Thèse n° 10 948

Topics in statistical physics of high-dimensional machine learning

Présentée le 24 juin 2024

Faculté des sciences de base
Laboratoire de physique statistique des systèmes computationnels (SB/IC)
Programme doctoral en physique

pour l'obtention du grade de Docteur ès Sciences

par

Hugo Chao CUI

Acceptée sur proposition du jury

Prof. L. Villard, président du jury
Prof. L. Zdeborová, directrice de thèse
Prof. G. Biroli, rapporteur
Prof. J. Bruna, rapporteur
Prof. N. Flammarion, rapporteur

École
polytechnique
fédérale
de Lausanne

2024

ABSTRACT

In the past few years, Machine Learning (ML) techniques have ushered in a paradigm shift, allowing the harnessing of ever more abundant sources of data to automate complex tasks. The technical workhorse behind these important breakthroughs arguably lies in the use of artificial neural networks to learn informative and actionable representations of data, from data. While the number of empirical successes accrues, a solid theoretical comprehension of the unreasonable effectiveness of ML methods in learning from high-dimensional data still proves largely elusive. This is the question addressed in this thesis, through the study of solvable models in high dimensions, satisfying the dual requirement of (a) capturing the key features of practical ML tasks while (b) remaining amenable to mathematical analysis. Borrowing ideas from statistical physics, this thesis presents sharp asymptotic incursions into a selection of central aspects of modern ML.

The remarkable versatility of ML models lies in their ability to extract informative features from data. The first part of the thesis delves into analyzing which structural characteristics of these features condition the learning of ML methods. Specifically, it highlights how, in several settings, a theory formulated in terms of two statistical descriptors can tightly capture the learning curves of simple real tasks. For kernel methods in particular, this insight enables one to relate the error scaling laws to the structure of the features.

The second part then refines the focus to study which features are extracted in multi-layer neural networks, both (a) when untrained and (b) when trained in the framework of Bayesian learning, or after one large gradient step. In particular, it delineates cases in which Gaussian universality holds and limits the network expressivity, and cases in which neural networks succeed in learning non-trivial features.

Finally, supervised learning tasks with fully-connected architectures constitute but a small part of the zoology of modern ML tasks. The last part of the thesis opens up the sharp asymptotic explorations to some modern aspects of the discipline, in particular transport-based generative models, and dot-product attention mechanisms.

Keywords – *Machine Learning, Statistical Physics, High-dimensional asymptotics, Deep Neural Networks, Random Features, Gaussian Universality, Kernels, Attention mechanisms, Generative models.*

RÉSUMÉ

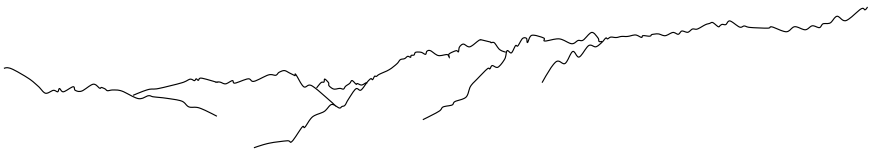
Les techniques d'Apprentissage Machine (AM) permettent d'exploiter des données toujours plus abondantes afin d'automatiser des tâches complexes. Elles reposent en grande partie sur l'utilisation de réseaux de neurones artificiels pour extraire des représentations informatives des données. Alors que le nombre de succès empiriques accroît, une compréhension théorique de la surprenante capacité des méthodes d'AM à apprendre à partir de données en hautes dimensions demeure élusive. C'est la question abordée dans cette thèse, à travers l'étude de modèles simplifiés en hautes dimensions, reflétant les caractéristiques clés des tâches d'AM , tout en demeurant analysables mathématiquement. S'inspirant d'idées de physique statistique, cette thèse présente des études asymptotiques de certains aspects clés de l'AM moderne.

La remarquable polyvalence des modèles d'AM réside dans leur capacité à construire des repésentations informatives des données. La première partie de cette thèse explore quels aspects structuraux de ces représentations conditionnent l'apprentissage des méthodes d'AM. En particulier, elle montre comment une théorie bâtie à partir de deux descripteurs statistiques seulement peut décrire quantitativement les performances des méthodes d'AM dans certains cas réels. Pour les méthodes à noyau en particulier, cette théorie permet de relier le taux de décroissance de l'erreur de généralisation à la structure des représentations.

La deuxième partie étudie quelles représentations sont extraites dans les réseaux neuronaux profonds, (a) à l'initialisation et (b) après entraînement, dans le cadre de l'apprentissage bayésien, ou après un unique pas de gradient. En particulier, elle délimite les cas où l'universalité gaussienne prévaut et limite l'expressivité des réseaux, et les cas où les réseaux sont en mesure d'apprendre des représentations non triviales.

Ces cas ne constituent cependant qu'une partie de la zoologie des applications de l'AM modernes. La dernière partie de la thèse couvre certains aspects modernes de la discipline, en particulier les modèles génératifs et les mécanismes d'attention.

Mots-clés – *Apprentissage automatique, Physique statistique, Hautes dimensions, Réseaux neuronaux profonds, Représentations aléatoires, Universalité gaussienne, Méthodes à noyaux, Mécanismes d'attention, Modèles génératifs.*



Mountainline from office 514

ACKNOWLEDGEMENTS

I would like to extend my thanks to Giulio Biroli, Joan Bruna and Nicolas Flammarion for their reading of this thesis, and Laurent Villard for presiding over the defense.

My deepest gratitude goes to Lenka for her steadfast and always kind guidance throughout this journey. Her bottomless wisdom has always been a source of inspiration, and taught me what it means to be a researcher. I would also like to address a big thank you to Florent for his mentorship and kindness.

Big thanks also go to my wise academic big brothers Bruno and Luca Saglietti for their scientific –and sometimes life– tips, and all the collaborators that made this thesis possible – Daniil, Dominik, Eric, Yue, Marc – from whom I learnt a lot. To all the office- and flat- mates – past present and future– who made these four years so lively, Luca Pesce, Freya, Vittorio, Christian, Lucas, Valentina, Anastasiia, Kathryn, thank you : for the memorable acai bowls and everyday adventures, for the colorful afternoon breaks and colorful plots, for the loopy lofi morning sessions, for being that soothing presence in the armchair, for the chill kitchen chats. To all the amazing people I got to know in Smile/Spoc, Sphinx/Idephics – Francesca, Ruben, Federica, Benjamin, Alia, Jonathan, Cédric, Sebastian, Marylou, Antoine Maillard, Stefano, Maria, Antoine Baker; Giovanni, Gabriele, Luca Biggio, Emanuele, Odilon, Frank, Cédric, Indaco, Barbora; Davide, Damien, Pierre, Yatin, Matteo, Adrien, Ludovic, Luca Arnaboldi, Guillaume– and to all the friends I met along the way, at cocktail and pull-up bars– Guanyu, Urte, Tiffany, Halla, Umberto, and so many others– : thank you for making those four years so colorful and memorable. I cannot wait to cross paths and go adventuring –or aperoing– together again.

To Claire, for your mellow and bright companionship, and for always giving me gentle pushes forward when I needed them : thank you so much.

Finally, to my parents, who had unconditional faith in me all these years, as they did my whole life: thank you so much for all you did, and just you watch as I keep going !

TABLE OF CONTENTS

I INTRODUCTION

II DATA STRUCTURE

II A. Features structure

II B. Kernel features

III MULTI-LAYER NETWORKS

III A. Random Features

V Bibliography

LIST OF ABBREVIATIONS

AE	Auto-Encoder
AI	Artificial Intelligence
AL	Active Learning
ANN	Artificial Neural Networks
1dCLT	one-dimensional Central Limit Theorem
CLT	Central Limit Theorem
CNN	Convolutional Neural Networks
DAE	Denoising Auto-Encoder
dcGAN	Deep Convolutional Generative Adversarial Network
DL	Deep Learning
DNN	Deep Neural Networks
dRF	deep Random Features
GAMP	Generalized Approximate Message Passing
GAN	Generative Adversarial Network
GCM	Gaussian Covariate Model
GD	Gradient-Descent
GEP	Gaussian Equivalence Property
GLM	Generalized Linear Model
seq-GLM	Sequence GLM
GP	Gaussian Process
i.i.d	independent and identically distributed
KRR	Kernel Ridge Regression
LSTM	Long Short-Term Memory
ML	Machine Learning
MSE	Mean Squared Error
NTK	Neural Tangent Kernel
ODE	Ordinary Differential Equation
PAC	Probably Approximately Correct
PCA	Principal Component Analysis
RAE	Reconstruction Auto-Encoder
RBF	Radial Basis Function
RF	Random Features
RKHS	Reproducing Kernel Hilbert Space
RS	Replica Symmetry
SE	State Evolution
sRF	spiked Random Features
SGD	Stochastic Gradient-Descent
SP	Saddle Point
SVM	Support Vector Machines
T-S	Teacher-Student

w.h.p	with high probability

LIST OF SYMBOLS

Algebra

a_i	A scalar
$\boldsymbol{a}$	A vector
A	A matrix
$\boldsymbol{a}^\top, A^\top$	Transpose of $\boldsymbol{a}, A$
$\mathbf{A} \otimes \mathbf{B}$	Tensorial product of $\mathbf{A}$ and $\mathbf{B}$
$\mathbf{A} \odot \mathbf{B}$	Hadamard product of $\mathbf{A}$ and $\mathbf{B}$
$\mathbb{I}_d$	Identity matrix of size $d \times d$
$\mathbf{1}_d$	Vector of ones of size d
$\det(\mathbf{A})$	Determinant of $\mathbf{A}$
$\mathrm{Tr}(\mathbf{A})$	Trace of $\mathbf{A}$

Probabilities

$\mathrm{X} \mid \mathrm{Y}$	The random variable X knowing the variable Y
$\mathbb{P}(\mathrm{X})$	The probability of the random variable X, shorthand for $\mathbb{P}(\mathrm{X} = x)$
$\mathbb{E}_{\mathrm{X}}[f(\mathrm{X})]$	Expectation with respect to the random variable X
$\mathrm{Var}(\mathrm{X})[f(\mathrm{X})]$	Variance with respect to the random variable X
$\mathcal{N}(\boldsymbol{\mu}, \Sigma)$	Gaussian distributionwith mean $\boldsymbol{\mu}$ and covariance matrix Σ

Calculus and functions

$\ln$	Natural logarithm
$\equiv$	Defined as
$\asymp$	Asymptotically equivalent to
$f : \mathbb{A} \mapsto \mathbb{B}$	A function f with domain $\mathbb{A}$ and range $\mathbb{B}$
$f_{\boldsymbol{w}}(\boldsymbol{x})$	A function of $\boldsymbol{x}$ parametrized by $\boldsymbol{w}$
$f \circ g$	Composition of the functions f and g
$\mathbb{1}_c$	Indicator function equals to 1 if c is true and 0 otherwise

FOREWORD

Recent years have witnessed a shift in paradigm, driven by the use of Machine Learning (ML) techniques to harness ever more abundant sources of data to automate complex tasks. Artificial Intelligence (AI) is now increasingly ubiquitous, permeating various aspects of our daily lives – communication, entertainment, work, even creative undertakings. The workhorse behind this technical upheaval arguably lies in the use of Artificial Neural Networks (ANN) to extract and learn informative representations *of* data, *from* data. The extracted representations – also called *features*– allow for simple downstream processing, using simple algorithms like linear or logistic regression, ultimately allowing the transfer and application of the learnt task to fresh data.

On the other hand, the tremendous practical success enjoyed by ML empirics stands in sharp contrast with the currently relatively sparse theoretical comprehension thereof. Prominent among the challenges faced by ML theory research is the need to analyze *high-dimensional*, *non-linear*, *non-convex* optimization problems. Given these difficulties, a reasonable research agenda consists of first seeking a sharp theoretical understanding of exactly solvable models– namely simplified mathematical models satisfying the dual requirement of (a) retaining key features of practical ML problems, while (b) remaining amenable to full mathematical treatment. This perspective has led to a rich body of works, investigating the asymptotic properties of ANNs in the particularly relevant *data-intensive*, *high-dimensional* limit. Particularly instrumental in these investigations are ideas inspired by statistical physics (Mézard et al., 2009; Zdeborová et al., 2016; Gabrié, 2020), which, by identifying the relevant statistics that govern the training, offer a particularly concise and insightful description of the learning. This viewpoint constitutes the broader framework of this thesis.

ORGANIZATION OF THE MANUSCRIPT

Part I sets up the necessary contextual and technical framework on high-dimensional ML and statistical physics approaches thereto. Chapter 1 delineates the motivations behind the thesis, and offers a concise and self-contained presentation of some of the technical tools employed. Chapter 2 discusses some perspectives on future research directions in the field, beyond the results discussed in this thesis. The rest of the dissertation then delves into three interconnected aspects of high-dimensional ML.

Part II first explores how the *structure of data* –or representations thereof– impacts the learning of linear methods, Deep Neural Networks (DNN)s or kernel methods. It highlights how for several tasks, including for some real setups, a small set of feature statistics suffice to provide a tight characterization of the test error which quantitatively captures the learning curves, and how scaling laws can be deduced therefrom.

In many cases, the features are shaped by the propagation of the data through the intermediate layers of a multi-layer, fully-connected DNN, which is the object of Part III. The purpose of Part III is to characterize the learning of (a) *deep* networks at initialization and then (b) *trained* networks, either via gradient-based or Bayesian learning methods.

The past decade has witnessed the popularization and success of novel learning paradigms, beyond the intensely studied exemplar of supervised learning with fully-connected, feed-forward DNNs. With these new methods come novel training procedures (for instance, generative models couple a transport problem to the training of a DNN), and novel architectures (e.g. attention layers). Part IV contributes to initiating the extension of statistical physics analyses of ML to these modern aspects, and offers studies of the learning of a flow-based generative model and of a dot-product attention mechanism, at finite sample complexities.

CONTRIBUTIONS

This thesis results from a collection of 10 published or pre-published works.

1. *'Learning curves of generic features maps for realistic datasets with a teacher-student model'*. Loureiro, Gerbelot, Cui, Goldt, Krzakala, Mézard, and Zdeborová (2021)
   Published in *Advances in Neural Information Processing Systems* **34** pp. 18137–18151*;* invited to the special Machine Learning issue of *Journal of Statistical Mechanics: Theory and Experiment,* **11** pp. 114001. Presented in Chap. 3.

   Summary: Teacher-Student (T-S) models provide a framework in which the typical-case performance of high-dimensional supervised learning can be described in closed form. The assumptions of Gaussian independent and identically distributed (i.i.d) input data underlying the canonical teacher-student model may, however, be perceived as too restrictive to capture the behaviour of realistic data sets. We introduce a Gaussian covariate generalisation of the model where the teacher and student can act on different spaces, generated with fixed, but generic feature maps. While still solvable in a closed form, this generalization is able to capture the learning curves for a broad range of realistic data sets, thus redeeming the potential of the teacher-student framework. First, we prove a rigorous formula for the asymptotic training loss and generalisation error for the Gaussian Covariate Model (GCM). Second, we present a number of situations where the learning curve of the model captures the one of a realistic data set learned with kernel regression and classification, with out-of-the-box feature maps such as random projections or scattering transforms, or with pre-learned ones - such as the features learned by training multi-layer neural networks.

Contributions: I contributed to parts of the replica derivation of the main results, and to parts of the numerical experiments.

2. *'Generalization error rates in kernel regression: The crossover from the noiseless to noisy regime'*. Cui, Loureiro, Krzakala, and Zdeborová (2021) Published in *Advances in Neural Information Processing Systems* **34** pp. 10131–10143*;* invited to the special Machine Learning issue of *Journal of Statistical Mechanics: Theory and Experiment,* **11** pp. 114004. Presented in Chap. 4.

   Summary: We consider Kernel Ridge Regression (KRR) under the Gaussian design. Exponents for the decay of the excess generalization error of KRR have been reported in various works under the assumption of power-law decay of eigenvalues of the features covariance. These decays were, however, provided for sizeably different setups, namely in the noiseless case with constant regularization and in the noisy optimally regularized case. Intermediary settings have been left substantially uncharted. We unify and extend this line of work, providing characterization of all regimes and excess error decay rates that can be observed in terms of the interplay of noise and regularization. In particular, we show the existence of a transition in the noisy setting between the noiseless exponents to its noisy values as the sample complexity is increased. Finally, we illustrate how this crossover can also be observed on real data sets.

   Contributions: I conducted the full theoretical analysis and implemented the corresponding numerical experiments.

3. *'Error scaling laws for kernel classification under source and capacity conditions'*. Cui, Loureiro, Krzakala, and Zdeborová (2023b) Published in *Machine Learning: Science and Technology* **4** 3 p. 035033. Presented in Chap. 5.

   Summary: We consider the problem of kernel classification. While worst-case bounds on the decay rate of the prediction error with the number of samples are known for some classifiers, they often fail to accurately describe the learning curves of real data sets. We consider the important class of data sets satisfying the standard source and capacity conditions, comprising a number of real data sets as we show numerically. Under the Gaussian design, we derive the decay rates for the misclassification (prediction) error as a function of the source and capacity coefficients. We do so for two standard kernel classification settings, namely margin-maximizing Support Vector Machines (SVM) and ridge classification, and contrast the two methods. We find that our rates tightly describe the learning curves for this class of data sets, and are also observed on real data. Our results can also be seen as an explicit prediction of the exponents of a scaling law for kernel classification that is accurate on some real datasets.

Contributions: I conducted the full theoretical analysis and implemented the corresponding numerical experiments.

4. *'Deterministic equivalent and error universality of deep random features learning'*. Schröder, Cui, Dmitriev, and Loureiro (2023)
Published in *International Conference on Machine Learning* **40** pp. 30285–30320. Presented in Chap. 6.

   Summary: We the problem of learning a random Gaussian ANN function using a Fully-connected Neural Network (FNN) with frozen intermediate layers and trainable readout layer – namely deep Random Features (dRF) models. This problem can be seen as a natural generalization of the widely studied random features model to deeper architectures. First, we prove Gaussian universality of the test error in a ridge regression setting where the learner and target networks share the same intermediate layers, and provide a sharp asymptotic formula for it. Establishing this result requires proving a deterministic equivalent for traces of the dRF sample covariance matrices which can be of independent interest. Second, we conjecture the asymptotic Gaussian universality of the test error in the more general setting of arbitrary convex losses and generic learner/target architectures. We provide extensive numerical evidence for this conjecture, which requires the derivation of closed-form expressions for the layer-wise post-activation population covariances. In light of our results, we investigate the interplay between architecture design and implicit regularization.

   Contributions: I contributed to the conception of the project, which was partly motivated by my derivation of the linearization formula for population covariances in deep random neural networks. I also proposed the tight generalization error characterization in the generic case, designed and conducted the investigation of architectural bias, and performed the numerical experiments.

5. *'Asymptotics of Learning with Deep Structured (Random) Features'*. Schröder, Dmitriev, Cui, and Loureiro (2024)
Published in *International Conference on Machine Learning* **41**. Presented in Chap. 7.

   Summary: For a large class of feature maps we provide a tight asymptotic characterisation of the test error associated with learning the readout layer, in the high-dimensional limit where the input dimension, hidden layer widths, and number of training samples are proportionally large. This characterization is formulated in terms of the population covariance of the features. Our work is partially motivated by the problem of learning with Gaussian rainbow ANNs, namely deep non-linear fully-connected networks with random but structured weights, whose row-wise covariances are further allowed to depend on the weights of previous layers. For such networks we also derive a closed-form formula for the feature covariance in terms of the weight matrices. We further find that in some cases our results can capture feature maps learned by deep, finite-width ANNs trained under gradient descent.

Contributions: I contributed to the initial heuristic derivation of the linearization formulae and sharp characterization of the test error in the uncorrelated case, and contributed to the numerical experiments.

6. *'Bayes-optimal learning of deep random networks of extensive-width'*. Cui, Krzakala, and Zdeborová (2023a)
Published in *International Conference on Machine Learning* **40** 6468–6521, (Oral). Presented in Chap. 8.

   Summary: We consider the problem of learning a target function corresponding to a deep, extensive-width, non-linear DNN with random Gaussian weights. We consider the asymptotic limit where the number of samples, the input dimension and the network width are proportionally large and propose a closed-form expression for the Bayes-optimal test error, for regression and classification tasks. We further compute closed-form expressions for the test errors of ridge regression, kernel and random features regression. We find, in particular, that optimally regularized ridge regression, as well as kernel regression, achieve Bayes-optimal performances, while the logistic loss yields a near-optimal test error for classification. We further show numerically that when the number of samples grows faster than the dimension, ridge and kernel methods become suboptimal, while ANNs achieve test error close to zero from quadratically many samples.

   Contributions: I conceived the project, and conducted the full theoretical analysis and corresponding numerical experiments.

7. *'Asymptotics of feature learning in two-layer networks after one gradient-step'*. Cui, Pesce, Dandi, Krzakala, Lu, Zdeborová, and Loureiro (2024d)
Published in *International Conference on Machine Learning* **41**. Presented in Chap. 9.

   Summary: We investigate the problem of how two-layer neural networks learn features from data, and improve over the kernel regime, after being trained with a single gradient descent step. Leveraging a connection from (Ba et al., 2022) with a non-linear spiked matrix model and recent progress on Gaussian universality (Dandi et al., 2023), we provide an exact asymptotic description of the generalization error in the high-dimensional limit where the number of samples n, the width p

and the input dimension d grow at a proportional rate. We characterize exactly how adapting to the data is crucial for the network to efficiently learn non-linear functions in the direction of the gradient – where at initialization it can only express linear functions in this regime. To our knowledge, our results provides the first tight description of the impact of feature learning in the generalization of two-layer neural networks in the large learning rate regime $\eta = \Theta_d(d)$, beyond perturbative finite width corrections of the conjugate and neural tangent kernels.

Contributions: I conducted the full theoretical analysis and a number of the corresponding numerical experiments.

8. *'High-dimensional asymptotics of denoising autoencoders'*. Cui and Zdeborová (2024c)
   Published in *Advances in Neural Information Processing Systems* **36** (Spotlight). Presented in Chap. 10.

   Summary: We address the problem of denoising data from a Gaussian mixture using a two-layer non-linear Auto-Encoder (AE) with tied weights and a skip connection. We consider the high-dimensional limit where the number of training samples and the input dimension jointly tend to infinity while the number of hidden units remains bounded. We provide closed-form expressions for the denoising Mean Squared Error (MSE). Building on this result, we quantitatively characterize the advantage of the considered architecture over the AE without the skip connection that relates closely to principal component analysis. We further show that our results accurately capture the learning curves on a range of real data sets.

   Contributions: I contributed to the design of the theoretical model, and conducted the full theoretical analysis and corresponding numerical experiments.

9. *'Analysis of learning a flow-based generative model from limited sample complexity'*. Cui, Krzakala, Vanden-Eijnden, and Zdeborová (2024b)
   Published in *International Conference on Learning Representations* **12**. Presented in Chap. 11.

   Summary: We study the problem of training a flow-based generative model, parametrized by a two-layer AE, to sample from a high-dimensional Gaussian mixture. We provide a sharp end-to-end analysis of the problem. First, we provide a tight closed-form characterization of the learnt velocity field, when parametrized by a shallow Denoising Auto-Encoder (DAE) trained on a finite number n of samples from the target distribution. Building on this analysis, we provide a sharp description of the corresponding generative flow, which pushes the base Gaussian density forward to an approximation of the target density. In particular, we provide closed-form formulae for the distance between the mean of the generated mixture and the mean of the target mixture, which we show decays as $\Theta_n({}^1/_n)$. Finally, this rate is shown to be in fact Bayes-optimal.

Contributions: I contributed to the design of the theoretical model, and conducted the full theoretical analysis and corresponding numerical experiments.

10. *'A phase transition between positional and semantic learning in a solvable model of dot-product attention'*. Cui, Behrens, Krzakala, and Zdeborová (2024a)
Published in *arXiv preprint arXiv:2402.03902*. Presented in Chap. 12.

    Summary: We investigate how a dot-product attention layer learns a positional attention matrix (with tokens attending to each other based on their respective positions) and a semantic attention matrix (with tokens attending to each other based on their meaning). For an algorithmic task, we experimentally show how the same simple architecture can learn to implement a solution using either the positional or semantic mechanism. On the theoretical side, we study the learning of a non-linear self-attention layer with trainable tied and low-rank query and key matrices. In the asymptotic limit of high-dimensional data and a comparably large number of training samples, we provide a closed-form characterization of the global minimum of the non-convex empirical loss landscape. We show that this minimum corresponds to either a positional or a semantic mechanism and evidence an emergent phase transition from the former to the latter with increasing sample complexity. Finally, we compare the dot-product attention layer to linear positional baseline, and show that it outperforms the latter using the semantic mechanism provided it has access to sufficient data.

    Contributions: I conceived the project, and conducted the full theoretical analysis of the exactly solvable model.

The two following works, which explore the sub-branch of ML theory known as Active Learning (AL), have also been completed in the framework of the PhD, but will not be the object of further discussion in the present thesis.

11. *'Large deviations in the perceptron model and consequences for active learning'*. Cui, Saglietti, and Zdeborová (2020)
Published in *Mathematical and Scientific Machine Learning* **1**, pp. 390–430 and *Machine Learning: Science and Technology* **2** 4 pp. 045001.

    Summary: AL is a branch of machine learning that deals with problems where unlabeled data is abundant yet obtaining labels is expensive. The learning algorithm has the possibility of querying a limited number of samples to obtain the corresponding labels, subsequently used for supervised learning. We consider the task of choosing the subset of samples to be labeled from a fixed finite pool of samples. We assume the pool of samples to be a random matrix and the ground truth labels to be generated by a single-layer teacher random neural network. We employ replica methods to analyze the large deviations for the accuracy achieved after supervised learning on a subset of the original pool. These large deviations then provide optimal achievable performance boundaries for any AL algorithm. We show that the optimal learning performance can be efficiently approached by simple message-passing AL algorithms. We also provide a comparison with the performance of some other popular AL strategies.

Contributions: I conducted the full theoretical analysis and a part of the numerical experiments.

12. *'Large deviations of semisupervised learning in the stochastic block model'*. Cui, Saglietti, and Zdeborová (2022)
Published in *Physical Review E* **105** 3 p. 034108.

Summary: In semisupervised community detection, the membership of a set of revealed nodes is known in addition to the graph structure and can be leveraged to achieve better inference accuracies. While previous works investigated the case where the revealed nodes are selected at random, this paper focuses on correlated subsets leading to atypically high accuracies. In the framework of the dense stochastic block model, we employ statistical physics methods to derive a large deviation analysis of the number of these rare subsets, as characterized by their free energy. We find theoretical evidence of a non-monotonic relationship between reconstruction accuracy and the free energy associated to the posterior measure of the inference problem. We further discuss possible implications for AL applications in community detection.

Contributions: I conducted the full theoretical analysis and a part of the numerical experiments.

LIST OF PUBLICATIONS

[1] Hugo Cui, Freya Behrens, Florent Krzakala, and Lenka Zdeborová. ‘A phase transition between positional and semantic learning in a solvable model of dot-product attention.’ In: *arXiv preprint arXiv:2402.03902* (2024) (cit. on p. xxiv).

[2] Hugo Cui, Florent Krzakala, Eric Vanden-Eijnden, and Lenka Zdeborová. ‘Analysis of learning a flow-based generative model from limited sample complexity.’ In: *International Conference on Learning Representations* 12 (2024) (cit. on p. xxiii).

[3] Hugo Cui, Florent Krzakala, and Lenka Zdeborová. ‘Bayes-optimal learning of deep random networks of extensive-width.’ In: *International Conference on Machine Learning* 40 (2023), 6468–6521, (Oral) (cit. on p. xxii).

[4] Hugo Cui, Bruno Loureiro, Florent Krzakala, and Lenka Zdeborová. ‘Error scaling laws for kernel classification under source and capacity conditions.’ In: *Machine Learning: Science and Technology* 4.3 (2023), p. 035033 (cit. on p. xx).

[5] Hugo Cui, Bruno Loureiro, Florent Krzakala, and Lenka Zdeborová. ‘Generalization error rates in kernel regression: The crossover from the noiseless to noisy regime.’ In: *Advances in Neural Information Processing Systems* **34** pp. 10131–10143; invited to the special Machine Learning issue of *Journal of Statistical Mechanics: Theory and Experiment,* **11** pp. 114004 (2021) (cit. on p. xx).

[6] Hugo Cui, Luca Saglietti, and Lenka Zdeborová. ‘Large deviations in the perceptron model and consequences for active learning.’ In: *Mathematical and Scientific Machine Learning* **1**, pp. 390–430 and *Machine Learning: Science and Technology* **2** 4 pp. 045001 (2020) (cit. on p. xxiv).

[7] Hugo Cui, Luca Saglietti, and Lenka Zdeborová. ‘Large deviations of semisupervised learning in the stochastic block model.’ In: *Physical Review E* 105.3 (2022), p. 034108 (cit. on p. xxv).

[8] Hugo Cui and Lenka Zdeborová. 'High-dimensional asymptotics of denoising autoencoders.' In: *Advances in Neural Information Processing Systems* 36 (2024), (Spotlight) (cit. on p. xxiii).

[9] Hugo Cui et al. 'Asymptotics of feature learning in two-layer networks after one gradient-step.' In: *International Conference on Machine Learning* 41 (2024) (cit. on p. xxii).

[10] Bruno Loureiro et al. 'Learning curves of generic features maps for realistic datasets with a teacher-student model.' In: *Advances in Neural Information Processing Systems* **34** pp. 18137–18151*;* invited to the special Machine Learning issue of *Journal of Statistical Mechanics: Theory and Experiment,* **11** pp. 114001 (2021) (cit. on p. xix).

[11] Dominik Schröder, Hugo Cui, Daniil Dmitriev, and Bruno Loureiro. 'Deterministic equivalent and error universality of deep random features learning.' In: *International Conference on Machine Learning* 40 (2023), pp. 30285–30320 (cit. on p. xxi).

[12] Dominik Schröder, Daniil Dmitriev, Hugo Cui, and Bruno Loureiro. 'Asymptotics of Learning with Deep Structured (Random) Features.' In: *International Conference on Machine Learning* 41 (2024) (cit. on p. xxi).

Part I

INTRODUCTION

1

INTRODUCTION

This first chapter offers a concise overview of the basic concepts in *machine learning theory*, and a companion rendition of common statistical physics analyses of ML models. A more exhaustive introduction to ML can be found for instance in (Goodfellow et al., 2016; Mohri et al., 2018).

1.1 BASIC CONCEPTS IN ML

1.1.1 WHY ML THEORY?

ML is, above all, a collection of techniques and tools in the statistical processing of large quantities of data. It thus constitutes a branch of engineering that has, as such, historically progressed empirically, by trial and error. With the unremitting pace of practical advances and the increasing ubiquity of ML tools in all fields – including sensitive ones such as health – such an approach is no longer sustainable, and calls for a more principled basis for the development of ML. A thorough mathematical theory of ML is, to that end, a requisite – both to ensure a safe use of current tools, and to guide and inspire the development of novel methods.

