particular arrangement of the saddles in linear networks: where the typical escape path of a saddle is within the attractive manifold of the next saddle. It remains an open question whether the hiererchy of the saddles follow the same specialized structure in non-linear networks. We present the conclusions in Section 3.7.

3.1 Main Results

We study deep linear networks $x \mapsto A_\theta x$ of depth $L \geq 1$ and widths $n_0, ..., n_L$, that is $A_\theta = W_L ... W_1$ where $W_1, ..., W_L$ are matrices such that $W_i \in \mathbb{R}^{n_i \times n_{i-1}}$ and θ is a vector that consists of all the (learnable) parameters of the DLN, i.e. the components of the matrices $W_1, ..., W_L$. For any general convex cost $C : \mathbb{R}^{n_L \times n_0} \to \mathbb{R}$ on matrices, we investigate the gradient flow minimizing the loss $\mathscr{L}(\theta) = C(A_\theta)$. To ease the notation, suppose that the hidden layers have the same size, that is $w = n_1 = ... = n_{L-1}$.

The variance of the parameters at initialization has a profound effect on the training dynamics. If the parameters are initialized with variance $\sigma^2 = w^{-\gamma}$, where w is the size of the hidden layers, we observe a transition in the infinite width limit as $w \to \infty$ as shown in Theorem 3.4.1:

- when $\gamma < 1$, the random initialization θ_0 is (with high probability) very close to a global minimum and very far from any saddle,
- when $\gamma > 1$, the initialization is very close to a saddle and far from any global minimum.

The case $\gamma < 1$ corresponds to the NTK regime (or kernel/lazy regime of Jacot, Gabriel, and Hongler, 2018a) and the case $\gamma = 1$ corresponds to the Mean-Field limit (or the Maximal Update parametrization of G. Yang, 2019). It appears that the case $\gamma > 1$ has been much less studied in previous works.

To understand this regime, we investigate in Section 3.5 the case $\gamma \to +\infty$. More precisely, we fix the width of the network and let the variance at initialization go to zero. We show in Theorem 3.5.1 that the gradient flow trajectory asymptotically goes from the saddle at the origin $\vartheta^0 = 0$ to a rank-one saddle ϑ^1, i.e. a saddle where the matrices $W_1, \dots, W_L$ are of rank 1. The proof is based on a new description, in the spirit of the Hartman-Grobman theorem, of the so-called fast escape paths at the origin. This theorem may be of independent interest.

We propose the Conjecture 3.5.2, backed by numerical experiments, describing the full gradient flow when the variance at initialization is very small, suggesting that it goes from saddle to saddle, visiting the neighborhoods of a sequence of critical points $\vartheta^0, \dots, \vartheta^K$ (the first K ones being saddle points, the last one being either a global minimum or a point at infinity) corresponding to matrices of increasing ranks. This is consistent with Gissin, Shalev-Shwartz, and Daniely, 2020 which shows that incremental learning occurs in a toy model of DLNs and that gradient-based optimization hence has an implicit bias towards simple (sparse) solutions.

In Section 3.5, we show how this Saddle-to-Saddle dynamics can be described using a greedy low-rank algorithm which bears similarities with that of Z. Li, Y. Luo, and Lyu, 2020 and leads to a low-rank bias of the final learned function. This is in stark contrast to the NTK regime which features no low-rank bias.

3.2 Related Works

The existence of distinct regimes in the training dynamics of DNNs has been explored in previous works, both theoretically (Chizat and Bach, 2018a; G. Yang, 2019) and empirically (Geiger, Spigler, Jacot, et al., 2020). The loss landscape of shallow linear neural networks has been characterized by Baldi and Hornik, 1989. The recent theoretical works (Chizat and Bach, 2018a; G. Yang, 2019) have mostly focused on the transition from the NTK regime ($\gamma < 1$) to the Mean-Field regime ($\gamma = 1$). This paper is focused on the regime beyond the critical one ($\gamma > 1$).

Our study of the Saddle-to-Saddle dynamics can also be understood as a generalization of the

works (Saxe, McClelland, and Ganguli, 2014; Advani and Saxe, 2017; Saxe, McClelland, and Ganguli, 2019; Gidel, Bach, and Lacoste-Julien, 2019; Arora, Cohen, W. Hu, et al., 2019) which describe a similar plateau effect in a very specific setting and with a very carefully chosen initialization.

Shortly after the initial publication of this article, we came aware of the paper of Z. Li, Y. Luo, and Lyu, 2020 which provides a similar description to our Saddle-to-Saddle dynamics. For shallow networks, the results are almost equivalent, although the techniques are very different, especially when dealing with the fact that the escape directions (and escape paths) are unique only up to rotations. Z. Li, Y. Luo, and Lyu, 2020 use a clever trick that allows them to both study the dynamics of the output matrix $A_{\theta(t)}$, without the need to keep track of the parameters, and obtain a unicity property for the asymptotic dynamics. Instead, we focus on the dynamics of the parameters, give an identification of all optimal escape paths, and show that the path followed by the parameters' dynamics is unique up to symmetries of the network. Note also that, as in our paper, Z. Li, Y. Luo, and Lyu, 2020 only prove the first step of the Saddle-to-Saddle regime: for the subsequent steps, it is assumed that the next saddle is not approached along a 'bad' direction. For deep networks, our results are more general as they hold for more general initializations than in Z. Li, Y. Luo, and Lyu, 2020. Indeed, in order to avoid the non-uniqueness problem of the escape paths in the space of parameters, their analysis relies heavily on the assumption that the network is balanced at initialization. Our analysis does not rely on this trick which is a crucial step to generalize this type of result to nonlinear networks, where balancedness cannot be used.

3.3 Setup

A DLN of depth L and widths $n_0, \dots, n_L$ is the composition of L matrices

$$A_\theta = W_L \cdots W_1$$

where $W_\ell \in \mathbb{R}^{n_\ell \times n_{\ell-1}}$. The number of parameters is $P = \sum_{\ell=1}^{L} n_{\ell-1} n_\ell$ and we denote by $\theta = (W_L, \dots, W_1) \in \mathbb{R}^P$ the vector of parameters. The input dimension, resp. the output dimension is n_0, resp. n_L. All parameters are initialized as i.i.d. $\mathcal{N}(0, \sigma^2)$ Gaussian random variables.

We will focus on the so-called rectangular networks, in which the number of neurons in all hidden layers is the same, i.e. $n_1 = \cdots = n_{L-1} = w$. Such rectangular network is called a (L, w)-DLN, and its number of parameters is denoted by $P = n_0 w + (L-2) w^2 + w n_L$. The proofs given in this article can be extended to the non-rectangular case.

We study the dynamics of gradient descent on the loss $\mathscr{L}(\theta) = C(A_\theta)$ for a general twice-differentiable and convex cost C on $n_L \times n_0$ matrices. To ensure a non-trivial minimisation problem, we assume that the null matrix is not a global minimum of C: in this case, the origin in the parameter space is a saddle of $\mathscr{L}$. Given a point $\theta_0 \in \mathbb{R}^P$, we denote by $t \mapsto \Gamma(t, \theta_0)$ the gradient flow path on the cost $L(\theta)$ starting from θ_0, i.e. $\Gamma(0, \theta_0) = \theta_0$ and $\partial_t \Gamma(t, \theta_0) =$

$-\nabla\mathscr{L}(\Gamma(t,\theta_0))$.

While our analysis applies to general costs, some typical costs used in practice are:

Mean-Squared Error (MSE): $C(A) = \frac{1}{N}\|AX - Y\|_F^2$ for some inputs $X \in \mathbb{R}^{n_0 \times N}$ and labels $Y \in \mathbb{R}^{n_L \times N}$, where $||\cdot||_F$ is the Frobenius norm.

Matrix Completion (MC): $C(A) = \frac{1}{N}\sum_{i=1}^{N}(A_{k_i,m_i} - A^*_{k_i,m_i})^2$ for some true matrix A^* of which we observe only the N entries $A^*_{k_1,m_1}$,..., $A^*_{k_N,m_N}$.

3.4 The Loss Landscape

Symmetries and Invariance

A key tool in this paper is the use of two important symmetries of the parametrization map $\theta \mapsto A_\theta$ in DLNs: rotations of hidden layers and inclusions in wider DLNs.

Rotations: A $L-1$ tuple $R = (O_1, \dots, O_{L-1})$ of orthogonal $w \times w$ matrices is called a w-width network rotation, or in short a rotation. A rotation R acts on a parameter vector $\theta = (W_L, \dots, W_1)$ as $R\theta = (W_L O_{L-1}^T, O_{L-1}W_{L-1}O_{L-2}^T, \dots, O_1 W_1)$. The space of rotations is an important symmetry of DLN: indeed, for any parameter θ and any cost C, the two following important properties hold:

$$A_{R\theta} = A_\theta, \quad \nabla_\theta C(A_{R\theta}) = R\nabla_\theta C(A_\theta),$$

where we considered $\nabla_\theta C(A_\theta) \in \mathbb{R}^P$ as another vector of parameters. This property implies that if $\theta(t) = \Gamma(t,\theta_0)$ is a gradient flow path, then so is $R\theta(t) = \Gamma(t, R\theta_0)$.

Inclusion: The inclusion $I^{(w'\to w)}(\theta) = (V_L, \dots, V_1)$ of a network of width w into a wider network ($w > w'$) simply adds zero neurons

$$V_1 = \begin{pmatrix} W_1 \\ 0 \end{pmatrix}, V_\ell = \begin{pmatrix} W_\ell & 0 \\ 0 & 0 \end{pmatrix}, V_L = \begin{pmatrix} W_L & 0 \end{pmatrix}.$$

For any parameters θ and any cost C, we have $A_{I^{(w'\to w)}\theta} = A_\theta$ and $\nabla C(A_{I^{(w'\to w)}\theta}) = I^{(w'\to w)}\nabla C(A_\theta)$. This property implies that if $\theta(t) = \Gamma(t,\theta_0)$ is a gradient flow path of the (w', L) network, then $I^{(w'\to w)}\theta(t) = \Gamma(t, I^{(w'\to w)}\theta_0)$ is a gradient flow path of the (w, L) network (as well as any rotation $RI^{(w'\to w)}\theta(t)$ thereof).

Proximity of Critical Points at Initialization

It has already been observed that in the infinite width limit, when the width w of the network grows to infinity, the scale at which the variance σ^2 of the parameters at initialization scales with the width can lead to very different behaviors (Chizat and Bach, 2018a; Geiger, Spigler, Jacot, et al., 2020; G. Yang, 2019). Let us consider scaling of the variance $\sigma^2 = w^{-\gamma}$ for $\gamma \geq 1 - \frac{1}{L}$.

The reason we lower bound γ is that any smaller γ would lead to an explosion of the variance of the matrix A_θ at initialization as the width w grows.

Let d_{m} and d_{s} be the Euclidean distances between the initialization θ and, respectively, the set of global minima and the set of all saddles. For random variables $f(w), g(w)$ which depend on w, we write $f \asymp g$ if both $f(w)/g(w)$ and $g(w)/f(w)$ are stochastically bounded as $w \to \infty$. The following theorem studies how d_{m} and d_{s} scale as $w \to \infty$:

Theorem 3.4.1. *Suppose that the set of matrices that minimize C is non-empty, has Lebesgue measure zero, and does not contain the zero matrix. Let θ be i.i.d. centered Gaussian r.v. of variance $\sigma^2 = w^{-\gamma}$ where $1 - \frac{1}{L} \leq \gamma < \infty$. Then:*

1. *if $\gamma < 1$, then $d_{\mathrm{m}} \asymp w^{-\frac{(1-\gamma)(L-1)}{2}}$ and $d_{\mathrm{s}} \asymp w^{\frac{1-\gamma}{2}}$,*
2. *if $\gamma = 1$, then $d_{\mathrm{m}}, d_{\mathrm{s}} \asymp 1$,*
3. *if $\gamma > 1$, then $d_{\mathrm{m}} \asymp 1$ and $d_{\mathrm{s}} \asymp w^{-\frac{\gamma-1}{2}}$.*

This theorem shows an important change of behavior between the case $\gamma < 1$ and $\gamma > 1$. When $\gamma < 1$, the network is initialized very close to a global minimum and far from any saddle. When $\gamma > 1$, the parameters are initialized very close to a saddle but far away from any global minimum. The critical case $\gamma = 1$ is the unique limit where both types of critical points are at the same distance from the initialization.

Hence, the landscape of the loss near the initialization displays distinct features in the three regimes highlighted in the previous theorem, and this leads to very different gradient dynamics. In Appendix B.1, we show that the largest initialization, corresponding to the choice $\gamma = 1 - \frac{1}{L}$, is equivalent to the so-called NTK parametrization of Jacot, Gabriel, and Hongler, 2018b, up to a rescaling of the learning rate. In the range $1 - \frac{1}{L} < \gamma < 1$, G. Yang and E. J. Hu, 2020 obtain a similar, yet slightly different, kernel regime. The initialization $\gamma = 1$ corresponds to the Mean-Field limit for shallow networks (Chizat and Bach, 2018b; G. Rotskoff and Vanden-Eijnden, 2018) or, more generally, to the Maximal Update parametrization of G. Yang and E. J. Hu, 2020 (see Appendix B.1). The case $\gamma > 1$ is however much less studied and is difficult to study since the initialization approaches a saddle as $w \to \infty$. Thus, in this regime, the wider the network, the longer it takes to escape this nearby saddle and, in the limit as $w \to \infty$, nothing happens over a finite number of gradient steps. With the right time parametrization, we will observe interesting Saddle-to-Saddle dynamics in this regime, leading to some low-rank bias. This is regime is related to the condensed regime identified in T. Luo et al., 2021.

3.5 Saddle-to-Saddle Training Dynamics

In contrast to the NTK regime ($\gamma < 1$) where gradient flow never approaches any saddle, we will see how in the Saddle-to-Saddle regime gradient flow is affected not only by the saddle it is

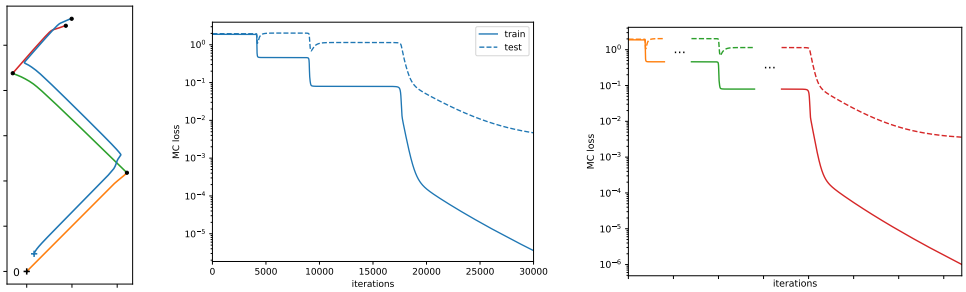


Figure 3.1 – *Saddle-to-Saddle dynamics:* A DLN ($L = 4, w = 100$) with a small initialization ($\gamma = 2$) trained on a MC loss fitting a 10×10 matrix of rank 3. *Left:* Projection onto a plane of the gradient flow path θ_α in parameter space (in blue) and of the sequence of 3 paths $\theta^1, \theta^2, \theta^3$ (in orange, green and red), described by Algorithm $\mathscr{A}_{\epsilon,T,\eta}$, starting from the origin (+) and passing through 2 saddles ($\cdot$) before converging. *Middle:* Train (solid) and test (dashed) MC costs through training. We observe three plateaus, corresponding to the three saddles visited. *Right:* The train (solid) and test (dashed) losses of the three paths plotted sequentially, in the saddle-to-saddle limit; the dots represent an infinite amount of steps separating these paths.

initialized close to but by a sequence of saddles. This leads to a bias towards learning low-rank matrices which is absent in the NTK regime (Woodworth et al., 2020).

We now study the dynamics of DLN during training as the variance at initialization goes to zero, under the assumption that it is representative of the whole Saddle-to-Saddle regime. Specifically, we sample some random parameters θ_0 with i.i.d. $\mathcal{N}(0,1)$ entries, consider the gradient flow $\theta_\alpha(t) = \Gamma(t, \alpha\theta_0)$, and let $\alpha \searrow 0$. Since the origin is a saddle, for all fixed times t, $\lim_{\alpha \searrow 0} \theta_\alpha(t) = 0$. We will show however that there is an escape time t_α, which grows to infinity as $\alpha \searrow 0$, such that the limit $\lim_{\alpha \searrow 0} \theta_\alpha(t_\alpha + t)$ is non-trivial for all $t \in \mathbb{R}$.

The study of shallow networks ($L = 2$) is facilitated by the fact that the saddle at the origin is strict: its Hessian has negative eigenvalues. For deeper networks ($L > 2$), the saddle is highly degenerate: the $L-1$ first order derivatives vanish.

First Path

It turns out that gradient flow paths naturally escape the saddle at the origin along so-called *optimal escape paths*. We say that a gradient flow path $\theta(\cdot) : \mathbb{R} \to \mathbb{R}^P$ is an *escape path* of a critical point θ^* if $\lim_{t \to -\infty} \theta(t) = \theta^*$. Informally, the optimal escape paths, are the escape paths that allow the fastest exit from a saddle. In DLNs, these optimal escape paths are of the form $RI^{(1 \to w)} \underline{\theta}^1(t)$ where $\underline{\theta}^1(t)$ is a path of a width 1 DLN which escapes from the origin:

Theorem 3.5.1. *Assume that the largest singular value s_1 of the gradient of C at the origin $\nabla C(0) \in \mathbb{R}^{n_L \times n_0}$ has multiplicity 1. There is a deterministic gradient flow path $\underline{\theta}^1$ in the space of width-1 DLNs such that, with probability 1 if $L \leq 3$, and probability at least $1/2$ if $L > 3$, there*

exists an escape time t_α^1 *and a rotation* R *such that*

$$\lim_{\alpha\to 0} \theta_\alpha(t_\alpha^1 + t) = RI^{(1\to w)}\underline{\theta}^1(t).$$

The unicity of the largest singular value of the gradient at the origin guarantees the unicity (up to rotation) of the optimal escape paths. For example, with the MSE loss, the gradient at the origin is $2YX^T$: for generic Y and X, the largest singular value of the gradient has a multiplicity of 1.

The reason why, for DLN with $L > 3$, we can only guarantee a probability of $\frac{1}{2}$ in the previous theorem, is that we need to ensure that gradient descent does not get stuck at the saddle at the origin or at other saddles connected to it. For $L = 2$, this follows from the fact that the saddle is strict. When $L > 2$, the saddle is not strict and we were only able to prove it in the case where $L = 3$. We conjecture that Theorem 3.5.1 is true with probability 1 for all $L \geq 2$.

As shown in the Appendix of our paper (Jacot, Ged, Şimşek, et al., 2022), the escape time t_α is of order $-\log\alpha$ for shallow networks and of order $\alpha^{-(L-2)}$ for networks of depth $L > 2$. Hence, the deeper the network, the slower the gradient flow escapes the saddle. Besides, the norm $\left\|\theta^1(t)\right\|$ of the limiting escape path $\theta^1(t) = RI^{(1\to w)}\underline{\theta}^1(t)$ grows at an optimal speed: as $e^{s^*(t+T)}$ for some T when $L = 2$; and as $(s^*(T-t))^{-\frac{1}{L-2}}$ for some T when $L > 2$, where s^* is the optimal escape speed $s^* = L^{-\frac{L-2}{2}}s_1$. These are optimal in the sense that given another gradient flow path $\theta(t)$ which exits from the origin, there exists a ball B centered at the origin such that, for any small ϵ, if t_1 and t_2 are the times such that $\left\|\theta^1(t_1)\right\| = \epsilon = \|\theta(t_2)\|$, then $\left\|\theta^1(t+t_1)\right\| \geq \|\theta(t+t_2)\|$ for any positive t, until one of the paths exits the ball B.

Subsequent Paths

What happens after this first path? The width-1 gradient flow path $\underline{\theta}^1(t)$ converges to a width-1 critical point $\underline{\vartheta}^1$ as $t \to \infty$. While $\underline{\vartheta}^1$ may be a local minimum amongst width-1 DLNs, its inclusion $\vartheta^1 = RI^{(1\to w)}(\underline{\vartheta}^1)$ will be a saddle assuming it is not a global minimum already and that the network is wide enough, since if $w \geq \min\{n_0, n_L\}$ all critical points are either global minima or saddles (Laurent and Brecht, 2018; Nouiehed and Razaviyayn, 2021).

Theorem 3.5.1 guarantees that, as $\alpha \searrow 0$, the gradient flow path $\theta_\alpha(t)$ will approach a saddle ϑ^1. It is then natural to assume that $\theta_\alpha(t)$ will escape this saddle along an optimal escape path (which is the inclusion of a width-2 path). Repeating this process, we expect gradient flow to converge as $\alpha \searrow 0$ to the concatenation of paths going from saddle to saddle of increasing width:

Conjecture 3.5.2. *With probability* 1*, there exist* $K+1$ *critical points* $\vartheta^0, \dots, \vartheta^K \in \mathbb{R}^{P_{L,w}}$ *(with* $\vartheta^0 = 0$*) and* K *gradient flow paths* $\theta^1, \dots, \theta^K : \mathbb{R} \to \mathbb{R}^{P_{L,w}}$ *connecting the critical points (i.e.* $\lim_{t\to-\infty}\theta^k(t) = \vartheta^{k-1}$ *and* $\lim_{t\to+\infty}\theta^k(t) = \vartheta^k$*) such that the path* $\theta_\alpha(t)$ *converges as* $\alpha \to 0$ *to the concatenation of* $\theta^1(t), \dots, \theta^K(t)$ *in the following sense: for all* $k < K$*, there exist times* t_α^k

(which depend on θ_0) such that

$$\lim_{\alpha\to 0} \theta_\alpha(t_\alpha^k + t) = \theta^k(t).$$

Furthermore, for all $k < K$, there is a deterministic path $\underline{\theta}^k(t)$ and a local minimum $\underline{\vartheta}^k$ of a width-k network such that for some rotation R (which depends on θ_0), $\theta^k(t) = RI^{(k\to w)}(\underline{\theta}^k(t))$ and $\vartheta^k = RI^{(k\to w)}(\underline{\vartheta}^k)$ for all k and t.

This Saddle-to-Saddle behavior explains why for small initialization scale, the train error gets stuck at plateaus during training (Figures 3.1 and 3.2). Conjecture 3.5.2 suggests that these plateaus correspond to the saddle visited.

Note that for losses such as the cross-entropy, the gradient descent may diverge towards infinity, as studied in Soudry et al., 2018; Gunasekar, J. Lee, et al., 2018. From now on, we focus on the case where ϑ^K is a finite global minimum. By the invariance under gradient flow of $\mathrm{Im}[I^{(k\to w)}]$ (the image of the inclusion map), the inclusion of a width-k local minimum $\underline{\vartheta}^k$ into a larger network is a saddle ϑ^k (if A_{ϑ^k} is not a global minimum of C). These types of saddles are closely related to the symmetry-induced saddles studied in Şimşek, Ged, et al., 2021 in nonlinear networks.

Greedy Low-Rank Algorithm

Conjecture 3.5.2 suggests that the gradient flow with vanishing initialization implements a greedy low-rank algorithm which performs a greedy search for a lowest-rank solution: it first tries to fit a width 1 network, then a width 2 network and so on until reaching a solution. Thus, we expect that as $\alpha \searrow 0$, the dynamics of gradient flow corresponds, up to inclusion and rotation, to the limit of the algorithm $\mathscr{A}_{\epsilon,T,\eta}$ as sequentially $T \to \infty$, $\eta \to 0$ and $\epsilon \to 0$. In particular, we used the Algorithm $\mathscr{A}_{\epsilon,T,\eta}$, with large T and small η and ϵ to approximate the paths $\underline{\theta}^k$ and points $\underline{\vartheta}^k$ in Figure 3.1. Note how this limiting algorithm is deterministic. This implies that even for finite widths the dynamics of gradient flow converge to a deterministic limit (up to random rotations R) as the variance at initialization goes to zero.

A similar algorithm has already been described in Z. Li, Y. Luo, and Lyu, 2020, however thanks to our different proof techniques, we are able to give a more precise description of the evolution of the parameters.

Remark. *To prove Conjecture 3.5.2, one needs to apply a similar argument to understand how gradient flow escapes the subsequent saddles $\vartheta^1, \dots, \vartheta^K$. There are two issues:*

First, even though Theorem 3.5.1 guarantees that gradient descent will come arbitrarily close to the next saddle ϑ^1, it may not approach it along a generic direction: it could approach along a "bad" direction. For the first path, we relied on the fact that θ_0 is Gaussian to guarantee that these bad directions are avoided with probability 1 *(or* 1/2*). Note that this problem could be addressed using the so-called perturbed stochastic gradient descent described in Jin et al., 2017a; S. S. Du, Jin, J. D. Lee, M. I. Jordan, Póczos, et al., 2017 since, in this learning algorithm, once in*

Algorithm 1 $\mathscr{A}_{\epsilon,T,\eta}$

Compute the first singular vectors of $\nabla C(0)$:
$u, s, v \leftarrow \mathrm{SVD}_1(\nabla C(0))$
$\theta \leftarrow (-\epsilon v^T, \epsilon, \dots, \epsilon u)$
$w \leftarrow 1$
while $C(A_\theta) < C_{min} + \epsilon$ **do**
  # T steps of GD on the loss of width-w DLN with lr η
  $\theta \leftarrow \mathrm{SGD}_{w,T,\eta}(\theta)$
  $u, s, v \leftarrow \mathrm{SVD}_1(\nabla C(A_\theta))$
  $\theta \leftarrow \left(\begin{pmatrix} W_1 \\ -\epsilon v^T \end{pmatrix}, \begin{pmatrix} W_2 & 0 \\ 0 & \epsilon \end{pmatrix}, \dots, \begin{pmatrix} W_L & \epsilon u \end{pmatrix} \right)$
  $w \leftarrow w + 1$
end while

the vicinity of the saddle, a small Gaussian noise is added to the parameters: as a consequence, they end up being in a generic position in the neighborhood of the saddle.

Second, for deep networks ($L > 2$), the saddle ϑ^1 has a different local structure to ϑ^0. Indeed, at the origin, the $L-1$ first derivatives vanish, leading to an (approximately) L-homogeneous saddle at the origin. On the contrary, at the rank 1 saddle $\vartheta^1 = RI^{(1\to w)}(\underline{\vartheta}^1)$, if $\underline{\vartheta}^1$ is a local minimum of the width 1 network, the Hessian is positive along the inclusion $\mathrm{Im}\left[RI^{(1\to w)}\right]$. This implies that the dynamics can only escape the saddle through the Hessian null-space, along which the first $L-1$ derivatives vanish. Although the loss restricted to this null-space around ϑ_1 has a similar structure to the loss around the origin, the fact that the Hessian at ϑ_1 is not null complexifies the analysis.

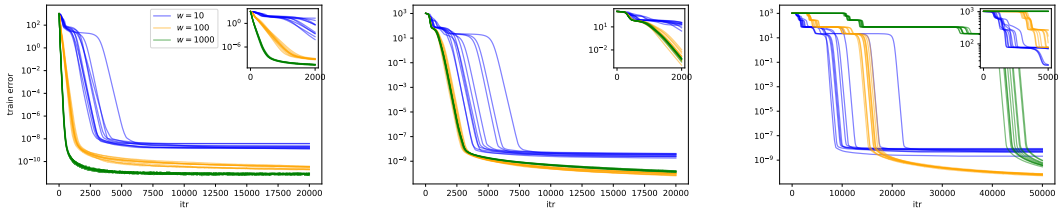
Description of the paths that escape a saddle

Our proof relies on a theorem which relates the escape paths of the saddle at the origin of the cost $\mathscr{L}$ and the escape paths of the L-th order Taylor approximation H of $\mathscr{L}$. This correspondence only applies to paths which escape the saddle sufficiently fast.

We define the set of fast escaping paths $F_{\mathscr{L}}(s)$ of the cost $\mathscr{L}$ with speed at least s as follows:

- for shallow networks ($L = 2$), it is the set of gradient flow paths that satisfy $\|\theta(t)\| = O(e^{st})$ as $t \to -\infty$,
- for deep networks ($L > 2$), it is the set of gradient flow paths that satisfy $\|\theta(t)\| \leq (s(T-t))^{-\frac{1}{L-2}}$ for some T and any small enough t.

The optimal escape speed is $s^* = L^{-\frac{L-2}{2}} s_1$ where s_1 is the largest singular value of $\nabla C(0)$. It is the optimal escape speed in the sense that there are no faster escape paths: $F_{\mathscr{L}}(s) = \emptyset$ if $s > s^*$. Escape paths which exit the saddle at the optimal escape speed are called optimal escape paths. There is a bijection between fast escaping paths of the loss $\mathscr{L}$ and those of its L-th order



(a) $\gamma = 0.75$ (NTK) (b) $\gamma = 1$ (MF) (c) $\gamma = 1.5$ (S-S)

Figure 3.2 – *Training in (a) NTK, (b) mean-field, (c) saddle-to-saddle regimes in deep linear networks for three widths* $w = 10, 100, 1000$*,* $L = 4$*, and* 10 *seeds.* Parameters are initialized with variance $\sigma^2 = w^{-\gamma}$. We observe that *(a)* in the NTK regime, the training loss shows typical linear convergence behavior for $w = 1000$ and $w = 100$; *(b)* in the mean-field regime, we observe that even the large width networks approach to a saddle at the beginning of the training and that the length of the plateaus remains constant between widths $w = 1000$ and $w = 100$; *(c)* in the saddle-to-saddle regime, the plateaus become longer as the width grows. In all cases, we see a reduction in the variation between the different seeds as $w \to \infty$.

Taylor approximation H in the paper of Jacot, Ged, Şimşek, et al., 2022 that is similar to the Hartman-Grobman Theorem and might be of independent interest.

3.6 Regimes of Training

In light of the results presented in this paper, we discuss the three regimes that can be obtained by varying the initialization. We expect these characteristics to translate to the nonlinear case.

The NTK limit ($\gamma = 1 - \frac{1}{L}$) (Jacot, Gabriel, and Hongler, 2018b; J. Lee, Xiao, et al., 2019) is representative of the other scalings $1 - \frac{1}{L} \leq \gamma < 1$ (G. Yang and E. J. Hu, 2020). The critical regime $\gamma = 1$ corresponds to the Mean-Field limit for shallow networks (Chizat and Bach, 2018b; G. Rotskoff and Vanden-Eijnden, 2018) or the Maximal Update parametrization for deep networks (G. Yang and E. J. Hu, 2020). Finally, we conjecture that the last regime where $\gamma > 1$, displays features very akin to the $\gamma = +\infty$ case studied in this article. Under this assumption, we obtain the following characterizations of the regimes:

In the **NTK regime** ($1 - \frac{1}{L} \leq \gamma < 1$):

1. During training, the parameters converge to a nearby global minimum, and do not approach any saddle (Figure 3.2a shows how the plateaus disappear as w grows).
2. If the cost on matrices C is strictly convex, one can guarantee exponential decay of the loss.
3. The NTK is asymptotically fixed during training.
4. No sparsity/low-rank bias in the learned matrix.

The **Saddle-to-Saddle regime** ($\gamma > 1$):

1. The parameters start in the vicinity of a saddle and visit a sequence of saddles during training. They come closer to each of these saddles as the width grows.
2. As the width grows, it takes longer to escape each saddle, leading to long plateaus for the training error. The training time is therefore asymptotically infinite (see Figure 3.2c).
3. The rate of change $\|\Theta(\theta_T) - \Theta(\theta_0)\|$ (where $T \in \mathbb{R}$ is the stopping time) of the NTK is infinitely larger than the NTK at initialization $\|\Theta(\theta_0)\|$. This follows from the fact that the NTK at initialization goes to zero, while it has finite size at the end of training.
4. The learned matrix is the result of a greedy algorithm that finds a low rank solution.

The **Mean-Field regime** $\gamma = 1$ lies at the transition between the two previous regimes and is more difficult to characterize:

1. In this critical regime, the constant factor c in the variance at initialization $\sigma^2 = cw^{-\gamma}$ can have a strong effect on the dynamics.
2. Plateaus can still be observed (see Figure 3.2b), however in contrast to the Saddle-to-Saddle regime, the length of the plateaus does not increase as the width grows, but remains roughly constant.
3. The NTK and its rate of change are of same order.

In general, we observe some tradeoff: the NTK regime leads to fast convergence without low-rank bias, while the Saddle-to-Saddle regime leads to some low-rank bias, but at the cost of an asymptotically infinite training time.

3.7 Conclusion

We propose a simple criterion to identify three regimes in the training of large DLNs: the distances from the initialization to the nearest global minimum and to the nearest saddle. The NTK regime ($1 - \frac{1}{L} \leq \gamma < 1$) is characterized by an initialization which is close to a global minimum and far from any saddle, the Saddle-to-Saddle regime ($\gamma > 1$) is characterized by an initialization which is close to a saddle and (comparatively) far from any global minimum and, finally, in the critical Mean-Field regime ($\gamma = 1$), these two distances are of the same order as the width grows.

While the NTK and Mean-Field limits are well-studied, the Saddle-to-Saddle regime is less understood. We therefore investigate the case $\gamma = +\infty$ (i.e. we fix the width and let the variance at initialization go to zero). In this limit, the initialization converges towards the saddle at

the origin $\vartheta^0 = 0$. We show that gradient flow naturally escapes this saddle along an 'optimal escape path' along which the network behaves as a width-1 network. This leads the gradient flow to subsequently visit a second saddle ϑ^1 which has the property that the matrix A_{ϑ^1} has rank 1. We conjecture that the gradient flow next visits a sequence of critical points $\vartheta^2, \ldots, \vartheta^K$ of increasing rank, implementing some form of greedy low-rank algorithm. These saddles explain the plateaus in the loss curve which are characteristic of the Saddle-to-Saddle regime.

Finite-Width Neural Networks Part II

4 The Loss Landscape of (Non-Linear) Neural Networks

In this chapter, we present a novel and comprehensive study of the loss landscape of neural networks from a symmetry point of view. This chapter uses some material from Şimşek, Ged, et al., 2021 but it largely is based on unpublished material. Typically, statistical physics and probability tools are used to study high-dimensional landscapes in the limit when one of the problem parameter goes to infinitity. Our approach is constructive, and it applies to any neural network of finite-width, any dataset, and any twice-differentiable cost function. We first present the main results in Section 4.1 and related works in Section 4.2. In Section 4.3, we discuss (permutation-)symmetric losses and introduce operations to grow the neural network size without changing the network function as well as the so-called symmetry-induced critical points. A symmetry-induced critical point is always degenarate in the sense that it is one of the points of a continuum (i.e. line) of critical points. In Section 4.4, we present a precise second-order characterization of the line of critical points which form exotic constellations. We introduce the notion of 'plateau saddles': a connected manifold of constant loss such that the points in its interior are local minima and each point on its boundary is a non-strict saddle which enable an escape via Langevin dynamics. Zooming out, we then study the scaling law of manifolds of critical points which give a lens to see the global structure of the loss landscape in Section 4.5. We then study the hiererchical organization of the symmetry-induced saddles forming loss-levels where each level corresponds to an optimal solution of the tiny network of width n in Section 4.6. In the loss landscape of neural networks, high-loss saddles correspong to splitting many neurons of an optimal solution a tiny network, which in turn yields many escape directions. This is consistent with the typical hiererchical organization of saddles (high-loss saddles have high index) in other high-dimensional loss landscapes. We close by conclusions in Section 4.7.

4.1 Main Results

1. We introduce the concept of an irreducible parameter in Section 4.3 and a neuron splitting which characterizes *all* symmetry-induced critical points.

2. In particular, splitting one neuron creates a line of critical points due to symmetries in the network parameterization. In Section 4.4, we study the Hessian of the critical points on the line and find exotic constellations of saddles and local minima on the line connected through non-strict saddles. Our main Theorem 4.4.1 gives the signs of the Hessian spectrum in terms of the signs of the original Hessian spectrum and the signs of a submatrix that is given by the splitted neuron and the network function. We present the proof and further discussions in Appendix C.3.

3. Next we focus on the global structure of the loss landscape. We give the scaling law of the manifolds of symmetry-induced critical points by counting partitions of neurons and permutation of the generated neurons in Section 4.5. We observe that the scaling law grows factorially that is slightly faster than the exponential growth.

4. In Section 4.6, we present the index of symmetry-induced critical points by applying Theorem 4.4.1 to *all* splitted neurons. This gives a rigorous account for the formation of the hiererchical organization of saddles: high-loss saddles have many escape directions thus allowing an escape; whereas low-loss saddles have a tiny fraction of escape directions hence posing a danger for training in practice.

4.2 Related Works

A large body of work focuses on the geometric investigation of neural network landscapes. Dauphin et al., 2014 suggested a proliferation of saddles in neural network landscapes through an analogy with high-dimensional Gaussian Processes. Other models have been proposed to understand the general structure of the loss landscape inspired by statistical physics (Geiger, Spigler, d'Ascoli, et al., 2019), and via high-dimensional wedges (Fort and Jastrzebski, 2019). These model-based empirical works focus mainly on the Hessian spectrum at the critical points.

The focus on symmetries in our work is similar to that of Fukumizu and Amari, 2000; Brea, Şimşek, et al., 2019; Fukumizu, Yamaguchi, et al., 2019 regarding the critical points coming from neuron replications. In an orthogonal direction, Kunin et al., 2020; Głuch and Urbanke, 2021 present a catalog of symmetries appearing in deep networks, which however does not include the permutation symmetry.

An orthogonal line of work studies the neuron splitting in neural networks from a small width into a large width. Fukumizu and Amari, 2000 study the splitting of a *single* neuron into two neurons at a critical point and the resulting line of critical points; Fukumizu, Yamaguchi, et al., 2019 generalizes it to the splitting into many neurons. We study the combinatorics of the problem by splitting *all* neurons into many others: the scaling law of the manifolds of symmetry-induced critical points. Y. Zhang, Z. Zhang, et al., 2021 provide some further discussions and numerics. Jacot, Ged, Şimşek, et al., 2021; Boursier, Pillaud-Vivien, and Flammarion, 2022 analyze a training regime where the gradient flow visits a sequence of

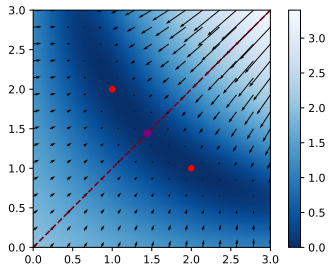


Figure 4.1 – *No gradient pointing outside of a symmetry subspace.* The gradient flow of a permutation-symmetric loss $L(w_1, w_2) = \log(\frac{1}{2}((w_1 + w_2 - 3)^2 + (w_1 w_2 - 2)^2) + 1)$. Red: permutation-symmetric global minima, purple: saddle, dashed line: the symmetry subspace.

saddles. More generally, flat saddles where the escape direction has a very low curvature form an obstacle for non-convex optimization (Pascanu et al., 2014; Y. Zhang, Qu, and Wright, 2020). We show the existence of a continuum of strict saddles of constant loss that may transit into local minima via non-strict saddles for any neural network with more than one neuron.

4.3 Setup

For $m \geq 1$, set $[m] = \{1, \ldots, m\}$ and let S_m denote the symmetric group on m symbols, i.e. the set of permutations of $[m]$. For a permutation $\pi \in S_m$ and $D \geq 1$, the map $P_\pi : \mathbb{R}^{Dm} \to \mathbb{R}^{Dm}$ permutes the units $\vartheta_i \in \mathbb{R}^D$ of a vector $\theta = (\vartheta_1, \ldots, \vartheta_m)$ according to π, i.e. $P_\pi \theta = (\vartheta_{\pi(1)}, \ldots, \vartheta_{\pi(m)})$; we sometimes use $\theta_\pi := P_\pi \theta$. With a slight abuse of notation, we will refer to permutations of affine subspaces defined as

$$P_\pi V = \{P_\pi \theta : \quad \theta \in V\}.$$

Symmetric Losses

Numerous machine learning models involve permutation-symmetric parameterizations: mixture models, matrix factorization, and neural networks. In this section, we abstract away the particular parameterization of these models and focus on the implications of permutation symmetry on the gradient flow. In particular, the discussion here is general and applies to neural networks.

Definition 4.3.1. *A loss function $L^m : \mathbb{R}^{Dm} \to \mathbb{R}$ is a* symmetric loss[1] *on m units if for any $\pi \in S_m$ and any $\theta = (\vartheta_1, \vartheta_2, \ldots, \vartheta_m)$ with $\vartheta_i \in \mathbb{R}^D$, we have*

$$L^m(\theta) = L^m(P_\pi \theta).$$

[1] When the units are 1-dimensional, symmetric losses are symmetric functions (Kung, Rota, and Yan, 2009; Sagan, 2013).

The term *unit* may refer to a Gaussian vector in the context of Gaussian mixture models, to a factor in the context of matrix factorization, or to a neuron in the context of neural networks. The symmetry subspaces are defined by the constraint that at least two units are identical:

Definition 4.3.2. *Let $i_1, \dots, i_k \in [m]$ be distinct indices. The* symmetry subspace $\mathscr{H}_{i_1,\dots,i_k}$ *is defined as*

$$\mathscr{H}_{i_1,\dots,i_k} := \{(\vartheta_1, \dots, \vartheta_m) \in \mathbb{R}^{Dm} : \vartheta_{i_1} = \dots = \vartheta_{i_k}\}.$$

As each constraint $\vartheta_i = \vartheta_j$ suppresses D degrees of freedom, we have $\dim(\mathscr{H}_{i_1,\dots,i_k}) = D(m-k+1)$. The largest symmetry subspaces are $\mathscr{H}_{i,j}$'s: any other symmetry subspace is the intersection of such subspaces. Let $\Gamma : \mathbb{R}_{\geq 0} \times \mathbb{R}^{Dm} \to \mathbb{R}^{Dm}$ denote the gradient flow

$$\partial_t \Gamma(t, \theta_0) = -\nabla L^m(\Gamma(t, \theta_0)) \tag{4.1}$$

for $t \geq 0$ and for a given initialization θ_0. The gradient on the symmetry subspace is tangent to it. In general, the gradient components of a symmetry subspace pointing to neighbor regions cancel out due to permutation symmetry

Lemma 4.3.1. *We assume that $L^m : \mathbb{R}^{Dm} \to \mathbb{R}$ is a symmetric loss on m units and a C^1 function. Let $\Gamma : \mathbb{R}_{\geq 0} \times \mathbb{R}^{Dm} \to \mathbb{R}^{Dm}$ be its gradient flow. If $\Gamma(0, \theta_0) \in \mathscr{H}_{i_1,\dots,i_k}$, the gradient flow stays inside the symmetry subspace, i.e. $\Gamma(t, \theta_0) \in \mathscr{H}_{i_1,\dots,i_k}$ for all $t > 0$. If $\Gamma(0, \theta_0) \notin \mathscr{H}_{i,j}$ for all $i \neq j \in [m]$, the gradient flow does not visit any symmetry subspace in finite time.*

Remark. *Lemma 4.3.1 does not exclude the following scenario: if there is a critical point on the symmetry subspace that is attractive in some directions orthogonal to the symmetry subspace, the gradient flow can reach it in infinite time (i.e. at convergence).*

Neural Network Losses

Let $f^{(2)} : \mathbb{R}^{d_0} \to \mathbb{R}^{d_{\text{out}}}$ be a two-layer neural network of width m

$$f_\theta^{(2)}(x) = \sum_{j=1}^{m} a_j \sigma(w_j \cdot x)$$

and $\theta = (w_1, a_1) \oplus \dots \oplus (w_m, a_m)$ is a point in the parameter space $\mathbb{R}^{Dm}$ with $w_i \in \mathbb{R}^{d_0}$, $a_i \in \mathbb{R}^{d_{\text{out}}}$, and $D = d_0 + d_{\text{out}}$.

The training dataset of size N is denoted by $\{(x_i, y_i)\}_{i=1}^N$ where $x_k \in \mathbb{R}^{d_0}$, $y_k \in \mathbb{R}^{d_{\text{out}}}$. The training loss $L^m : \mathbb{R}^{Dm} \to \mathbb{R}$ is

$$L^m(\theta) = \frac{1}{N} \sum_{i=1}^{N} c(f_\theta^{(2)}(x_i), y_i) \tag{4.2}$$

where $c : \mathbb{R}^{d_{\text{out}}} \times \mathbb{R}^{d_{\text{out}}} \to [0, +\infty)$ is a twice differentiable cost in its first component. We assume that the activation function σ is twice differentiable.

Since $f^{(2)}$ is invariant under the permutation of *neurons* $\vartheta_i := [w_i, a_i] \in \mathbb{R}^D$, the concatenation of the incoming and outgoing weight vectors, L^m is a symmetric loss (Def. 4.3.1). Therefore the symmetry subspaces $\vartheta_i = \vartheta_j$ are invariant under the gradient flow (Lemma 4.3.1). Neural network functions exhibit further invariances.

Definition 4.3.3. *We call a point* $\theta \in \mathbb{R}^{Dm}$ irreducible *no two incoming vectors are equal to each other.*

We are interested in the simplest case of irreducibility in this chapter in relation to the study of symmetry-induced critical points. Note that in general, groups of neurons can be reduced to a single neuron, which is the general form of irreducibility. We will work with the general definition in Chapter 5.

Symmetry-Induced Critical Points

Let us assume that θ is a critical point of the loss of a neural network with n neurons, i.e. $\nabla L^n(\theta) = 0$, and that it does not achive zero-loss. Let us define the neuron splitting matrix $\oplus^{\mu,j} : \mathbb{R}^{Dn} \to \mathbb{R}^{D(n+1)}$

$$\oplus^{\mu,j}\theta = (w_1, a_1) \oplus \ldots \oplus (w_j, \mu a_j) \oplus \ldots \oplus (w_n, a_n) \oplus (w_j, (1-\mu)a_j). \tag{4.3}$$

where μ is the mixing ratio and j is the index of the neuron splitted.

The seminal paper of Fukumizu and Amari, 2000 shows that $\nabla L^{(n+1)}(\oplus^{\mu,j}(\theta)) = 0$ with a simple gradient calculation; hence $\oplus^{\mu,j}$ maps a critical point to another critical point. Moreover, by varying the (arbitrary) mixing ratio μ, we get a *line of critical points* with the same loss as the original network with parameter θ. More recent paper of Fukumizu, Yamaguchi, et al., 2019 studied the splitting of a single neuron into multiple ones. We are intested in the most general form of neuron splitting: splitting many neuron types into multiple others with arbitrary mixing ratios.

We call all critical points that can be represented as $\oplus^{\mu,j}\theta$ *symmetry-induced critical points*. Note that all symmetry-induced critical points are reducible (that is, not irreducible) according to definition 4.3.3.

We first focus on splitting a single neuron and studying the second-order derivatives of the loss at the critical points on the line in Section 4.4. We then derive the scaling law of the affine subspaces of symmetry-induced critical points in Section 4.5.

4.4 Second-Order Characterization

We reviewed that a critical point of the loss of the network with n neurons turns, after neuron splitting, into a critical point of the loss of the network with $n+1$ neurons. In this section, we consider the question of whether this critical point is a saddle or a local minimum.

Classification of Critical Points and Plateau Saddles

We consider a critical point θ of the loss of a neural network with n neurons. Since both the activation function and cost are twice differentiable, the Hessian of the loss is well-defined and denoted by $HL^n(\theta)$ which is a matrix of size $Dn \times Dn$. In Section 4.3, we reviewed that splitting any of the neurons of θ creates symmetry-induced critical points denoted by $\oplus^{j,\mu}\theta$ that form a line as we vary the mixing ratio $\mu \in \mathbb{R}$. To identify whether these points are local minima, strict saddles, or non-strict saddles, we study their second-order derivatives. First, let us lay out the definitions to classify critical points with different characteristics:

1. *Strict saddle:* A critical point with at least one positive and one negative eigenvalue in the Hessian (which assures an escape direction in a neighborhood towards decreasing loss).
2. *Local minimum:* A critical point with no negative eigenvalues in the Hessian and the loss is non-decreasing in all directions of small perturbations.
3. *Non-strict saddle:* A critical point with no negative eigenvalues in the Hessian and there is an escape direction towards decreasing loss in a neighborhood.

If the Hessian does not have a negative eigenvalue at the critical point, it can be either a non-strict saddle or a local minimum; thus a higher-order analysis is needed for classification (Anandkumar and Ge, 2016). This is important because first- and second-order optimization algorithms may get stuck at non-strict saddles (Anandkumar and Ge, 2016).

More generally, due to symmetries of the neural network parameterization, symmetry-induce critical points are always degenerate in the sense that they form affine subspaces of critical points. As we will see, it is possible to have exotic structures in the neural network loss functions with transitions from local minima to strict saddles through non-strict saddles. In particular, we define the following critical manifold of interest

Definition 4.4.1 (Plateau saddle)**.** *A connected line segment of constant loss, possibly infinite on one side, such that each point in its interior is a local minimum and each point on its boundary is a non-strict saddle.*

The non-strict saddles at the boundary provide escape directions pointing outside of the local minima manifold. When initialized near a one-dimensional plateau saddle that is a line segment (Fig. 4.2-A), a deterministic algorithm such as gradient descent gets stuck at a local minimum in its interior whereas a stochastic algorithm such as Langevin dynamics escapes it eventually with probability one (Mertikopoulos et al., 2020; Kamalaruban et al., 2020). An analogous case occurs when the plateau saddle is a half-line (Fig. 4.2-B): if initialized near the plateau saddle a random walk eventually escapes it, too.

For one output neuron (i.e. $d_{\text{out}} = 1$), Fukumizu and Amari, 2000 proved that the second-order characteristics of the symmetry-induced critical points is determined by the mixing ratio μ and

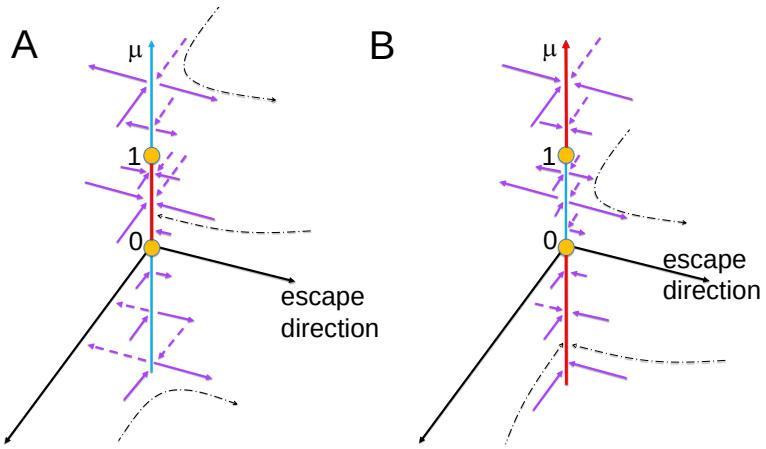


Figure 4.2 – *Gradient flow in the neighborhood of plateau saddles (red ∪ orange) in the parameter space of a neural network.* Strict saddles (blue) transit into local minima (red) through non-strict saddles (orange). Plateau saddles induced by neuron splitting may arrive *(A)* either on the line segment $[0,1]$ *(B)* or on two half-lines of the mixing coefficient μ (vertical axis). The loss along the vertical axis is constant. Only two other directions are shown, one of them being the direction of escape. At the local minima (red), the loss increases in all other directions since the Hessian has, apart from the zero eigenvalue along the vertical line, only positive eigenvalues so that the flow is towards the vertical line. Sample trajectories are shown in dot-dashed black lines; flow arrows in magenta.

a second-order derivative matrix Y (for its definition see Theorem 4.4.1 below). They show that depending on the eigenvalue signs of the matrix $\mu(1-\mu)Y$, a symmetry-induced critical point $\oplus^{j,\mu}\theta$ is either a local minimum or a strict saddle for $\mu \notin \{0,1\}$, whereas the symmetry-induced critical points for $\mu \in \{0,1\}$ remain unclassified.

The Hessian Spectrum

Using a novel decomposition of the Hessian, we generalize the result of Fukumizu and Amari, 2000 in two ways:

i. For $d_{\text{out}} = 1$ and $\mu \in \{0,1\}$, we show that the symmetry-induced critical points have at least $d+1$ zero eigenvalues of the Hessian for any distribution of the eigenvalues of the matrix Y.

ii. For arbitrary d_{out}, we explicitly give the number of positive, negative, and zero eigenvalues of the Hessian at a symmetry-induced critical point which depends on two matrices of second-order derivatives, namely Y and V (see Eq. 4.5). This generalization to arbitrary d_{out} is the relevant case for deep networks.

Theorem 4.4.1. *Let $\theta \in \mathbb{R}^{Dn}$ be a critical point of the loss of a neural network with n neurons. Let $\oplus^{j,\mu}\theta$ be a symmetry-induced critical point of the network with $n+1$ neurons. The spectrum of $HL(\oplus^{j,\mu}\theta)$ is composed of two parts: (i) the bulk of Dn eigenvalues has the same signs as the*

eigenvalues of $HL(\theta)$, (ii) the remaining D eigenvalues have the same signs as the eigenvalues of the following matrix

$$\begin{bmatrix} \mu(1-\mu)Y(w_j,a_j) & -V(w_j,a_j) \\ -V(w_j,a_j)^T & 0 \end{bmatrix} \tag{4.4}$$

where

$$\begin{aligned} Y(w_j,a_j) &= \frac{1}{N}\sum_{i=1}^{N} \sigma''(w_j\cdot x_i)x_i x_i^T a_j \cdot c'(f_\theta^{(2)}(x_i),y_i) \in \mathbb{R}^{d\times d}, \\ V(w_j,a_j)_{k\ell} &= \frac{1}{N}\sum_{i=1}^{N} \sigma'(w_j\cdot x_i)(x_i)_k c'(f_\theta^{(2)}(x_i),y_i)_\ell \text{ with } k\in[d], \ell\in[d_{out}]. \end{aligned} \tag{4.5}$$

See Appendix Section C.3 for the proof sketch and the complete proof.