*Why theory in ML? "**Comfort**:We knew it worked, but it's nice to have a proof; **Insight**: Aha! So that's why it works!; **Innovation**: At last, a mathematically proven idea that applies to data; **Suggestion**: Something like this might work with data." (Breiman, 1995)*

Section 1.1.2 provides a concise introduction to ML as a *mathematical object* of study, highlighting in particular how it can be formulated as a random optimization problem in high dimensions. Section 1.2 then illustrates, on a simple case study, how such problems can be analyzed using ideas borrowed from statistical physics.

1.1.2 THE ML PIPELINE

ML is a discipline concerned with *automating complex tasks* – by which it is understood tasks that admit no direct simple mathematical formulation –, through the statistical processing of *data*. In its barest form, the ML pipeline consists of approximating a mapping $f_\star : \mathscr{X} \to \mathscr{Y}$ from the *input data* $\boldsymbol{x} \in \mathscr{X}$ to its *target value* (label) $f_\star(\boldsymbol{x}) \in \mathscr{Y}$. In practice, for example, the input/label pair $\boldsymbol{x}, f_\star(\boldsymbol{x})$ can be a sentence and its translation, an image and its resolution-enhanced version, an image and its caption, or a pair of physical attributes. Because $f_\star$ admits in most settings no known closed-form mathematical or algorithmic formulation which can be readily coded and implemented by a computer, the ML pipeline rather aims at *approximately* implementing

In the particular case of $f_\star(\boldsymbol{x}) = \boldsymbol{x}$, the ML tasks are qualified as self-supervised*. In other cases, we speak of* supervised *tasks.*

the target function $f_\star$. To that end, ML techniques seek to leverage an informative representation (*feature map*) $\varphi : \mathscr{X} \to \mathscr{Z}$, which transforms and processes the original data point $\boldsymbol{x}$ into a more informative set of *features* $\varphi(\boldsymbol{x})$, and subsequently approximating the target $f_\star$ typically as a linear combination$f_w = \boldsymbol{w}^\top \varphi(\boldsymbol{x})$, with the *weights vector* $\boldsymbol{w}$ collecting the corresponding coefficients. On an intuitive level, the features $\varphi(\boldsymbol{x})$ should thus regroup the important characteristics and attributes of the data point $\boldsymbol{x}$ in the context of the task.

The choice of the feature transformation φ is, for some applications, natural. Consider for instance the case of images, for which wavelet transforms offer a concise and informative representation, capturing the information encoded at multiple scales (Mallat, 2016). In practice, on the other hand, there could exist more efficient feature extraction maps φ; further, in some other cases, as for natural language data, there exists no natural off-the-shelf transform candidate. It thus proves convenient to also seek an efficient representation φ inside a *parametric family* $\{\varphi_\theta\}_\theta$. Allowing the feature extractor φ_θ to be *trainable* (*learnable*) is the driving paradigm behind modern successes of Deep Learning (DL) techniques. The learning problem thus consists in finding the best function $f_{\hat{w},\hat{\theta}}$ among the parametric family $\{f_{w,\theta} = \boldsymbol{w}^\top \varphi_\theta(\cdot)\}_{w,\theta}$. In the context of DL, the vector $\boldsymbol{w}$ is referred to as the *readout weights*, and we shall sometimes refer to the feature map parameters $\boldsymbol{\theta}$ as the *internal weights* of the model. The selected values of these parameters $\hat{\boldsymbol{w}}, \hat{\boldsymbol{\theta}}$ are usually called the *trained* –or *learnt*– weights.

The parametric family of functions $\{f_{w,\theta}\}_{w,\theta}$ *is called the* hypothesis class *in learning theory.*

DL is the branch of ML which leverages DNNs as an hypothesis class.

To find satisfactory weights, ML techniques first seek to find a good approximation of the target $f_\star$ on a set of points $\mathscr{D} = \{\boldsymbol{x}^\mu, f_\star(x^\mu)\}_{\mu=1}^n$ for which the target value is known. The set $\mathscr{D}$ is called the *training set*, and is used as an empirical proxy for the true data distribution P_x. Mathematically, this procedure boils down to minimizing a notion of distance between the parametric function $f_{w,\theta}$ and the target $f_\star$ at all points of $\mathscr{D}$, which is referred to as the empirical *risk* (or *loss*)

$$\mathscr{R}(\boldsymbol{w}, \boldsymbol{\theta}) = \sum_{\mu=1}^{n} \ell\left(f_\star(\boldsymbol{x}^\mu), f_{\boldsymbol{w},\theta}(\boldsymbol{x}^\mu)\right) + g\left(\boldsymbol{w}, \boldsymbol{\theta}\right). \tag{1}$$

$\ell(\cdot)$ is a function that generically increases when its two arguments are dissimilar; $g(\cdot)$ is a –typically convex– *regularizer*, which penalizes too large weight parameters. Satisfactory values for the model weights can then be selected as minimizers of the risk (1)

$$\hat{\boldsymbol{w}}, \hat{\boldsymbol{\theta}} = \underset{\boldsymbol{w},\theta}{\operatorname{argmin}}\, \mathscr{R}(\boldsymbol{w}, \boldsymbol{\theta}). \tag{2}$$

Popular choices for the loss function include the square loss $\ell(y,z) = {}^1\!/\!{}_2(y-z)^2$, *the hinge loss* $\ell(y,z) = (0, 1-yz)_+$. *Typically regularizers include the* ℓ_2 *regularization* $g(z) = {}^\lambda\!/\!{}_2 \|z\|^2$, *where* λ *is then referred to as the* regularization strength.

In practice, this Empirical Risk Minimization (ERM) is numerically carried out using first-order Gradient-Descent (GD) based methods. The quality of the minimization (2) is measured by the *training error*

$$\varepsilon_t = \frac{1}{n}\sum_{\mu=1}^{n} \ell_{\text{tr.}}\left(f_\star(\boldsymbol{x}^\mu), f_{\hat{w},\hat{\theta}}(\boldsymbol{x}^\mu)\right), \tag{3}$$

for some metric $\ell_{\text{tr.}}(\cdot)$, which quantifies the distance between the target $f_\star$ and the fitted parametric function $f_{\hat{\boldsymbol{w}},\hat{\theta}}$ on the points of the train set $\mathscr{D}$. Typically, in *regression* settings where the target $f_\star$ takes continuous values – e.g. $\mathscr{Y} = \mathbb{R}$ –, a popular choice is the squared error $\ell_{\text{tr.}}(y,z) = {}^1\!/_2(y-z)^2$. In classification settings – e.g. $\mathscr{Y} = \{-1,+1\}$–, a natural choice is the misclassification error $\ell_{\text{tr.}}(y,z) = 1 - \delta_{y,\text{sign}(z)}$, which measures the fraction of misclassified training samples. At the end of the training, the statistician thus has access to an approximate implementation $f_{\hat{w},\hat{\theta}}$ of the target map $f_\star$, which can be in turn readily applied to fresh data. The discrepancy between the true target and its learnt approximation is quantified by the *test error*

$$\varepsilon_g = \mathbb{E}_{\boldsymbol{x}\sim P_x}[\ell_{\text{ts.}}\left(f_\star(\boldsymbol{x}), f_{\hat{w},\hat{\theta}}(\boldsymbol{x})\right)], \tag{4}$$

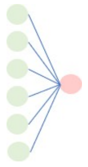
for some choice of metric $\ell_{\text{ts.}}$. The test error (4) constitutes a central metric in ML to evaluate the generalization ability of the learning models.

1.1.3 SOME ML MODELS

Choosing – and parametrizing– a suitable feature transformation φ is thus the centerpiece of the ML pipeline. The development of the field has thus unsurprisingly gone hand-in-hand with a swift expansion of the zoology of possible feature map options. In this subsection, we offer a selected digest of some choices relevant to the present thesis, starting first from off-the-shelf fixed transforms, and secondly tunable ANN feature maps.

1.1.3.A OFF-THE-SHELF FEATURE MAPS

No feature map– The simplest case is when the input $\boldsymbol{x}$ is used as is, with no further transformation, namely $\varphi(\boldsymbol{x}) = \boldsymbol{x}$. The corresponding models $f_w(\boldsymbol{x}) = \boldsymbol{w}^\top\boldsymbol{x}$ are commonly regrouped under the umbrella of *linear models* in ML, and include canonical algorithms such as



Linear methods can also be viewed as single-layer ANNs.

- Ridge regression $\ell(y,z) = {}^1\!/_2(y-z)^2, g(\boldsymbol{w}) = {}^\lambda\!/_2\|\boldsymbol{w}\|^2$,
- Logistic regression $\ell(y,z) = \ln(1+e^{-yz})$,
- Hinge regression $\ell(y,z) = (1-yz)_+$.

Because of their simplicity and their convexity, linear methods are very easy to train. A key limitation however lies in that they can only implement linear functions of the data, and thus suffer from poor *expressivity*. In other words, they do not provide a versatile enough framework to approximate complex targets $f_\star$.

Kernel feature maps– Kernel methods constitute another centerpiece of traditional ML. Given a kernel $K : \mathscr{X} \times \mathscr{X} \to \mathbb{R}$, the Moore–Aronszajn theorem ensures that the bilinear operation it defines can be rewritten in scalar product form

By the Moore–Aronszajn theorem, any symmetric, positive semi-definite kernel K can be simply written as a scalar product in some Hilbert space $\mathscr{H}$, called the RKHS of K.

$$K(\boldsymbol{x}_1,\boldsymbol{x}_2) = \langle \varphi(\boldsymbol{x}_1), \varphi(\boldsymbol{x}_2) \rangle_{\mathscr{H}} \tag{5}$$

where the *kernel feature map* $\varphi : \mathscr{X} \to \mathscr{H}$ maps the data in non-linear fashion to a typically large (or infinite) dimensional Reproducing Kernel Hilbert Space (RKHS) space $\mathscr{H}$. Despite their relative simplicity, kernels thus provide a versatile framework to learn using non-linear features, while remaining in the realm of convex optimization. Furthermore, representer theorems such as (Kimeldorf et al., 1971) imply that kernel methods can be efficiently trained, even as they tap into the expressivity of an infinite feature space.

Random Features – The closely related class of Random Features (RF) models were first introduced in (Rahimi et al., 2007b) as an efficient way to approximate kernel methods. They can alternatively be seen as FNN at initialization. Mathematically, a depth L RF feature map is defined as the composition of maps

$$\varphi = \psi_L \circ \cdots \circ \psi_1 \tag{6}$$

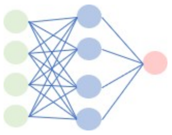
where the $\ell-$*th layer* ψ_ℓ is the map

$$\psi_\ell(\boldsymbol{x}_{\ell-1}) = \sigma_\ell(W_\ell \boldsymbol{x}_{\ell-1}). \tag{7}$$

In (7), $\sigma_\ell(\cdot)$ is a non-linear function, and W_ℓ, referred to as the $\ell - th$ layer weights, is a *fixed* –not trained– random matrix. The first dimension of W_ℓ defines the *width* of the $\ell-$th hidden layer. Aside from their connection to kernel methods, RF models provide stylized proxies for FNNs, and thus afford an ideal theoretical sandbox to analytically probe some properties of the latter.

1.1.3.B TUNABLE FEATURE MAPS

While ready-to-use feature maps provide computationally efficient pathways to enhance the model expressivity, the features they extract are not tailored to the data and task, and may prove sub-optimal. On the other hand, trainable maps such as DNN feature maps are themselves parametric and thus allow for additional tunability and versatility.



Conventional graphical representation of a $2-$layer FNN.

Fully-connected feed-forward networks– Historically, the first instance of ANN is provided by the work of (Rosenblatt, 1958) on the *perceptron* –

namely, a single-layer neural network in modern nomenclature. DNNs (LeCun et al., 2015) build expressive feature maps by essentially stacking perceptron neurons into layers, and subsequently stacking several layers to construct deep *architectures*. A depth L feed-forward FNN is thus defined as the composition of maps

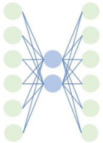
$$\varphi_{W_1,\ldots,W_L} = \psi^{(L)}_{W_L} \circ \cdots \circ \psi^{(1)}_{W_1} \tag{8}$$

where the $\ell-$th layer $\psi^{(\ell)}_{W_\ell}$ is defined as

$$\psi^{(\ell)}_{W_\ell}(\boldsymbol{x}_{\ell-1}) = \sigma_\ell(W_\ell \boldsymbol{x}_{\ell-1}). \tag{9}$$

Here, the weights $W_1, \ldots, W_L$ are importantly *learnable* parameters. While FNNs are typically well suited to process vector data $\mathscr{X} = \mathbb{R}^d$, Convolutional Neural Networks (CNN) architectures (Fukushima, 1980) are specifically designed to extract features from image data, leveraging convolutional and downsampling layers to take into account invariances inherent to this type of data.

Autoencoders – AEs represent specific instances of DNNs designed for self-supervised tasks, typically when $\mathscr{Y} = \mathscr{X}$. In its simplest two-layer instance, an AE $f_{w,\theta}(\boldsymbol{x}) = \boldsymbol{w}\varphi_\theta(\boldsymbol{x})$ is the succession of an *encoder* feature map $\varphi_\theta : \mathbb{R}^d \to \mathbb{R}^b$ and a *decoder* layer with weights $\boldsymbol{w} \in \mathbb{R}^{d\times b}$. Typically, an AE displays a *bottleneck* structure, with the hidden layer width b being small. This enforces that the encoder learns a concise low-dimensional representation of the data, which can subsequently be mapped back into the original space by the decoder. Because AEs learn compact, thus a priori robust, latent representations, they are popular choices in denoising applications (Vincent et al., 2010). AEs, and related modern denoiser architectures such as U-nets (Ronneberger et al., 2015a), have further enjoyed a recent regain in interest as they find themselves at the heart of diffusion-based generative models (Sohl-Dickstein et al., 2015; Ho et al., 2020a).



Graphical representation of a 2−layer AE with two hidden units.

Transformers – Transformers (Vaswani et al., 2017) offer an efficient way of extracting features from sequential data – such as language. Given an input $\boldsymbol{x} \in \mathbb{R}^{L\times d}$ of length L, a transformer feature map usually corresponds to a composition of a FNN and an *attention layer*, defined as the map

$$\varphi_{W_Q,W_K,W_V} = \mathrm{softmax}\left(\boldsymbol{x}W_Q W_K^\top \boldsymbol{x}^\top\right)\boldsymbol{x}W_v, \tag{10}$$

parametrized by the three trainable matrices W_Q, W_K, W_V. As for FNNs, the value matrix W_V acts at the level of the token representations to build more informative features therefrom. Simultaneously, the $L \times L$ attention score matrix contextualizes each token, by mixing the sequence in an input-aware fashion. Crucially, transformer architectures are scalable, requiring less training time than recurrent network architectures such as Long Short-Term Memory (LSTM) (Bahdanau et al., 2014).

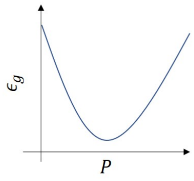
W_Q, W_K, W_V are respectively known as the query, key *and* value *matrices.*

1.1.4 TWO CONUNDRUMS IN HIGH DIMENSIONS

Despite the evident successes of day-to-day DL empirics, the field still lacks a solid theoretical foundation. In fact, fundamental questions regarding the unreasonable effectiveness of statistical models such as DNNs in approximating complex, typically high-dimensional functions have remained unanswered for decades. The interrogations raised by (Breiman, 1995) in the 90s – *Why don't heavily parametrized neural networks overfit the data? [...] Why doesn't backpropagation head for poor local minima?* – still remain of striking relevance in modern ML theory.

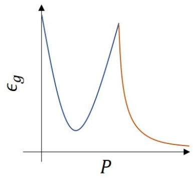
1.1.4.A PUZZLE 1 – THE CURSE OF DIMENSIONALITY

Approximating a function in d dimension generically requires an exponential number $\log n \asymp d$ of samples. This intuitive limitation of learning in large-dimensional spaces is often referred to as *the curse of dimensionality*. More even: there exist data sets for which training even a three-layer architecture is NP-complete (Blum et al., 1988). Learning high-dimensional functions is thus generically a computationally hard task. Yet, this intuition is blatantly belied, in daily DL practice, by the observed effectiveness of DNNs in performing such tasks. From whence the discrepency? Among the many possible explanations is the fact that real data sets are *structured* (Hein et al., 2005; Mallat, 2016), and there thus exists a lower-dimensional embedding that retains most of the information contained in the data, whilst emancipating the statistical model from the curse of high-dimensional learning. As we further discuss in section 1.2 and Chapter 2, considering *typical*, *realistic* data structures – as opposed to worst-case – is therefore a first requisite towards reaching an answer to this first high-dimensional puzzle.



The U-shaped curve associated to the bias-variance tradeoff.

1.1.4.B PUZZLE 2 –BENIGN OVERPARAMETRIZATION



Double descent.

A statistical model which does not have a sufficient number of parameters *underfits* the data. Conversely, a model with too many parameters will *overfit* the dataset and consequently also yield poor generalization. The *bias-variance tradeoff* thus prescribes to balance the number of parameters P of the model, in order to locate the minimum of the generically observed U-shaped test error curve in conventional statistics. DNNs, however, defy this piece of classical statistical wisdom. In fact, they often generalize *better* beyond the interpolation threshold, in heavily overparametrized regimes (Geman et al., 1992). Rather than a U-shaped curve, the learning curves of complex architectures typically thus take the form of a *double-descent curve* (Belkin et al., 2019; Geiger et al., 2020). That heavily parametrized DNNs do not overfit, while they are expressive enough to fit even random labels (Zhang et al., 2021), constitutes another fundamental – and largely open– puzzle in DL theory.

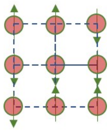
The interpolation regime designates the regime in which the DNN achieves zero training error.

These puzzles make the compelling case that a thorough investigation of the learning of DNNs in overparametrized, high-dimensional regimes warrants *tight* and *typical-case* explorations – questions that fall in the orbit of *statistical physics*.

1.2 STATISTICAL PHYSICS OF ML

1.2.1 STATISTICAL PHYSICS IN THE ML RESEARCHSCAPE

The birth of statistical physics can be traced back to the seminal works of Maxwell, Gibbs and Boltzmann in the 19th century, which were concerned with the study of emergent collective behaviours arising from the microscopic interaction of large assemblies of particles. More formally, the statistical physicist aims at a concise description of high-dimensional probability measures arising from large systems in interaction, in terms of a compact set of *macroscopic* observables. Connecting to the previous section, learning in ML systems is also a result of the complex interactions between a large number of variables –namely the parameters of the learning model– as they are jointly optimized to minimize the ERM loss. Perhaps then unsurprisingly, ML, as a field concerned with random optimization problems in high dimensions, naturally falls in the orbit of statistical physics techniques.



Historically, some important objects of study in statistical physics include gases, and interacting spin systems as stylized models of magnetism or glasses.

Statistical physics ideas were first successfully applied to sharply characterize the learning curve of perceptron models by Gardner (Gardner et al., 1988; Gardner et al., 1989), giving the initial impulse to a long and rich line of works (Seung et al., 1992; Györgyi et al., 1990; Schwarze, 1993; Sompolinsky et al., 1990) (see (Mézard et al., 2009; Zdeborová et al., 2016; Gabrié, 2020) for reviews). These early works made it patent that the worst case, Probably Approximately Correct (PAC) viewpoint on ML (Valiant, 1984), then predominant in computer science, was too coarse to capture some aspects of the learning in high-dimensional models. While these bounds describe a continuous and smooth improvement of the generalization with the number of training samples, abrupt *phase transitions* to perfect generalization were found and highlighted. Such observations made the compelling case that distribution-free, worst-case analyses could sometimes prove insufficient, and that a theory of ML also required tight, typical-case analyses.

As a part of theoretical physics, statistical physics analyses of ML are almost always model-driven, taking as a triple starting point:

(a) A specific data distribution;

(b) A specific target function generating the labels in supervised settings;

(c) A specific learning model and training procedure;

and aiming at tightly computing – down to the constant – the typical generalization learning curves of the model. Such *exactly solvable models* should be simple enough to be amenable to mathematical scrutiny, while capturing key aspects of practical ML tasks. This perspective and approach now well supersede the boundaries of statistical physics, sparking works in e.g. mathematical optimization (Thrampoulidis et al., 2014; Thrampoulidis et al., 2018; Oymak et al., 2013) or random matrix theory (Louart et al., 2018b; Liao et al., 2020; Xiao et al., 2022). Together with statistical physics-inspired analyses (Seung et al., 1992; Sompolinsky et al., 1990; Barbier et al., 2019b; Aubin et al., 2018b; Aubin et al., 2020a), these lines of research are sometimes gathered under the umbrella of *exact asymptotics* in ML theory.

In the rest of this introduction, we provide a self-contained technical illustration of some of these ideas, borrowed from statistical physics, which mainly constitute the technical backbone of the present manuscript.

1.2.2 A CASE STUDY: LEARNING A SEQUENCE GENERALIZED LINEAR MODEL

To illustrate some of the ideas and techniques employed in most of the works gathered in this thesis, we present in the following a concise and self-contained asymptotic analysis of a ML problem in a T-S setting – namely the learning of a variant of a Generalized Linear Model (GLM) acting on sequential data, which we subsequently refer to as a Sequence GLM (seq-GLM). The analysis of ERM for linear methods was first detailed in (Aubin et al., 2020a), for non-sequential isotropic inputs and common loss functions. We propose in this section a rendition in a more generic case, closer to the setting of (Cui et al., 2024a).

The teacher in statistical physics is also referred to as the oracle *in computer science. The denomination T-S most often refers to settings where the target function is linear, or of ANN form.*

Consider the ERM loss over $\boldsymbol{w} \in \mathbb{R}^{d\times r}$

$$\hat{w} = \underset{\boldsymbol{w}}{\mathrm{argmin}} \left[\sum_{\mu=1}^{n} \ell\left(\frac{\boldsymbol{x}^\mu \boldsymbol{w}_\star}{\sqrt{d}}, \frac{\boldsymbol{x}^\mu \boldsymbol{w}}{\sqrt{d}}, \frac{\boldsymbol{w}^\top \boldsymbol{w}}{d}, c^\mu \right) + \frac{\lambda}{2}\|\boldsymbol{w}\|^2 \right], \tag{11}$$

where $\boldsymbol{w}_\star \in \mathbb{R}^{d\times t}$ parametrizes the target function and $\ell : \mathbb{R}^{L\times t} \times \mathbb{R}^{L\times r} \times \mathbb{R}^{r\times r} \times \mathbb{R} \to \mathbb{R}_+$ is a –not necessarily convex– function. We assume the training data $\boldsymbol{x}^\mu \in \mathbb{R}^{L\times d}$ are i.i.d as stacks of L independent rows (tokens), drawn from a Gaussian mixture

$$\boldsymbol{x}_\ell, c_\ell \sim \sum_{k=1}^{K_\ell} \rho_{\ell,k} \delta_{c_\ell,k} \mathcal{N}(\boldsymbol{\mu}_{\ell,k}, \Sigma_{\ell,k}), \tag{12}$$

with the random variable c_ℓ indicating the cluster assignment. This setting can serve as a simple model for several data distributions of interest:

- For $L = 1$, it simply models a single vector input $\boldsymbol{x} = \boldsymbol{x}_1$, which is relevant for the study of e.g. simple linear models (see Part II) or FNNs (Part III). In this case, the seq-GLM coincides with a usual GLM.

- For $L > 1$, it models length-L sequences with uncorrelated tokens embedded in dimension d, which can serve as a model for inputs for an attention layer (see Chapter 12);
- for $L = 2$ and when the second row corresponds to white noise $x_2 \sim \mathcal{N}(0, \mathbb{I}_d)$, it models an input with a signal $\boldsymbol{x}_1$ component and a corrupting noise $\boldsymbol{x}_2$, relevant for the study of denoising and generative tasks (see Chapters 10 and 11).

Finally, note that the function ℓ in (11) is a compact and generic way to write the final loss function directly in terms of the overlaps between the parameters $\boldsymbol{w}, \boldsymbol{w}_\star$ and the data $\boldsymbol{x}$. In particular, model specifications such as the activation functions are kept implicit, and subsumed within ℓ. Finally, the main object we seek to characterize is the average test error (4)

$$\varepsilon_g = \varepsilon_g = \mathbb{E}_{\boldsymbol{x}} \ell_{\text{tst.}} \left(\frac{\boldsymbol{x}\boldsymbol{w}_\star}{\sqrt{d}}, \frac{\boldsymbol{x}\hat{\boldsymbol{w}}}{\sqrt{d}}, \frac{\hat{\boldsymbol{w}}^\top \hat{\boldsymbol{w}}}{d}, c \right). \tag{13}$$

The ERM problem (11) is a *high-dimensional, non-linear, non-convex* optimization problem and thus challenging on several levels. The next subsections highlights how the the minimizers of the empirical risk $\mathscr{R}(\boldsymbol{w})$ (11) can be tightly characterized, deploying ideas borrowed from statistical physics, in the joint asymptotic limit where the number of samples n and the dimension d jointly tend to infinity $d, n \to \infty$, while staying comparably large $\alpha \equiv n/d = \Theta_d(1)$. All other dimensions of the problem L, r, t, as well as the norm of the means $\|\boldsymbol{\mu}_{\ell,k}\|$, are on the hand assumed to remain $\Theta_d(1)$. This particular asymptotic limit was considered in a stream of works – e.g. (Aubin et al., 2018b; Maillard et al., 2020a; Donoho et al., 2009; Gardner et al., 1988; Gardner et al., 1989; El Karoui et al., 2010; Goldt et al., 2020c; Seung et al., 1992; Sompolinsky et al., 1990) – and is often referred to as the *proportional regime.*

α is called the sample complexity.

1.2.3 THE REPLICA METHOD

The replica method (Parisi, 1979a; Parisi, 1983b) (see also (Mézard et al., 2009) for a review) starts from the simple observation that for any test function (observable) $\phi(\hat{\boldsymbol{w}})$ of the trained weights $\hat{\boldsymbol{w}}$ – such as the test error (13)–, one can write

$$\mathbb{E}_{\mathscr{D}} \phi(\hat{w}) = \lim_{\beta \to \infty} \mathbb{E}_{\mathscr{D}} \frac{1}{Z} \int d\boldsymbol{w} e^{-\beta \mathscr{R}(\boldsymbol{w})} \phi(\boldsymbol{w}), \tag{14}$$

where we introduced the *partition function* (normalization factor)

$$Z = \int d\boldsymbol{w} e^{-\beta \mathscr{R}(\boldsymbol{w})}. \tag{15}$$

Characterizing the average $\mathbb{E}_{\mathscr{D}} \phi(\hat{w})$ is thus tantamount to studying the family of $\beta-$ parametrized measures $\mathbb{P}_\beta(\boldsymbol{w}) = e^{-\beta \mathscr{R}(\boldsymbol{w})}/Z$, for different values

In statistical physics, $\mathbb{P}_\beta$ is called a Boltzmann distribution. It discribes the equilibrium distribution of the particles w_i when their interaction energy is given by $\mathscr{R}(\boldsymbol{w})$.

of the *inverse temperature* $\beta > 0$. To that end, it is natural to focus on studying the cumulant-generating function

$$f = -\lim_{\beta \to \infty} \frac{1}{\beta d} \mathbb{E}_{\mathscr{D}} \ln Z. \tag{16}$$

In statistical physics, f is called the *free energy*. Analytically evaluating the logarithm of a random variable Z is, however, usually a complex enterprise. The *replica method* builds on the simplifying identity

$$\mathbb{E}_{\mathscr{D}} \ln Z = \lim_{s \to 0} \frac{\mathbb{E}_{\mathscr{D}} Z^s - 1}{s} \tag{17}$$

to map it back to the simpler computation of the moment Z^s, for a parameter $s \to 0$. Note that Z^s corresponds to the partition function (normalization) of the product measure of s copies (the eponymous replicas) of the original problem. The centerpiece of the replica method then lies in the computation of $\mathbb{E}_{\mathscr{D}} Z^s$, which we detail below.

The replicated partition function Z^s reads

$$\begin{aligned}\mathbb{E}_{\mathscr{D}} Z^s &= \int \prod_{a=1}^{s} d\boldsymbol{w}_a e^{-\beta \sum\limits_{a=1}^{s} \lambda \|\boldsymbol{w}_a\|^2} \prod_{\mu=1}^{n} \mathbb{E}_x e^{-\beta \sum\limits_{a=1}^{s} \ell\left(\frac{\boldsymbol{x}\boldsymbol{w}_\star}{\sqrt{d}}, \frac{\boldsymbol{x}\boldsymbol{w}_a}{\sqrt{d}}, \frac{\|\boldsymbol{w}_a\|^2}{d}, c\right)} \\ &= \int \prod_{a=1}^{s} d\boldsymbol{w}_a e^{-\beta \sum\limits_{a=1}^{s} \lambda \|\boldsymbol{w}_a\|^2} \prod_{\mu=1}^{n} \mathbb{E}_c \left[\mathbb{E}_{x|c} e^{-\beta \sum\limits_{a=1}^{s} \ell\left(\frac{\boldsymbol{x}\boldsymbol{w}_\star}{\sqrt{d}}, \frac{\boldsymbol{x}\boldsymbol{w}_a}{\sqrt{d}}, \frac{\boldsymbol{w}_a^\top \boldsymbol{w}_a}{d}, c\right)}\right]\end{aligned} \tag{18}$$

To simplify the expression, let us introduce the random variables

$$h^a \equiv \frac{(\boldsymbol{x} - \boldsymbol{\mu}_c)\boldsymbol{w}_a}{\sqrt{d}} \in \mathbb{R}^{L \times r}, \qquad h^\star \equiv \frac{(\boldsymbol{x} - \boldsymbol{\mu}_c)\boldsymbol{w}_\star}{\sqrt{d}} \in \mathbb{R}^{L \times t}. \tag{19}$$

and the parameters

$$m_a^c \equiv \frac{\boldsymbol{\mu}_c \boldsymbol{w}_a}{\sqrt{d}} \in \mathbb{R}^{L \times r}, \qquad m_\star^c \equiv \frac{\boldsymbol{\mu}_c \boldsymbol{w}_\star}{\sqrt{d}} \in \mathbb{R}^{L \times t}, \tag{20}$$

with rows $m_a^\ell, m_\star^\ell$. The variables (19) are Gaussian with statistics

$$\mathbb{E}_{x|c}[h_\ell^a (h_\kappa^b)^\top] = \delta_{\ell\kappa} \frac{\boldsymbol{w}_a^\top \Sigma_{\ell,c_\ell} \boldsymbol{w}_b}{d} \equiv q_{ab}^{\ell,c_\ell}, \tag{21}$$

$$\mathbb{E}_{x|c}[h_\ell^\star (h_\kappa^\star)^\top] = \delta_{\ell\kappa} \frac{\boldsymbol{w}_\star^\top \Sigma_{\ell,c_\ell} \boldsymbol{w}_\star}{d} \equiv \rho_{\ell,c_\ell}, \tag{22}$$

$$\mathbb{E}_{x|c}[h_\ell^a (h_\kappa^\star)^\top] = \delta_{\ell\kappa} \frac{\boldsymbol{w}_a^\top \Sigma_{\ell,c_\ell} \boldsymbol{w}_\star}{d} \equiv \theta_a^{\ell,c_\ell}. \tag{23}$$

Let us also define

$$v_a = \frac{\boldsymbol{w}_a^\top \boldsymbol{w}_a}{d}. \tag{24}$$

Therefore, the replicated partition function (18) can be rewritten as

$$
\begin{aligned}
\mathbb{E}_{\mathscr{D}} Z^s = & \int \prod_{a} dv_a d\hat{v}_a \prod_{\ell=1}^{L} \prod_{k=1}^{K_\ell} \prod_{a=1}^{s} dm_a^{\ell,k} d\hat{m}_a^{\ell,k} d\theta_a^{\ell,k} d\hat{\theta}_a^{\ell,k} \prod_{a\leq b}^{s} dq_{ab}^{\ell,k} d\hat{q}_{ab}^{\ell,k} \\
& \underbrace{e^{-d\sum\limits_a\sum\limits_\ell\sum\limits_k\left[\hat{m}_a^{\ell,k\top} m_a^{\ell,k} + \mathrm{Tr}\left(\theta_a^{\ell,k}\hat{\theta}_a^{\ell,k\top}\right)\right] - d\sum\limits_\ell\sum\limits_k\sum\limits_{1\leq a\leq b\leq s}\mathrm{Tr}\left(q_{ab}^{\ell,k}\hat{q}_{ab}^{\ell,k\top}\right) - d\sum\limits_a \mathrm{Tr}[v_a\hat{v}_a]}}_{e^{sd\beta\Psi_t}} \\
& \int \prod_{a=1}^{s} d\boldsymbol{w}_a e^{\sum\limits_{a=1}^{s} -\beta\lambda\|\boldsymbol{w}_a\|^2 + \mathrm{Tr}\left[\hat{v}_a \boldsymbol{w}_a^\top \boldsymbol{w}_a\right]} \\
& \underbrace{e^{+\sum\limits_a\sum\limits_\ell\sum\limits_k(\sqrt{d}\hat{m}_a^{\ell,k\top}\boldsymbol{w}_a^\top\boldsymbol{\mu}_{\ell,k} + \mathrm{Tr}\left[\hat{\theta}_a^{\ell,k}\boldsymbol{w}_\star^\top\Sigma_{\ell,k}\boldsymbol{w}_a\right]) + \sum\limits_{1\leq a\leq b\leq s}\sum\limits_\ell\sum\limits_k \mathrm{Tr}\left[\hat{q}_{ab}^{\ell,k}\boldsymbol{w}_b^\top\Sigma_{\ell,k}\boldsymbol{w}_a\right]}}_{e^{sd\beta\Psi_w}} \\
& \underbrace{\left[\mathbb{E}_c \mathbb{E}_{h^\star,\{h_a\}_{a=1}^s|c} e^{-\beta\sum\limits_{a=1}^{s}\ell(h_\star + m_\star^c, h^a + m_a^c, v_a, c)}\right]^{\alpha d}}_{e^{s\alpha d\beta\Psi_y}}.
\end{aligned}
\tag{25}
$$

We decomposed the replicated free entropy into the trace, entropic and energetic potentials Ψ_t, Ψ_w, Ψ_y, which we shall study in turn in the following. Importantly, note that all exponents are scaling with $d \to \infty$. Therefore the integral in (25) can be computed using a Laplace saddle-point approximation, and reduces to an extremization problem.

1.2.3.A REPLICA-SYMMETRIC ANSATZ

We have thus rephrased the analysis of the average (14) as an optimization problem over the order parameters $\{q_{ab}^{\ell,k}, \theta_a^{\ell,k}, m_a^{\ell,k}, v_a\}$, and the associated conjugate variables. While conceptually simpler, this optimization still bears over $2L(s^2+1)+2s$ variables. Besides, one needs to further deal with the analytical continuation $s \to 0$. In order to make progress, one can look for the extremizer of the exponent of (25) in a specific form. A particular prescription is the Replica Symmetry (RS) ansatz (Parisi, 1983b; Parisi, 1979a)

$$q_{ab}^{\ell,k} = (r_{\ell,k} - q_{\ell,k})\delta_{ab} + q_{\ell,k}, \tag{26}$$

$$m_a^{\ell,k} = m_{\ell,k}, \tag{27}$$

$$\theta_a^{\ell,k} = \theta_{\ell,k}, \tag{28}$$

$$v_a = v, \tag{29}$$

$$\hat{q}_{ab}^{\ell,k} = -\left({\scriptstyle \hat{r}_{\ell,k}}/2 + \hat{q}_{\ell,k}\right) + \hat{q}_{\ell,k}, \tag{30}$$

$$\hat{m}_a^{\ell,k} = \hat{m}_{\ell,k}, \tag{31}$$

$$\hat{\theta}_a^{\ell,k} = \hat{\theta}_{\ell,k}, \tag{32}$$

$$\hat{v}_a = -\frac{1}{2}\hat{v}. \tag{33}$$

In words, the RS ansatz assumes that the overlaps between any two distinct replicas are identical, and that all replicas further share the same overlap with the target weights. The RS ansatz (26) is in particular always correct for convex problems (Zdeborová et al., 2016). One is now in a position to sequentially simplify the expressions of the potentials Ψ_t, Ψ_w, Ψ_y. Crucially, motivated by the definition of these quantities, we assume that $r_{\ell,k}, q_{\ell,k}, v, \hat{r}_{\ell,k}, \hat{q}_{\ell,k}, \hat{v}$ are symmetric matrices.

1.2.3.B TRACE POTENTIAL

To leading order in s, under the RS ansatz (26), the trace potential Ψ_t can be compactly written as

$$\begin{aligned}\beta\Psi_t = &-\sum_\ell \sum_k \left(\hat{m}_{\ell,k}^\top m_{\ell,k} + \mathrm{Tr}\left[\theta_{\ell,k}\hat{\theta}_{\ell,k}^\top + \frac{(V_{\ell,k}+q_{\ell,k})(\hat{V}_{\ell,k}-\hat{q}_{\ell,k})^\top}{2} + \frac{q_{\ell,k}\hat{q}_{\ell,k}^\top}{2} \right] \right) \\ &+ \frac{1}{2}\mathrm{Tr}[v\hat{v}], \end{aligned} \tag{34}$$

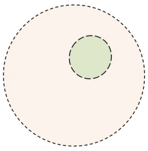
where we introduced the variance order parameters

$$V_{\ell,k} \equiv r_{\ell,k} - q_{\ell,k}, \qquad \hat{V}_{\ell,k} \equiv \hat{r}_{\ell,k} + \hat{q}_{\ell,k}. \tag{35}$$

1.2.3.C ENTROPIC POTENTIAL

We now turn to the entropic potential Ψ_w, which can be expressed as

$$\begin{aligned} e^{\beta s d \Psi_w} = &\int \prod_{a=1}^s d\boldsymbol{w}_a e^{\sum_{a=1}^s -\beta \frac{\lambda}{2}\|\boldsymbol{w}_a\|^2 + \mathrm{Tr}\left[\hat{v}\boldsymbol{w}_a^\top \boldsymbol{w}_a\right]} \\ &e^{\sum_a \sum_\ell \sum_k (\sqrt{d}\hat{m}_a^{\ell,k\top} \boldsymbol{w}_a^\top \boldsymbol{\mu}_{\ell,k} + \mathrm{Tr}\left[\hat{\theta}_a^{\ell,k} \boldsymbol{w}_\star^\top \Sigma_{\ell,k} \boldsymbol{w}_a\right]) + \sum_{1\le a\le b\le s} \sum_\ell \sum_k \mathrm{Tr}\left[\hat{q}_{ab}^{\ell,k} \boldsymbol{w}_b^\top \Sigma_{\ell,k} \boldsymbol{w}_a\right]} \\ = &\mathbb{E}_\Xi \left[\int d\boldsymbol{w} e^{H(\boldsymbol{w},\Xi)} \right]^s . \end{aligned} \tag{36}$$



The entropic potential Ψ_w measures the volume in weight space corresponding to the overlaps $q_\ell, m_\ell, \theta_\ell$.

The expectation bears over a tensor $\Xi \in \mathbb{R}^{L\times r\times d}$ with i.i.d standard Gaussian entries, and we introduced the shorthand

$$\begin{aligned} &H(\boldsymbol{w}, \Xi) \\ &\equiv -\beta g(\boldsymbol{w}) - \frac{1}{2}\boldsymbol{w} \odot \left[\hat{v} \otimes \mathbb{I}_d + \sum_\ell \sum_k \hat{V}_{\ell,k} \otimes \boldsymbol{\Sigma}_{\ell,k} \right] \odot \boldsymbol{w} \\ &+ \left(\sum_\ell \sum_k \sqrt{d}\hat{m}_{\ell,k}\boldsymbol{\mu}_{\ell,k}^\top + \hat{\theta}_{\ell,k}\boldsymbol{w}_\star^\top \Sigma_{\ell,k} + \Xi_{\ell,k} \odot (\hat{q}_{\ell,k} \otimes \boldsymbol{\Sigma}_{\ell,k})^{\frac{1}{2}} \right) \odot \boldsymbol{w}. \end{aligned} \tag{37}$$

Therefore,

$$\beta\Psi_w = \frac{1}{d}\int \mathbb{E}_\Xi \ln \left[\int d\boldsymbol{w} e^{H(\boldsymbol{w},\Xi)} \right]. \tag{38}$$

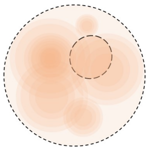
For a matrix $\boldsymbol{\xi} \in \mathbb{R}^{r\times d}$ and tensors $\boldsymbol{A}, \boldsymbol{B} \in \mathbb{R}^{r\times d} \otimes \mathbb{R}^{r\times d}$, we denoted $(\boldsymbol{\xi} \odot \boldsymbol{A})_{kl} = \sum_{ij} \boldsymbol{\xi}^{ij} \boldsymbol{A}_{ij,kl}$ and $(\boldsymbol{A} \odot \boldsymbol{B})_{ij,kl} = \sum_{rs} \boldsymbol{A}_{ij,rs} \boldsymbol{B}_{rs,kl}$.

1.2.3.D ENERGETIC POTENTIAL

The computation of the energetic potential Ψ_y requires more lengthy, albeit straightforward, steps. For the sake of the conciseness of the presentation, we do not exhaustively reproduce all of them here and refer the interested reader to e.g. Appendix IV.2.1 of (Aubin et al., 2020a). We only report here the last steps:

$$\begin{aligned}&\beta\Psi_y\\&=\underbrace{\mathbb{E}_c\int_{\mathbb{R}^{L\times t}}dY\,\mathbb{E}_\Xi\prod_{\ell=1}^{L}\frac{e^{-\frac{1}{2}\left(y_\ell-\theta_{\ell,c_\ell}^\top q_{\ell,c_\ell}^{\frac{1}{2}}\xi_\ell\right)^\top(\rho_{\ell,c_\ell}-\theta_{\ell,c_\ell}^\top q_{\ell,c_\ell}^{-1}\theta_{\ell,c_\ell})^{-1}\left(y_\ell-\theta_{\ell,c_\ell}^\top q_{\ell,c_\ell}^{\frac{1}{2}}\xi_\ell\right)}}{\sqrt{\det\left(2\pi(\rho_{\ell,c_\ell}-\theta_{\ell,c_\ell}^\top q_{\ell,c_\ell}^{-1}\theta_{\ell,c_\ell})\right)}}}_{\equiv\mathbb{E}_{c,Y,\Xi}}\\&\times\ln\left[\int_{\mathbb{R}^{L\times r}}dX\prod_{\ell=1}^{L}\frac{e^{-\frac{1}{2}\left(x_\ell-q_{\ell,c_\ell}^{\frac{1}{2}}\xi_\ell\right)^\top V_{\ell,c_\ell}^{-1}\left(x_\ell-q_{\ell,c_\ell}^{\frac{1}{2}}\xi_\ell\right)}}{\sqrt{\det\left(2\pi V_{\ell,c_\ell}\right)}}e^{-\beta\ell(Y+m_\star^c,X+m^c,v,c)}\right].\end{aligned}\tag{39}$$

The expectation again bears over a tensor $\Xi\in\mathbb{R}^{L\times r\times d}$ with i.i.d standard Gaussian entries.



The energetic potential Ψ_y measures the average loss corresponding to the overlaps $q_\ell, m_\ell, \theta_\ell$.

1.2.3.E ZERO-TEMPERATURE LIMIT

We now take the limit $\beta\to\infty$. Rescaling

$$\begin{aligned}&\beta\hat{V}_{\ell,k}\leftarrow\hat{V}_{\ell,k}, && \frac{1}{\beta}V_{\ell,k}\leftarrow V_{\ell,k}, && \beta\hat{m}_{\ell,k}\leftarrow\hat{m}_{\ell,k},\\&\beta\hat{\theta}_{\ell,k}\leftarrow\hat{\theta}_{\ell,k}, && \beta^2\hat{q}_{\ell,k}\leftarrow\hat{q}_{\ell,k}, && \beta\hat{v}\leftarrow\hat{v}\end{aligned}\tag{40}$$

the entropic potential Ψ_w (38) then reduces to

$$\begin{aligned}\Psi_w=&\frac{1}{2d}\mathbb{E}_\Xi\,\mathrm{Tr}\left[\check{V}^{-1}\odot\left(\sum_\ell\sum_k\sqrt{d}\hat{m}_{\ell,k}\boldsymbol{\mu}_{\ell,k}^\top+\hat{\theta}_{\ell,k}\boldsymbol{w}_\star^\top\Sigma_{\ell,k}+\Xi_{\ell,k}\odot(\hat{q}_{\ell,k}\otimes\boldsymbol{\Sigma}_{\ell,k})^{\frac{1}{2}}\right)^{\otimes 2}\right]\\&-\frac{1}{d}\mathbb{E}_\Xi\mathscr{M}_g(\Xi),\end{aligned}\tag{41}$$

where we defined the entropic Moreau enveloppe

$$\begin{aligned}&M_g(\Xi)\\&\equiv\inf_{\boldsymbol{w}}\left[\frac{1}{2}\left\|\check{V}^{1/2}\left(\boldsymbol{w}-\check{V}^{-1}\left(\sum_\ell\sum_k\sqrt{d}\hat{m}_{\ell,k}\boldsymbol{\mu}_{\ell,k}^\top+\hat{\theta}_{\ell,k}\boldsymbol{w}_\star^\top\Sigma_{\ell,k}+\Xi_{\ell,k}\odot(\hat{q}_{\ell,k}\otimes\boldsymbol{\Sigma}_{\ell,k})^{\frac{1}{2}}\right)\right)\right\|^2+\frac{\lambda}{2}\|\boldsymbol{w}\|^2\right].\end{aligned}\tag{42}$$

and used the shorthand

$$\check{V}\equiv\hat{v}\otimes\mathbb{I}_d+\sum_\ell\sum_k\hat{V}_{\ell,k}\otimes\Sigma_{\ell,k}.\tag{43}$$

The energetic potential Ψ_y (39) can be similarly recast into a more compact form

$$\Psi_y = -\mathbb{E}_{c,Y,\Xi}\mathscr{M}(c,Y,\Xi), \tag{44}$$

where the Moreau envelope is defined as

$$\begin{aligned}&\mathscr{M}(c,Y,\Xi)\\&=\inf_X\left\{\frac{1}{2}\sum_{\ell=1}^{L}\mathrm{Tr}\left[V_{\ell,c_\ell}^{-1}\left(x_\ell-q_{\ell,c_\ell}^{1/2}\xi_\ell-m_{\ell,c_\ell}\right)^{\otimes 2}\right]+\ell\left(Y+m_\star^c,X,v,c\right)\right\}.\end{aligned} \tag{45}$$

1.2.3.F REPLICA FREE ENERGY

The RS free energy

$$f=-\lim_{\beta\to\infty}\frac{1}{\beta d}\mathbb{E}_{\mathscr{D}}\ln Z=-\Psi_t-\Psi_w-\alpha\Psi_y \tag{46}$$

can finally be written as the solution of the low-dimensional optimization problem

$$f=-\mathrm{extr}\ \Phi \tag{47}$$

where the extremization bears on the variables $q_{\ell,k}, V_{\ell,k}, m_{\ell,k}, \theta_{\ell,k}, v, \hat{q}_{\ell,k}, \hat{V}_{\ell,k}, \hat{m}_{\ell,k}, \hat{\theta}_{\ell,k}, \hat{v}$ and the *free entropy* Φ reads

$$\begin{aligned}&\Phi\\&=\sum_\ell\sum_k\left(\frac{1}{2}\mathrm{Tr}\left[q_{\ell,k}\hat{V}_{\ell,k}^\top-V_{\ell,k}\hat{q}_{\ell,k}^\top\right]-\mathrm{Tr}\left[\theta_{\ell,k}\hat{\theta}_{\ell,k}^\top\right]-\hat{m}_{\ell,k}^\top m_{\ell,k}\right)+\frac{1}{2}\mathrm{Tr}[v\hat{v}]\end{aligned} \tag{48}$$

$$\begin{aligned}&+\frac{1}{2d}\mathbb{E}_\Xi\mathrm{Tr}\left[\left(\hat{v}\otimes\mathbb{I}_d+\sum_\ell\sum_k\hat{V}_{\ell,k}\otimes\Sigma_{\ell,k}\right)^{-1}\odot\left(\sum_\ell\sum_k\sqrt{d}\hat{m}_{\ell,k}\boldsymbol{\mu}_{\ell,k}^\top+\hat{\theta}_{\ell,k}\boldsymbol{w}_\star^\top\Sigma_{\ell,k}+\Xi_{\ell,k}\odot(\hat{q}_{\ell,k}\otimes\boldsymbol{\Sigma}_{\ell,k})^{\frac{1}{2}}\right)^{\otimes 2}\right]\\&-\frac{1}{d}\mathbb{E}_\Xi\mathscr{M}_g(\Xi)-\alpha\mathbb{E}_{c,Y,\Xi}\mathscr{M}(c,Y,\Xi).\end{aligned}$$

Note that these equations are not yet fully asymptotic, in the sense that they still involve a high-dimensional optimization problem in the form of the entropic Moreau envelope (42). For many regularizers $g(\cdot)$ of interest however, prox_g admits a simple fully asymptotic closed-form expression. We first need the additional assumption:

Assumption 1.2.1. *The set of matrices $\{\{\Sigma_{\ell,k}\}_{k=1}^{K_\ell}\}_{\ell=1}^{L}$ admits a common set of eingenvectors $\{\boldsymbol{e}_i\}_{i=1}^{d}$. We denote $\{\lambda_i^{\ell,k}\}_{i=1}^{d}$ the eigenvalues of$\Sigma_{\ell,k}$. The eigenvalues $\{\lambda_i^{\ell,k}\}_{\ell,k,i}$ and the projection of the cluster means $\{\boldsymbol{\mu}_{\ell,k}\}_{\ell,k}$ and the teacher columns $\{(\boldsymbol{w}_\star)_i\}_{i=1}^{t}$ on these eigenvectors are assumed to admit* *a well-defined joint distribution ν as $d \to \infty$ – namely, for $\gamma = (\gamma_{\ell,k})_{\ell,k}$, $\pi = (\pi_1, \dots, \pi_t) \in \mathbb{R}^t$ and $\tau = (\tau_{\ell,k})_{\ell,k}$:*

$$\frac{1}{d}\sum_{i=1}^{d}\prod_{\ell=1}^{L}\prod_{k=1}^{K_\ell}\delta\left(\lambda_i^{\ell,k} - \gamma_{\ell,k}\right)\delta\left(\sqrt{d}\boldsymbol{e}_i^\top \boldsymbol{\mu}_{\ell,k} - \tau_{\ell,k}\right)\prod_{j=1}^{t}\delta\left(\boldsymbol{e}_i^\top(\boldsymbol{w}_\star)_j - \pi_j\right) \xrightarrow{d\to\infty} \nu\left(\gamma, \tau, \pi\right). \tag{49}$$

This assumption is satisfied e.g. when all covariances Σ_ℓ are jointly diagonalizable (or equal), and $\boldsymbol{w}_\star, \boldsymbol{\mu}_\ell$ are i.i.d (or just independent) Gaussian vectors, which is the setting in a large number of past studies of ML with statistical physics ideas, see e.g. (Zdeborová et al., 2016; Gabrié, 2020) for reviews.

The Saddle Point (SP) equations, expressing the extremization conditions, can then be written as

$$\begin{cases}
\hat{q}_{\ell,k} = \alpha\mathbb{E}_c\delta_{c_\ell,k}\mathbb{E}_{\Xi,Y}V_{\ell,k}^{-1}\left(\mathrm{prox}_\ell^c - q_{\ell,k}^{\frac{1}{2}}\xi_\ell - m_{\ell,k}\right)^{\otimes 2}V_{\ell,k}^{-1} \\
\hat{V}_{\ell,k} = \hat{\theta}_{\ell,k}\theta_{\ell,k}^\top q_{\ell,k}^{-1} - \alpha\mathbb{E}_c\delta_{c_\ell,k}\mathbb{E}_{\Xi,Y}V_{\ell,k}^{-1}\left(\mathrm{prox}_\ell^c - q_{\ell,k}^{\frac{1}{2}}\xi_\ell - m_{\ell,k}\right)\xi_\ell^\top q_{\ell,k}^{-\frac{1}{2}} \\
\hat{m}_{\ell,k} = \alpha\mathbb{E}_c\delta_{c_\ell,k}\mathbb{E}_{\Xi,Y}V_{\ell,k}^{-1}\left(\mathrm{prox}_\ell^c - q_{\ell,k}^{\frac{1}{2}}\xi_\ell - m_{\ell,k}\right) \\
\hat{\theta}_{\ell,k} = \alpha\mathbb{E}_c\delta_{c_\ell,k}\mathbb{E}_{\Xi,Y}V_{\ell,k}^{-1}\left(\mathrm{prox}_\ell^c - q_{\ell,k}^{\frac{1}{2}}\xi_\ell - m_{\ell,k}\right) \\
\qquad \left(y_\ell - \theta_{\ell,k}^\top q_{\ell,k}^{-1/2}\xi_\ell\right)^\top\left(\rho_{\ell,k} - \theta_{\ell,k}^\top q_{\ell,k}^{-1}\theta_{\ell,k}\right)^{-1} \\
\hat{v} = 2\alpha\mathbb{E}_c\mathbb{E}_{\Xi,Y}\partial_3\ell(Y + m_\star^c, \mathrm{prox}^c, v, c)
\end{cases} \tag{50}$$

$$\begin{cases}
q_{\ell,k} = \int d\nu(\gamma,\tau,\pi)\gamma_{\ell,k}\left(\lambda\mathbb{I}_r + \hat{v} + \sum_\kappa\sum_j\gamma_{\kappa,j}\hat{V}_{\kappa,j}\right)^{-1} \\
\qquad \left[\left(\sum_\kappa\sum_j\hat{m}_{\kappa,j}\tau_{\kappa,j} + \gamma_{\kappa,j}\hat{\theta}_{\kappa,j}\pi\right)^{\otimes 2} + \sum_\kappa\sum_j\gamma_{\kappa,j}\hat{q}_{\kappa,j}\right] \\
\qquad \left(\lambda\mathbb{I}_r + \hat{v} + \sum_\kappa\sum_j\gamma_{\kappa,j}\hat{V}_{\kappa,j}\right)^{-1} \\
V_{\ell,k} = \int d\nu(\gamma,\tau,\pi)\gamma_{\ell,k}\left(\lambda\mathbb{I}_r + \hat{v} + \sum_\kappa\sum_j\gamma_{\kappa,j}\hat{V}_{\kappa,j}\right)^{-1} \\
m_{\ell,k} = \int d\nu(\gamma,\tau,\pi)\tau_{\ell,k}\left(\lambda\mathbb{I}_r + \hat{v} + \sum_\kappa\sum_j\gamma_{\kappa,j}\hat{V}_{\kappa,j}\right)^{-1} \\
\qquad \left(\sum_\kappa\sum_j\hat{m}_{\kappa,j}\tau_{\kappa,j} + \gamma_{\kappa,j}\hat{\theta}_{\kappa,j}\pi\right) \\
\theta_{\ell,k} = \int d\nu(\gamma,\tau,\pi)\gamma_{\ell,k}\left(\lambda\mathbb{I}_r + \hat{v} + \sum_\kappa\sum_j\gamma_{\kappa,j}\hat{V}_{\kappa,j}\right)^{-1} \\
\qquad \left(\sum_\kappa\sum_j\hat{m}_{\kappa,j}\tau_{\kappa,j} + \gamma_{\kappa,j}\hat{\theta}_{\kappa,j}\pi\right)\pi^\top \\
v = \int d\nu(\gamma,\tau,\pi)\left(\lambda\mathbb{I}_r + \hat{v} + \sum_\kappa\sum_j\gamma_{\kappa,j}\hat{V}_{\kappa,j}\right)^{-1} \\
\qquad \left[\left(\sum_\kappa\sum_j\hat{m}_{\kappa,j}\tau_{\kappa,j} + \gamma_{\kappa,j}\hat{\theta}_{\kappa,j}\pi\right)^{\otimes 2} + \sum_\kappa\sum_j\gamma_{\kappa,j}\hat{q}_{\kappa,j}\right] \\
\qquad \left(\lambda\mathbb{I}_r + \hat{v} + \sum_\kappa\sum_j\gamma_{\kappa,j}\hat{V}_{\kappa,j}\right)^{-1}
\end{cases} \tag{51}$$

How to interpret the parameters $q_{\ell,k}, m_{\ell,k}, \theta_{\ell,k}$? Consider the summary statistics

$$\check{q}_{\ell,k}(\hat{\boldsymbol{w}}) = \frac{\hat{\boldsymbol{w}}^\top \Sigma_{\ell,k} \hat{\boldsymbol{w}}}{d}, \quad \check{m}_{\ell,k}(\hat{\boldsymbol{w}}) = \frac{\boldsymbol{\mu}_{\ell,k}^\top \hat{\boldsymbol{w}}}{\sqrt{d}}, \quad \check{\theta}_{\ell,k}(\hat{\boldsymbol{w}}) = \frac{\hat{\boldsymbol{w}}^\top \Sigma_{\ell,k} \boldsymbol{w}_\star}{d}, \tag{52}$$

of the trained weights $\hat{\boldsymbol{w}}$. The average value of any test function $\phi(\hat{\boldsymbol{w}}) \equiv \phi(\{\check{q}_{\ell,k}(\hat{\boldsymbol{w}}), \check{m}_{\ell,k}(\hat{\boldsymbol{w}}), \check{\theta}_{\ell,k}(\hat{\boldsymbol{w}})\}_{\ell,k})$ of these statistics can be rewritten similarly to (14) as

$$\begin{aligned}
\mathbb{E}_{\mathscr{D}} \phi(\hat{\boldsymbol{w}}) &= \lim_{\beta\to\infty} \mathbb{E}_{\mathscr{D}} \frac{1}{Z} \int d\boldsymbol{w} e^{-\beta \mathscr{R}(\boldsymbol{w})} \phi(\boldsymbol{w}) \\
&= \lim_{\beta\to\infty, s\to 0} \int d\boldsymbol{w} e^{-\beta \mathscr{R}(\boldsymbol{w})} \phi(\boldsymbol{w}) Z^{s-1} \\
&= \lim_{\beta\to\infty, s\to 0} \int \prod_{a=1}^{s} d\boldsymbol{w}^a e^{-\beta \mathscr{R}(\boldsymbol{w}^a)} \phi(\boldsymbol{w}^1) \\
&= \lim_{\beta\to\infty, s\to 0} \int \prod_{\ell=1}^{L} \prod_{k=1}^{K_\ell} \prod_{a=1}^{s} dm_a^{\ell,k} d\hat{m}_a^{\ell,k} d\theta_a^{\ell,k} d\hat{\theta}_a^{\ell,k} \\
&\qquad \prod_{a\le b}^{s} dq_{ab}^{\ell,k} d\hat{q}_{ab}^{\ell,k} \phi(\{q_{11}^{\ell,k}, m_1^{\ell,k}, \theta_1^{\ell,k}\}_{\ell,k}) e^{sd\beta(\Psi_t + \Psi_w + \Psi_y)} \\
&\asymp \lim_{\beta\to\infty, s\to 0} \phi(\{q_{\ell,k}, m_{\ell,k}, \theta_{\ell,k}\}_{\ell,k}) e^{-s\beta d f} \\
&= \phi(\{q_{\ell,k}, m_{\ell,k}, \theta_{\ell,k}\}_{\ell,k}),
\end{aligned} \tag{53}$$

where the last equality results from first taking the $s \to 0$ limit. In words, the average of any function of the summary statistics $\check{q}_{\ell,k}(\hat{\boldsymbol{w}}), \check{m}_{\ell,k}(\hat{\boldsymbol{w}}), \check{\theta}_{\ell,k}(\hat{\boldsymbol{w}})$ (52), and in particular the average of the summary statistics themselves, is asymptotically given by (a function of) the quantities $q_{\ell,k}, m_{\ell,k}, \theta_{\ell,k}$ characterized by the SP equations (50). The replica method thus provides a powerful framework to *tightly* describe the minimizer $\hat{\boldsymbol{w}}$ of the ERM problem (11), in terms of a set of summary statistics.

1.2.3.G TEST ERROR

We remind the expression for the average test error associated to the trained seq-GLM:

$$\epsilon_g = \mathbb{E}_{\boldsymbol{x}} \ell_{\text{tst.}} \left(\frac{\boldsymbol{x} \boldsymbol{w}_\star}{\sqrt{d}}, \frac{\boldsymbol{x} \hat{\boldsymbol{w}}}{\sqrt{d}}, \frac{\hat{\boldsymbol{w}}^\top \hat{\boldsymbol{w}}}{d}, c \right). \tag{54}$$

Explicating this expression in terms of the correlated Gaussian variables $\boldsymbol{x}\hat{\boldsymbol{w}}, \boldsymbol{x}\boldsymbol{w}_\star$ allows to derive the following compact asymptotic characterization:

$$\epsilon_g = \mathbb{E}_{c,X,Y} \ell_{\text{ts.}} (Y, X, \nu, c), \tag{55}$$

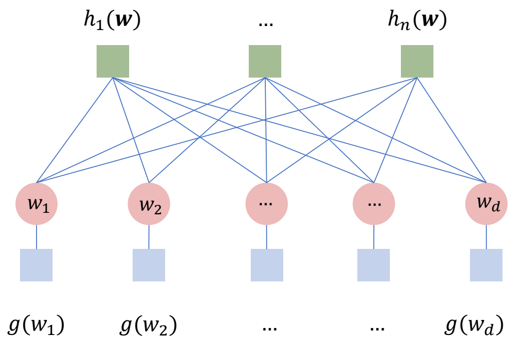


Figure 1: Graphical model associated to the measure $\mathbb{P}_\beta$ (15). We used the shorthands $h_\mu(\boldsymbol{w}) \equiv \exp(\beta \ell(\boldsymbol{x}^\mu \boldsymbol{w}_\star/\sqrt{d}, \boldsymbol{x}^\mu \boldsymbol{w}/\sqrt{d}, \boldsymbol{w}^\top \boldsymbol{w}/d, c^\mu)), g(w_i) = \exp(\beta\lambda/2\|w_i\|^2)$. Iterative schemes such as GAMP (2) (Bayati et al., 2011a; Rangan et al., 2016) can be used to estimate marginals from such distributions.

where, conditioned on the class assignments c, the average bears on $X \in \mathbb{R}^{L\times r}, Y \in \mathbb{R}^{L\times t}$ with independent rows with statistics

$$(x_\ell, y_\ell) \sim \mathcal{N}\left[\begin{pmatrix} m_{\ell,c_\ell} \\ m^\star_{\ell,c_\ell} \end{pmatrix}, \left(\begin{array}{c|c} q_{\ell,c_\ell} & \theta_{\ell,c_\ell} \\ \hline \theta^\top_{\ell,c_\ell} & \rho_{\ell,c_\ell} \end{array}\right)\right], \tag{56}$$

where the summary statistics $q_{\ell,c_\ell}, \theta_{\ell,c_\ell}, \rho_{\ell,c_\ell}$ are characterized by (50).