Remark (Deep Networks)**.** *Theorem 4.4.1 can be generalized to deep networks by considering neuron splitting within one of the hidden layers. The submatrices of Y and V need to be updated by (i) replacing x with the post-activation vector coming from the previous hidden layer and (ii) the derivative of the cost needs to be calculated w.r.t the pre-activation vector of the next hidden layer. We discuss the case of multiple output neurons in detail in Appendix Section C.3.1.*

We need to unpack the submatrix in Eq. 4.4 to understand the second-order characteristics of the symmetry-induced critical points. First, note that

$$V(w_j,a_j)a_j = \frac{\partial L}{\partial w_j}(\theta) = 0 \in \mathbb{R}^d. \tag{4.6}$$

If we have a single output node ($d_{\text{out}} = 1$) and a_j is non-zero, the submatrix in Eq. 4.4 reduces to

$$\begin{bmatrix} \mu(1-\mu)Y(w_j,a_j) & 0 \\ 0 & 0 \end{bmatrix} \in \mathbb{R}^{(d+1)\times(d+1)}. \tag{4.7}$$

A scenario of interest is when the parameter θ is the optimal solution of an n-neuron network and its Hessian is positive definite. What are the characteristics of the critical points on the line after splitting one of the neurons of θ? Note that the two symmetry-induced critical points $\oplus^{j,\mu}\theta$ and $\oplus^{j,1-\mu}\theta$ have the same set of neurons hence it is enough to discuss the half-line $\mu \in [0.5,\infty)$. There are three main scenarios[2]

i. *Positive definite Y:* strict saddles on $\mu > 1$ transit into local minima on $\mu \in [0.5,1)$ via a non-strict saddle at $\mu = 1$. Because of the mirror symmetry around 0.5, the line segment $\mu \in [0,1]$ is a plateau saddle.

[2] An exception to the three scenarios above is the case of vanishing $Y = 0$ which is the case for the linear activation function.

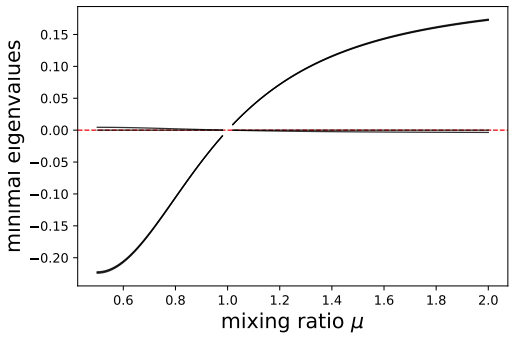


Figure 4.3 – *The five smallest eigenvalues of the Hessian on the line of critical points* as a function of the mixing ratio μ. The line corresponds to the splitting of the optimal solution of the one-neuron network learning from an orthogonal teacher network with $k = 4$ neurons into two neurons (input dimension $d = 4$). On both sides of the line, we have strict saddles. The three eigenvalue curves are very close to each other, hence they virtually seem to be overlapping.

ii. *Negative definite Y :* local minima on $\mu > 1$ transit into strict saddles on $\mu \in [0.5, 1)$ via a non-strict saddle at $\mu = 1$. The half-line $\mu \in [1, \infty)$ is a plateau saddle.

iii. *Y has at least one positive and one negative eigenvalue:* strict saddles on $\mu > 1$ transit into other strict saddles on $\mu \in [0.5, 1)$ via a non-strict saddle at $\mu = 1$.

Although the Hessian spectrum does not suffice to classify the critical point at $\mu = 1$, we conclude that it is a non-strict saddle with a one or two-sided escape route(s) towards lower loss, since there is a strict saddle in its neighborhood on at least one side. Thus, the loss of neural networks with more than one neuron violates the commonly studied 'strict' saddle property (Ge et al., 2015; Sun, Qu, and Wright, 2015; Jin et al., 2017b) due to the existence of symmetry-induced 'non-strict' saddles.

Remark (Application to Two-Neuron Network)**.** *Because the optimal solution of a one-neuron network induces symmetry-induced critical points of a two-neuron network, we can apply results from the one-neuron case to characterize some properties of the loss of the two-neuron network. In particular, the two-neuron case is prone to containing plateau saddles for some mixing ratio, either for $\mu \in (0, 1)$ or for $\mu \in \mathbb{R}/[0, 1]$. If there are no plateau saddles (see Fig. 4.3), then the line of critical points in the two-neuron case (that is induced by the solution of the one-neuron case) consists of strict saddles except the non-strict saddles at $\mu \in \{0, 1\}$.*

The Minimal Hessian Eigenvalue as a Function of the Mixing Ratio

In this section, we dig deeper into the second-order analysis of the symmetry-induced critical points for one output neuron by studying the smallest non-trivial eigenvalue of the Hessian denoted by $\lambda^\dagger$. At a strict saddle, $\lambda^\dagger$ is the smallest eigenvalue ($\lambda^\dagger < 0$); otherwise, it denotes the second smallest eigenvalue ($\lambda^\dagger \geq 0$) excluding the trivial zero. This is relevant because

(i) at a strict saddle, the magnitude of the most negative eigenvalue gives the escape speed (Jin et al., 2017b; J. D. Lee, Panageas, et al., 2019b) (ii) at a local minimum, the smallest non-trivial eigenvalue gives a measure of flatness. Since the Hessian has mixed-sign eigenvalues, techniques developed for positive definite matrices are not directly applicable here. We instead manually plug in unit vectors to obtain an upper bound on the Rayleigh coefficient (see Appendix C.3.2)

Lemma 4.4.2. *The smallest non-trivial eigenvalue of the Hessian of a symmetry-induced critical point can be bounded as follows where* $u(\mu) = \frac{\mu(1-\mu)}{(1-\mu)^2+\mu^2}$

- $\lambda^\dagger(HL(\oplus^{j,\mu}\theta)) \leq u(\mu)\lambda_{\min}(Y) \quad$ *for* $\mu \in (0,1)$,
- $\lambda^\dagger(HL(\oplus^{j,\mu}\theta)) \leq u(\mu)\lambda_{\max}(Y) \quad$ *for* $\mu \in \mathbb{R}/[0,1]$,

As a sanity check, let us consider Y positive definite: in this case $\lambda^\dagger$ is positive for $\mu \in (0,1)$ and so is the upper bound. Similarly, for Y negative definite, $\lambda^\dagger$ is negative for $\mu \in (0,1)$ and so is the upper bound. Finally, if Y has at least one negative and one positive eigenvalue, $\lambda^\dagger$ is negative and so is the upper bound in both cases. We also provide a lower bound in Appendix C.3.2.

If Y is negative definite, the symmetry-induced critical points on $\mu \in \mathbb{R}/[0,1]$ are local minima. The upper bound $u(\mu)$ crosses zero at $\mu = 1$ and it decreases in the interval $\mu \in [1,\infty)$ (same for $\mu \in (-\infty, 0]$ due to mirror symmetry). In the limit, we have $\lim_{\mu\to\infty} u(\mu) = -0.5$, hence

$$-\frac{1}{2}\lambda_{\max}(Y) \geq \lim_{\mu\to\infty} \lambda^\dagger(HL(\oplus^{j,\mu}\theta)). \tag{4.8}$$

This suggests that the loss increases near (but orthogonal to) the half-line $[1,\infty)$ of local minima; it gets steeper as μ grows larger at least initially at $\mu = 1$ (see Figure C.3 in the Appendix).

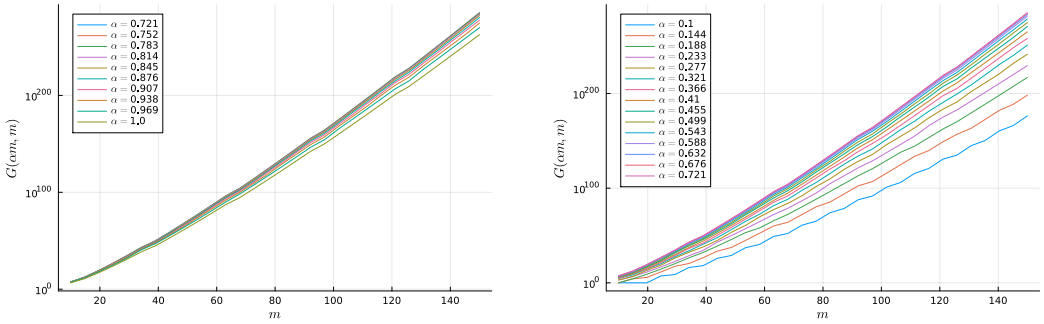
4.5 Scaling Law of the Critical Manifolds

In this section, we consider all possible partitions of neuron splittings of a parameter from a neural network of width n into a neural network of width m. Let us first give a formal definition of an affine subspace that is generated by splitting the neurons of a parameter θ.

Definition 4.5.1. *For* $n < m$*, let us pick a partition of* m *neurons such that* $s_1 + ... + s_n = m$ *and* $s_i \geq 1$*. The affine subspace of parameters* $V_s(\theta)$ *corresponding to splitting the neurons of a parameter* $\theta \in \mathbb{R}^{Dn}$ *with the partition* $s = (s_1, ..., s_n)$ *is defined recursively as*

$$V^\ell_{s_\ell} = \{\theta_\ell : \theta_\ell = \oplus^{\mu_{s_\ell-1},\ell} ... \oplus^{\mu_1,\ell} \theta_{\ell-1} \quad \text{for all } \mu_1, ..., \mu_{s_\ell-1} \in \mathbb{R},\ \theta_{\ell-1} \in V^{\ell-1}_{s_{\ell-1}}\} \quad \text{if } s_\ell \geq 2,$$
$$V^\ell_{s_\ell} = V^{\ell-1}_{s_\ell-1} \quad \text{if } s_\ell = 1;$$

for $\ell = 1, ..., n$ *where the initial subspace is* $V^0_{s_0} = \{\theta\}$ *and* $V_s(\theta) := V^n_{s_n}$.



(a) for $\alpha \in \left(\frac{1}{2log2}, 1\right)$ (b) for $\alpha \in \left(0, \frac{1}{2log2}\right)$

Figure 4.4 – *Scaling law of the manifolds of symmetry-induced critical points.* In the left panel, we plot the number $G(\alpha m, m)$ as a function of the network width m for $\alpha \in (\alpha^*, 1)$; we observe that $\alpha^* = \frac{1}{2log2}$ is the maximum for fixed m. In the right, we plot the same number for $\alpha \in (0, \alpha^*)$; we observe that again α^* is the maximum. Overall, visually, the scaling law is slightly faster than exponential as the curves seem to have positive curvature as opposed to straight line. In fact the curves follow the same trend as $\alpha = 1$ which we know is the usual factorial $G(m, m) = m!$. The jitter for small α is due to finite-size effect.

Let us consider a critical point θ of the loss of the neural network of width n, for example an optimal solution. If θ is reducible, we merge its neurons by summing the outgoing weights until we find an irreducible critical point θ which has distinct incoming vectors.

The symmetry-induced critical points of the loss of the neural network with width m that are generated from the irreducible critical point θ are collected in the union of permutations of affine subspaces of all possible partitions from n to m

$$S_{n\to m}(\theta) = \bigcup_{\substack{\pi\in S_m,\ s=(s_1,\dots,s_n)\\ s_1+\dots+s_n=m,\ s_i\geq 1}} P_\pi V_s(\theta).$$

We note that no two affine subspace written above interect each other since there is always a mismatch in the position of the incoming weight vectors. The scaling law of the manifolds of critical points is given by

Proposition 4.5.1. *Let $\theta \in \mathbb{R}^{Dn}$ be an irreducible critical point. For $n \leq m$, the number of distinct affine subspaces in $S_{n\to m}(\theta)$ is given by*

$$G(n, m) = \sum_{\ell=1}^{n} \binom{n}{\ell} (-1)^{n-\ell} \ell^m.$$

We first investigate the scaling law in simple limits when either n is fixed and $m \to \infty$, the number of manifolds induced by tiny networks of width n onto infinite width networks; or the number of new neurons h is fixed and $m \to \infty$, the number of manifolds induced by very large

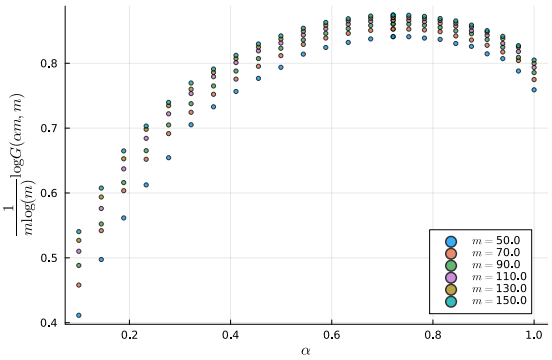


Figure 4.5 – *The scaling law as a function of the shrinkage factor α in the large m limit.* We observe that the curves approach to a limit when the scaling law $G(\alpha m, m)$ is normalized with the leading term m^m of the factorial according to Stirling's approximation. We observe that $\log G(\alpha m, m)/(m \log m)$ converges to a unimodal curve $c(\alpha)$ as a function of α.

networks onto slightly wider networks.

Lemma 4.5.2. *For any $h \geq 0$ fixed, we have,*

$$G(m-h, m) \sim \frac{m^h}{2^h h!} m! \quad \text{as } m \to \infty.$$

For any fixed $n \geq 0$, we have $G(n, m) \sim n^m$ as $m \to \infty$.

The scaling law G can in fact be mapped onto an enumarative combinatorics problem that can be phrased as: "How many surjective functions are from a set of n items onto another set of $m \geq n$ items?". Every element of the domain can be mapped onto several items of the codomain in parallel to the splitting of every neuron onto several neurons. The natural scaling of the number of surjective functions, that is $G(n, m)$ is linear, i.e. $m/n = \alpha$ fixed and $m, n \to \infty$. Moreover, the maximum of $G(\alpha m, m)$ is attained at $\alpha = \frac{1}{2log2}$ as $m \to \infty$ (details in the mathoverflow discussion). We plot the scaling laws $G(\alpha m, m)$ as a function of m for $\alpha \in (0, 1)$ values in Figure 4.4.

Based on the apperance of the curves in Figure 4.4, we make a guess that in the limit with appropriate normalization, we have

$$\lim_{m \to \infty} \frac{1}{m \log m} \log G(\alpha m, m) \to c(\alpha). \tag{4.9}$$

Based on this case, we plotted the scaling law $G(\alpha m, m)$ as a function of α. We observe that as m grows, the curves seem to converge to a limit suggesting that this is an appropriate scaling. Unfortunately, we are not able to plot for m values larger than 150 as the numbers grow factorially large (see fig. 4.5). This is to be constrasted with other high-dimensional loss

landscapes such as spherical spin glasses (Auffinger, Arous, and Černy, 2013)

$$\lim_{m\to\infty}\frac{1}{m}\log N(\text{given energy } \epsilon, \text{index idx}) \to c(\epsilon, \text{idx}), \tag{4.10}$$

where the landscape complexity grows exponentially.

4.6 Hierarchical Organization of Saddles

In this section, we explore the hierarchy between symmetry-induced critical points in the loss landscape of a neural network of width m. The *first-level* saddles refer to symmetry-induced critical points that are equivalent to a minimum of a network of width $m-1$; more generally, *k-th level* saddles refer to those equivalent to a minimum of a network of width $m-k$. Adding neurons enables the network to reach a lower loss minimum thus higher-level symmetry-induced saddles usually attain higher losses. We notice a similarity with Gaussian Process (Bray and Dean, 2007) and spherical spin glass (Auffinger, Arous, and Černy, 2013) landscapes, where the higher-index saddle points typically attain higher losses. Index is the ratio of the number of negative eigenvalues to the total number of eigenvalues in the Hessian.

We developed in Section 4.4 that the Hessian spectrum of a symmetry-induced critical point is given by in the case of $d_{\text{out}} = 1$ (written informally)

$$\sigma_{\text{sgn}}(HL(\oplus^{j,\mu}\theta)) = \sigma_{\text{sgn}}(HL(\theta)) \oplus \sigma_{\text{sgn}}(\mu(1-\mu))\sigma_{\text{sgn}}(Y(w_j, a_j)). \tag{4.11}$$

where $\sigma_{\text{sgn}}(x) = 1$ if $x > 0; 0$ if $x = 0; -1$ if $x < 0$. What happens if we split another neuron? Applying the above rule once again, we get

$$\begin{aligned}\sigma_{\text{sgn}}(HL(\oplus^{i,\gamma}\oplus^{j,\mu}\theta)) = \sigma_{\text{sgn}}(HL(\theta)) &\oplus \sigma_{\text{sgn}}(\mu(1-\mu))\sigma_{\text{sgn}}(Y(w_j, a_j)) \\ &\oplus \sigma_{\text{sgn}}(\gamma(1-\gamma))\sigma_{\text{sgn}}(Y(w_i, a_i)).\end{aligned} \tag{4.12}$$

Care should be taken if we split the same neuron twice, say in the following order

$$(w_j, a_j) \to (w_j, \mu a_j) \oplus (w_j, (1-\mu)a_j) \to (w_j, \gamma\mu a_j) \oplus (w_j, (1-\mu)a_j) \oplus (w_j, (1-\gamma)\mu a_j);$$

then we obtain in this case the following sign spectrum

$$\begin{aligned}\sigma_{\text{sgn}}(HL(\oplus^{j,\gamma}\oplus^{j,\mu}\theta)) = \sigma_{\text{sgn}}(HL(\theta)) &\oplus \sigma_{\text{sgn}}(\mu(1-\mu))\sigma_{\text{sgn}}(Y(w_j, a_j)) \\ &\oplus \sigma_{\text{sgn}}(\gamma(1-\gamma)\mu)\sigma_{\text{sgn}}(Y(w_j, a_j))\end{aligned} \tag{4.13}$$

where we used that $Y(w_j, \mu a_j) = \mu Y(w_j, a_j)$. As a sanity check, let us check the symmetry of the final formula. Let us set $\gamma \leftarrow (1-\gamma)/\mu$ for $\mu \neq 0$. γ and μ are then permutation-symmetric which is reflected in the final formula.

In general, we observe constellations of different-index saddles connected together into affine

subspaces. This is a generalization of plateau saddle discussed in Section 4.4. Globally, adding neurons add at maximum d new negative eigenvalues at the cost of having d new positive eigenvalues on the complementary part of the line. Let us assume that a typical neuron-splitting matrix $Y(w_j, a_j)$ has $d/2$ negative and $d/2$ positive eigenvalues. In this case adding h neurons gives the index

$$\frac{(d/2)h}{(d/2)h+D} \sim \frac{h}{h+2n}. \tag{4.14}$$

where n is the initial number of neurons.

4.7 Conclusion

We studied the atypical structure of critical points of the neural network loss landscapes: symmetry-induced critical points are not isolated but form manifolds. For example, in the landscape of a neural network with more than one neuron, there exists a line of symmetry-induced critical points induced by the optimal solution of the one-neuron network. Therefore, the celebrated Kac-Rice formula describing the scaling of the number of isolated critical points of complex landscapes (Auffinger, Arous, and Černy, 2013; Ros et al., 2019; Maillard, Arous, and Biroli, 2020) does not directly apply to neural network landscapes with more than one neuron. Moreover, we described the characteristics of critical points on the line, in particular the plateau saddle that is composed of local minima enclosed by non-strict saddle(s). An under-parameterized network of n neurons contains at least $2(n-1)(n-1)!$ non-strict saddles due to permutation symmetry. Therefore, the loss functions of neural networks with more than one neuron ($n \geq 2$) violate the strict saddle property (Ge et al., 2015; Sun, Qu, and Wright, 2015; Jin et al., 2017b) due to the existence of the (many) non-strict saddles.

We studied the scaling law in detail and showed that the number of critical manifolds grows factorially. In particular, in the loss landscape of a neural network of width m, the most dominant manifold of saddles is the one originated in the neural network of width $\frac{m}{2\log 2}$. However, these saddles typically have a non-negligible fraction of escape directions, so it remains an open question whether they pose danger to training dynamics or not. More generally, the scaling law of the saddles that originate in a neural network of width αm for $\alpha \in \left(\frac{1}{2\log 2}, 1\right)$ dominate the ground-level configurations since we have $G(\alpha m, m) \gg G(m, m) = m!$. Overall, there is a trade-off between the scaling law of saddles and their index: as α increases in the interval $\left(\frac{1}{2\log 2}, 1\right)$, the number of manifolds decrease (making them less attractive); at the cost of lowering the index of the saddles (making them harder to escape).

5 Overparameterized Networks

In this chapter, we re-present our ICML paper (Şimşek, Ged, et al., 2021) with a new notation and numerics from our preprint (Martinelli et al., 2023) that are intimately linked to the loss landscape and training dynamics of overparameterized networks. We present the main results in Section 5.1, related works in Section 5.2, and problem setup in Section 5.3. We present the main theorem on the geometry and topology of the zero-loss solutions in overparameterized networks in Section 5.4. The proof is presented in Appendix D.1.

Leveraging the scaling laws of the manifolds of symmetry-induced critical points and manifolds of zero-loss points, we propose a landscape complexity measure for overparameterized networks in Subsection 5.5.1. The complexity gives us lenses to discuss difficulty of training the overparameterized networks of finite-width and a quantitative discussion between mild vs. vast overparameterization regimes in Section 5.5. Our predictions are supported by numerical experiments in Section 5.5.2. We discuss generalizations to deep neural networks in Section 5.6 and close by conclusion and future directions in Section 5.7.

5.1 Main Results

1. Suppose an L-layer Artificial Neural Network (ANN) with hidden layer widths $k_1, ..., k_{L-1}$ reaches a unique (up to permutation) zero-loss global minimum (we call such a network *minimal* if it cannot achieve zero loss if any neuron is removed). The permutation symmetries give rise to $k_1!...k_{L-1}!$ equivalent discrete global minima. We show that adding one neuron to each layer is sufficient to connect these global minima into a single zero-loss manifold.

2. For a two-layer overparameterized network of width $m = k+h$, we describe the geometry of the global minima manifold precisely: it consists of a union of a number $T(k,m)$ of affine subspaces of dimension $\geq h$ and it is piecewise linearly connected. Furthermore, we show that the global minima manifold contains *all* zero-loss parameters for a broad class of activation functions and with input distribution of full support.

3. We propose a landscape complexity measure coming from the scaling laws of manifolds of symmetry-induced critical points and zero-loss points. We find that there is a cross-over in between the onset of overparameterization until a factor of 1.2 to 1.6 times more neurons, where the landscape complexity decreases rapidly from large to low values; indicating that the complexity of training decreases rapidly at the onset of overparameterization.

    - When the number of additional neurons satisfies $h \ll k$ (i.e. at the beginning of the overparameterized regime), the scaling law of the global minima manifold is much *smaller* than the scaling law of the low-index symmetry-induced critical points. In this sense, there is a proliferation of saddles and the global minima manifold is 'tiny'. The landscape complexity decays exponentially in this regime therefore the large-complexity regime is avoided after a factor $m/k \in [1.5, 2]$ of overparameterization.
    - Conversely, when $h \gg k$ (i.e. we are far into or within the overparameterized regime), the landscape complexity goes down to zero; which shows that the loss landscape simplies significantly in the regime when the training provably converges to zero-loss.

5.2 Related Works

A number of recent works have explored the typical path taken by a gradient-based optimizer. For very wide neural networks, the gradient flow converges to a global minimum in spite of the non-convexity of the loss (Jacot, Gabriel, and Hongler, 2018b; S. S. Du, Zhai, et al., 2018; Chizat and Bach, 2018c; Arora, S. Du, et al., 2019; S. Du, J. Lee, et al., 2019; J. Lee, Xiao, et al., 2019; J. Lee, Schoenholz, et al., 2020). First-order gradient algorithms provably escape strict saddles (Jin et al., 2017b; J. D. Lee, Panageas, et al., 2019b), although they can face an exponential slowdown around these saddles (S. S. Du, Jin, J. D. Lee, M. I. Jordan, A. Singh, et al., 2017). For pruned neural networks, the training with typical (random) initialization does not reach any global minimum, in spite of their presence in the landscape (Frankle and Carbin, 2018).

Another line of work suggests that global minima found by stochastic gradient descent are connected (i.e. there is a path linking arbitrary two minima along which the loss increases only negligibly) via simply parameterized low-loss curves (Draxler et al., 2018; Garipov et al., 2018) or line segments (Sagun, Evci, et al., 2017; Frankle, Dziugaite, et al., 2020; Fort, Dziugaite, et al., 2020). Theoretical work limited to ReLU-type activation functions, showed that in overparameterized networks, all global minima lie in a connected manifold (Freeman and Bruna, 2016; Q. Nguyen, 2019), however without giving a geometrical description of this manifold. Cooper, 2020 studied the geometry of a subset of the manifolds of critical points. Kuditipudi et al., 2019 showed that the global minima for ReLU networks, for which *half* of the neurons can be dropped without incurring a significant increase in loss, are connected via piecewise linear paths of minimal cost.

5.3 Setup

Let $f^{(2)} : \mathbb{R}^{d_0} \to \mathbb{R}^{d_{\text{out}}}$ be a two-layer neural network of width m

$$f_\theta^{(2)}(x) = \sum_{j=1}^{m} a_j \sigma(w_j \cdot x)$$

and $\theta = (w_1, a_1) \oplus ... \oplus (w_m, a_m) \in \mathbb{R}^{Dm}$ is a parameter with $w_i \in \mathbb{R}^{d_0}$, $a_i \in \mathbb{R}^{d_{\text{out}}}$, and $D = d_0 + d_{\text{out}}$. We are interested in describing the equal-loss parameter manifolds in the loss landscape coming from network parameterization. Let us introduce the definition of irreducible parameter which will play a critical role in our analysis.

Definition 5.3.1. *We call a parameter* $\theta \in \mathbb{R}^{Dm}$ irreducible *if any of its* $k \geq 2$ *neurons, say* $i_1, ..., i_k$ *(distinct indices) cannot be merged into a single neuron, that is for all* $x \in \mathbb{R}^{d_0}$

$$\sum_{\ell=1}^{k} a_{i_\ell} \sigma(w_{i_\ell} \cdot x) = a^* \sigma(w^* \cdot x).$$

We assume that there is a neural network function f^* of finite width k and a parameter $\theta^* \in \mathbb{R}^{Dk}$ that generates the targets given the input points, i.e. $f^*(x_i) = y_i$ for $i = 1, ..., N$. It is also known as the teacher network. Importantly, we assume that k is minimal in the sense that there is no narrower network that can generate the same function $f^* = f_{\theta^*}^{(2)}$. If θ^* is irreducible, k is minimal by definition[1].

We call wider networks with width $m > k$ **overparameterized** and the narrower networks with width $n < k$ **underparameterized**.

Conversely, a parameter $\theta \in \mathbb{R}^{Dm}$ is trivially *reducible* if it has two neurons that share the same incoming vector or if it has a zero neuron since then we can merge two neurons into one as follows

1. $(w_1, a_1) \oplus (w_1, a_2) =_f (w_1, a_1 + a_2)$ or
2. $(w_1, a_1) \oplus (w_1, 0) =_f (w_1, a_1)$

where $\theta_1 =_f \theta_2$ denotes the functional equivalence between parameters of possibly different dimensions, that is $f_{\theta_1}^{(2)}(x) = f_{\theta_2}^{(2)}(x)$ for all $x \in \mathbb{R}^{d_0}$. We can continue dropping neurons as above until we find an irreducible point θ_0 that is functionally equivalent to θ. Equivalently (going in the opposite direction), an irreducible parameter

$$\theta_0 = (w_1, a_1) \oplus ... \oplus (w_n, a_n)$$

[1] We note that there may be other scenarios due to the symmetries in the dataset or simply finite-size effects for which θ^* is irreducible but the network function $f_{\theta^*}^{(2)}$ is not of minimal width (i.e. that there is a narrower network that interpolates the training dataset).

yields an affine subspace of equal loss points in a network with width $m \geq n$. Let us introduce the two essential operations of neuron addition that preserves the network function

1. *Neuron splitting:* $A^j_{a'}(\theta) = (w_1, a_1) \oplus ... \oplus (w_j, a_j - a') \oplus ... \oplus (w_n, a_n) \oplus (w_j, a')$,

2. *Zero neuron addition:* $A^0_{w'}(\theta) = (w_1, a_1) \oplus ... \oplus (w_n, a_n) \oplus (w', 0)$.

Definition 5.3.2. *For $n \geq 1$, $j \geq 0$ with $n+j \leq m$, let $s = (s_1, ..., s_n)$ and $z = (z_1, ..., z_j)$ be tuples of integers with $s_i, z_i \geq 1$. Let us define their union $\bar{s} = (s_1, ..., s_n, s_{n+1} = z_1, ..., s_{n+j} = z_j)$ which satisfies sum$(\bar{s}) := s_1 + ... + s_{n+j} = m$. The affine subspace $V_{s,z}(\theta)$ of parameters that are functionally equivalent to the point $\theta \in \mathbb{R}^{Dn}$ is defined recursively as*

$$V^0_{s,z} = \{\theta_0 : \quad \theta_0 = A^0_{w'_j} \circ ... \circ A^0_{w'_1}(\theta^n), \quad w'_i \in \mathbb{R}^d \text{ for } i \in [j]\}, \tag{5.1}$$

$$V^k_{s,z} = \{\theta_k : \quad \theta_k = A^k_{a'_{s_k}} \circ ... \circ A^k_{a'_1}(\theta^{k-1}), \quad a'_i \in \mathbb{R}^{d_{out}} \text{ for } i \in [s_k], \quad \theta_{k-1} \in V^{k-1}_{s,z}\}, \tag{5.2}$$

for all $k = 1, ..., n+j$, and finally $V_{s,z}(\theta) := V^{n+j}_{s,z}$.

Neurons that share an incoming weight vector w' for which the corresponding outgoing weight vectors add up to zero are called **'zero-type'** neurons. Moreover, the network function remains invariant under any permutation of neurons. Each permutation defines another affine subspace

$$P_\pi V_{s,z}(\theta) := \{P_\pi \theta : \quad \theta \in V_{s,z}(\theta) \quad \text{and} \quad \pi \in S_m\}$$

where P_π permutes the neurons $\vartheta_i = (w_i, a_i)$ of θ. We call the union of such affine subspaces (corresponding to different partitions (s, z) of neurons and their permutations) the expansion manifold of θ:

Definition 5.3.3. *For $n \leq m$, the* expansion manifold $\Theta_{n\to m}(\theta) \subset \mathbb{R}^{Dm}$ *of a parameter $\theta \in \mathbb{R}^{Dn}$ is defined as*

$$\Theta_{n\to m}(\theta) := \bigcup_{\substack{(s,z) \\ \pi \in S_m}} P_\pi V_{s,z}(\theta),$$

where $s = (s_1, ..., s_n)$ and $z = (z_1, ..., z_j)$ ($j \geq 0$) are two tuples with $s_i, z_i \geq 1$ that satisfy $s_1 + ... + s_n + z_1 + ... + z_j = m$.

Using zero neuron addition, neuron splitting, and permutation, we constructured a union manifolds (i.e. affine subspaces) that produce the same network function and denote it by $\Theta_{n\to m}$. The curious question is whether this construction represents all parameters that represent the same function. In the next section we give a positive answer for a broad class of activation functions.

5.4 Geometry and Topology of Zero-Loss Solutions

In this section, we first show that for a broad class of activation functions, a network function can be represented by a larger network only if the set of neurons in the larger net correspond to zero neurons groups and splitted neurons of the original network. We will then apply this theorem to the loss landscape of overparameterized neural networks which gives an exact description of the geometry and topology of the zero-loss solutions.

Assumption 5.4.1. *The activation function $\sigma : \mathbb{R} \to \mathbb{R}$ is C^∞, $\sigma(0) \neq 0$, and $\sigma^{(\ell)}(0) \neq 0$ for infinitely many even and odd values of ℓ (where $\sigma^{(\ell)}$ denotes the ℓ-th derivative of σ).*

In general, there are additional symmetries such as the mirror symmetry of tanh

$$(w_1, a_1) =_f (-w_1, -a_1) \tag{5.3}$$

which is an odd activation function that we can address with our construction methodology. For the ReLU activation function, there is at least the positive scaling symmetry

$$(\alpha w_1, \frac{1}{\alpha} a_1) =_f (w_1, a_1) \quad \text{for} \quad \alpha > 0, \tag{5.4}$$

but we cannot directly address this case since it is not differentiable at zero.

Theorem 5.4.2. *Let us assume that the activation function satisfies Assumption 5.4.1 and $m \geq k$. If a parameter $\theta \in \mathbb{R}^{Dm}$ produces the same function as the true parameter $\theta_* \in \mathbb{R}^{Dk}$, i.e. $\theta =_f \theta_*$, then $\theta \in \Theta_{n\to m}(\theta^*)$.*

Now it is time to introduce a loss function so that we can apply Theorem 5.4.2 to the zero-loss configurations. The single sample loss $c : \mathbb{R}^{d_{\text{out}}} \times \mathbb{R}^{d_{\text{out}}} \to \mathbb{R}_{\geq 0}$ satisfies $c(\hat{y}, y) = 0$ if and only if $\hat{y} = y$, for example for the least-squares or the logistic loss. We consider the loss function $L^m : \mathbb{R}^{Dm} \to \mathbb{R}$

$$L^m(\theta) = \int_{\mathbb{R}^{d_0}} c(f_\theta^{(2)}(x), f^*(x)) \mathscr{D}(dx),$$

where $\mathscr{D}$ is an input data distribution with support $\mathbb{R}^{d_0}$. We note that $L^m(\theta) = 0$ if and only if the network function matches the true function $f_\theta^{(2)}(x) = f^*(x)$ for all $x \in \mathbb{R}^{d_0}$.

Combining the pieces together, we conclude that all zero-loss solutions of the loss L^m in an overparameterized network with width $m \geq k$ (we also allow for zero-overparameterization) are identical up to symmetries for a certain class of activation functions.

Remark. *The function $\sigma_{\alpha,\gamma}(x) = \sigma_{soft}(x) + \alpha\sigma_{sig}(\gamma x)$ with $\alpha, \gamma > 0$ (Figure 5.1) satisfies the Assumption 5.4.1, but the standard softplus $\sigma_{soft}(x) = \log(1 + e^x)$ or sigmoidal $\sigma_{sig}(x) = 1/(1 + e^{-x})$ functions do not. The analysis needs to include other forms of neuron groupings.*

A very interesting question is whether there are other zero-loss solutions outside of the expansion manifold $\Theta_{n\to m}(\theta_*)$ for finite-size datasets. For example, Kuditipudi et al. (2019)

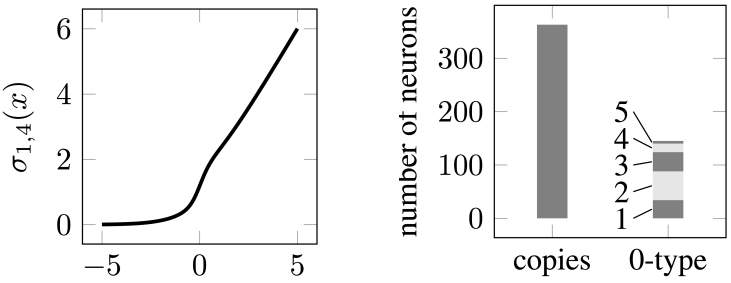


Figure 5.1 – *Left:* The function $\sigma_{\alpha,\gamma}(x) = \sigma_{\text{soft}}(x) + \alpha\sigma_{\text{sig}}(\gamma x)$ satisfies the Assumption 5.4.1. With this activation function, data is generated by a teacher network of width 4. All 50 student networks with width 10 find a global minimum by reaching loss values below 10^{-16}. *Right:* The $500 = 50 \times 10$ hidden neurons of all the 50 student networks are classified as copies of teacher neurons or zero-type neurons with vanishing sum of output weights. The zero-type neurons are further classified according to group size: there are 34 neurons with vanishing output weight (group size 1), 54 neurons that have a partner neuron with the same input weights and the sum of output weights equal to 0 (group size 2) etc.

construct an example of a finite-size dataset for two-layer overparameterized ReLU networks where they find *discrete* global minima points. On the other hand, we show numerically in Martinelli et al., 2023 for a wide range of problems that the neurons of an overparameterized network that has converged to zero-loss form groups and can be pruned away.

5.4.1 Piecewise Linear Connectivity

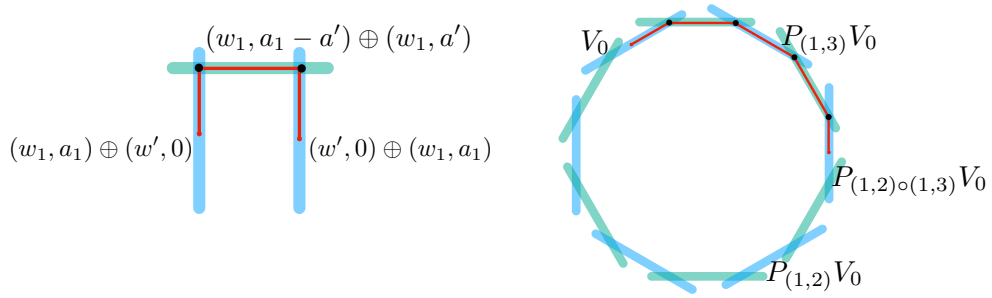
We next focus on describing the precise geometry of the expansion manifolds. First we give the number of affine subspaces in the expansion manifold and a piecewise linear connectivity result. When applied to the expansion manifold of the true parameter, the number of affine subspaces gives the number of zero-loss manifolds and how this number scales with overparameterization. The piecewise linear connectivity comes from the fact that the affine subspaces have non-empty intersections with one another which we will examine in detail in Subsection 5.4.2.

Theorem 5.4.3. *For $m \geq n$, the expansion manifold $\Theta_{n\to m}(\theta)$ of an irreducible point θ consists of exactly*[2]

$$T(n,m) := \sum_{j=0}^{m-n} \sum_{\substack{s_1+\ldots+s_n+z_1+\ldots+z_j=m \\ s_i,z_i\geq 1}} \binom{m}{s_1,\ldots,s_n,z_1,\ldots,z_j} \frac{1}{c_1!\ldots c_{m-n}!}$$

distinct affine subspaces (none is including another one) of dimension at least $\min(d_0, d_{out})(m-n)$*, where c_i is the number of occurences of i among $(z_1,\ldots,z_j)$. For $m > n$, $\Theta_{n\to m}(\theta)$ is connected: any pair of distinct points $\theta, \theta' \in \Theta_{n\to m}$ is connected via a union of line segments $\gamma : [0,1] \to$*

[2] $\binom{n_1+\cdots+n_r}{n_1,\ldots,n_r}$ denotes the coefficient $\frac{(n_1+\cdots+n_r)!}{n_1!\ldots n_r!}$.



(a) (b)

Figure 5.2 – *Piecewise-linear connectivity of the expansion manifold.* The arrangement of the affine subspaces is demonstrated geometrically. Blue subspaces have one vanishing output weight, green subspaces have two identical incoming weight vectors. *(a)* $\Theta_{1\to 2}(\theta^1)$; case of a network with two hidden neurons with parameters $(w_1, a_1) \oplus (w', 0)$. The base subspace $V_0 = (w_1, a_1) \oplus (w', 0)$ is connected to a neighbor subspace $P_{(1,2)}V_0$ via three line segments: we first shift w' towards w_1 while keeping the other parameters fixed and then move a_1 to a' while keeping the summation of the outgoing weights fixed. *(b)* $\Theta_{2\to 3}(\theta^2)$; case of a network with three hidden neurons with the base subspace $V_0 = (w_1, a_1) \oplus (w_2, a_2) \oplus (w', 0)$. V_0 is connected to any other blue subspace $P_\pi V_0$ through transitions from one neighbor to the next. Note that there are $T(2,3) = 12$ subspaces.

$\Theta_{n\to m}$ *such that* $\gamma(0) = \theta$ *and* $\gamma(1) = \theta'$.

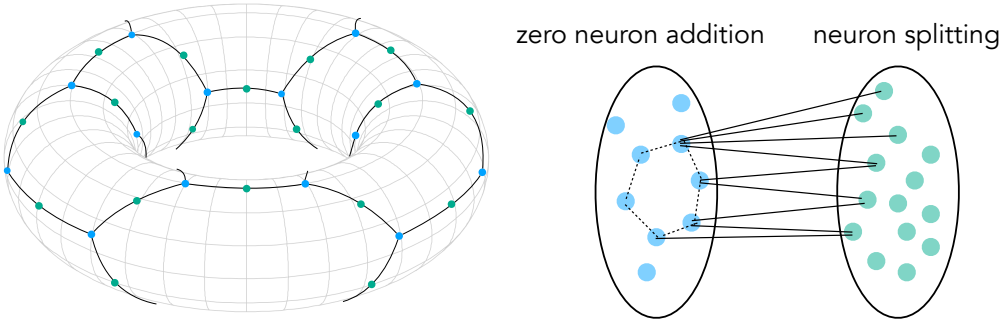
Proof (Sketch). The number of affine subspaces T is equal to the distinct permutations of the incoming weight vectors $(w_1, \ldots, w_n, w'_1, \ldots, w'_j)$ for all possible partitions represented by (s, z) where w_i's are distinct and w'_i's are dummy variables representing *zero-type* neurons. The normalization factor $1/c_1!c_2!\cdots c_{m-n}!$ cancels the repetitions coming from the zero-type neurons $(w'_1, \ldots, w'_j)$. For example for the simplest case $m = n$, there is no room for zero-type neurons. As a result we have

$$T(n,n) = \sum_{\substack{s_1+\ldots+s_n=n \\ s_i \geq 1}} \binom{n}{k_1, \ldots, k_n} = \binom{n}{1, \ldots, 1} = n!$$

distinct subspaces of dimension $\min(d_0, d_{\text{out}})(m - r) = 0$.

For the general case $m > n$, the proof for connectivity follows from the following observations. We start from a base subspace $V_0 = V_{s,z}(\theta)$, where there is a zero-type neuron with outgoing weight vector exactly zero[3] at position i^*. The neighbor subspaces $P_{(i^*,i)}V_0$, where $(i^*, i) \in S_m$ is a transposition that swaps the two neurons, are connected to the base subspace via three line segments (Figure 5.2-a). Since any permutation is a composition of transpositions, permuted subspaces $P_\pi V_0$ can be reached via a union of line segments by going from one neighbor to

[3] If all zero-type neurons are part of a group with more than one neuron, we can choose the first neuron in a group and set its outgoing weight vector to zero while respecting the condition in Eq. 5.1.



(a) $\Theta_{3\to 4}(\theta^3)$ (b) $\Theta_{k\to k+1}(\theta^k)$

Figure 5.3 – *Connectivity graph of the affine subspaces in the expansion manifold.* Blue vertices represent the affine subspaces where the extra neuron is a zero neuron, green vertices represent the affine subspaces where the extra neuron is splitted from one of the teacher neurons. *(a)* The exact connectivity graph for $k = 3$. There are $T(3,4) = 60$ subspaces (24 blue and 36 green), where each blue subspace is connected to three green subspaces and each green subspace is connected to two blue subspaces. There are 12 cliques made of 12 vertices (one blue followed by another green) which is identical to the clique in Figure 5.2-b in the sense that the minimum number egdes (i.e. line segments) needed to get back to the same vertex requires swapping two neurons of the teacher network and back. *(b)* Structure of the connectivity graph for $m = k+1$. There are $(k+1)!$ blue dots and $(k+1)!k/2$ green dots. Each blue dot is connected to k green dots, and each green dot is connected to two blue dots; forming $(k+1)!k$ edges in the graph. Blue and green dots form cliques of 12 vertices (shown as a clique of 6 blue vertices connected with dashed lines). Each blue vertex participates in $\binom{k}{2}$ cliques.

the next (Figure 5.2-b). ■

5.4.2 Connectivity Graph of Affine Subspaces

Given the number of affine subspaces in an expansion manifold, it remains to study how they are connected to one another. This can be phrased as a graph problem: each vertex represents an affine subspace and we draw an edge if two affine subspaces intersect each other. We call this the connectivity graph and give its properties for the case when one neuron is added. Since there is only one room for the extra neuron, this can either be a zero neuron

$$(w_1, a_1) \oplus \ldots \oplus (w_k, a_k) \oplus (w', 0) \tag{5.5}$$

which intersects k affine subspaces corresponding to the splitting of one of the k neurons, for example,

$$(w_1, a_1 - a') \oplus \ldots \oplus (w_k, a_k) \oplus (w_1, a'). \tag{5.6}$$

We note that the affine subspace above intersects 2 affine subspaces of the zero-neuron type through its first and $(k+1)$-th neuron. No type of affine subspaces intersect with their own type, hence forming a bipartite graph as shown in Figure 5.3-b. We leave the exact description of the connectivity graph in terms of formal graph theory for future. Moreover, describing the connectivity graph starting with the number of edges for the expansion manifolds when more than one neuron is added remains an open question. An intruiging question is whether the connectivity graph can give some insight into training dynamics.

5.5 Mild vs. Vast Overparameterization

5.5.1 The Landscape Complexity

We showed that the global minima manifold grows with overparameterization due to numerous arrangements of hidden neurons representing a zero-loss solution. Factoring this in, the appropriate landscape complexity measure for overparameterized networks needs to be normalized with the number of zero-loss manifolds. While it is true that we have manifolds of critical points instead of discrete points, they are all 'tiny' compared to the ambient dimensionality of the parameter space. Therefore we focus on the comparison of the number of critical subspaces and that of global minima subspaces. We propose a landscape complexity measure that compares the scaling law of the critical manifolds at the lowest energy level (low-index saddles) with the scaling law of the zero-loss manifolds

$$C(k,m) := \frac{G(k-1,m)}{T(k,m)}. \tag{5.7}$$

To study the scaling of the landscape complexity $C(k,m)$ we first give a closed-form formula for the scaling law T in terms of the scaling law G. This is proven in Appendix-??????? using Newton's series for finite differences Milne-Thomson, 2000 and a counting argument:

Proposition 5.5.1. *For $k \leq m$, we have*

$$T(k,m) = G(k,m) + \sum_{\ell=1}^{m-k} \binom{m}{\ell} G(k,m-\ell)\, g(\ell)$$

where $g(\ell) = \sum_{n=1}^{\ell} \frac{1}{n!} G(n,\ell)$. Moreover, we have that the scaling law T has the same growth as the scaling law G in the following limit for fixed k

$$T(m-k,m) \sim G(m-k,m) \quad \text{as } m \to \infty.$$

Using the limit rates of the scaling law G (Lemma 4.5.2) and Proposition D.2.2, we get the following insightful limiting behaviors for the landscape complexity

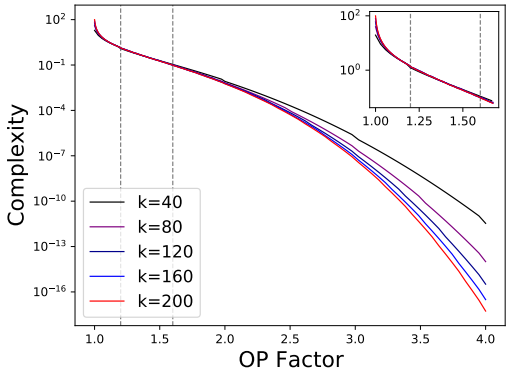


Figure 5.4 – The landscape complexity gradually decreases with overparameterization (OP) factor $\rho = m/k$. A fast decay takes place at the very onset of overparameterization (until the first dashed line at $\rho = 1.2$) which is followed by an exponential decay (until the second dashed line at $\rho = 1.6$); shown in the inset. Afterward, there is even a faster than exponential decay kicking in which pushes the landscape complexity down to zero rapidly. We expect this decay to slow down eventually to exponential decay and match the infinite-width limit rate in eq. (5.8). In the infinite teacher width limit at the onset of overparameterization, we observe that complexity grows, however slowly; it is not overly visible in log-scale (better seen in the inset; linear growth in the limit $k \to \infty$ in eq. (5.9)).

- *In the infinite-width limit,* for fixed k, we have that

$$C(k,m) \leq \frac{G(k-1,m)}{G(k,m)} \sim \left(\frac{k-1}{k}\right)^m \quad \text{as} \quad m \to \infty \tag{5.8}$$

  which goes to zero exponentially fast. This is the well-studied limit where the gradient flow converges to zero loss in commonly studied regimes of training such as the so-called lazy and mean-field regimes (Jacot, Gabriel, and Hongler, 2018b; Chizat and Bach, 2018c).

- *In the infinite data complexity limit & for mild overparameterization,* i.e. as $k \to \infty$, $m = k + h$ with fixed $h \geq 0$, we have that

$$C(k,m) \sim c_0 \frac{m^{h+1} m!}{m^h m!} = c_0 m \quad \text{as} \quad m \to \infty \tag{5.9}$$

  which grows linearly as a function of overparameterization m. As a result, the landscape complexity approaches to infinity.

In general, we observe that the landscape complexity gradually decreases with overparameterization for arbitrary widths (see Figure 5.4). The decrease is exponential at the onset of overparameterization (until about a factor of 1.6) which is followed by an even faster decay.

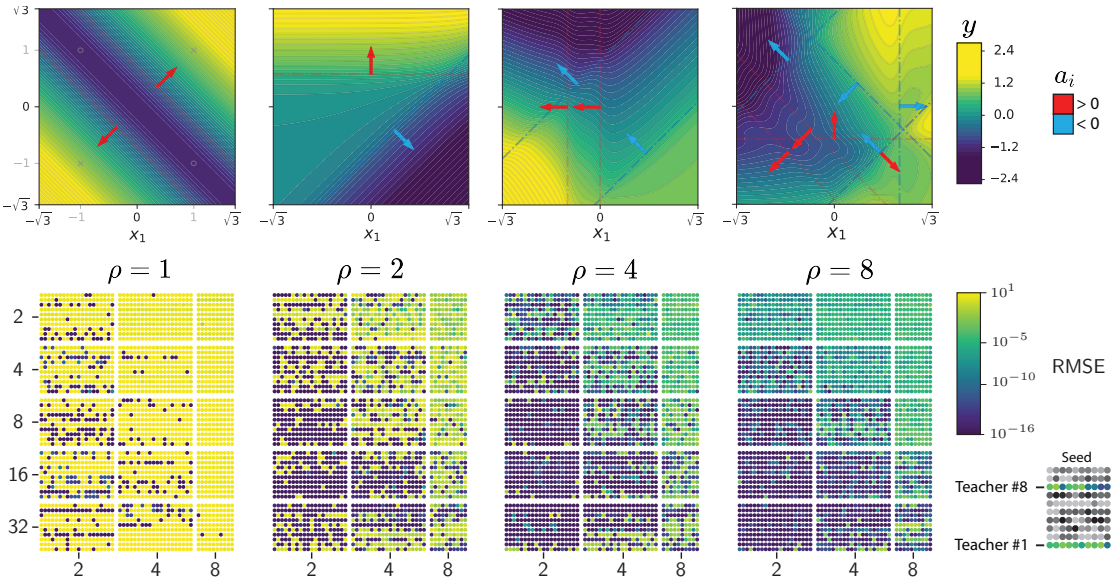


Figure 5.5 – *Top: Teacher complexity increases with number k of hidden neurons;* contourplot of the teacher network output for $d_0 = 2$ input dimensions. Each hidden neuron generates a hyperplane, ${w_j}^T x + b_j = 0$ (dashed lines); the direction of the feature vector w_j is indicated by an arrow and the sign of the output weight a_j by its color. Top left: generalisation of the XOR or parity-bit problem to a regression setting. From left to right: As the number of hidden neuron increases, the level lines become more intricated. *Bottom: Effects of overparametrization on convergence;* for each combination of $d_0 = 2, 4, 8, 16, 32$ and $k = 2, 4, 8$ we generated 10 teachers; for each teacher we trained 20 or 10 students with different seeds and hidden layer size $\rho \cdot k$. Each dot corresponds to one seed (see inset bottom right). Dark blue dots indicate loss below 10^{-14}. Student networks with overparameterization $\rho = 4$ or larger are more likely to converge to near-zero loss than those without ($\rho = 1$). The general trend is that overparameterization helps to converge to zero-loss parameters. For difficult teachers, $r/d_0 \geq 1$, training is very slow and convergence to zero-loss is not guaranteed in finite time.

5.5.2 Numerics

In contrast to random teachers (Saad and Solla, 1995; Goldt et al., 2019; Raman, Rotondo, and O'Leary, 2019) our approach contains regression problems with an XOR-like structure (Fig. 5.5, top) which is encouraged by the selection of biases from a relatively small set of significantly different values. Moreover, randomly initialized networks tend to behave as constant random function as depth increases (Jakub and Nica, 2023), yielding uninteresting data-generator models. In contrast to pure checker-board problems (Rumelhart, G. E. Hinton, and Williams, 1986), only a subset of hyperplanes is aligned with one of the axis. We use the symmetry-free activation function $\sigma = \sigma_{\text{soft}}(x) + \sigma_{\text{sig}}(4x)$ in this subsection.

We trained overparameterized students on this family of teachers. In order to be closest to

the theoretical setting of near-zero loss and to obtain perfect parameter recovery, we used the package MLPGradientFlow.jl (Brea, Martinelli, et al., 2023) that allowed us to find global and local minima with machine precision accuracy for tiny networks (Fig. 5.5, $\rho = 1$). However, for slightly larger networks it becomes challenging to converge fully to global minima within a reasonable amount of time (Fig. 5.5, $\rho \in \{4, 8\}$) and methods to deal with imperfectly trained students are needed. Since training is full-batch, the only source of randomness is in the initialisation. Figure 5.5 shows a beneficial trend as overparameterization increases.