1.2.4 AN ALGORITHMIC PERSPECTIVE

The precedent section 1.2.3.a showed how the learning of a seq-GLM, could be asymptotically characterized in terms of a set of low-dimensional equations (50). In this section, we complement this discussion by providing an alternative *algorithmic* viewpoint on the SP equations (50). More precisely, we will first show that (50) describe the fixed points of an iterative Generalized Approximate Message Passing (GAMP) algorithm. Furthermore, the set of fixed point of GAMP will be shown to coincide with critical (zero gradient) points of the empirical ERM landscape. As a result, it follows that aside from the global minimizer, the other solutions of (50) may describe non-global critical points of the empirical landscape. Finally, note that under Assumption 1.2.1, one can assume without loss of generality all covariances $\Sigma_{\ell,k}$ to be diagonal.

1.2.4.A GAMP ALGORITHM

The measure $\mathbb{P}_\beta$ (15) associated to the ERM problem (11) can be represented as a graphical model, see Fig. 1. For such classes of distributions, *message-passing* algorithms provide a versatile framework to evaluate the marginal $\hat{\boldsymbol{w}}$ (here the trained weights), for any given sample of the train set $\mathscr{D}$. We refer the interested reader to (Mézard et al., 2009; Zdeborová et al., 2016; Gabrié, 2019) for introductions and reviews. For the measure $\mathbb{P}_\beta$ (15), the corresponding relaxed Belief Propagation (rBP) algorithm reads, in the sought-after $\beta \to \infty$ limit:

Algorithm 1 rBP

```
Inputs : $\{X_\ell \in \mathbb{R}^{n\times d}\}_{\ell=1}^{L}, \boldsymbol{y} \in \mathbb{R}^{n\times L\times t}$
Initialize $\forall 1 \le \mu \le n,\ 1 \le i \le d,\ \hat{w}^0_{i\to\mu} = 0_r, \hat{c}^0_{i\to\mu} = \mathbb{I}_r, \{f^0_{\ell\mu\to i} = 0_r\}_{\ell=1}^{L}$

for $t \le t_{\max}$ do
    $\forall 1 \le \ell, \kappa \le L, 1 \le \mu \le n, 1 \le i \le d,\ (V^t_{\mu\to i})_{\ell\kappa} = \frac{1}{d}\sum_{j\ne i}(x^\mu_{\ell j})(x^\mu_{\kappa j})\hat{c}^t_{j\to\mu}$
    $1 \le \mu \le n, 1 \le i \le d,\ \Gamma^t_{\mu\to i} = \frac{1}{d}\sum_{j\ne i}\hat{w}^t_{j\to\mu}(\hat{w}^t_{j\to\mu})^\top$
    $\forall 1 \le \ell, 1 \le \mu \le n, 1 \le i \le d,\ \omega^t_{\ell,\mu\to i} = \frac{1}{\sqrt{d}}\sum_{j\ne i} x^\mu_{\ell,j}\hat{w}^t_{j\to\mu}$
    $\forall 1 \le \ell, 1 \le \mu \le n, 1 \le i \le d,$
    $f^t_{\ell,\mu\to i} = \left[(V^t_{\mu\to i})^{-1}\left(\mathrm{prox}(y_\mu, \omega^t_{\mu\to i}, V^t_{\mu\to i}, \Gamma^t_{\mu\to i}, c^\mu) - \omega^t_{\mu\to i}\right)\right]_\ell$
    $\forall 1 \le \mu \le n, 1 \le i \le d,$
    $\eta^t_{\mu\to i} = \partial_3\ell\left(y_\mu, \mathrm{prox}(y_\mu, \omega^t_{\mu\to i}, V^t_{\mu\to i}, \Gamma^t_{\mu\to i}, c^\mu), \Gamma^t_{\mu\to i}, c^\mu\right)$
    $\forall 1 \le \ell, \kappa \le L, 1 \le \mu \le n, 1 \le i \le d,\ g^t_{\mu\to i} = \nabla_\omega f^t_{\mu\to i}$
    $\forall 1 \le \mu \le n, 1 \le i \le d,\ A^t_{i\to\mu} = -\frac{1}{d}\sum_{\ell,\kappa=1}^{L}\sum_{\nu\ne\mu}(x^\nu_{\ell i})(x^\nu_{\kappa i})g^t_{\ell\kappa,\nu\to i}$
    $\forall 1 \le \mu \le n, 1 \le i \le d,\ C^t_{i\to\mu} = \frac{2}{d}\sum_{\nu\ne\mu}\eta^t_{\nu\to i}$
    $\forall 1 \le \mu \le n, 1 \le i \le d,\ b^t_{i\to\mu} = \frac{1}{\sqrt{d}}\sum_{\ell=1}^{L}\sum_{\nu\ne\mu} x^\nu_{\ell i} f^t_{\ell,\nu\to i}$
    $\forall 1 \le \mu \le n, 1 \le i \le d,\ \hat{w}^{t+1}_{i\to\mu} = (\lambda\mathbb{I}_r + C^t_{i\to\mu} + A^t_{i\to\mu})^{-1} b^t_{i\to\mu}$
    $\forall 1 \le \mu \le n, 1 \le i \le d,\ \hat{c}^{t+1}_{i\to\mu} = (\lambda\mathbb{I}_r + C^t_{i\to\mu} + A^t_{i\to\mu})^{-1}$
end for

return Estimator $\hat{\boldsymbol{w}}$
```

We noted, for $1 \le \ell \le L$, $X_\ell \in \mathbb{R}^{n\times d}$ the matrix of stacked row $\boldsymbol{x}_\ell \in \mathbb{R}^d$, and $x^\mu_{\ell i}$ the $\mu, i-$th element thereof. Above, $V^t_{\mu\to i} \in \mathbb{R}^{rL\times rL}$ is viewed as a matrix and $\omega^t_{\mu\to i} \in \mathbb{R}^{Lr}$ is viewed as a block vector of size $L \times r$, so that $(\omega^t_{\mu\to i})_\ell \in \mathbb{R}^r$ for $1 \le \ell \le L$. We also introduced the resolvent

$$\mathrm{prox}(y, \omega, V, \Gamma, c) \equiv \underset{X\in\mathbb{R}^{Lr}}{\mathrm{arginf}}\left\{\frac{1}{2}(X-\omega)V^{-1}(X-\omega) + \ell(y, X, \Gamma, c)\right\}. \tag{57}$$

Finally, we remind that $y_\mu = {}^{x^\mu w_\star}/\sqrt{d} \in \mathbb{R}^{L\times t}$. The rBP iterations can be simplified into the GAMP equations

Algorithm 2 GAMP

```
**Inputs** : $\{X_\ell \in \mathbb{R}^{n\times d}\}_{\ell=1}^{L}, \boldsymbol{y} \in \mathbb{R}^{n\times L\times t}$
**Initialize** $\forall 1 \le \mu \le n,\ \ 1 \le i \le d,\ \ \hat{w}_i^0 = 0_r, \hat{c}_i^0 = \mathbb{I}_r, \{f_{\ell\mu}^0 = 0_r\}_{\ell=1}^{L}$
**for** $t \le t_{\max}$ **do**
    $\forall 1 \le \ell, \kappa \le L, 1 \le \mu \le n,\ \ (V_\mu^t)_{\ell\kappa} = \frac{1}{d}\sum_i (x_{\ell i}^\mu)(x_{\kappa i}^\mu)\hat{c}_i^t$
    $\Gamma^t = \frac{1}{d}\sum_i \hat{w}_i^t(\hat{w}_i^t)^\top$
    $\forall 1 \le \ell, 1 \le \mu \le n,\ \ \omega_{\ell,\mu}^t = \frac{1}{\sqrt{d}}\sum_i x_{\ell,i}^\mu \hat{w}_i^t - \sum_\kappa (V_\mu^t)_{\ell\kappa} f_{\kappa\mu}$
    $\forall 1 \le \ell, 1 \le \mu \le n,\ f_{\ell,\mu}^t = \left[(V_\mu^t)^{-1}\left(\mathrm{prox}(y_\mu, \omega_\mu^t, V_\mu^t, \Gamma^t, c^\mu) - \omega_{\mu\to i}^t\right)\right]_\ell$
    $\forall 1 \le \mu \le n,\ \ \eta_\mu^t = \partial_3 \ell\left(y_\mu, \mathrm{prox}(y_\mu, \omega_\mu^t, V_\mu^t, \Gamma^t, c^\mu), \Gamma_\mu^t, c^\mu\right)$
    $\forall 1 \le \ell, \kappa \le L, 1 \le \mu \le n,\ \ g_\mu^t = \nabla_\omega f_\mu^t$
    $1 \le i \le d,\ \ A_i^t = -\frac{1}{d}\sum_{\ell,\kappa=1}^{L}\sum_\mu (x_{\ell i}^\mu)(x_{\kappa i}^\mu) g_{\ell\kappa,\mu}^t$
    $C^t = \frac{2}{d}\sum_\mu \eta_\mu^t$
    $1 \le i \le d,\ \ b_i^t = \frac{1}{\sqrt{d}}\sum_{\ell=1}^{L}\sum_\mu x_{\ell i}^\mu f_{\ell,\mu}^t + A_i^t \hat{w}_i^t$
    $\forall 1 \le \mu \le n, 1 \le i \le d,\ \ \hat{w}_i^{t+1} = (\lambda \mathbb{I}_r + C^t + A_i^t)^{-1} b_i^t$
    $\forall 1 \le \mu \le n, 1 \le i \le d,\ \ \hat{c}_i^{t+1} = (\lambda \mathbb{I}_r + C^t + A_i^t)^{-1}$
**end for**

**return** Estimator $\hat{\boldsymbol{w}}$
```

The rBP (1) and GAMP (2) algorithms are in fact asymptotically equivalent, see e.g. (Zdeborová et al., 2016) for an overview. In the next paragraphs, we will show that the equations (50) describe the fixed points of the rBP and GAMP algorithms.

1.2.4.B STATE EVOLUTION

In this section we show that the dynamics of GAMP (2) can be fully tracked by the same summary statistics appearing in the replica SP equations (50). In particular, the equations (50) describe the statistics of the GAMP fixed points. To see this, it is convenient to rather take as a starting point the equivalent rBP equations (1). In the following, we examine each of the variables $V_{\mu\to i}^t$, $\omega_{\mu\to i}^t$, $f_{\mu\to i}^t$, $g_{\mu\to i}^t$, $A_{i\to\mu}^t$, $b_{i\to\mu}^t$, $\hat{w}_{i\to\mu}^t$, $\hat{c}_{i\to\mu}^t$ involved in the rBP iterations, and ascertain their probability distribution. As a convention, we note $\cdot_\mu$ the version of a variable $\cdot_{\mu\to i}$ where the summation also encompasses the index i, and $\cdot_i$ the version of a variable $\cdot_{i\to\mu}$ where the summation also encompasses the index μ. Note that in all cases above the two variables $\cdot_\mu, \cdot_{\mu\to i}$ or $\cdot_i, \cdot_{i\to\mu}$ differ by at most $\Theta_d({}^1/\sqrt{d})$.

For example, $\boldsymbol{\omega}_{\ell,\mu}^t$ is defined as ${}^1/\sqrt{d}\sum_j x_{\ell j}^\mu \hat{\boldsymbol{w}}_{j\to\mu}$, namely with the index $j = i$ included in the summation.

Concentration of $(V^t_{\mu\to i})_{\ell\kappa}$**,** $(\Gamma^t_{\mu\to i})_{\ell\kappa}$ We first show that the variables $V^t_{\mu\to i}$ concentrate to a deterministic value:

$$\begin{aligned}
(V^t_{\mu\to i})_{\ell\kappa} &= \frac{1}{d}\sum_{j\neq i}(x^\mu_{\ell j})(x^\mu_{\kappa j})\hat{c}^t_{j\to\mu} \\
&= \underbrace{\frac{1}{d}\sum_{j\neq i}(\tilde{x}^\mu_\ell)_j(\tilde{x}^\mu_\kappa)_j\hat{c}^t_{j\to\mu}}_{\delta_{\ell\kappa}\Theta_d(1)+(1-\delta_{\ell\kappa})\Theta_d(1/\sqrt{d})} \\
&\qquad \underbrace{+\frac{1}{d}\sum_{j\neq i}(\tilde{x}^\mu_\ell)_j(\mu_{\kappa,c^\mu_\kappa})_j\hat{c}^t_{j\to\mu}+(\ell\leftrightarrow\kappa)}_{\Theta_d(1/d)} \\
&\qquad \underbrace{+\frac{1}{d}\sum_{j\neq i}(\mu_{\ell,c^\mu_\ell})_j(\mu_{\kappa,c^\mu_\kappa})_j\hat{c}^t_{j\to\mu}}_{\Theta_d(1/d)} \\
&= \delta_{\ell\kappa}\frac{1}{d}\sum_j(\Sigma_{\ell,c_\ell})_{jj}\hat{c}^t_j \equiv V^t_{\ell,c^\mu_\ell},
\end{aligned} \tag{58}$$

where we denoted $\tilde{\boldsymbol{x}}_\ell = \boldsymbol{x}_\ell - \boldsymbol{\mu}_\ell$ the centered data, and we introduced the summary statistic V^t_{ℓ,c_ℓ}. By the same token, $(\Gamma^t_{\mu\to i})_{\ell\kappa}$ concentrates to

$$\Gamma^t_{\mu\to i} = \frac{1}{d}\sum_i \hat{w}^t_i(\hat{w}^t_i)^\top \equiv v^t, \tag{59}$$

where we introduced the summary statistic v^t

Distribution of $\omega^t_{\ell,\mu\to i}$ We now move to examine the probability distribution of $\omega^t_{\ell,\mu\to i}$. Let us first introduce the auxiliary random variable

$$\tilde{y}_{\mu,\ell} = \frac{1}{\sqrt{d}}\sum_i(\tilde{x}^\mu_\ell)_i w^\star_i = y_{\mu\ell} - m^\star_{\ell,c^\mu_\ell}, \tag{60}$$

Further, remark that it is reasonable to expect the random variables $\hat{w}^t_{j\to\mu}$ involved in the sum defining $\omega^t_{\ell,\mu\to i}$ in the rBP updates (1) to be asymptotically weakly correlated – this is in fact a standard assumption in the derivation and analysis of AMP algorithms, see (Zdeborová et al., 2016). As a sum of asymptotically independent variables, $\omega^t_{\ell,\mu\to i}$ is thus Gaussian-distributed according to the CLT. We can now ascertain the joint distribution of $\tilde{y}_{\mu,\ell}, \omega^t_{\ell,\mu\to i}$. These variables have mean

$$\mathbb{E}[\omega^t_{\ell,\mu\to i}] = \frac{\boldsymbol{\mu}^\top_{\ell,c^\mu_\ell}\hat{\boldsymbol{w}}^t}{\sqrt{d}} \equiv m^t_{\ell,c^\mu_\ell}, \tag{61}$$

and respective variance

$$
\begin{aligned}
\mathbb{E}[(\omega^t_{\ell,\mu\to i}-m^t_{\ell,c^\mu_\ell})(\omega^t_{\kappa,\nu\to j}-m^t_{\kappa,c^\mu_\kappa})^\top] &= \delta_{\mu\nu}\delta_{\ell\kappa}\frac{1}{d}\sum_{i,j}\hat{w}^t_i(\Sigma_{\ell,c^\mu_\ell})_{ij}(\hat{w}^t_j)^\top \\
&\equiv \delta_{\mu\nu}\delta_{\ell\kappa}q^t_{\ell,c^\mu_\ell}, && (62)\\
\mathbb{E}[\tilde{y}_{\mu\ell}\tilde{y}^\top_{\nu\kappa}] = \delta_{\mu\nu}\delta_{\ell\kappa}\frac{1}{d}\sum_{i,j}w^\star_i(\Sigma_{\ell,c^\mu_\ell})_{ij}(w^\star_j)^\top &\equiv \delta_{\mu\nu}\delta_{\ell\kappa}\rho_{\ell,c^\mu_\ell}, && (63)\\
\mathbb{E}[(\omega^t_{\kappa,\nu\to j}-m^t_{\kappa,c^\mu_\kappa})\tilde{y}^\top_{\mu\ell}] = \delta_{\mu\nu}\delta_{\ell\kappa}\frac{1}{d}\sum_{i,j}(\Sigma_{\ell,c^\mu_\ell})_{ij}\hat{w}^t_j(w^\star_i)^\top &\equiv \delta_{\mu\nu}\delta_{\ell\kappa}\theta^t_{\ell,c^\mu_\ell}. && (64)
\end{aligned}
$$

We introduced the summary statistics $q^t_{\ell,k},\rho_{\ell,k},\theta^t_{\ell,k},m^t_{\ell,k}$.

Distribution of $b^t_{i\to\mu}$ Let us now study the distribution of $b^t_{i\to\mu}$. Expanding the resolvent inside the summand yields

$$
\begin{aligned}
b^t_{i\to\mu} &= \frac{1}{\sqrt{d}}\sum_\ell\sum_{\nu\neq\mu}(x^\nu_\ell)_i f^t_{\ell,\nu\to i}\\
&= \frac{1}{\sqrt{d}}\sum_\ell\sum_{\nu\neq\mu}((\tilde{x}^\nu_\ell)_i+(\mu_{\ell,c^\nu_\ell})_i)\left(1+{}^1/\!\sqrt{d}\sum_\gamma(\tilde{x}^\nu_\gamma)_i(w^\star_i\cdot\nabla_{y_\gamma})\right)\\
&\qquad\left[(V^t_{\nu\to i})^{-1}\left(\mathrm{prox}(y_{\nu\to i},\omega^t_{\nu\to i},V^t_{\nu\to i},\Gamma^t_{\nu\to i},c_\nu)-\omega^t_{\nu\to i}\right)\right]_\ell\\
&= \frac{1}{\sqrt{d}}\sum_\ell\sum_{\nu\neq\mu}((\tilde{x}^\nu_\ell)_i+(\mu_{\ell,c^\nu_\ell})_i)\left(1+{}^1/\!\sqrt{d}\sum_\gamma(\tilde{x}^\nu_\gamma)_i(w^\star_i\cdot\nabla_{y_\gamma})\right)\\
&\qquad(V^t_{\ell,c^\nu_\ell})^{-1}\left(\mathrm{prox}(y_{\nu\to i},\omega^t_{\nu\to i},V^t_{\nu\to i},\Gamma^t_{\nu\to i},c_\nu)-\omega^t_{\nu\to i}\right)_\ell \qquad (65)
\end{aligned}
$$

We denoted $y_{\mu\to i}\equiv y_\mu - x^\mu_i w^\star_i$, and used in the last line the block-diagonal structure of $V^t_{\nu\to i}$ that follows from (58). As for $\omega^t_{\ell,\mu\to i}$, it follows from the CLT that $b^t_{i\to\mu}$ asymptotically follows a Gaussian distribution with mean

$$
\begin{aligned}
&\mathbb{E}[b^t_{i\to\mu}]\\
&=\sum_\ell\sum_k(\sqrt{d}\mu_{\ell,k})_i\underbrace{\alpha\mathbb{E}_c\delta_{c_\ell,k}\mathbb{E}_{y^c,\Xi^c}(V^t_{\ell,k})^{-1}\left[\mathrm{prox}(y^c,m^t_c+\Xi^c,V^t_c,v^t,c)_\ell-(m^t_{\ell,k}+\Xi^c_\ell)\right]}_{\equiv\hat{m}^t_{\ell,k}}\\
&+\sum_\ell\sum_k(\Sigma_{\ell,k})_{ii}\underbrace{\alpha\mathbb{E}_c\delta_{c_\ell,k}\mathbb{E}_{y^c,\Xi^c}(V^t_{\ell,k})^{-1}\nabla_{y_\ell}\left[\mathrm{prox}(y^c,m^t_c+\Xi^c,V^t_c,v^t,c)_\ell-(m^t_{\ell,k}+\Xi^c_\ell)\right]}_{\equiv\hat{\theta}^t_{\ell,k}}w^\star_i,
\end{aligned}
\qquad (66)
$$

where the expectations bear over $\Xi^c\in\mathbb{R}^{L\times r}$ with colored Gaussian rows $(q^t_{\ell,c_\ell})^{1/2}\xi_\ell$, where $\xi_\ell\sim\mathcal{N}(0_r,\mathbb{I}_r)$, and $y\in\mathbb{R}^{L\times t}$ with rows $y^c_\ell\sim\mathcal{N}(m^\star_{\ell,c_\ell}+(\theta^t_{\ell,c_\ell})^\top(q^t_{\ell,c_\ell})^{-1/2}\xi_\ell,\rho_{\ell,c_\ell}-(\theta^t_{\ell,c_\ell})^\top(q^t_{\ell,c_\ell})^{-1}\theta^t_{\ell,c_\ell})$. We further denoted by $V^t\in$

$\mathbb{R}^{rL\times rL}$ the block-diagonal matrix with blocks V^t_ℓ. The variance of $b^t_{i\to\mu}$ can similarly be evaluated as

$$\begin{aligned}&\mathbb{V}[b^t_i, b^t_j]\\ &= \delta_{ij}\sum_\ell\sum_k(\Sigma_{\ell,k})_{ii}\\ &\underbrace{\alpha\mathbb{E}_c\delta_{c_\ell,k}\mathbb{E}_{y^c,\Xi^c}(V^t_{\ell,k})^{-1}\left[\mathrm{prox}(y^c, m^t_c+\Xi^c, V^t_c, v^t, c)_\ell-(m^t_{\ell,k}+\Xi^c_\ell)\right]^{\otimes 2}(V^t_{\ell,k})^{-1}}_{\equiv\hat{q}^t_{\ell,k}}.\end{aligned} \tag{67}$$

We introduced the summary statistics $\hat{q}^t_{\ell,k}, \hat{m}^t_{\ell,k}, \hat{\theta}^t_{\ell,k}$.

Concentration of $A^t_{i\to\mu}, C^t_{i\to\mu}$ Finally, like $V^t_{\mu\to i}$, $A^t_{i\to\mu}$ concentrates to a deterministic value

$$\begin{aligned}&A^t_{i\to\mu}\\ &=\sum_\ell\sum_k(\Sigma_{\ell,k})_{ii}\underbrace{-\alpha\mathbb{E}_c\delta_{c_\ell,k}\left(\mathbb{E}_{y^c,\Xi^c}(V^t_{\ell,k})^{-1}\nabla_{\omega_\ell}\mathrm{prox}(y^c, m^t_c+\Xi^c, V^t_c, v^t, c)_\ell-1\right)}_{\equiv\hat{V}^t_{\ell,k}}\end{aligned} \tag{68}$$

We introduced the summary statistics $\hat{V}^t_{\ell,c_\ell}$.Similarly, $C^t_{i\to\mu}$ concentrates to

$$C^t_{i\to\mu}=2\alpha\mathbb{E}_c\mathbb{E}_{y^c,\Xi^c}\partial_3\ell\left(y^c, \mathrm{prox}(y^c, m^t_c+\Xi^c, V^t_c, v^t, c), V^t_c\right)\equiv\hat{v}^t \tag{69}$$

All the variables involved in the rBP iterations are thus either Gaussian-distributed or deterministic, making it possible to concisely capture their asymptotic dynamics with a small set of summary statistics $q^t_{\ell,k}$, $\theta^t_{\ell,k}$, $m^t_{\ell,k}$, $V^t_{\ell,k}$, v^t, $\hat{q}^t_{\ell,k}$, $\hat{m}^t_{\ell,k}$, $\hat{\theta}^t_{\ell,k}$, $\hat{V}^t_{\ell,k}$, $\hat{v}^t$. In the following paragraph, we derive the update equations obeyed by these statistics, and show that they coincide with a time-indexed version of the SP equations (50) previously derived from the replica method. In particular, the set of self-consistent equations (50) is satisfied at convergence by the infinite-time iterates $q^\infty_{\ell,k}$, $\theta^\infty_{\ell,k}$, $m^\infty_{\ell,k}$, $V^\infty_{\ell,k}$, v^∞, $\hat{q}^\infty_{\ell,k}$, $\hat{m}^\infty_{\ell,k}$, $\hat{\theta}^\infty_{\ell,k}$, $\hat{V}^\infty_{\ell,k}$, $\hat{v}^\infty$.

Recovering equations (50) Wrapping up, we now massage these equations to recover equations (50) derived from the replica method, as discussed in section 1.2.3.a. Starting from $V^t_{\ell,k}$ (58):

$$\begin{aligned}V^t_{\ell,k}&=\frac{1}{d}\sum_i(\Sigma_{\ell,k})_{ii}\left(\lambda\mathbb{I}_r+\hat{v}^t+\sum_\kappa\sum_j\hat{V}^{t-1}_{\kappa,j}(\Sigma_{\kappa,j})_{ii}\right)^{-1}\\ &=\int d\nu(\gamma,\tau)\gamma_{\ell,k}\left(\lambda\mathbb{I}_r+\hat{v}^t+\sum_\kappa\sum_j\hat{V}^{t-1}_{\kappa,j}\gamma_{\kappa,j}\right)^{-1}.\end{aligned} \tag{70}$$

Next, for v^t (62):

$$\begin{aligned}
v^t = \frac{1}{d}\sum_i &\left(\lambda \mathbb{I}_r + \hat{v}^t + \sum_\kappa \sum_j \hat{V}_{\kappa,j}^{t-1}(\Sigma_{\kappa,j})_{ii}\right)^{-1} \\
&\left[\left(\sum_\kappa \sum_j \sqrt{d}(\mu_{\kappa,j})_i \hat{m}_{\kappa,j}^{t-1} + (\Sigma_{\kappa,j})_{ii}\hat{\theta}_{\kappa,j}^{t-1} w_i^\star\right)^{\otimes 2} + \sum_\kappa \sum_j (\Sigma_{\kappa,j})_{ii}\hat{q}_{\kappa,j}^{t-1}\right] \\
&\left(\lambda \mathbb{I}_r + \hat{v}^t + \sum_\kappa \sum_j \hat{V}_{\kappa,j}^{t-1}(\Sigma_{\kappa,j})_{ii}\right)^{-1} \\
= \int d\nu(\gamma,\tau,\pi) &\left(\lambda \mathbb{I}_r + \hat{v}^t + \sum_\kappa \sum_j \hat{V}_{\kappa,j}^{t-1}\gamma_{\kappa,j}\right)^{-1} \\
&\left[\left(\sum_\kappa \sum_j (\tau_{\kappa,j}\hat{m}_{\kappa,j}^{t-1} + \gamma_{\kappa,j}\hat{\theta}_{\kappa,j}^{t-1}\pi\right)^{\otimes 2} + \sum_\kappa \sum_j \gamma_{\kappa,j}\hat{q}_{\kappa,j}^{t-1}\right] \\
&\left(\lambda \mathbb{I}_r + \hat{v}^t + \sum_\kappa \sum_j \hat{V}_{\kappa,j}^{t-1}\gamma_{\kappa,j}\right)^{-1}.
\end{aligned} \tag{71}$$

Next, for $q_{\ell,k}^t$ (62):

$$\begin{aligned}
q_{\ell,k}^t = \frac{1}{d}\sum_i (\Sigma_{\ell,k})_{ii} &\left(\lambda \mathbb{I}_r + \hat{v}^t + \sum_\kappa \sum_j \hat{V}_{\kappa,j}^{t-1}(\Sigma_{\kappa,j})_{ii}\right)^{-1} \\
&\left[\left(\sum_\kappa \sum_j \sqrt{d}(\mu_{\kappa,j})_i \hat{m}_{\kappa,j}^{t-1} + (\Sigma_{\kappa,j})_{ii}\hat{\theta}_{\kappa,j}^{t-1} w_i^\star\right)^{\otimes 2} + \sum_\kappa \sum_j (\Sigma_{\kappa,j})_{ii}\hat{q}_{\kappa,j}^{t-1}\right] \\
&\left(\lambda \mathbb{I}_r + \hat{v}^t + \sum_\kappa \sum_j \hat{V}_{\kappa,j}^{t-1}(\Sigma_{\kappa,j})_{ii}\right)^{-1} \\
= \int d\nu(\gamma,\tau,\pi)\gamma_{\ell k} &\left(\lambda \mathbb{I}_r + \hat{v}^t + \sum_\kappa \sum_j \hat{V}_{\kappa,j}^{t-1}\gamma_{\kappa,j}\right)^{-1} \\
&\left[\left(\sum_\kappa \sum_j (\tau_{\kappa,j}\hat{m}_{\kappa,j}^{t-1} + \gamma_{\kappa,j}\hat{\theta}_{\kappa,j}^{t-1}\pi\right)^{\otimes 2} + \sum_\kappa \sum_j \gamma_{\kappa,j}\hat{q}_{\kappa,j}^{t-1}\right] \\
&\left(\lambda \mathbb{I}_r + \hat{v}^t + \sum_\kappa \sum_j \hat{V}_{\kappa,j}^{t-1}\gamma_{\kappa,j}\right)^{-1}.
\end{aligned} \tag{72}$$

For $\theta^t_{\ell,k}$(62):

$$\begin{aligned}\theta^t_{\ell,k} &= \frac{1}{d}\sum_i (\Sigma_{\ell,k})_{ii}\left(\lambda\mathbb{I}_r + \hat{v}^t + \sum_\kappa\sum_j \hat{V}^{t-1}_{\kappa,j}\gamma_{\kappa,j}\right)^{-1} \\ &\qquad \left(\sum_\kappa\sum_j (\sqrt{d}(\mu_{\kappa,j})_i\hat{m}^{t-1}_{\kappa,j} + (\Sigma_{\kappa,j})_{ii}\hat{\theta}^{t-1}_{\kappa,j}w^\star_i\right)(w^\star_i)^\top \\ &= \int d\nu(\gamma,\tau,\pi)\gamma_{\ell,k}\left(\lambda\mathbb{I}_r + \hat{v}^t + \sum_\kappa\sum_j \hat{V}^{t-1}_{\kappa,j}\gamma_{\kappa,j}\right)^{-1} \\ &\qquad \left(\sum_\kappa\sum_j (\tau_{\kappa,j}\hat{m}^{t-1}_{\kappa,j} + \gamma_{\kappa,j}\hat{\theta}^{t-1}_{\kappa,j}\pi\right)\pi^\top.\end{aligned} \tag{73}$$

For $m^t_{\ell,k}$ (61):

$$\begin{aligned}m^t_{\ell,k} &= \frac{1}{d}\sum_i (\sqrt{d}\mu_{\ell,k})_i\left(\lambda\mathbb{I}_r + \hat{v}^t + \sum_\kappa\sum_j \hat{V}^{t-1}_{\kappa,j}\gamma_{\kappa,j}\right)^{-1} \\ &\qquad \left(\sum_\kappa\sum_j (\sqrt{d}(\mu_{\kappa,j})_i\hat{m}^{t-1}_{\kappa,j} + (\Sigma_{\kappa,j})_{ii}\hat{\theta}^{t-1}_{\kappa,j}w^\star_i\right) \\ &= \int d\nu(\gamma,\tau,\pi)\tau_{\ell,k}\left(\lambda\mathbb{I}_r + \hat{v}^t + \sum_\kappa\sum_j \hat{V}^{t-1}_{\kappa,j}\gamma_{\kappa,j}\right)^{-1} \\ &\qquad \left(\sum_\kappa\sum_j (\tau_{\kappa,j}\hat{m}^{t-1}_{\kappa,j} + \gamma_{\kappa,j}\hat{\theta}^{t-1}_{\kappa,j}\pi\right).\end{aligned} \tag{74}$$

For $\hat{m}^t_{\ell,k}$ (66):

$$\hat{m}^t_\ell = \alpha\mathbb{E}_c\delta_{c_\ell,k}\mathbb{E}_{y,\Xi}(V^t_{\ell,k})^{-1}\left[\mathrm{prox}^c_\ell - (q^t_{\ell,k})^{1/2}\xi_\ell - m^t_{\ell,k}\right], \tag{75}$$

while for $\hat{\theta}^t_{\ell,k}$ (66):

$$\begin{aligned}\hat{\theta}^t_{\ell,k} &= \alpha\mathbb{E}_c\delta_{c_\ell,k}\mathbb{E}_{y,\Xi}(V^t_{\ell,k})^{-1}\nabla_{y_\ell}\left[\mathrm{prox}^c_\ell - (q^t_{\ell,k})^{1/2}\xi_\ell - m^t_{\ell,k}\right] \\ &= \alpha\mathbb{E}_c\delta_{c_\ell,k}\mathbb{E}_{y,\Xi}(V^t_{\ell,k})^{-1}\left[\mathrm{prox}^c_\ell - (q^t_{\ell,k})^{1/2}\xi_\ell - m^t_{\ell,k}\right] \\ &\qquad \left(y_\ell - m^\star_{\ell,k} - (\theta^t_{\ell,k})^\top(q^t_{\ell,k})^{-1/2}\xi_\ell\right)^\top\left(\rho_{\ell,k} - (\theta^t_{\ell,k})^\top(q^t_{\ell,k})^{-1}\theta^t_{\ell,k}\right)^{-1}.\end{aligned} \tag{76}$$

Now turning to $\hat{q}^t_{\ell,k}$:

$$\hat{q}^t_{\ell,k} = \alpha\mathbb{E}_c\delta_{c_\ell,k}\mathbb{E}_{y,\Xi}\left[(V^t_{\ell,k})^{-1}\left[\mathrm{prox}^c_\ell - (q^t_{\ell,k})^{1/2}\xi_\ell - m^t_{\ell,k}\right]^{\otimes 2}(V^t_{\ell,k})^{-1}\right]. \tag{77}$$

For $\hat{v}^t$:

$$\hat{v}^t = 2\alpha \mathbb{E}_c \mathbb{E}_{y,\Xi} \partial_3 \ell\left(y^c, \mathrm{prox}^c, V_c^t, v^t, c\right) \equiv \hat{v}^t \tag{78}$$

Finally, for $\hat{V}^t_{\ell,k}$ (68):

$$\begin{aligned}
\hat{V}^t_{\ell,k} &= -\alpha \mathbb{E}_c \delta_{c_\ell,k} \mathbb{E}_{y,\Xi} (V^t_{\ell,k})^{-1} \left[\nabla_{\omega_\ell} \mathrm{prox}^c_\ell - 1\right] \\
&= -\alpha \mathbb{E}_c \delta_{c_\ell,k} \mathbb{E}_{y,\Xi} (V^t_{\ell,k})^{-1} \left[\nabla_{\xi_\ell} (\mathrm{prox}^c_\ell - (q^t_{\ell,k})^{1/2} \xi_\ell - m^t_{\ell,k}) (q^t_{\ell,k})^{-1/2}\right] \\
&= \alpha \mathbb{E}_c \delta_{c_\ell,k} \mathbb{E}_{y,\Xi} \left[(V^t_{\ell,k})^{-1} (\mathrm{prox}^c_\ell - (q^t_{\ell,k})^{1/2} \xi_\ell - m^t_{\ell,k}) . \right. \\
&\quad \cdot \qquad \left[\left(y_\ell - m^\star_{\ell,k} - (\theta^t_{\ell,k})^\top (q^t_{\ell,k})^{-1/2} \xi_{\ell,k}\right)^\top \right. \\
&\qquad\qquad \left. \left. \left(\rho_{\ell,k} - (\theta^t_{\ell,k})^\top (q^t_{\ell,k})^{-1} \theta^t_{\ell,k}\right)^{-1} (\theta^t_{\ell,k})^\top (q^t_{\ell,k})^{-1/2} - \xi^\top_{\ell,k} \right] (q^t_{\ell,k})^{-1/2} \right] \\
&= \hat{\theta}^t_{\ell,k} (\theta^t_{\ell,k})^\top (q^t_{\ell,k})^{-1} \\
&\quad - \alpha \mathbb{E}_c \delta_{c_\ell,k} \mathbb{E}_{y,\Xi} (V^t_{\ell,k})^{-1} (\mathrm{prox}^c_\ell - (q^t_{\ell,k})^{1/2} \xi_\ell - m^t_{\ell,k}) \xi^\top_\ell (q^t_{\ell,k})^{-1/2}.
\end{aligned} \tag{79}$$

This concludes the derivation of the update equations satisfied by the set of summary statistics $q^t_{\ell,k}, \theta^t_{\ell,k}, m^t_{\ell,k}, V^t_{\ell,k}, \hat{q}^t_{\ell,k}, \hat{m}^t_{\ell,k}, \hat{\theta}^t_{\ell,k}, \hat{V}^t_{\ell,k}$, called the SE equations. These equations concisely describe the macroscopic asymptotic behaviour of the rBP (equivalently GAMP) iterates, thus abstracting away the precise dynamics of the $\Theta_d(d^2)$ variables $V^t_{\mu\to i}, \omega^t_{\mu\to i}, f^t_{\mu\to i}, g^t_{\mu\to i}, A^t_{i\to\mu}, b^t_{i\to\mu}, \hat{w}^t_{i\to\mu}, \hat{c}^t_{i\to\mu}$. This reductionist viewpoint is, much like the replica approach of section 1.2.3.a, very characteristic of statistical physics.

Summary : State evolution equations We now regroup the State Evolution (SE) equations derived in the previous paragraph:

$$\begin{cases} V^t_{\ell,k} = \int d\nu(\gamma,\tau)\gamma_{\ell,k}\left(\lambda\mathbb{I}_r + \hat{v}^t + \sum_\kappa\sum_j \hat{V}^{t-1}_{\kappa,j}\gamma_{\kappa,j}\right)^{-1} \\ q^t_{\ell,k} = \int d\nu(\gamma,\tau,\pi)\gamma_{\ell,k}\left(\lambda\mathbb{I}_r + \hat{v}^t + \sum_\kappa\sum_j \hat{V}^{t-1}_{\kappa,j}\gamma_{\kappa,j}\right)^{-1} \\ \qquad \left[\left(\sum_\kappa\sum_j(\tau_{\kappa,j}\hat{m}^{t-1}_{\kappa,j} + \gamma_{\kappa,j}\hat{\theta}^{t-1}_{\kappa,j}\pi\right)^{\otimes 2} + \sum_\kappa\sum_j\gamma_{\kappa,j}\hat{q}^{t-1}_{\kappa,j}\right] \\ \qquad \left(\lambda\mathbb{I}_r + \hat{v}^t + \sum_\kappa\sum_j \hat{V}^{t-1}_{\kappa,j}\gamma_{\kappa,j}\right)^{-1} \\ \theta^t_{\ell,k} = \int d\nu(\gamma,\tau,\pi)\gamma_{\ell,k}\left(\lambda\mathbb{I}_r + \hat{v}^t + \sum_\kappa\sum_j \hat{V}^{t-1}_{\kappa,j}\gamma_{\kappa,j}\right)^{-1} \\ \qquad \left(\sum_\kappa\sum_j(\tau_{\kappa,j}\hat{m}^{t-1}_{\kappa,j} + \gamma_{\kappa,j}\hat{\theta}^{t-1}_{\kappa,j}\pi\right)\pi^\top \\ m^t_{\ell,k} = \int d\nu(\gamma,\tau,\pi)\tau_{\ell,k}\left(\lambda\mathbb{I}_r + \hat{v}^t + \sum_\kappa\sum_j \hat{V}^{t-1}_{\kappa,j}\gamma_{\kappa,j}\right)^{-1} \\ \qquad \left(\sum_\kappa\sum_j(\tau_{\kappa,j}\hat{m}^{t-1}_{\kappa,j} + \gamma_{\kappa,j}\hat{\theta}^{t-1}_{\kappa,j}\pi\right) \\ v^t = \int d\nu(\gamma,\tau,\pi)\left(\lambda\mathbb{I}_r + \hat{v}^t + \sum_\kappa\sum_j \hat{V}^{t-1}_{\kappa,j}\gamma_{\kappa,j}\right)^{-1} \\ \qquad \left[\left(\sum_\kappa\sum_j(\tau_{\kappa,j}\hat{m}^{t-1}_{\kappa,j} + \gamma_{\kappa,j}\hat{\theta}^{t-1}_{\kappa,j}\pi\right)^{\otimes 2} + \sum_\kappa\sum_j\gamma_{\kappa,j}\hat{q}^{t-1}_{\kappa,j}\right] \\ \qquad \left(\lambda\mathbb{I}_r + \hat{v}^t + \sum_\kappa\sum_j \hat{V}^{t-1}_{\kappa,j}\gamma_{\kappa,j}\right)^{-1} \end{cases} \tag{80}$$

$$\begin{cases} \hat{V}^t_{\ell,k} = \hat{\theta}^t_{\ell,k}(\theta^t_{\ell,k})^\top(q^t_{\ell,k})^{-1} \\ \qquad -\alpha\mathbb{E}_c\delta_{c_\ell,k}\mathbb{E}_{y,\Xi}(V^t_{\ell,k})^{-1}(\mathrm{prox}^c_\ell - (q^t_{\ell,k})^{1/2}\xi_\ell - m^t_{\ell,k})\xi_\ell^\top(q^t_{\ell,k})^{-1/2} \\ \hat{q}^t_{\ell,k} = \alpha\mathbb{E}_c\delta_{c_\ell,k}\mathbb{E}_{y,\Xi}\left[(V^t_{\ell,k})^{-1}\left[\mathrm{prox}^c_\ell - (q^t_{\ell,k})^{1/2}\xi_\ell - m^t_{\ell,k}\right]^{\otimes 2}(V^t_{\ell,k})^{-1}\right] \\ \hat{\theta}^t_{\ell,k} = \alpha\mathbb{E}_c\delta_{c_\ell,k}\mathbb{E}_{y,\Xi}(V^t_{\ell,k})^{-1}\left[\mathrm{prox}^c_\ell - (q^t_{\ell,k})^{1/2}\xi_\ell - m^t_{\ell,k}\right] \\ \qquad \left(y_\ell - m^\star_{\ell,k} - (\theta^t_{\ell,k})^\top(q^t_{\ell,k})^{-1/2}\xi_\ell\right)^\top\left(\rho_{\ell,k} - (\theta^t_{\ell,k})^\top(q^t_{\ell,k})^{-1}\theta^t_{\ell,k}\right)^{-1} \\ \hat{m}^t_{\ell,k} = \alpha\mathbb{E}_c\delta_{c_\ell,k}\mathbb{E}_{y,\Xi}(V^t_{\ell,k})^{-1}\left[\mathrm{prox}^c_\ell - (q^t_{\ell,k})^{1/2}\xi_\ell - m^t_{\ell,k}\right] \\ \hat{v}^t = 2\alpha\mathbb{E}_c\mathbb{E}_{y,\Xi}\partial_3\ell\left(y^c, \mathrm{prox}^c, V^t_c, v^t, c\right) \end{cases} \tag{81}$$

which exactly recovers the replica SP equations (50) derived in section 1.2.3.a, with the difference that in the dynamical SE equations (80), the summary statistics bear time indices. This subsection has thus established that the equations (50) describe the summary statistics capturing the dynamics of GAMP iterations (2), provided the relevant time indices are included. In particular, the equations (50) – without time indices – describe the fixed points of GAMP. The next subsection finally shows that critical (zero-gradient) points of the empirical landscape (11), i.e. fixed points of GD, also correspond to fixed points of GAMP, and are thus solutions of the replica SP equations (50).

1.2.4.c FIXED POINTS OF GAMP ARE FIXED POINTS OF GD

In this subsection, we show that critical (zero gradient) points of the empirical ERM landscape (11) coincide with fixed points of GAMP (2), as asymptotically described by the replica equations (50) derived in section 1.2.3.a. Let us first observe that the zero-gradient condition on the empirical loss $\partial_{w_{ik}}\mathscr{R}(\boldsymbol{w}) = 0$ can be expounded as

$$\sum_{\mu=1}^{n}\sum_{\ell=1}^{L} x^{\mu}_{\ell i}\partial_{z_{\ell k}}\ell(y_\mu, z_\mu, \Gamma, c_\mu) + \frac{2}{d}\sum_{\mu=1}^{n}\sum_{j=1}^{r} w_{ij}\partial_{\Gamma_{ij}}\ell(y_\mu, z_\mu, \Gamma, c_\mu) + \lambda w_{ik} = 0. \tag{82}$$

In (82), we directly considered the case of a ℓ_2 regularizer, and denoted $z^\mu \equiv \boldsymbol{x}^\mu \boldsymbol{w}/\sqrt{d}$, which we shall conveniently view as a vector in $\mathbb{R}^{Lr}$. Finally, we remind the notation $y_\mu = \boldsymbol{x}^\mu \boldsymbol{w}_\star/\sqrt{d}$. Let us introduce a family $\{V^{\mu}_{\ell\kappa}\}_{\mu,\ell,\kappa}$ of $\mathbb{R}^{r\times r}$ positive definite matrices, and $V_\mu \in \mathbb{R}^{Lr\times Lr}$ the block-diagonal with blocks $\{V^{\mu}_{\ell\kappa}\}_{\ell,\kappa}$. We are now in a position to introduce the variable $\omega_\mu \in \mathbb{R}^{Lr}$ as

$$\omega_{\mu,\ell} = V_\mu \nabla_z \ell(y_\mu, z_\mu, \Gamma, c_\mu) + z_\mu. \tag{83}$$

The relationship between ω_μ, z^μ can be equivalently rewritten with a resolvent as

$$\begin{aligned} z_\mu &= \underset{x\in\mathbb{R}^{Lr}}{\operatorname{argmin}}\left[\frac{1}{2}(x-\omega_\mu)^\top V_\mu^{-1}(x-\omega_\mu) + \ell(y_\mu, x, \Gamma, c_\mu)\right] \\ &\equiv \operatorname{prox}(y_\mu, \omega_\mu, V_\mu, \Gamma, c_\mu). \end{aligned} \tag{84}$$

These manipulations let us introduce the variables ω_μ, V_μ, which as we later show will match the corresponding variables in the GAMP algorithm 2. Let us now choose a family positive definite symmetric matrix $\{A \in \mathbb{R}^{r\times r}\}_{i=1}^{d}$, and introduce the variables $b_i \in \mathbb{R}^r$ as

$$b_i = \lambda w_i + C w_i + A_i w_i, \tag{85}$$

or equivalently

$$w_i = (\lambda \mathbb{I}_r + C + A_i)^{-1} b_i. \tag{86}$$

We defined

$$C = \frac{2}{d}\sum_{\mu=1}^{n} \partial_\Gamma \ell(y_\mu, z_\mu, \Gamma, c_\mu). \tag{87}$$

The zero gradient condition (82) and the definition of the variable z_μ can then be rewritten as the system of equations

$$\begin{cases} \sum\limits_{\mu=1}^{n}\sum\limits_{\ell=1}^{L} x^{\mu}_{\ell i}\left[V_\mu^{-1}\left(\omega_\mu - \mathrm{prox}(y_\mu,\omega_\mu,V_\mu,\Gamma,c_\mu)\right)\right]_\ell \\ \qquad +\lambda(\lambda\mathbb{I}_r + C + A_i)^{-1} b_i, \\ \mathrm{prox}(y_\mu,\omega_\mu,V_\mu,\Gamma,c_\mu)_\ell = \sum\limits_{i=1}^{d} x^{\mu}_{\ell i}(\lambda\mathbb{I}_r + C + A_i)^{-1} b_i. \end{cases} \tag{88}$$

Let us introduce the variables $f_\mu \in \mathbb{R}^{Lr}$ and $\hat{w}_i \in \mathbb{R}^r$ as

$$f_{\ell,\mu} = \left[V_\mu^{-1}\left(\mathrm{prox}(y_\mu,\omega_\mu,V_\mu,\Gamma,c_\mu) - \omega_\mu\right)\right]_\ell, \quad \hat{w}_i = (\lambda\mathbb{I}_r + C + A_i)^{-1} b_i. \tag{89}$$

The equations (88) can then be rewritten as

$$\begin{cases} \sum\limits_{\mu=1}^{n}\sum\limits_{\ell=1}^{L} x^{\mu}_{\ell i} f_{\mu,\ell} = b_i - A_i \hat{w}_i, \\ \omega_\mu = \sum\limits_{i=1}^{d} x^{\mu}_{\ell i}\hat{w}_i - \sum\limits_{\kappa=1}^{L} (V_\mu)_{\ell\kappa} f_{\mu,\kappa}. \end{cases} \tag{90}$$

which correspond to the fixed-point equations of GAMP (Algorithm 2). Thus, critical points of the empirical landscape (11) are also fixed points of GAMP. To summarize, we have shown that equations (50) describe the zero-gradient points of the ERM landscape (11), i.e. fixed points of GD.

1.2.5 SUMMARY

This section detailed the asymptotic analysis of the simple, yet rather general, example of learning with a seq-GLM in a T-S setting. In subsection 1.2.3.a, we demonstrated how, using the replica method from statistical physics, a sharp characterization of the minimizer of the ERM could be reached in terms of a sufficient set of finite-dimensional summary statistics. These statistics are solutions of a system of self-consistent SP equations (50). The analysis thus importantly enables the reduction of the original high-dimensional optimization problem into a set of equations in *finite* dimensions. Subsection 1.2.4.b then provided an algorithmic viewpoint on the SP equations as the fixed point conditions of a GAMP algorithm. Furthermore, critical points of the ERM landscape – i.e. fixed points of GD – were further shown coincide with fixed points of GAMP, implying that the set of solutions of the SP equations is also descriptive of such critical points. The study of the set of finite-dimensional SP equations thus offers an informative and insightful perspective on the ERM landscape, and affords a particularly powerful framework to analyze ML learning tasks in high dimensions. Similar techniques as those illustrated in this section underlie most of the analyses presented in this thesis.

2

PERSPECTIVES

This chapter, the last of this first introductory part, gathers an overview of recent progress – including those reported in this thesis – and envisioned future directions in the field of statistical physics of high-dimensional ML. Its purpose is to propose and delineate future research axes beyond current knowledge. We however deliberately choose to place it here, at the beginning of the thesis, rather than as a conclusive chapter, so that the reader may, as they read the thesis, understand the reported results in the light of this broader perspective.

A USER GUIDE TO SOLVABLE HIGH DIMENSIONAL MODELS

Statistical physics of ML, as a field, makes use of the study of *exactly solvable models* – namely simplified models capturing the essential aspects of a ML task while remaining amenable to analysis– as a gateway towards better theoretical comprehension of ML empirics. The aim of such a line of research is arguably twofold:

Purpose: to reach a better understanding of random non-convex high-dimensional optimizations that naturally arise in ML settings, *as mathematical problems*. This aspect is of primal and theoretical interest.

Ambition: to construct an effective theory of ML *descriptive* and predictive of *real, practical* ML tasks, with the long-reaching ambition of constituting an actionable theory able to guide ML practice.

These two goals in fact align to a large extent, in the sense that pursuing either warrants the same thing – developing more complex, and realistic, models. In the following, we highlight three important levers that can be actioned to construct such models, which we believe constitute possible axes for the development of the research effort in the field. We detail how the work reported in the thesis fits in each of these axes, and further discuss possible future steps.

2.1 THREE LEVERS

2.1.1 THE FIRST LEVER: MODELS

Broadly construed, DL theory could be defined as the study of parametric families of feature extractors, and more particularly ANNs. In an asymptotic

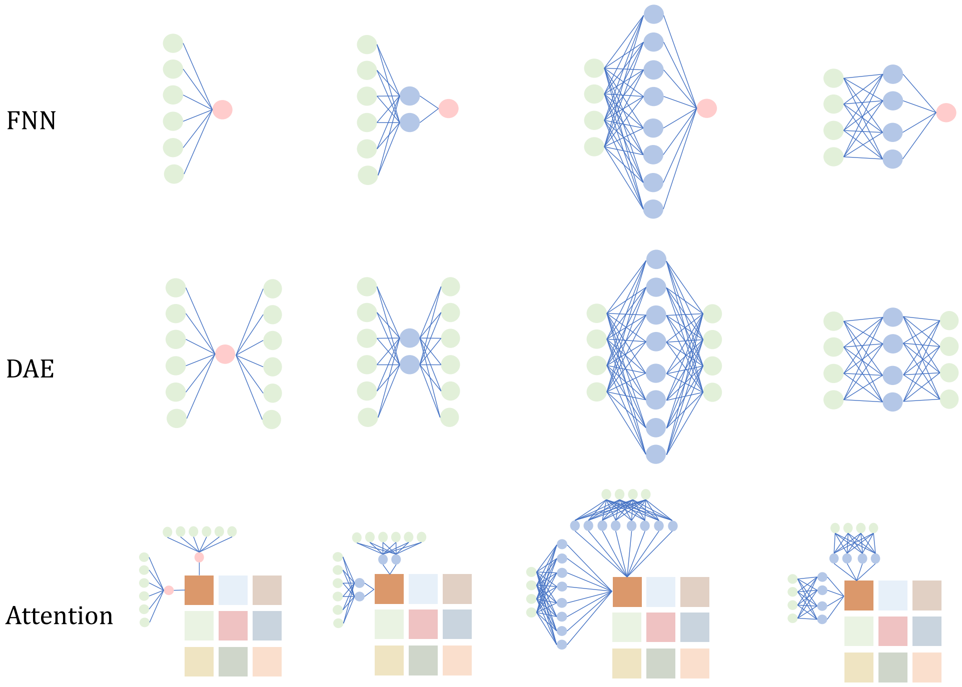


Figure 2: Graphical representation of some existing or possible models in asymptotic ML theories for FNNs (top), AEs (middle), or attention mechanisms (bottom). Each row corresponds to a different asymptotic limit for the ANN architecture. From left to right: single hidden unit models, models with a finite number of hidden units, infinite-width models, and extensive-width models.

theory, an ANN model is specified by the relative scaling of its width p with respect to the input dimension d and the number of training samples n. For a given family of models, a whole spectrum of different asymptotic limits can thus be studied, yielding distinct theoretical models with diverse phenomenology. Some examples are represented in Fig 2, with each row corresponding to a family of ANN –FNNs, AEs, attention mechanisms– and each column to a different scaling limit – $p = 0$, $p = \Theta_d(1)$, $p \gg d$, $p = \Theta_d(d)$–. The plurality of asymptotic limits of interest for a given type of ANN is perhaps most apparent for FNNs, which have been the object of sustained theoretical scrutiny over the past decades. The learning of simple models with no hidden units (GLMs) or a finite number thereof $p = \Theta_d(1)$ has been well characterized in a series of works (Gardner et al., 1988; Gardner et al., 1989; Györgyi et al., 1990; Seung et al., 1992; Sompolinsky et al., 1990; Barbier et al., 2019b; Aubin et al., 2020a; Aubin et al., 2018b; Schwarze, 1993; Saad et al., 1995) (see (Gabrié, 2019; Zdeborová et al., 2016) for reviews). Recent years have further witnessed a realization by the community that, on the opposite end of the spectrum, infinite-width networks are also amenable to relatively easy analytical characterization due to their connection to kernel methods (Jacot et al., 2018a; Chizat et al., 2019a; Geiger et al., 2019; Mei et al., 2019a). Part III of the present thesis rather puts emphasis on the intermediate *extensive-width limit* $p = \Theta_d(d)$ (rightmost in Fig. 2) – a regime which should allow to probe the learning of overparametrized models, while not reducing to some kernel limit. The extension of the zoology of existing limits represented in Fig. 2 along its horizontal axis, namely studying more scaling regimes, is a natural first actionable lever toward building more complex and expressive models of ANNs. With a large number of finite width studies set in the so-called *proportional regime* $n = \Theta_d(d)$, perhaps one salient aspect of such an endeavour should consist in considering other scalings of the number of samples n, in particular polynomial scalings $n = \Theta_d(d^k)$ with $k > 1$, as initiated e.g. in (Xiao et al., 2022) in the infinite-width case. Considering more data-intensive regimes should open the door to richer learning regimes, as exmplified in Chapter 8.

Naturally, other scalings of p, d are of equal theoretical interest, as considered e.g. in (Camilli et al., 2023).

To broaden the family of asymptotically solvable models, the zoology of Fig. 2 should also be extended its *vertical* direction – by diversifying the types of analyzable ANN families. While supervised tasks with FNN architectures have hitherto been the primary focal point of ML theory, the fast pace of recent practical breakthroughs in DL has arguably shifted the emphasis to considering other parametric families, such as denoiser-type networks or attention mechanisms. This constitutes the second direction of this first axis, and should be concurrently pursued. Part IV of the present thesis extends the study of narrow architectures $p = \Theta_d(1)$ to AE and attention models. Again, as for FNNs, a whole spectrum of asymptotic limits can be considered for a given model. Let us cite for illustration the work of (Nguyen, 2021), which initiates the study of infinite-width limits in AEs. In fact, the bottom right sector of Fig. 2 is arguably a still largely uncharted research territory,

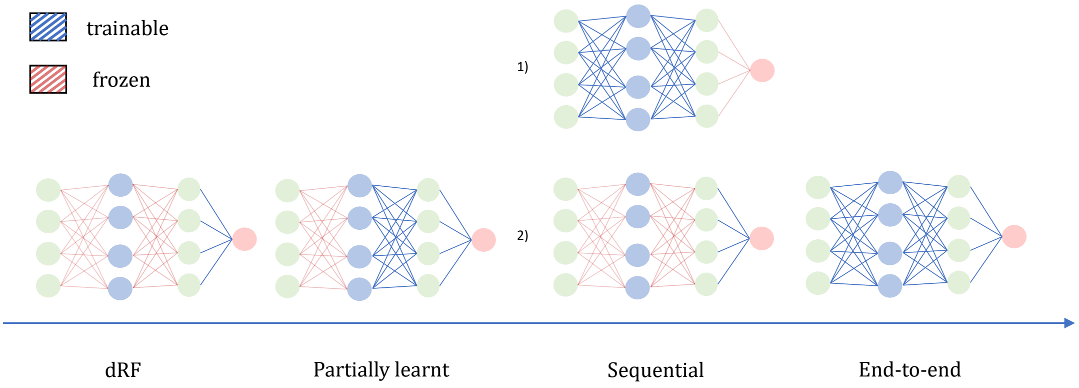


Figure 3: For a given model, different training protocols may be considered, allowing for different learning phenomenologies. In particular, some layers of the networks may be frozen and left untrained. From left to right : dRF; partially trained ANN considered e.g. in (Chen et al., 2022b); two-step training, covered e.g. in Chapter 9; full end-to-end training.

and populating it with exactly solvable model should be an important step towards the construction of a theory truly representative of the diversity of modern ML practice.

Finally, ANNs are by design *modular* in nature. The models of Fig. 2 can be combined into composite models. The most natural such composition consists in stacking them into *multilayer, deep* architectures. Gaining a firm analytical grasp on the learning of such deep models is a crucial pursuit of DL theory. Some key theoretical insights for FNNs such as the mean-field training limit (Rotskoff et al., 2022; Mei et al., 2018; Chizat et al., 2018) are hitherto limited to shallow architectures. Chapters 6, 7 and 8 inscribe themselves in this endeavour of pushing the theoretical understanding to deeper architectures.

In the related context of statistical inference, deep random ANN priors have been considered in e.g. (Gabrié et al., 2018a; Manoel et al., 2017a).

2.1.2 THE SECOND LEVER: TRAINING PROCEDURES

While the extensive width limit (rightmost column of Fig. 2) represents a promising direction to model finite-width, yet expressive FNNs, the analysis of fully trained networks in this regime is a notoriously challenging open question. In the face of this hurdle, an alternative research route consists in considering simplified partial training procedures, and incrementally nudging the theoretical understanding forward towards eventually analyzing the complete end-to-end training. This constitutes the second actionable lever in the construction of realistic ANN models. Fig. 3 gathers some examples of simplified training procedures. One popular simplification consists in allowing only for a trainable readout layer, while freezing the intermediate

On a technical level, the hardness of the problem lies in the fact that there are $\Theta_d(d^2)$ learnable parameters (see e.g. (Maillard et al., 2022) for a discussion), while the standard theoretical toolbox of statistical physics (see e.g. section 1.2) can, so far, only address $\Theta_d(d)$ trainable parameters.

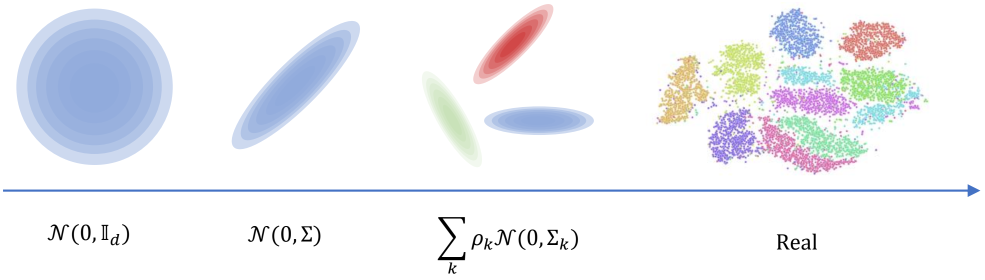

$\mathcal{N}(0,\mathbb{I}_d)$
$\mathcal{N}(0,\Sigma)$
$\sum_k \rho_k \mathcal{N}(0,\Sigma_k)$
Real

Figure 4: Towards more realistic data distributions. From left to right, unimodal isotropic Gaussian density, colored Gaussian density, colored Gaussian mixture density, and t-distributed Stochastic Neighbour Embedding (tSNE) (Van der Maaten et al., 2008) visualization of the MNIST train set (LeCun et al., 1998a).

weights at initialization (corresponding to dRF models (Rahimi et al., 2007b), see Chapter 6), or after some amount of training (see e.g. (Ba et al., 2022a; Damian et al., 2022; Abbe et al., 2023; Berthier et al., 2023) and Chapters 7, 9). These sequential training protocols allow to consider the training of the readout and intermediate weights in isolation, and thus ease the technical study thereof. Future works in this direction should aim either at (a) reducing the number of frozen weights, in an effort to tend towards a fully trainable model (rightmost in Fig. 3) or (b) training the weights as much as possible before freezing. An example of a step in the direction of (b) consists in freezing the intermediate weights only after a number of gradient steps have been performed. Chapter 9, which presents an analysis in the case of a single step, would for instance benefit from being extended to cover multiple steps.