Across all the problems considered, the change in convergence rate (the ratio of dark blue dots to the number of trials) in going from the OP factor $\rho = 1$ to $\rho = 2$ seems more significant compared to the the improvement in the convergence rate at $\rho = 4$ and $\rho = 8$ – which is supported by the exponential decay of the landscape complexity in the mild overparameterization regime (until about a factor of 1.6). The landscape complexity decreases even faster in the vast overparameterization regime (after about the factor 1.6), however this further growth of the number of zero-loss manifolds does not seem to facililate training as much. Our landscape complexity therefore is not sufficient to explain convergence rate trends, say after a factor of $\rho = 2$. This might be either a limitation of our landscape complexity measure, or an artifact due to the small scale of the toy problems considered here ($k \in \{2, 4, 8\}$). We also observe a significant dependence on the dataset (or teacher) complexity: as the teacher expansion ratio r/d_0 increases, it becomes harder to train overparameterized students to global minima.

5.6 Deep Neural Networks

In this section, we introduce the expansion manifold for multi-layer networks that enables obtaining connectivity and counting results on the global minima manifold for multi-layer networks (i.e., generalizing Theorem 5.4.3). A neural network with L layers $f^{(L)} : \mathbb{R}^{d_0} \to \mathbb{R}^{d_{\text{out}}}$ with widths $\mathbf{n} = (n_1, n_2, \ldots, n_{L-1})$ is

$$f^{(L)}(x) = W^{(L)}\sigma(W^{(L-1)} \cdots \sigma(W^{(1)}x))) \tag{5.10}$$

where $W^{(\ell)} \in \mathbb{R}^{n_\ell \times n_{\ell-1}}$ for $\ell = 1, \ldots, L$ with $n_0 = d_0$ and $n_L = d_{\text{out}}$, the non-linearity σ is applied element-wise, and $\theta = (W^{(L)}, \ldots, W^{(1)})$ is the vector of parameters. Observing that any pair of weight matrices $(W^{(\ell)}, W^{(\ell+1)})$ for $\ell = 1, \ldots, L-1$ forms a two-layer network within the multi-layer network, we say that a multi-layer network is irreducible if all pairs $(W^{(\ell)}, W^{(\ell+1)})$ are irreducible.

We define the expansion manifold of an irreducible network with widths $\mathbf{n}$ into larger widths $\mathbf{m} = (m_1, m_2, \ldots, m_{L-1})$ by taking the sequential expansion manifolds of all pairs $(W^{(\ell)}, W^{(\ell+1)})$. More precisely, we define the multi-layer expansion manifold as follows

$$\Theta_{\mathbf{n}\to\mathbf{m}}(\theta) := \{\phi_1 : \phi_{L-1} \in \Theta^{(L-1)}_{\mathbf{n}\to\mathbf{m}}(\theta), ..., \phi_1 \in \Theta^{(1)}_{\mathbf{n}\to\mathbf{m}}(\phi_2)\} \tag{5.11}$$

where $\Theta^{(\ell)}_{\mathbf{n}\to\mathbf{m}}(\phi)$ substitutes the pair $(W^{(\ell)}, W^{(\ell+1)})$ with those of a point in the usual expansion

manifold (Def. 5.3.3). Since each expansion leaves the output of the network unchanged, all points in this expansion have the same loss. Note that the order in which we take these expansions affects the final manifold; expanding from the last layer to the first one gives the largest final manifold. The same final manifold can be obtained via a 'forward pass' if one considers expansion up to an equivalence of the incoming weight vectors.

Assume that a minimal L-layer network achieves a unique (up to permutation) global minimum point θ_* with widths $(k_1, k_2, \ldots, k_{L-1})$. In an overparameterized network of widths $(m_1, \ldots, m_{L-1})$ with $m_\ell > k_\ell$ for all $\ell \in [L-1]$ (i.e. at least one extra neuron at every hidden layer), we find a connected manifold of global minimum, which is simply the multi-layer expansion manifold $\Theta_{\mathbf{k}\to\mathbf{m}}(\theta_*)$ of the minimum point θ_*.

Similarly, we can consider the symmetry-induced critical points for multi-layer networks by applying neuron splitting to the neurons of all hidden layers. The number of affine subspaces of the symmetry-induced critical points is exponential in depth since the permutation-symmetry applies to every hidden layer.

Landscape Complexity. We consider the case where a minimal L-layer network with k neurons at each hidden layer reaches a global minimum point θ_* and an overparameterization of $m = k + h$ neurons at each hidden layer. The landscape complexity is then

$$C^{(L)}(k, m) = C(k, m)^{L-1}$$

which is exponential in depth. Therefore in the mildly overparameterized regime, i.e. when h is small, we see that the ratio of the scaling law of low-index saddles to that of global minima grows exponentially with depth. For the vastly overparameterized regime, i.e. when h is large, we observe the opposite effect: the dominance of the scaling law of global minima is stronger in the multi-layer case. Finally, we observe a width-depth trade-off in reaching a dominance of the global minima: one can either increase the width of a two-layer network so that the complexity goes down to zero; or increase the depth in a network where each layer is just large enough to guarantee that the two-layers complexity is smaller than one which eventually decreases the total ratio down to zero.

5.7 Conclusion & Future Directions

We showed that the addition of a single neuron connects affine subspaces of zero-loss points. This gives a simple explanation of 'linear mode connectivity' observed in practice. Beyond linear mode connectivity, we developed the machinery to describe paths and the number of piecewise linear segments connecting two arbitrary solutions. When an arbitrary number of neurons is added, we gave the scaling law of the zero-loss manifolds (affine subspaces) as a function of the data complexity (i.e. teacher width) and overparameterization. Our mathematical result is that for input distributions with full-support and a broad class of activation functions, any zero-loss parameter is equivalent to the teacher parameter up to

zero neuron addition, neuron splitting, and permutation therefore the scaling law is exact.

Using the scaling laws of the manifolds of symmetry-induced critical points and zero-loss points, we proposed a landscape complexity measure to study the difficulty of training in overparameterized networks of finite-width. In mildly overparameterized networks, the landscape complexity is large ($\gg 1$), so that in practice, the gradient trajectories may get influenced by these saddles or even get either transiently or effectively stuck in their neighborhood for a fraction of typical initializations. However this regime is rather transient thanks to the fast decay of the landscape complexity, and for a small factor of overparameterization (around 1.5), the lanscape complexity goes below 1 and into a rapidly decaying regime. We observe empirical signatures of the fast then slow decay of landscape complexity numerically for a large number of toy problems. From a practical point of view, our theoretical results pave the way to applications in optimization of non-convex neural network loss landscapes via a combination of overparameterization and pruning (Martinelli et al., 2023).

6 Neural Networks with Few Neurons

In this chapter, we present recent unpublished results on the analysis of shallow neural networks with few neurons. We study the interesting problem of learning in the classic student-teacher setup, when the student has only a few neurons hence not enough capacity to match the teacher network function. We present the main results in Section 6.1, related works in Section 6.2, and problem setup in Section 6.3. The main contribution is the reformulation of the problem as a constrained optimization problem that applies to the students of arbitrary widths which we present in Section 6.4. We apply this to the one-neuron network in Section 6.5, getting closed-form expressions for the optimal solution. We close with the conclusions and discussion of future work in Section 6.6. The proofs and further discussions are presented in the Appendix E.

6.1 Main Results

- We propose a reparameterization of the two-layer teacher-student problem which enables a constrained optimization formulation. We assume that the teacher network has orthogonal incoming weights and the input data is standard Gaussian. The optimal solution of the constrained optimization problem is equivalent to the optimal solution of the underparamerized student network (for a student network that is at least as wide as the teacher, the optimal loss is trivially zero).

- For the one neuron network, we identify necessary conditions on the so-called interactions which yields that at the optimal neuron is equally aligned with all teacher incoming vectors. We show that the common monotonic activation functions satisfy this condition.

- For ReLU activation function, we give the closed-form expression of the optimal solution of the one neuron network using the analytic formula of the interaction.

- The optimal solution of a one-neuron network for general monotonic activation functions has a simple interpretation: the incoming vector of the one-neuron network

implements a damped average of all incoming teacher vectors while its outgoing weight is larger than the sum of outgoing weights of the teacher network.

6.2 Related Works

Although the teacher-student framework is a commonly studied model of neural networks, neither global landscape nor gradient flow dynamics are fully understood even in the simplest cases. In the limit when the input dimension goes to infinity, a well-known result is that online stochastic gradient descent converges to a deterministic limit (Saad and Solla, 1995; Goldt et al., 2019). Recent work of Veiga et al., 2022; Arous, Gheissari, and Jagannath, 2022 extended the analysis to the non-vanishing learning rates. Another line of work studies the case of finite input dimension. When the numbers of student neurons, teacher neurons (that are orthogonal), and input dimension are equal, for the ReLU activation function, I. Safran and Shamir, 2018 show that local minima are prevalent and some families are characterized by Arjevani and Field, 2021. Despite the vast literature, there is no rigorous result on the optimal solution of the one-neuron network approximating multiple teacher neurons, beyond the case of a one-neuron teacher network (Tian, 2017; Mei, Y. Bai, and Montanari, 2018; Yehudai and Ohad, 2020; Vardi, Yehudai, and Shamir, 2021; Wu, 2022).

6.3 Setup

Network function: Consider a (student) two-layer network function $f : \mathbb{R}^d \to \mathbb{R}$ with n neurons

$$f(x) = \sum_{j=1}^{n} a_j \sigma\left(w_j \cdot x\right) \tag{6.1}$$

where $w_j \in \mathbb{R}^d$ is the incoming vector to and $a_j \in \mathbb{R}$ the outgoing weight of the neuron j. The activation function σ is twice differentiable unless it is specified to be ReLU.

Parameter: We use the following notation for the parameter $\theta \in \mathbb{R}^P$ with $P = (d+1)n$

$$\theta = (w_1, a_1) \oplus \cdots \oplus (w_n, a_n) \tag{6.2}$$

where $\oplus$ concatenates the neuron vectors (w_1, a_1) back to back. Sometimes θ is made explicit in the network function $f(x|\theta) = f(x)$.

Orthogonal teacher network (with $b_i = 1$): For the study of the under-parameterized networks, we assume that the target function is a (teacher) two-layer neural network

$$f^*(x) = \sum_{i=1}^{k} b_i \sigma(v_i \cdot x) \tag{6.3}$$

where all output weights b_i are equal to unity and incoming vectors $v_1, \ldots, v_k \in \mathbb{R}^d$ are or-

thonormal vectors. This implies that the input dimension is at least as large as the number of teacher neurons, i.e. $d \geq k$.

Loss function: We consider a squared loss and assume that the input distribution is standard Gaussian unless otherwise stated which yields the following non-convex loss function

$$L_{\text{orig}}((w_j, a_j)_{j=1}^n) = \mathbb{E}_{x\sim\mathcal{N}(0,I_d)}\left[\left(\sum_{j=1}^n a_j\sigma(w_j\cdot x) - \sum_{i=1}^k b_i\sigma(v_i\cdot x)\right)^2\right]. \tag{6.4}$$

Using the linearity of expectation, the loss function in Eq. 6.4 can be expanded to a weighted sum of the following type of Gaussian integral terms (see Section 6.4)

$$\mathbb{E}_{x\sim\mathcal{N}(0,I_d)}[\sigma(w_1\cdot x)\sigma(w_2\cdot x)]. \tag{6.5}$$

Since the input distribution is assumed to be standard Gaussian, both $w_1\cdot x$ and $w_2\cdot x$ are centered Gaussian random variables. Hence the above integral can be expressed in terms of the covariance of the two-dimensional Gaussian $(w_1\cdot x, w_2\cdot x)$ which reduces the dimensionality of the problem. This is a standard trick in teacher-student problems and the covariance parameters after reduction are called summary statistics in probability (Arous, Gheissari, and Jagannath, 2022) or order parameters in statistical physics (Goldt et al., 2019). We express the covariance as follows

$$\mathbb{E}_{x\sim\mathcal{N}(0,I_d)}\begin{bmatrix}(w_1\cdot x)^2 & (w_1\cdot x)(w_2\cdot x)\\ (w_1\cdot x)(w_2\cdot x) & (w_2\cdot x)^2\end{bmatrix} = \begin{bmatrix}r_1^2 & r_1r_2u\\ r_1r_2u & r_2^2\end{bmatrix} \tag{6.6}$$

where $r_i = \|w_i\|$ for $i = 1,2$ and $u = w_1\cdot w_2/(r_1r_2)$ which allows us to explicitly bound the *correlation* $u\in[-1,1]$ thanks to the Cauchy-Schwarz inequality. The key to solving the one-neuron network lies in the novel study of the *interaction* function $g:\mathbb{R}^2_{\geq 0}\times[-1,1]\to\mathbb{R}$, i.e. the Gaussian integral term in Eq. 6.5

$$g(r_1,r_2,u) = \mathbb{E}_{x\sim\mathcal{N}(0,I_d)}[\sigma(w_1\cdot x)\sigma(w_2\cdot x)]. \tag{6.7}$$

This formalism sets the groundwork for a fundamental study of under-parameterized networks and how they express the optimal solution.

6.4 Risk Minimization as a Constrained Optimization Problem

Expanding the loss in terms of the interaction functions (see Eq. 6.7), we get

$$L = \sum_{j=1}^n a_j^2 g(r_j,r_j,1) + 2\sum_{j\neq j'} a_ja_{j'}g(r_j,r_{j'},\tilde{u}_{jj'}) - 2\sum_{j=1}^n\sum_{i=1}^k a_jg(r_j,1,u_{ji}) + \text{const} \tag{6.8}$$

where $r_j = \|w_j\|$ is the norm of an incoming vector, $\tilde{u}_{jj'} = w_j\cdot w_{j'}/(r_jr_{j'})$ is the correlation between two normalized student incoming vectors, and $u_{ji} = w_j\cdot v_i/r_j$ is the correlation

between a normalized student and a teacher incoming vector for all $j \neq j' \in [n]$ and $i \in [k]$. The constant does not depend on the problem parameters and represents the squared target function integrated over the Gaussian input distribution, i.e. $\mathbb{E}_{x\sim N(0,I_d)}[f^*(x)^2]$.

All correlations are by definition bounded by 1 in absolute value. However, there are tighter geometric constraints on the correlations between student and teacher incoming vectors which we make explicit next. Note that we can expand any normalized incoming vector of the student in the basis of the teacher's incoming vectors as follows

$$\frac{w_j}{r_j} = \sum_{i=1}^{k} u_{ji} v_i + v_\perp \tag{6.9}$$

where $v_\perp$ is orthogonal to all v_i. The normalized vector has unit norm, hence the expansion also has a unit norm which yields the constraint on the *student-teacher* correlations u_{ji}

$$\sum_{i=1}^{k} u_{ji}^2 = 1 - \|v_\perp\|^2 \leq 1 \quad \forall j \in [n]. \tag{6.10}$$

Next, we can express the *student-student* correlations $\tilde{u}_{jj'}$ using Eq. 6.9 as

$$\tilde{u}_{jj'} = (\sum_{i=1}^{k} u_{ji} v_i + v_\perp)(\sum_{i=1}^{k} u_{j'i} v_i + v'_\perp) = \sum_{i=1}^{k} u_{ji} u_{j'i} + v_\perp \cdot v'_\perp \tag{6.11}$$

which yields the second constraint on the optimization problem after noting $|v_\perp \cdot v'_\perp| \leq \|v_\perp\| \|v'_\perp\|$

$$\left| \tilde{u}_{jj'} - \sum_{i=1}^{k} u_{ji} u_{j'i} \right| \leq \sqrt{1 - \sum_{i=1}^{k} u_{ji}^2} \sqrt{1 - \sum_{i=1}^{k} u_{j'i}^2} \quad \forall j' \neq j \in [n]. \tag{6.12}$$

Remark. *We can relax the assumption of orthogonality between $v_1, \ldots, v_k$ to linear independence. Let us collect the incoming vectors into a matrix $V = [v_1, \ldots, v_k] \in \mathbb{R}^{d\times k}$. The expansion in Eq. 6.9 can be rewritten as*

$$\frac{w_j}{r_j} = \sum_{i=1}^{k} \gamma_{ji} v_i + v_\perp = V\Gamma_j + v_\perp \tag{6.13}$$

where $\Gamma_j = [\gamma_{j1}, \ldots, \gamma_{jk}] \in \mathbb{R}^k$. The normalized vector has unit norm, hence we have

$$\|V\Gamma_j + v_\perp\|^2 = \Gamma_j^T V^T V \Gamma_j + \|v_\perp\|^2 = 1. \tag{6.14}$$

For the correlation vector $U_j = [u_{j1}, \ldots, u_{jk}] \in \mathbb{R}^k$, we have

$$U_j = V^T V \Gamma_j \tag{6.15}$$

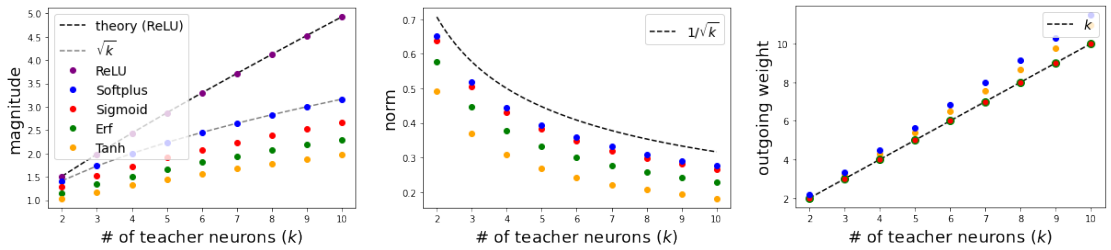


Figure 6.1 – *Structure of the optimal solution of the one-neuron network for various activation functions.* We trained 20 seeds of one-neuron students learning from the orthogonal teacher networks with $k = 2, \dots, 10$ neurons. All students converge to the same solution except for tanh and erf for which there is also a sign-symmetric solution[1]. *(a)* For ReLU, the magnitude $\|w^*\| a^*$ exactly matches with the result of Theorem 6.5.3. For softplus, the magnitude is very close to $\sqrt{k}$; for sigmoid, tanh, and erf, it is below $\sqrt{k}$. *(b)* The norm of the incoming vector is smaller than $1/\sqrt{k}$ not only for softplus, but also for sigmoid, tanh, and erf. *(c)* The outgoing weight is larger than k for softplus and tanh; and it is virtually k for sigmoid and erf.

which yields the following constraint due to Eq. 6.14 for all $j \in [n]$

$$U_j^T (V^T V)^{-1} U_j \leq 1. \tag{6.16}$$

6.5 The Optimal Solution of the One-Neuron Network

As a first step towards characterizing the optimal solution in the networks with few neurons, we focus on the case of a single neuron ($n = 1$) approximating an orthogonal teacher with k neurons. The loss in Eq. 6.8 simplifies

$$L(r, a, (u_i)_{i=1}^k) = a^2 g(r, r, 1) - 2a \sum_{i=1}^k g(r, 1, u_i) + \text{const}, \quad \text{subject to } r \geq 0, \sum_{i=1}^k u_i^2 \leq 1. \tag{6.17}$$

We assume that the interaction g is twice differentiable in the correlation for all $u \in (-1, 1)$ throughout the paper. Let us introduce some properties of g that we will use in the following

Assumption 6.5.1. *The interaction satisfies the following properties for all* $r_1, r_2 > 0$ *and* $u \in (-1, 1)$

$$(i)\ \frac{d}{du} g(r_1, r_2, u) > 0, \qquad (ii)\ \frac{d^2}{du^2} g(r_1, 1, u) u < \frac{d}{du} g(r_1, 1, u). \tag{6.18}$$

[1] For odd functions such as tanh and erf, i.e. $\sigma(x) = \text{erf}(x/\sqrt{2})$, there are two solutions that are sign-symmetric: the usual one where the correlations are all $1/\sqrt{k}$, with a norm and outgoing weight denoted by (r, a) with $a > 0$, and its symmetric solution where the correlations are all $-1/\sqrt{k}$, with a norm and outgoing weight $(r, -a)$. Only the first solution is plotted in the figure for finer comparison on the positive scale.

Applying Lemma E.1.1 for $\sigma_1 = \sigma_2 = \sigma$ that is differentiable, we obtain

$$\frac{d}{du} g(r_1, r_2, u) = r_1 r_2 \mathbb{E}[\sigma'(r_1 x)\sigma'(r_2 y)]. \tag{6.19}$$

Hence if σ is increasing (or decreasing),[2] the integrand on the right-hand side is positive; satisfying Assumption 6.5.1 (i). The relationship between Assumption 6.5.1 (ii) and the activation function is more subtle. We show in Lemma E.1.2 that the interactions of the common activation functions such as softplus, sigmoid, tanh, and erf (respectively)

$$\sigma(x) = \frac{1}{\beta}\log(e^{\beta x}+1) \text{ with } \beta \in (0,2], \quad \frac{1}{1+e^{-x}}, \quad \frac{1-e^{-x}}{1+e^{-x}}, \quad \frac{2}{\sqrt{\pi}}\int_0^x e^{-t^2/2} dt,$$

satisfy Assumption 6.5.1 (ii). Finally, for the ReLU activation function, i.e. $\sigma(x) = \max(0, x)$, the interaction has an analytical expression (Cho and L. Saul, 2009; I. Safran and Shamir, 2018),

$$g(r_1, r_2, u) = r_1 r_2 h(u) \text{ where } h(u) = \frac{1}{2\pi}\left(\sqrt{1-u^2} + (\pi - \arccos(u))u\right).$$

With a simple calculation, we show that h satisfies Assumption 6.5.1 (i) and (ii) (with a slight modification in the domain; see the proof of Theorem 6.5.3). For the activation functions for which the corresponding interaction satisfies Assumption 6.5.1 (such as the ones listed above), we show that at any non-trivial critical point of the loss in Eq. 6.17, the correlations satisfy the constraint $u_1^2 + \ldots + u_k^2 = 1$ and they are equal.

Theorem 6.5.2. *Assume that the interaction $g(r_1, r_2, u)$ satisfies Assumption 6.5.1. At any non-trivial critical point of the loss in Eq. 6.17, that satisfies $a \neq 0$ and $r \neq 0$, all correlations u_i are identical (for $k > 1$) and equal to either $1/\sqrt{k}$ or $-1/\sqrt{k}$ (for all $k \geq 1$).*

Intuitively, for positive outgoing weight, the incoming vector of the one-neuron network is pulled toward the teacher's incoming vectors to minimize the loss. Hence, the optimal solution has to be in the span of the teacher's incoming vectors which means that the constraint in Eq. 6.17 is satisfied. Since the teacher's incoming vectors are orthogonal to each other and they are equal in strength – with unit norm and unit outgoing weight – the incoming vector should align with each of them equally to be stationary (see Appendix Section E.2.1 for the proof). In particular, since the optimal solution of the original problem in the weight space is a stationary point in Eq. 6.17, it aligns with all incoming vectors of the teacher equally.

Thanks to Theorem 6.5.2, the loss in Eq. 6.17 can now be reduced to a two-dimensional loss parameterized by a and r. First, let us study it analytically for the ReLU activation function.

Theorem 6.5.3. *Assume that the activation function is $\sigma(x) = \max(0, x)$. Any global minima (w^*, a^*) of the loss in Eq. 6.4 for the one-neuron network ($n = 1$) satisfies*

$$\|w^*\| a^* = k\frac{h(1/\sqrt{k})}{h(1)}, \quad w^* = \frac{\|w^*\|}{\sqrt{k}}\sum_{i=1}^{k} v_i, \tag{6.20}$$

[2] Increasing (decreasing) means strictly increasing (decreasing) everywhere in this paper.

forming an equal-loss hyperbola. The optimal loss is given by

$$L^* = k^2\left(h(0) - \frac{h(1/\sqrt{k})^2}{h(1)}\right) + k(h(1) - h(0)). \tag{6.21}$$

For ReLU, thanks to the analytical formula of the interaction, we characterize the global landscape of the one-neuron network (see the proof of Theorem 6.5.3 in Section E.2.2). In particular, we give the closed-form formula of the optimal solution that is unique up to the scaling symmetry of ReLU in Theorem 6.5.3.

For general activation functions (with the exception of erf), there is no analytical expression for the interaction. In Lemma E.1.2, we showed that the common activation functions listed above satisfy the Assumption 6.5.1, therefore we know that the correlations are equal and the constraint on the correlations is satisfied at the optimal solution. What remains is solving a fixed point equation on the norm (see Appendix Section E.2.1) which we do numerically (see Figure E.1). Numerically we observe that there is a unique solution to the fixed point equation where $r \leq 1/\sqrt{k}$ which yields a lower bound on the outgoing weight. We propose the following conjecture which is also supported by Figure 6.1 for softplus, sigmoid, erf, and tanh activation functions.

Conjecture 6.5.4. *Assume that the interaction of the activation function satisfies Assumption 6.5.1. There is a unique critical point, that is the global minimum, of the loss in Eq. 6.17 which satisfies*

$$u_i = 1/\sqrt{k}, \;\; r \leq \frac{1}{\sqrt{k}}, \;\; \textit{and} \;\; k \leq a. \tag{6.22}$$

For the teacher network with one neuron, the unique critical point is given by $u_1 = 1$, $r = 1$, $a = 1$.

According to the conjecture, the incoming vector of the optimal solution of the one-neuron network can be expressed as

$$w^* = \frac{r}{\sqrt{k}} \sum_{i=1}^{k} v_i,$$

with $r \leq 1/\sqrt{k}$. Hence, the incoming vector implements a damped average of the incoming vectors of the teacher with a damping factor of $r/\sqrt{k} \leq 1/k$. The outgoing weight at the optimal solution is at least k since the one neuron should compensate for approximating k teacher neurons.

6.6 Conclusion & Generalizations

One-Neuron Network. We proposed a novel proof to study the optimal solution of the one-neuron network when it learns from a teacher network with an arbitrary number of neurons. The key proof idea relies on a reparameterization of the order parameters/summary statistics

in terms of correlations and norms. We then study the interaction as a function of three parameters (two norms and a correlation) without relying on the analytic expression. For the ReLU activation function, using the analytic expression of the interaction, we gave the closed-form formula of the optimal solution. For some general activation functions, we numerically showed that the optimal solution implements a damped average in its incoming vector and a compensating outgoing weight.

Two-Neuron Network/Future Directions. What about the optimal solution of the two-neuron network? The characterization of this solution and the question of other non-trivial (i.e. irreducible) critical points in under-parameterized neural networks remain a substantial challenge for future research. The properties of the interactions may play a key role in the precise study of the whole regime of shallow networks; the one-neuron case being a non-trivial application. The idea of decoupling the norm and correlations is strong as it yields an equivalent constrained optimization problem (see the constraints in Eq.11 and Eq.13; valid in general). We came up with Lemma E.1.1, which is a derivative law with respect to correlation; it will likely play a key role in solving the underparamerized networks beyond the one-neuron limit.

Another interesting direction is generalizing the proofs to non-orthogonal incoming vectors, and also for arbitrary outgoing weights of the teacher. The general setting for the teacher network is exhaustive because any continuous target function can be approximated by a (teacher) network thanks to the universal approximation theorem (Funahashi, 1989; Cybenko, 1989; Hornik, Stinchcombe, and White, 1989).

Bibliography

Abbé, E., E. B. Adserà, and T. Misiakiewicz (2022). "The merged-staircase property: a necessary and nearly sufficient condition for sgd learning of sparse functions on two-layer neural networks". In: *Conference on Learning Theory*. PMLR, pp. 4782–4887.

Abbé, E., E. Boix-Adserà, M. S. Brennan, et al. (2021). "The staircase property: How hierarchical structure can guide deep learning". In: *Advances in Neural Information Processing Systems*. Ed. by A. Beygelzimer et al.

Abbé, E., E. Boix-Adserà, and T. Misiakiewicz (2023). "SGD learning on neural networks: leap complexity and saddle-to-saddle dynamics". In: *arXiv preprint arXiv:2302.11055*.

Advani, M. S. and A. M. Saxe (2017). *High-dimensional dynamics of generalization error in neural networks*.

Advani, M. S., A. M. Saxe, and H. Sompolinsky (2020). "High-dimensional dynamics of generalization error in neural networks". In: *Neural Networks* 132, pp. 428–446.

Ainsworth, S. K., J. Hayase, and S. Srinivasa (2022). "Git re-basin: Merging models modulo permutation symmetries". In: *arXiv preprint arXiv:2209.04836*.

Anandkumar, A. and R. Ge (2016). "Efficient approaches for escaping higher order saddle points in non-convex optimization". In: *Conference on learning theory*. PMLR, pp. 81–102.

Arjevani, Y. and M. Field (2021). "Analytic study of families of spurious minima in two-layer ReLU neural networks: a tale of symmetry II". In: *Advances in Neural Information Processing Systems* 34, pp. 15162–15174.

Arora, S., N. Cohen, N. Golowich, et al. (2019). "A Convergence Analysis of Gradient Descent for Deep Linear Neural Networks". In: *International Conference on Learning Representations*.

Arora, S., N. Cohen, W. Hu, et al. (2019). "Implicit regularization in deep matrix factorization". English (US). In: *Advances in Neural Information Processing Systems* 32.

Arora, S., S. S. Du, et al. (2019). "On exact computation with an infinitely wide neural net". In: *Advances in Neural Information Processing Systems* 32.

Arora, S., S. Du, et al. (2019). "Fine-grained analysis of optimization and generalization for overparameterized two-layer neural networks". In: *International Conference on Machine Learning*. PMLR, pp. 322–332.

Arous, G. B., R. Gheissari, and A. Jagannath (2022). "High-dimensional limit theorems for SGD: Effective dynamics and critical scaling". In: *arXiv preprint arXiv:2206.04030*.

Arous, G. B., S. Mei, et al. (2019). "The landscape of the spiked tensor model". In: *Communications on Pure and Applied Mathematics* 72.11, pp. 2282–2330.

Au, B. et al. (2018). *Large permutation invariant random matrices are asymptotically free over the diagonal*. To appear in Annals of Probability.

Auffinger, A., G. B. Arous, and J. Černy (2013). "Random matrices and complexity of spin glasses". In: *Communications on Pure and Applied Mathematics* 66.2, pp. 165–201.

Ba, J. et al. (2022). "High-dimensional asymptotics of feature learning: How one gradient step improves the representation". In: *arXiv preprint arXiv:2205.01445*.

Bahri, Y. et al. (2021). "Explaining neural scaling laws". In: *arXiv preprint arXiv:2102.06701*.

Bai, Z. and Z. Wang (2008). "Large sample covariance matrices without independence structures in columns". In: *Statistica Sinicia* 18, pp. 425–442.

Baity-Jesi, M. et al. (2018). "Comparing dynamics: Deep neural networks versus glassy systems". In: *International Conference on Machine Learning*. PMLR, pp. 314–323.

Baldi, P. and K. Hornik (1989). "Neural networks and principal component analysis: Learning from examples without local minima". In: *Neural networks* 2.1, pp. 53–58.

Bartlett, P. L. et al. (2019). "Benign overfitting in linear regression". In: *arXiv preprint arXiv:1906.11300*.

Belkin, M., D. Hsu, S. Ma, et al. (2018). "Reconciling modern machine learning and the bias-variance trade-off". In: *arXiv preprint arXiv:1812.11118*.

Belkin, M., D. Hsu, and J. Xu (2019). "Two models of double descent for weak features". In: *arXiv preprint arXiv:1903.07571*.

Belkin, M., S. Ma, and S. Mandal (Feb. 2018). "To understand deep learning we need to understand kernel learning". In: *arXiv preprint*.

Benigni, L. and S. Péché (2019). "Eigenvalue distribution of nonlinear models of random matrices". In: *arXiv preprint arXiv:1904.03090*.

Benzing, F. et al. (2022). "Random initialisations performing above chance and how to find them". In: *arXiv preprint arXiv:2209.07509*.

Bishop, C. M. (2006). *Pattern recognition and machine learning*. springer.

Bordelon, B., A. Canatar, and C. Pehlevan (2020a). "Spectrum Dependent Learning Curves in Kernel Regression and Wide Neural Networks". In: *arXiv preprint arXiv:2002.02561*.

— (2020b). "Spectrum dependent learning curves in kernel regression and wide neural networks". In: *International Conference on Machine Learning*. PMLR, pp. 1024–1034.

Boursier, E., L. Pillaud-Vivien, and N. Flammarion (2022). "Gradient flow dynamics of shallow ReLU networks for square loss and orthogonal inputs". In: *arXiv preprint arXiv:2206.00939*.

Boyd, S., S. P. Boyd, and L. Vandenberghe (2004). *Convex optimization*. Cambridge university press.

Bray, A. J. and D. S. Dean (2007). "Statistics of critical points of Gaussian fields on large-dimensional spaces". In: *Physical review letters* 98.15, p. 150201.

Brea, J., F. Martinelli, et al. (2023). "MLPGradientFlow: going with the flow of multilayer perceptrons (and finding minima fast and accurately)". In: *arXiv preprint arXiv:2301.10638*.

Brea, J., B. Şimşek, et al. (2019). "Weight-space symmetry in deep networks gives rise to permutation saddles, connected by equal-loss valleys across the loss landscape". In: *arXiv preprint arXiv:1907.02911*.

Brown, T. et al. (2020). "Language models are few-shot learners". In: *Advances in neural information processing systems* 33, pp. 1877–1901.

Caponnetto, A. and E. De Vito (2007). "Optimal rates for the regularized least-squares algorithm". In: *Foundations of Computational Mathematics* 7.3, pp. 331–368.

Chizat, L. and F. Bach (Sept. 2020). "Implicit Bias of Gradient Descent for Wide Two-layer Neural Networks Trained with the Logistic Loss". In: *Proceedings of Thirty Third Conference on Learning Theory*. Ed. by J. Abernethy and S. Agarwal. Vol. 125. Proceedings of Machine Learning Research. PMLR, pp. 1305–1338.

Chizat, L. and F. Bach (2018a). "A note on lazy training in supervised differentiable programming". In: *arXiv preprint arXiv:1812.07956*.

— (2018b). "On the Global Convergence of Gradient Descent for Over-parameterized Models using Optimal Transport". In: *Advances in Neural Information Processing Systems 31*. Curran Associates, Inc., pp. 3040–3050.

— (2018c). "On the global convergence of gradient descent for over-parameterized models using optimal transport". In: *Advances in neural information processing systems* 31.

Cho, Y. and L. Saul (2009). "Kernel methods for deep learning". In: *Advances in neural information processing systems* 22.

Cho, Y. and L. K. Saul (2009). "Kernel Methods for Deep Learning". In: *Advances in Neural Information Processing Systems 22*. Curran Associates, Inc., pp. 342–350.

Choromanska, A. et al. (2015). "The loss surfaces of multilayer networks". In: *Artificial intelligence and statistics*. PMLR, pp. 192–204.

Cooper, Y. (2020). "The critical locus of overparameterized neural networks". In: *arXiv preprint arXiv:2005.04210*.

Cybenko, G. (1989). "Approximation by superpositions of a sigmoidal function". In: *Mathematics of control, signals and systems* 2.4, pp. 303–314.

d'Ascoli, S. et al. (2020). "Double Trouble in Double Descent: Bias and Variance (s) in the Lazy Regime". In: *arXiv preprint arXiv:2003.01054*.

Dauphin, Y. N. et al. (2014). "Identifying and attacking the saddle point problem in high-dimensional non-convex optimization". In: *Advances in neural information processing systems*, pp. 2933–2941.

Devlin, J. et al. (2018). "Bert: Pre-training of deep bidirectional transformers for language understanding". In: *arXiv preprint arXiv:1810.04805*.

Dobriban, E. and S. Wager (Feb. 2018). "High-dimensional asymptotics of prediction: Ridge regression and classification". In: *Ann. Statist.* 46.1, pp. 247–279. DOI: 10.1214/17-AOS1549.

Dosovitskiy, A. et al. (2020). "An image is worth 16x16 words: Transformers for image recognition at scale". In: *arXiv preprint arXiv:2010.11929*.

Draxler, F. et al. (2018). "Essentially no barriers in neural network energy landscape". In: *arXiv preprint arXiv:1803.00885*.

Du, S. S., C. Jin, J. D. Lee, M. I. Jordan, B. Póczos, et al. (2017). "Gradient descent can take exponential time to escape saddle points". In: *Proceedings of the 31st International Conference on Neural Information Processing SystemsDecember 2017, NIPS'17*. Curran Associates, Inc., pp. 1067–1077.

Du, S. S., X. Zhai, et al. (2019). "Gradient Descent Provably Optimizes Over-parameterized Neural Networks". In: *International Conference on Learning Representations*.

Du, S. S., C. Jin, J. D. Lee, M. I. Jordan, A. Singh, et al. (2017). "Gradient descent can take exponential time to escape saddle points". In: *Advances in neural information processing systems*, pp. 1067–1077.

Du, S. S., X. Zhai, et al. (2018). "Gradient descent provably optimizes over-parameterized neural networks". In: *arXiv preprint arXiv:1810.02054*.

Du, S. and J. Lee (2018). "On the power of over-parametrization in neural networks with quadratic activation". In: *International conference on machine learning*. PMLR, pp. 1329–1338.

Du, S., J. Lee, et al. (2019). "Gradient descent finds global minima of deep neural networks". In: *International Conference on Machine Learning*. PMLR, pp. 1675–1685.

Eaton, M. (Jan. 2007). "Multivariate Statistics: A Vector Space Approach". In: *Journal of the American Statistical Association* 80. DOI: 10.2307/20461449.

Eftekhari, A. (13–18 Jul 2020). "Training Linear Neural Networks: Non-Local Convergence and Complexity Results". In: *Proceedings of the 37th International Conference on Machine Learning*. Ed. by H. D. III and A. Singh. Vol. 119. Proceedings of Machine Learning Research. PMLR, pp. 2836–2847.

Elhage, N. et al. (2022). "Toy Models of Superposition". In: *arXiv preprint arXiv:2209.10652*.

Entezari, R. et al. (2021). "The role of permutation invariance in linear mode connectivity of neural networks". In: *arXiv preprint arXiv:2110.06296*.

Fan, Z. and Z. Wang (2020). "Spectra of the conjugate kernel and neural tangent kernel for linear-width neural networks". In: *Advances in neural information processing systems* 33, pp. 7710–7721.

Fort, S., G. K. Dziugaite, et al. (2020). "Deep learning versus kernel learning: an empirical study of loss landscape geometry and the time evolution of the Neural Tangent Kernel". In: *arXiv preprint arXiv:2010.15110*.

Fort, S. and S. Jastrzebski (2019). "Large scale structure of neural network loss landscapes". In: *Advances in Neural Information Processing Systems*, pp. 6709–6717.

Frankle, J. and M. Carbin (2018). "The lottery ticket hypothesis: Finding sparse, trainable neural networks". In: *arXiv preprint arXiv:1803.03635*.

Frankle, J., G. K. Dziugaite, et al. (2020). "Linear mode connectivity and the lottery ticket hypothesis". In: *International Conference on Machine Learning*. PMLR, pp. 3259–3269.

Freeman, C. D. and J. Bruna (2016). "Topology and geometry of half-rectified network optimization". In: *arXiv preprint arXiv:1611.01540*.

Fukumizu, K. and S.-i. Amari (2000). "Local minima and plateaus in hierarchical structures of multilayer perceptrons". In: *Neural networks* 13.3, pp. 317–327.

Fukumizu, K., S. Yamaguchi, et al. (2019). "Semi-flat minima and saddle points by embedding neural networks to overparameterization". In: *Advances in Neural Information Processing Systems* 32, pp. 13868–13876.

Funahashi, K.-I. (1989). "On the approximate realization of continuous mappings by neural networks". In: *Neural networks* 2.3, pp. 183–192.

G. Matthews, A. G. de et al. (2018). "Gaussian Process Behaviour in Wide Deep Neural Networks". In: *International Conference on Learning Representations*.

Gabriel, F. (2015). “Combinatorial Theory of Permutation-Invariant Random Matrices II: Cumulants, Freeness and Levy Processes”. In: *arXiv preprint arXiv:1507.02465*.

Gardner, E. and B. Derrida (1989). “Three unfinished works on the optimal storage capacity of networks”. In: *Journal of Physics A: Mathematical and General* 22.12, p. 1983.

Garipov, T. et al. (2018). “Loss surfaces, mode connectivity, and fast ensembling of dnns”. In: *Advances in Neural Information Processing Systems* 31, pp. 8789–8798.

Ge, R. et al. (2015). “Escaping from saddle points—online stochastic gradient for tensor decomposition”. In: *Conference on learning theory*. PMLR, pp. 797–842.

Geiger, M., A. Jacot, S. Spigler, F. Gabriel, L. Sagun, S. d’Ascoli, et al. (2019). “Scaling description of generalization with number of parameters in deep learning”. In: *arXiv preprint arXiv:1901.01608*.

Geiger, M., A. Jacot, S. Spigler, F. Gabriel, L. Sagun, S. d’Ascoli, et al. (2020). “Scaling description of generalization with number of parameters in deep learning”. In: *Journal of Statistical Mechanics: Theory and Experiment* 2020.2, p. 023401.

Geiger, M., S. Spigler, S. d’Ascoli, et al. (2019). “Jamming transition as a paradigm to understand the loss landscape of deep neural networks”. In: *Physical Review E* 100.1, p. 012115.

Geiger, M., S. Spigler, A. Jacot, et al. (2020). “Disentangling feature and lazy training in deep neural networks”. In: *Journal of Statistical Mechanics: Theory and Experiment* 2020.11, p. 113301.

Geman, S., E. Bienenstock, and R. Doursat (1992a). “Neural networks and the bias/variance dilemma”. In: *Neural computation* 4.1, pp. 1–58.

— (1992b). “Neural networks and the bias/variance dilemma”. In: *Neural computation* 4.1, pp. 1–58.

Gerfo, L. L. et al. (2008). “Spectral algorithms for supervised learning”. In: *Neural Computation* 20.7, pp. 1873–1897.

Gidel, G., F. Bach, and S. Lacoste-Julien (2019). “Implicit Regularization of Discrete Gradient Dynamics in Linear Neural Networks”. In: *Advances in Neural Information Processing Systems*. Ed. by H. Wallach et al. Vol. 32. Curran Associates, Inc.

Girosi, F., M. Jones, and T. Poggio (1995). “Regularization theory and neural networks architectures”. In: *Neural computation* 7.2, pp. 219–269.

Gissin, D., S. Shalev-Shwartz, and A. Daniely (2020). “The Implicit Bias of Depth: How Incremental Learning Drives Generalization”. In: *International Conference on Learning Representations*.

Glorot, X. and Y. Bengio (13–15 May 2010). “Understanding the difficulty of training deep feedforward neural networks”. In: *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*. Ed. by Y. W. Teh and M. Titterington. Vol. 9. Proceedings of Machine Learning Research. Chia Laguna Resort, Sardinia, Italy: JMLR Workshop and Conference Proceedings, pp. 249–256.

Głuch, G. and R. Urbanke (2021). “Noether: The More Things Change, the More Stay the Same”. In: *arXiv preprint arXiv:2104.05508*.

Goldt, S. et al. (2019). “Dynamics of stochastic gradient descent for two-layer neural networks in the teacher-student setup”. In: *Advances in neural information processing systems* 32.

Gunasekar, S., J. D. Lee, et al. (2018). “Implicit Bias of Gradient Descent on Linear Convolutional Networks”. In: *Advances in Neural Information Processing Systems*. Ed. by S. Bengio et al. Vol. 31. Curran Associates, Inc.

Gunasekar, S., J. Lee, et al. (Oct. 2018). “Characterizing Implicit Bias in Terms of Optimization Geometry”. In: *Proceedings of the 35th International Conference on Machine Learning*. Ed. by J. Dy and A. Krause. Vol. 80. Proceedings of Machine Learning Research. PMLR, pp. 1832–1841.

Hastie, T. et al. (2022). “Surprises in high-dimensional ridgeless least squares interpolation”. In: *The Annals of Statistics* 50.2, pp. 949–986.

Horn, R. A. and C. R. Johnson (2012). *Matrix analysis*. Cambridge university press.

Hornik, K., M. Stinchcombe, and H. White (1989). “Multilayer feedforward networks are universal approximators”. In: *Neural networks* 2.5, pp. 359–366.

Huang, J. and H.-T. Yau (13–18 Jul 2020). “Dynamics of Deep Neural Networks and Neural Tangent Hierarchy”. In: *ICML*. Proceedings of Machine Learning Research 119. Ed. by H. D. III and A. Singh, pp. 4542–4551.

Jacot, A., F. Gabriel, and C. Hongler (2018a). “Neural Tangent Kernel: Convergence and Generalization in Neural Networks”. In: *NeurIPS*.

— (2018b). “Neural tangent kernel: Convergence and generalization in neural networks”. In: *Advances in neural information processing systems* 31.

Jacot, A., F. Ged, B. Şimşek, et al. (2021). “Deep linear networks dynamics: Low-rank biases induced by initialization scale and l2 regularization”. In: *arXiv preprint arXiv:2106.15933*.

— (2022). *Saddle-to-Saddle Dynamics in Deep Linear Networks: Small Initialization Training, Symmetry, and Sparsity*.

Jacot, A., B. Şimşek, et al. (2020a). *Implicit Regularization of Random Feature Models*.

— (2020b). “Implicit regularization of random feature models”. In: *International Conference on Machine Learning*. PMLR, pp. 4631–4640.

— (2020c). “Kernel Alignment Risk Estimator: Risk Prediction from Training Data”. In: *Advances in Neural Information Processing Systems*. Ed. by H. Larochelle et al. Vol. 33. Curran Associates, Inc., pp. 15568–15578.

— (2020d). “Kernel alignment risk estimator: Risk prediction from training data”. In: *Advances in Neural Information Processing Systems* 33, pp. 15568–15578.

Jakub, C. and M. Nica (2023). “Depth Degeneracy in Neural Networks: Vanishing Angles in Fully Connected ReLU Networks on Initialization”. In: *arXiv preprint arXiv:2302.09712*.

Ji, Z. and M. Telgarsky (2018). “Gradient descent aligns the layers of deep linear networks”. In: *CoRR* abs/1810.02032.

— (2020). “Directional convergence and alignment in deep learning”. In: *Advances in Neural Information Processing Systems*. Ed. by H. Larochelle et al. Vol. 33. Curran Associates, Inc., pp. 17176–17186.

Jin, C. et al. (2017a). “How to Escape Saddle Points Efficiently”. In: *Proceedings of the 34th International Conference on Machine Learning - Volume 70*. ICML’17. Sydney, NSW, Australia: JMLR.org, pp. 1724–1732.

— (2017b). “How to escape saddle points efficiently”. In: *International Conference on Machine Learning*. PMLR, pp. 1724–1732.

Jordan, K. et al. (2022). “REPAIR: REnormalizing Permuted Activations for Interpolation Repair”. In: *arXiv preprint arXiv:2211.08403*.

Kamalaruban, P. et al. (2020). “Robust reinforcement learning via adversarial training with langevin dynamics”. In: *Advances in Neural Information Processing Systems* 33, pp. 8127–8138.

Kaplan, J. et al. (2020). “Scaling laws for neural language models”. In: *arXiv preprint arXiv:2001.08361*.

Kawaguchi, K. (2016). “Deep Learning without Poor Local Minima”. In: *Advances in Neural Information Processing Systems*. Ed. by D. Lee et al. Vol. 29. Curran Associates, Inc.

Kuditipudi, R. et al. (2019). “Explaining landscape connectivity of low-cost solutions for multilayer nets”. In: *Advances in Neural Information Processing Systems*, pp. 14601–14610.

Kung, J. P., G.-C. Rota, and C. H. Yan (2009). *Combinatorics: the Rota way*. Cambridge University Press.

Kunin, D. et al. (2020). “Neural Mechanics: Symmetry and Broken Conservation Laws in Deep Learning Dynamics”. In: *arXiv preprint arXiv:2012.04728*.

Laurent, T. and J. Brecht (2018). “Deep linear networks with arbitrary loss: All local minima are global”. In: *International conference on machine learning*. PMLR, pp. 2902–2907.

LeCun, Y., Y. Bengio, and G. Hinton (2015). “Deep learning”. In: *nature* 521.7553, pp. 436–444.

Lee, J., Y. Bahri, et al. (2017a). “Deep neural networks as gaussian processes”. In: *arXiv preprint arXiv:1711.00165*.

— (2017b). “Deep neural networks as gaussian processes”. In: *arXiv preprint arXiv:1711.00165*.

Lee, J., S. Schoenholz, et al. (2020). “Finite versus infinite neural networks: an empirical study”. In: *Advances in Neural Information Processing Systems* 33, pp. 15156–15172.

Lee, J., L. Xiao, et al. (2019). “Wide neural networks of any depth evolve as linear models under gradient descent”. In: *Advances in neural information processing systems* 32.

Lee, J. D., M. Simchowitz, et al. (23–26 Jun 2016). “Gradient Descent Only Converges to Minimizers”. In: *29th Annual Conference on Learning Theory*. Ed. by V. Feldman, A. Rakhlin, and O. Shamir. Vol. 49. Proceedings of Machine Learning Research. Columbia University, New York, New York, USA: PMLR, pp. 1246–1257.

Lee, J. D., I. Panageas, et al. (2019a). “First-order methods almost always avoid strict saddle points”. In: *Mathematical programming* 176.1, pp. 311–337.

— (2019b). “First-order methods almost always avoid strict saddle points”. In: *Mathematical programming* 176.1, pp. 311–337.

Li, Z., Y. Luo, and K. Lyu (2020). “Towards resolving the implicit bias of gradient descent for matrix factorization: Greedy low-rank learning”. In: *arXiv preprint arXiv:2012.09839*.

Li, Z., R. Wang, et al. (2019). “Enhanced Convolutional Neural Tangent Kernels”. In: *arXiv preprint arXiv:1911.00809*.

Liang, T., A. Rakhlin, and X. Zhai (2020). “On the Multiple Descent of Minimum-Norm Interpolants and Restricted Lower Isometry of Kernels”. In: *arXiv preprint arXiv:1908.10292 [cs, math, stat]*.

Liu, S. and E. Dobriban (2020). "Ridge Regression: Structure, Cross-Validation, and Sketching". In: *International Conference on Learning Representations*.

Louart, C., Z. Liao, and R. Couillet (Feb. 2017). "A Random Matrix Approach to Neural Networks". In: *The Annals of Applied Probability* 28. DOI: 10.1214/17-AAP1328.

Luo, T. et al. (2021). "Phase Diagram for Two-layer ReLU Neural Networks at Infinite-width Limit". In: *Journal of Machine Learning Research* 22.71, pp. 1–47.

Lyu, K. and J. Li (2020). "Gradient Descent Maximizes the Margin of Homogeneous Neural Networks". In: *International Conference on Learning Representations*.

Maillard, A., G. B. Arous, and G. Biroli (2020). "Landscape complexity for the empirical risk of generalized linear models". In: *Mathematical and Scientific Machine Learning*. PMLR, pp. 287–327.

Mallat, S. (2012). "Group invariant scattering". In: *Communications on Pure and Applied Mathematics* 65.10, pp. 1331–1398.

Marteau-Ferey, U. et al. (2019). "Beyond Least-Squares: Fast Rates for Regularized Empirical Risk Minimization through Self-Concordance". In: *CoRR* abs/1902.03046.

Martinelli, F. et al. (2023). "Mild Overparameterization for Network Parameter Identification". In: arXiv preprint arXiv:2304.12794.

Mei, S., Y. Bai, and A. Montanari (2018). "The landscape of empirical risk for nonconvex losses". In: *The Annals of Statistics* 46.6A, pp. 2747–2774.

Mei, S., T. Misiakiewicz, and A. Montanari (2019). "Mean-field theory of two-layers neural networks: dimension-free bounds and kernel limit". In: *arXiv preprint arXiv:1902.06015*.

Mei, S. and A. Montanari (2019). "The generalization error of random features regression: Precise asymptotics and double descent curve". In: *arXiv preprint arXiv:1908.05355*.

Mei, S., A. Montanari, and P.-M. Nguyen (2018a). "A mean field view of the landscape of two-layer neural networks". In: *Proceedings of the National Academy of Sciences* 115.33, E7665–E7671.

— (2018b). "A mean field view of the landscape of two-layer neural networks". In: *Proceedings of the National Academy of Sciences* 115.33, E7665–E7671.

Mertikopoulos, P. et al. (2020). "On the almost sure convergence of stochastic gradient descent in non-convex problems". In: *Advances in Neural Information Processing Systems* 33, pp. 1117–1128.

Milne-Thomson, L. M. (2000). *The calculus of finite differences*. American Mathematical Soc.

Moroshko, E. et al. (2020). "Implicit Bias in Deep Linear Classification: Initialization Scale vs Training Accuracy". In: *Advances in Neural Information Processing Systems*. Ed. by H. Larochelle et al. Vol. 33. Curran Associates, Inc., pp. 22182–22193.

Mousavi-Hosseini, A. et al. (2022). "Neural Networks Efficiently Learn Low-Dimensional Representations with SGD". In: *arXiv preprint arXiv:2209.14863*.

Nakkiran, P. et al. (2019). "Deep double descent: Where bigger models and more data hurt". In: *arXiv preprint arXiv:1912.02292*.

Neal, B. et al. (2018). "A Modern Take on the Bias-Variance Tradeoff in Neural Networks". In: *arXiv preprint arXiv:1810.08591*.

Neal, R. M. (1996). *Bayesian Learning for Neural Networks*. Secaucus, NJ, USA: Springer-Verlag New York, Inc.