A second viable route would be to consider alternative learning paradigms to ERM, such as Bayesian learning. A series of works (Li et al., 2021c; Ariosto et al., 2022a; Zavatone-Veth et al., 2022a), and (Cui et al., 2023a) which is the object of Chapter 8, have studied the proportional, extensive-width limit of Bayesian FNNs. While these work succeed in formulating a theory of fully trained Bayesian FNNs, as evidenced in Chapter 8, such networks are expected to be more expressive in polynomial regimes $n \gg d$, which remain largely uncharted. The exploration thereof constitutes a challenging, yet important, aspect of the second lever.

2.1.3 THE THIRD LEVER: REALISTIC DATA MODELS

In the building of a model of a ML task, the modeling of the data distribution poses a singular challenge. In contrast to ANNs, which are by design mathematical constructs and can thus be formally modelled straightforwardly, it is to a large extent unclear how to model the distribution of real data. What indeed is the probability distribution of e.g. images of cats and dogs? Of natural language? Satisfyingly answering these interrogations is to a large extent an open question in ML theory. As of the end of the 2010s, a sizeable part of the research effort in sharp asymptotic studies of ML methods was in fact set under the assumption of isotropic, unimodal Gaussian data (Aubin et al., 2018a; Aubin et al., 2020a; Barbier et al., 2019b; Zdeborová et al., 2016), or random orthogonal design (Kabashima, 2008; Shinzato et al., 2008a; Shinzato et al., 2008b). Overcoming those somewhat restrictive assumptions to consider colored (or bimodal) Gaussian densities was an endeavour initiated in e.g. (Mignacco et al., 2020a; Goldt et al., 2020a; D'Ascoli et al., 2021a). Further morphing these simple densities into more *realistic* distributions, along the steps represented in Fig. 4, is an essential step towards building *realistic* models which can ultimately be descriptive of day-to-day ML practice. This constitutes the third, and last, actionable lever.

A natural route is to consider simple data distributions capturing key structural features of the real data distributions. This endeavour is in part carried out in Chapters 3 and 10, which discuss how theoretical characterizations formulated under the assumption of Gaussian (mixture) data densities can in a number of cases capture the learning curves of tasks on *real* data, provided the second-order statistics of the surrogate Gaussian distribution match those of the real dataset. In Chapters 4 and 5, in the context of kernel learning, we further show that in some real settings, the error rates only depend on two scalar structural descriptors, and are thus amenable to full analytical characterization. These observations signal a form of *universality* : namely, only a limited number of structural characteristics of the distribution actually matter for the learning, and can thus be captured by simple, surrogate analytical distributions.

The main theoretical challenge thus lies in ascertaining, for a given model, which statistical or structural descriptors are relevant for the learning. As discussed above, a stream of recent works in the proportional regime $n = \Theta_d(d)$ –including some presented in this thesis– has evidenced empirically (Loureiro et al., 2021b; Goldt et al., 2020a) or rigorously (Hu et al., 2022a; Mei et al., 2022c; Goldt et al., 2020b) that, in some settings, these descriptors simply coincide with the first two moments of the density. In other words, these setups fall under *Gaussian universality.* On the other hand, Chapter 8 shows that DNNs are able to learn more complex statistical features in polynomial regimes $n \gg d$, suggesting that higher-order statistics should become relevant in these limits. Further, even in the proportional regime, works such as (Chung et al., 2018) demonstrate that in some cases the relevant universality class is not Gaussian, but rather based on more complex geometric descriptors. Chapter 9 discusses another instance where simple Gaussian universality does not directly hold in the proportional regime, and needs to be refined. Ascertaining the universality class –if any– of any given model defined by levers 2.1.2 and 2.1.1 thus represents the third, and arguably most fundamental, axis of the research endeavour.

CONCLUSION: ON BUILDING MODELS IN HIGH DIMENSIONS

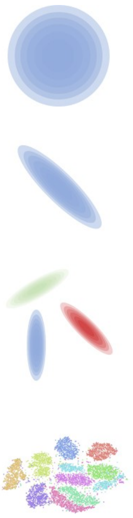
The three axes 2.1.2, 2.1.1 and 2.1.3 outline research directions for a –hopefully– methodological, albeit not systematic, exploration and probing of high-dimensional machine learning with solvable models. The works gathered in this thesis chart some parts of the thus delineated research agenda. The route to a solvable, tight, and ultimately actionable ML theory, however, is long, and extends beyond this manuscript, into exciting future researchscapes.

Part II

DATA STRUCTURE

OUTLINE AND MOTIVATIONS

How to account for the unreasonable effectiveness of DNN in learning from high-dimensional data? The seminal work of (Blum et al., 1988) teaches us that in the *worst case*, learning can be NP hard. The successes of day-to-day DL practice, on the other hand, clearly signal that real data sets do not share this complexity, and must be somehow simpler. In fact, real data distributions are typically *structured*, and the information encoded therein therefore admits a more concise, simpler description. A central challenge in ML theory thus lies in understanding *which structural information encoded in the data, or a representation thereof, is identified and exploited by learning algorithms*. For instance, in the context of computer vision, the informative features are known to lie in a space whose dimensionality is much smaller than that of the image embedding (Hein et al., 2005).



On the other hand, as of the end of the past decade, a sizeable portion of the asymptotic theoretical explorations of ML tasks (Barbier et al., 2019b; Donoho et al., 2011; Maillard et al., 2020a; Aubin et al., 2018b; Gabrié, 2020; Gardner et al., 1989; Watkin et al., 1993; Seung et al., 1992; Hosaka et al., 2002; Mignacco et al., 2020a; Aubin et al., 2020a; Sompolinsky et al., 1990), stemming from the seminal work of (Gardner et al., 1989) in the 80's, shared the assumption of *unstructured data* $\boldsymbol{x}^\mu$, assumed to be i.i.d from a high-dimensional isotropic Gaussian distribution $\mathcal{N}(0, \mathbb{I}_d)$. A sizeable theoretical effort has been devoted in recent years to the study of more structured colored Gaussian (mixture) densities (Mignacco et al., 2020b; D'Ascoli et al., 2021a; Goldt et al., 2020a). As discussed in the introductory Chapter 2, this endeavour needs to be taken one step further to cover *realistic* data distribution, with the long-term objective of tending to a ML theory descriptive of real, practical tasks. The first part of this thesis presents some contributions in this direction.

Considering more complex and realistic models of data structure is key towards building a theory of ML descriptive of practical tasks, see Chapter 2 in Part I.

REALISTIC FEATURES

Which structural or statistical characteristics of data are picked up, learnt, and exploited by ML methods? Chapter 3 first empirically finds that, in a number of cases, when the number of training samples n is comparably large to the dimension of the features $\varphi(\boldsymbol{x})$, the test and train errors largely depend on only *second-order statistics* of the features distribution. Such cases encompass some real datasets, data generated from Generative Adversarial Network (GAN)s, and FNN features. Because of this Gaussian *universality*, these ML tasks display the same learning metrics as an equivalent T-S GCM with matching second-order statistics, and can be precisely asymptotically characterized. Chapter 3 thus offers a powerful and versatile framework to begin to apprehend the effect of *realistic* features structure on the learning, and thus opens the door to building effective asymptotic theories for simple real tasks. The subsequent Chapters 4 and 5 examine one such case of particular interest – namely *kernel* methods.

KERNEL FEATURES

Kernel methods are a mainstay in the ML toolbox, as they allow to leverage expressive non-linear features without leaving the realm of convex optimization. A recent regain in interest has further been fueled by the theoretical realization that kernels also coincide with some infinite-width limits of DNNs (Neal, 1996b; Williams, 1996a; Jacot et al., 2018a; Chizat et al., 2018; Geiger et al., 2019). A central question in the theoretical study of kernels – further motivated by recent theoretical interest in error scaling laws (Hestness et al., 2017; Kaplan et al., 2020; Rosenfeld et al., 2019; Henighan et al., 2020)– is to ascertain the *rate of decay of the test error with the number of training samples*, and further, how this rate depends on the data structure.

In the context of KRR, a long line of work has been devoted to this question, dating back to the seminal works of (Caponnetto et al., 2005; Caponnetto et al., 2007). These works have evidenced that, for a sizeable class of data sets, the error rates could be concisely characterized in terms of only *two* structural descriptors, namely the relative decay of the kernel eigenvalues (a.k.a the *capacity*), and of the target decomposition in the eigenbasis (a.k.a the *source*). As of the end of the past decade however, this body of works was rather disparate, and several expressions for the error decay rates could be found in the literature, with no clear explanation for the discrepancies. Chapter 4 leverages the framework of Chapter 3 to provide a unifying and exhaustive characterization of all observable regimes in KRR, whenever the data structure can be captured by these two descriptors. Importantly, this includes a number of real datasets. Chapter 5 addresses the more complex case of *classification*, and presents the first tight rates for such tasks, again encompassing a number of real setups.

Part II A.

FEATURES STRUCTURE

3 REALISTIC LEARNING CURVES FROM STRUCTURED FEATURES

T-S models are a popular framework to study the high-dimensional asymptotic performance of learning problems with synthetic data, and have been the subject of intense investigations spanning three decades (Seung et al., 1992; Watkin et al., 1993; Engel et al., 2001; Donoho et al., 2009; El Karoui et al., 2013; Zdeborová et al., 2016; Donoho et al., 2016). In the wake of understanding the limitations of classical statistical learning approaches (Zhang et al., 2017; Belkin et al., 2019; Belkin et al., 2020), this direction is witnessing a renewal of interest (Mei et al., 2019b; Hastie et al., 2022; Belkin et al., 2020; Candès et al., 2020; Aubin et al., 2020a; Salehi et al., 2020). However, this framework is often assuming the input data to be Gaussian i.i.d., which is arguably too simplistic to be able to capture properties of realistic data. In this paper, we redeem this line of work by defining a Gaussian covariate model where the teacher and student act on different Gaussian correlated spaces with arbitrary covariance. We derive a rigorous asymptotic solution of this model generalizing the formulas found in the above mentioned classical works.

We then put forward a theory, supported by universality arguments and numerical experiments, that this model captures learning curves, i.e. the dependence of the training and test errors on the number of samples, for a generic class of feature maps applied to realistic datasets. These maps can be deterministic, random, or even learnt from the data. This analysis thus gives a unified framework to describe the learning curves of, for example, kernel regression and classification, the analysis of feature maps – random projections (Rahimi et al., 2008), neural tangent kernels (Jacot et al., 2018a), scattering transforms (Andreux et al., 2020) – as well as the analysis of transfer learning performance on data generated by a GAN (Goodfellow et al., 2014). We also discuss limits of applicability of our results, by showing concrete situations where the learning curves of the Gaussian covariate model differ from the actual ones.

Model definition — The Gaussian covariate T-S model is defined via two vectors $\boldsymbol{u} \in \mathbb{R}^p$ and $\boldsymbol{v} \in \mathbb{R}^d$, with correlation matrices $\Psi \in \mathbb{R}^{p\times p}, \Omega \in \mathbb{R}^{d\times d}$ and $\Phi \in \mathbb{R}^{p\times d}$, from which we draw n independent samples:

$$\begin{bmatrix} \boldsymbol{u}^\mu \\ \boldsymbol{v}^\mu \end{bmatrix} \in \mathbb{R}^{p+d} \underset{\text{i.i.d.}}{\sim} \mathcal{N}\left(0, \begin{bmatrix} \Psi & \Phi \\ \Phi^\top & \Omega \end{bmatrix}\right), \qquad \mu = 1, \cdots, n. \tag{91}$$

The *l*abels y^μ are generated by a **teacher** function that is only using the vectors $\boldsymbol{u}^\mu$:

$$y^\mu = f_0\left(\frac{1}{\sqrt{p}} \boldsymbol{\theta}_0^\top \boldsymbol{u}^\mu\right), \tag{92}$$

where $f_0 : \mathbb{R} \to \mathbb{R}$ is a function that may include randomness such as, for instance, an additive Gaussian noise, and $\boldsymbol{\theta}_0 \in \mathbb{R}^p$ is a vector of teacher-weights with finite norm which can be either random or deterministic. Learning is performed by the **student** with weights $\boldsymbol{w}$ via empirical risk minimization that has access only to the features $\boldsymbol{v}^\mu$:

$$\hat{\boldsymbol{w}} = \operatorname{argmin}_{\boldsymbol{w}\in\mathbb{R}^d} \left[\sum_{\mu=1}^n g\left(\frac{\boldsymbol{w}^\top \boldsymbol{v}^\mu}{\sqrt{d}}, y^\mu\right) + r(\boldsymbol{w})\right], \tag{93}$$

where r and g are proper, convex, lower-semicontinuous functions of $\boldsymbol{w} \in \mathbb{R}^d$ (e.g. g can be a logistic or a square loss and r a ℓ_p $(p{=}1,2)$ regularization). The key quantities we want to compute in this model are the *averaged training and generalisation errors* for the estimator $\boldsymbol{w}$,

$$\mathcal{E}_{\text{train.}}(\boldsymbol{w}) \equiv \frac{1}{n}\sum_{\mu=1}^n g\left(\frac{\boldsymbol{w}^\top \boldsymbol{v}^\mu}{\sqrt{d}}, y^\mu\right) \tag{94}$$

$$\mathcal{E}_{\text{gen.}}(\boldsymbol{w}) \equiv \mathbb{E}\left[\hat{g}\left(\hat{f}\left(\frac{\boldsymbol{v}_{\text{new}}^\top \boldsymbol{w}}{\sqrt{d}}\right), f_0\left(\frac{\boldsymbol{u}_{\text{new}}^\top \boldsymbol{\theta}_0}{\sqrt{p}}\right)\right)\right]. \tag{95}$$

where g is the loss function in eq. (93), $\hat{f}$ is a prediction function (e.g. $\hat{f} = \text{sign}$ for a classification task), $\hat{g}$ is a performance measure (e.g. $\hat{g}(\hat{y}, y) = (\hat{y} - y)^2$ for regression or $\hat{g}(\hat{y}, y) = \mathbb{P}(\hat{y} \neq y)$ for classification) and $(\boldsymbol{u}_{\text{new}}, \boldsymbol{v}_{\text{new}})$ is a fresh sample from the joint distribution of $\boldsymbol{u}$ and $\boldsymbol{v}$.

Our two **main technical contributions** are:

(C1) In Theorems 3.1.1 & 3.1.2, we give a rigorous closed-form characterisation of the properties of the estimator $\hat{\boldsymbol{w}}$ for the Gaussian covariate model (91), and the corresponding training and generalisation errors in the high-dimensional limit. We prove our result using Gaussian comparison inequalities (Gordon, 1985);

(C2) We show how the same expression can be obtained using the replica method from statistical physics (Mézard et al., 1987). This is of additional interest given the wide range of applications of the replica

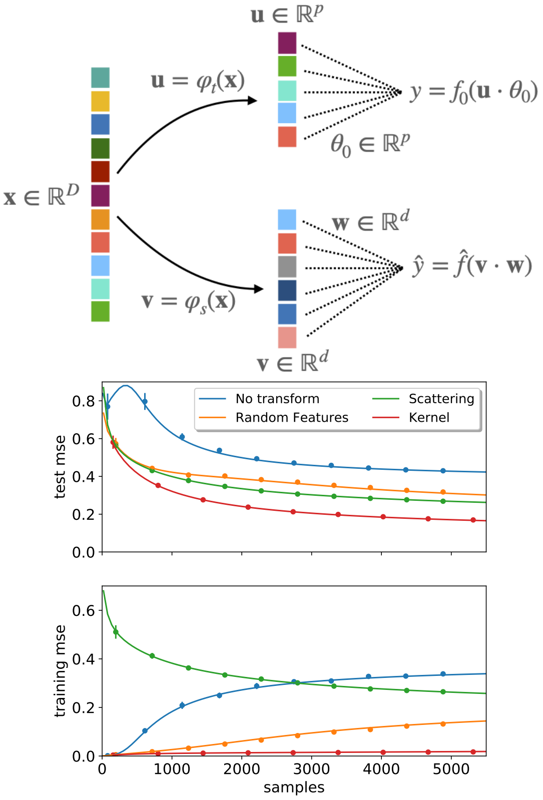


Figure 5: **T**op: Given a data set $\{\boldsymbol{x}^\mu\}_{\mu=1}^n$, teacher $\boldsymbol{u} = \boldsymbol{\varphi}_t(\boldsymbol{x})$ and student maps $\boldsymbol{v} = \boldsymbol{\varphi}_t(\boldsymbol{x})$, we assume $[\boldsymbol{u}, \boldsymbol{v}]$ to be jointly Gaussian random variables and apply the results of the Gaussian covariate model (91). **B**ottom: Illustration on real data, here ridge regression on even vs odd MNIST digits, with regularisation $\lambda = 10^{-2}$. Full line is theory, points are simulations. We show the performance with no feature map (blue), random feature map with $\sigma = \text{erf}$ & Gaussian projection (orange), the scattering transform with parameters $J = 3, L = 8$ (Andreux et al., 2020) (green), and of the limiting kernel of the random map (Williams, 1996b) (red). The covariance Ω is empirically estimated from the full data set, while the other quantities appearing in the Theorem 3.1.1 are expressed directly as a function of the labels, see Section 3.2.4. Simulations are averaged over 10 independent runs.

approach in machine learning and computer science (Mézard et al., 2009). In particular, this allows to put on a rigorous basis many results previously derived with the replica method.

Towards realistic data — In the second part of our paper, we argue that the above Gaussian covariate model (91) is generic enough to capture the learning behaviour of a broad range of realistic data. Let $\{\boldsymbol{x}^\mu\}_{\mu=1}^n$ denote a data set with n independent samples on $\mathscr{X} \subset \mathbb{R}^D$. Based on this input, the **features** $\boldsymbol{u}, \boldsymbol{v}$ are given by (potentially) elaborated transformations of $\boldsymbol{x}$, i.e.

$$\boldsymbol{u} = \boldsymbol{\varphi}_t(\boldsymbol{x}) \in \mathbb{R}^p \quad \text{and} \quad \boldsymbol{v} = \boldsymbol{\varphi}_s(\boldsymbol{x}) \in \mathbb{R}^d \tag{96}$$

for given centred feature maps $\boldsymbol{\varphi}_t : \mathscr{X} \to \mathbb{R}^p$ and $\boldsymbol{\varphi}_s : \mathscr{X} \to \mathbb{R}^d$, see Fig. 5. Uncentered features can be taken into account by shifting the covariances, but we focus on the centred case to lighten notation.

The Gaussian covariate model (91) is exact in the case where $\boldsymbol{x}$ are Gaussian variables and the feature maps $(\boldsymbol{\varphi}_s, \boldsymbol{\varphi}_s)$ preserve the Gaussianity, for example linear features. In particular, this is the case for $\boldsymbol{u}=\boldsymbol{v}=\boldsymbol{x}$, which is the widely-studied vanilla T-S model (Gardner et al., 1989). The interest of the model (91) is that it also captures a range of cases in which the feature maps $\boldsymbol{\varphi}_t$ and $\boldsymbol{\varphi}_s$ are deterministic, or even learnt from the data. The covariance matrices Ψ, Φ, and Ω then represent different aspects of the data-generative process and learning model. The student (93) then corresponds to the last layer of the learning model. These observation can be distilled into the following conjecture:

Conjecture 3.0.1. *(Gaussian equivalent model) For a wide class of data distributions $\{\boldsymbol{x}^\mu\}_{\mu=1}^n$, and features maps $\boldsymbol{u} = \boldsymbol{\varphi}_t(\boldsymbol{x}), \boldsymbol{v} = \boldsymbol{\varphi}_s(\boldsymbol{x})$, the generalisation and training errors of estimator (93) are asymptotically captured by the equivalent Gaussian model (91), where $[\boldsymbol{u}, \boldsymbol{v}]$ are jointly Gaussian variables, and thus by the closed-form expressions of Theorem 3.1.1.*

The second part of our **main contributions** are:

(C3) In Sec. 3.2.3 we show that the theoretical predictions from (C1) captures the learning curves in non-trivial cases, e.g. when input data are generated using a trained generative adversarial network, while extracting both the feature maps from a neural network trained on real data.

(C4) In Sec. 3.2.4, we show empirically that for ridge regression the asymptotic formula of Theorem 3.1.1 can be applied *d*irectly to real data sets, even though the Gaussian hypothesis is not satisfied. This universality-like property is a consequence of Theorem 3.2.1 and is illustrated in Fig. 5 (right) where the real learning curve of several features maps learning the odd-versus-even digit task on MNIST is compared to the theoretical prediction.

3.0.1 RELATED WORK —

Rigorous results for T-S models–: The Gaussian covariate model (91) contains the vanilla T-S model as a special case where one takes $\boldsymbol{u}$ and $\boldsymbol{v}$ *i*dentical, with unique covariance matrix Ω. This special case has been extensively studied in the statistical physics community using the heuristic replica method (Gardner et al., 1989; Opper et al., 1996; Seung et al., 1992; Watkin et al., 1993; Engel et al., 2001). Many recent rigorous results for such models can be rederived as a special case of our formula, e.g. refs. (Mei et al., 2019b; Hastie et al., 2022; Ghorbani et al., 2020b; Belkin et al., 2020; Candès et al., 2020; Thrampoulidis et al., 2018; Montanari et al., 2019; Aubin et al., 2020a; Salehi et al., 2020; Celentano et al., 2020). Numerous of these results are based on the same proof technique as we employed here: the Gordon's

Gaussian min-max inequalities (Gordon, 1985; Stojnic, 2013a; Oymak et al., 2013). The asymptotic analysis of kernel ridge regression (Bordelon et al., 2020), of margin-based classification (Huang et al., 2020) also follow from our theorem. Other examples include models of the double descent phenomenon (Mitra, 2019). Closer to our work is the recent work of (Dhifallah et al., 2020) on the random feature model. For ridge regression, there are also precise predictions thanks to random matrix theory (Dobriban et al., 2018a; Hastie et al., 2022; Wu et al., 2020a; Liao et al., 2020; Liu et al., 2020; Bartlett et al., 2020a; Jacot et al., 2020b). A related set of results was obtained in (Gerbelot et al., 2020) for orthogonal random matrix models. The main technical novelty of our proof is the handling of a generic loss and regularisation, not only ridge, representing convex empirical risk minimization, for both classification and regression, with the generic correlation structure of the model (91).

Gaussian equivalence– A similar Gaussian conjecture has been discussed in a series of recent works, and some authors proved partial results in this direction (Hastie et al., 2022; Hu et al., 2022b; Mei et al., 2019b; Montanari et al., 2019; Gerace et al., 2020b; Goldt et al., 2020a; Goldt et al., 2021b; Dhifallah et al., 2020). Ref. (Goldt et al., 2021b) analyses a special case of the Gaussian model (corresponding to $\boldsymbol{\varphi}_t = \text{id}$ here), and proves a Gaussian equivalence theorem (GET) for feature maps $\boldsymbol{\varphi}_s$ given by single-layer neural networks with fixed weights. They also show that for Gaussian data $\boldsymbol{x} \sim \mathcal{N}(\mathbf{0}, \mathrm{I}_D)$, feature maps of the form $\boldsymbol{v} = \sigma(\mathrm{W}\boldsymbol{x})$ (with some technical restriction on the weights) led to the jointly-Gaussian property for the two scalars $(\boldsymbol{v} \cdot \boldsymbol{w}, \boldsymbol{u} \cdot \boldsymbol{\theta}_0)$ for *almost* any vector $\boldsymbol{w}$. However, their stringent assumptions on random teacher weights limited the scope of applications to unrealistic label models. A related line of work discussed similar universality through the lens of random matrix theory (El Karoui et al., 2010; Pennington et al., 2017; Louart et al., 2018a). In particular, Seddik et al. (Seddik et al., 2020) showed that, in our notations, vectors $[\boldsymbol{u}, \boldsymbol{v}]$ obtained from Gaussian inputs $\boldsymbol{x} \sim \mathcal{N}(\mathbf{0}, \mathrm{I}_D)$ with Lipschitz feature maps satisfy a concentration property. In this case, again, one can expect the two scalars $(\boldsymbol{v} \cdot \boldsymbol{w}, \boldsymbol{u} \cdot \boldsymbol{\theta}_0)$ to be jointly Gaussian with high-probability on $\boldsymbol{w}$. Remarkably, in the case of random feature maps, (Hu et al., 2022b) could go beyond this central-limit-like behavior and established the universality of the Gaussian covariate model (91) for the actual learned weights $\hat{\boldsymbol{w}}$.

3.1 MAIN TECHNICAL RESULTS

Our main technical result is a closed-form expression for the asymptotic training and generalisation errors (94) of the Gaussian covariate model introduced above. We start by presenting our result in the most relevant setting for the applications of interest in Section 3.2, which is the case of the ℓ_2 regularization. Next, we briefly present our result in larger generality, which includes non-asymptotic results for non-separable losses and regularizations.

We start by defining key quantities that we will use to characterize the estimator $\hat{\boldsymbol{w}}$. Let $\Omega = \mathrm{S}^\top \mathrm{diag}(\omega_i)\mathrm{S}$ be the spectral decomposition of Ω. Let:

$$\rho \equiv \frac{1}{d}\boldsymbol{\theta}_0^\top \Psi \boldsymbol{\theta}_0 \in \mathbb{R}, \qquad \bar{\boldsymbol{\theta}} \equiv \frac{\mathrm{S}\Phi^\top \boldsymbol{\theta}_0}{\sqrt{\rho}} \in \mathbb{R}^d \tag{97}$$

and define the joint empirical density $\hat{\mu}_d$ between $(\omega_i, \bar{\theta}_i)$:

$$\hat{\mu}_d(\omega, \bar{\theta}) \equiv \frac{1}{d}\sum_{i=1}^{d} \delta(\omega - \omega_i)\delta(\bar{\theta} - \bar{\theta}_i). \tag{98}$$

Note that $\Phi^\top \boldsymbol{\theta}_0$ is the projection of the teacher weights on the student space, and therefore $\bar{\boldsymbol{\theta}}$ is the rotated projection on the basis of the student covariance, rescaled by the teacher variance. Together with the student eigenvalues ω_i, these are relevant statistics of the model, encoded here in the joint distribution $\hat{\mu}_d$.

Assumptions — Consider the *high-dimensional* limit in which the number of samples n and the dimensions p, d go to infinity with fixed ratios:

$$\alpha \equiv \frac{n}{d}, \text{ and } \gamma \equiv \frac{p}{d}. \tag{99}$$

Assume that the covariance matrices Ψ, Ω are positive-definite and that the Schur complement of the block covariance in equation (91) is positive semi-definite. Additionally, the spectral distributions of the matrices Φ, Ψ and Ω converge to distributions such that the limiting joint distribution μ is well-defined, and their maximum singular values are bounded with high probability as $n, p, d \to \infty$. Finally, regularity assumptions are made on the loss and regularization functions mainly to ensure feasibility of the minimization problem. We assume that the cost function $r + g$ is coercive, i.e. $\lim_{\|\boldsymbol{w}\|_2 \to +\infty}(r+g)(\boldsymbol{w}) = +\infty$ and that the following scaling condition holds : for all $n, d \in \mathbb{N}, \boldsymbol{z} \in \mathbb{R}^n$ and any constant $c > 0$, there exist a finite, positive constant C, such that, for any standard normal random vectors $\boldsymbol{h} \in \mathbb{R}^d$ and $\boldsymbol{g} \in \mathbb{R}^n$:

$$\begin{aligned} \|\boldsymbol{z}\|_2 \leqslant c\sqrt{n} \implies \sup_{\boldsymbol{x} \in \partial g(z)} \|\boldsymbol{x}\|_2 \leqslant C\sqrt{n}, \\ \frac{1}{d}\mathbb{E}\left[r(\boldsymbol{h})\right] < +\infty, \qquad \frac{1}{n}\mathbb{E}\left[g(\boldsymbol{g})\right] < +\infty \end{aligned} \tag{100}$$

We are now in a position to state our result.

Theorem 3.1.1. *(Closed-form asymptotics for ℓ_2 regularization) In the asymptotic limit defined above, the training and generalisation errors* (94) *of the* *estimator* $\hat{\boldsymbol{w}} \in \mathbb{R}^d$ *solving the empirical risk minimisation problem in eq.* (93) *with* ℓ_2 *regularization* $r(\boldsymbol{w}) = \frac{\lambda}{2}||\boldsymbol{w}||_2^2$ *verify:*

$$\begin{aligned}
&\mathscr{E}_{\text{train.}}(\hat{\boldsymbol{w}}) \xrightarrow[d\to\infty]{P} \mathbb{E}_{s,h\sim\mathcal{N}(0,1)}\left[g\left(prox_{V^\star g(.,f_0(\sqrt{\rho}s))}\left(\tfrac{m^\star}{\sqrt{\rho}}s+\sqrt{q^\star-\tfrac{m^{\star 2}}{\rho}}h\right),f_0(\sqrt{\rho}s)\right)\right] \\
&\mathscr{E}_{\text{gen.}}(\hat{\boldsymbol{w}}) \xrightarrow[d\to\infty]{P} \mathbb{E}_{(\nu,\lambda)}\left[\hat{g}\left(\hat{f}(\lambda),f_0(\nu)\right)\right]
\end{aligned} \tag{101}$$

where prox stands for the proximal operator defined as

$$prox_{Vg(.,y)}(x) = \text{argmin}_z\{g(z,y) + \frac{1}{2V}(x-z)^2\} \tag{102}$$

and where (ν,λ) *are jointly Gaussian scalar variables:*

$$(\nu,\lambda) \sim \mathcal{N}\left(0,\begin{bmatrix}\rho & m^\star \\ m^\star & q^\star\end{bmatrix}\right), \tag{103}$$

and the overlap parameters $(V^\star,q^\star,m^\star)$ *are prescribed by the unique fixed point of the following set of self-consistent equations:*

$$\begin{cases}
V = \mathbb{E}_{(\omega,\bar{\theta})\sim\mu}\left[\frac{\omega}{\lambda+\hat{V}\omega}\right] \\
m = \frac{\hat{m}}{\sqrt{\gamma}}\mathbb{E}_{(\omega,\bar{\theta})\sim\mu}\left[\frac{\bar{\theta}^2}{\lambda+\hat{V}\omega}\right] \\
q = \mathbb{E}_{(\omega,\bar{\theta})\sim\mu}\left[\frac{\hat{m}^2\bar{\theta}^2\omega+\hat{q}\omega^2}{\left(\lambda+\hat{V}\omega\right)^2}\right]
\end{cases}$$

$$\begin{cases}
\hat{V} = \frac{\alpha}{V}(1-\mathbb{E}_{s,h\sim\mathcal{N}(0,1)}[f_g'(V,m,q)]) \\
\hat{m} = \frac{1}{\sqrt{\rho\gamma}}\frac{\alpha}{V}\mathbb{E}_{s,h\sim\mathcal{N}(0,1)}\left[sf_g(V,m,q)-\frac{m}{\sqrt{\rho}}f_g'(V,m,q)\right] \\
\hat{q} = \frac{\alpha}{V^2}\mathbb{E}_{s,h\sim\mathcal{N}(0,1)}\left[\left(\frac{m}{\sqrt{\rho}}s+\sqrt{q-\frac{m^2}{\rho}}h-f_g(V,m,q)\right)^2\right]
\end{cases} \tag{104}$$

where we defined the scalar random functions

$$f_g(V,m,q) = prox_{Vg(.,f_0(\sqrt{\rho}s))}(\rho^{-1/2}ms+\sqrt{q-\rho^{-1}m^2}h)$$

and $f_g'(V,m,h) = prox'_{Vg(.,f_0(\sqrt{\rho}s))}(\rho^{-1/2}ms+\sqrt{q-\rho^{-1}m^2}h)$ *as the first derivative of the proximal operator.*

Proof: This result is a consequence of Theorem 3.1.2.