Neyshabur, B., S. Bhojanapalli, et al. (2017). "Exploring generalization in deep learning". In: *Advances in neural information processing systems* 30.

Neyshabur, B., Z. Li, et al. (2018). "Towards understanding the role of over-parametrization in generalization of neural networks". In: *arXiv preprint arXiv:1805.12076*.

Neyshabur, B., R. Tomioka, and N. Srebro (2014). "In search of the real inductive bias: On the role of implicit regularization in deep learning". In: *arXiv preprint arXiv:1412.6614*.

Nguyen, Q. (2019). "On connected sublevel sets in deep learning". In: *arXiv preprint arXiv:1901.07417*.

Nouiehed, M. and M. Razaviyayn (2021). "Learning deep models: Critical points and local openness". In: *INFORMS Journal on Optimization*.

Oymak, S. and M. Soltanolkotabi (2020). "Toward moderate overparameterization: Global convergence guarantees for training shallow neural networks". In: *IEEE Journal on Selected Areas in Information Theory* 1.1, pp. 84–105.

Pascanu, R. et al. (2014). "On the saddle point problem for non-convex optimization". In: *arXiv preprint arXiv:1405.4604*.

Poggio, T. and F. Girosi (1990). "Networks for approximation and learning". In: *Proceedings of the IEEE* 78.9, pp. 1481–1497.

Rahimi, A. and B. Recht (2008a). "Random Features for Large-Scale Kernel Machines". In: *Advances in Neural Information Processing Systems 20*. Curran Associates, Inc., pp. 1177–1184.

— (2008b). "Random features for large-scale kernel machines". In: *Advances in neural information processing systems*, pp. 1177–1184.

— (2009). "Weighted sums of random kitchen sinks: Replacing minimization with randomization in learning". In: *Advances in neural information processing systems*, pp. 1313–1320.

Raman, D. V., A. P. Rotondo, and T. O'Leary (2019). "Fundamental bounds on learning performance in neural circuits". In: *Proceedings of the National Academy of Sciences* 116.21, pp. 10537–10546.

Ros, V. et al. (2019). "Complex energy landscapes in spiked-tensor and simple glassy models: Ruggedness, arrangements of local minima, and phase transitions". In: *Physical Review X* 9.1, p. 011003.

Rotskoff, G. M. and E. Vanden-Eijnden (2018a). "Neural networks as interacting particle systems: Asymptotic convexity of the loss landscape and universal scaling of the approximation error". In: *stat* 1050, p. 22.

— (2018b). "Trainability and accuracy of neural networks: An interacting particle system approach". In: *arXiv preprint arXiv:1805.00915*.

Rotskoff, G. and E. Vanden-Eijnden (2018). "Parameters as interacting particles: long time convergence and asymptotic error scaling of neural networks". In: *Advances in Neural Information Processing Systems 31*. Curran Associates, Inc., pp. 7146–7155.

Rudi, A. and L. Rosasco (2017). "Generalization properties of learning with random features". In: *Advances in Neural Information Processing Systems*, pp. 3215–3225.

Rumelhart, D. E., G. E. Hinton, and R. J. Williams (1986). "Learning Internal Representations by Error Propagation". In: *Parallel Distributed Processing*. Ed. by D. E. Rumelhart, J. L. McClelland, and P. R. Group. Vol. 1. MIT press Cambridge, MA. Chap. 8, pp. 318–362.

Saad, D. and S. A. Solla (1995). "On-line learning in soft committee machines". In: *Physical Review E* 52.4, p. 4225.

Safran, I. M., G. Yehudai, and O. Shamir (2021). "The effects of mild over-parameterization on the optimization landscape of shallow ReLU neural networks". In: *Conference on Learning Theory*. PMLR, pp. 3889–3934.

Safran, I. and O. Shamir (2018). "Spurious local minima are common in two-layer relu neural networks". In: *International conference on machine learning*. PMLR, pp. 4433–4441.

Sagan, B. E. (2013). *The symmetric group: representations, combinatorial algorithms, and symmetric functions*. Vol. 203. Springer Science & Business Media.

Sagun, L., U. Evci, et al. (2017). "Empirical analysis of the hessian of over-parametrized neural networks". In: *arXiv preprint arXiv:1706.04454*.

Sagun, L., V. U. Guney, et al. (2014). "Explorations on high dimensional landscapes". In: *arXiv preprint arXiv:1412.6615*.

Sarao Mannelli, S., E. Vanden-Eijnden, and L. Zdeborová (2020). "Optimization and generalization of shallow neural networks with quadratic activation functions". In: *Advances in Neural Information Processing Systems* 33, pp. 13445–13455.

Saxe, A. M., J. L. McClelland, and S. Ganguli (2019). "A mathematical theory of semantic development in deep neural networks". In: *Proceedings of the National Academy of Sciences* 116.23, pp. 11537–11546. DOI: 10.1073/pnas.1820226116.

— (2014). *Exact solutions to the nonlinear dynamics of learning in deep linear neural networks*.

Schölkopf, B., A. Smola, and K.-R. Müller (1998a). "Nonlinear component analysis as a kernel eigenvalue problem". In: *Neural Computation* 10.5, pp. 1299–1319.

— (1998b). "Nonlinear component analysis as a kernel eigenvalue problem". In: *Neural Computation* 10.5, pp. 1299–1319.

Seung, H. S., H. Sompolinsky, and N. Tishby (1992). "Statistical mechanics of learning from examples". In: *Physical review A* 45.8, p. 6056.

Shankar, V. et al. (2020). "Neural Kernels Without Tangents". In: *arXiv preprint arXiv:2003.02237*.

Shawe-Taylor, J. and N. Cristianini (2004). *Kernel Methods for Pattern Analysis*. New York, NY, USA: Cambridge University Press.

Silverstein, J. (1995). "Strong Convergence of the Empirical Distribution of Eigenvalues of Large Dimensional Random Matrices". In: *Journal of Multivariate Analysis* 55.2, pp. 331–339. DOI: https://doi.org/10.1006/jmva.1995.1083.

Şimşek, B., A. Bendjeddou, et al. (2023). "Should Under-parameterized Student Networks Copy or Average Teacher Weights?" Under Review for NeurIPS 2023.

Şimşek, B., F. Ged, et al. (2021). "Geometry of the loss landscape in overparameterized neural networks: Symmetries and invariances". In: *International Conference on Machine Learning*. PMLR, pp. 9722–9732.

Singh, S. P. and M. Jaggi (2020). "Model fusion via optimal transport". In: *Advances in Neural Information Processing Systems* 33, pp. 22045–22055.

Sirignano, J. and K. Spiliopoulos (2020). “Mean field analysis of neural networks: A law of large numbers”. In: *SIAM Journal on Applied Mathematics* 80.2, pp. 725–752.

Soudry, D. et al. (2018). “The implicit bias of gradient descent on separable data”. In: *The Journal of Machine Learning Research* 19.1, pp. 2822–2878.

Speicher, R. (2017). “Free Probability and Random Matrices”. In: *Free Probability and Random Matrices*.

Spigler, S. et al. (2018). “A jamming transition from under-to over-parametrization affects loss landscape and generalization”. In: *arXiv preprint arXiv:1810.09665*.

Sridharan, K., S. Shalev-Shwartz, and N. Srebro (2009). “Fast rates for regularized objectives”. In: *Advances in neural information processing systems*, pp. 1545–1552.

Sriperumbudur, B. and Z. Szabó (2015). “Optimal rates for random fourier features”. In: *Advances in Neural Information Processing Systems*, pp. 1144–1152.

Sun, J., Q. Qu, and J. Wright (2015). “When are nonconvex problems not scary?” In: *arXiv preprint arXiv:1510.06096*.

Tian, Y. (2017). “An analytical formula of population gradient for two-layered relu network and its applications in convergence and critical point analysis”. In: *International conference on machine learning*. PMLR, pp. 3404–3413.

Vardi, G., G. Yehudai, and O. Shamir (2021). “Learning a single neuron with bias using gradient descent”. In: *Advances in Neural Information Processing Systems* 34, pp. 28690–28700.

Veiga, R. et al. (2022). “Phase diagram of Stochastic Gradient Descent in high-dimensional two-layer neural networks”. In: *arXiv preprint arXiv:2202.00293*.

Vershynin, R. (2010). “Introduction to the non-asymptotic analysis of random matrices”. In: *arXiv preprint arXiv:1011.3027*.

Vyas, N., Y. Bansal, and P. Nakkiran (2022). “Limitations of the ntk for understanding generalization in deep learning”. In: *arXiv preprint arXiv:2206.10012*.

Woodworth, B. et al. (2020). *Kernel and Rich Regimes in Overparametrized Models*.

Wu, L. (2022). “Learning a Single Neuron for Non-monotonic Activation Functions”. In: *International Conference on Artificial Intelligence and Statistics*. PMLR, pp. 4178–4197.

Yang, G. (Feb. 2019). “Scaling Limits of Wide Neural Networks with Weight Sharing: Gaussian Process Behavior, Gradient Independence, and Neural Tangent Kernel Derivation”. In: *arXiv e-prints*, arXiv:1902.04760, arXiv:1902.04760.

Yang, G. and E. J. Hu (2020). *Feature Learning in Infinite-Width Neural Networks*.

Yang, T. et al. (2012). “Nyström method vs random Fourier features: A theoretical and empirical comparison”. In: *Advances in neural information processing systems*, pp. 476–484.

Yehudai, G. and S. Ohad (2020). “Learning a single neuron with gradient methods”. In: *Conference on Learning Theory*. PMLR, pp. 3756–3786.

Yu, F. X. X. et al. (2016). “Orthogonal Random Features”. In: *Advances in Neural Information Processing Systems*, pp. 1975–1983.

Yun, C., S. Krishnan, and H. Mobahi (2021). “A unifying view on implicit bias in training linear neural networks”. In: *International Conference on Learning Representations*.

Zhang, C. et al. (2016). “Understanding deep learning requires rethinking generalization”. In: *arXiv preprint arXiv:1611.03530*.

Zhang, T. (2003). “Effective dimension and generalization of kernel learning”. In: *Advances in Neural Information Processing Systems*, pp. 471–478.

Zhang, Y., Z. Zhang, et al. (2021). “Embedding principle of loss landscape of deep neural networks”. In: *Advances in Neural Information Processing Systems* 34, pp. 14848–14859.

Zhang, Y., Q. Qu, and J. Wright (2020). “From symmetry to geometry: Tractable nonconvex problems”. In: *arXiv preprint arXiv:2007.06753*.

A Gaussian Random Features Model

A.1 Gaussian Random Features

Proposition A.1.1. *Let $\hat{f}_\lambda^{(RF)}$ be the λ-RF predictor and let $\hat{y} = F\hat{\theta}$ be the prediction vector on training data, i.e. $\hat{y}_i = \hat{f}_\lambda^{(RF)}(x_i)$. The process $\hat{f}_\lambda^{(RF)}$ is a mixture of Gaussians: conditioned on F, we have that $\hat{f}_\lambda^{(RF)}$ is a Gaussian process. The mean and covariance of $\hat{f}_\lambda^{(RF)}$ conditioned on F are given by*

$$\mathbb{E}[\hat{f}_\lambda^{(RF)}(x)|F] = K(x,X)K(X,X)^{-1}\hat{y}, \tag{A.1}$$

$$\mathrm{Cov}[\hat{f}_\lambda^{(RF)}(x), \hat{f}_\lambda^{(RF)}(x')|F] = \frac{\|\hat{\theta}\|^2}{P}\tilde{K}(x,x') \tag{A.2}$$

where $\tilde{K}(x,x') = K(x,x') - K(x,X)K(X,X)^{-1}K(X,x')$ denotes the posterior covariance kernel.

Proof. Let $F = (\frac{1}{\sqrt{P}} f^{(j)}(x_i))_{i,j}$ be the $N \times P$ matrix of values of the random features on the training set. By definition, $\hat{f}_\lambda^{(RF)} = \frac{1}{\sqrt{P}}\sum_{p=1}^P \hat{\theta}_p f^{(p)}$. Conditioned on the matrix F, the optimal parameters $(\hat{\theta}_p)_p$ are not random and $(f^{(p)})_p$ is still Gaussian, hence, conditioned on the matrix F, the process $\hat{f}_\lambda^{(RF)}$ is a mixture of Gaussians. Moreover, conditioned on the matrix F, for any p, p', $f^{(p)}$ and $f^{(p')}$ remain independent, hence

$$\begin{aligned}
\mathbb{E}\left[\hat{f}_\lambda^{(RF)}(x) \mid F\right] &= \frac{1}{\sqrt{P}}\sum_{p=1}^P \hat{\theta}_p \mathbb{E}\left[f^{(p)}(x) \mid f_N^{(p)}\right] \\
\mathrm{Cov}\left[\hat{f}_\lambda^{(RF)}(x), \hat{f}_\lambda^{(RF)}(x') \mid F\right] &= \frac{1}{P}\sum_{p=1}^P \hat{\theta}_p^2 \mathrm{Cov}\left[f^{(p)}(x), f^{(p)}(x') \mid f_N^{(p)}\right].
\end{aligned}$$

where we have set $f_N^{(p)} = (f^{(p)}(x_i))_i \in \mathbb{R}^N$. The value of $\mathbb{E}\left[f^{(p)}(x) \mid f_N^{(p)}\right]$ and $\mathrm{Cov}\left[f^{(p)}(x), f^{(p)}(x') \mid f_N^{(p)}\right]$ are obtained from classical results on Gaussian conditional distributions Eaton, 2007:

$$\begin{aligned}\mathbb{E}\left[f^{(p)}(x) \mid f_N^{(p)}\right] &= K(x,X)K(X,X)^{-1}f_N^{(p)},\\ \mathrm{Cov}\left[f^{(p)}(x), f^{(p)}(x') \mid f_N^{(p)}\right] &= \tilde{K}(x,x'),\end{aligned}$$

where $\tilde{K}(x,x') = K(x,x') - K(x,X)K(X,X)^{-1}K(X,x')$. Thus, conditioned on F, the predictor $\hat{f}_\lambda^{(RF)}$ has expectation:

$$\mathbb{E}\left[\hat{f}_\lambda^{(RF)}(x) \mid F\right] = K(x,X)K(X,X)^{-1}\frac{1}{\sqrt{P}}\sum_{p=1}^{P}\hat{\theta}_p f_N^{(p)} = K(x,X)K(X,X)^{-1}\hat{y}$$

and covariance:

$$\mathrm{Cov}\left[\hat{f}_\lambda^{(RF)}(x), \hat{f}_\lambda^{(RF)}(x') \mid F\right] = \frac{1}{P}\sum_{p=1}^{P}\hat{\theta}_p^2\tilde{K}(x,x') = \frac{\|\hat{\theta}\|^2}{P}\tilde{K}(x,x').$$

□

Using Proposition A.1.1, in order to have a better description of the distribution of the predictor $\hat{f}_{\lambda,\gamma}^{(RF)}$, it remains to study the distributions of both the final labels $\hat{y}$ on the training set and the parameter norm $\|\hat{\theta}\|^2$. In Section A.3, we first study the expectation of the final labels $\hat{y}$: this allows us to study the loss of the average predictor $\mathbb{E}\left[\hat{f}_{\lambda,\gamma}^{(RF)}\right]$. Then in Section A.5, a study of the variance of the predictor allows us to study the average loss of the RF predictor.

A.2 Generalized Wishart Matrix

Setup. In this section, we consider a fixed deterministic matrix K of size $N \times N$ which is diagonal positive semi-definite, with eigenvalues $d_1, \dots, d_N$. We also consider a $P \times N$ random matrix W with i.i.d. standard Gaussian entries.

The key object of study is the $P \times P$ generalized Wishart random matrix $F^TF = \frac{1}{P}WKW^T$ and in particular its Stieltjes transform defined on $z \in \mathbb{C} \setminus \mathbb{R}^+$, where $\mathbb{R}^+ = [0, +\infty[$:

$$m_P(z) = \frac{1}{P}\mathrm{Tr}\left[\left(F^TF - z\mathrm{I}_P\right)^{-1}\right] = \frac{1}{P}\mathrm{Tr}\left[\left(\frac{1}{P}WKW^T - z\mathrm{I}_P\right)^{-1}\right],$$

where K is a fixed positive semi-definite matrix.

Since F^TF has positive real eigenvalues $\lambda_1, \dots, \lambda_P \in \mathbb{R}_+$, and

$$m_P(z) = \frac{1}{P}\sum_{p=1}^{P}\frac{1}{\lambda_p - z},$$

we have that for any $z \in \mathbb{C} \setminus \mathbb{R}^+$,

$$|m_P(z)| \leq \frac{1}{d(z,\mathbb{R}_+)},$$

where $d(z,\mathbb{R}_+) = \inf\{|z-y|, y \in \mathbb{R}^+\}$ is the distance of z to the positive real line. More precisely, $m_P(z)$ lies in the convex hull $\Omega_z = \mathrm{Conv}\left(\{\frac{1}{d-z} : d \in \mathbb{R}_+\}\right)$. As a consequence, the argument $\arg(m_P(z)) \in (-\pi,\pi)$ lies between 0 and $\arg\left(-\frac{1}{z}\right)$, i.e. $m_P(z)$ lies in the cone spanned by 1 and $-\frac{1}{z}$.

Our first lemma implies that the Stieljes transform concentrates around its mean as N and P go to infinity with $\gamma = \frac{P}{N}$ fixed.

Lemma A.2.1. *For any integer $m \in \mathbb{N}$ and any $z \in \mathbb{C} \setminus \mathbb{R}^+$, we have*

$$\mathbb{E}\left[|m_P(z) - \mathbb{E}[m_P(z)]|^m\right] \leq \mathbf{c} P^{-\frac{m}{2}},$$

where $\mathbf{c}$ depends on z, γ, and m only.

Proof. The proof follows Step 1 of Z. Bai and Z. Wang, 2008. Let $w_1, ..., w_N$ be the columns of W from left to right. Let us introduce the $P \times P$ matrices $B(z) = \frac{1}{P} WKW^T - z\mathrm{I}_P$ and $B_{(i)}(z) = \frac{1}{P} W_{(i)} K_{(i)} W_{(i)}^T - z\mathrm{I}_P$ where $W_{(i)}$ is the $P \times (N-1)$ submatrix of W obtained by removing its i-th column w_i, and $K_{(i)}$ is the $(N-1) \times (N-1)$ submatrix of K obtained by removing both its i-th column and i-th row. Since the eigenvalues of WKW^T and $W_{(i)} K_{(i)} W_{(i)}^T$ are all real and positive, $B(z)$ and $B_{(i)}(z)$ are invertible matrices for $z \notin \mathbb{R}^+$.

Noticing that

$$B(z) = \frac{1}{P} WKW^T - z\mathrm{I}_P = \frac{1}{P} W_{(i)} K_{(i)} W_{(i)}^T - z\mathrm{I}_P + \frac{d_i}{P} w_i w_i^T$$

is a rank one perturbation of the matrix $B_{(i)}(z)$, by the Sherman–Morrison's formula, the inverse of $B(z)$ is given by:

$$B(z)^{-1} = \left(B_{(i)}(z)\right)^{-1} - \frac{d_i}{P} \frac{1}{1 + \frac{d_i}{P} w_i^T \left(B_{(i)}(z)\right)^{-1} w_i} \left(B_{(i)}(z)\right)^{-1} w_i w_i^T \left(B_{(i)}(z)\right)^{-1}.$$

We denote $\mathbb{E}_i$ the conditional expectation given $w_{i+1}, ..., w_N$. We have $\mathbb{E}_0[m_P(z)] = m_P(z)$ and $\mathbb{E}_N[m_P(z)] = \mathbb{E}[m_P(z)]$. As a consequence, we get:

$$\begin{aligned} m_P(z) - \mathbb{E}[m_P(z)] &= \sum_{i=1}^{N} \left(\mathbb{E}_{i-1}[m_P(z)] - \mathbb{E}_i[m_P(z)]\right) \\ &= \frac{1}{P} \sum_{i=1}^{N} \left(\mathbb{E}_{i-1} - \mathbb{E}_i\right)\left[\mathrm{Tr}\left(B(z)^{-1}\right)\right] \\ &= \frac{1}{P} \sum_{i=1}^{N} \left(\mathbb{E}_{i-1} - \mathbb{E}_i\right)\left[\mathrm{Tr}\left(B(z)^{-1}\right) - \mathrm{Tr}\left(B_{(i)}(z)^{-1}\right)\right]. \end{aligned}$$

The last equality comes from the fact that $\mathrm{Tr}\left(B_{(i)}(z)^{-1}\right)$ does not depend on w_i, hence

$$\mathbb{E}_{i-1}\left[\mathrm{Tr}\left(B_{(i)}(z)^{-1}\right)\right] = \mathbb{E}_i\left[\mathrm{Tr}\left(B_{(i)}(z)^{-1}\right)\right].$$

Let $g_i : \mathbb{C}\setminus\mathbb{R}^+ \to \mathbb{C}$ be the holomorphic function given by $g_i(z) := \frac{1}{P} w_i^T \left(B_{(i)}(z)\right)^{-1} w_i$. Its derivative is given by $g_i'(z) = \frac{1}{P} w_i^T \left(B_{(i)}(z)\right)^{-2} w_i$. Hence

$$\begin{aligned}\mathrm{Tr}\left(B(z)^{-1}\right) - \mathrm{Tr}\left(B_{(i)}(z)^{-1}\right) &= -\frac{\frac{d_i}{P}\mathrm{Tr}\left(\left(B_{(i)}(z)\right)^{-1} w_i w_i^T \left(B_{(i)}(z)\right)^{-1}\right)}{1+d_i g_i(z)} \\ &= -\frac{d_i g_i'(z)}{1+d_i g_i(z)},\end{aligned}$$

where we used the cyclic property of the trace. We can now bound this difference:

$$\begin{aligned}\left|\mathrm{Tr}\left(B(z)^{-1}\right) - \mathrm{Tr}\left(B_{(i)}(z)^{-1}\right)\right| &= \left|\frac{d_i g_i'(z)}{1+d_i g_i(z)}\right| \\ &\leq \left|\frac{w_i^T \left(B_{(i)}(z)\right)^{-2} w_i}{w_i^T \left(B_{(i)}(z)\right)^{-1} w_i}\right| \\ &\leq \max_w \left|\frac{w^T \left(B_{(i)}(z)\right)^{-2} w}{w^T \left(B_{(i)}(z)\right)^{-1} w}\right| \\ &\leq \| \left(B_{(i)}(z)\right)^{-1} \|_{op} = \max_j |\frac{1}{\nu_j - z}| \leq \frac{1}{d(z,\mathbb{R}^+)},\end{aligned}$$

where ν_j are the eigenvalues of $\frac{1}{P} W_{(i)} K_{(i)} W_{(i)}^T$.

The sequence

$$\left((\mathbb{E}_{N-i} - \mathbb{E}_{N-i+1})\left[\mathrm{Tr}\left(B(z)^{-1}\right) - \mathrm{Tr}\left(B_{(N-i+1)}(z)^{-1}\right)\right]\right)_{i=1,\ldots,N}$$

is a martingale difference sequence. Hence, by Burkholder's inequality, there exists a positive constant K_m such that

$$\begin{aligned}\mathbb{E}\left[|m_P(z) - \mathbb{E}[m_P(z)]|^m\right] &\leq K_m \frac{1}{P^m} \mathbb{E}\left[\left(\sum_{i=1}^N \left|[\mathbb{E}_{i-1} - \mathbb{E}_i]\left(\mathrm{Tr}\left(B(z)^{-1}\right) - \mathrm{Tr}\left(B_{(i)}(z)^{-1}\right)\right)\right|^2\right)^{\frac{m}{2}}\right] \\ &\leq K_m \frac{1}{P^m}\left(N\left(\frac{2}{d(z,\mathbb{R}_+)}\right)^2\right)^{\frac{m}{2}} \\ &\leq K_m \gamma^{-\frac{m}{2}}\left(\frac{2}{d(z,\mathbb{R}_+)}\right)^m P^{-\frac{m}{2}},\end{aligned}$$

hence the desired result with $\mathbf{c} = K_m \gamma^{-\frac{m}{2}}\left(\frac{2}{d(z,\mathbb{R}_+)}\right)^m$. □

The following lemma, which is reminiscent of Lemma 4.5 in Au et al., 2018, is a consequence of Wick's formula for Gaussian random variables and is key to prove Lemma C.4.

Lemma A.2.2. *If $A^{(1)}, \dots, A^{(k)}$ are k square random matrices of size P independent from a standard Gaussian vector w of size P,*

$$\mathbb{E}\left[w^T A^{(1)} w w^T A^{(2)} w \dots w^T A^{(k)} w \right] = \sum_{p \in \boldsymbol{P}_2(2k)} \sum_{p \leq \mathrm{Ker}(i_1,\dots,i_{2k})} \sum_{i_1,\dots,i_{2k} \in \{1,\dots,P\}} \mathbb{E}\left[A^{(1)}_{i_1 i_2} \dots A^{(k)}_{i_{2k-1} i_{2k}} \right] \tag{A.3}$$

where $\boldsymbol{P}_2(2k)$ is the set of pair partitions of $\{1,\dots,2k\}$, $\leq$ is the coarser (i.e. $p \leq q$ if q is coarser than p), and for any $i_1,\dots,i_{2k}$ in $\{1,\dots,P\}$, $\mathrm{Ker}(i_1,\dots,i_{2k})$ is the partition of $\{1,\dots,2k\}$ such that two elements u and v in $\{1,...,2k\}$ are in the same block (i.e. pair) of $\mathrm{Ker}\,(i_1,\dots,i_{2k})$ if and only if $i_u = i_v$.

Furthermore,

$$\begin{aligned} &\mathbb{E}\left[\left(w^T A^{(1)} w - \mathrm{Tr}\left(A^{(1)}\right) \right) \left(w^T A^{(2)} w - \mathrm{Tr}\left(A^{(2)}\right) \right) \dots \left(w^T A^{(k)} w - \mathrm{Tr}\left(A^{(k)}\right) \right) \right] \\ &\qquad = \sum_{p \in : \boldsymbol{P}_2(2k) :} \sum_{p \leq \mathrm{Ker}(i_1,\dots,i_{2k})} \sum_{i_1,\dots,i_{2k} \in \{1,\dots,P\}} \mathbb{E}\left[A^{(1)}_{i_1 i_2} \dots A^{(k)}_{i_{2k-1} i_{2k}} \right], \end{aligned} \tag{A.4}$$

where $: \boldsymbol{P}_2(2k) :$ is the subset of partitions p in $\boldsymbol{P}_2(2k)$ for which $\{2j-1, 2j\}$ is not a block of p for any $j \in \{1,\dots,k\}$.

Proof. Expanding the left-hand side of Equation (A.3), we obtain:

$$\mathbb{E}\left[\sum_{i_1,\dots,i_{2k} \in \{1,\dots,P\}} w_{i_1} A^{(1)}_{i_1 i_2} w_{i_2} w_{i_3} A^{(2)}_{i_3 i_4} w_{i_4} \dots w_{i_{2k-1}} A^{(k)}_{i_{2k-1} i_{2k}} w_{i_{2k}} \right].$$

Using Wick's formula, we get:

$$\sum_{i_1,\dots,i_{2k} \in \{1,\dots,P\}} \sum_{p \leq \mathrm{Ker}(i_1,\dots,i_{2k})} \sum_{p \in \boldsymbol{P}_2(2k),} \mathbb{E}\left[A^{(1)}_{i_1 i_2} A^{(2)}_{i_3 i_4} \dots A^{(k)}_{i_{2k-1} i_{2k}} \right],$$

hence, interchanging the order of summation, we recover the left-hand side of Equation (A.3):

$$\sum_{p \in \boldsymbol{P}_2(2k)} \sum_{p \leq \mathrm{Ker}(i_1,\dots,i_{2k})} \sum_{i_1,\dots,i_{2k} \in \{1,\dots,P\}} \mathbb{E}\left[A^{(1)}_{i_1 i_2} \dots A^{(k)}_{i_{2k-1} i_{2k}} \right].$$

We now prove Equation (A.4). Expanding the product, the left-hand side is equal to:

$$\sum_{I \subset \{1,\dots,k\}} (-1)^{k - \#I} \mathbb{E}\left[\prod_{i \in I} w^T A^{(i)} w \prod_{i \notin I} \mathrm{Tr}(A^{(i)}) \right].$$

Expanding the product and the trace, and using Wick's equation, we obtain: a

$$\sum_{I\subset\{1,\dots,k\}}(-1)^{k-\#I}\sum_{i_1,\dots,i_{2k}\in\{1,\dots,P\}}\sum_{p\leq \mathrm{Ker}(i_1,\dots,i_{2k})]p\in \boldsymbol{P}_2(2k),p\leq p_I}\mathbb{E}\left[A^{(1)}_{i_1i_2}\dots A^{(k)}_{i_{2k-1}i_{2k}}\right].$$

where p_I is the partition composed of blocks of size 2 given by $\{2l,2l+1\}$ with $l\notin I$ and the rest of the indices contained in a single block. Interchanging the order of summation, we get:

$$\sum_{i_1,\dots,i_{2k}\in\{1,\dots,P\}}\sum_{p\leq \mathrm{Ker}(i_1,\dots,i_{2k})]p\in \boldsymbol{P}_2(2k),}\mathbb{E}\left[A^{(1)}_{i_1i_2}\dots A^{(k)}_{i_{2k-1}i_{2k}}\right]\left[\sum_{p\leq p_I]I\subset\{1,\dots,k\},}(-1)^{k-\#I}\right].$$

Since $\left[\sum_{I\subset\{1,\dots,k\},p\leq p_I}(-1)^{\#I}\right]=\delta_{\{I\subset[k],p\leq p_I\}=\{\{1,\dots,k\}\}}$ and $\{I\subset[k],p\leq p_I\}=\{\{1,\dots,k\}\}$ if and only if $p\in\boldsymbol{P}_2(2k)$:, interchanging a last time the order of summation, we recover the left-hand side of Equation (A.4):

$$\sum_{p\in:\boldsymbol{P}_2(2k):}\sum_{p\leq \mathrm{Ker}(i_1,\dots,i_{2k})]i_1,\dots,i_{2k}\in\{1,\dots,P\}}\mathbb{E}\left[A^{(1)}_{i_1i_2}\dots A^{(k)}_{i_{2k-1}i_{2k}}\right].$$

□

For any $z\in\mathbb{C}\setminus\mathbb{R}^+$, we define the holomorphic function $g_i:\mathbb{C}\setminus\mathbb{R}^+\to\mathbb{C}$ by

$$g_i(z)=\frac{1}{P}w_i^T\left(\frac{1}{P}W_{(i)}K_{(i)}W_{(i)}^T-z\,I_P\right)^{-1}w_i,$$

where $W_{(i)}$ is the $P\times(N-1)$ submatrix of W obtained by removing its i-th column w_i, and $K_{(i)}$ is the $(N-1)\times(N-1)$ submatrix of K obtained by removing both its i-th column and i-th row. In the following lemma, we bound the distance of $g_i(z)$ to its mean. Then we prove that $\mathbb{E}[g_i(z)]$ is close to the expected Stieljes transform of K.

Lemma A.2.3. *The random function $g_i(z)$ satisfies:*

$$\begin{aligned}\left|\mathbb{E}\left[g_i(z)\right]-\mathbb{E}\left[m_P(z)\right]\right| &\leq \frac{\mathbf{c_0}}{P},\\ \mathrm{Var}\left(g_i(z)\right) &\leq \frac{\mathbf{c_1}}{P},\\ \mathbb{E}\left[\left(g_i(z)-\mathbb{E}\left[g_i(z)\right]\right)^4\right] &\leq \frac{\mathbf{c_2}}{P^2},\\ \mathbb{E}\left[\left(g_i(z)-\mathbb{E}\left[g_i(z)\right]\right)^8\right] &\leq \frac{\mathbf{c_3}}{P^4},\end{aligned}$$

where $\mathbf{c_0}$, $\mathbf{c_1}$, $\mathbf{c_2}$, and $\mathbf{c_3}$ depend on γ and z only.

Proof. The random variable w_i is independent from $B_{(i)}(z)=\frac{1}{P}W_{(i)}K_{(i)}W_{(i)}^T-z\mathrm{I}_P$ since the i-th column of W does not appear in the definition of $B_{(i)}(z)$. Using Lemma A.2.2, since there

exists a unique pair partition $p \in \boldsymbol{P}_2(2)$, namely $\{\{1,2\}\}$, the expectation of $g_i(z)$ is given by

$$\mathbb{E}\left[g_i(z)\right] = \frac{1}{P}\mathbb{E}\left[\mathrm{Tr}\left[B_{(i)}(z)^{-1}\right]\right].$$

Recall that $\mathbb{E}[m_P(z)] = \frac{1}{P}\mathbb{E}\left[\mathrm{Tr}\left[B(z)^{-1}\right]\right]$ and $\left|\mathrm{Tr}\left(B(z)^{-1}\right) - \mathrm{Tr}\left(B_{(i)}(z)^{-1}\right)\right| \leq \frac{1}{d(z,\mathbb{R}_+)}$ (from the proof of Lemma A.2.1). Hence

$$\left|\mathbb{E}\left[g_i(z)\right] - \mathbb{E}\left[m_P(z)\right]\right| \leq \frac{1}{P}\mathbb{E}\left[\left|\mathrm{Tr}\left(B(z)^{-1}\right) - \mathrm{Tr}\left(B_{(i)}(z)^{-1}\right)\right|\right] \leq \frac{1}{P}\frac{1}{d(z,\mathbb{R}_+)}.$$

which proves the first assertion with $\mathbf{c_0} = \frac{1}{d(z,\mathbb{R}_+)}$.

Now, let us consider the variance of $g_i(z)$. Using our previous computation of $\mathbb{E}\left[g_i(z)\right]$, we have

$$\mathrm{Var}(g_i(z)) \quad = \quad \mathbb{E}\left[w_i^T \frac{\left(B_{(i)}(z)\right)^{-1}}{P} w_i\, w_i^T \frac{\left(B_{(i)}(z)\right)^{-1}}{P} w_i\right] - \mathbb{E}\left[\frac{1}{P}\mathrm{Tr}\left[B_{(i)}(z)^{-1}\right]\right]^2.$$

The first term can be computed using the first assertion of Lemma A.2.2: there are 2 matrices involved, thus we have to sum over 3 pair partitions. A simplification arises since $\frac{\left(B_{(i)}(z)\right)^{-1}}{P}$ is symmetric: the partition $\{\{1,2\},\{3,4\}\}$ yields $\mathbb{E}\left[\left(\mathrm{Tr}\left[\frac{\left(B_{(i)}(z)\right)^{-1}}{P}\right]\right)^2\right]$ whereas both $\{\{1,3\},\{2,4\}\}$ and $\{\{1,4\},\{2,4\}\}$ yield $\mathbb{E}\left(\mathrm{Tr}\left[\frac{\left(B_{(i)}(z)\right)^{-2}}{P^2}\right]\right)$.

Thus, the variance of $g_i(z)$ is given by:

$$\mathrm{Var}(g_i(z)) = 2\mathbb{E}\left(\mathrm{Tr}\left[\frac{\left(B_{(i)}(z)\right)^{-2}}{P^2}\right]\right) + \mathbb{E}\left[\left(\frac{1}{P}\mathrm{Tr}\left[\left(B_{(i)}(z)\right)^{-1}\right]\right)^2\right] - \mathbb{E}\left[\frac{1}{P}\mathrm{Tr}\left[\left(B_{(i)}(z)\right)^{-1}\right]\right]^2$$

hence is given by a sum of two terms:

$$\mathrm{Var}(g_i(z)) = \frac{2}{P}\mathbb{E}\left(\frac{1}{P}\mathrm{Tr}\left[\left(B_{(i)}(z)\right)^{-2}\right]\right) + \mathrm{Var}\left(\frac{1}{P}\mathrm{Tr}\left[\left(B_{(i)}(z)\right)^{-1}\right]\right).$$

Using the same arguments as those explained for the bound on the Stieltjes transform, the first term is bounded by $\frac{2}{Pd(z,\mathbb{R}_+)^2}$. In order to bound the second term, we apply Lemma A.2.1 for $W_{(i)}$ and $K_{(i)}$ in place of W and K. The second term is bounded by $\frac{\mathbf{c}}{P}$, hence the bound $\mathrm{Var}\left(g_i(z)\right) \leq \frac{\mathbf{c_1}}{P}$.

Finally, we prove the bound on the fourth moment of $g_i(z) - \mathbb{E}\left[g_i(z)\right]$. We denote $m_{(i)}(z) = \frac{1}{P}\mathrm{Tr}\left[\left(B_{(i)}(z)\right)^{-1}\right]$. Recall that $\mathbb{E}\left[g_i(z)\right] = \mathbb{E}\left[m_{(i)}(z)\right]$. Using the convexity of $t \mapsto t^4$, we have

$$\begin{aligned}\mathbb{E}\left[\left(g_i(z) - \mathbb{E}[g_i(z)]\right)^4\right] &= \mathbb{E}\left[\left(g_i(z) - m_{(i)}(z) + m_{(i)}(z) - \mathbb{E}\left[m_{(i)}(z)\right]\right)^4\right] \\ &\leq 8\mathbb{E}\left[\left(g_i(z) - m_{(i)}(z)\right)^4\right] + 8\mathbb{E}\left[\left(m_{(i)}(z) - \mathbb{E}\left[m_{(i)}(z)\right]\right)^4\right].\end{aligned}$$

We bound the second term using the concentration of the Stieljes transform (Lemma A.2.1): it

is bounded by $\frac{8\mathbf{c}}{P^2}$. The first term is bounded using the second assertion of Lemma A.2.2. Using the symmetry of $B_{(i)}(z)$, the partitions in $:\boldsymbol{P}_2(4):$ yield two different terms, namely:

1. $\frac{1}{P^2}\mathbb{E}\left[\left(\frac{1}{P}\mathrm{Tr}\left[\left(B_{(i)}(z)\right)^{-2}\right]\right)^2\right]$, for example if $p=\{\{1,3\},\{2,4\},\{5,7\},\{6,8\}\}$
2. $\frac{1}{P^3}\mathbb{E}\left[\frac{1}{P}\mathrm{Tr}\left[\left(B_{(i)}(z)\right)^{-4}\right]\right]$, for example if $p=\{\{2,3\},\{4,5\},\{6,7\},\{8,1\}\}$.

We bound the two terms using the same arguments as those explained for the bound on the Stieljes transform at the beginning of the section. The first term is bounded by $\frac{d(z,\mathbb{R}^+)^{-4}}{P^2}$ and the second term by $\frac{d(z,\mathbb{R}^+)^{-4}}{P^3}$ hence the bound $\mathbb{E}\left[\left(g_i(z)-\mathbb{E}\left[g_i(z)\right]\right)^4\right]\leq\frac{\mathbf{c_2}}{P^2}$.

The bound $\mathbb{E}[\left(g_i(z)-\mathbb{E}\left[g_i(z)\right]\right)^8]\leq\frac{\mathbf{c_3}}{P^4}$ is obtained in a similar way, using the second assertion of Lemma A.2.2 and simple bounds on the Stieljes transform. □

In the next proposition we show that the Stieltjes transform $m_P(z)$ is close in expectation to the solution of a fixed point equation.

Proposition A.2.4. *For any $z\in\mathbb{H}_{<0}=\{z:\mathrm{Re}(z)<0\}$,*

$$|\mathbb{E}[m_P(z)]-\tilde{m}(z)|\leq\frac{\mathbf{e}}{P},$$

where $\mathbf{e}$ *depends on z, γ, and $\frac{1}{N}\mathrm{Tr}(K)$ only and where $\tilde{m}(z)$ is the unique solution in the cone $\mathscr{C}_z:=\{u-\frac{1}{z}v:u,v\in\mathbb{R}_+\}$ spanned by 1 and $-\frac{1}{z}$ of the equation*

$$\gamma=\frac{1}{N}\sum_{i=1}^{N}\frac{d_i\tilde{m}(z)}{1+d_i\tilde{m}(z)}-\gamma z\tilde{m}(z).$$

Proof. We use the same notation as in the previous proofs, namely $B(z)=\frac{1}{P}WKW^T-z\mathrm{I}_P$, $B_{(i)}(z)=\frac{1}{P}W_{(i)}K_{(i)}W_{(i)}^T-z\mathrm{I}_P$ and $g_i(z)=\frac{1}{P}w_i^T\left(B_{(i)}(z)\right)^{-1}w_i$. Let $\nu_j\geq 0,\ j=1,\dots,P$ be the spectrum of the positive semi-definite matrix $\frac{1}{P}W_{(i)}K_{(i)}W_{(i)}^T$. After diagonalization, we have

$$B_{(i)}(z)^{-1}=O^T\mathrm{diag}(\frac{1}{\nu_1-z},\dots,\frac{1}{\nu_P-z})O,$$

with O an orthogonal matrix. Then

$$g_i(z)=\frac{1}{P}\mathrm{Tr}\left(\left(B_{(i)}(z)\right)^{-1}w_iw_i^T\right)=\frac{1}{P}\sum_{j=1}^{P}\frac{((Ow_i)_{jj})^2}{\nu_j-z}.\tag{A.5}$$

Since $z\in\mathbb{H}_{<0}$, we conclude that $\Re[g_i(z)]\geq 0$ for all $i=1,\dots,P$.

In order to prove the proposition, the key remark is that, since $\mathrm{Tr}\left((\frac{1}{P}WKW^T-z\mathrm{I}_P)(B(z))^{-1}\right)=$

P, the Stieltjes transform $m_P(z)$ satisfies the following equation:

$$P = \mathrm{Tr}\left(\frac{1}{P} K W^T B(z)^{-1} W\right) - zP m_P(z).$$

From the proof of Lemma A.2.1, recall that $B^{-1}(z) = B_{(i)}^{-1}(z) - \frac{d_i}{P} \frac{1}{1+\frac{d_i}{P} w_i^T B_{(i)}^{-1}(z) w_i} B_{(i)}^{-1}(z) w_i w_i^T B_{(i)}^{-1}(z)$, hence:

$$\begin{aligned} \frac{1}{P} w_i^T B^{-1}(z) w_i &= g_i(z) - \frac{d_i g_i(z)^2}{1 + d_i g_i(z)} \\ &= \frac{g_i(z)}{1 + d_i g_i(z)}. \end{aligned} \tag{A.6}$$

Expanding the trace,

$$\mathrm{Tr}\left(\frac{1}{P} K W^T B(z)^{-1} W\right) = \sum_{i=1}^{N} d_i \frac{1}{P} w_i^T B^{-1}(z) w_i = \sum_{i=1}^{N} \frac{d_i g_i(z)}{1 + d_i g_i(z)}.$$

Thus, the Stieljes transform $m_P(z)$ satisfies the following equation $P = \sum_{i=1}^{N} \frac{d_i g_i(z)}{1+d_i g_i(z)} - zP m_P(z)$, or equivalently

$$\gamma = \frac{1}{N} \sum_{i=1}^{N} \frac{d_i g_i(z)}{1 + d_i g_i(z)} - z\gamma m_P(z).$$

Recall that $\gamma > 0$ and $\mathrm{Re}(z) < 0$. The Stieljes transform $m_P(z)$ can be written as a function of $g_i(z)$ for $i = 1, \ldots, n$: $m_P(z) = f(g_1(z), ..., g_N(z))$ where

$$f(g_1, \ldots, g_N) = \frac{1}{\gamma z N} \sum_{i=1}^{N} \frac{d_i g_i}{1 + d_i g_i} - \frac{1}{z} = -\frac{1}{z}\left(1 - \frac{1}{\gamma} + \frac{1}{\gamma} \frac{1}{N} \sum_{i=1}^{N} \frac{1}{1 + d_i g_i}\right).$$

From Lemma A.2.5, the map $f(m) = f(m, ..., m)$ has a unique non-degenerate fixed point $\tilde{m}(z)$ in the cone $\mathscr{C}_z$. We will show that $\mathbb{E}[m_P(z)]$ is close to $\tilde{m}(z)$ using the following two steps: we show a non-tight bound $|\mathbb{E}[m_P(z)] - \tilde{m}(z)| \leq \frac{\mathbf{e}'}{\sqrt{P}}$ and use it to obtain the tighter bound $|\mathbb{E}[m_P(z)] - \tilde{m}(z)| \leq \frac{\mathbf{e}}{P}$.

Let us prove the $\frac{\mathbf{e}'}{\sqrt{P}}$ bound. From Lemma A.2.5, the distance between $m_P(z)$ and the fixed point $\tilde{m}(z)$ of f is bounded by the distance between $f(m_P(z), \ldots, m_P(z))$ and $m_P(z)$. Using the fact that $m_P(z) = f(g_1(z), ..., g_N(z))$, we obtain

$$|\mathbb{E}[m_P(z)] - \tilde{m}(z)| \leq \mathbb{E}[|m_P(z) - \tilde{m}(z)|] \leq \mathbb{E}\left[\left|f(m_P(z), \ldots, m_P(z)) - f(g_1(z), ..., g_N(z))\right|\right].$$

Recall that for any $z \in \mathbb{H}_{<0}$, $\Re(g_i(z)) \geq 0$: we need to study the function f on $\mathbb{H}_{\geq 0}^N$ where

$\mathbb{H}_{\geq 0} = \{z \in \mathbb{C} | \Re(z) \geq 0\}$. On $\mathbb{H}_{\geq 0}^N$, the function f is Lipschitz:

$$\left|\partial_{g_i} f(g_1, .., g_N)\right| = \left|\frac{1}{\gamma z N} \frac{d_i}{(1+d_i g_i)^2}\right| \leq \frac{d_i}{\gamma |z| N}.$$

Thus,

$$\mathbb{E}\left[\left|f\left(m_P(z), ..., m_P(z)\right) - f\left(g_1(z), ..., g_N(z)\right)\right|\right] \leq \sum_{i=1}^{N} \frac{d_i}{\gamma |z| N} \mathbb{E}\left[\left|m_P(z) - g_i(z)\right|\right].$$

Since

$$\mathbb{E}\left[\left|m_P(z) - g_i(z)\right|\right] \leq \mathbb{E}\left[\left|m_P(z) - \mathbb{E}\left[m_P(z)\right]\right|\right] + \left|\mathbb{E}\left[m_P(z)\right] - \mathbb{E}\left[g_i(z)\right]\right| + \mathbb{E}\left[\left|g_i(z) - \mathbb{E}\left[g_i(z)\right]\right|\right],$$

using Lemmas A.2.1 and A.2.3, we get that $\mathbb{E}\left[\left|m_P(z) - g_i(z)\right|\right] \leq \frac{\mathbf{d}}{\sqrt{P}}$, where $\mathbf{d}$ depends on γ and z only. This implies that

$$\mathbb{E}\left[\left|f\left(m_P(z), ..., m_P(z)\right) - f\left(g_1(z), ..., g_N(z)\right)\right|\right] \leq \frac{1}{\sqrt{P}} \frac{\mathbf{d}}{N} \mathrm{Tr}\left(K\right),$$

which allows to conclude that $|\mathbb{E}[m_P(z)] - \tilde{m}(z)| \leq \frac{\mathbf{e}'}{\sqrt{P}}$ where $\mathbf{e}'$ depends on γ, z and $\frac{1}{N}\mathrm{Tr}(K)$ only.

We strengthen this inequality and show the $\frac{\mathbf{e}}{P}$ bound. Using again Lemma A.2.5, we bound the distance between $\mathbb{E}[m_P(z)]$ and the fixed point $\tilde{m}(z)$ by

$$|\mathbb{E}[m_P(z)] - \tilde{m}(z)| \leq \left|\mathbb{E}[f(g_1(z), \ldots, g_N(z))] - f(\mathbb{E}[m_P(z)], \ldots, \mathbb{E}[m_P(z)])\right|$$

and study the r.h.s. using a Taylor approximation of f near $\mathbb{E}[m_P(z)]$. For $i = 1, \ldots, N$ and $m_0 \in \mathbb{H}_{\geq 0}$, let $\mathrm{T}_{m_0} h_i$ be the first order Taylor approximation of the map $h_i : m \mapsto \frac{1}{1+d_i m}$ at a point m_0. The error of the first order Taylor approximation is given by

$$h_i(m) - \mathrm{T}_{m_0} h_i(m) = \frac{1}{1+d_i m} - \left(\frac{1}{1+d_i m_0} - \frac{d_i (m - m_0)}{(1+d_i m_0)^2}\right) = \frac{d_i^2 (m_0 - m)^2}{(1+d_i m)(1+d_i m_0)^2},$$

which, for $m \in \mathbb{H}_{\geq 0}$ can be upper bounded by a quadratic term:

$$\left|h_i(m) - \mathrm{T}_{m_0} h_i(m)\right| = \left|\frac{d_i^2}{(1+d_i m)(1+d_i m_0)^2}\right| |m_0 - m|^2 \leq \frac{1}{|m_0|^2} |m_0 - m|^2. \tag{A.7}$$

The first order Taylor approximation $\mathrm{T}f$ of f at the N-tuple $(\mathbb{E}[m_P(z)], ..., \mathbb{E}[m_P(z)])$ is

$$\mathrm{T}f(g_1, .., g_N) = -\frac{1}{z}\left(1 - \frac{1}{\gamma} + \frac{1}{\gamma}\frac{1}{N}\sum_{i=1}^{N} \mathrm{T}_{\mathbb{E}[m_P(z)]} h_i(g_i)\right).$$

Using this Taylor approximation, $\mathbb{E}[f(g_1(z),\dots,g_N(z))] - f(\mathbb{E}[m_P(z)],\dots,\mathbb{E}[m_P(z)])$ is equal to:

$$\mathbb{E}\left[\mathrm{T}f(g_1(z),..,g_N(z))\right] - f(\mathbb{E}[m_P(z)],\dots,\mathbb{E}[m_P(z)]) + \mathbb{E}\left[f(g_1(z),...,g_N(z)) - \mathrm{T}f(g_1(z),..,g_N(z))\right].$$

Using Lemma A.2.3, we get

$$\begin{aligned}\left|\mathbb{E}\left[f(g_1(z),...,g_N(z)) - \mathrm{T}f(g_1(z),..,g_N(z))\right]\right| &\leq \frac{1}{|z|\gamma}\frac{1}{N}\sum_{i=1}^{N}\frac{1}{|\mathbb{E}[m_P(z)]|^2}\mathbb{E}\left[\left|g_i(z) - \mathbb{E}[m_P(z)]\right|^2\right] \\ &\leq \frac{1}{P}\frac{\alpha}{|\mathbb{E}[m_P(z)]|^2}\end{aligned}$$

and

$$\begin{aligned}\left|\mathbb{E}\left[\mathrm{T}f(g_1(z),..,g_N(z))\right] - f(\mathbb{E}[m_P(z)],...,\mathbb{E}[m_P(z)])\right| &\leq \frac{1}{|z|\gamma}\frac{1}{N}\sum_{i=1}^{N}\frac{d_i\left|\mathbb{E}\left[g_i\right] - \mathbb{E}[m_P(z)]\right|}{|1+d_i\mathbb{E}[m_P(z)]|^2} \\ &\leq \frac{\beta\left(\frac{1}{N}\mathrm{Tr}K\right)}{P}\end{aligned}$$

where α and β depends on z and γ only. From the bounds $|\mathbb{E}[m_P(z)] - \tilde{m}(z)| \leq \frac{\mathbf{e}'}{\sqrt{P}}$ and $|\tilde{m}(z)| \geq (|z| + \frac{1}{N\gamma}\mathrm{Tr}(K))^{-1}$ (Lemma A.2.5), the bound $\frac{1}{P}\frac{\alpha}{|\mathbb{E}[m_P(z)]|^2}$ yields a $\frac{\tilde{\alpha}}{P}$ bound. This implies that $\left|\mathbb{E}[m_P(z)] - f(\mathbb{E}[m_P(z)],\dots,\mathbb{E}[m_P(z)])\right| \leq \frac{\mathbf{e}}{P}$, hence the desired inequality $|\mathbb{E}[m_P(z)] - \tilde{m}(z)| \leq \frac{\mathbf{e}}{P}$. □

For the proof of Proposition A.2.4, we have used the fact that the map f_z introduced therein has a unique non-degenerate fixed point in the cone $\mathscr{C}_z := \{u - \frac{1}{z}v : u, v \in \mathbb{R}_+\}$. We now proceed with proving this statement.

Lemma A.2.5. *Let $d_1,\dots,d_n \geq 0$ and let $\gamma \geq 0$. For any fixed $z \in \mathbb{H}_{<0}$, let $f_z : \mathbb{H}_{\geq 0} \to \mathbb{C}$ be the function $t \mapsto f_z(t) = -\frac{1}{z}\left(1 - \frac{1}{\gamma}\frac{1}{N}\sum_{i=1}^{N}\frac{d_i t}{1+d_i t}\right)$. Let $\mathscr{C}_z := \{u - \frac{1}{z}v : u, v \in \mathbb{R}_+\}$ be the convex region spanned by the half-lines $\mathbb{R}_+$ and $-\frac{1}{z}\mathbb{R}_+$. Then for every $z \in \mathbb{H}_{<0}$ there exists a unique fixed point $\tilde{t}(z) \in \mathscr{C}_z$ such that $\tilde{t}(z) = f_z(\tilde{t}(z))$. The map $\tilde{t} : z \mapsto \tilde{t}(z)$ is holomorphic in $\mathbb{H}_{<0}$ and*

$$|\tilde{t}(z)| \geq \left(|z| + \frac{\sum_i d_i}{\gamma N}\right)^{-1}.$$

Furthermore for every $z \in \mathbb{H}_{<0}$ and any $t \in \mathbb{H}_{\geq 0}$, one has

$$|t - \tilde{t}(z)| \leq |t - f_z(t)|.$$

Proof. By means of Schwarz reflection principle, we can assume that $\Im(z) \geq 0$. Let $z \in \mathbb{H}_{<0}$ and let $\Pi_z := \{-\frac{w}{z} : \Im(w) \leq 0\}$ and let $\mathscr{C}_z$ be the wedged region $\mathscr{C}_z := \Pi_z \cap \{w \in \mathbb{C} : \Im(w) \geq 0\}$. To show the existence of a fixed point in $\mathscr{C}_z$ we show that 0 is in the image of the function $\psi : t \mapsto f_z(t) - t$. Note that since $d_i \geq 0$, the eventual poles of f_z are all strictly negative real numbers, hence $\psi : \mathscr{C}_z \to \mathbb{C}$ is an holomorphic function.