The parameters of the model $(\boldsymbol{\theta}_0,\Omega,\Phi,\Psi)$ only appear trough ρ, eq. (97), and the asymptotic limit μ of the joint distribution eq. (98) and $(f_0,\hat{f},g,\lambda)$. One can easily iterate the above equations to find their fixed point, and extract (q^*,m^*) which appear in the expressions for the training and generalisation errors $(\mathscr{E}^\star_{\text{train}},\mathscr{E}^\star_{\text{gen}})$, see eq. (94). Note that $(q^\star,m^\star)$ have an intuitive interpretation in terms of the estimator $\hat{\boldsymbol{w}} \in \mathbb{R}^d$:

$$q^\star \equiv \frac{1}{d}\hat{\boldsymbol{w}}^\top\Omega\hat{\boldsymbol{w}}, \qquad m^\star \equiv \frac{1}{\sqrt{dp}}\boldsymbol{\theta}_0^\top\Phi\hat{\boldsymbol{w}} \tag{105}$$

Or in words: $m^\star$ is the correlation between the estimator projected in the teacher space, while $q^\star$ is the reweighted norm of the estimator by the covariance Ω. The parameter V^* also has a concrete interpretation : it parametrizes the deformation that must be applied to a Gaussian field specified by the solution of the fixed point equations to obtain the asymptotic behaviour of $\hat{\mathbf{z}}$. It prescribes the degree of non-linearity given to the linear output by the chosen loss function. This is coherent with the robust regression viewpoint, where one introduces non-square losses to deal with the potential non-linearity of the generative model. $\hat{V}^*$ plays a similar role for the estimator $\hat{\mathbf{w}}$ through the proximal operator of the regularisation. Two cases are of particular relevance for the experiments that follow. The first is the case of *ridge regression*, in which $f_0(x) = \hat{f}(x)$ and both the loss g and the performance measure $\hat{g}$ are taken to be the *mean-squared error* $\mathrm{mse}(y,\hat{y}) = \frac{1}{2}(y-\hat{y})^2$, and the asymptotic errors are given by the simple closed-form expression:

$$\mathcal{E}^\star_{\text{gen}} = \rho + q^\star - 2m^\star, \qquad \mathcal{E}^\star_{\text{train}} = \frac{\mathcal{E}^\star_{\text{gen}}}{(1+V^\star)^2}, \tag{106}$$

The second case of interest is the one of a binary classification task, for which $f_0(x) = \hat{f}(x) = \mathrm{sign}(x)$, and we choose the performance measure to be the *classification error* $\hat{g}(y,\hat{y}) = \mathbb{P}(y \neq \hat{y})$. In the same notation as before, the asymptotic generalisation error in this case reads:

$$\mathcal{E}^\star_{\text{gen}} = \frac{1}{\pi}\cos^{-1}\left(\frac{m^\star}{\sqrt{\rho q^\star}}\right), \tag{107}$$

while the training error $\mathcal{E}^\star_{\text{train}}$ depends on the choice of g - which we will take to be the logistic loss $g(y,x) = \log\left(1+e^{-xy}\right)$ in all of the binary classification experiments.

As mentioned above, this paper includes stronger technical results including finite size corrections and precise characterization of the distribution of the estimator $\hat{\boldsymbol{w}}$, for generic, non-separable loss and regularization g and r. This type of distributional statement is encountered for special cases of the model in related works such as (Miolane et al., 2018; Celentano et al., 2020; Montanari et al., 2019). Define $\mathcal{V} \in \mathbb{R}^{n\times d}$ as the matrix of concatenated samples used by the student. Informally, in high-dimension, the estimator $\hat{\boldsymbol{w}}$ and $\hat{\boldsymbol{z}} = \frac{1}{\sqrt{d}}\mathcal{V}\hat{\boldsymbol{w}}$ roughly behave as non-linear transforms of Gaussian random variables centered around the teacher vector $\boldsymbol{\theta}_0$ (or its projection on the covariance spaces) as follows:

$$\boldsymbol{w}^* = \Omega^{-1/2}\mathrm{prox}_{\frac{1}{\hat{V}^*}r(\Omega^{-1/2}.)}\left(\frac{1}{\hat{V}^*}(\hat{m}^*\boldsymbol{t} + \sqrt{\hat{q}^*}\boldsymbol{g})\right),$$

$$\boldsymbol{z}^* = \mathrm{prox}_{V^*g(.,\boldsymbol{z})}\left(\frac{m^*}{\sqrt{\rho}}\boldsymbol{s} + \sqrt{q^* - \frac{(m^*)^2}{\rho}}\boldsymbol{h}\right).$$

where $\boldsymbol{s},\boldsymbol{h} \sim \mathcal{N}(0,\mathrm{I}_n)$ and $\boldsymbol{g} \sim \mathcal{N}(0,\mathrm{I}_d)$ are random vectors independent of the other quantities, $\boldsymbol{t} = \Omega^{-1/2}\Phi^\top\boldsymbol{\theta}_0$, $\boldsymbol{y} = \boldsymbol{f}_0\left(\sqrt{\rho}\boldsymbol{s}\right)$, and $(V^*,\hat{V}^*,q^*,\hat{q}^*,m^*,\hat{m}^*)$

is the unique solution to the fixed point equations. The formal concentration of measure result can then be stated in the following way:

Theorem 3.1.2. *(Non-asymptotic version, generic loss and regularization)(*informal*) Consider any optimal solution $\hat{\boldsymbol{w}}$ to 93. Then, there exist constants $C, c, c' > 0$ such that, for any Lipschitz function $\phi_1 : \mathbb{R}^d \to \mathbb{R}$, and separable, pseudo-Lipschitz function $\phi_2 : \mathbb{R}^n \to \mathbb{R}$ and any $0 < \varepsilon < c'$:*

$$\mathbb{P}\left(\left|\phi_1\left(\frac{\hat{\boldsymbol{w}}}{\sqrt{d}}\right) - \mathbb{E}\phi_1\left(\frac{\boldsymbol{w}^*}{\sqrt{d}}\right)\right| \geqslant \varepsilon\right) \leqslant \frac{C}{\varepsilon^2} e^{-cn\varepsilon^4}$$
$$\mathbb{P}\left(\left|\phi_2\left(\frac{\hat{\boldsymbol{z}}}{\sqrt{n}}\right) - \mathbb{E}\phi_2\left(\frac{\boldsymbol{z}^*}{\sqrt{n}}\right)\right| \geqslant \varepsilon\right) \leqslant \frac{C}{\varepsilon^2} e^{-cn\varepsilon^4}.$$

Note that in this form, the dimensions n, p, d still appear explicitly, as we are characterizing the convergence of the estimator's distribution for large but finite dimension. The clearer, one-dimensional statements are recovered by taking the $n, p, d \to \infty$ limit with separable functions and an ℓ_2 regularization. Other simplified formulas can also be obtained from our general result in the case of an ℓ_1 penalty, but since this breaks rotational invariance, they do look more involved than the ℓ_2 case. From Theorem 3.1.2, one can deduce the expressions of a number of observables, represented by the test functions ϕ_1, ϕ_2, characterizing the performance of $\hat{\boldsymbol{w}}$, for instance the training and generalization error.

3.2 APPLICATIONS OF THE GAUSSIAN MODEL

We now discuss how the theorems above are applied to characterise the learning curves for a range of concrete cases. We present a number of cases – some rather surprising – for which Conjecture 3.0.1 seems valid, and point out some where it is not.

3.2.1 RANDOM KITCHEN SINK WITH GAUSSIAN DATA

If we choose RF maps $\boldsymbol{\varphi}_s(\boldsymbol{x}) = \sigma(\mathrm{F}\boldsymbol{x})$ for a random matrix F and a chosen scalar function σ acting component-wise, we obtain the random kitchen sink model (Rahimi et al., 2008). This model has seen a surge of interest recently, and a sharp asymptotic analysis was provided in the particular case of uncorrelated Gaussian data $\boldsymbol{x} \sim \mathcal{N}(\boldsymbol{0}, \mathrm{I}_D)$ and $\boldsymbol{\varphi}_t(\boldsymbol{x}) = \boldsymbol{x}$ in (Mei et al., 2019b; Hastie et al., 2022) for ridge regression and generalised by (Gerace et al., 2020b; Hu et al., 2022b) for generic convex losses. Both results can be framed as a Gaussian covariate model with:

$$\Psi = \mathrm{I}_p, \qquad \Phi = \kappa_1 \mathrm{F}^\top, \qquad \Omega = \kappa_0^2 \boldsymbol{1}_d \boldsymbol{1}_d^\top + \kappa_1^2 \frac{\mathrm{F}\mathrm{F}^\top}{d} + \kappa_\star^2 \mathrm{I}_d, \tag{108}$$

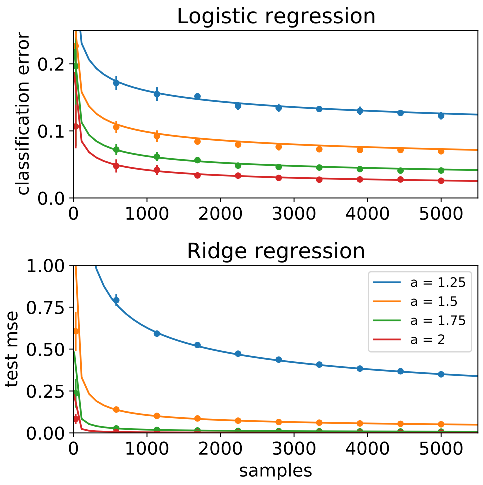


Figure 6: Learning in kernel space: Teacher and student live in the same (Hilbert) feature space $\boldsymbol{v} = \boldsymbol{u} \in \mathbb{R}^d$ with $d \gg n$, and the performance only depends on the relative decay between the student spectrum $\omega_i = d\, i^{-2}$ (the capacity) and the teacher weights in feature space $\theta_{0i}^2 \omega_i = d\, i^{-a}$ (the source). Top: a task with sign teacher (in kernel space), fitted with a max-margin support vector machine (logistic regression with vanishing regularisation (Rosset et al., 2003)). Bottom: a task with linear teacher (in kernel space) fitted via kernel ridge regression with vanishing regularisation. Points are simulation that matches the theory (lines). Simulations are averaged over 10 independent runs.

where $\mathbf{1}_d \in \mathbb{R}^d$ is the all-one vector and the constants $(\kappa_0, \kappa_1, \kappa_\star)$ are related to the non-linearity σ:

$$\begin{aligned}
\kappa_0 &= \mathbb{E}_{z \sim \mathcal{N}(0,1)}\left[\sigma(z)\right], \\
\kappa_1 &= \mathbb{E}_{z \sim \mathcal{N}(0,1)}\left[z\sigma(z)\right], \\
\kappa_\star &= \sqrt{\mathbb{E}_{z \sim \mathcal{N}(0,1)}\left[\sigma(z)^2\right] - \kappa_0^2 - \kappa_1^2}\,.
\end{aligned} \tag{109}$$

In this case, the averages over μ in eq. (104) can be directly expressed in terms of the Stieltjes transform associated with the spectral density of $\mathrm{FF}^\top$. Note, however, that our present framework can accommodate more involved random sinks models, such as when the teacher features are also a RF model or multi-layer random architectures. Deep RF are the object of further in-depth discussion in Chapters 6 and 7 in Part III.

3.2.2 KERNEL METHODS WITH GAUSSIAN DATA

Another direct application of our formalism is to kernel methods, which shall be the object of more detailed discussion in Chapters 4 and 5. Kernel methods admit a dual representation in terms of optimization over feature space (Scholkopf et al., 2018). The connection is given by Mercer's theorem, which provides an eigen-decomposition of the kernel and of the target function in the feature basis, effectively mapping kernel regression to a T-S problem on feature space. The classical way of studying the performance of kernel methods (Steinwart et al., 2009; Caponnetto et al., 2007) is then to directly analyse the performance of convex learning in this space. In our notation, the teacher and student feature maps are equal, and we thus set $p = d, \Psi = \Phi = \Omega = \text{diag}(\omega_i)$ where ω_i are the eigenvalues of the kernel and we take the teacher weights $\boldsymbol{\theta}_0$ to be the decomposition of the target function in the kernel feature basis.

There are many results in classical learning theory on this problem for the case of ridge regression (where the teacher is usually called "the source" and the eigenvalues of the kernel matrix the "capacity", see e.g. (Steinwart et al., 2009; Pillaud-Vivien et al., 2018)). However, these are worst case approaches, where no assumption is made on the true distribution of the data. In contrast, here we follow a *t*ypical case analysis, assuming Gaussianity in feature space. Through Theorem 3.1.1, this allows us to go beyond the restriction of the ridge loss. An example for logistic loss is in Fig. 6.

For the particular case of kernel ridge regression, Th. 3.1.1 provides a rigorous proof of the formula conjectured in (Bordelon et al., 2020). Hard-margin Support Vector Machines (SVMs) have also been studied using the heuristic replica method from statistical physics in (Dietrich et al., 1999a; Opper et al., 2001a). In our framework, this corresponds to the *hinge loss* $g(x,y) = \max(0, 1-yx)$ when $\lambda \to 0^+$. Our theorem thus puts also these works on rigorous grounds, and extends them to more general losses and regularization. We refer the interested reader to Chapters 4 and 5 for further in-depth discussion.

3.2.3 GAN-GENERATED DATA AND LEARNED TEACHERS

To approach more realistic data sets, we now consider the case in which the input data $\boldsymbol{x} \in \mathscr{X}$ is given by a GAN $\boldsymbol{x} = \mathscr{G}(\boldsymbol{z})$, where $\boldsymbol{z}$ is a Gaussian i.i.d. latent vector. Therefore, the covariates $[\boldsymbol{u}, \boldsymbol{v}]$ are the result of the following Markov chain:

$$\boldsymbol{z} \underset{\mathscr{G}}{\mapsto} \boldsymbol{x} \in \mathscr{X} \underset{\boldsymbol{\varphi}_t}{\mapsto} \boldsymbol{u} \in \mathbb{R}^p, \qquad \boldsymbol{z} \underset{\mathscr{G}}{\mapsto} \boldsymbol{x} \in \mathscr{X} \underset{\boldsymbol{\varphi}_s}{\mapsto} \boldsymbol{v} \in \mathbb{R}^d. \tag{110}$$

With a model for the covariates, the missing ingredient is the teacher weights $\boldsymbol{\theta}_0 \in \mathbb{R}^p$, which determine the label assignment: $y = f_0(\boldsymbol{u}^\top \boldsymbol{\theta}_0)$. In the experiments that follow, we fit the teacher weights *from the original data set in which the generative model $\mathscr{G}$ was trained.* Different choices for the fitting yield different teacher weights, and the quality of label assignment can be accessed by the performance of the fit on the test set. The set $(\boldsymbol{\varphi}_t, \boldsymbol{\varphi}_s, \mathscr{G}, \boldsymbol{\theta}_0)$ defines the data generative process. For predicting the learning curves from the iterative eqs. (104) we need to sample from the spectral measure μ, which amounts to estimating the *population* covariances (Ψ, Φ, Ω). This is done from the generative process in eq. (110) with a Monte Carlo sampling algorithm.

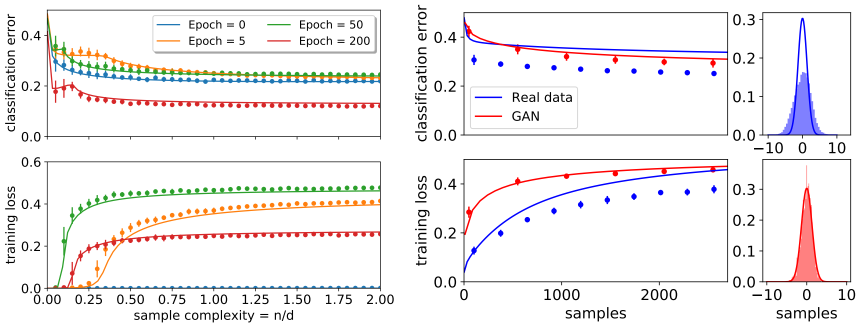


Figure 7: **Left**: generalisation classification error (top) and (unregularised) training loss (bottom) vs the sample complexity $\alpha = n/d$ for logistic regression on a learned feature map trained on Deep Convolutional Generative Adversarial Network (dcGAN)-generated CIFAR10-like images labelled by a teacher fully-connected neural network, with vanishing ℓ_2 regularisation. The different curves compare featured maps at different epochs of training. The theoretical predictions based on the Gaussian covariate model (full lines) are in very good agreement with the actual performance (points). **Right**: Test classification error (top) and (unregularised) training loss, (bottom) for logistic regression as a function of the number of samples n for an animal vs not-animal binary classification task with ℓ_2 regularization $\lambda = 10^{-2}$, comparing real CIFAR10 grey-scale images (blue) with dcGAN-generated CIFAR10-like gray-scale images (red). The real-data learning curve was estimated, just as in Figs. 8 from the population covariances on the full data set, and it is not in agreement with the theory in this case. On the very right we depict the histograms of the variable $\frac{1}{\sqrt{d}}\boldsymbol{v}^\top\hat{\boldsymbol{w}}$ for a fixed number of samples $n = 2d = 2048$ and the respective theoretical predictions (solid line). Simulations are averaged over 10 independent runs.

Fig. 7 shows an example of the learning curves resulting from the pipeline discussed above in a logistic regression task on data generated by a GAN trained on CIFAR10 images. More concretely, we used a pre-trained five-layer deep convolutional GAN (dcGAN) from (Radford et al., 2016), which maps 100 dimensional i.i.d. Gaussian noise into $k = 32 \times 32 \times 3$ realistic looking CIFAR10-like images: $\mathscr{G} : \boldsymbol{z} \in \mathbb{R}^{100} \mapsto \boldsymbol{x} \in \mathbb{R}^{32\times32\times3}$. To generate labels, we trained a simple fully-connected four-layer neural network on the *real* CIFAR10 data set, on a odd ($y = +1$) vs. even ($y = -1$) task, achieving $\sim 75\%$ classification accuracy on the test set. The teacher weights $\boldsymbol{\theta}_0 \in \mathbb{R}^p$ were taken from the last layer of the network, and the teacher feature map $\boldsymbol{\varphi}_t$ from the three previous layers. For the student model, we trained a completely independent fully connected 3-layer neural network on the dcGAN-generated CIFAR10-like images and took snapshots of the feature maps $\boldsymbol{\varphi}_s^i$ induced by the 2-first layers during the first $i \in \{0, 5, 50, 200\}$ epochs of training. Finally, once $\left(\mathscr{G}, \boldsymbol{\varphi}_t, \boldsymbol{\varphi}_s^i, \boldsymbol{\theta}_0\right)$ have been fixed, we estimated the covariances (Ψ, Φ, Ω) with a Monte Carlo algorithm.

Fig. 7 depicts the resulting learning curves obtained by training the last layer of the student. Interestingly, the performance of the feature map at epoch 0 (random initialisation) beats the performance of the learned features during early phases of training in this experiment. Another interesting behaviour is given by the separability threshold of the learned features, i.e. the number of samples for which the training loss becomes larger than 0 in logistic regression. At epoch 50 the learned features are separable at lower sample complexity $\alpha = n/d$ than at epoch 200 - even though in the later the training and generalisation performances are better.

3.2.4 LEARNING FROM REAL DATA SETS

Applying T-S to a real data set — Given that the learning curves of realistic-looking inputs can be captured by the Gaussian covariate model, it is fair to ask whether the same might be true for *real data sets*. To test this idea, we first need to cast the real data set into the T-S formalism, and then compute the covariance matrices Ω, Ψ, Φ and teacher vector $\boldsymbol{\theta}_0$ required by model (91).

Let $\{\boldsymbol{x}^\mu, y^\mu\}_{\mu=1}^{n_{\text{tot}}}$ denote a real data set, e.g. MNIST or Fashion-MNIST for concreteness, where $n_{\text{tot}} = 7 \times 10^4$, $\boldsymbol{x}^\mu \in \mathbb{R}^D$ with $D = 784$. Without loss of generality, we can assume the data is centred. To generate the teacher, let $\boldsymbol{u}^\mu = \boldsymbol{\varphi}_t(\boldsymbol{x}^\mu) \in \mathbb{R}^p$ be a feature map such that data is invertible in feature space, i.e. that $y^\mu = \boldsymbol{\theta}_0^\top \boldsymbol{u}^\mu$ for some teacher weights $\boldsymbol{\theta}_0 \in \mathbb{R}^p$, which should be computed from the samples. Similarly, let $\boldsymbol{v}^\mu = \boldsymbol{\varphi}_s(\boldsymbol{x}^\mu) \in \mathbb{R}^d$ be a feature map we are interested in studying. Then, we can estimate the population covariances (Ψ, Φ, Ω) empirically from the *entire* data set as:

$$\Psi = \sum_{\mu=1}^{n_{\text{tot}}} \frac{\boldsymbol{u}^\mu \boldsymbol{u}^{\mu\top}}{n_{\text{tot}}}, \qquad \Phi = \sum_{\mu=1}^{n_{\text{tot}}} \frac{\boldsymbol{u}^\mu \boldsymbol{v}^{\mu\top}}{n_{\text{tot}}}, \qquad \Omega = \sum_{\mu=1}^{n_{\text{tot}}} \frac{\boldsymbol{v}^\mu \boldsymbol{v}^{\mu\top}}{n_{\text{tot}}}. \tag{111}$$

At this point, we have all we need to run the self-consistent equations (104). The issue with this approach is that there is not a unique teacher map $\boldsymbol{\varphi}_t$ and teacher vector $\boldsymbol{\theta}_0$ that fit the true labels. However, we can show that *all interpolating linear teachers are equivalent*:

Theorem 3.2.1. *(Universality of linear teachers) For any teacher feature map $\boldsymbol{\varphi}_t$, and for any $\boldsymbol{\theta}_0$ that interpolates the data so that $y^\mu = \boldsymbol{\theta}_0^\top \boldsymbol{u}^\mu \, \forall \mu$, the asymptotic predictions of model (91) are equivalent.*

Proof. It follows from the fact that the teacher weights and covariances only appear in eq. (104) through $\rho = \frac{1}{p}\boldsymbol{\theta}_0^\top \Psi \boldsymbol{\theta}_0$ and the projection $\Phi^\top \boldsymbol{\theta}_0$. Using the estimation (111) and the assumption that it exists $y^\mu = \boldsymbol{\theta}_0^\top \boldsymbol{u}^\mu$, one can write these quantities directly from the labels y^μ:

$$\rho = \frac{1}{n_{\text{tot}}} \sum_{\mu=1}^{n_{\text{tot}}} (y^\mu)^2, \qquad \Phi^\top \boldsymbol{\theta}_0 = \frac{1}{n_{\text{tot}}} \sum_{\mu=1}^{n_{\text{tot}}} y^\mu \boldsymbol{v}^\mu. \tag{112}$$

For linear interpolating teachers, results are thus independent of the choice of the teacher. □

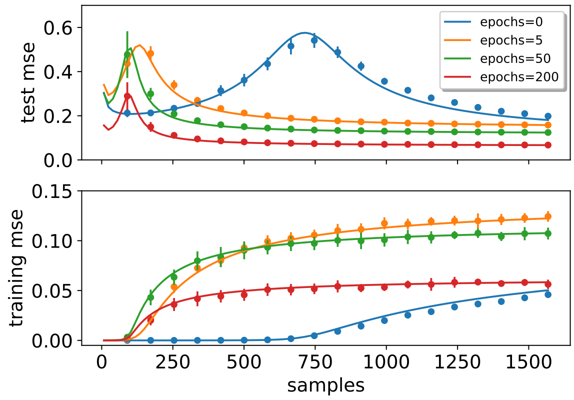


Figure 8: Test and training mean-squared errors eqs. (106) as a function of the number of samples n for ridge regression. The Fashion-MNIST data set, with vanishing regularisation $\lambda = 10^{-5}$. In this plot, the student feature map $\boldsymbol{\varphi}_s$ is a 3-layer fully-connected neural network with $d = 2352$ hidden neurons trained on the full data set with the square loss. Different curves correspond to the feature map obtained at different stages of training. Simulations are averaged over 10 independent runs.

Although this result might seen surprising at first sight, it is quite intuitive. Indeed, the information about the teacher model only enters the Gaussian covariate model (91) through the statistics of $\boldsymbol{u}^\top \boldsymbol{\theta}_0$. For a linear teacher $f_0(x) = x$, this is precisely given by the labels.

Ridge Regression with linear teachers — We now test the prediction of model (91) on real data sets, and show that it is surprisingly effective in predicting the learning curves, at least for the ridge regression task. We have trained a 3-layer fully connected neural network with ReLU activations on the full Fashion-MNIST data set to distinguish clothing used above vs. below the waist (Xiao et al., 2017). The student feature map $\boldsymbol{\varphi}_s : \mathbb{R}^{784} \to \mathbb{R}^d$ is obtained by removing the last layer. In Fig. 8 we show the test and training errors of the ridge estimator on a sub-sample of $n < n_{\text{tot}}$ on the Fashion-MNIST images. We observe remarkable agreement between the learning curve obtained from simulations and the theoretical prediction by the matching Gaussian covariate model. Note that for the square loss and for $\lambda \ll 1$, the worst performance peak is located at the point in which the linear system becomes invertible. Curiously, Fig. 8 shows that the fully-connected network progressively learns a low-rank representation of the data as training proceeds. This can be directly verified by counting the number of zero eigenvalues of Ω, which go from a full-rank matrix to a matrix of rank 380 after 200 epochs of training.

Fig. 5 (right) shows a similar experiment on the MNIST data set, but for different out-of-the-box feature maps, such as random features and the scattering transform (Bruna et al., 2013), and we chose the number of random features $d = 1953$ to match the number of features from the scattering transform. Note the characteristic double-descent behaviour (Opper et al., 1996; Spigler et al., 2019; Belkin et al., 2019), and the accurate prediction of the peak where the interpolation transition occurs. For both Figs. 8 and 5, for a number of samples n closer to n_{tot} we start to see deviations between the real learning curve and the theory. This is to be expected since in the T-S framework the student can, in principle, express the same function as the teacher if it recovers its weights exactly. Recovering the teacher weights becomes possible with a large training set. In that case, its test error will be zero. However, in our setup the test error on real data remains finite even if more training data is added, leading to the discrepancy between T-S learning curve and real data.

Why is the Gaussian model so effective for describing learning with data that are *n*ot Gaussian? The point is that ridge regression is sensitive only to second order statistics, and not to the full distribution of the data. It is a classical property that the training and generalisation errors are only a function of the spectrum of the *empirical* and *population* covariances, and of their products. Random matrix theory teaches us that such quantities are very robust, and their asymptotic behaviour is universal for a broad class of distributions of $[\boldsymbol{u}, \boldsymbol{v}]$ (Bai et al., 2008b; Ledoit et al., 2011; El Karoui et al., 2009; Louart et al., 2018a). The asymptotic behavior of kernel matrices has indeed been the subject of intense scrutiny (El Karoui et al., 2010; Cheng et al., 2013; Pennington et al., 2017; Mei et al., 2019b; Fan et al., 2019; Seddik et al., 2020). Indeed, a universality result akin to Theorem 3.2.1 was noted in (Jacot et al., 2020b) in the specific case of kernel methods. We thus expect the validity of model (91) for ridge regression, with a linear teacher, to go way beyond the Gaussian assumption.

Beyond ridge regression — The same strategy fails beyond ridge regression and mean-squared test error. This suggests a limit in the application of model (91) to real (non-Gaussian) data to the universal linear teacher. To illustrate this, consider the setting of Figs. 8, and compare the model predictions for the binary classification error instead of the ℓ_2 one. There is a clear mismatch between the simulated performance and prediction given by the theory due to the fact that the classification error does not depends only on the first two moments.

We present an additional experiment in Fig. 7. We compare the learning curves of logistic regression on a classification task on the *real* CIFAR10 images with the real labels versus the one on dcGAN-generated CIFAR10-like images and teacher generated labels from Sec. 3.2.3. While the Gaussian theory captures well the behaviour of the later, it fails on the former. A histogram of the distribution of the product $\boldsymbol{u}^\top \hat{\boldsymbol{w}}$ for a fixed number of samples illustrates well the deviation from the prediction of the theory with the real case, in particular on the tails of the distribution. The difference between GAN generated data (that fits the Gaussian theory) and real data is clear. Given that for classification problems there exists a number of choices of "sign" teachers and feature maps that give the exact same labels as in the data set, an interesting open question is: *is there a teacher that allows to reproduce the learning curves more accurately*? This question is left for future works.

Part II B.

KERNEL FEATURES

4

SCALING LAWS FOR KERNEL REGRESSION

Kernel methods are among the most popular models in machine learning. Despite their relative simplicity, they define a powerful framework in which non-linear features can be exploited without leaving the realm of convex optimisation. Kernel methods in machine learning have a long and rich literature dating back to the 60s (Nadaraya, 1964; Watson, 1964), but have recently made it back to the spotlight as a proxy for studying neural networks in different regimes, e.g. the infinite width limit (Neal, 1996b; Williams, 1996a; Jacot et al., 2018a; Lee et al., 2018a) and the lazy regime of training (Chizat et al., 2019b). Despite being defined in terms of a non-parametric optimisation problem, kernel methods can be mathematically understood as a standard parametric linear problem in a (possibly infinite) Hilbert space spanned by the kernel eigenvectors (a.k.a *features*). This dual picture fully characterizes the asymptotic performance of kernels in terms of a trade-off between two key quantities: the relative decay of the eigenvalues of the kernel (a.k.a. its *capacity*) and the coefficients of the target function when expressed in feature space (a.k.a. the *source*). Indeed, a sizeable body of work has been devoted to understanding the decay rates of the excess error as a function of these two relative decays, and investigated whether these rates are attained by algorithms such as Stochastic Gradient-Descent (SGD) (Pillaud-Vivien et al., 2018; Berthier et al., 2020).