To prove that $0 \in \psi(\mathscr{C}_z)$ we proceed with a geometrical reasoning: the image $\psi(\mathscr{C}_z)$ is (one of) the region of the plane confined by $\psi(\partial\mathscr{C}_z)$, so we only need to "draw" $\psi(\partial\mathscr{C}_z)$ and show that 0 belongs to the "good" connected component confined by it.

The boundary of C_z is made up of two half-lines $\mathbb{R}_+$ and $-\frac{1}{z}\mathbb{R}_+$. Under the map f_z, 0 is mapped to $-\frac{1}{z}$ and ∞ is mapped to $-\frac{1-\frac{1}{\gamma}}{z}$, the two half-lines are hence mapped to paths from $-\frac{1}{z}$ to $-\frac{1-\frac{1}{\gamma}}{z}$. Now under ψ the half-lines will be mapped to paths going $-\frac{1}{z}$ to ∞ because by our assumption $-\frac{1}{z}$ lies in the upper right quadrant, we will show that the image of $\mathbb{R}_+$ under ϕ goes 'above' the origin while the image of $-\frac{1}{z}\mathbb{R}_+$ goes 'under' the origin:

- $\mathbb{R}_+$ is mapped under f_z to the segment $-\frac{1}{z}[1, 1-\frac{1}{\gamma}]$, as a result, its map under ψ lies in the Minkowski sum $-\frac{1}{z}[1, 1-\frac{1}{\gamma}] + (-\mathbb{R}_+)$ which is contained in $\overline{\mathbb{C} \setminus \Pi_z}$.
- For any $t \in -\frac{1}{z}\mathbb{R}_+$ we have for all d_i

$$\Im\left(\frac{d_i t}{1+d_i t}\right) = \Im\left(1 - \frac{1}{1+d_i t}\right) = \Im\left(\frac{1}{1+d_i t}\right) \leq 0,$$

  since $\Im(t) \geq 0$. As a result the image of $-\frac{1}{z}\mathbb{R}_+$ under f_z lies in Π_z and its image under ψ lies in the Minkovski sum $\Pi_z + (-\frac{1}{z}\mathbb{R}_+) = \Pi_z$.

Thus we can conclude that $0 \in \psi(\mathscr{C}_z)$, which shows that there exists at least a fixed point $\tilde{m}$ in $\mathscr{C}_z$.

We observe that, for every $t \in \mathscr{C}_z$, the derivative of f has negative real part:

$$\begin{aligned}
\mathrm{Re}\left(f_z'(t)\right) &= \frac{1}{\gamma}\frac{1}{N}\sum_{i=1}^{N} \mathrm{Re}\left(\frac{d_i}{z(1+d_i t)^2}\right) \\
&= \frac{1}{\gamma}\frac{1}{N}\sum_{i=1}^{N} \frac{d_i\left[\Re(z) + 2d_i\Re(z)\Re(t) - 2d_i\Im(z)\Im(t) + d_i^2\Re(zt^2)\right]}{|z|^2\,|1+d_i t|^4} \leq 0,
\end{aligned}$$

where we concluded the last inequality by using that $\Re(z) \leq 0$, $\Re(t) \geq 0$, $\Im(z)\Im(t) \geq 0$ and $\Re(zt^2) \leq 0$. Thus, since for no point $t \in \mathscr{C}_z$ has $f_z'(t) = 1$, any fixed point of f_z is a simple fixed point.

We now proceed to show the uniqueness of the fixed point in the region $\mathscr{C}_z$. Suppose there are two fixed points t_1 and t_2, then

$$\begin{aligned}
t_1 - t_2 &= f_z(t_1) - f_z(t_2) \\
&= (t_1 - t_2)\,\frac{1}{z}\frac{1}{\gamma N}\sum_{i=1}^{N} \frac{d_i}{(1+d_i t_1)(1+d_i t_2)}.
\end{aligned}$$

Again, since $\Re(z) \leq 0$, $\Re(t_1), \Re(t_2) \geq 0$, $\Im(z)\Im(t_1), \Im(z)\Im(t_2), \geq 0$ and $\Re(z t_1 t_2) \leq 0$, the factor $\frac{1}{z}\frac{1}{N}\sum_{i=1}^{N} \frac{d_i}{(1+d_i t_1)(1+d_i t_2)}$ has negative real part, and thus the identity is possible only if $t_1 = t_2$.

Let's then $\tilde{t}(z)$ be the only fixed point in $\mathscr{C}_z$.

We proceed now to show that $|t - f_z(t)| \geq |t - \tilde{t}(z)|$, i.e. if t and its image are close, then t is not too far from being a fixed point, and so it is close to $\tilde{t}(z)$.

For any $t \in \mathscr{C}_z$, we have

$$\begin{aligned}
|t - f_z(t)| &= |t - \tilde{t}(z) + f_z(\tilde{t}(z)) - \tilde{f}_z(t)| \\
&= \left| (t - \tilde{t}(z)) - \left(t - \tilde{t}(z)\right) \left(\frac{1}{z} \frac{1}{\gamma N} \sum_{i=1}^{N} \frac{d_i}{(1 + d_i t)(1 + d_i \tilde{t}(z))} \right) \right| \\
&= \left| t - \tilde{t}(z) \right| \left| 1 - \frac{1}{z} \frac{1}{\gamma N} \sum_{i=1}^{N} \frac{d_i}{(1 + d_i t)(1 + d_i \tilde{t}(z))} \right| \\
&\geq \left| t - \tilde{t}(z) \right|
\end{aligned}$$

where we have used again that $\frac{1}{z}\frac{1}{N}\sum_{i=1}^{N} \frac{d_i}{(1+d_i t)(1+d_i \tilde{t}(z))}$ has negative real part.

We provide a lower bound on the norm of the fixed point:

$$\left| \tilde{t}(z) \right| = \frac{1}{|z|} \left| 1 - \frac{1}{\gamma} \frac{1}{N} \sum_{i=1}^{N} \frac{d_i \tilde{t}(z)}{1 + d_i \tilde{t}(z)} \right| \geq \frac{1}{|z|} \left(1 - \frac{1}{\gamma} \frac{1}{N} \sum_{i=1}^{N} \left| \frac{d_i \tilde{t}(z)}{1 + d_i \tilde{t}(z)} \right| \right) \geq \frac{1}{|z|} \left(1 - \frac{\left| \tilde{t}(z) \right|}{\gamma N} \sum_{i=1}^{N} d_i \right).$$

hence

$$|\tilde{t}(z)| \geq \left(|z| + \frac{\sum_i d_i}{\gamma N} \right)^{-1}.$$

Finally, note that z can be expressed from the fixed point $\tilde{m}$, hence defining an inverse for the map $\tilde{t}$:

$$\tilde{t}^{-1}(\tilde{m}) = z = -\frac{1}{\tilde{m}} \left(1 - \frac{1}{\gamma} \frac{1}{N} \sum_{i=1}^{N} \frac{d_i \tilde{m}}{1 + d_i \tilde{m}} \right)$$

because the inverse is holomorphic, so is $\tilde{t}$. □

A.3 Expectation of the Predictor

The optimal parameters $\hat{\theta}$ which minimize the regularized MSE loss is given by $\hat{\theta} = F^T(FF^T + \lambda \mathrm{I}_N)^{-1} y$, or equivalently by $\hat{\theta} = (F^T F + \lambda)^{-1} F^T y$. Thus, the final labels take the form $\hat{y} = A(-\lambda) y$ where $A(z)$ is the random matrix defined as

$$\begin{aligned}
A(z) &:= F \left(F^T F - z \mathrm{I}_P \right)^{-1} F^T \\
&= \frac{1}{P} K^{\frac{1}{2}} W^T \left(\frac{1}{P} W K W^T - z \mathrm{I}_P \right)^{-1} W K^{\frac{1}{2}}.
\end{aligned}$$

Note that the matrix A_λ defined in the proof sketch of Theorem 4.1 in the main text is given by $A_\lambda = A(-\lambda)$.

Proposition A.3.1. *For any $\gamma > 0$, any $z \in \mathbb{H}_{<0}$, and any symmetric positive definite matrix K,*

$$\|\mathbb{E}[A(z)] - K(K + \tilde{\lambda}(-z) I_N)^{-1}\|_{op} \leq \frac{c}{P}, \tag{A.8}$$

where $\tilde{\lambda}(z) := \frac{1}{\tilde{m}(-z)}$ and $c > 0$ depends on z, γ and $\frac{1}{N} Tr(K)$ only.

Proof. Since the distribution of W is invariant under orthogonal transformations, by applying a change of basis, in order to prove Inequality (A.8), we may assume that K is diagonal with diagonal entries $d_1, \dots, d_N$. Denoting $w_1, \dots, w_N$ the columns of W, for any $i, j = 1, \dots, N$,

$$(A(z))_{ij} = \frac{1}{P} \sqrt{d_i d_j}\, w_i^T \left(\frac{1}{P} W K W^T - z I_P \right)^{-1} w_j,$$

where $WKW^T = \sum_{i=1}^N d_i w_i w_i^T$. Replacing w_i by $-w_i$ does not change the law W hence does not change the law of $(A(z))_{ij}$. Since WKW^T is invariant under this change of sign, we get that for $i \neq j$, $\mathbb{E}\left[(A(z))_{ij}\right] = -\mathbb{E}\left[(A(z))_{ij}\right]$, hence the off-diagonal terms of $\mathbb{E}[A(z)]$ vanish.

Consider a diagonal term $(A(z))_{ii}$. From Equation (A.6), we get

$$(A(z))_{ii} = \frac{d_i}{P} w_i^T B^{-1}(z) w_i = \frac{d_i g_i(z)}{1 + d_i g_i(z)}. \tag{A.9}$$

By Lemma A.2.3, g_i lies close to $m_P(z)$ which itself is approximatively equal to $\tilde{m}(z)$ by Proposition A.2.4. Therefore, we expect $\mathbb{E}[(A(z))_{ii}] = \mathbb{E}\left[\frac{d_i g_i}{1+d_i g_i}\right]$ to be at short distance from $\frac{d_i \tilde{m}(z)}{1+d_i \tilde{m}(z)}$.

In order to make rigorous this heuristic and to prove that $\mathbb{E}[(A(z))_{ii}]$ is within $\mathcal{O}(\frac{1}{P})$ distance to $\frac{d_i \tilde{m}(z)}{1+d_i \tilde{m}(z)}$, we consider the first order Taylor approximation $\mathrm{T}_{\tilde{m}(z)} h_i$ of the map $h_i : g \mapsto \frac{1}{1+d_i g}$ (as in the proof Proposition A.2.4 but this time centered at $\tilde{m}(z)$). Using the fact that $\frac{d_i t}{1+d_i t} = 1 - \frac{1}{1+d_i t} = 1 - h_i(t)$, and inserting the Taylor approximation, $\mathbb{E}[(A(z))_{ii}] - \frac{d_i \tilde{m}(z)}{1+d_i \tilde{m}(z)}$ is equal to:

$$h_i(\tilde{m}(z)) - h_i(g_i(z)) = \frac{1}{1 + d_i \tilde{m}(z)} - \mathbb{E}\left[\mathrm{T}_{\tilde{m}(z)} h(g_i(z))\right] + \mathbb{E}\left[\mathrm{T}_{\tilde{m}(z)} h(g_i(z)) - h(g_i(z))\right].$$

Thus,

$$\left| \mathbb{E}[(A(z))_{ii}] - \frac{d_i \tilde{m}(z)}{1 + d_i \tilde{m}(z)} \right| \leq \left| \frac{1}{1 + d_i \tilde{m}(z)} - \mathbb{E}\left[\mathrm{T}_{\tilde{m}(z)} h(g_i(z))\right] \right| + \left| \mathbb{E}\left[\mathrm{T}_{\tilde{m}(z)} h(g_i(z)) - h(g_i(z))\right] \right|.$$

Using Lemma A.2.3 and Proposition A.2.4, the first term $\left| \frac{1}{1+d_i \tilde{m}(z)} - \mathbb{E}\left[\mathrm{T}_{\tilde{m}(z)} h(g_i(z))\right] \right| = \frac{d_i |\mathbb{E}[g_i(z)] - \tilde{m}(z)|}{|1+d_i \tilde{m}(z)|^2}$ can be bounded by $\frac{\delta}{P} \frac{d_i}{|1+d_i \tilde{m}(z)|^2}$ where δ depends on z, γ and $\frac{1}{N}\mathrm{Tr}(K)$ only. Since $\mathrm{Re}[\tilde{m}(z)] \geq 0$ thus $|1 + d_i \tilde{m}(z)| \geq \max(1, |d_i \tilde{m}(z)|)$, and $|\tilde{m}(z)| \geq \frac{1}{|z| + \frac{1}{\gamma} \frac{1}{N} \mathrm{Tr} K}$ (Lemma A.2.5), the denominator

can be lower bounded:

$$|1+d_i\tilde{m}(z)|^2 \geq |d_i\tilde{m}(z)| \geq \frac{d_i}{|z|+\frac{1}{\gamma}\frac{1}{N}\mathrm{Tr}K},$$

yielding the upper bound:

$$\left|\frac{1}{1+d_i\tilde{m}(z)} - \mathbb{E}\left[\mathrm{T}_{\tilde{m}(z)}h(g_i(z))\right]\right| \leq \frac{1}{P}\delta\left[|z|+\frac{1}{\gamma}\frac{1}{N}\mathrm{Tr}K\right].$$

For the second term, using the same arguments as for the proof of Proposition A.2.4, we have:

$$\left|\mathbb{E}\left[\mathrm{T}_{\tilde{m}(z)}h(g_i(z)) - h(g_i(z))\right]\right| \leq \frac{\mathbb{E}\left[\left|\tilde{m}(z)-g_i(z)\right|^2\right]}{|\tilde{m}(z)|^2}.$$

Recall that $|\tilde{m}(z)| \geq \frac{1}{|z|+\frac{1}{\gamma}\frac{1}{N}\mathrm{Tr}K}$ and that, by Lemma A.2.3 and Proposition A.2.1, $\mathbb{E}\left[\left|\tilde{m}(z)-g_i(z)\right|^2\right] \leq \frac{\tilde{\delta}}{P}$ where $\tilde{\delta}$ depends on z,γ and $\frac{1}{N}\mathrm{Tr}(K)$ only. This implies that

$$\left|\mathbb{E}\left[\mathrm{T}_{\tilde{m}(z)}h(g_i(z)) - h(g_i(z))\right]\right| \leq \frac{\tilde{\delta}}{P}\left[|z|+\frac{1}{\gamma}\frac{1}{N}\mathrm{Tr}K\right]^2.$$

As a consequence, there exists a constant c which depends on z,γ and $\frac{1}{N}\mathrm{Tr}(K)$ only such that:

$$\left|\mathbb{E}\left[(A(z))_{ii}\right] - \frac{d_i\tilde{m}(z)}{1+d_i\tilde{m}(z)}\right| \leq \frac{c}{P}.$$

Using the effective ridge $\tilde{\lambda}(z) := \frac{1}{\tilde{m}(-z)}$, the term $\frac{d_i\tilde{m}(z)}{1+d_i\tilde{m}(z)} = \frac{d_i}{d_i+\tilde{\lambda}(-z)}$ is equal to $(K(K+\tilde{\lambda}I_N)^{-1})_{ii}$ since, in the basis considered, $K(K+\tilde{\lambda}I_N)^{-1}$ is a diagonal matrix. Hence, we obtain:

$$\left\|\mathbb{E}\left[A(z)\right] - K(K+\tilde{\lambda}I_N)^{-1}\right\|_{op} \leq \frac{c}{P}$$

which allows us to conclude. □

Using the above proposition, we can bound the distance between the expected λ-RF predictor and the $\tilde{\lambda}$-RF predictor.

Theorem A.3.2. *For $N,P>0$ and $\lambda>0$, we have*

$$\left|\mathbb{E}[\hat{f}^{(RF)}_{\lambda,\gamma}(x)] - \hat{f}^{(K)}_{\lambda}(x)\right| \leq \frac{c\sqrt{K(x,x)}\|y\|_{K^{-1}}}{P} \tag{A.10}$$

where the effective ridge $\tilde{\lambda}(\lambda,\gamma)>\lambda$ is the unique positive number satisfying

$$\tilde{\lambda} = \lambda + \frac{\tilde{\lambda}}{\gamma}\frac{1}{N}\sum_{i=1}^{N}\frac{d_i}{\tilde{\lambda}+d_i}, \tag{A.11}$$

and where $c > 0$ depends on λ, γ, and $\frac{1}{N}\mathrm{Tr}K(X,X)$ only.

Proof. Recall that $\tilde{m}(-\lambda)$ is the unique non negative real such that $\gamma = \frac{1}{N}\sum_{i=1}^{N} \frac{d_i \tilde{m}(-\lambda)}{1+d_i\tilde{m}(-\lambda)} + \gamma\lambda\tilde{m}(-\lambda)$. Dividing this equality by $\gamma\tilde{m}(-\lambda)$ yields Equation (A.11). From now on, let $\tilde{\lambda} = \tilde{\lambda}(\lambda,\gamma)$.

We now bound the l.h.s. of Equation (A.10). By Proposition A.1.1, since $\hat{y} = A(-\lambda)y$, the average λ-RF predictor is $\mathbb{E}\left[f^{(RF)}_{\lambda,\gamma}(x)\right] = K(x,X)K^{-1}\mathbb{E}[A(-\lambda)]\,y$. The $\tilde{\lambda}$-KRR predictor is $f^{(K)}_{\tilde{\lambda}}(x) = K(x,X)\left(K+\tilde{\lambda}I_N\right)^{-1}y$. Thus:

$$\left|\mathbb{E}[f^{(RF)}_{\lambda,\gamma}(x)] - f^{(K)}_{\tilde{\lambda}}(x)\right| = \left|K(x,X)K^{-1}\left[\mathbb{E}[A(-\lambda)] - K\left(K+\tilde{\lambda}I_N\right)^{-1}\right]y\right|.$$

The r.h.s. can be expressed as the absolute value of the scalar product $\left|\langle w, v\rangle_{K^{-1}}\right| = \left|v^T K^{-1} w\right|$ where $v = K(x,X)$ and $w = [\mathbb{E}[A(-\lambda)] - K(K+\tilde{\lambda}I_N)^{-1}]y$. By Cauchy-Schwarz inequality, $\left|\langle v, w\rangle_{K^{-1}}\right| \le \|v\|_{K^{-1}}\|w\|_{K^{-1}}$.

For a general vector v, the K^{-1}-norm $\|v\|_{K^{-1}}$ is equal to the norm mininum Hilbert norm (for the RKHS associated to the kernel K) interpolating function:

$$\|v\|_{K^{-1}} = \min_{f\in\mathscr{H}, f(x_i)=v_i} \|f\|_{\mathscr{H}}.$$

Indeed the minimal interpolating function is the kernel regression given by $f^{(K)}(\cdot) = K(\cdot,X)K(X,X)^{-1}v$ which has norm (writing $\beta = K^{-1}v$):

$$\left\|f^{(K)}\right\|_{\mathscr{H}} = \left\|\sum_{i=1}^{N}\beta_i K(\cdot,x_i)\right\|_{\mathscr{H}} = \sqrt{\sum_{i,j=1}^{N}\beta_i\beta_j K(x_i,x_j)} = \sqrt{v^T K^{-1}KK^{-1}v} = \|v\|_{K^{-1}}.$$

We can now bound the two norms $\|v\|_{K^{-1}}$ and $\|w\|_{K^{-1}}$. For $v = K(x,X)$, we have

$$\|v\|_{K^{-1}} = \min_{f\in\mathscr{H}, f(x_i)=v_i} \|f\|_{\mathscr{H}} \le \|K(x,\cdot)\|_{\mathscr{H}} = K(x,x)^{\frac{1}{2}}. \tag{A.12}$$

since $K(x,\cdot)$ is an interpolating function for v.

It remains to bound $\|w\|_{K^{-1}}$. Recall that $K = UDU^T$ with D diagonal, and that, from the previous proposition, $\mathbb{E}[A(-\lambda)] = UD_AU^T$ where $D_A = \mathrm{diag}\left(\frac{d_1 g_1(-\lambda)}{1+d_1 g_1(-\lambda)},\dots,\frac{d_N g_N(-\lambda)}{1+d_N g_N(-\lambda)}\right)$. The norm $\|w\|_{K^{-1}}$ is equal to

$$\sqrt{\tilde{y}^T\left[D_A - D\left(D+\tilde{\lambda}(\lambda)I_N\right)^{-1}\right]^T D^{-1}\left[D_A - D\left(D+\tilde{\lambda}(\lambda)I_N\right)^{-1}\right]\tilde{y}},$$

where $\tilde{y} = U^T y$. Expanding the product, $\|w\|_{K^{-1}} = \sqrt{\sum_{i=1}^{N}\frac{\tilde{y}_i^2}{d_i}\left((D_A)_{ii} - \frac{d_i}{\tilde{\lambda}(\lambda)+d_i}\right)^2}$, hence by Proposition A.3.1, $\|w\|_{K^{-1}} \le \frac{c}{P}\sqrt{\sum_{i=1}^{N}\frac{\tilde{y}^2}{d_i}}$. The result follows from noticing that $\sum_{i=1}^{N}\frac{\tilde{y}^2}{d_i} =$

$\tilde{y}^T D^{-1} \tilde{y} = \|y\|^2_{K^{-1}}$:

$$\left|\mathbb{E}[f^{(RF)}_{\lambda,\gamma}(x)] - f^{(K)}_{\tilde{\lambda}}(x)\right| \leq \|v\|_{K^{-1}} \|w\|_{K^{-1}} \leq \frac{cK(x,x)^{\frac{1}{2}} \|y\|_{K^{-1}}}{P}.$$

which allows us to conclude. □

Corollary A.3.3. *If $\mathbb{E}_{\mathscr{D}}[K(x,x)] < \infty$, we have that the difference of errors $\delta_E = \left|L(\mathbb{E}[\hat{f}^{(RF)}_{\lambda,\gamma}]) - L(\hat{f}^{(K)}_{\tilde{\lambda}})\right|$ is bounded from above by*

$$\delta_E \leq \frac{C\|y\|_{K^{-1}}}{P}\left(2\sqrt{L\left(\hat{f}^{(K)}_{\tilde{\lambda}}\right)} + \frac{C\|y\|_{K^{-1}}}{P}\right),$$

where C is given by $c\sqrt{\mathbb{E}_{\mathscr{D}}[K(x,x)]}$, with c the constant appearing in (A.10) above.

Proof. For any function $f : \mathbb{R}^d \to \mathbb{R}$, we denote by $\|f\| = (\mathbb{E}_{\mathscr{D}}\left[f(x)^2\right])^{\frac{1}{2}}$ its $L^2(\mathscr{D})$-norm. Integrating $\left|\mathbb{E}[f^{(RF)}_{\lambda,\gamma}(x)] - f^{(K)}_{\tilde{\lambda}}(x)\right|^2 \leq \frac{c^2 K(x,x)\|y\|^2_{K^{-1}}}{P^2}$ over $x \sim \mathscr{D}$, we get the following bound:

$$\|\mathbb{E}[f^{(RF)}_{\lambda,\gamma}] - f^{(K)}_{\tilde{\lambda}}\| \leq \frac{c\left[\mathbb{E}_{\mathscr{D}}\left[K(x,x)\right]\right]^{\frac{1}{2}} \|y\|_{K^{-1}}}{P}.$$

Hence, if f^* is the true function, by the triangular inequality,

$$\left|\|\mathbb{E}[f^{(RF)}_{\lambda,\gamma}] - f^*\| - \|f^{(K)}_{\tilde{\lambda}} - f^*\|\right| \leq \frac{c\left[\mathbb{E}_{\mathscr{D}}\left[K(x,x)\right]\right]^{\frac{1}{2}} \|y\|_{K^{-1}}}{P}.$$

Notice that $L(\mathbb{E}[\hat{f}^{(RF)}_{\gamma,\lambda}]) = \|\mathbb{E}[f^{(RF)}_{\lambda,\gamma}] - f^*\|^2$ and $L(\hat{f}^{(K)}_{\tilde{\lambda}}) = \|f^{(K)}_{\tilde{\lambda}} - f^*\|^2$. Since $\left|a^2 - b^2\right| \leq |a-b|\,(|a-b| + 2|b|)$, we obtain

$$\left|L\left(\mathbb{E}[\hat{f}^{(RF)}_{\gamma,\lambda}]\right) - L\left(\hat{f}^{(K)}_{\tilde{\lambda}}\right)\right| \leq \frac{c\left[\mathbb{E}_{\mathscr{D}}\left[K(x,x)\right]\right]^{\frac{1}{2}} \|y\|_{K^{-1}}}{P}\left(2\sqrt{L\left(\hat{f}^{(K)}_{\tilde{\lambda}}\right)} + \frac{c\left[\mathbb{E}_{\mathscr{D}}\left[K(x,x)\right]\right]^{\frac{1}{2}} \|y\|_{K^{-1}}}{P}\right),$$

which allows us to conclude. □

A.4 Properties of the Effective Ridge

Thanks to the implicit definition of the effective ridge $\tilde{\lambda}$, we obtain the following:

Proposition A.4.1. *The effective ridge $\tilde{\lambda}$ satisfies the following properties:*

1. *for any $\gamma > 0$, we have $\lambda < \tilde{\lambda}(\lambda,\gamma) \leq \lambda + \frac{1}{\gamma}T$;*
2. *the function $\gamma \mapsto \tilde{\lambda}(\lambda,\gamma)$ is decreasing;*
3. *for $\gamma > 1$, we have $\tilde{\lambda} \leq \frac{\gamma}{\gamma-1}\lambda$;*

4. *for $\gamma < 1$, we have $\tilde{\lambda} \geq \frac{1-\sqrt{\gamma}}{\sqrt{\gamma}} \min_i d_i$.*

Proof. **(1)** The upper bound in the first statement follows directly from Lemma A.2.5 where it was shown that $\tilde{m}(-\lambda) \geq \frac{1}{\lambda + \frac{1}{\gamma}\frac{1}{N}\mathrm{Tr}K}$ and from the fact that $\tilde{\lambda}(\lambda,\gamma) = \frac{1}{\tilde{m}(-\lambda)}$. For the lower bound, remark that Equation (A.11) can be written as:

$$\tilde{\lambda}(\lambda,\gamma) = \lambda + \frac{1}{\gamma}\frac{1}{N}\mathrm{Tr}[\tilde{\lambda}(\lambda,\gamma)K(\tilde{\lambda}(\lambda,\gamma)I_N + K)^{-1}].$$

Since $\tilde{\lambda}(\lambda,\gamma) \geq 0$ and K is a positive symmetric matrix, $\mathrm{Tr}[K[\tilde{\lambda}(\lambda,\gamma)I_N + K]^{-1}] \geq 0$: this yields $\tilde{\lambda}(\lambda,\gamma) \geq \lambda$.

(2) We show that $\gamma \mapsto \tilde{\lambda}(\lambda,\gamma)$ is decreasing by computing the derivative of the effective ridge with respect to γ. Differentiating both sides of Equation (A.11), $\partial_\gamma \tilde{\lambda} = \partial_\gamma \left[\lambda + \frac{\tilde{\lambda}}{\gamma}\frac{1}{N}\sum_{i=1}^N \frac{d_i}{\tilde{\lambda}+d_i}\right]$. The r.h.s. is equal to:

$$\frac{\partial_\gamma \tilde{\lambda}}{\gamma}\frac{1}{N}\sum_{i=1}^N \frac{d_i}{\tilde{\lambda}+d_i} - \frac{\tilde{\lambda}}{\gamma^2}\frac{1}{N}\sum_{i=1}^N \frac{d_i}{\tilde{\lambda}+d_i} - \frac{\tilde{\lambda}}{\gamma}\frac{1}{N}\sum_{i=1}^N \frac{d_i \partial_\gamma \tilde{\lambda}}{(\tilde{\lambda}+d_i)^2}.$$

Using Equation (A.11), $\frac{1}{\gamma}\frac{1}{N}\sum_{i=1}^N \frac{d_i}{\tilde{\lambda}+d_i} = \frac{\tilde{\lambda}-\lambda}{\tilde{\lambda}}$ and thus:

$$\partial_\gamma \tilde{\lambda}\left[\frac{\lambda}{\tilde{\lambda}} + \frac{\tilde{\lambda}}{\gamma}\frac{1}{N}\sum_{i=1}^N \frac{d_i}{\left(\tilde{\lambda}+d_i\right)^2}\right] = -\frac{\tilde{\lambda}-\lambda}{\gamma}.$$

Since $\tilde{\lambda} \geq \lambda \geq 0$, the derivative of the effective ridge with respect to γ is negative: the function $\gamma \mapsto \tilde{\lambda}(\lambda,\gamma)$ is decreasing.

(3) Using the bound $\frac{d_i}{\tilde{\lambda}+d_i} \leq 1$ in Equation (A.11), we obtain $\tilde{\lambda} \leq \lambda + \frac{\tilde{\lambda}}{\gamma}$ which, when $\gamma \geq 1$, implies that $\tilde{\lambda} \leq \lambda\frac{\gamma}{\gamma-1}$.

(4) Recall that $\lambda > 0$ and that the effective ridge $\tilde{\lambda}$ is the unique fixpoint of the map $f(t) = \lambda + \frac{t}{\gamma}\frac{1}{N}\sum_i \frac{d_i}{t+d_i}$ in $\mathbb{R}_+$. The map is concave and, at $t = 0$, we have $f(t) = \lambda > 0 = t$: this implies that $f'(\tilde{\lambda}) < 1$ otherwise by concavity, for any $t \leq \tilde{\lambda}$ one would have $f(t) \leq t$. The derivative of f is $f'(t) = \frac{1}{\gamma}\frac{1}{N}\sum_{i=1}^N \frac{d_i^2}{(t+d_i)^2}$, thus $\frac{1}{\gamma}\frac{1}{N}\sum_{i=1}^N \frac{d_i^2}{(\tilde{\lambda}+d_i)^2} < 1$. Using the fact that d_0 is the smallest eigenvalue of $K(X,X)$, i.e. $d_i \geq d_0$, we get $1 > \frac{1}{\gamma}\frac{d_0^2}{(\tilde{\lambda}+d_0)^2}$ hence $\tilde{\lambda} \geq d_0\frac{1-\sqrt{\gamma}}{\sqrt{\gamma}}$. □

Similarily, we gather a number of properties of the derivative $\partial_\lambda \tilde{\lambda}(\lambda,\gamma)$.

Proposition A.4.2. *For $\gamma > 1$, as $\lambda \to 0$, the derivative $\partial_\lambda \tilde{\lambda}$ converges to $\frac{\gamma}{\gamma-1}$. As $\lambda\gamma \to \infty$, we have $\partial_\lambda \tilde{\lambda}(\lambda,\gamma) \to 1$.*

Proof. Differentiating both sides of Equation (A.11),

$$\partial_\lambda \tilde{\lambda} = 1 + \partial_\lambda \tilde{\lambda} \frac{1}{\gamma} \frac{1}{N} \sum_{i=1}^{N} \frac{d_i}{\tilde{\lambda} + d_i} - \tilde{\lambda} \partial_\lambda \tilde{\lambda} \frac{1}{\gamma} \frac{1}{N} \sum_{i=1}^{N} \frac{d_i}{(\tilde{\lambda} + d_i)^2}.$$

Hence the derivative $\partial_\lambda \tilde{\lambda}$ satisfies the following equality

$$\partial_\lambda \tilde{\lambda} \left(1 - \frac{1}{\gamma} \frac{1}{N} \sum_{i=1}^{N} \frac{d_i}{\tilde{\lambda} + d_i} + \tilde{\lambda} \frac{1}{\gamma} \frac{1}{N} \sum_{i=1}^{N} \frac{d_i}{(\tilde{\lambda} + d_i)^2} \right) = 1. \tag{A.13}$$

(1) Assuming $\gamma > 1$, from the point 3. of Proposition A.4.1, we already know that $\tilde{\lambda}(\lambda, \gamma) \leq \lambda \frac{\gamma}{\gamma - 1}$ hence $\tilde{\lambda}(0, \gamma) = 0$. Actually, using similar arguments as in the proof of point 3., this holds also for $\gamma = 1$. Using the fact that $\tilde{\lambda}(0, \gamma) = 0$, we get $\partial_\lambda \tilde{\lambda}(0, \gamma) = 1 + \frac{\partial_\lambda \tilde{\lambda}(0, \gamma)}{\gamma}$, hence $\partial_\lambda \tilde{\lambda}(0, \gamma) = \frac{\gamma}{\gamma - 1}$.

(2) From the first point of Proposition A.4.1, $\tilde{\lambda} \sim \lambda$ as $\lambda\gamma \to \infty$. Since Equation (A.13) can be expressed as:

$$\partial_\lambda \tilde{\lambda} \left(1 - \frac{1}{\gamma\lambda} \frac{1}{N} \sum_{i=1}^{N} \frac{d_i}{\frac{\tilde{\lambda}}{\lambda} + d_i} + \frac{1}{\gamma\lambda} \frac{\tilde{\lambda}}{\lambda} \frac{1}{N} \sum_{i=1}^{N} \frac{d_i}{(\frac{\tilde{\lambda}}{\lambda} + d_i)^2} \right) = 1,$$

we obtain that $\partial_\lambda \tilde{\lambda} \to 1$ as $\lambda \to \infty$. □

A.5 Variance of the Predictor

By the bias-variance decomposition, in order to bound the difference between $\mathbb{E}[L(\hat{f}^{(RF)}_{\gamma,\lambda})]$ and $L(\hat{f}^{(K)}_{\tilde{\lambda}})$, we have to bound $\mathbb{E}_{\mathscr{D}}[\mathrm{Var}(f(x))]$. The law of total variance yields $\mathrm{Var}(\hat{f}(x)) = \mathrm{Var}(\mathbb{E}[\hat{f}(x)|F]) + \mathbb{E}[\mathrm{Var}[\hat{f}(x)|F]]$. By Proposition A.1.1, we have $\mathbb{E}[\hat{f}(x)|F] = K(x,X)K(X,X)^{-1}\hat{y}$ and $\mathrm{Var}[\hat{f}(x)|F] = \frac{1}{P}\|\hat{\theta}\|^2 \tilde{K}(x,x)$. Hence, it remains to study $\mathrm{Var}\left(K(x,X)K(X,X)^{-1}\hat{y}\right)$ and $\mathbb{E}[\|\hat{\theta}\|^2]$. Recall that we denote $T = \frac{1}{N}\mathrm{Tr}K(X,X)$.

This section is dedicated to the proof of the variance bound of Theorem 5.1 of the paper:

Theorem 5.1 *There are constants $c_1, c_2 > 0$ depending on λ, γ, T only such that*

$$\mathrm{Var}\left(K(x,X)K(X,X)^{-1}\hat{y}\right) \leq \frac{c_1 K(x,x)\|y\|^2_{K^{-1}}}{P}$$
$$\left|\mathbb{E}\|[\hat{\theta}\|^2] - \partial_\lambda \tilde{\lambda} y^T M_{\tilde{\lambda}} y\right| \leq \frac{c_2 \|y\|^2_{K^{-1}}}{P},$$

where $\partial_\lambda \tilde{\lambda}$ is the derivative of $\tilde{\lambda}$ with respect to λ and for $M_{\tilde{\lambda}} = K(X,X)(K(X,X) + \tilde{\lambda} I_N)^{-2}$. As a result

$$\mathrm{Var}\left(\hat{f}^{(RF)}_{\lambda}(x)\right) \leq \frac{c_3 K(x,x)\|y\|^2_{K^{-1}}}{P},$$

where $c_3 > 0$ depends on λ, γ, T.

• **Bound on** $\mathrm{Var}\left(K(x,X)K(X,X)^{-1}\hat{y}\right)$**.** We first study the covariance of the entries of the matrix

$$A_\lambda = \frac{1}{P}K^{\frac{1}{2}}W^T\left(\frac{1}{P}WKW^T + \lambda \mathrm{I}_P\right)^{-1}WK^{\frac{1}{2}},$$

where $K = \mathrm{diag}(d_1,\dots,d_N)$ is a positive definite diagonal matrix and W is a $P\times N$ matrix with i.i.d. Gaussian entries. In the next proposition we show a $\frac{c_1}{P}$ bound for the covariance of the entries of A_λ, then we exploit this result in order to prove the bound on the variance of $K(x,X)K(X,X)^{-1}\hat{y}$.

Proposition A.5.1. *There exists a constant $c_1' > 0$ depending on λ, γ, and $\frac{1}{N}\mathrm{Tr}(K)$ only, such that the following bounds hold:*

$$\left|\mathrm{Cov}\left((A_\lambda)_{ii},(A_\lambda)_{jj}\right)\right| \leq \frac{c_1'}{P}$$
$$\mathrm{Var}\left((A_\lambda)_{ij}\right) \leq \min\left\{\frac{d_i}{d_j},\frac{d_j}{d_i}\right\}\frac{c_1'}{P}.$$

For all other cases (i.e. if i,j,k and l take more than two different values), $\mathrm{Cov}\left((A_\lambda)_{ij},(A_\lambda)_{kl}\right) = 0$.

Proof. We want to study the covariances $\mathrm{Cov}\left((A_\lambda)_{ij},(A_\lambda)_{kl}\right)$ for any i,j,k,l. Using the same symmetry argument as in the proof of Proposition A.3.1, $\mathbb{E}\left[(A_\lambda)_{ij}(A_\lambda)_{kl}\right] = 0$ whenever each value in $\{i,j,k,l\}$ does not appear an even number of times in (i,j,k,l). Using the fact that A_λ is symmetric, it remains to study $\mathrm{Cov}\left((A_\lambda)_{ii},(A_\lambda)_{jj}\right)$, $\mathrm{Var}\left((A_\lambda)_{ii}\right)$ and $\mathrm{Var}\left[(A_\lambda)_{ij}\right]$ for all $i\neq j$. By the Cauchy-Schwarz inequality, any bound on $\mathrm{Var}\left((A_\lambda)_{ii}\right)$ will imply a similar bound on $\mathrm{Cov}\left((A_\lambda)_{ii},(A_\lambda)_{jj}\right)$. Besides, as we have seen in the proof of Proposition A.3.1, $\mathbb{E}\left[(A_\lambda)_{ij}\right] = 0$ for any $i\neq j$. Thus, we only have to study $\mathrm{Var}\left((A_\lambda)_{ii}\right)$ and $\mathbb{E}\left[(A_\lambda)_{ij}^2\right]$.

• Bound on $\mathrm{Var}\left((A_\lambda)_{ii}\right)$: From Equation (A.9),

$$\mathrm{Var}\left((A_\lambda)_{ii}\right) = \mathrm{Var}\left(\frac{d_i g_i}{1+d_i g_i}\right) = \mathrm{Var}\left(1-\frac{1}{1+d_i g_i}\right) = \mathrm{Var}\left(\frac{1}{1+d_i g_i}\right) \leq \mathbb{E}\left[\left(\frac{1}{1+d_i g_i}-\frac{1}{1+d_i\tilde{m}}\right)^2\right],$$

where $g_i := g_i(-\lambda)$. Again, we use the first order Taylor approximation $\mathrm{T}h$ of $h: x\to\frac{1}{1+d_i x}$ centered at $\tilde{m} := \tilde{m}(-\lambda)$, as well as the bound (A.7), to obtain

$$\begin{aligned}\mathbb{E}\left[\left(\frac{1}{1+d_i g_i}-\frac{1}{1+d_i\tilde{m}}\right)^2\right] &= \mathbb{E}\left[\left(-\frac{d_i}{(1+d_i\tilde{m})^2}(g_i-\tilde{m}) + h(g_i) - \mathrm{T}h(g_i)\right)^2\right]\\ &\leq \frac{2d_i^2}{(1+d_i\tilde{m})^4}\mathbb{E}\left[\left(g_i-\tilde{m}\right)^2\right] + 2\mathbb{E}\left[\left(h(g_i)-\mathrm{T}h(g_i)\right)^2\right]\\ &\leq \frac{2}{6\tilde{m}^2}\mathbb{E}\left[\left(g_i-\tilde{m}\right)^2\right] + \frac{2}{\tilde{m}^4}\mathbb{E}\left[\left(g_i-\tilde{m}\right)^4\right].\end{aligned}$$

Using Lemma A.2.3, we get $\mathrm{Var}\left((A_\lambda)_{ii}\right) \leq \frac{c_1'}{P}$, where $c_1' > 0$ depends on λ,γ, and $\frac{1}{N}\mathrm{Tr}(K)$ only.

• Bound on $\mathbb{E}\left((A_\lambda)_{ij}\right)$ for $i \neq j$: Following the same arguments as for Equation (A.9), $(A_\lambda)_{ij}$ is equal to

$$(A_\lambda)_{ij} = \frac{\sqrt{d_i d_j}}{P}\left[w_i^T B_{(i)}^{-1} w_j - \frac{d_i g_i}{1+d_i g_i} w_i^T B_{(i)}^{-1} w_j\right] = \frac{\sqrt{d_i d_j}}{1+d_i g_i}\frac{1}{P} w_i^T B_{(i)}^{-1} w_j,$$

where we set $B_{(i)} := B_i(-\lambda)$. Since w_i and $B_{(i)}$ are independent, $\mathbb{E}\left[\left(w_i^T B_{(i)}^{-1} w_j\right)^2\right] = \mathbb{E}\left[w_j^T B_{(i)}^{-2} w_j\right]$, and thus, by the Cauchy-Schwarz inequality, we have

$$\mathbb{E}\left[(A_\lambda)_{ij}^2\right] \leq \frac{1}{P^2}\sqrt{\mathbb{E}\left[\frac{d_i^2 d_j^2}{\left(1+d_i g_i\right)^4}\right]}\sqrt{\mathbb{E}\left[\left(w_j^T B_{(i)}^{-2} w_j\right)^2\right]}. \tag{A.14}$$

Recall that $\tilde{m} := \tilde{m}(-\lambda)$. Using the fact that $\frac{1}{1+d_i g_i} = \frac{1}{1+d_i\tilde{m}} + \frac{1}{1+d_i g_i} - \frac{1}{1+d_i\tilde{m}}$ and inserting the first Taylor approximation $\mathrm{T}h$ of $h : x \to \frac{1}{1+d_i x}$ centered at $\tilde{m}$, we get:

$$\mathbb{E}\left[\left(\frac{1}{1+d_i g_i}\right)^4\right] = \mathbb{E}\left[\left(\frac{1}{1+d_i\tilde{m}} - \frac{d_i}{(1+d_i\tilde{m})^2}(g_i-\tilde{m}) + h(g_i) - \mathrm{T}h(g_i)\right)^4\right].$$

Using a convexity argument, the bound (A.7), and the lower bound on $\tilde{m}$ given by Lemma A.2.5, there exists three constants $\tilde{c}_1$, $\tilde{c}_2$, $\tilde{c}_3$, which depend on λ, γ and $\frac{1}{N}\mathrm{Tr}(K)$ only, such that $\mathbb{E}\left[\left(\frac{1}{1+d_i g_i}\right)^4\right]$ is bounded by

$$\frac{\tilde{c}_1}{(1+d_i\tilde{m})^4} + \frac{\tilde{c}_2 d_i^4}{(1+d_i\tilde{m})^8}\mathbb{E}\left[\left(g_i-\tilde{m}\right)^4\right] + \tilde{c}_3\mathbb{E}\left[\left(g_i-\tilde{m}\right)^8\right].$$

Thanks to Lemma A.2.3 and Proposition A.2.4, this last expression can be bounded by an expression of the form $\frac{\tilde{e}_1}{d_i^4} + \frac{\tilde{e}_2}{P^2 d_i^4} + \frac{\tilde{e}_3}{P^4}$. Note that $\frac{\tilde{e}_2}{P^2 d_i^4} \leq \frac{\tilde{e}_2}{d_i^4}$ and $\frac{\tilde{e}_3}{P^4} \leq \frac{\tilde{e}_3}{\gamma^4}\frac{(\frac{1}{N}\mathrm{Tr}(K))^4}{d_i^4}$. Hence, we obtain the bound:

$$\mathbb{E}\left[\left(\frac{1}{1+d_i g_i}\right)^4\right] \leq \frac{\tilde{c}}{d_i^4},$$

where $\tilde{c} = \tilde{e}_1 + \tilde{e}_2 + \frac{\tilde{e}_3(\frac{1}{N}\mathrm{Tr}(K))^4)}{\gamma^4}$ depends on λ, γ and $\frac{1}{N}\mathrm{Tr}(K)$ only.

Let us now consider the second term in the r.h.s. of (A.14). Using the fact that $\|B_{(i)}\|_{op} \geq \frac{1}{\lambda}$, we get

$$\sqrt{\mathbb{E}\left[\left(w_j^T B_{(i)}^{-2} w_j\right)^2\right]} \leq \sqrt{\frac{1}{\lambda^4}\mathbb{E}\left[\left(w_j^T w_j\right)^2\right]} = \sqrt{\frac{1}{\lambda^4}N(N+2)} \leq \frac{N+1}{\lambda^2},$$

where we have used the fact that the second moment of a $\chi^2(N)$ distribution is $N(N+2)$.

Together, we obtain

$$\begin{aligned}\mathbb{E}\left[(A)_{ij}^2\right] &\leq \frac{1}{P^2}\sqrt{\mathbb{E}\left[\frac{d_i^2 d_j^2}{\left(1+d_i g_i\right)^4}\right]}\sqrt{\mathbb{E}\left[\left(w_j^T B_{(i)}^{-2} w_j\right)^2\right]} \\ &\leq \frac{\tilde{c} d_i d_j}{d_i^2}\frac{N+1}{P^2\lambda^2} \\ &\leq \frac{\tilde{c} d_j}{P d_i \lambda^2 \gamma}\frac{N+1}{N} \leq \frac{c_1'}{P}\frac{d_i}{d_j},\end{aligned}$$

for $c_1' = 2\frac{\tilde{c}}{\lambda^2\gamma}$. Since the matrix A_λ is symmetric, we finally conclude that

$$\mathbb{E}\left[(A_\lambda)_{ij}^2\right] \leq \frac{c_1'}{P}\min\left\{\frac{d_i}{d_j}, \frac{d_j}{d_i}\right\}.$$

Note that c_1' is a constant related to the bounds constructed in Lemma A.2.1 and Proposition A.2.4 and as such it depends on $\frac{1}{N}\mathrm{Tr}(K)$, γ and λ only. □

Proposition A.5.2. *There exists a constant $c_1 > 0$ (depending on λ, γ, T only) such that the variance of the estimator is bounded by*

$$\mathrm{Var}\left(K(x,X)K(X,X)^{-1}\hat{y}\right) \leq \frac{c_1 \|y\|_{K^{-1}}^2 K(x,x)}{P}.$$

Proof. As in the proof of Theorem A.3.2, with the right change of basis, we may assume the Gram matrix $K(X,X)$ to be diagonal.

We first express the covariances of $\hat{y} = A(-\lambda)y$. Using Proposition Proposition A.5.1, for $i \neq j$ we have

$$\mathrm{Cov}\left(\hat{y}_i, \hat{y}_j\right) = \sum_{k,l=1}^{N} \mathrm{Cov}\left((A_\lambda)_{ik}, (A_\lambda)_{lj}\right) y_k y_l = \mathrm{Cov}\left((A_\lambda)_{ii}, (A_\lambda)_{jj}\right) y_i y_j + \mathbb{E}\left[(A_\lambda)_{ij}^2\right] y_j y_i,$$

whereas for $i = j$ we have

$$\mathrm{Cov}\left(\hat{y}_i, \hat{y}_i\right) = \sum_{k=1}^{N} \mathrm{Cov}\left((A_\lambda)_{ik}, (A_\lambda)_{ki}\right) y_k^2 = \mathrm{Var}\left((A_\lambda)_{ii}\right) y_i^2 + \sum_{k\neq i} \mathbb{E}\left[(A_\lambda)_{ik}^2\right] y_k^2.$$

We decompose $K^{-\frac{1}{2}}\mathrm{Cov}(\hat{y},\hat{y})K^{-\frac{1}{2}}$ into two terms: let C be the matrix of entries

$$C_{ij} = \frac{\mathrm{Cov}((A_\lambda)_{ii}, (A_\lambda)_{jj}) + \delta_{i\neq j}\mathbb{E}\left[(A_\lambda)_{ij}^2\right]}{\sqrt{d_i d_j}} y_i y_j,$$

and let D the diagonal matrix with entries

$$D_{ii} = \frac{\sum_{k\neq i} \mathbb{E}\left[(A_\lambda)_{ik}^2\right] y_k^2}{d_i}.$$

We have the decomposition $K^{-\frac{1}{2}}\mathrm{Cov}(\hat{y},\hat{y})K^{-\frac{1}{2}} = C + D$.

Proposition A.5.1 asserts that $\mathrm{Cov}((A_\lambda)_{ii},(A_\lambda)_{jj}) \leq \frac{c_1'}{P}$ and $\mathbb{E}\left[(A_\lambda)_{ij}^2\right] \leq \frac{c_1'}{P}$, and thus the operator norm of C is bounded by

$$\begin{aligned}
\|C\|_{op} &\leq \|C\|_F \\
&= \sqrt{\sum_{i,j} \frac{\left(\mathrm{Cov}((A_\lambda)_{ii},(A_\lambda)_{jj}) + \delta_{i\neq j}\mathbb{E}\left[(A_\lambda)_{ij}^2\right]\right)^2}{d_i d_j} y_i^2 y_j^2} \\
&\leq \frac{2c_1'}{P}\sqrt{\sum_{ij}\frac{1}{d_i d_j} y_i^2 y_j^2} = \frac{2c_1'\|y\|_{K^{-1}}^2}{P}
\end{aligned}$$

For the matrix D, we use the bound $\mathbb{E}\left[(A_\lambda)_{ik}^2\right] \leq \frac{c_1'}{P}\frac{d_i}{d_k}$ to obtain

$$D_{ii} = \frac{\sum_{k\neq i}\mathbb{E}\left[(A_\lambda)_{ik}^2\right] y_k^2}{d_i} \leq \frac{c_1'}{P}\sum_{k\neq i}\frac{y_k^2}{d_k} \leq \frac{c_1'\|y\|_{K^{-1}}^2}{P},$$

which implies that $\|D\|_{op} \leq \frac{c_1'\|y\|_{K^{-1}}^2}{P}$. As a result

$$\begin{aligned}
\mathrm{Var}\left(K(x,X)K^{-1}\hat{y}\right) &= K(x,X)K^{-1}\mathrm{Cov}(\hat{y},\hat{y})K^{-1}K(X,x) \\
&\leq K(x,X)K^{-\frac{1}{2}}\|C+D\|_{op}K^{-\frac{1}{2}}K(X,x) \\
&\leq \frac{3c_1'\|y\|_{K^{-1}}^2}{P}\|K(x,X)\|_{K^{-1}}^2 \\
&\leq \frac{3c_1'K(x,x)\|y\|_{K^{-1}}^2}{P},
\end{aligned}$$

where we used Inequality (A.12). This yields the result with $c_1 = 3c_1'$. □

• **Bound on $\mathbb{E}_\pi\left[\|\hat{\theta}\|^2\right]$.** To understand the variance of the λ-RF estimator $\hat{f}_\lambda^{(RF)}$, we need to describe the distribution of the squared norm of the parameters:

Proposition A.5.3. *For $\gamma,\lambda > 0$ there exists a constant $c_2 > 0$ depending on λ,γ,T only such that*

$$\left|\mathbb{E}[\|\hat{\theta}\|^2] - \partial_\lambda\tilde{\lambda}\, y^T K(X,X)\left(K(X,X)+\tilde{\lambda}I_N\right)^{-2} y\right| \leq \frac{c_2\|y\|_{K^{-1}}^2}{P}. \tag{A.15}$$

Proof. As in the proof of Theorem A.3.2, with the right change of basis, we may assume the

Gram matrix $K(X,X)$ to be diagonal. Recall that $\hat{\theta} = \frac{1}{\sqrt{P}}\left(\frac{1}{P}WK(X,X)W^T + \lambda I_N\right)^{-1} WK(X,X)^{\frac{1}{2}} y$, thus we have:

$$\|\hat{\theta}\|^2 \;=\; \frac{1}{P} y^T K(X,X)^{\frac{1}{2}} W^T \left(\frac{1}{P} WK(X,X)W^T + \lambda I_P\right)^{-2} WK(X,X)^{\frac{1}{2}} y \;=\; y^T A'(-\lambda) y, \qquad \text{(A.16)}$$

where $A'(-\lambda)$ is the derivative of

$$A(z) = \frac{1}{P} K(X,X)^{\frac{1}{2}} W^T \left(\frac{1}{P} WK(X,X)W^T - z\mathrm{I}_P\right)^{-1} WK(X,X)^{\frac{1}{2}}$$

with respect to z evaluated at $-\lambda$. Let

$$\tilde{A}(z) = K(X,X)(K(X,X) + \tilde{\lambda}(-z)\mathrm{I}_N)^{-1}.$$

Remark that the derivative of $\tilde{A}(z)$ is given by $\tilde{A}'(z) = \tilde{\lambda}'(-z)K(X,X)(K(X,X) + \tilde{\lambda}(-z)I_N)^{-2}$. Thus, from Equation (A.16), the l.h.s. of (A.15) is equal to:

$$\left| y^T \left(\mathbb{E}[A'(-\lambda)] - \tilde{A}'(-\lambda) \right) y \right|. \qquad \text{(A.17)}$$

Using a classical complex analysis argument, we will show that $\mathbb{E}[A'(-\lambda)]$ is close to $\tilde{A}'(-\lambda)$ by proving a bound of the difference between $\mathbb{E}[A(z)]$ and $\tilde{A}(z)$ for any $z \in \mathbb{H}_{<0}$.

Note that the proof of Proposition A.3.1 provides a bound on the diagonal entries of $\mathbb{E}[A(z)]$, namely that for any $z \in \mathbb{H}_{<0}$,

$$\left| \mathbb{E}[(A(z))_{ii}] - (\tilde{A}(z))_{ii} \right| \leq \frac{c}{P},$$

where $\hat{c}$ depends on z, γ and T only. Actually, in order to prove (A.15), we will derive the following slightly different bound: for any $z \in \mathbb{H}_{<0}$,

$$\left| \mathbb{E}[(A(z))_{ii}] - (\tilde{A}(z))_{ii} \right| \leq \frac{\hat{c}}{d_i P}, \qquad \text{(A.18)}$$

where $\hat{c}$ depends on z, γ and T only. Let $g_i := g_i(z)$ and $\tilde{m} := \tilde{m}(z)$. Recall that for $h_i : x \mapsto \frac{d_i x}{1+d_i x}$, one has $(A(z))_{ii} = h_i(g_i)$, $(\tilde{A}(z))_{ii} = h_i(\tilde{m})$ and

$$\mathrm{T}_{\tilde{m}} h_i(g_i) = \frac{d_i \tilde{m}}{1 + d_i \tilde{m}} - \frac{d_i \left(g_i - \tilde{m}\right)}{(1 + d_i \tilde{m})^2},$$

$$h_i(g_i) - \mathrm{T}_{\tilde{m}} h_i(g_i) = \frac{d_i^2 \left(g_i - \tilde{m}\right)^2}{\left(1 + d_i g_i\right)(1 + d_i \tilde{m})^2},$$

where $\mathrm{T}_{\tilde{m}} h_i$ is the first order Taylor approximation of h_i centered at $\tilde{m}$. Using this first order

Taylor approximation, we can bound the difference $\left|\mathbb{E}[h_i(g_i)] - h_i(\tilde{m})\right|$:

$$\begin{aligned}\left|\mathbb{E}[h_i(g_i)] - h_i(\tilde{m})\right| &\leq \frac{d_i\left|\mathbb{E}[g_i] - \tilde{m}\right|}{(1+d_i\tilde{m})^2} + \frac{d_i^2}{(1+d_i\tilde{m})^2}\mathbb{E}\left[\frac{\left|g_i - \tilde{m}\right|^2}{1+d_i g_i}\right] \\ &\leq \frac{\mathbf{a}}{d_i P} + \mathbf{a}\sqrt{\mathbb{E}\left[\frac{1}{\left(1+d_i g_i\right)^2}\right]\mathbb{E}\left[\left|g_i - \tilde{m}\right|^4\right]},\end{aligned}$$

where $\mathbf{a}$ depends on z, γ and T. We need to bound $\mathbb{E}\left[\frac{1}{(1+d_i g_i)^2}\right]$. Recall that in the proof of Proposition A.5.1, we bounded $\mathbb{E}\left[\frac{1}{(1+d_i g_i)^4}\right]$. Using similar arguments, one shows that

$$\mathbb{E}\left[\frac{1}{\left(1+d_i g_i\right)^2}\right] \leq \frac{\hat{e}^2}{d_i^2},$$

where $\hat{e}$ depends on z, γ and $\frac{1}{N}\mathrm{Tr}(K(X,X))$ only. The term $\mathbb{E}\left[\left|g_i - \tilde{m}\right|^4\right]$ is bounded using Lemmas A.2.3, A.2.1 and Proposition A.2.4. This allows us to conclude that:

$$\left|\mathbb{E}[h_i(g_i)] - h_i(\tilde{m})\right| \leq \frac{\hat{c}}{d_i P},$$

where $\hat{c}$ depends on z, γ and $\frac{1}{N}\mathrm{Tr}(K(X,X))$ only, hence we obtain the Inequality (A.18).

We can now prove Inequality A.15. We bound the difference of the derivatives of the diagonal terms of $A(z)$ and $\tilde{A}(z)$ by means of Cauchy formula. Consider a simple closed path $\phi : [0,1] \to \mathbb{H}_{<0}$ which surrounds z. Since

$$\mathbb{E}[(A'(z))_{ii}] - (\tilde{A}'(z))_{ii} = \frac{1}{2\pi i}\oint_{\phi}\frac{\mathbb{E}[(A(z))_{ii}] - (\tilde{A}(z))_{ii}}{(w-z)^2}dw,$$

using the bound (A.18), we have:

$$\left|\mathbb{E}[(A'(z))_{ii}] - (\tilde{A}'(z))_{ii}\right| \leq \frac{\hat{c}}{d_i P}\frac{1}{2\pi}\oint_{\phi}\frac{1}{|w-z|^2}dw \leq \frac{c_2}{d_i P},$$

where c_2 depends on z, γ, and T only. This allows one to bound the operator norm of $K(X,X)(\mathbb{E}[A'(z)] - \tilde{A}'(z))$:

$$\|K(X,X)(\mathbb{E}[A'(z)] - \tilde{A}'(z))\|_{op} \leq \frac{c_2}{P}.$$

Using this bound and (A.17), we have

$$\left|\mathbb{E}[\|\hat{\theta}\|^2] - \partial_\lambda\tilde{\lambda}\; y^T K(X,X)\left(K(X,X)+\tilde{\lambda}I_N\right)^{-2}y\right| = \left|y^T\left(\mathbb{E}[A'(-\lambda)] - \tilde{A}'(-\lambda)\right)y\right| \leq \frac{c_2\|y\|_{K^{-1}}^2}{P},$$

which allows us to conclude. □

• **Bound on** $\mathrm{Var}\left(\hat{f}_\lambda^{(RF)}(x)\right)$**.** We have shown all the bounds needed in order to prove the following proposition.

Proposition A.5.4. *For any $x \in \mathbb{R}^d$, we have*

$$\mathrm{Var}\left(\hat{f}_\lambda^{(RF)}(x)\right) \leq \frac{c_3 K(x,x)\|y\|_{K^{-1}}^2}{P},$$

where $c_3 > 0$ depends on λ, γ, T.

Proof. Recall that for any $x \in \mathbb{R}^d$,

$$\begin{aligned}\mathrm{Var}(\hat{f}_\lambda^{(RF)}(x)) &= \mathrm{Var}\left(\mathbb{E}\left[\hat{f}_\lambda^{(RF)}(x) \mid F\right]\right) + \mathbb{E}\left[\mathrm{Var}\left[\hat{f}_\lambda^{(RF)}(x) \mid F\right]\right] \\ &= \mathrm{Var}\left(K(x,X)K(X,X)^{-1}\hat{y}\right) + \frac{1}{P}\mathbb{E}\left[\|\hat{\theta}\|^2\right]\left[K(x,x) - K(x,X)K(X,X)^{-1}K(X,x)\right].\end{aligned}$$

From Proposition A.5.2,

$$\mathrm{Var}\left(K(x,X)K(X,X)^{-1}\hat{y}\right) \leq \frac{c_1 K(x,x)\|y\|_{K^{-1}}^2}{P},$$

and from Proposition A.5.3, we have:

$$\mathbb{E}\left[\|\hat{\theta}\|^2\right] \leq \partial_\lambda \tilde{\lambda}\; y^T K\left(K + \tilde{\lambda} I_N\right)^{-2} y + \frac{c_2\|y\|_{K^{-1}}^2}{P} \leq \partial_\lambda \tilde{\lambda}\; \|y\|_{K^{-1}}^2 + \frac{c_2\|y\|_{K^{-1}}^2}{P} \leq \alpha\|y\|_{K^{-1}}^2,$$

where $\alpha = \partial_\lambda \tilde{\lambda} + c_2$. Using the fact that $\tilde{K}(x,x) \leq K(x,x)$, we get

$$\begin{aligned}\mathbb{E}\left[\mathrm{Var}\left[\hat{f}(x) \mid F\right]\right] &= \frac{1}{P}\mathbb{E}\left[\|\hat{\theta}\|^2\right]\left[K(x,x) - K(x,X)K(X,X)^{-1}K(X,x)\right] \\ &\leq \frac{\alpha\|y\|_{K^{-1}}^2 K(x,x)}{P}.\end{aligned}$$

This yields

$$\mathrm{Var}\left(\hat{f}_\lambda^{(RF)}(x)\right) \leq \frac{c_3\|y\|_{K^{-1}}^2 K(x,x)}{P},$$

where $c_3 = \alpha + c_1$. □

A.6 Corollaries

Putting the pieces together, we obtain the following bound on the difference $\Delta_E = |\mathbb{E}[L(\hat{f}_{\lambda,\gamma}^{(RF)})] - L(\hat{f}_\lambda^{(K)})|$ between the expected RF loss and the KRR loss:

Corollary A.6.1. *If $\mathbb{E}_{\mathscr{D}}[K(x,x)] < \infty$, we have*

$$\Delta_E \leq \frac{C_1\|y\|_{K^{-1}}}{P}\left(2\sqrt{L(\hat{f}_\lambda^{(K)})} + C_2\|y\|_{K^{-1}}\right),$$

where C_1 and C_2 depend on λ, γ, T and $\mathbb{E}_{\mathscr{D}}[K(x,x)]$ only.

Proof. Using the bias/variance decomposition, Corollary A.3.3, and the bound on the variance of the predictor, we obtain

$$\begin{aligned}\left|\mathbb{E}\left[L\left(\hat{f}^{(RF)}_{\gamma,\lambda}\right)\right] - L\left(\hat{f}^{(K)}_{\tilde{\lambda}}\right)\right| &\leq \left|L\left(\mathbb{E}\left[\hat{f}^{(RF)}_{\gamma,\lambda}\right]\right) - L\left(\hat{f}^{(K)}_{\tilde{\lambda}}\right)\right| + \mathbb{E}_{\mathscr{D}}\left[\mathrm{Var}\left(\hat{f}(x)\right)\right] \\ &\leq \frac{C\|y\|_{K^{-1}}}{P}\left(2\sqrt{L\left(\hat{f}^{(K)}_{\tilde{\lambda}}\right)} + \frac{C\|y\|_{K^{-1}}}{P}\right) + \frac{c_3\|y\|^2_{K^{-1}}\mathbb{E}_{\mathscr{D}}[K(x,x)]}{P} \\ &\leq \frac{C_1\|y\|_{K^{-1}}}{P}\left(2\sqrt{L\left(\hat{f}^{(K)}_{\tilde{\lambda}}\right)} + C_2\|y\|_{K^{-1}}\right),\end{aligned}$$

where C_1 and C_2 depends on λ, γ, T and $\mathbb{E}_{\mathscr{D}}[K(x,x)]$ only. □

Recall that for any $\tilde{\lambda}$, we denote $M_{\tilde{\lambda}} = K(X,X)(K(X,X)+\tilde{\lambda}I_N)^{-2}$. A direct consequence of Proposition A.5.3 is the following lower bound on the variance of the predictor.

Corollary A.6.2. *There exists $c_4 > 0$ depending on λ, γ, T only such that* $\mathrm{Var}\left(\hat{f}^{(RF)}_{\lambda}(x)\right)$ *is bounded from below by*

$$\partial_\lambda\tilde{\lambda}\frac{y^T M_{\tilde{\lambda}} y}{P}\tilde{K}(x,x) - \frac{c_4 K(x,x)\|y\|^2_{K^{-1}}}{P^2}.$$

Proof. By the law of total cumulance,

$$\mathrm{Var}\left(\hat{f}^{(RF)}_{\lambda}(x)\right) \geq \mathbb{E}\left[\mathrm{Var}\left[\hat{f}^{(RF)}_{\lambda}(x) \mid F\right]\right] \geq \frac{1}{P}\mathbb{E}\left[\|\hat{\theta}\|^2\right]\tilde{K}(x,x).$$

From Proposition A.5.3, $\mathbb{E}[\|\hat{\theta}\|^2] \geq \partial_\lambda\tilde{\lambda}\ y^T M_{\tilde{\lambda}} y - \frac{c_2\|y\|^2_{K^{-1}}}{P}$, hence

$$\mathrm{Var}\left(\hat{f}^{(RF)}_{\lambda}(x)\right) \geq \partial_\lambda\tilde{\lambda}\frac{y^T M_{\tilde{\lambda}} y}{P}\tilde{K}(x,x) - \frac{c_4\tilde{K}(x,x)\|y\|^2_{K^{-1}}}{P^2}.$$

The result follows from the fact that $\tilde{K}(x,x) \leq K(x,x)$. □

B Deep Linear Networks

B.1 Equivalence of Parametrizations/Initializations

NTK Parametrization

Let us show that the NTK parametrization corresponds to a scaling of $\gamma = 1 - \frac{1}{L}$.

The NTK parametrization Jacot, Gabriel, and Hongler, 2018b for linear networks is

$$A_\theta^{NTK} = \frac{W_L}{\sqrt{n_{L-1}}} \cdots \frac{W_1}{\sqrt{n_0}} = \frac{1}{\sqrt{n_0 \cdots n_{L-1}}} W_L \cdots W_1$$

with all parameters initialized with a variance of 1. One can show that gradient flow $\theta^{NTK}(t)$ with the NTK parametrization, initialized at some parameters θ_0^{NTK} is equivalent (up to a rescaling of the learning rate) to gradient flow $\theta(t)$ with the classical parametrization with an initialization of $\theta_0 = (n_0 \cdots n_{L-1})^{-\frac{1}{2L}} \theta_0^{NTK}$:

Proposition B.1.1. *Let $\theta^{NTK}(t)$ be gradient flow on the loss $\mathscr{L}^{NTK}(\theta) = C(A_\theta^{NTK})$ initialized at some parameters θ_0^{NTK} and $\theta(t)$ be gradient flow on the cost $\mathscr{L}(\theta) = C(A_\theta)$ initialized at $\theta_0 = (n_0 \cdots n_{L-1})^{-\frac{1}{2L}} \theta_0^{NTK}$. We have*

$$A_{\theta(t)} = A^{NTK}_{\theta^{NTK}(\sqrt{n_0 \cdots n_{L-1}}\, t)}.$$

Proof. We will show that $\theta(t) = (n_0 \cdots n_{L-1})^{-\frac{1}{2L}} \theta^{NTK}(\sqrt{n_0 \cdots n_{L-1}}\, t)$ which implies that $A_{\theta(t)} = A^{NTK}_{\theta^{NTK}(t)}$. This is obviously true at $t = 0$. Now assuming it is true at a time t, we show that the time derivatives of $\theta(t)$ and $(n_0 \cdots n_{L-1})^{-\frac{1}{2L}} \theta^{NTK}(\sqrt{n_0 \cdots n_{L-1}}\, t)$ match:

$$\partial_t \theta^{NTK}(\sqrt{n_0 \cdots n_{L-1}}\, t) = \frac{\sqrt{n_0 \cdots n_{L-1}}}{\sqrt{n_0 \cdots n_{L-1}}} \partial_t \theta(t) = \partial_t \theta(t).$$

□

This implies that the NTK parametrization with $\mathcal{N}(0,1)$ initialization is equivalent to the classical parametrization with $\mathcal{N}(0,(n_0\cdots n_{L-1})^{-\frac{1}{L}})$ initialization, which for rectangular networks corresponds to a $\mathcal{N}(0,n_0^{-\frac{1}{L}}w^{-\frac{L-1}{L}})$ initialization with scaling $\gamma=\frac{L-1}{L}=1-\frac{1}{L}$.

Maximal Update Parametrization

The Maximal Update parametrization (or μ-parametrization) G. Yang and E. J. Hu, 2020 is equivalent to $\gamma=1$. The μ-parametrization for linear rectangular networks is the same the classical one, since

$$A_\theta^\mu=\frac{W_L}{\sqrt{w}}W_{L-1}\cdots W_2\left(\sqrt{w}W_1\right)=W_L\cdots W_1$$

and the parameters are initialized with variance w^{-1}, i.e. $\gamma=1$.

B.2 Distance to Different Critical Points

Let d_{m} and d_{s} be the Euclidean distances between the initialization θ and, respectively, the set of global minima and the set of all saddles. For random variables $f(w),g(w)$ which depend on w, we write $f\asymp g$ if both $f(w)/g(w)$ and $g(w)/f(w)$ are stochastically bounded as $w\to\infty$. The following theorem studies how d_{m} and d_{s} scale as $w\to\infty$:

Theorem B.2.1 (Theorem 3.4.1 in the main)**.** *Suppose that the set of matrices that minimize C is non-empty, has Lebesgue measure zero, and does not contain the zero matrix. Let θ be i.i.d. centered Gaussian r.v. of variance $\sigma^2=w^{-\gamma}$ where $1-\frac{1}{L}\leq\gamma<\infty$. Then:*

1. *if $1-\frac{1}{L}\leq\gamma<1$, we have $d_{\mathrm{m}}\asymp w^{-\frac{(1-\gamma)(L-1)}{2}}$ and $d_{\mathrm{s}}\asymp w^{\frac{1-\gamma}{2}}$,*
2. *if $\gamma=1$, we have $d_{\mathrm{m}},d_{\mathrm{s}}\asymp 1$,*
3. *if $\gamma>1$ we have $d_{\mathrm{m}}\asymp 1$ and $d_{\mathrm{s}}\asymp w^{-\frac{\gamma-1}{2}}$.*

To prove this result, we require a few Lemmas:

Lemma B.2.2. *Let θ be the vector of parameters of a DLN with i.i.d. $\mathcal{N}(0,w^{-\gamma})$ Gaussian entries, and let $\mathscr{A}_{\min}=\{A\in\mathbb{R}^{n_L\times n_0}:C(A)=0\}$ be the set of global minimizers of C. Under the same assumptions on the cost C as Proposition B.2.1, we have $d(A_\theta,\mathscr{A}_{\min})\asymp 1$ as $w\to\infty$.*

Proof. If $\gamma>1-\frac{1}{L}$ then A_θ converges in distribution to the zero matrix as $w\to\infty$, the distance $d(A_\theta,\mathscr{A}_{\min})$ therefore converges to the finite value $d(0,\mathscr{A}_{\min})\neq 0$.

If $\gamma=1-\frac{1}{L}$, then A_θ converges in distribution to random Gaussian matrix with iid $\mathcal{N}(0,1)$ entries (this can seen as a consequence of the more general results for nonlinear networks J. Lee, Bahri, et al., 2017b; G. Matthews et al., 2018). As a result the distribution of $d(A_\theta,\mathscr{A}_{\min})$ converges to the distribution of $d(B,\mathscr{A}_{\min})$ for a matrix B with iid Gaussian $\mathcal{N}(0,1)$ entries.

Since $\mathbb{P}[d(B,\mathscr{A}_{\min})=0]=0$ and $\mathbb{P}[d(B,\mathscr{A}_{\min})>b]\to 0$ as $b\to\infty$ we have that $d(A_\theta,\mathscr{A}_{\min})\asymp 1$ as needed. □

Lemma B.2.3. *Let θ be the vector of parameters of a DLN with iid $\mathcal{N}(0,w^{-\gamma})$ Gaussian entries. For all ϵ, there is a constant $C_{\epsilon,L}$ that does not depend on w s.t. with prob. $1-\epsilon$, we have for all $\theta'\in\mathbb{R}^P$ that*

$$\|A_{\theta'}-A_\theta\|_F^2\le C_{\epsilon,L}\sum_{k=1}^{L}\left\|\theta-\theta'\right\|^{2k}w^{(1-\gamma)(L-k)}.$$

Proof. By Corollary 5.35 in Vershynin, 2010, reformulated as Theorem B.2.4 below, we know that for all ϵ, there is a constant c_ϵ that does not depend on w s.t. with prob. $1-\epsilon$, we have for all ℓ

$$\|W_\ell\|_{op}^2\le c_\epsilon w^{1-\gamma}.$$

We now write $d\theta=\theta'-\theta$ (and the corresponding matrices $dW_\ell=W_\ell'-W_\ell$) so that we may write the difference $A_{\theta+d\theta}-A_\theta$ as the following sum

$$\sum_{\substack{a_1,\dots,a_L\in\{0,1\}\\ \exists\ell,a_\ell\neq 0}}\left(\begin{cases}W_L & \text{if } a_L=0\\ dW_L & \text{if } a_L=1\end{cases}\right)\cdots\left(\begin{cases}W_1 & \text{if } a_1=0\\ dW_1 & \text{if } a_1=1\end{cases}\right)$$

where the indicator a_ℓ determines whether we take W_ℓ or dW_ℓ in the product. We can therefore bound

$$\begin{aligned}\|A_{\theta+d\theta}-A_\theta\|_F^2&\le\left(\sum_{\substack{a_1,\dots,a_L\in\{0,1\}\\ \exists\ell,a_\ell\neq 0}}\left\|\left(\begin{cases}W_L & \text{if } a_L=0\\ dW_L & \text{if } a_L=1\end{cases}\right)\cdots\left(\begin{cases}W_1 & \text{if } a_1=0\\ dW_1 & \text{if } a_1=1\end{cases}\right)\right\|_F\right)^2\\ &\le(2^L-1)\sum_{\substack{a_1,\dots,a_L\in\{0,1\}\\ \exists\ell,a_\ell\neq 0}}\left\|\left(\begin{cases}W_L & \text{if } a_L=0\\ dW_L & \text{if } a_L=1\end{cases}\right)\cdots\left(\begin{cases}W_1 & \text{if } a_1=0\\ dW_1 & \text{if } a_1=1\end{cases}\right)\right\|_F^2\\ &\le(2^L-1)\sum_{\substack{a_1,\dots,a_L\in\{0,1\}\\ \exists\ell,a_\ell\neq 0}}\left(\begin{cases}\|W_L\|_{op}^2 & \text{if } a_L=0\\ \|dW_L\|_F^2 & \text{if } a_L=1\end{cases}\right)\cdots\left(\begin{cases}\|W_1\|_{op}^2 & \text{if } a_1=0\\ \|dW_1\|_F^2 & \text{if } a_1=1\end{cases}\right)\end{aligned}$$

We now bound $\|W_L\|_{op}^2$ by $c_\epsilon w^{1-\gamma}$ and $\|dW_L\|_F^2$ by $\|d\theta\|^2$ so that we obtain the bound

$$\|A_{\theta+d\theta}-A_\theta\|_F^2\le(2^L-1)\sum_{k=1}^{L}\binom{L}{k}\|d\theta\|^{2k}c_\epsilon^{L-k}w^{(1-\gamma)(L-k)}\le C_{\epsilon,L}\sum_{k=1}^{L}\|d\theta\|^{2k}w^{(1-\gamma)(L-k)}$$

for $C_{\epsilon,L} = (2^L - 1)\max_{k=1,\dots,L}\begin{pmatrix} L \\ k \end{pmatrix} c_\epsilon^{L-k}$. □

Let us now prove Theorem B.2.1:

Proof. **(1) Distance to minimum:** Let us first give an lower bound on the distance from initialization to a global minimum. Let θ be the intialization and $\theta + d\theta$ be the closest minimum. By Lemma B.2.3, we obtain

$$\|A_{\theta+d\theta} - A_\theta\|_F^2 \leq C'_L \sum_{k=1}^{L} \|d\theta\|^{2k} w^{(1-\gamma)(L-k)}.$$

If $\gamma > 1$, the term with $k = L$ dominates, in which case $\|A_{\theta+d\theta} - A_\theta\|_F^2 \leq \|d\theta\|^{2L}$ which implies that $\|d\theta\| \geq \|A_{\theta+d\theta} - A_\theta\|_F^{\frac{1}{L}} \geq d(A_\theta, \mathscr{A}_{\min})^{\frac{1}{L}} \asymp 1$ by Lemma B.2.2.

If $\gamma < 1$, the term $k = 1$ dominates, which implies $\|A_{\theta+d\theta} - A_\theta\|_F^2 \leq \|d\theta\|^2 w^{(1-\gamma)(L-1)}$ which implies that $\|d\theta\| \geq \|A_{\theta+d\theta} - A_\theta\|_F w^{-\frac{(1-\gamma)(L-1)}{2}} = O(w^{-\frac{(1-\gamma)(L-1)}{2}})$, which decreases with width.

Let us now show upper bounds on $\|d\theta\|$. When $\gamma > 1$, we will construct a closeby minimum. Let us first define the parameters $\bar{\theta} = (\bar{W}_1, \dots, \bar{W}_L)$ where $\bar{W}_1 = 0$ and $\bar{W}_L = 0$ and $\bar{W}_{\ell,ij} = \begin{cases} W_{\ell,ij} & \text{if } i,j > \min\{n_0, n_L\} \\ 0 & \text{otherwise} \end{cases}$. Since we have set only $O(w)$ parameters to zero, we have $\left\|\theta - \bar{\theta}\right\|^2 = O(\sigma^2 w) = O(w^{1-\gamma})$. Now let the matrix A be a global minimum of the cost C with SVD $A = USV^T$ (with inner dimension equal to the rank k of A), we then set $\theta^* = \bar{\theta} + I^{(k\to w)}(S^{\frac{1}{L}}V^T, S^{\frac{1}{L}}, \dots, S^{\frac{1}{L}}, US^{\frac{1}{L}})$. The parameters θ^* are a global minimum since $A_{\theta^*} = A$ and $\|\theta^* - \theta\| \leq \left\|\theta^* - \bar{\theta}\right\| + \left\|\bar{\theta} - \theta\right\| = O(1) + O(w^{\frac{1-\gamma}{2}}) = O(1)$.

When $\gamma < 1$, with prob. $1 - \epsilon$, we have $s_{min}(W_{L-1}\cdots W_1) > \frac{1}{2}\sigma^{(L-1)} w^{\frac{L-1}{2}} = w^{\frac{(1-\gamma)(L-1)}{2}}$, we can reach a global minimum by only changing W_L, we need $dW_L W_{L-1}\cdots W_1 = A^* - A_\theta$ hence we take $dW_L = (A^* - A_\theta)(W_L\cdots W_1)^+$ with norm $\|d\theta\| = \|dW_L\|_F \leq \frac{\|A^* - A_\theta\|}{s_{min}(W_{L-1}\cdots W_1)} = O(w^{-\frac{(1-\gamma)(L-1)}{2}})$.

(2) Distance to saddles: Given parameters $\theta = (W_1, \dots, W_L)$, we can obtain a saddle θ^* by setting all entries of W_1 and W_L to zero. We have

$$\mathbb{E}\left[\left\|\theta - \theta^*\right\|^2\right] = \mathbb{E}\left[\|W_1\|_F^2\right] + \mathbb{E}\left[\|W_L\|_F^2\right] = \sigma^2(n_0 + n_L)w = O(w^{1-\gamma}).$$

This gives an upper bound of order $w^{1-\gamma}$ on the distance between θ and the set of saddles θ^*.

Now let $\theta^* = \theta + d\theta$ be the saddle closest to θ, we know that

$$0 = \partial_{W_L}\mathscr{L}(\theta^*) = \nabla C(A_{\theta^*})\left(W_1^*\right)^T \cdots \left(W_{L-1}^*\right)^T.$$

Since A_{θ^*} is not a global minimum, $\nabla C(A_{\theta^*}) \neq 0$, for the above to be zero, we therefore need $\left(W_1^*\right)^T \cdots \left(W_{L-1}^*\right)^T$ to not have full column rank, i.e. $\text{Rank}\left(W_1^*\right)^T \cdots \left(W_{L-1}^*\right)^T = n_0$.

We will show that at initialization $(W_1)^T \cdots (W_{L-1})^T$ has rank n_0 and its smallest non-zero singular value s_{min} is of order $w^{\frac{(1-\gamma)(L-1)}{2}}$. We will use the fact that $\left\| (W_1)^T \cdots (W_{L-1})^T - \left(W_1^*\right)^T \cdots \left(W_{L-1}^*\right)^T \right\|_F \geq s_{min}$ to lower bound the distance $\|\theta - \theta^*\|$ using Lemma B.2.3.

The singular values of $W_1^T \cdots W_{L-1}^T$ are the squared root of the eigenvalues of the $n_0 \times n_0$ matrix $W_1^T \cdots W_{L-1}^T W_{L-1} \cdots W_1$. One can show that as $w \to \infty$ this matrix concentrates in its expectation

$$\mathbb{E}\left[W_1^T \cdots W_{L-1}^T W_{L-1} \cdots W_1\right] = \sigma^{2(L-1)} w^{L-1} = w^{(1-\gamma)(L-1)}.$$

which implies that $s_{\min}$ concentrates in $w^{\frac{(1-\gamma)(L-1)}{2}}$ and therefore $s_{\min} \asymp w^{\frac{(1-\gamma)(L-1)}{2}}$.

Now by Lemma B.2.3 (applied to the depth $L-1$ this time), we have with prob. $1-\epsilon$

$$\begin{aligned} s_{\min}^2 &\leq \left\| (W_1)^T \cdots (W_{L-1})^T - \left(W_1^*\right)^T \cdots \left(W_{L-1}^*\right)^T \right\|_F^2 \\ &\leq C_{\epsilon,L-1} \sum_{k=1}^{L-1} \left\|\theta - \theta'\right\|^{2k} w^{(1-\gamma)(L-1-k)} \end{aligned}$$

and $\left\|\theta - \theta'\right\|$ needs to be at least of order $w^{\frac{(1-\gamma)}{2}}$ for any of the terms in the sum to be at least of order $w^{(1-\gamma)(L-1)}$ (actually all these become of the right order at the same time). □

B.2.1 Spectrum Bounds

An important tool in our analysis is the following Theorem (which is a reformulation of Corollary 5.35 in Vershynin, 2010)

Theorem B.2.4. *Let A be a $m \times n$ matrix with i.i.d. $\mathcal{N}(0,\sigma^2)$ entries. For all $t \geq 0$, with probability at least $1 - 2e^{-\frac{t^2}{2}}$, it holds that*

$$\sigma(-\sqrt{m} - \sqrt{n} - t) \leq s_{min}(A) \leq s_{max}(A) \leq \sigma\left(\sqrt{m} + \sqrt{n} + t\right).$$

Corollary B.2.5. *If the parameters θ are independent centered Gaussian with variance σ^2, for all $t \geq 0$, with probability at least $1 - 2Le^{-\frac{t^2}{2}}$, it holds that*

$$\|A_\theta\|_{op} \leq (1+t)^L \sigma^L \left(\sqrt{n_0} + \sqrt{w}\right) (4w)^{\frac{L-2}{2}} \left(\sqrt{w} + \sqrt{n_L}\right).$$

Proof. By Theorem B.2.4, with probability greater than $1 - 2Le^{-\frac{t^2}{2}}$, for all $\ell = 1, \ldots, L$, $\|W_\ell\|_{op} \leq \sigma\left(\sqrt{n_{\ell-1}} + \sqrt{n_\ell} + t\right)$, where $n_\ell = w$ for $\ell \in \{1, \cdots, L-1\}$. Hence

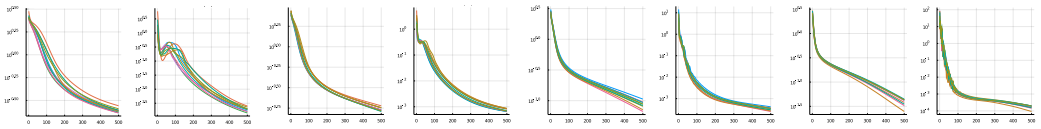
$$\|A_\theta\|_{op} \leq \|W_L\|_{op} \cdots \|W_1\|_{op} \leq \sigma^L \prod_{\ell=1}^{L} \left(\sqrt{n_{\ell-1}} + \sqrt{n_\ell} + t\right) \leq (1+t)^L \sigma^L \prod_{\ell=1}^{L} \left(\sqrt{n_{\ell-1}} + \sqrt{n_\ell}\right).$$

□

C The Loss Landscape of (Non-Linear) Neural Networks

C.1 Numerics

C.1.1 Saddle-to-Saddle Training Dynamics



(a) $m = 10$ (b) $m = 20$ (c) $m = 100$ (d) $m = 1000$

Figure C.1 – *Network width m impacts whether gradient trajectories approach a saddle or not.* For all a-b-c-d, the loss curves are demonstrated on the left and the norm of the gradient is demonstrated on the right. We observe that the norm of the gradient decreases and then increases in narrow networks *(a-b)*, indicating an approach to a saddle and then escaping it. We do not observe a sharp non-monotonicity in the norm of the gradient for wider networks *(c-d)*. Instead we observe short decrease and increase periods in the norm of the gradient (see the zigzag) *(d)*, which indicates that the gradient trajectories move from one saddle to the next in this regime, yet without getting very close these saddles.

Experimental Details

The training set consisted of the standard MNIST test set, i.e. 10'000 grayscale images of 28x28 pixels with corresponding labels. The networks had a single hidden layer of width m with the softplus non-linearity $\sigma(x) = \log(e^x + 1)$. The networks were initialised with the Glorot uniform initialisation (Glorot and Bengio, 2010) and trained on the cross-entropy loss with Adam and gradients always computed on the full dataset. We measured the squared norm of the gradients and the squared norm of the parameter updates.

Interpretation

We observe that the gradient trajectories visit a saddle in a narrow network and the duration of the visit to the saddles becomes shorter as we increase the width (i.e. in (a), we see a longer plateau in the loss curve compared to (b)). In the overparameterized regime, we observe another behavior change, i.e. we observe a zigzag behavior on the norm of the gradient, possibly indicating many short visits to the saddles.

C.2 Further Properties of Symmetric Losses

The most well known property of symmetric losses is the $m!$ multiplicity of the critical points: for a critical point $\theta = (\vartheta_1, \vartheta_2, \dots, \vartheta_m)$ with distinct units $\vartheta_i \neq \vartheta_j$ for all $i \neq j$, there are $m!$ equivalent critical points induced by permutations $\pi \in S_m$. Similarly, every point θ with distinct units has $m! - 1$ partner points with equal loss. For a symmetric loss function, a fundamental region

$$\mathscr{R}_0 := \{(\vartheta_1, \dots, \vartheta_m) \in \mathbb{R}^{Dm} : \vartheta_1 \geq \dots \geq \vartheta_m\}$$

has $m! - 1$ partner regions where the landscape of the loss is the same up to permutations. Note that above and elsewhere we use the lexicographic order: for two units $\vartheta, \vartheta' \in \mathbb{R}^D$, we write $\vartheta > \vartheta'$ if there exists $j \in [D]$ such that $\vartheta_i = \vartheta'_i$ for all $i \in [j-1]$ and $\vartheta_j > \vartheta'_j$; and $\vartheta = \vartheta'$, if $\vartheta_i = \vartheta'_i$ for all $i \in [D]$.

Definition C.2.1. *For a permutation $\pi \in S_m$, a* replicant region $\mathscr{R}_\pi$ *is defined by*

$$\mathscr{R}_\pi := \{(\vartheta_1, \dots, \vartheta_m) \in \mathbb{R}^{Dm} : \vartheta_{\pi(1)} \geq \dots \geq \vartheta_{\pi(m)}\}. \tag{C.1}$$

We denote by $\mathring{\mathscr{R}}_\pi$ the interior of the replicant region.

Any two partner points $\theta_\pi \in \mathscr{R}_\pi$ and $\theta_{\pi'} \in \mathscr{R}_{\pi'}$ have the same loss $L^m(\theta_\pi) = L^m(\theta_{\pi'})$ and they are linked with a permutation matrix $\mathscr{P}_{\pi' \circ \pi^{-1}} : \mathscr{P}_{\pi' \circ \pi^{-1}} \theta_\pi = \theta_{\pi'}$.

Note that the lexicographic order is a total order thus it allows to compare any two D-dimensional units. Therefore every point $\theta \in \mathbb{R}^{Dm}$ falls in at least one replicant region, i.e.

$$\mathbb{R}^{Dm} = \cup_{\pi \in S_m} \mathscr{R}_\pi.$$

The intersection of all these regions $\mathscr{R}_\pi$ corresponds to the D-dimensional linear subspace $\vartheta_1 = \vartheta_2 = \dots = \vartheta_m$; more generally intersections of replicant regions define symmetry subspaces.

As each constraint $\vartheta_i = \vartheta_j$ suppresses D degrees of freedom, we have $\dim(\mathscr{H}_{i_1,\dots,i_k}) = D(m - k + 1)$. Observe that the largest symmetry subspaces are $\mathscr{H}_{i,j}$'s since any other symmetry subspace is included in one of these $\binom{m}{2}$ subspaces.

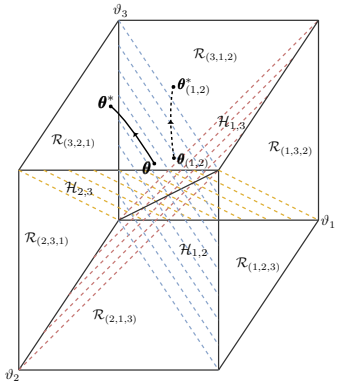


Figure C.2 – *Replicant regions $\mathscr{R}_\pi$ and symmetry subspaces $\mathscr{H}_{i,j}$ for the* 3*-dimensional parameter space* $\mathbb{R}^3$. An example gradient flow trajectory starting at $\theta \in \mathscr{R}_{(3,2,1)}$ and arriving at a minimum θ^* (solid curve) and its partner trajectory starting at a partner point $\theta_{(1,2)} \in R_{(3,1,2)}$ thus arriving at a partner minimum $\theta^*_{(1,2)}$ (dashed curve) are shown.

For $D = 1$, the largest symmetry subspaces have codimension 1. As a result, any path from $\mathscr{R}_\pi$ to any another replicant region has to cross a symmetry subspace (see Figure C.2). However, for $D > 1$, the symmetry subspaces have codimension at least D; thus there exist paths connecting replicant regions without crossing symmetry subspaces.

Lemma C.2.1 (Lemma 4.3.1 in the main)**.** *We assume that $L^m : \mathbb{R}^{Dm} \to \mathbb{R}$ is a symmetric loss on m units and a C^1 function. Let $\Gamma : \mathbb{R}_{\geq 0} \times \mathbb{R}^{Dm} \to \mathbb{R}^{Dm}$ be its gradient flow. If $\Gamma(0, \theta_0) \in \mathscr{H}_{i_1,\ldots,i_k}$, the gradient flow stays inside the symmetry subspace, i.e. $\Gamma(t, \theta_0) \in \mathscr{H}_{i_1,\ldots,i_k}$ for all $t > 0$. If $\Gamma(0, \theta_0) \notin \mathscr{H}_{i,j}$ for all $i \neq j \in [m]$, the gradient flow does not visit any symmetry subspace in finite time.*

Proof. We will write the gradient of L^m in the block form

$$\nabla L^m(\theta) = (\nabla_1 L^m(\theta), \ldots, \nabla_m L^m(\theta)) \quad \text{where for all} \quad j \in [m]$$
$$\nabla_j L^m(\theta) = (\partial_{D(j-1)+1} L^m(\theta), \ldots, \partial_{D(j-1)+D} L^m(\theta))$$

is a D-dimensional vector. We will use the identity that comes from chain rule $\nabla L^m(P_\pi \theta) = P_\pi \nabla L^m(\theta)$. We will show that if $\theta = (\vartheta_1, \ldots, \vartheta_m) \in \mathscr{H}_{i_1,\ldots,i_k}$ where $\vartheta_{i_1} = \cdots = \vartheta_{i_k}$, its gradient satisfies $\nabla_{i_1} L^m(\theta) = \ldots = \nabla_{i_k} L^m(\theta)$ therefore the gradient flow remains on the symmetry subspace for all times.

We denote a transposition by $(i, j) \in S_m$, which is a permutation that only swaps the units i and j. Assume $\theta \in \mathscr{H}_{i,j}$, that is $\theta = P_{(i,j)}\theta$, and thus

$$\nabla L^m(\theta) = \nabla L^m(P_{(i,j)}\theta) = P_{(i,j)} \nabla L^m(\theta),$$

and in particular $\nabla_i L^m(\theta) = \nabla_j L^m(\theta)$. This entails that for $\theta \in \mathscr{H}_{i_1,\ldots,i_k}$, we have $\nabla L^m(\theta) \in \mathscr{H}_{i_1,\ldots,i_k}$ as well, which completes the first part of the proof.

We now prove the second part of the claim by contradiction. Let $\gamma(t) = \Gamma(t, \theta_0)$. Suppose now that $\gamma(0) \notin \mathscr{H}_{i,j}$ for any $i \neq j \in [k]$ and $t_0 < \infty$ be the first time such that $\gamma(t_0) \in \mathscr{H}_{i',j'}$ for some $i' \neq j' \in [k]$. Let $\widetilde{\gamma}(t) = P_{(i',j')}\gamma(t)$, that is the symmetric path with respect to $\mathscr{H}_{i',j'}$. Then one sees that γ and $\widetilde{\gamma}$ intersect for the first time at t_0 on $\mathscr{H}_{i',j'}$ and then $\gamma(t) = \widetilde{\gamma}(t) \in \mathscr{H}_{i',j'}$ for all $t > t_0$, as we showed in the first part of the proof. Since ∇L^m is continuous, Picard-Lindelöf Theorem applies on a neighbourhood of $\gamma(t_0)$, which ensures the unicity of the gradient flow on $[t_0 - \epsilon, t_0]$ for some $\epsilon > 0$. Thus, $\gamma(t_0 - \epsilon) = \widetilde{\gamma}(t_0 - \epsilon)$, which contradicts the fact that t_0 is the first time when γ intersects $\widetilde{\gamma}$. □

Remark. *Let $\Gamma(0, \theta_0) \in \mathscr{R}_\pi$ for some $\pi \in S_m$. In the case of 1-dimensional units, $D = 1$, we have $\Gamma(t, \theta_0) \in \mathscr{R}_\pi$ for all $t \in \mathbb{R}_+$. Hence, in this case, the gradient flow can only be affected by the critical points of a single replicant region.*

Proof. Indeed, assume that $\Gamma(0, \theta_0) = (\vartheta_1(0), \ldots, \vartheta_m(0)) \in \mathscr{R}_\pi$, i.e. $\vartheta_{\pi_1}(0) \geq \cdots \geq \vartheta_{\pi_m}(0)$ and $\Gamma(1, \theta_0) = (\vartheta_1(1), \ldots, \vartheta_m(1)) \in \mathscr{R}_{\pi'}$ for another permutation π', i.e. $\vartheta_{\pi'_1}(0) \geq \cdots \geq \vartheta_{\pi'_m}(0)$. Since $\pi \neq \pi'$, there exists a pair (i, j) such that $\vartheta_i(0) \geq \vartheta_j(0)$ and $\vartheta_j(1) \geq \vartheta_i(1)$. Thus we have

$$(\vartheta_i - \vartheta_j)(0) \geq 0 \geq (\vartheta_i - \vartheta_j)(1).$$

Because the gradient flow is continuous (since L^m is C^1) there exists a time t_0 such that $(\vartheta_i - \vartheta_j)(t_0) = 0$, i.e. $\Gamma(t_0, \theta_0) \in \mathscr{H}_{i,j}$, which yields a contradiction. □

Remark. *In the case of 1-dimensional units, $D = 1$, if $\Gamma(0, \theta_0) \in \mathscr{R}_\pi$ for some $\pi \in S_m$, we have $\Gamma(t, \theta_0) \in \mathscr{R}_\pi$ for all $t \in \mathbb{R}_+$. Hence, in this case, the gradient flow can only be affected by the critical points of a single replicant region.*

C.3 Second-Order Analysis of the Symmetry-Induced Critical Points

In this section, we first present a proof sketch for Theorem 4.4.1, then prove it. In Appendix C.3.1, we study the matrix in Eq. 4.4 in detail for the case of multiple output neurons to characterize the critical points on the line. In Appendix C.3.2, we will present the proof of Lemma 4.4.2.

Proof Sketch. We decompose the Hessian of a symmetry-induced critical point $\oplus^{j,\mu}\theta$ using a specific invertible linear transformation $A(\mu)$ as follows

$$HL(\oplus^{j,\mu}\theta) = A(\mu)^T \begin{bmatrix} \mu(1-\mu)Y & -V & 0 \\ -V & 0 & 0 \\ 0 & 0 & HL(\theta) \end{bmatrix} A(\mu)$$

where the matrix in the middle of the RHS is denoted by $\widetilde{H}L(\oplus^{j,\mu}\theta)$. In our decomposition, $A(\mu)$ is invertible for all μ. Thanks to Sylvester's law of inertia, the number of positive, negative, and zero eigenvalues of $H(\mu) = HL(\oplus^{j,\mu}\theta)$ are the same as those of $\widetilde{H}(\mu) = \widetilde{H}L(\oplus^{j,\mu}\theta)$ which is

a congruent matrix.

Therefore it suffices to study the eigenvalue signs of $\widetilde{H}(\mu)$ which is composed of two block matrices on the diagonal –the matrix in Eq. 4.4 and the Hessian of the original local minimum $HL(\theta)$– with off-diagonal blocks being all-zero. Thus the eigenvalues of $\widetilde{H}(\mu)$ are identical to the union of eigenvalues of its block-diagonal matrices. Hence, the Hessian of the loss at the irreducible critical point $HL(\theta)$ gives the bulk of the sign spectrum. This completes the proof of the first part of the statement. Finally, the eigenvalue signs in the new D directions are determined by the matrix in Eq. 4.4 which is the second part of the statement. *End of Proof Sketch.*

Let us now present the full Hessian of a symmetry-induced critical point

$$HL(\oplus^{j,\mu}\theta) = \begin{bmatrix} \mu^2 X + \mu Y & \mu(1-\mu)X & \mu U + V & \mu U & 0 \\ \mu(1-\mu)X & (1-\mu)^2 X + (1-\mu)Y & (1-\mu)U & (1-\mu)U + V & 0 \\ \mu U^T + V^T & (1-\mu)U^T & Z & Z & 0 \\ \mu U^T & (1-\mu)U^T + V^T & Z & Z & 0 \\ 0 & 0 & 0 & 0 & HL(\ominus^j\theta) \end{bmatrix} \tag{C.2}$$

where X and Y are $d \times d$; U and V are $d \times d_{\text{out}}$; and Z is $d_{\text{out}} \times d_{\text{out}}$ and $HL(\ominus^j\theta)$ is the Hessian corresponding to the parameter θ except for the j-th neuron, which we denote by $\ominus^j\theta$. We need to compute the second-order derivatives to write out the submatrices explicitly. First let us compute the first-order derivatives

$$\partial_{a_j} L(\theta) = \frac{1}{N}\sum_{i=1}^{N} \sigma(w_j \cdot x_i) c'(f_\theta^{(2)}(x_i), y_i)$$
$$\partial_{w_j} L(\theta) = \frac{1}{N}\sum_{i=1}^{N} \sigma'(w_j \cdot x_i) x a_j^T c'(f_\theta^{(2)}(x_i), y_i).$$

Then the second-order derivatives follow

$$\partial^2_{w_j w_k} L(\theta) = \frac{1}{N}\sum_{i=1}^{N} \sigma'(w_j \cdot x_i)\sigma'(w_k \cdot x_i) x_i (x_i)^T a_j^T c''(f_\theta^{(2)}(x_i), y_i) a_k$$
$$\partial^2_{w_j^2} L(\theta) = \frac{1}{N}\sum_{i=1}^{N} \sigma'(w_j \cdot x_i)^2 x_i x_i^T a_j^T c''(f_\theta^{(2)}(x_i), y_i) a_j + \sigma''(w_j \cdot x_i) x_i x_i^T a_j^T c'(f_\theta^{(2)}(x_i), y_i)$$
$$\partial^2_{w_j a_k} L(\theta) = \frac{1}{N}\sum_{i=1}^{N} \sigma(w_j \cdot x_i)\sigma'(w_k \cdot x_i) x_i a_j^T c''(f_\theta^{(2)}(x_i), y_i)$$
$$\partial^2_{w_j a_j} L(\theta) = \frac{1}{N}\sum_{i=1}^{N} \sigma(w_j \cdot x_i)\sigma'(w_j \cdot x_i) x_i a_j^T c''(f_\theta^{(2)}(x_i), y_i) + \sigma'(w_j \cdot x_i) x_i c'(f_\theta^{(2)}(x_i), y_i)^T$$
$$\partial^2_{a_j a_k} L(\theta) = \frac{1}{N}\sum_{i=1}^{N} \sigma(w_j \cdot x_i)\sigma(w_k \cdot x) c''(f_\theta^{(2)}(x_i), y_i)$$
$$\partial^2_{a_j^2} L(\theta) = \frac{1}{N}\sum_{i=1}^{N} \sigma(w_j \cdot x)^2 c''(f_\theta^{(2)}(x_i), y_i).$$

We introduce the following submatrices to ease the notation and to allow noticing the recurrence at the symmetry-induced critical points

$$
\begin{aligned}
X((w_j,a_j),(w_k,a_k)) &= \frac{1}{N}\sum_{i=1}^{N} \sigma'(w_j\cdot x_i)\sigma'(w_k\cdot x_i)x_i x_i^T a_j^T c''(f_\theta^{(2)}(x_i),y_i)a_k \\
X(w_j,a_j) &= \frac{1}{N}\sum_{i=1}^{N} \sigma'(w_j\cdot x_i)^2 x_i x_i^T a_j^T c''(f_\theta^{(2)}(x_i),y_i)a_j \\
Y(w_j,a_j) &= \frac{1}{N}\sum_{i=1}^{N} \sigma''(w_j\cdot x_i)x_i x_i^T a_j^T c'(f_\theta^{(2)}(x_i),y_i) \\
U((w_j,a_j),(w_k,a_k)) &= \frac{1}{N}\sum_{i=1}^{N} \sigma(w_j\cdot x_i)\sigma'(w_k\cdot x_i)x_i a_k^T c''(f_\theta^{(2)}(x_i),y_i) \\
U(w_j,a_j) &= \frac{1}{N}\sum_{i=1}^{N} \sigma(w_j\cdot x_i)\sigma'(w_j\cdot x_i)x_i a_j^T c''(f_\theta^{(2)}(x_i),y_i) \\
V(w_j) &= \frac{1}{N}\sum_{i=1}^{N} \sigma'(w_j\cdot x_i)x_i c'(f_\theta^{(2)}(x_i),y_i)^T \\
Z(w_j,w_k) &= \frac{1}{N}\sum_{i=1}^{N} \sigma(w_j\cdot x_i)\sigma(w_k\cdot x_i)c''(f_\theta^{(2)}(x_i),y_i) \\
Z(w_j) &= \frac{1}{N}\sum_{i=1}^{N} \sigma(w_j\cdot x_i)^2 c''(f_\theta^{(2)}(x_i),y_i)
\end{aligned}
$$

which reduces the second-order derivatives into

$$
\begin{aligned}
\partial^2_{w_i w_j} L(\theta) &= X((w_i,a_i),(w_j,a_j)) \\
\partial^2_{w_i^2} L(\theta) &= X(w_i,a_i) + Y(w_i,a_i) \\
\partial^2_{w_j a_i} L(\theta) &= U((w_i,a_i),(w_j,a_j)) \\
\partial^2_{w_i a_i} L(\theta) &= U(w_i,a_i) + V(w_i) \\
\partial^2_{a_i a_j} L(\theta) &= Z(w_i,w_j) \\
\partial^2_{a_i^2} L(\theta) &= Z(w_i).
\end{aligned}
$$

Note that $X(w_i,a_i)$ is positive definite if the cost c is convex. Moreover $Y(w_i,a_i)$ is a symmetric matrix thus it has real eigenvalues.

Next, we change the basis via an invertible matrix A. We obtain the following transformed

Hessian denoted by $\tilde{H}$ which has an approximate block-diagonal structure and $P^- = P - D$

$$\tilde{H}(\mu) = \begin{bmatrix} \mu(1-\mu)Y & 0 & 0 & -V & 0 \\ 0 & X+Y & U+V & 0 & 0 \\ 0 & U+V & Z & 0 & 0 \\ -V & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & HL(\ominus^j\theta) \end{bmatrix}$$

$$\tilde{H}(\mu) = \begin{bmatrix} (1-\mu)I_d & -\mu I_d & 0 & 0 & 0 \\ I_d & I_d & 0 & 0 & 0 \\ 0 & 0 & \mu I_{d_{\text{out}}} & (1-\mu)I_{d_{\text{out}}} & 0 \\ 0 & 0 & -I_{d_{\text{out}}} & I_{d_{\text{out}}} & 0 \\ 0 & 0 & 0 & 0 & I_{P^-} \end{bmatrix} H(\mu) \begin{bmatrix} (1-\mu)I_d & I_d & 0 & 0 & 0 \\ -\mu I_d & I_d & 0 & 0 & 0 \\ 0 & 0 & \mu I_{d_{\text{out}}} & -I_{d_{\text{out}}} & 0 \\ 0 & 0 & (1-\mu)I_{d_{\text{out}}} & I_{d_{\text{out}}} & 0 \\ 0 & 0 & 0 & 0 & I_{P^-} \end{bmatrix}$$

$$\tilde{H}(\mu) = (A(\mu)^{-1})^T H(\mu) A(\mu)^{-1}; \tag{C.3}$$

where $A(\mu)$ is given by

$$A(\mu) = \begin{bmatrix} I_d & -I_d & 0 & 0 & 0 \\ \mu I_d & (1-\mu)I_d & 0 & 0 & 0 \\ 0 & 0 & I_{d_{\text{out}}} & I_{d_{\text{out}}} & 0 \\ 0 & 0 & -(1-\mu)I_{d_{\text{out}}} & \mu I_{d_{\text{out}}} & 0 \\ 0 & 0 & 0 & 0 & I_{P^-} \end{bmatrix}.$$

Finally, after a change of block rows and block columns, we recover the statement of the Thm. 4.4.1 due to the following observation

$$HL(\theta) = \begin{bmatrix} X+Y & U+V & 0 \\ U+V & Z & 0 \\ 0 & 0 & HL(\ominus^j\theta) \end{bmatrix}. \tag{C.4}$$

In the case of biases, the decomposition applies in the same way. The only important thing to take into account is the update in the submatrices of Y and V

$$Y(w_j, b_j, a_j) = \frac{1}{N}\sum_{i=1}^{N} \sigma''([w_j, b_j] \cdot [x_i, 1])[x_i, 1][x_i, 1]^T a_j^T c'(f_\theta^{(2)}(x_i), y_i) \in \mathbb{R}^{(d+1)\times(d+1)}$$

$$V(w_j, b_j, a_j)_{k\ell} = \frac{1}{N}\sum_{i=1}^{N} \sigma'([w_j, b_j] \cdot [x_i, 1])[x_i, 1]_k c'(f_\theta^{(2)}(x_i), y_i)_\ell \text{ with } k \in [d+1], \ell \in [d_{\text{out}}].$$

C.3.1 Multiple Output Neurons

In this section, we study the matrix in Eq. 4.4 to characterize the symmetry-induced critical points in the case of multiple output neurons. Under a minor assumption, we first show that all critical points on the line are strict saddles (Lemma C.3.1). Then for the case $\mu = 0$ (and $\mu = 1$), we give the number of zero eigenvalues in the Hessian of the loss (Lemma C.3.2).

Lemma C.3.1. *For multiple output neurons with $d_{out} \geq 2$, if the matrix V is non-vanishing, the submatrix in Eq. 4.4 has at least one negative eigenvalue.*

Proof. We will show that $\tilde{H}$ has a negative eigenvalue as long as at least one entry of V is non-vanishing, i.e. $V_{k\ell} \neq 0$. It suffices to show that a submatrix of the submatrix, say 2×2 in Eq. 4.4 has one negative eigenvalue since we can that construct an vector such as $[a_1, a_2, 0]$ which returns a negative direction by picking out 2×2 submatrix. We pick the following 2×2 submatrix

$$\begin{bmatrix} Y_{kk} & -V_{k\ell} \\ -V_{k\ell} & 0 \end{bmatrix}. \tag{C.5}$$

Note that the determinant of the above matrix is $-V_{k\ell}^2 < 0$ since $V_{k\ell} \neq 0$. This completes the proof. □

Lemma C.3.1 implies that for multiple number of output neurons, if $V \neq 0$, then all symmetry-induced critical points on the line are strict saddles.

For the mixing ratio $\mu = 0$, changing the corresponding incoming vector does not change the network function. Therefore we obtain a d-dimensional subspace that goes through the SI critical point $\oplus^{j,0}\theta$ where the loss remains constant. We also have an additional direction of constant loss in the span of the outgoing vectors which is the one pointing towards the line of symmetry-induced critical points. However, this does not guarantee $d+1$ zero eigenvalues in its Hessian since this subspace may correspond to the directions that are not eigenvectors, nevertheless the second-order derivatives vanish. This happens for the Hessians that have positive and negative eigenvalues and where the second-order derivative vanish on the directions between the eigenvectors. A simple example is $L(w_1, w_2) = w_1^2 - w_2^2$ where the Hessian is

$$HL = \begin{bmatrix} 2 & 0 \\ 0 & -2 \end{bmatrix}.$$

The Hessian HL has no zero eigenvalues, however in the direction $w_1 = w_2$, the loss remains constant, which lies in between the two eigenvectors $[1,0]$ and $[0,1]$.

Next we investigate the eigenvalues signs of

$$\begin{bmatrix} 0 & V \\ V^T & 0 \end{bmatrix}. \tag{C.6}$$

to determine the eigenvalue signs of the Hessian of SI critical points at $\mu \in \{0,1\}$ in the new directions. Note that the dimensionality of the null space of V is at least one due to the constraint in Eq. 4.6.

Lemma C.3.2. *Let V be a matrix of size $d \times d_{out}$ such that $dim(Null(V)) = n \geq 1$. Then the number of zero-eigenvalues of the following matrix*

$$\begin{bmatrix} 0 & V \\ V^T & 0 \end{bmatrix} \tag{C.7}$$

of size $(d + d_{out}) \times (d + d_{out})$ is at least $|d - d_{out}|$. If $d > d_{out}$, then at least $d - d_{out} + n$ zero-eigenvalues are guaranteed in particular for $n = 1$, the exact number of zero-eigenvalues is $d - d_{out} + 2$.

Proof. Non-zero eigenvalues. First, observe that for every non-zero eigenvalue λ with the eigenvector $[a, b]$

$$\begin{bmatrix} 0 & V \\ V^T & 0 \end{bmatrix} \begin{bmatrix} a \\ b \end{bmatrix} = \begin{bmatrix} Vb \\ V^T a \end{bmatrix} = \begin{bmatrix} \lambda a \\ \lambda b \end{bmatrix},$$

$-\lambda$ is an eigenvalue corresponding to the eigenvector $[-a, b]$ due to the following

$$\begin{bmatrix} Vb \\ -V^T a \end{bmatrix} = \begin{bmatrix} -\lambda(-a) \\ -\lambda b \end{bmatrix}.$$

In short, the non-zero eigenvalues of the matrix in Eq. C.6 come in pairs $(\lambda, -\lambda)$.

Zero eigenvalues. We search for the number of different solutions (up to sign and scaling) of the following equation

$$\begin{bmatrix} Vb \\ V^T a \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}.$$

For V, recall that we have $Va_j = 0$.

Case1: $d_{out} \leq d$. In this case V^T is $d_{\text{out}} \times d$ so it has a null-space of the dimension at least $d - d_{\text{out}}$. We choose $d - d_{\text{out}}$ orthogonal vectors spanning it, say $v_1, ..., v_{d-d_{\text{out}}}$. Concatenating each one of $d - d_{\text{out}}$ vectors with 0 gives orthogonal eigenvectors, i.e. $[v_1, 0], ..., [v_{d-d_{\text{out}}}, 0]$ of the matrix in Eq. C.6 with zero eigenvalues. In addition, we can concatenate the 0 vector with a_j which is in the null space of V, i.e. $[0, a_j]$, which is orthogonal to the others. In general, if $\dim(N(V)) = n$, by concatenating all with zero vectors, we get $[0, v]$ eigenvectors that are orthogonal to each other and others of the form $[v, 0]$. Therefore, we constructed $d - d_{\text{out}} + n$ orthogonal eigenvectors with zero eigenvalues.

Finally we know that the number of non-zero eigenvalues should be even. If n is odd, so is $(d + d_{\text{out}}) - (d - d_{\text{out}} + n) = 2d_{\text{out}} - n$, therefore there has to be at least one more zero eigenvalue.

In this case, there are at least $d - d_{\text{out}} + n + 1$ zero eigenvalues.

On the other hand, the rank of $V^T V$ is $d_{\text{out}} - n$. Let v be an eigenvector with a non-zero

eigenvalue λ^2. Therefore, $[\frac{1}{\lambda}Vv, v]$ is an eigenvector of the matrix in Eq. C.6 with the eigenvalue λ. Following this construction, overall we get $2(d_{\text{out}} - n)$ non-zero eigenvalues.

If $n = 1$, there are exactly $2(d_{\text{out}} - 1)$ non-zero and $d - d_{\text{out}} + 2$ zero eigenvalues.

Case2: $d < d_{out}$. In this case V is $d \times d_{\text{out}}$, therefore it has a null-space of the dimension at least $d_{\text{out}} - d$. Concatenating each one of them with 0 vectors, we find at least $d_{\text{out}} - d$ zero eigenvalues.

Note the asymmetry between the two cases: a_j is a non-zero vector in the null space of V, but we do not have such a knowledge for V^T thus its null space may be 0-dimensional. □

C.3.2 Bounding the Minimal Hessian Eigenvalue

Now using the decomposition, we will provide an upper bound for the minimum eigenvalue of the Hessian. In this section we denote the Hessian by H to ease notation (i.e. dropping the loss L) or by $H(\mu)$ where it makes sense to emphasize μ.

Negative Minimum Eigenvalue

The Rayleigh quotient for any $u \neq 0$ tightly upper bounds the minimum eigenvalue Horn and Johnson, 2012

$$\frac{u^T H u}{u^T u} \geq \lambda_{\min}(H).$$

Plugging in the decomposition (Eq. C.4), we get

$$\frac{u^T A^T \tilde{H} A u}{u^T u} = \frac{v^T H v}{(A^{-1}v)^T A^{-1} v} \geq \lambda_{\min}(H)$$

for any vector v. We can assume that it is a unit vector. Let us choose $v = [v_0, 0, 0, 0, 0]$ where v_0 is a unit eigenvector of Y with an eigenvalue λ_0. Thus we have

$$\begin{aligned} v^T \tilde{H} v &= \begin{bmatrix} v_0^T & 0 & 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} \mu(1-\mu)Y & 0 & 0 & -V & 0 \\ 0 & X+Y & U+V & 0 & 0 \\ 0 & U+V & Z & 0 & 0 \\ -V & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & HL(\ominus^j \theta) \end{bmatrix} \begin{bmatrix} v_0 \\ 0 \\ 0 \\ 0 \\ 0 \end{bmatrix} \\ &= \begin{bmatrix} v_0^T & 0 & 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} \mu(1-\mu)\lambda_0 v_0 \\ 0 \\ 0 \\ -V v_0 \\ 0 \end{bmatrix} \\ &= \mu(1-\mu)\lambda_0 \end{aligned} \tag{C.8}$$

We need to check

$$A^{-1} v = \begin{bmatrix} (1-\mu)I_d & I_d & 0 & 0 & \\ -\mu I_d & I_d & 0 & 0 & 0 \\ 0 & 0 & \mu I_{d_{\text{out}}} & -I_{d_{\text{out}}} & 0 \\ 0 & 0 & (1-\mu)I_{d_{\text{out}}} & I_{d_{\text{out}}} & 0 \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} v_0 \\ 0 \\ 0 \\ 0 \\ 0 \end{bmatrix} = \begin{bmatrix} (1-\mu)v_0 \\ -\mu v_0 \\ 0 \\ 0 \\ 0 \end{bmatrix}$$

so the norm of $A^{-1} v$ is

$$\| A^{-1} v \|^2 = (1-\mu)^2 + \mu^2.$$

Therefore by choosing a specific unit eigenvector v_k, we obtained the following upper bound on the minimum eigenvalue

$$\frac{\mu(1-\mu)}{(1-\mu)^2+\mu^2} \lambda_k \geq \lambda_{\min}(H)$$

which is valid for every μ and every eigenvalue λ_k of Y. To make the bound tightest using this form of v, we need to choose v_k as the extreme eigenvectors of Y. We obtain (see Fig. C.3):

- *for* $\mu \in (0,1)$*:*
$$\frac{\mu(1-\mu)}{(1-\mu)^2+\mu^2} \lambda_{\min}(Y) \geq \lambda_{\min}(H(\mu)) \text{ for } \mu \in (0,1),$$

- *for* $\mu \in \mathbb{R}/[0,1]$*:*
$$\frac{\mu(1-\mu)}{(1-\mu)^2+\mu^2} \lambda_{\max}(Y) \geq \lambda_{\min}(H(\mu)) \text{ for } \mu \in \mathbb{R}/[0,1].$$

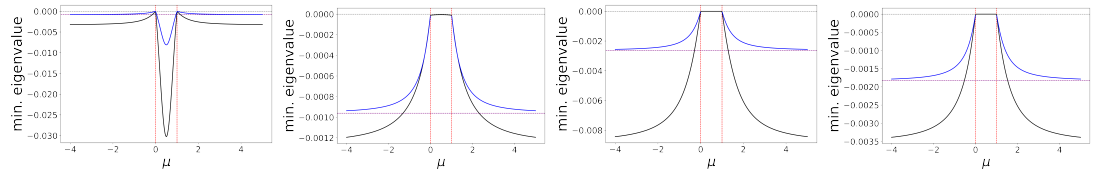


Figure C.3 – *The minimal Hessian eigenvalue of the sym. ind. strict saddles as a function of μ (black) and the upper bound (blue).* We analyze the case where both the student and the teacher have 4 neurons. We observe that the upper bound on the most negative eigenvalue of the Hessian qualitatively captures the behavior of the most negative eigenvalue. In the cases (b-c-d), the matrix Y is positive-definite, the upper bound for the line segment $\mu \in (0,1)$ is positive. Since we already know that the min. eigenvalue for this line segment is zero, the upper bound is not plotted.

In particular, in the limits as $\mu \to \pm\infty$, we get

$$-\frac{1}{2}\lambda_{\max}(Y) \geq \lim_{\mu\to\pm\infty} \lambda_{\min}(H(\mu))$$

which is the statement of Eq. 4.8.

Similarly, using an additive decomposition, we will give a lower bound on the minimum eigenvalue. Note that the entries of the Hessian are quadratic in μ. We use an additive decomposition of the Hessian

$$\mu^2 \underbrace{\begin{bmatrix} X & -X & 0 & 0 & 0 \\ -X & X & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix}}_{M_2} + \mu \underbrace{\begin{bmatrix} Y & X & U & U & 0 \\ X & -2X-Y & -U & -U & 0 \\ U^T & -U^T & 0 & 0 & 0 \\ U^T & -U^T & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix}}_{M_1} + \underbrace{\begin{bmatrix} 0 & 0 & V & 0 & 0 \\ 0 & X+Y & U & U+V & 0 \\ V^T & U^T & Z & Z & 0 \\ 0 & U^T+V^T & Z & Z & 0 \\ 0 & 0 & 0 & 0 & HL(\ominus^j\theta) \end{bmatrix}}_{M_0}$$

since X is positive definite. Moreover, $\lambda_{\min}(M_2) = 0$ since we can choose an eigenvector $[v, v, 0, 0]$ where v is an eigenvector of X.

Using the following general inequality twice Horn and Johnson, 2012

$$\lambda_{\min}(A+B) = \min_{\|u\|=1} u^T(A+B)u \geq \min_{\|u\|=1} u^T Au + \min_{\|u\|=1} u^T Bu = \lambda_{\min}(A) + \lambda_{\min}(B),$$

we obtain the following bound on the min. eigenvalue for $\mu \geq 0$

$$\lambda_{\min}(H(\mu)) \geq \mu^2\lambda_{\min}(M_2) + \mu\lambda_{\min}(M_1) + \lambda_{\min}(M_0) \geq \mu\lambda_{\min}(M_1) + \lambda_{\min}(M_0),$$

and for $\mu < 0$

$$\lambda_{\min}(H(\mu)) \geq \mu\lambda_{\max}(M_1) + \lambda_{\min}(M_0).$$

In particular, as $\mu \to \infty$, this shows that $\lambda_{\min}(H(\mu))$ is at most linearly decreasing.

Minimum Positive Eigenvalue & One Output Neuron

If the minimum eigenvalue is not negative, we know that it is zero since all symmetry-induced critical points have a zero eigenvalue in the Hessian. In this case, we want to bound the minimum positive eigenvalue to get a measure of sharpness for the symmetry-induced local minimum. Recalling the decomposition in the case of one output neuron, we have

$$H = A^T \begin{bmatrix} \mu(1-\mu)Y & 0 & 0 & 0 & 0 \\ 0 & X+Y & U+V & 0 & 0 \\ 0 & U+V & Z & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & H(\ominus^j \theta) \end{bmatrix} A. \tag{C.9}$$

We know the eigenvector of $\tilde{H}$ corresponding to the trivial-zero eigenvalue, let's denote it by $e = [0,0,0,1,0]$. First, note that

$$HA^{-1}e = A^T \tilde{H} A A^{-1} e = 0. \tag{C.10}$$

Let's denote the minimum non-negative eigenvalue of H by $\lambda^+_{\min}$ excluding the trivial zero corresponding to the eigenvector $A^{-1}e$. We have the following upper bound for all $u \perp A^{-1}e$

$$\frac{u^T H u}{u^T u} \geq \lambda^+_{\min}(H). \tag{C.11}$$

Plugging in the decomposition we get

$$\frac{u^T A^T \tilde{H} A u}{u^T u} = \frac{v^T \tilde{H} v}{(A^{-1}v)^T A^{-1} v} \geq \lambda^+_{\min}(H) \tag{C.12}$$

where for any $v \perp e$. We can choose $v = [v_0, 0, 0, 0, 0]$ which is orthogonal to e where v_0 is an eigenvector of Y as in the previous case (Sec. C.3.2) which gives us the following upper bound

$$\frac{\mu(1-\mu)}{(1-\mu)^2+\mu^2}\lambda_0 \geq \lambda^+_{\min}(H).$$

Therefore the tightest bounds are as follows:

- *positive definite Y, for $\mu \in (0,1)$:*

$$\frac{\mu(1-\mu)}{(1-\mu)^2+\mu^2}\lambda_{\min}(Y) \geq \lambda^+_{\min}(H);$$

- *negative definite Y, for $\mu \in \mathbb{R}/[0,1]$:*

$$\frac{-\mu(1-\mu)}{(1-\mu)^2+\mu^2}\lambda_{\min}(|Y|) \geq \lambda^+_{\min}(H).$$

C.4 Scaling Law

It will be convenient to use Newton's series for finite differences (Milne-Thomson, 2000):

Definition C.4.1. *Let p be a polynomial of degree d, we define the k-th forward difference of the polynomial $p(x)$ at 0 as*

$$\Delta^k[p](0) = \sum_{i=0}^{k} \binom{k}{i} (-1)^{k-i} p(i).$$

Hence, we can write $p(x)$ as

$$p(x) = \sum_{k=0}^{d} \binom{x}{k} \Delta^k[p](0). \tag{C.13}$$

Rearranging the summands in Equation C.13, one observes that Newton's series for finite differences is a discrete analog of Taylor's series

$$p(x) = \sum_{k=0}^{d} \frac{\Delta^k[p](0)}{k!} [x]_k$$

where $(x)_k = x(x-1)\dots(x-k+1)$ is the falling factorial.

We now proceed with proving Proposition 4.5 in the main.

Proposition C.4.1. *For $n \leq m$, we have*

$$G(n,m) = \sum_{\ell=1}^{n} \binom{n}{\ell} (-1)^{n-\ell} \ell^m, \tag{C.14}$$

Proof. Let us first write out explicitly the scaling law as a sum of permutations for each partition s

$$G(n,m) := \sum_{\substack{s_1+\dots+s_n=m \\ s_i \geq 1}} \binom{m}{s_1,\dots,s_n}.$$

The above can be restated by using the identity

$$\sum_{\substack{s_1+\dots+s_n=m \\ s_i \geq 0}} \binom{m}{s_1,\dots,s_n} = \sum_{\ell=0}^{n} \binom{n}{\ell} \sum_{\substack{s_1+\dots+s_n=m \\ s_i \geq 0}} \binom{m}{s_1,\dots,s_n} 1_{I_\ell}(s_1,\dots,s_n) \tag{C.15}$$

where $I_\ell := \{(0,\dots,0,s_{\ell+1},\dots,s_n) : s_i \geq 1 \text{ for } \ell+1 \leq i \leq n\}$. Equation (C.15) is equivalent to

$$n^m = \sum_{\ell=0}^{n} \binom{n}{\ell} G(n-\ell, m) \tag{C.16}$$

$$= \sum_{\ell=0}^{n} \binom{n}{\ell} G(\ell, m), \tag{C.17}$$

with the convention that $G(0,m) = 0$. Newton's series for finite differences (Equation (C.13)), applied to the polynomial $p(x) = x^m$ at $x = n$, yields

$$n^m = \sum_{\ell=0}^{n} \binom{n}{\ell} \sum_{i=0}^{\ell} \binom{\ell}{i} (-1)^{\ell-i} i^m. \tag{C.18}$$

Note that the outer summation goes up to n instead of m since the terms with a factor $\binom{n}{k}$ for $k \geq n+1$ are zero. Hence we have

$$\sum_{\ell=0}^{n} \binom{n}{\ell} \left[\sum_{i=0}^{\ell} \binom{\ell}{i} (-1)^{\ell-i} i^m - G(\ell, m) \right] = 0. \tag{C.19}$$

Indeed, with m fixed, the solution

$$G(\ell, m) = \sum_{i=0}^{\ell} \binom{\ell}{i} (-1)^{\ell-i} i^m \tag{C.20}$$

is the unique solution for the Equation (C.19) with initial value given by the condition $1^m = 1$. The uniqueness follows from an immediate induction argument: since

$$G(1,m) = \sum_{k_1=m} \binom{m}{k_1} = 1 = \sum_{i=0}^{1} \binom{1}{i} (-1)^{1-i} i^m,$$

the initial step of induction is verified. Then, for the induction hypothesis, for $k = 1,\dots,n-1$, the first $n-1$ term in the summation in Equation (C.19) are null, leaving us with the condition

$$G(n,m) = \sum_{i=0}^{n} \binom{n}{i} (-1)^{n-i} i^m.$$

□

The Proposition above, which holds for $n < m$ shows that $G(n,m)$ are the forward finite difference at 0 for $p(x) = x^m$, i.e. $G(n,m) = \Delta^n[p](0)$. We now comment on the meaning of the formula for $n \geq m$. For a given polynomial $p(x)$ define the *rescaled* Newton's finite differences $\Delta_h^r[p](0)$ as Newton's finite differences (at 0) for the polynomial $p(hx)$; hence, we can write the n-th derivative of the polynomial p as the $h \to 0$ limit of the h-the n-th Newton's finite

difference:

$$p^{(n)}(0) = \lim_{h\to 0^+} \frac{\Delta_h^n[p](0)}{h^n} = \lim_{h\to 0^+} \frac{1}{h^n} \sum_{i=0}^{n} \binom{n}{i} (-1)^{n-i} (hi)^m = \lim_{h\to 0^+} \frac{1}{h^{n-m}} G(n,m).$$

Hence for $n = m$ we obtain $G(m,m) = m!$, whereas for $n > m$ we find $G(n,m) = 0$.

Lemma C.4.2. *For any $k \geq 0$ fixed, we have,*

$$G(m-k,m) \sim \frac{m^k}{2^k k!} m!, \text{ as } m \to \infty.$$

For any fixed $n \geq 0$, we have $G(n,m) \sim n^m$ as $m \to \infty$.

Proof. We begin to show that

$$\lim_{n\to\infty} \frac{1}{(n+k)!n^k} G(n,n+k) = \frac{1}{2^k k!}. \tag{C.21}$$

In particular, we observe that for $k = 1$ we have that

$$G(n,n+1) = \sum_{\substack{s_1+\ldots+s_n=n+1 \\ s_i\geq 1}} \binom{n+1}{s_1,\ldots,s_n} = \binom{n}{1}\binom{n+1}{2,1,\ldots,1} = n\frac{(n+1)!}{2!}.$$

We find that the asymptotic in Equation (C.21) is in fact an exact equality for any $n > 0$.

For a generic $k \geq 0$, we divide the summation in G according to the number of 1's in $(s_1,\ldots,s_n)$

$$\begin{aligned} G(n,n+k) &= \sum_{\substack{s_1+\cdots+s_n=n+k \\ s_i\geq 1}} \binom{n+k}{s_1,\ldots,s_n} \\ &= \binom{n}{k}\binom{n+k}{\underbrace{2,\ldots,2}_{k},\underbrace{1,\ldots,1}_{n-k}} + \sum_{\ell=1}^{k-1}\binom{n}{\ell} \sum_{\substack{s_1+\cdots+s_\ell=\ell+k \\ s_i\geq 2}} \binom{n+k}{s_1,\ldots,s_\ell,\underbrace{1,\ldots,1}_{n-\ell}}. \end{aligned} \tag{C.22}$$

For a given tuple $(s_1,\ldots,s_n)$, let $c = (c_2,\ldots,c_\ell)$, with $\sum_{i=2}^{\ell} c_i = \ell$ and c_i is the number of occurrences of i among $(s_1,\ldots,s_n)$, hence we have

$$\binom{n+k}{s_1,\ldots,s_\ell,1,\ldots,1} = \frac{(n+k)!}{2!^{c_2}\cdots\ell!^{c_\ell}}.$$

Since for a given $c = (c_2,\ldots,c_\ell)$ there are $\binom{\ell}{c_2,\ldots,c_\ell}$ ℓ-tuples $(s_1,\ldots,s_\ell)$ with such occurrences, we rewrite Equation (C.22) as

$$G(n,n+k) = \binom{n}{k}\frac{(n+k)!}{2^k} + \sum_{\ell=1}^{k-1}\binom{n}{\ell} \sum_{\substack{2c_2+\cdots+\ell c_\ell=\ell+k \\ c_2+\cdots+c_\ell=\ell}} \binom{\ell}{c_2,\ldots,c_\ell}\frac{(n+k)!}{2!^{c_2}\cdots\ell!^{c_\ell}}.$$

Dividing both sides by $(n+k)!n^k$, we find

$$\frac{G(n,n+k)}{(n+k)!n^k} = \frac{1}{2^k k!}\frac{n(n-1)\dots(n-k+1)}{n^k} + \sum_{n=1}^{k-1} \sum_{\substack{2c_2+\dots+\ell c_\ell=\ell+k \\ c_2+\dots+c_\ell=\ell}} \frac{n(n-1)\dots(n-\ell+1)}{n^k} C_c, \quad \text{(C.23)}$$

where $C_c := 1/(c_2!\cdots c_\ell!\cdot 2!^{c_2}\cdots \ell!^{c_\ell})$. For $\ell \le k$, we have the following immediate double inequality:

$$n^{\ell-k}\left(\frac{n-\ell+1}{n}\right)^\ell \le \frac{n(n-1)\dots(n-\ell+1)}{n^k} \le n^{\ell-k}.$$

Together with Equation (C.23), the above double inequality leads to

$$\frac{1}{2^k k!}\left(\frac{n-k+1}{n}\right)^k + \sum_{\ell=1}^{k-1} \sum_{\substack{2c_2+\dots+\ell c_\ell=\ell+k \\ c_2+\dots+c_\ell=\ell}} n^{\ell-k}\left(\frac{n-\ell+1}{n}\right)^\ell C_c$$

$$\le \frac{1}{(n+k)!n^k} G(n,n+k) \le \frac{1}{2^k k!} + \sum_{\ell=1}^{k-1} \sum_{\substack{2c_2+\dots+\ell c_\ell=\ell+k \\ c_2+\dots+c_\ell=\ell}} n^{\ell-k} C_c.$$

In the limit $n\to\infty$, both the lower and the upper bound converge to $\frac{1}{2^k k!}$, hence giving

$$G(n,n+k) \sim \frac{n^k(n+k)!}{2^k k!} \sim \frac{(n+k)^k(n+k)!}{2^k k!};$$

finally, by choosing $n = m-k$, we recover the first asymptotics in the statement.

For the second asymptotics, with an induction argument, we show that $G(n,m) \sim n^m$ for fixed n and $m \gg n$. For $n=1$, we have $G(1,m)=1$. For $n=2$, we have $G(2,n) = 2^n - 2 \sim 2^n$. We assume that for all $\ell = 1,\dots,n-1$, we have $G(\ell,m) \sim \ell^m$. Normalizing Equation (C.16) by $1/n^m$, as $m\to\infty$ we have

$$1 = \frac{1}{n^m}G(n,m) + \frac{1}{n^m}\sum_{\ell=1}^{n-1}\binom{n}{\ell}G(\ell,m) \sim \frac{1}{n^m}G(n,m) + \sum_{\ell=1}^{n-1}\frac{n^\ell}{\ell!}\left(\frac{\ell}{n}\right)^m \sim \frac{1}{n^m}G(n,m),$$

which completes the induction step, thus the Lemma. □

D Overparameterized Networks

D.1 Exact Characterization of the Zero-Loss Solutions

The following assumption ensures that the activation function σ has no specificity that yields other invariances than the symmetries between units, e.g. σ cannot be even or odd.

Assumption A. *Let σ be a smooth activation function. We suppose that $\sigma(0) \neq 0$, that $\sigma^{(n)}(0) \neq 0$ for infinitely many even and odd values of $n \geq 0$, where $\sigma^{(n)}$ denotes the n-th derivative.*

The next lemma contains the main argument to prove that when considering an overparametrized 2-layers neural network, no new global minima are created besides those coming from invariances.

Lemma D.1.1. *Suppose that the activation function σ satisfies the Assumption A. If for some pairwise distinct nonzero $\beta_1, \dots, \beta_k \in \mathbb{R}$ and some constant $c \in \mathbb{R}$ we have $g(\alpha) := \sum_{\ell=1}^{k} a_\ell \sigma(\alpha\beta_\ell) = c$ for all $\alpha \in \mathbb{R}$, then $a_\ell = 0$ for all $\ell \in [k]$.*

Proof. We reorder the indices such that for all $\ell \in [k-1]$, either $|\beta_\ell| > |\beta_{\ell+1}|$, or $\beta_\ell = -\beta_{\ell+1}$ such that $|a_\ell| \geq |a_{\ell+1}|$ (if the equality holds the labelling between the two is not important). We distinguish the four following cases:

1. $|\beta_1| > |\beta_2|$,
2. $\beta_1 = -\beta_2$ and $|a_1| > |a_2|$,
3. $\beta_1 = -\beta_2$ and $a_1 = a_2$,
4. $\beta_1 = -\beta_2$ and $a_1 = -a_2$.

Note that there cannot be more that two indices ℓ with same $|\beta_\ell|$ and that 1. 2. 3. and 4. above are disjoint and cover all the possible cases.

Suppose that 1. holds. Note that

$$g^{(n)}(0) = \sum_{\ell=1}^{k} a_\ell \beta_\ell^n \sigma^{(n)}(0) = 0,$$

for all $n \geq 1$, by assumption. On the other hand, the triangle inequality yields that

$$|g^{(n)}(0)| \geq \left(|a_1 \beta_1^n| - \left| \sum_{\ell \neq 1} a_\ell \beta_\ell^n \right| \right) |\sigma^{(n)}(0)| \geq \left(|a_1 \beta_1^n| - |\beta_2^n| \sum_{\ell \neq 1} |a_\ell| \right) |\sigma^{(n)}(0)|.$$

One can always choose $n_0 \geq 1$ large enough such that $\sigma^{(n_0)}(0) \neq 0$ and

$$|\beta_1| > |a_1|^{-1/n_0} |\beta_2| \left(\sum_{\ell \neq \ell_1} |a_\ell| \right)^{1/n_0},$$

so that $|g^{(n)}(0)| > 0$, which is a contradiction with the fact that $g \equiv c$. Hence $a_1 = 0$. This shows the claim in the particular situation where all $|\beta_\ell|$'s are distinct.

One can deal with case 2. using that $|a_1| > |a_2|$, writing

$$|g^{(n)}(0)| \geq \left((|a_1| - |a_2|) |\beta_1^n| - |\beta_3| \sum_{\ell \neq 1,2} |a_\ell| \right) |\sigma^{(n)}(0)|.$$

The reasoning is then identical to 1.

In the case 3., since σ has infinitely many non-zero even derivatives at 0, we use that $a_1 \beta_1^{2n} + a_2 \beta_2^{2n} = 2a_1 \beta_1^{2n}$ to write

$$|g^{(2n)}(0)| \geq \left((2|a_1|) |\beta_1^{2n}| - \sum_{\ell \neq 1,2} |a_\ell \beta_\ell^{2n}| \right) |\sigma^{(2n)}(0)|,$$

then choose n large enough to argue as above that $a_1 = a_2 = 0$. We can thus eliminates these terms from the definition of g and go on with the argument.

In the case 4., if σ has infinitely many non-zero odd derivatives at 0, we apply the same reasoning as in 3. to show that $a_1 = a_2 = 0$.

Since σ has infinitely many even and infinitely many odd non-zero derivatives at 0, we can iterate the argument and the proof is over since the four cases above cover all possible cases.

□

When σ does not satisfy Assumption A, the proof above allows us to derive the following results:

Lemma D.1.2. *If σ is analytic such that $\sigma^{(n)}(0) \neq 0$ for infinitely many even $n \geq 0$ but only*

finitely many odd $n \geq 1$, then the function g in Lemma D.1.1 can be written as

$$g(\alpha) = \sum_{\ell=1}^{\widetilde{k}} \widetilde{a}_\ell \widetilde{\sigma}(\alpha \widetilde{\beta}_\ell),$$

where $\widetilde{\sigma}$ is an odd polynomial, the $\widetilde{a}_\ell$'s are nonzero and the $|\beta_\ell|$'s are pairwise distinct.

Similarly, if $\sigma^{(n)}(0) \neq 0$ for infinitely many odd $n \geq 1$ but only finitely many even $n \geq 0$, then the function g in Lemma D.1.1 can be written as

$$g(\alpha) = \sum_{\ell=1}^{\widetilde{k}} \widetilde{a}_\ell \widetilde{\sigma}(\alpha \widetilde{\beta}_\ell),$$

where $\widetilde{\sigma}$ is an even polynomial, the $\widetilde{a}_\ell$'s are nonzero and the $|\beta_\ell|$'s are pairwise distinct.

Proof. Suppose that $\sigma^{(2n+1)}(0) \neq 0$ for only finitely many $n \geq 0$. In the proof of Lemma D.1.1, the only problematic situation is 4., that is $\beta_1 = -\beta_2$ and $a_1 = -a_2$. In particular, they cancel out in the even derivatives of g, that is

$$g^{(2n)}(0) = \sigma^{(2n)}(0) \sum_{\ell \neq 1,2} a_\ell \beta_\ell^{2n}.$$

If $\beta_3, a_3, \beta_4, a_4$ do not fall into case 4. from the proof of Lemma D.1.1, then one can show with the same argument therein that $a_3 = a_4 = 0$. Therefore, the problem reduces to the situation where k is even, $\beta_{2\ell-1} = -\beta_{2\ell}$ and $a_{2\ell+1} = -a_{2\ell+2}$ for all $\ell \in [k/2]$. We can then rewrite g as

$$g(\alpha) = \sum_{\ell=1}^{\widetilde{k}} \widetilde{a}_\ell \widetilde{\sigma}(\alpha \widetilde{\beta}_\ell),$$

where $\widetilde{k} \leq k/2$, $\widetilde{a}_\ell := a_{2\ell-1}$, $\widetilde{\beta}_\ell := \beta_{2\ell-1}$ and $\widetilde{\sigma}(x) := \sigma(x) - \sigma(-x)$. The function $\widetilde{\sigma}$ is analytic and locally polynomial around 0, therefore is a polynomial on $\mathbb{R}$ and the $|\widetilde{\beta}_\ell|$'s are pairwise distinct.

When the even derivatives eventually vanish at 0 instead, then the problematic situation is the 3. from Lemma D.1.1 and the function becomes

$$g(\alpha) = \sum_{\ell=1}^{k/2} \widetilde{a}_\ell \widetilde{\sigma}(\alpha \widetilde{\beta}_\ell),$$

where $\widetilde{a}_\ell := a_{2\ell-1}$, $\widetilde{\beta}_\ell := \beta_{2\ell-1}$ and $\widetilde{\sigma}(x) := \sigma(x) + \sigma(-x)$ with $\widetilde{\sigma}$ polynomial as above. □

The case of the sigmoid activation $\sigma(x) = 1/(1+e^{-x})$. In this case, $\sigma(x) = 1/2 + \tanh(x)$ and tanh is an odd function, i.e. $\sigma^{(2n)}(0) = 0$ for all $n \geq 1$. Hence, $\widetilde{\sigma}(x) = \sigma(x) + \sigma(-x) = 1$ for all $x \in \mathbb{R}$ and one can construct the null function with already four β's satisfying the constraints: $a_1\sigma(\beta_1 x) + a_1\sigma(-\beta_1 x) + a_3\sigma(\beta_3 x) + a_3\sigma(-\beta_3 x) = 0$ as soon as $a_1 = -a_3$, such that $|\beta_1| \neq |\beta_3|$.

(One could then also achieve this for any even $p \geq 4$ such functions by tuning the a_ℓ's.)

The case of the softplus activation $\sigma(x) = \ln(1+e^x)$. The Softplus function is the primitive of the sigmoid such that $\sigma(x) = \int_{-\infty}^{x} \frac{1}{1+e^{-u}} \mathrm{d}u$. Therefore, $\sigma^{(2n+1)}(0) = 0$ when $n \geq 1$. In particular, $\widetilde{\sigma}(x) = \sigma(x) - \sigma(-x) = x$ for all $x \in \mathbb{R}$. One can thus obtain the null function with four (or a strictly greater even number) β's satisfying the constraints: $a_1\sigma(\beta_1 x) - a_1\sigma(-\beta_1 x) + a_3\sigma(\beta_3 x) - a_3\sigma(-\beta_3 x) = 0$, as soon as $a_1\beta_1 + a_3\beta_3 = 0$, where $|\beta_1| \neq |\beta_3|$ are pairwise distinct.

The case of the tanh activation function $\sigma(x) = (e^x - e^{-x})/(e^x + e^{-x})$. Since σ is an odd function, $\widetilde{\sigma}(x) = \sigma(x) + \sigma(-x) = 0$ for all $x \in \mathbb{R}$ and therefore one can achieve the null function with two (or a strictly greater even number) β's satisfying the constraints: $a_1\sigma(\beta_1 x) - a_1\sigma(-\beta_1 x)$.

We stress that for the three functions above, there is no other way to obtain the null function (i.e. the coefficients β_ℓ's and a_ℓ's have to be all in case 3. or case 4. depicted in the proof of Lemma D.1.1, according to the derivatives of σ).

Recall that we consider the loss L^m_μ where μ is an input data distribution with support $\mathbb{R}^{d_0}$.

Theorem D.1.3 (Theorem 4.2 in the main)**.** *Suppose that the activation function σ satisfies the Assumption A. For $m > k$, let θ be an m-neuron point, and θ_* be a unique k-neuron global minimum up to permutation, i.e. $L^k(\theta_*) = 0$. If $L^m(\theta) = 0$, then $\theta \in \Theta_{r^* \to m}(\theta_*)$.*

Proof. For $x \in \mathbb{R}^{d_0}$, let $h(x) := \sum_{j=1}^{m} a_j\sigma(w_j \cdot x) - \sum_{j=1}^{m^*} a_j^*\sigma(w_j^* \cdot x)$ and note that this function is zero on $\mathbb{R}$. Since θ_* is irreducible, we know that the w_j^*'s are pairwise distinct, and the a_j^*'s are nonzero. We can always group terms such that, wlog, the w_j's are nonzero, pairwise distinct and the a_j's are nonzero, and we remain in the expansion manifold, as we now argue: we have that

$$h(x) = \sum_{j=1}^{m+m^*} a_j\sigma(w_j \cdot x),$$

where we set $a_j = -a^*_{j-m}$ and $w_j = w^*_{j-m}$ for $j \in \{m+1, \ldots, m+m^*\}$. If some of the w_j's appear several times, we group them together and if some are zero vectors, we summarize them in a constant $c \in \mathbb{R}$ and arrive at

$$h(x) = \sum_{j=1}^{M} A_j\sigma(W_j \cdot x) = c,$$

with $M \leq m + m^*$, such that $W_i \neq W_j$ for all $i \neq j \in [M]$ with $W_j \neq (0, \ldots, 0)^{\mathrm{T}}$. Proving the claim, i.e. that $\theta \in \Theta_{r^* \to m}(\theta_*)$, is now equivalent to showing that $A_j = 0$ for all $j \in M$.

If $d_0 = 1$, we simply apply Lemma D.1.1 which shows that $A_j = 0$ for all $j \in [M]$.

Suppose now that $d_0 > 1$. Let $\epsilon > 0$ and let $t_\epsilon = (1, \epsilon, \epsilon^2, \dots, \epsilon^M)^{\mathrm{T}}$. We define

$$h_\epsilon(\alpha) := \sum_{j=1}^{M} A_j \sigma(\alpha W_j \cdot t_\epsilon), \qquad \alpha \in \mathbb{R}.$$

We claim that Lemma D.1.1 applies to h_ϵ, that is, the elements in $\{W_j \cdot t_\epsilon;\, j \in [M]\}$ are pairwise distinct for all $\epsilon > 0$ small enough. Indeed, by contradiction, suppose that there exists a positive decreasing sequence $(\epsilon_n)_{n\geq 1}$ such that $\lim_{n\to\infty} \epsilon_n = 0$ and $W_1 \cdot t_{\epsilon_n} = W_2 \cdot t_{\epsilon_n}$. Then $(W_1)_1 + \mathcal{O}(\epsilon_n) = (W_2)_1 + \mathcal{O}(\epsilon_n)$ where $(W_j)_k$ denotes the k-th component of W_j. Choosing n large enough enforces $(W_1)_1 = (W_2)_1$. It suffices then to explicit the terms of order ϵ_n in the identity and to reason identically since the rest is $\mathcal{O}(\epsilon_n^2)$. This implies that $W_1 = W_2$, which is a contradiction with the assumption that the vectors W_j are pairwise distinct.

Hence, by Lemma D.1.1 applied on h_ϵ, we have that $A_j = 0$ for all $j \in [M]$, which concludes the proof. □

Remark. *The theorem above does not apply to the sigmoid, the softplus and the* tanh *activation functions, since none of these satisfy Assumption A. Nonetheless, we discussed above the theorem how to reconstruct a neural network function with these activations, with parameters that have to satisfy some explicit constraints depending on the activation (in particular, every w' in the bigger network has to be either equal to w or $-w$ of the smaller network). By considering the extended expansion manifolds of these activation functions, comprised of the classical expansion manifold and these new points, Theorem D.1.3 holds true, that is, the extended expansion manifold is exactly the set of global minima.*

D.1.1 Piecewise Linear Connectivity

Theorem D.1.4. *For $m > k$, $\Theta_{n\to m}(\theta_0)$ is connected: any pair of distinct points $\theta, \theta' \in \Theta_{n\to m}(\theta_0)$ is connected via a union of line segments $\gamma : [0,1] \to \Theta_{n\to m}(\theta_0)$ such that $\gamma(0) = \theta$ and $\gamma(1) = \theta'$.*

Proof. We first prove the case $m = k+1$. Let $\theta_0 = (w_1, a_1) \oplus \dots \oplus (w_k, a_k)$ and consider the following set of points

$$\widetilde{\Theta}_{k\to k+1}(\theta_0) := \{P_\pi \theta^{k+1} : \theta^{k+1} = \theta_0 \oplus (w', 0);\ \pi \in S_r,\ w_0 \in \mathbb{R}^{d_0}\}$$

which is a subset of the expansion manifold $\Theta_{k\to k+1}(\theta_0)$. We will show that by construction that a point $\theta \in \widetilde{\Theta}_{k\to k+1}(\theta_0)$ such that $\theta = \theta_0 \oplus (w', 0)$ is connected to any other point $\widetilde{\Theta}_{k\to k+1}(\theta_0) = P_\pi \theta_0 \in \widetilde{\Theta}_{k\to k+1}(\theta_0)$ via a path in $\Theta_{k\to k+1}(\theta_0)$. To do so we first show that a neighbor where the neuron $\vartheta_0 = (w', 0)$ is swapped with $\vartheta_i = (w_i, a_i)$

$$\theta_1 = (w_1, a_1) \oplus \dots \oplus (w_{i-1}, a_{i-1}) \oplus (w', 0) \oplus \dots \oplus (w_i, a_i)$$

can be reached in three steps using the following line segments $\gamma_1^{(1)}, \gamma_2^{(1)}, \gamma_3^{(1)} : [0,1] \to \Theta_{k\to k+1}(\theta_0)$

$$\gamma_1^{(1)}(\alpha) = (w_1, a_1) \oplus \ldots \oplus (w_k, a_k) \oplus (\alpha(w_i - w_0), 0)$$
$$\gamma_2^{(1)}(\alpha) = (w_i, a_i) \oplus (w_1, a_1) \oplus \ldots \oplus (w_{i-1}, a_{i-1}) \oplus (w_i, \alpha a_i) \oplus \ldots \oplus (w_k, a_k)$$
$$\gamma_3^{(1)}(\alpha) = (w_i, a_i) \oplus (w_1, a_1) \oplus \ldots \oplus (w_{i-1}, a_{i-1}) \oplus (\alpha(w_0 - w_i) + w_i, 0) \oplus (w_{i+1}, a_{i+1}) \oplus \ldots \oplus \ldots (w_k, a_k)$$

where we have $\gamma_1^{(1)}(0) = \theta_0$, $\gamma_1^{(1)}(1) = \gamma_2^{(1)}(0)$, $\gamma_2^{(1)}(1) = \gamma_3^{(1)}(0)$, and $\gamma_3^{(1)}(1) = \theta_1$. In particular, we constructed a path $\gamma^{(1)}$ by glueing three line segments at their end points

$$\gamma^{(1)}(t) = \gamma_1^{(1)}(3t)1_{t\in[0,1/3)} + \gamma_2^{(1)}(3(t-1/3))1_{t\in[1/3,2/3)} + \gamma_3^{(1)}(3(t-2/3))1_{t\in[2/3,1]}$$

where $\gamma^{(1)}(0) = \theta_0$ and $\gamma^{(1)}(1) = \theta_1$. Note that going from $\theta_0 \to \theta_1$, we swapped the neurons ϑ_0 and ϑ_i. Moreover, it is well known that any permutation can be written as a composition of transpositions (permutations leaving all elements unchanged but two) and that $(i\ j) = (0\ j) \circ (0\ i) \circ (0\ j)$. In particular, we can reach $\widetilde{\theta}$ only by swapping ϑ_0 with other neurons, which corresponds to some other paths $\gamma^{(2)}, \ldots, \gamma^{(r)}$ made of three line segments. Glueing these paths, we observe that $\widetilde{\Theta}_{k\to k+1}(\theta_0)$ is connected via paths in $\Theta_{k\to k+1}(\theta_0)$. To finish the case for $m = k+1$, it is enough to show that any point $\theta \in \Theta_{k\to k+1}(\theta_0) \setminus \widetilde{\Theta}_{k\to k+1}(\theta_0)$

$$\theta = P_\pi(w_i, \alpha a_i) \oplus (\alpha a_i, (1-\alpha)a_i) \oplus (w_1, \alpha a_1) \oplus \ldots \oplus (w_k, \alpha a_k)$$

is connected (via a line segment) to a point in $\widetilde{\Theta}_{k\to k+1}(\theta_0)$ which is simply

$$\widetilde{\Theta}_{k\to k+1}(\theta_0) = P_\pi(w', 0) \oplus (w_i, a_i) \oplus (w_1, a_1) \oplus \ldots \oplus (w_k, a_k).$$

Next we will prove for the general case $m \geq k+1$ by induction. We assume that $\Theta_{r\to m}(\theta_0)$ is connected and we will show that $\Theta_{r\to m+1}(\theta_0)$ is also connected. First we show the connectivity of the points in the following set

$$\widetilde{\Theta}_{r\to m+1}(\theta_0) := \{P_\pi \theta^{m+1} : \theta^{m+1} = (\underbrace{w_1, \ldots, w_1}_{k_1}, \ldots, \underbrace{w_r, \ldots, w_r}_{k_r}, \underbrace{w_1', \ldots, w_j'}_{j+1}, \underbrace{a_1^1, \ldots a_1^{k_1}}_{k_1}, \ldots, \underbrace{a_r^1, \ldots a_r^{k_r}}_{k_r}, \underbrace{0, \ldots, 0}_{j+1})$$
$$\text{where } k_i \geq 1, j \geq 0, k_1 + \ldots + k_r + j = m, \sum_{i=1}^{k_j} a_j^i = a_j, \text{and } \pi \in S_{m+1}\}$$

which is a subset of $\Theta_{r\to m+1}(\theta_0)$. From the induction hypothesis, we have the connectivity of the manifold $\Theta_{r\to m}(\theta_0)$.

An element $\widetilde{\theta_0} \in \widetilde{\Theta}_{r\to m+1}(\theta_0)$ can be written as

$$\widetilde{\theta_0} = P_{\tilde{\pi}}(\underbrace{w_1, \ldots, w_1}_{k_1}, \ldots, \underbrace{w_r, \ldots, w_r}_{k_r}, \underbrace{w_1', \ldots, w_j'}_{j}, \underbrace{w_0}_{1}, \underbrace{a_1^1, \ldots a_1^{k_1}}_{k_1}, \ldots, \underbrace{a_r^1, \ldots a_r^{k_r}}_{k_r}, \underbrace{0, \ldots, 0}_{j+1}),$$

for some $j \geq 0$ and $\tilde{\pi} \in S_{m+1}$. For a fixed w_0 at a fixed position, there is a bijection $\widetilde{\Theta}_{r\to m+1}(\theta_0) \to$

$\Theta_{r\to m}(\theta_0)$ that sends $\widetilde{\theta}$ to

$$\theta = P_\pi(\underbrace{w_1,\dots,w_1}_{k_1},\dots,\underbrace{w_r,\dots,w_r}_{k_r},\underbrace{w'_1,\dots,w'_j}_{j},\underbrace{a_1^1,\dots a_1^{k_1}}_{k_1},\dots,\underbrace{a_r^1,\dots a_r^{k_r}}_{k_r},\underbrace{0,\dots,0}_{j})$$

for some $\pi \in S_m$, i.e. $\widetilde{\theta}$ where w_0 and its associated 0 outgoing weight vector have been dropped. In particular, any two points of $\widetilde{\Theta}_{k\to m+1}(\theta_0)$ with the same w_0 component at the same position are connected as a consequence of this correspondence and the connectivity of $\Theta_{k\to m}(\theta_0)$. Moreover, we note that $\widetilde{\theta_0} \in \widetilde{\Theta}_{k\to m+1}(\theta_0)$ is connected via a line segment in $\widetilde{\Theta}_{k\to m+1}(\theta_0)$ to every other point in $\widetilde{\Theta}_{k\to m+1}(\theta_0)$ whose components are the same as $\widetilde{\theta_0}$ except for w_0. This straightforwardly generalizes for different positions of w_0 and this establishes the connectivity of $\widetilde{\Theta}_{k\to m+1}(\theta_0)$.

Finally, we pick a point $\theta \in \Theta_{k\to m+1}(\theta_0)$ that is

$$\theta = P_\pi(\underbrace{w_1,\dots,w_1}_{k_1},\dots,\underbrace{w_r,\dots,w_r}_{k_r},\underbrace{w'_1,\dots,w'_1}_{b_1},\dots,\underbrace{w'_j,\dots,w'_j}_{b_j},\underbrace{a_1^1,\dots a_1^{k_1}}_{k_1},\dots,\underbrace{a_r^1,\dots a_r^{k_r}}_{k_r},\underbrace{\alpha_1^1,\dots,\alpha_1^{b_1}}_{b_1},\dots,\underbrace{\alpha_j^1,\dots,\alpha_j^{b_j}}_{b_j}).$$

for some $\pi \in S_{m+1}$. Note that θ is connected to

$$\widetilde{\theta} = P_\pi(\underbrace{w_1,\dots,w_1}_{k_1},\dots,\underbrace{w_r,\dots,w_r}_{k_r},\underbrace{w'_1,\dots,w'_1}_{b_1},\dots,\underbrace{w'_j,\dots,w'_j}_{b_j},\underbrace{a_1^1,\dots a_1^{k_1}}_{k_1},\dots,\underbrace{a_r^1,\dots a_r^{k_r}}_{k_r},\underbrace{0,\dots,0}_{b_1},\dots,\underbrace{0,\dots,0}_{b_j}),$$

which is in $\widetilde{\Theta}_{k\to m+1}(\theta_0)$. We have shown that all points in $\Theta_{k\to m+1}(\theta_0)$ are connected, which completes the induction step thus the proof.

□

D.2 Scaling Law of the Zero-Loss Manifolds

In order to prove Proposition D.2.2, we introduce the following Lemma D.2.1, which is in fact a counting of the same number in two ways.

Lemma D.2.1. *For $j \le n$, we have*

$$\frac{1}{j!}G(j,n) = \sum_{\substack{c_1+2c_2+\cdots+nc_n=n\\ c_1+c_2+\cdots+c_n=j\\ c_i\ge 0}} \frac{n!}{1!^{c_1}2!^{c_2}\cdots n!^{c_n}}\frac{1}{c_1!\cdots c_n!}.$$

Proof. By definition, we have

$$G(j,n) = \sum_{\substack{b_1+\ldots+b_j=n\\ b_i\ge 1}} \binom{n}{b_1,\dots,b_j}.$$

Starting from a tuple $(b_1,\dots,b_j)$, consider the tuple $(c_1,\dots,c_n)$ where c_i is the number of occurence of i in $(b_1,\dots,b_j)$. Therefore we have

$$\binom{n}{b_1,\dots,b_j} = \binom{n}{\underbrace{1,\dots,1}_{c_1},\underbrace{2,\dots,2}_{c_2},\dots,\underbrace{n}_{c_n}} = \frac{n!}{1!^{c_1}\cdots n!^{c_n}}. \tag{D.1}$$

Moreover, any c-tuple $(c_1,\dots,c_n)$ appears in

$$\binom{j}{c_1,\dots,c_n} = \frac{j!}{c_1!\cdots c_n!} \tag{D.2}$$

b-tuples that are exactly $(b_1,\dots,b_j)$. From Equation (D.1) and Equation (D.2) and summing over all tuples $(c_1,\dots,c_n)$ we conclude. □

Proposition D.2.2. *For $k \le m$, we have*

$$T(k,m) = G(k,m) + \sum_{\ell=1}^{m-k} \binom{m}{\ell} G(k,m-\ell)\, g(\ell) \tag{D.3}$$

where $g(\ell) = \sum_{n=1}^{\ell} \frac{1}{n!} G(n,\ell)$. Moreover, we have that the scaling law T has the same growth as the scaling law G in the following limit for fixed k

$$T(m-k,m) \sim G(m-k,m) \quad \text{as } m \to \infty.$$

Proof. Let $u = b_1 + \cdots + b_j$ and let c_i be, as in Lemma D.2.1, the number of occurrences of i among $(b_1,\dots,b_j)$. Recall that for T we have the identity

$$T(r,m) := \sum_{j=0}^{m-r} \sum_{\substack{(s)=m \\ k_i\ge 1, b_i \ge 1}} \binom{m}{k_1,\dots,k_r,b_1,\dots,b_j} \frac{1}{c_b}.$$

We rewrite the outer summation in T from the number of b_i's to the summation of b_i's and we obtain

$$T(r,m) = \sum_{u=0}^{m-r} \sum_{j=0}^{u} \binom{m}{u} \sum_{\substack{k_1+\cdots+k_r=m-u \\ b_1+\cdots+b_j=u \\ k_i\ge 1, b_i\ge 1}} \binom{m-u}{k_1,\dots,k_r} \binom{u}{b_1,\dots,b_j} \frac{1}{c_1! c_2! \cdots c_{m-r}!}$$

where we split the inner summation and the multinomial coefficient into two parts: one that comes from the incoming weight vectors and the others come from the zero-type neurons $(w_1',\dots,w_j')$. Using the formula for G on $(k_1,\dots,k_r)$, we simplify as follows

$$T(r,m) = \sum_{u=0}^{m-r} \binom{m}{u} G(r,m-u) \sum_{j=0}^{u} \sum_{\substack{b_1+\cdots+b_j=u \\ b_i\ge 1}} \binom{u}{b_1,\dots,b_j} \frac{1}{c_1! c_2! \cdots c_{m-r}!}.$$

Finally using Lemma D.2.1, we find

$$T(r,m) = \sum_{u=0}^{m-r} \binom{m}{u} G(r, m-u) \sum_{j=0}^{u} \frac{1}{j!} G(j,u)$$

where $G(0,0) = 1$. Splitting the case $u = 0$, we derive the closed form formula

$$T(r,m) = G(r,m) + \sum_{u=1}^{m-r} \binom{m}{u} G(r, m-u) \sum_{j=1}^{u} \frac{1}{j!} G(j,u).$$

In order to prove the asymptotic for $T(m-k,m)$ we divide both sides in Equation D.3 (with $r = m-k$) by $G(m-k,m)$:

$$\frac{T(m-k,m)}{G(m-k,m)} = 1 + \sum_{u=1}^{k} \binom{m}{u} \frac{G(m-k,m-u)}{G(m-k,m)} g(u).$$

The limit of $T(m-k,m)$ as $m \to \infty$, is then obtained from the asymptotic of $G(m-k,m)$ above:

$$1 + \sum_{u=1}^{k} \binom{m}{u} \frac{G(m-k,m-u)}{G(m-k,m)} g(u) \sim 1 + \sum_{u=1}^{k} \frac{m^u}{u!} c_u \frac{m^{k-u}(m-u)!}{m^k m!} g(u) \sim 1 + \sum_{u=1}^{k} \frac{g(u)}{u!} \frac{c_u}{m^u} \sim 1$$

hence, for large m, $T(m-k,m)$ and $G(m-k,m)$ grows at the same rate.

□

D.3 Teacher Construction

All tasks with artificial data have d_0-dimensional uniformly distributed input data in the range $x_i \in [-\frac{1}{\sqrt{3}}, \frac{1}{\sqrt{3}}]$. A specific task is defined by the parameters of a teacher network. Each hidden neuron i of the teacher is randomly sampled from a set of input weights $w_i \in \{-1,0,1\}^{d_0}$, output weigths $a_i \in \{-1,1\}$ and biases $b_i \in \{-\frac{2}{3}\sqrt{3}, -\frac{1}{3}\sqrt{3}, 0, \frac{1}{3}\sqrt{3}, \frac{2}{3}\sqrt{3}\}$. We repeat the sampling if two hidden neurons are identical up to the signs of output weights to avoid that two hidden neurons cancel each other. The input weight vectors w are then normalised to unity, then, both w and b are multiplied by a factor of 3. The above procedure yields hyperplanes in direction w located at a distance $|b|/||w||$ from the origin, and a steeply rising (or falling) activation on the positive side of the hyperplane. Finally, analogous to batch normalization, the output weights and bias are scaled such that the output has zero-mean and unit variance when averaged over the input distribution: $a \leftarrow a/\mathrm{std}(y)$ and $b_2 = -\langle y \rangle / \mathrm{std}(y)$, where y is the output vector of the network. We study teachers with input dimensionality $d_0 \in \{2,4,8,16,32\}$ and hidden layer size $k \in \{2,4,8\}$. The above construction of hidden neuron parameter vectors can be generalized to multi-layer teachers by stacking the procedure.

D.4 Deep Neural Networks

In the case of multi-layers, the equivalence of two incoming weight vectors in the intermediate layers should be understood in the general sense, i.e. all incoming weight vectors of layer ℓ are the outgoing weight vectors of layer $\ell-1$ that can be written as

$$\{(\underbrace{(a_1^1)_d,\dots,(a_1^{k_1})_d}_{k_1}),\dots,\underbrace{(a_r^1)_d,\dots,(a_r^{k_r})_d}_{k_r},\underbrace{(\alpha_1^1)_d,\dots,(\alpha_1^{b_1})_d}_{b_1},\dots,\underbrace{(\alpha_1^1)_d,\dots,(\alpha_r^{b_j})_d}_{b_j}):$$
$$\sum_{i=1}^{k_t}(a_t^i)_d=(a_t)_d \text{ and } \sum_{i=1}^{b_t}(\alpha_t^i)_d=0\}$$

where $d\in[k_\ell]$. All weight vectors in this set are equivalent in the sense that they produce the same neuron in layer ℓ.

For the general shape of the multi-layer expansion manifold, let us consider first a three-layer network. If we add one neuron to the first hidden layer, we have that $\Theta^{(1)}_{\mathbf{k}\to\mathbf{m}}(\theta)$ is connected. If we do not add a new neuron in the second hidden layer, the permutations of the neurons in the second hidden layer would bring $k_2!$ disconnected components where each one of the disconnected components have $T(k_1,k_1+1)$ affine subspaces that are connected to each other. Note that in this case the overall manifold $\Theta_{\mathbf{k}\to\mathbf{m}}(\theta)$ is disconnected. However, adding one neuron to the second hidden layer, every $k_2!$ disconnected components get connected through the parameters of the neurons in the second hidden layer, which yields a connected multi-layer expansion manifold $\Theta_{\mathbf{k}\to\mathbf{m}}(\theta)$.

In general, adding h_1 neurons to the first hidden layer results in $T(k_1,k_1+h_1)$ connected affine subspaces instead of the usual $k_1!$ discrete (i.e. disconnected) points. Adding h_2 neurons to the second hidden layer brings $T(k_2,k_2+h_2)$ affine subspaces instead of the usual $k_2!$ points, for each one of the $T(k_1,k_1+h_1)$ affine subspaces. Note that this is multiplicative because every combination of the parameters in the first hidden layer can be paired with every combination of the parameters in the second hidden layer which results in a distinct affine subspace. Similarly, via induction, if $h_\ell\geq 1$ for all $\ell\in[L-1]$, adding $(h_1,\dots,h_{L-1})$ neurons to each one of the hidden layers make a connected manifold of $\prod_{\ell=1}^{L-1}T(k_\ell,k_\ell+h_\ell)$ affine subspaces.

E Neural Networks with Few Neurons

E.1 General Properties of the Interactions

In Appendix E.1 and E.2, the input distribution is standard Gaussian and we will write $\mathbb{E}[\cdot] = \mathbb{E}_{\mathcal{N}(0,I)}[\cdot]$. In this Section, we introduce some general properties of the interactions. We use these only for the one-neuron network in this paper (see Section E.2). However, we expect these properties to play a role in studying the networks with two or more neurons.

We first present the partial derivative of a general interaction function, i.e. two activation functions may be different, for example, if the student activation function does not match the teacher, with respect to the correlation in a simple expression in Lemma E.1.1. In the second part, we present a property of the activation function sufficient for Assumption 6.5.1 (ii), and show that the differentiable activation functions mentioned in this paper satisfy this property in Lemma E.1.2.

Lemma E.1.1. *Assume that functions σ_1 and σ_2 are differentiable. The partial derivative of the following Gaussian integral term $\mathbb{E}[\sigma_1(r_1 x)\sigma_2(r_2 y)]$ with respect to the correlation $\mathbb{E}[xy] = u$ is*

$$\frac{d}{du}\mathbb{E}[\sigma_1(r_1 x)\sigma_2(r_2 y)] = r_1 r_2 \mathbb{E}[\sigma_1'(r_1 x)\sigma_2'(r_2 y)]. \tag{E.1}$$

Proof. We compute the derivative of $\mathbb{E}[\sigma_1(r_1 x)\sigma_2(r_2 y)]$ by making the correlation u explicit. Denote $u' = \sqrt{1-u^2}$ and $y = ux + u'z$. After the computation, we use Stein's lemma to reach the desired formula.

$$\partial_u \mathbb{E}[\sigma_1(r_1 x)\sigma_2(r_2 y)] = r_2 \mathbb{E}[\sigma_1(r_1 x)\sigma_2'(r_2 y)x] - \frac{r_2 u}{u'}\mathbb{E}[\sigma_1(r_1 x)\sigma_2'(r_2 y)z] \tag{E.2}$$

where x and z are independent standard Gaussians. Here is a reminder for Stein's Lemma for a standard Gaussian z

$$\mathbb{E}[v(z)z] = \mathbb{E}[v'(z)]. \tag{E.3}$$

To remove x in the first term, we apply Stein's formula for $v(x) = \sigma_1(r_1 x)\sigma_2'(r_2(ux + u'z))$ yielding

$$r_1 r_2 \mathbb{E}[\sigma_1'(r_1 x)\sigma_2'(r_2 y)] + r_2^2 u \mathbb{E}[\sigma_1(r_1 x)\sigma_2''(r_2 y)]. \tag{E.4}$$

To remove z in the second term, we apply Stein's formula for $v(z) = \sigma_2'(r_2(ux + u'z))$ yielding

$$-r_2^2 u \mathbb{E}[\sigma_1(r_1 x)\sigma_2''(r_2 y)]. \tag{E.5}$$

Summing up the two terms completes the proof. □

For softplus that is increasing and convex, using Lemma E.1.1 for $\sigma_1 = \sigma_2 = \sigma$ twice, we infer that the interaction g is also increasing and convex in u. Hence, for $u < 0$, Assumption 6.5.1 (ii) holds for softplus. However, for the other activation functions, using convexity does not help to show that the assumption holds. We will propose a new property of the activation function that implies that the interaction satisfies Assumption 6.5.1 (ii) and prove that softplus with $\beta \leq 2$, sigmoid, tanh, and erf satisfy this property.

Lemma E.1.2. *If the activation function σ is thrice-differentiable and it satisfies*

$$\sigma'(x) - x\sigma''(x) + \sigma'''(x) > 0, \tag{E.6}$$

then its interaction satisfies Assumption 6.5.1 (ii) for all $u \in (-1, 1)$. Softplus with $\beta \in (0, 2]$, sigmoid, tanh, and erf activation functions satisfy the above inequality.

Proof. Let us first write out Assumption 6.5.1 (ii) explicitly using Lemma E.1.1

$$r_1 u \mathbb{E}[\bar{\sigma}'(r_1 x)\bar{\sigma}'(y)] < \mathbb{E}[\bar{\sigma}(r_1 x)\bar{\sigma}(y)]. \tag{E.7}$$

where $\bar{\sigma}(x) = \sigma'(x)$. Using Stein's Lemma for $v(x) = \bar{\sigma}(r_1 x)\bar{\sigma}'(y)$, we get

$$\mathbb{E}[\bar{\sigma}(r_1 x)\bar{\sigma}'(y)x] = \mathbb{E}[\bar{\sigma}'(r_1 x)\bar{\sigma}'(y)]r_1 + \mathbb{E}[\bar{\sigma}(r_1 x)\bar{\sigma}''(y)]u. \tag{E.8}$$

The desired inequality is equivalent to

$$\mathbb{E}[\bar{\sigma}(r_1 x)(\bar{\sigma}(y) - \bar{\sigma}'(y)xu + \bar{\sigma}''(y)u^2)] > 0. \tag{E.9}$$

Let us introduce $f(x) = \bar{\sigma}(x) - x\bar{\sigma}'(x) + \bar{\sigma}''(x)$. For $y = ux + u'z$ where $u' = \sqrt{1 - u^2}$, we have the conditional average of y fixing x (we drop conditioning on the right-hand terms for

convenience)

$$\begin{aligned}\mathbb{E}[f(y)|x] &= \mathbb{E}[\bar{\sigma}(y)] - \mathbb{E}[y\bar{\sigma}'(y)] + \mathbb{E}[\bar{\sigma}''(y)] \\ &= \mathbb{E}[\bar{\sigma}(y)] - ux\mathbb{E}[\bar{\sigma}'(y)] - \mathbb{E}[u'z\bar{\sigma}'(y)] + \mathbb{E}[\bar{\sigma}''(y)] \\ &= \mathbb{E}[\bar{\sigma}(y)] - ux\mathbb{E}[\bar{\sigma}'(y)] - (u')^2\mathbb{E}[\bar{\sigma}''(y)] + \mathbb{E}[\bar{\sigma}''(y)] \\ &= \mathbb{E}[\bar{\sigma}(y)] - ux\mathbb{E}[\bar{\sigma}'(y)] - u^2\mathbb{E}[\bar{\sigma}''(y)], \end{aligned} \tag{E.10}$$

where the last equality comes from Stein's Lemma for $v(z) = \bar{\sigma}'(ux + u'z)$. Hence the desired inequality is equivalent to

$$\mathbb{E}[\bar{\sigma}(r_1 x) f(y)] > 0. \tag{E.11}$$

By straightforward calculus, we will show that $f(x) > 0$ first for the sigmoid and tanh activation functions, for which we have

$$\bar{\sigma}(x) = \frac{e^x}{(e^x+1)^2},\ \bar{\sigma}'(x) = \frac{e^x(1-e^x)}{(e^x+1)^3},\ \bar{\sigma}''(x) = \frac{e^x(e^{2x}-4e^x+1)}{(e^x+1)^4}. \tag{E.12}$$

Hence, we can explicitly write f as

$$f(x) = \frac{e^x}{(e^x+1)^2} - x\frac{e^x(1-e^x)}{(e^x+1)^3} + \frac{e^x(e^{2x}-4e^x+1)}{(e^x+1)^4} \tag{E.13}$$

$$= \frac{e^x}{(e^x+1)^4}((e^x+1)^2 - x(1-e^x)(e^x+1) + (e^{2x}-4e^x+1)). \tag{E.14}$$

Therefore showing $f(x) > 0$ is equivalent to showing that the factor on the right, that is,

$$2e^x(1-e^x) + 2 - x(1-e^{2x}) \tag{E.15}$$

is positive. For $x < 0$, we have $e^x < 1$ which implies $-x(1-e^{2x}) > 0$ and $(1-e^x)e^x \leq 1/4$ due to the inequality of arithmetic and geometric means hence the first term is upper bounded by $-1/2$ and since we have $+2$, the whole term is positive. For $x \geq 0$, we have $e^x \geq 1$, hence we can rewrite the inequality as a sum of non-negative terms

$$2e^x(e^x-1) + 2 + x(e^{2x}-1) > 0. \tag{E.16}$$

Let us now handle the case of erf. Its first three derivatives are given by

$$\bar{\sigma}(x) = \frac{2}{\sqrt{\pi}}e^{-x^2/2}, \bar{\sigma}'(x) = -\frac{2}{\sqrt{\pi}}xe^{-x^2/2}, \bar{\sigma}''(x) = \frac{2}{\sqrt{\pi}}(x^2e^{-x^2/2} - e^{-x^2/2}) \tag{E.17}$$

Hence, we can explicitly write f as

$$f(x) = \frac{2}{\sqrt{\pi}}e^{-x^2/2}(1 + xx + x^2 - 1) = \frac{4}{\sqrt{\pi}}e^{-x^2/2}x^2 \tag{E.18}$$

that is non-negative for all x and zero iff $x = 0$. Hence, the expectation in Eq. E.11 is positive since $f(y) > 0$ for some y values in the support of $ux + u'z$.

Finally, for the softplus activation function with $\beta \in (0,2]$, we have the following derivatives

$$\bar{\sigma}(x) = \frac{e^{\beta x}}{(e^{\beta x}+1)},\ \bar{\sigma}'(x) = \frac{\beta e^{\beta x}}{(e^{\beta x}+1)^2},\ \bar{\sigma}''(x) = \frac{\beta^2 e^{\beta x}(1-e^{\beta x})}{(e^{\beta x}+1)^3}. \tag{E.19}$$

Plugging in the function f, we get

$$f(x) = \frac{e^{\beta x}}{(e^{\beta x}+1)} - x\frac{\beta e^{\beta x}}{(e^{\beta x}+1)^2} + \frac{\beta^2 e^{\beta x}(1-e^{\beta x})}{(e^{\beta x}+1)^3} \tag{E.20}$$

$$= \frac{e^{\beta x}}{(e^{\beta x}+1)^3}((e^{\beta x}+1)^2 - x\beta(e^{\beta x}+1) + \beta^2(1-e^{\beta x})) \tag{E.21}$$

Therefore showing $f(x) > 0$ is equivalent to showing that the factor on the right, that is,

$$e^{2\beta x} + e^{\beta x}(2 - x\beta - \beta^2) + 1 - x\beta + \beta^2 \tag{E.22}$$

is positive. For $x \leq 0$, we have that $-x\beta > 0$ and $2 - \beta^2 \geq -2$ since $\beta \leq 2$, hence it is sufficient to show that the following is positive

$$e^{2\beta x} - 2e^{\beta x} + 1 + \beta^2 = (e^{\beta x} - 1)^2 + \beta^2 \tag{E.23}$$

which is a sum of squares. For $x > 0$, in the rest of the proof we will show that

$$e^{\beta x}(e^{\beta x} + 2 - x\beta - \beta^2) + 1 - x\beta + \beta^2 > 0, \tag{E.24}$$

for $\beta \in (0,2]$. Using $e^{\beta x} \geq (\beta x)^2/2 + \beta x + 1$, it suffices to show that

$$e^{\beta x}((\beta x)^2/2 + 3 - \beta^2) + 1 - x\beta + \beta^2 > 0. \tag{E.25}$$

If $(\beta x)^2/2 + 3 - \beta^2 \geq 1$, then the first term is bigger than $\beta x + 1$ hence the above term is positive. The remaining possibility is that we have

$$\frac{x^2}{2} < 1 - \frac{2}{\beta^2}. \tag{E.26}$$

$\beta \leq 2$ implies $x < 1$ and $x^2 > 0$ implies $\beta > \sqrt{2}$. Hence we have $-x\beta + \beta^2 > 0$ since $\beta > x$. Therefore, if we have $(\beta x)^2/2 + 3 - \beta^2 \geq 0$, Eq. E.24 is positive. Assuming the opposite, we get

$$\frac{x^2}{2} < 1 - \frac{3}{\beta^2}, \tag{E.27}$$

$\beta \leq 2$ implies $x < 1/\sqrt{2}$ and $x^2 > 0$ implies $\beta > \sqrt{3}$.

Going back to Eq. E.24, what remains to show is that it is positive in the domain $x < 1/\sqrt{2}$,

$\beta \in (\sqrt{3}, 2]$. It suffices to show that $e^{\beta x} + 2 - x\beta - \beta^2 > 0$. Assuming the contrary implies $e^{\beta x} < x\beta + 2$ since $\beta \leq 2$. We can then deduce that $x\beta < c = 1.2$ since otherwise we would have

$$e^{\beta x} = 1 + \beta x + \frac{(\beta x)^2}{2!} + \frac{(\beta x)^3}{3!} + \ldots \tag{E.28}$$

$$\geq 1 + \beta x + \frac{c^2}{2!} + \frac{c^3}{3!} + \ldots = 1 + \beta x + (e^c - c - 1) > 1 + \beta x + 1 \tag{E.29}$$

which implies a contradiction. c can be chosen smaller but this will be enough for our purposes.

Assuming $e^{\beta x} + 2 - x\beta - \beta^2 \leq 0$, let us expand Eq. E.24

$$e^{\beta x}(e^{\beta x} + 2 - x\beta - \beta^2) + 1 - x\beta + \beta^2 \geq \text{ (using } e^{\beta x} < \beta x + 2) \tag{E.30}$$

$$(x\beta + 2)e^{\beta x} + (x\beta + 2)(2 - x\beta - \beta^2) + 1 - x\beta + \beta^2 = \tag{E.31}$$

$$(x\beta + 2)e^{\beta x} - (x\beta)^2 - (1 + \beta^2)x\beta + 5 - \beta^2 > \text{ (using } e^{\beta x} > \beta x + 1) \tag{E.32}$$

$$7 - \beta^2 + (2 - \beta^2)x\beta \geq 3 - 2x\beta > 0 \tag{E.33}$$

where in the last inequality we used $x\beta < 1.2$. We note that this inequality holds for slightly larger β using the same technique, however, for significantly larger β, the property breaks down. □

E.2 The One-Neuron Network

For the one-neuron network, we will characterize all candidate critical points and in particular the optimal solution for the following loss (repeating Eq. 6.17)

$$L = a^2 g(r, r, 1) - 2a \sum_{i=1}^{k} g(r, 1, u_i) + \text{const}, \quad \text{subject to } r \geq 0, \sum_{i=1}^{k} u_i^2 \leq 1, \tag{E.34}$$

where the constant represents the sum of the teacher-teacher interactions. Let us denote the unit ball by $B = \{(u_1, \ldots, u_k) \mid u_1^2 + \ldots + u_k^2 \leq 1\}$. Its interior is denoted by $\text{int}\, B$ and its boundary is denoted by ∂B. By characterization, we either mean finding a closed-form expression when the interaction g has an analytic formula (for ReLU), or finding the exact formula for the correlations and bounding the incoming vector norm r and the outgoing weight a (for softplus). For general activation functions, we will numerically show that there is a unique critical point (up to a sign flip for the odd activation functions and up to scaling for ReLU) in Subsection E.2.1.

In this paper, we work with the following necessary conditions for a critical point of a loss in Eq. 6.17 defined on the domain $D = \mathbb{R}_{\geq 0} \times \mathbb{R} \times B$. The conditions we describe hold generally for a domain that is differentiable, bounded, and closed in some coordinates such as a unit ball.

Necessary Condition E.2.1. *We say p is a candidate critical point if for any path $\gamma(t) \in D$ for*

$t \in (-\epsilon, \epsilon)$ for some $\epsilon > 0$ such that $\gamma(0) = p$ and $L(\gamma(t))$ is differentiable, we have

$$\frac{d}{dt} L(\gamma(t))\big|_{t=0} = 0. \tag{E.35}$$

This condition gives a set of candidate critical points of the loss in Eq. 6.17. Reversing the argument, a point $p = (a, r, (u_i)_{i=1}^k)$ is not a critical point, if there exists a path $\gamma(t) \in B$ for $t \in (-\epsilon, \epsilon)$ for some $\epsilon > 0$ such that $\gamma(0) = p$ and

$$\frac{d}{dt} L(\gamma(t))\big|_{t=0} \neq 0, \tag{E.36}$$

that the derivative exists and is non-zero. This implies that any equivalent weight space parameter

$$\theta = (w, a) = \left(r \left(\sum_{i=1}^{k} u_i v_i + v_\perp \right), a \right) \tag{E.37}$$

is not a critical point of L_{orig}, where $v_\perp \in \mathbb{R}^d$ is an arbitrary vector perpendicular to all teacher incoming vectors $(v_1, ..., v_k)$. To see this, let us consider the following path in the weight space

$$\theta(t) = \left(r(t) \left(\sum_{i=1}^{k} u_i(t) v_i + v_\perp \right), a(t) \right). \tag{E.38}$$

Thanks to the equivalence of the losses, we have that

$$\frac{d}{dt} L_{\text{orig}}(\theta(t))\big|_{t=0} = \frac{d}{dt} L(\gamma(t))\big|_{t=0} \neq 0, \tag{E.39}$$

which implies that $\theta(0) = \theta$ is not a critical point in the weight space. Overall, the only possible critical points of the weight space are those that are equivalent to the points that satisfy the necessary conditions for the critical points of the loss in Eq. 6.17.

E.2.1 General Activation Functions

In Subsection E.2.1, we will prove Theorem 6.5.2, that is, for general activation functions satisfying Assumption 6.5.1, any non-trivial critical point of the one-neuron network attains equal correlations. In Subsection E.2.1, we will apply Theorem 6.5.2 and obtain a two-dimensional loss. From the derivative constraints of the two-dimensional loss, we get a fixed point equation that needs to be satisfied by the incoming vector norm r at any critical point. Finally, we numerically show that there is a unique solution to the fixed point equation for general activation functions.

Before we present the proof, let us take a detour to check the applicability of the convex optimization framework. For a convex and twice-differentiable activation function such as softplus, applying Lemma E.1.1 twice implies that the interaction $g(r_1, r_2, \cdot)$ is a convex

function of the correlation $u \in (-1,1)$ for $r_1, r_2 > 0$. Let us consider a fixed $a < 0$ and $r > 0$ and consider the loss parameterized by u_i's. It is convex since its Hessian is a diagonal matrix with entries

$$\frac{d^2}{du_i^2} L = -2a \frac{d^2}{du_i^2} g(r, 1, u_i) > 0. \tag{E.40}$$

Since the constraint on the correlations (Eq. 6.17) is also convex, we get a convex optimization problem that has a unique global minimum (see S. Boyd, S. P. Boyd, and Vandenberghe, 2004 Section 4.2). Swapping a pair of u_i does not change the loss, thus it is permutation symmetric. If any two u_i were distinct from each other at the minimum, then its permutation would also be a minimum which would violate the unicity. We conclude that at the unique minimum point, the correlations are equal to each other. However, for the case $a > 0$, and for other activation functions, the loss is not convex and there can be arbitrarily many minima. We instead use Lagrange multipliers for proving Theorem 6.5.2.

Proof of Theorem 6.5.2

Proof. (Step 1) We first show that given $a, r \neq 0$, there is no critical point in $\operatorname{int} B$. Assume that $u = (u_1, ..., u_k) \in \operatorname{int} B$. Then we have $u_i \in (-1,1)$. Any critical point inside the boundary should have zero gradients hence

$$\frac{d}{du_i} g(r, 1, u_i) = 0 \tag{E.41}$$

since $a \neq 0$. From Assumption 6.5.1 (i) we have that $\partial_u g(r, 1, u) > 0$ for $u \in (-1,1)$ that yields a contradiction. Thus, any critical point of the loss in Eq. 6.17 is on the boundary, i.e. $u_1^2 + ... + u_k^2 = 1$. If $k = 1$, this implies that $u_1 = -1$ or $u_1 = 1$. Next, we consider the case $k > 1$.

(Step 2) Any critical point of the loss in Eq. 6.17 on the boundary should be a critical point of the loss projected on the boundary too. We can see this as a consequence of the necessary conditions E.2.1 as follows. Let us consider fixed $a \neq 0$ and $r \neq 0$. If $p \in \partial B$ satisfies the necessary conditions, we have that, for any differentiable path on the boundary, i.e. $\gamma(t) \in \partial B$ for $t \in (-\epsilon, \epsilon)$ for some $\epsilon > 0$

$$\frac{d}{dt} L(\gamma(t))\big|_{t=0} = \nabla L(p) \cdot \gamma'(0) = 0. \tag{E.42}$$

which implies that $\nabla L(p)$ is orthogonal to all $\gamma'(0)$. The vector that is orthogonal to all $\gamma'(0)$ is the gradient of the surface, that is $2(u_1, ..., u_k)$. Hence we get that $\nabla L(p) \parallel p$ which is equivalent to the Lagrange multiplier condition. In particular, we get the following Lagrangian

$$\mathscr{L}(u, \lambda) = -2a \sum_{i=1}^{k} g(r, 1, u_i) + \lambda \left(\sum_{i=1}^{k} u_i^2 - 1 \right) \tag{E.43}$$

which implies the following condition at any critical point of $\mathscr{L}$

$$-2a\partial_u g(r,1,u_i) + 2\lambda u_i = 0 \;\; \forall i \in [k], \quad \sum_{i=1}^{k} u_i^2 = 1. \tag{E.44}$$

If $u_i = 0$, we get $\partial_u g(r,1,0) = 0$ which is not possible since $g(r,1,u)$ is increasing at $u = 0$ due to Assumption 6.5.1 (i). Hence we have

$$\frac{\partial_u g(r,1,u_i)}{u_i} = \frac{\lambda}{a}. \tag{E.45}$$

Let us observe that $\partial_u g(r,1,u)/u$ is decreasing for $u \in (-1,1) \setminus \{0\}$ if

$$\frac{d}{du}\left(\frac{1}{u}\frac{d}{du}g(r,1,u)\right) = \frac{1}{u}\frac{d^2}{du^2}g(r,1,u) - \frac{1}{u^2}\frac{d}{du}g(r,1,u) < 0, \tag{E.46}$$

which is equivalent to Assumption 6.5.1 (ii) for $u \in (-1,1) \setminus \{0\}$ (we included $u = 0$ in Assumption 6.5.1 (ii) for a simpler statement which is already implied from Assumption 6.5.1 (i) at $u = 0$). We note that $\partial_u g(r,1,u)/u$ is negative for $u < 0$ and positive for $u > 0$ due to Assumption 6.5.1 (i).

Taken together, we conclude that $\partial_u g(r,1,u)/u$ is injective in $u \in (-1,1) \setminus \{0\}$. We need to consider the remaining case $u_i \in \{-1,1\}$. In this case, necessarily, we have $u_j = 0$ for $j \neq i$, which is not possible since $\partial_u g(r,1,u) > 0$ at $u = 0$, yielding a contradiction in Eq. E.44. Thus, Eq. E.45 implies that all correlations are equal. Combining it with the boundary condition, we get $u_1 = ... = u_k = u$ with $ku^2 = 1$, which completes the proof. □

Two-Dimensional Loss, The Derivative Constraints, Uniqueness

At any non-trivial critical point with $a \neq 0$ and $r \neq 0$, we proved in Theorem 6.5.2 that all correlations are equal. Let us denote it by u that is either $1/\sqrt{k}$ or $-1/\sqrt{k}$ as shown in Theorem 6.5.2. Therefore, the loss in Eq. 6.17 at a critical point reduces to

$$L = a^2 g(r,r,1) - 2kag(r,1,u) + \text{const.} \tag{E.47}$$

Moreover, at a critical point, the partial derivatives with respect to the outgoing weight and norm should also be zero which gives the following two constraints

$$\partial_a L = 2ag(r,r,1) - 2kg(r,1,u) = 0, \tag{E.48}$$

$$\partial_r L = a^2 \partial_r g(r,r,1) - 2ka\partial_r g(r,1,u) = 0, \tag{E.49}$$

which can be rearranged into the following (assuming $g(r,r,1) \neq 0$ and $\partial_r g(r,r,1) \neq 0$)

$$\frac{a}{k} = \frac{g(r,1,u)}{g(r,r,1)} = \frac{2\partial_r g(r,1,u)}{\partial_r g(r,r,1)}. \tag{E.50}$$

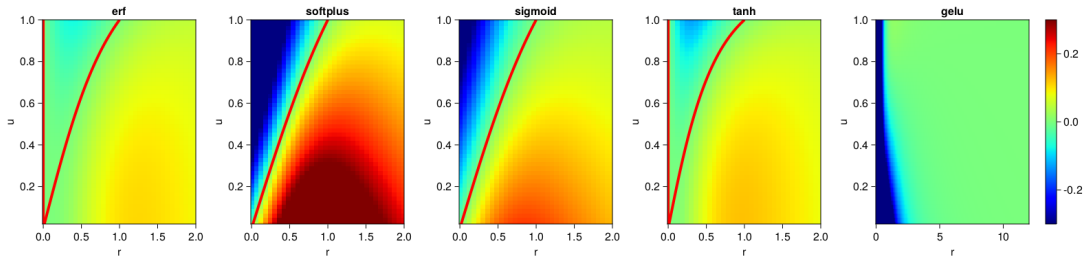


Figure E.1 – *The graph of* $f(r,u) = \frac{d}{dr}\left(\frac{1}{2}\log g(r,r,1) - \log g(r,1,u)\right)$ *for activation functions erf, softplus with* $\beta = 1$*, sigmoid, tanh, and gelu respectively. Zero crossings of* f *are shown in red.* For softplus and sigmoid, we observe that f is negative for $r = 0, u \in (0,1)$, positive for $r = 1, u \in (0,1)$, and increasing in $r \in [0,1]$ for any fixed u, thus satisfying the sufficient condition in Eq. E.54. However, for tanh and erf, f shows non-monotonic behavior in r when u is close to 1. For the GeLU activation function $\sigma(x) = x\Phi(x)$, which is non-monotonic, we observe that f does not cross zero for any (u,r) pair in the plotted domain. It approaches zero from below when $r \to \infty$ thus showing very different behavior from the other activation functions.

The second equality between the two ratios of Gaussian integral terms gives a fixed point equation on the norm r. Rearranging the terms in Eq. E.50 and writing the interactions explicitly, we get

$$f(u,r) = \frac{\mathbb{E}[\sigma'(rx)\sigma(rx)x]}{\mathbb{E}[\sigma(rx)^2]} - \frac{\mathbb{E}[\sigma'(rx)\sigma(y)x]}{\mathbb{E}[\sigma(rx)\sigma(y)]} \tag{E.51}$$

where x and y are standard Gaussians with correlation $\mathbb{E}[xy] = u$. Let us define the following helper functions

$$\begin{aligned} G(r) &= \frac{\mathbb{E}[\sigma'(rx)\sigma(rx)x]}{\mathbb{E}[\sigma(rx)^2]} = \frac{1}{2}\frac{d}{dr}\log(\mathbb{E}[\sigma(rx)^2]), \\ \tilde{G}(u,r) &= \frac{\mathbb{E}[\sigma'(rx)\sigma(y)x]}{\mathbb{E}[\sigma(rx)\sigma(y)]} = \frac{d}{dr}\log(\mathbb{E}[\sigma(rx)\sigma(y)]), \end{aligned} \tag{E.52}$$

which yields

$$f(u,r) = G(r) - \tilde{G}(u,r) = \frac{d}{dr}\log\left(\frac{\mathbb{E}[\sigma(rx)^2]^{\frac{1}{2}}}{\mathbb{E}[\sigma(rx)\sigma(y)]}\right). \tag{E.53}$$

Consider the case $u > 0$. We would like to show that for any given $u \in (0,1)$ there is a unique $r \in (0,1)$ such that $f(u,r) = 0$. A sufficient condition is that for any $u \in (0,1)$,

$$\text{(i)}\ \frac{\sigma'(0)}{\sigma(0)}\frac{\mathbb{E}[\sigma(y)x]}{\mathbb{E}[\sigma(y)]} > 0, \quad \text{(ii)}\ \frac{\mathbb{E}[\sigma'(x)\sigma(x)x]}{\mathbb{E}[\sigma(x)^2]} > \frac{\mathbb{E}[\sigma'(x)\sigma(y)x]}{\mathbb{E}[\sigma(x)\sigma(y)]}, \tag{E.54}$$

$$\text{(iii)}\ \frac{d^2}{dr^2}\log\left(\frac{\mathbb{E}[\sigma(rx)^2]^{\frac{1}{2}}}{\mathbb{E}[\sigma(rx)\sigma(y)]}\right) > 0. \tag{E.55}$$

Note that the first two conditions are equivalent to $f(u,0) < 0$ and $f(u,1) > 0$, respectively. The tricky part is the third condition which is equivalent to showing that

$$\frac{\mathbb{E}[\sigma(rx)\sigma(y)]}{\mathbb{E}[\sigma(rx)^2]^{\frac{1}{2}}} \tag{E.56}$$

is log-concave in r. We note that marginalization properties of log-concave functions may be helpful here. In this paper, we were not able to prove the sufficient conditions listed above, even for softplus which we studied in detail. Instead, we are presenting the numerical integration results, which show that for any given $u \in (0,1)$, there is a unique $r \in (0,1)$ such that $f = 0$ (see Fig. E.1). Once r is shown to be unique, then the matching outgoing weight a follows from Eq. E.50.

In the rest of Section E.2, we will focus on ReLU (Subsection E.2.2). We can fully characterize the critical point of the loss in Eq. 6.17 using the analytic expression of the interaction.

E.2.2 Exact Closed-Form Solution for the ReLU Activation

We will first show that the interaction of ReLU satisfies

$$\begin{aligned}
&\text{(i) } h'(u) > 0 \text{ for } u \in (-1,1),\\
&\text{(ii) } h''(u)u < h'(u) \text{ for } u \in (-1,u_0],\\
&\text{(iii) } \frac{h'(u_0)}{u_0} > \frac{h'(u)}{u} \text{ for } u \in (u_0,1),
\end{aligned} \tag{E.57}$$

where $u_0 = 1/\sqrt{2}$ (note that u_0 can be chosen bigger but this will be sufficient for our purposes). Note that property (i) is the same as Assumption 6.5.1 (i), property (ii) is almost the same as Assumption 6.5.1 (ii) except that it holds in the interval $(-1,u_0]$. Finally, property (iii) covers up for the missing piece of the interval in property (ii).

ReLU interaction satisfies Properties E.57; Proof. Let us write the first two derivatives of h:

$$h'(u) = \frac{\pi - \arccos(u)}{2\pi}, \quad h''(u) = \frac{1}{2\pi\sqrt{1-u^2}}. \tag{E.58}$$

Property (i) easily comes from noting that the derivative of h is positive for $u \in (-1,1)$. Property (ii) holds for $u \in (-1,0]$ since both the first and second derivatives are positive. Let us show that Property (ii) holds for $u \in (0,u_0]$, that is equivalent to

$$\frac{u}{\sqrt{1-u^2}} < \pi - \arccos(u) = \frac{\pi}{2} + \arcsin(u). \tag{E.59}$$

Let us note that the left-hand side is smaller than 1 since

$$\frac{u^2}{1-u^2} \leq 1 \tag{E.60}$$

due to $u^2 \leq 1/2$. Note that $\arcsin(u) > 0$ for $u > 0$ and $\pi/2 > 1$. This completes the proof of Property (ii). To show Property (iii), we first show that $h'(u)/u$ is convex in $u \in [0,1)$. The first two derivatives are

$$\frac{d}{du}\left(\frac{h'(u)}{u}\right) = \frac{h''(u)}{u} - \frac{h'(u)}{u^2}, \quad \frac{d^2}{du^2}\left(\frac{h'(u)}{u}\right) = \frac{h'''(u)}{u} - \frac{2h''(u)}{u^2} + \frac{2h'(u)}{u^3}. \tag{E.61}$$

Thus it is equivalent to showing

$$h'''(u)u - 2h''(u) + \frac{2h'(u)}{u} = \frac{u^2}{(1-u^2)^{3/2}} - \frac{2}{(1-u^2)^{1/2}} + \frac{\pi + 2\arcsin(u)}{u} > 0. \tag{E.62}$$

Using the Taylor series of arcsin and since $u \geq 0$, we have that $\arcsin(u) \geq u$. Hence, it suffices to show

$$\frac{1}{(1-u^2)^{1/2}}\left(-3 + \frac{1}{1-u^2} + 2(1-u^2)^{1/2}\right) \geq 0. \tag{E.63}$$

This holds due to the inequality of arithmetic and geometric means

$$\frac{1}{1-u^2} + (1-u^2)^{1/2} + (1-u^2)^{1/2} \geq 3. \tag{E.64}$$

Let us assume the contrary of Property (iii), that there exists $u \in (u_0, 1)$ such that

$$\frac{h'(u_0)}{u_0} \leq \frac{h'(u)}{u}. \tag{E.65}$$

Note that $h'(u_0)/u_0 > h'(1)$ because $\pi(1-u_0) - \arccos(u_0) \geq 0$ holds at $u_0 = 1/\sqrt{2}$. Since $h'(u)/u$ is continuous at $u = 1$, there exists an $\epsilon > 0$ such that

$$\frac{h'(u_0)}{u_0} > \frac{h'(1-\epsilon)}{1-\epsilon}. \tag{E.66}$$

Finally, using the convexity of $h'(u)/u$, there exists $\alpha \in [0,1]$ such that

$$\alpha\frac{h'(1-\epsilon)}{1-\epsilon} + (1-\alpha)\frac{h'(u_0)}{u_0} \geq \frac{h'(u)}{u}, \tag{E.67}$$

which yields a contradiction since the left-hand side is strictly smaller than $h'(u_0)/u_0$. This completes the proof of Property (iii). *ReLU interaction satisfies Properties E.57; End of Proof.*

Next, we will replicate the proof steps of Theorem 6.5.2 to show that any non-trivial critical point must be on the boundary and attain equal correlations. From Property E.57 (i), we get that there is no non-trivial critical point in int B (see the proof of Theorem 6.5.2). For $k = 1$, this implies that $u_1 = -1$ or $u_1 = 1$. For general k, let us recall that we get the following conditions from Lagrange multipliers (equivalently, from the necessary conditions E.2.1) for non-trivial

critical points, i.e. $ar \neq 0$,

$$-2arh'(u_i) + 2\lambda u_i = 0 \ \ \forall i \in [k], \quad \sum_{i=1}^{k} u_i^2 = 1. \tag{E.68}$$

Note that this is equivalent to Eq. E.44 if the activation function is ReLU. $u_i = 0$ is not possible since we have $h'(0) \neq 0$. Hence, we get

$$\frac{h'(u_i)}{u_i} = \frac{\lambda}{ar}. \tag{E.69}$$

Property E.57 (ii) implies that $f(u) = h'(u)/u$ is decreasing for $u \in (-1, u_0) \setminus \{0\}$. Moreover, f is negative for $u < 0$ and positive for $u > 0$. Therefore, if $\lambda/(ar) < 0$, we get that all u_i are equal and negative, hence they are all $-1/\sqrt{k}$. If $\lambda/(ar) = 0$, we get $u_i = -1$ for all i which implies that $k = 1$ which is already covered above.

The remaining case is $\lambda/(ar) > 0$. Property E.57 (iii) gives that $f(u_0) > f(u)$ for $u \in (u_0, 1]$. Since f is decreasing we have also $f(u) > f(u_0)$ for $u \in (0, u_0)$. Therefore, $f(u_i)$ are equal to each other only when all $u_i < u_0$ or $u_i > u_0$. However, the latter case is not possible for $k \geq 2$ since it breaks the ball constraint, i.e. $u_1^2 + u_2^2 > 1$. Hence, we get that all $u_i \in (0, u_0]$ and they are equal since f is decreasing in this interval, implying that they are all $1/\sqrt{k}$. This completes the proof of replica of Theorem 6.5.2 for the ReLU interaction.

Trivial critical points. Consider the loss in Eq. 6.17. Setting the partial derivative with respect to a to zero, we get

$$ar^2 h(1) = r \sum_{i=1}^{k} h(u_i) \tag{E.70}$$

Let us consider the case $a = 0$. This implies either (i) $r = 0$ or (ii) $k = 1$ and $u_1 = -1$ since $h(u) > 0$ for $u > -1$. Using the factorization property of ReLU interactions and setting the partial derivative with respect to r to zero, we get

$$a^2 rh(1) = a \sum_{i=1}^{k} h(u_i) \tag{E.71}$$

which is satisfied if $a = 0$ hence both (i) and (ii) are critical points of the loss in Eq. 6.17. Since a and r are symmetric to each other, we get that (iii) $r = 0$, $k = 1$ and $u_1 = -1$ is also a critical point of the loss in Eq. 6.17. For general k, we have a candidate critical point at $a = 0$ and $r = 0$, and for the case $k = 1$, we have a candidate critical point at $a = 0$ (or $r = 0$) and $u_1 = -1$. In either case, the loss is equivalent to

$$L_{\text{orig}} = \mathbb{E}[(\sum_{i=1}^{k} \sigma(v_i \cdot x))^2] = kh(1) + k(k-1)h(0). \tag{E.72}$$

Non-trivial critical points with $u_i = u$ that is either $-1/\sqrt{k}$ or $1/\sqrt{k}$.

Let us first show that $(-1/\sqrt{k})_{i=1}^k$ and $(1/\sqrt{k})_{i=1}^k$ are the global minimum and the global maximum of the following constrained loss function

$$\sum_{i=1}^k h(u_i) \quad \text{subject to} \quad \sum_{i=1}^k u_i^2 \leq 1. \tag{E.73}$$

Note that there is no other critical point due to the Lagrange multiplier analysis, hence these are the only two critical points of the constrained loss function. Since h is strictly convex, we have that

$$h(u_1) + h(u_2) > 2h\left(\frac{u_1 + u_2}{2}\right) \tag{E.74}$$

if $u_1 \neq u_2$. Note that we have $u_1^2 + u_2^2 > ((u_1 + u_2)/2)^2$. Therefore, at the global minimum, all correlations must be equal. Then the minimum of $kh(u)$ is attained at the minimum of the u that is feasible, that is $-1/\sqrt{k}$. We will next show that the loss is decreasing locally on any path from $(1/\sqrt{k}, ..., 1/\sqrt{k})$ to another point $(u_1, ..., u_k) \in B$. More precisely, let

$$\gamma(t) = (1/\sqrt{k}, ..., 1/\sqrt{k}) + t(u_1 - 1/\sqrt{k}, ..., u_k - 1/\sqrt{k}). \tag{E.75}$$

We have that

$$\frac{d}{dt}\sum_{i=1}^k h(u_i(t))\bigg|_{t=0} = h'(1/\sqrt{k})\sum_{i=1}^k (u_i - 1/\sqrt{k}) \tag{E.76}$$

which is negative since $(u_1 + ... + u_k)^2 < (u_1^2 + ... + u_k^2)/k = 1/k$ unless u_i are all identical and equal to $-1/\sqrt{k}$. In either case, the above derivative is negative implying that $(1/\sqrt{k}, ..., 1/\sqrt{k})$ is a local maximum. Since there is no other local maximum, it is also the global maximum.

Next, we will give the closed-form solution of the other parameters. Plugging in the correlation in the loss and using the factorization of the interaction in Eq. 6.17, we get

$$L = a^2 r^2 \cdot h(1) - 2kar \cdot h(u) + \text{const.} \tag{E.77}$$

Let us set $\tilde{a} = ar$. The loss is a second-order polynomial in $\tilde{a}$

$$L = h(1)\left(\tilde{a}^2 - 2\tilde{a}k\frac{h(u)}{h(1)} + k + k(k-1)\frac{h(0)}{h(1)}\right) \tag{E.78}$$

where we also made the constant explicit. Since the coefficient of the leading term is positive, there is a minimizer and it is the only critical point. Taking the derivative, the minimum is attained at

$$\tilde{a}_* = k\frac{h(u)}{h(1)} \geq 0. \tag{E.79}$$

Finally, plugging in $\tilde{a}_*$, we get

$$L = -k^2 \frac{h(u)^2}{h(1)} + kh(1) + k(k-1)h(0). \tag{E.80}$$

For $u = 1/\sqrt{k}$, $h(u)$ is non-zero hence the above loss is smaller than the loss at the trivial critical points. We conclude that this is the optimal solution for the one-neuron network.

Berfin Şimşek

bsimsek.com | berfin.simsek@epfl.ch | bsimsek13 | berfinsimsek

Education

Ph.D. in Computer and Communication Sciences 2018-July 2023
École Polytechnique Fédérale de Lausanne (EPFL)
under the supervision of Prof. Clément Hongler & Prof. Wulfram Gerstner

B.S. in Electrical-Electronics Engineering & Mathematics (Double Major) 2013-2018
Koç University, Istanbul
with an exchange period at University of California, Los Angeles (UCLA) Jan-June 2017

Experience

Meta AI Paris June-Oct 2021
Research Intern, supervised by Levent Sagun

UC Berkeley, Helen Wills Neuroscience Institute Jan-Aug 2017
Undergraduate intern, supervised by Vivek Athalye & Jose Carmena

Awards

G-Research PhD Thesis Award, 1st prize EPFL Lausanne 2023
International Mathematical Olympiad (IMO), Bronze Medal Argentina 2012, Colombia 2013
European Girls Mathematical Olympiad, Gold Medal England 2012, Luxembourg 2013
Junior Balkan Mathematical Olympiad, Silver Medal Bosnia and Herzegovina 2019

Publications (Google Scholar)

1. Berfin Şimşek, Amire Bendjeddou, Wulfram Gerstner, and Johanni Brea. *Should Under-parameterized Student Networks Copy or Average Teacher Weights?*. under review at NeurIPS 2023.
2. Flavio Martinelli, Berfin Şimşek, Johanni Brea, and Wulfram Gerstner. *Expand-and-Cluster: Exact Parameter Recovery of Neural Networks*. under review at NeurIPS 2023.
3. Johanni Brea, Flavio Martinelli, Berfin Şimşek, and Wulfram Gerstner. *MLPGradientFlow: Going with the Flow of Multilayer Perceptrons (and Finding Minima Fast and Accurately)*. [arXiv link]
4. Berfin Şimşek, Melissa Hall, and Levent Sagun. *Understanding Out-of-distribution Accuracies through Quantifying Difficulty of Test Samples*. [arXiv link]
5. Berfin Şimşek, François Ged, Arthur Jacot, Francesco Spadaro, Clément Hongler, Wulfram Gerstner, and Johanni Brea. *Geometry of the Loss Landscape in Overparameterized Neural Networks: Symmetries and Invariances*. ICML 2021. [conference paper]
6. Hugo Fabregues, Berfin Şimşek. Overfitting of Polynomial Regression with Overparameterization. Overparameterization: Pitfalls and Opportunities Workshop, ICML 2021.
7. Arthur Jacot, François Ged, Berfin Şimşek, Clément Hongler, and Franck Gabriel. *Saddle-to-Saddle Dynamics in Deep Linear Networks: Small Initialization Training, Symmetry, and Sparsity*. [arXiv link]
8. Arthur Jacot*, Berfin Şimşek*, Francesco Spadaro, Clément Hongler, and Franck Gabriel. *Implicit Regularization of Random Feature Models*. ICML 2020. [conference paper]

9. Arthur Jacot, Berfin Şimşek, Francesco Spadaro, Clément Hongler, and Franck Gabriel. *Kernel Alignment Risk Estimator: Risk Prediction from Training Data.* NeurIPS 2020. [conference paper]

10. Johanni Brea*, Berfin Şimşek*, Bernd Illing, and Wulfram Gerstner. *Weight-space Symmetry in Deep networks Gives Rise to Permutation Saddles, Connected by Equal-loss Valleys across the Loss Landscape.* [arXiv link]

11. Berfin Şimşek, Alper Erdoğan. *Online Bounded Component Analysis: A Simple Recurrent Neural Network with Local Update Rule for Unsupervised Separation of Dependent and Independent Sources.* Asilomar 2019. [conference paper]

Talks & Presentations

Youth in High Dimensions, *ICTP Trieste*, invited speaker	June 2023
Theoretical Physics for Machine Learning, *Aspen Center for Physics*, poster	Feb 2023
Towards a Theory of Artificial and Biological Neural Networks Workshop, *Les Houches*, poster	Feb 2023
Workshop on Spin Glasses, *SwissMAP Research Station*, poster	Sept 2022
Joint Workshop on Machine Learning, *EPFL + RIKEN*, invited talk	Sept 2022
Meta AI Research Visit, *Paris*, group seminar	Aug 2022
Summer School on Statistical Physics and Machine Learning, *Les Houches*	July 2022
Math Machine Learning seminar, *MPI MiS + UCLA*, invited talk	April 2022
Loss Landscape of Neural Networks Symposium, *EPFL*, organizer & talk	Feb 2022
Mathematics of Deep Learning reading group, *ELLIS*, invited talk	Feb 2022
International Conference of Machine Learning, *online*, poster & short talk	July 2021
Neural Net Theory Group, *EPFL*, talk	Feb 2020
Theoretical Advances in Deep Learning Workshop, *Istanbul*, talk	Aug 2019
Machine Learning Summer School (MLSS), *London*, poster	July 2019

Reviewing

Transactions on Machine Learning Research (TMLR)	active
ICML	2021, 2022, 2023
ICLR	2022
NeurIPS	2021, 2023
Science and Engineering of Deep Learning Workshop ICLR	2021
Science meets Engineering of Deep Learning Workshop NeurIPS	2019

Supervision & Teaching

Master Thesis, Oğuz Kaan Yüksel [website], joint supervision with Arthur Jacot	Spring 2022
Master Thesis, Theodor Stoican	Spring 2022
Semester Project, Zhengqing Wu	Spring 2022
Bachelor Thesis, Hugo Fabregues	Spring 2021
Master Thesis, Manu Halvagal [website], joint supervision with Bernd Illing	Fall 2020
Artificial Neural Networks, Master, TA	Spring 2019-20
Artificial Neural Networks/Reinforcement Learning, Master, head TA	Spring 2021-22
Linear Algebra, TA	Fall 2019-20

Skills & Other Activies

- **Languages:** Turkish (native), English (fluent), French & Spanish (A2-B1)
- **Programming:** Python, Pytorch, Julia, LaTeX, Matlab, Java
- **Argentine Tango:** (2013-present) Social Dancer, Instructor & Performer between 2017-2019
- **Non-profit, Crossing Paths:** (2017-2019) created a program for Turkish undergrads in need of funding for an internship abroad & provided career mentorship

Last updated: June 19, 2023