Rigorous optimal rates for the excess generalization error in KRR are well-known since the seminal works of (Caponnetto et al., 2005; Steinwart et al., 2009). However, recent interesting works (Spigler et al., 2020; Bordelon et al., 2020) surprisingly reported very different - and actually better - rates supported by numerical evidences. These papers appeared to either not comment on this discrepancy (Bordelon et al., 2020), or to attribute this apparent contradiction to a difference between typical and worse-case analysis (Spigler et al., 2020). As we shall see, the key difference between these works stems instead from the fact that most of classical works considered *n*oisy data and fine-tuned regularization, while (Spigler et al., 2020; Bordelon et al., 2020) focused on noiseless data sets. This observation raises a number of questions: is there a connection between both sets of exponents? Are Gaussian design exponents actually different from worst-case ones? What about intermediary setups (for instance noisy labels with generic regularization, noiseless labels with varying regularization) and regimes (intermediary sample complexities)? How does infinitesimal noise differ from no noise at all?

Main contributions — In this manuscript, we answer all the above questions, and redeem the apparent contradiction by reconsidering the Gaussian design analysis. We provide a unifying picture of the decay rates for the excess generalization error, along a more exhaustive characterization of the regimes in which each is observed, evidencing the interplay of the role of regularization, noise and sample complexity. We show in particular that typical-case analysis with a Gaussian design is actually in perfect agreement with the statistical learning worst-case data-agnostic approach. We also show how the optimal excess error decay can transition from the recently reported noiseless value to its well known noisy value as the number of samples is increased. We illustrate this crossover from the *noiseless* regime to the *noisy* regime also in a variety of KRR experiments on real data.

Related work — The analysis for kernel methods and ridge regression is a classical topic in statistical learning theory (Caponnetto et al., 2005; Caponnetto et al., 2007; Steinwart et al., 2009; Fischer et al., 2020; Lin et al., 2018; Bartlett et al., 2020b; Lin et al., 2018). In this classical setting, decay exponents for optimally regularized *n*oisy linear regression on features with power-law co-variance spectrum have been provided. Interestingly, it has been shown that such optimal rates can be obtained in practice by SGD, without explicit regularization, with single-pass (Polyak et al., 1992; Nemirovskij et al., 1983) or multi-pass (Pillaud-Vivien et al., 2018), as well as by randomized algorithms (Jun et al., 2019). Closed-form bounds for the prediction error have been provided in a number of worst-case analyses (Jun et al., 2019; Lin et al., 2018).

The recent line of work on the noiseless setting includes contributions from statistical learning theory (Berthier et al., 2020; Varre et al., 2021) and statistical physics (Spigler et al., 2020; Bordelon et al., 2020). This much more recent second line of work proved decay rates for a given, constant regularization. An example of noise-induced crossover is furthermore mentioned in (Berthier et al., 2020). The interplay between noisy and noiseless regimes has also been investigated in the related Gaussian Process literature (Kanagawa et al., 2018).

The study of ridge regression with Gaussian design is also a classical topic. Ref. (Dicker et al., 2016) considered a model in which the covariates are isotropic Gaussian in $\mathbb{R}^p$, and computed the exact asymptotic generalization error in the high-dimensional asymptotic regime $p, n \to \infty$ with dimension-to-sample-complexity ratio p/n fixed. This result was generalised to arbitrary co-variances (Hsu et al., 2012; Dobriban et al., 2018a) using fundamental results from random matrix theory (Ledoit et al., 2011). Non-asymptotic rates of convergence for a related problems were given in Chapter 3. Previous results also existed in the statistical physics literature, e.g. (Dietrich et al., 1999b; Opper et al., 1996; Opper et al., 2001b; Kabashima, 2008). Gaussian models for regression have seen a surge of popularity recently, and have been used in particular to study over-parametrization and the double-descent phenomenon, e.g. in (Advani et al., 2020; Belkin et al., 2020; Hastie et al., 2022; Mei et al., 2019b; Gerace et al., 2020b; Ghorbani et al., 2019b; Kobak et al., 2020; Wu et al., 2020a; Bartlett et al., 2020b; Richards et al., 2021; Liao et al., 2020; Jacot et al., 2020b; Ghorbani et al., 2020b; Liu et al., 2020).

4.1 SETTING

Consider a data set $\mathscr{D} = \{x^\mu, y^\mu\}_{\mu=1}^n$ with n independent samples from a probability measure ν on $\mathscr{X} \times \mathscr{Y}$, where $\mathscr{X} \subset \mathbb{R}^d$ is the input and $\mathscr{Y} \subset \mathbb{R}$ the response space. Let K be a kernel and $\mathscr{H}$ denote its associated RKHS. KRR corresponds to the following non-parametric minimisation problem:

$$\min_{f \in \mathscr{H}} \frac{1}{n} \sum_{\mu=1}^{n} (f(x^\mu) - y^\mu)^2 + \lambda ||f||_{\mathscr{H}}^2. \tag{113}$$

where $||\cdot||_{\mathscr{H}}$ is the norm associated with the scalar product in $\mathscr{H}$, and $\lambda \geq 0$ is the regularisation. The convenience of KRR is that it admits a dual representation in terms of a standard parametric problem. Indeed, the kernel K can be diagonalized in an orthonormal basis $\{\phi_k\}_{k=1}^\infty$ of $L^2(\mathscr{X})$:

$$\int_{\mathscr{X}} \nu_x(\mathrm{d}x') K(x,x') \phi_k(x') = \eta_k \phi_k(x) \tag{114}$$

where $\{\eta_k\}_{k=1}^\infty$ are the corresponding (non-negative) kernel eigenvalues and ν_x is the marginal distribution over $\mathscr{X}$. Note that the kernel $\{\phi_k\}_{k=1}^\infty$ eigenvectors form an orthonormal basis of $L^2(\mathscr{X})$. It is convenient to define the re-scaled basis of *kernel features* $\psi_k(x) = \sqrt{\eta_k}\phi_k(x)$ and to work in matrix notation in feature space: define $\phi(x) \equiv \{\phi_k(x)\}_{k=1}^p$ (with p possibly infinite)

$$\begin{aligned}
&\psi(x) = \Sigma^{\frac{1}{2}} \phi(x) \\
&\mathbb{E}_{x \sim \nu_x} \left[\phi(x)\phi(x)^\top\right] = \mathbb{1}_p, \\
&\mathbb{E}_{x' \sim \nu_x} [K(x,x')\phi(x')] = \Sigma \phi(x),
\end{aligned} \tag{115}$$

where $\Sigma \equiv \mathbb{E}_{x \sim \nu_x}\left[\psi(x)\psi(x)^\top\right] = \mathrm{diag}(\eta_1, \eta_2, ..., \eta_p)$ is the features co-variance (a diagonal operator in feature space). In this notation, the RKHS $\mathscr{H}$ can be formally written as $\mathscr{H} = \{f = \psi^\top \theta : \theta \in \mathbb{R}^p, \quad ||\theta||_2 < \infty\}$, i.e. the space of functions for which the coefficients in the feature basis are square summable. With this notation, we can rewrite eq. (116) in feature space as a standard parametric problem for the following empirical risk:

$$\hat{\mathscr{R}}_n(w) = \frac{1}{n} \sum_{\mu=1}^{n} \left(w^\top \psi(x^\mu) - y^\mu\right)^2 + \lambda\, w^\top w. \tag{116}$$

Our main results concern the typical averaged performance of the KRR estimator, as measured by the typical prediction (out-of-sample) error

$$\epsilon_g = \mathbb{E}_{\mathscr{D}} \mathbb{E}_{(x,y) \sim \nu} (\hat{f}(x) - y)^2, \tag{117}$$

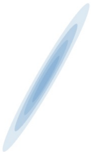
where the first average is over the data $\mathscr{D}=\{x^\mu,y^\mu\}$ and the second over a fresh sample $(x,y)\sim\nu$.

In what follows we assume the labels $y^\mu\in\mathscr{Y}$ were generated, up to an independent additive Gaussian noise with variance σ^2, by a target function $f^\star$ (not necessarily belonging to $\mathscr{H}$):

$$y^\mu \overset{d}{=} f^\star(x^\mu)+\sigma\mathcal{N}(0,1), \tag{118}$$

and we denote by $\theta^\star$ the coefficients of the target function in the features basis $f^\star(x)=\psi(x)^\top\theta^\star$. As we will characterize below, whether the target function $f^\star$ belongs or not to $\mathscr{H}$ depends on the relative decay coefficients $\theta^\star$ with respect to the eigenvalues of the kernel. We often refer to $\theta^\star$ as the *teacher.* While the present results and discussion are provided for additive gaussian noise for simplicity, our method are not restricted to this particular noise, and a more complete extension of the results for other noise settings is left for future work.

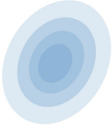
*A large **capacity** α characterizes an effectively low-dimensional features distribution*



We are then interested in the evolution of the *excess error* $\varepsilon_g-\sigma^2$ as the number of samples n is increased.

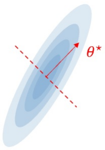
Capacity and source coefficients — Motivated by the discussion above, we focus on ridge regression in an infinite dimensional ($p\to\infty$) space $\mathscr{H}$ with Gaussian design $u^\mu \overset{\text{def}}{=} \psi(x^\mu) \overset{d}{=} \mathcal{N}(0,\Sigma)$ with (without loss of generality) diagonal co-variance $\Sigma=\mathrm{diag}(\eta_1,\eta_2,...)$. We expect however the results of this manuscript to be universal for a large class of distribution beyond the Gaussian one. In particular, we anticipate the gaussianity assumption should be amenable to being relaxed to sub-gaussians (Tsigler et al., 2020) or even any concentrated distribution (Talagrand, 1994; Louart et al., 2018b).

A small α conversely signals a high dimensional distribution



Following the statistical learning terminology, we introduce two parameters $\alpha>1, r\geq 0$, herefrom referred to as the *capacity* and *source* conditions (Caponnetto et al., 2007), to parametrize the difficulty of the target function and the learning capacity of the kernel

*A large **source** r means the teacher vector is well aligned with the main direction of the distribution*

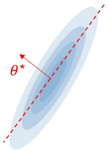


$$\mathrm{tr}\,\Sigma^{\frac{1}{\alpha}}<\infty, \qquad\qquad ||\Sigma^{\frac{1}{2}-r}\theta^\star||_{\mathscr{H}}<\infty. \tag{119}$$

As in (Dobriban et al., 2018a; Spigler et al., 2020; Bordelon et al., 2020; Berthier et al., 2020), we consider the particular case where both the spectrum of Σ and the teacher components $\theta_k^\star$ have exactly a power-law form satisfying the limiting source/capacity conditions (119):

$$\eta_k=k^{-\alpha}, \qquad\qquad \theta_k^\star=k^{-\frac{1+\alpha(2r-1)}{2}}. \tag{120}$$

A small r characterizes a harder to recover target



The power law ansatz (120) is empirically observed to be a rather good approximation for some real simple datasets and kernels. The parameters α, r introduced in (120) control the complexity of the data the teacher respectively.

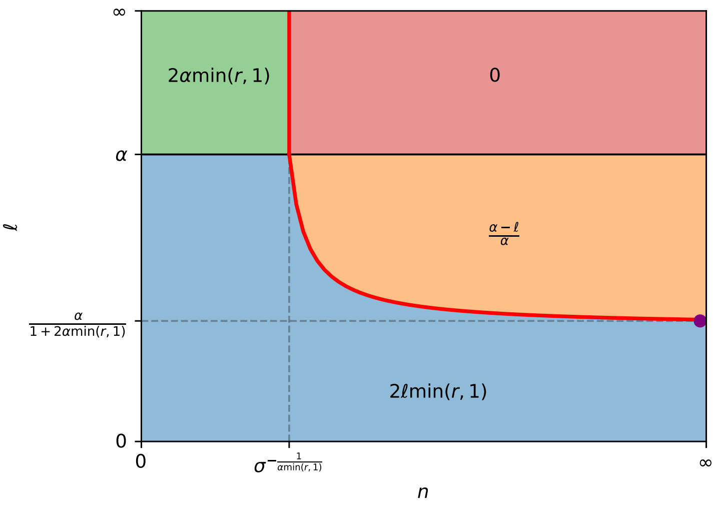


Figure 9: Different decays for the excess generalization error $\epsilon_g - \sigma^2$ for different values of n and different decays ℓ of the regularization $\lambda \sim n^{-\ell}$, at given noise variance σ. The red solid line represents the noise-induced crossover line, separating the effectively noiseless regime (green and blue) on its left from the effectively noisy regime (red and orange) on its right. Any KRR experiment at fixed regularization decay ℓ (corresponding to drawing a horizontal line at ordinate ℓ) crosses the crossover line if $\ell > \alpha/(1+2\alpha\min(r,1))$. The corresponding learning curve will accordingly exhibit a crossover from a fast decay (noiseless regime) to a slow decay (noisy regime).

A large α can be loosely seen as characterizing a effectively low dimensional (and therefore easy to fit) data distribution. By the same token, a large r signals a good alignment of the teacher with the important directions in the data covariance, and therefore an a priori simpler learning task.

The regularization λ is allowed to vary with n according to a power-law $\lambda = n^{-\ell}$. This very general form allows us to encompass both the zero regularization case (corresponding to $\ell = \infty$) and the case where $\lambda = \lambda^\star$ is optimized, with some optimal decay rate $\ell^\star$. Note that this power law form implies that λ is assumed positive. While this is indeed the assumption of (Caponnetto et al., 2005; Caponnetto et al., 2007) with which we intend to make contact, (Wu et al., 2020a) have shown that the optimal λ may in some settings be negative. Some numerical experiments suggest that removing the positivity constraint on λ while optimizing does not affect the results presented in this manuscript. A more detailed investigation is left to future work.

4.2 MAIN RESULTS

Depending on the regularization decay strength ℓ, capacity α, source r and noise variance σ^2, four regimes can be observed. The derivation of these decays from the asymptotic solution of the Gaussian design problem is sketched in Section 4.3, and here we concentrate on the key results. The different observable decays for the excess error $\epsilon_g - \sigma^2$ are summarized in Fig. 9, and are given by:

- If $\ell \geq \alpha$ (weak regularization $\lambda = n^{-\ell}$),

  $$\epsilon_g - \sigma^2 = \mathscr{O}\left(\max\left(\sigma^2, n^{-2\alpha\min(r,1)}\right)\right). \tag{121}$$

  The excess error transitions from a fast decay $2\alpha\min(r,1)$ (green region in Fig. 9 and green dashed line in Fig. 10) to a plateau (red region in Fig. 9 and red dashed line in Fig. 10) with no decay as n increases. This corresponds to a crossover from the green region to the red region in the phase diagram Fig. 9.

- If $\ell \leq \alpha$ (strong regularization $\lambda = n^{-\ell}$) ,

  $$\epsilon_g - \sigma^2 = \mathscr{O}\left(\max\left(\sigma^2, n^{1-2\ell\min(r,1)-\frac{\ell}{\alpha}}\right) n^{\frac{\ell-\alpha}{\alpha}}\right). \tag{122}$$

  The excess error transitions from a fast decay $2\ell\min(r,1)$ (blue region in Fig. 9) to a slower decay $(\alpha-\ell)/\alpha$ (orange region in Fig. 9) as n is increased and the effect of the additive noise kicks in, see Fig. 11. The crossover disappears for too slow decays $l \leq \alpha/(1+2\alpha\min(r,1))$, as the regularization λ is always sufficiently large to completely mitigate the effect of the noise. This corresponds to the max in (122) being realized by its second argument for all n.

Given these four different regimes as depicted in Fig. 9, one may wonder about the optimal learning solution when the regularization is fine tuned to its best value. To answer this question, we further define the *asymptotically optimal* regularization decay $\ell^\star$ as the value leading to fastest decay of the typical excess error $\epsilon_g - \sigma^2$. We find that two different optimal rates exist, depending on the quantity of data available.

- If $n \ll n_1^* \approx \sigma^{-\frac{1}{\alpha\min(r,1)}}$, any $\ell^\star \in (\alpha, \infty)$ yields excess error decay

  $$\epsilon_g^\star - \sigma^2 \sim n^{-2\alpha\min(r,1)}. \tag{123}$$

- If $n \gg n_2^* \approx \sigma^{-\max\left(2, \frac{1}{\alpha\min(r,1)}\right)}$,

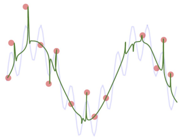
  $$\epsilon_g^\star - \sigma^2 \sim n^{\frac{1}{1+2\alpha\min(r,1)}-1}, \quad \text{by choosing} \quad \lambda^\star \sim n^{-\frac{\alpha}{1+2\alpha\min(r,1)}}. \tag{124}$$

The optimal decay for the excess error $\epsilon_g^\star - \sigma^2$ thus transitions from a fast decay $2\alpha\min(r,1)$ when $n \ll n_1^*$ – corresponding to, effectively, the optimal rates expected in a "noiseless" situation – to a slower decay $2\alpha\min(r,1)/(1+$

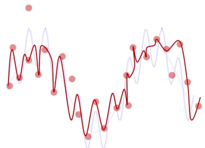
$2\alpha \min(r, 1))$ when $n \gg n_2^*$ corresponding to the classical "noisy" optimal rate, depicted with the purple point in Fig. 9. This is illustrated in Fig. 12 where the two rates are observed in succession for the same data as the number of points is increased.

We can now finally clarify the apparent discrepancy in the recent literature discussed in the introduction. The exponent recently reported in (Spigler et al., 2020; Bordelon et al., 2020) actually corresponds to the "noiseless" regime. In contrast, the rate described in (124) is the classical result (Caponnetto et al., 2005) for the non-saturated case $r < 1$ for generic data. We see here that the same rate is also achieved with Gaussian design, and that there are no differences between fixed and Gaussian design as long as the capacity and source condition are matching. We unveiled, however, the existence of two possible sets of optimal rate exponents depending on the number of data samples.

Benign overfitting *can be illustrated by the following 1d analogy. When learning a function (light blue lines) from a small number of training data (red points), high-variance features (in this illustration slow frequencies) are learnt well, whereas the less relevant features (higher frequencies) are used to overfit the noise. This is the analog of the green regime.*



With more samples, higher frequencies start to be learnt, but have to be disentangled from the noise, making the learning task harder : this is the analog of the red regime.



All setups (effectively non-regularized KRR (121), effectively regularized KRR (122) or optimally regularized KRR (123), (124)) can therefore exhibit a crossover from an effectively *noiseless* regime (green or blue in Fig. 9), to an effectively *noisy* regime (red, orange in Fig. 9) depending on the quantity of data available. We stress that while the noise is indeed present in the green and blue "noiseless" regimes, its presence is effectively not felt, and noiseless rates are observed. In fact, if the noise is small, one will not observed the classical noisy rates unless an astronomical amount of data is available. This can be intuitively understood as follows: for small sample size n, low-variance dimensions are used to overfit the noise, while the spiked subspace of large-variance dimensions is well fitted. In noiseless regions, the excess error is thus characterized by a fast decay. This phenomenon, where the noise variance is diluted over the dimensions of lesser importance, is connected to the *benign overfitting* discussed by (Bartlett et al., 2020b) and (Tsigler et al., 2020). Benign overfitting is possible due to the decaying structure of the co-variance spectrum (120). As more samples are accessed, further decrease of the excess error requires good generalization also over the low-variance subspace, and the overfitting of the noise results in a slower decay.

While our analysis is for the optimal full-batch learning, we note that a similar crossover in the case of SGD in the effectively non-regularized case (from green to red) has been discussed in (Berthier et al., 2020; Varre et al., 2021). It would be interesting to further explore how SGD can behave in the different regimes discussed here.

When $\lambda = \lambda_0 n^{-\ell}$ for a prefactor λ_0 that is allowed to be very small, a *regularization-induced* crossover, similar to the one reported in (Bordelon et al., 2020), can also be observed on top of the noise-induced crossover which is the focus of the present work.

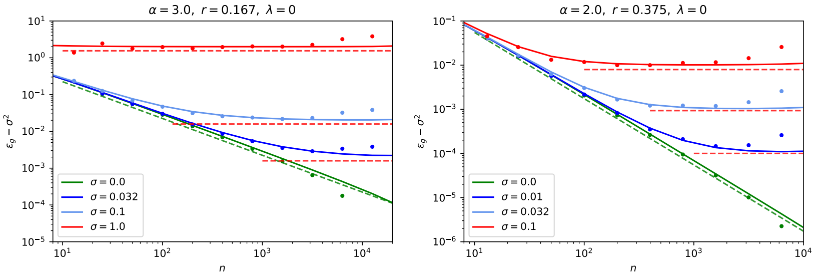


Figure 10: Kernel ridge regression on synthetic data sets with capacity α and source coefficient r i.e. the idealized gaussian setting (120), with no regularization $\lambda = 0$. Solid lines correspond to the theoretical prediction of eq. (126) using the GCM package associated with Chapter 3. Points are simulations conducted using the python `scikit-learn KernelRidge` package (Pedregosa et al., 2011a), where the feature space dimension has been cut off to $p = 10^4$ for the simulations, and to 10^5 for the theoretical curves. Dashed lines represent the slopes predicted by eq. (121), with the color (red and green) in correspondence to the regime from Fig. 9.

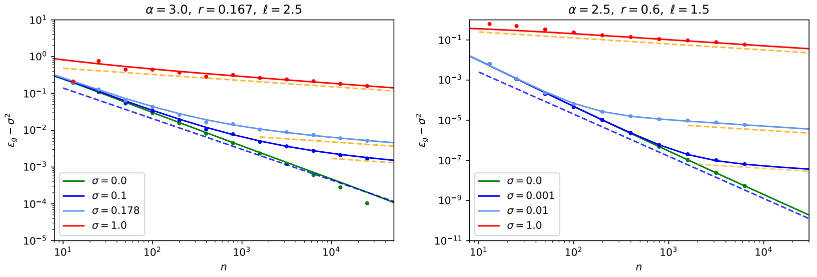


Figure 11: Kernel ridge regression on synthetic data sets with capacity α and source coefficient r, with regularization $\lambda = n^{-\ell}$. Solid lines correspond to the theoretical prediction of eq. (126) using the GCM package associated with Chapter 3. Points are simulations conducted using the python `scikit-learn KernelRidge` package (Pedregosa et al., 2011a), where the feature space dimension has been cut off to $p = 10^4$ for the simulations, and to 10^5 for the theoretical curves. Dashed lines represent the slopes predicted by eq. (122), with the color (blue and orange) in correspondence to the regime from Fig. 9.

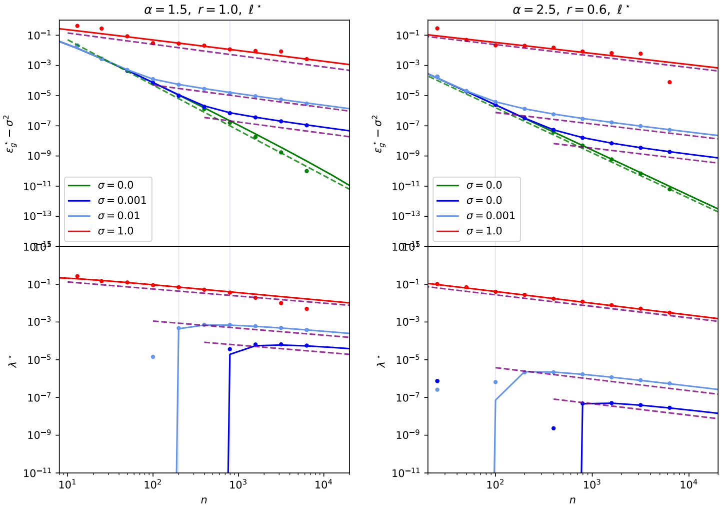


Figure 12: Kernel ridge regression on synthetic data sets with capacity α and source coefficient r. The regularization λ is chosen as the one minimizing the theoretical prediction for the excess generalization error, deduced from eq. (126) using the GCM package associated with Chapter 3. Solid lines correspond to the theoretical prediction of eq. (126). Points are simulations conducted with the python `scikit-learn KernelRidge` package (Pedregosa et al., 2011a), where the feature space dimension has been cut off to $p = 10^4$ for the simulations, and to 10^5 for the theoretical curves. In simulations, the best $\lambda^\star$ was determined using python `scikit-learn GridSearchCV` cross validation package (Pedregosa et al., 2011a). Note that because cross validation is not adapted to small training sets, a few discrepancies are observed for smaller n. Dashed lines represent the slopes predicted by theory, with the colors in correspondence to the regimes in Fig. 9, purple for the purple point in Fig. 9. Top: excess error. Bottom: optimal $\lambda^\star$. Note the noiseless case has $\lambda^* = 0$.

4.3 SKETCH OF THE DERIVATION

We provide in this section the main ideas underlying the derivation of the main results exposed in section 4.2 and summarized in Fig. 9.

Closed-form solution for Gaussian design — Closed-form, rigorous solution of the risk of ridge regression with Gaussian data of arbitrary covariance in the high-dimensional asymptotic regime have been studied in (Dobriban et al., 2018a) (Wu et al., 2020a; Richards et al., 2021). We shall use here the equivalent notations of Chapter 3, who have the advantage of having rigorous non-asymptotic rates guarantees. Using these characterizations as a starting point, we shall sketch how the crossover phenomena (121) (122)(123) and (124), which are the main contribution of this paper, can be derived. Within the framework of Chapter 3, with high-probability when n, p are large the excess prediction error is expressed as

$$\epsilon_g - \sigma^2 = \rho - 2m^\star + q^\star, \tag{125}$$

with $\rho = \theta^{\star\top}\Sigma\theta^\star$, and $(m^\star, q^\star)$ are the unique fixed-points of the following self-consistent equations:

$$\begin{cases} \hat{V} = \frac{\frac{n}{p}}{1+V} \\ \hat{q} = \frac{n}{p}\frac{\rho+q-2m+\sigma^2}{(1+V)^2} \\ q = p\sum_{k=1}^{p} \frac{\hat{q}\eta_k^2+\theta_k^{\star 2}\eta_k^2\hat{m}^2}{(n\lambda+p\hat{V}\eta_k)^2} \\ m = p\hat{V}\sum_{k=1}^{p} \frac{\theta_k^{\star 2}\eta_k^2}{n\lambda+p\hat{V}\eta_k} \\ V = \frac{1}{p}\sum_{k=1}^{p} \frac{p\eta_k}{n\lambda+p\hat{V}\eta_k} \end{cases} . \tag{126}$$

We recall the reader that $\lambda > 0$ is the regularisation strength and $\{\eta_k\}_{k=1}^{p}$ are the kernel eigenvalues. The next step is thus to insert the power-law decay (120) for the eigenvalues into (126), and to take the limit $n, p \to \infty$. We note, however, that this last step is not completely justified rigorously. Indeed, (Dobriban et al., 2018a) assumes $p/n = O(1)$ as $n, p \to \infty$ while here we first send $p \to \infty$ and then take the large n limit, thus working effectively with $p/n \to 0$. While the non-asymptotic rates guarantees of Chapter 3 are reassuring in this respect, a finer control of the limit would be needed for a fully rigorous justification. Nevertheless, we observed in our experiments that the agreement between theory and numerical simulations for the excess prediction error (117) is perfect (see Figs. 10, 11 and 12). In the large n limit, one can finally close the equation for the excess prediction error into

$$\epsilon_g - \sigma^2 = \frac{\sum_{k=1}^{\infty} \frac{k^{-1-2r\alpha}}{(1+nz^{-1}k^{-\alpha})^2}}{1-\frac{n}{z^2}\sum_{k=1}^{\infty}\frac{k^{-2\alpha}}{(1+nz^{-1}k^{-\alpha})^2}} + \sigma^2 \frac{\frac{n}{z^2}\sum_{k=1}^{\infty}\frac{k^{-2\alpha}}{(1+nz^{-1}k^{-\alpha})^2}}{1-\frac{n}{z^2}\sum_{k=1}^{\infty}\frac{k^{-2\alpha}}{(1+nz^{-1}k^{-\alpha})^2}}. \tag{127}$$

with z being a solution of

$$z \approx n\lambda + \left(\frac{z}{n}\right)^{1-\frac{1}{\alpha}} \int_{\left(\frac{z}{n}\right)^{1/\alpha}}^{\infty} \frac{\mathrm{d}x}{1+x^{\alpha}}. \tag{128}$$

We note that this equation was observed with heuristic arguments from statistical physics (using the non-rigorous cavity method) in (Canatar et al., 2021a).

The different regimes of excess generalization error rates discussed in Section 4.2 are derived from this self-consistent equation. Note that the excess error (127) decomposes over a sum of two contributions, respectively accounting for the sample variance and the noise-induced variance. In contrast to a typical bias-variance decomposition, the effect of the bias introduced in the task for non-vanishing λ is subsumed in both terms.

Derivation of the four regimes — If the second term in (128) dominates, then $z \sim n^{1-\alpha}$, which is self consistent if $\ell \geq \alpha$. This is the *effectively non-regularized regime*, where the regularization λ is not sensed, and corresponds to the green and red regimes in the phase diagram in Fig. 9. This scaling of z can then be used to estimate the asymptotic behaviour of the sample and noise induced variance in the decomposition on the excess error (127), yielding

$$\varepsilon_g - \sigma^2 = \mathcal{O}(n^{-2\alpha \min(r,1)}) + \sigma^2 \mathcal{O}(1), \tag{129}$$

which can be rewritten more compactly as (121). Therefore, for small sample sizes the sample variance drives the decay of the excess prediction error, while for larger samples sizes the noise variance dominates and causes the error to plateau. The crossover happens when both variance terms in (129) are balanced, around

$$n \sim \sigma^{-\frac{1}{\alpha \min(r,1)}}, \tag{130}$$

which corresponds to the vertical part of the crossover line in Fig. 9.

If the first term $n\lambda$ dominates in (128), then $z \sim n\lambda$, which is consistent provided that $\ell < \alpha$. This is the *effectively regularized regime* (blue, orange regions in Fig. 9). The two variances in (127) are found to asymptotically behave like

$$\varepsilon_g - \sigma^2 = \mathcal{O}(n^{-2\ell \min(r,1)}) + \sigma^2 \mathcal{O}(n^{\frac{\ell-\alpha}{\alpha}}), \tag{131}$$

which can be rewritten more compactly as (122). If the decay of the noise variance term $(\alpha - \ell)/\alpha$ is faster than the $2\ell \min(r,1)$ decay of the sample variance term, then the latter always dominates and no crossover is observed. This is the case for $\ell < \alpha/(1+2\alpha \min(r,1))$. If on the contrary the decay of the noise variance term is the slowest, then this term dominates at larger n, with a crossover when both terms in (131) are balanced, around

$$n \sim \sigma^{\frac{2}{1-\frac{\ell}{\alpha}(1+2\alpha\min(r,1))}} \tag{132}$$

Eqs. (129) and (131) are respectively equivalent to (121) and (122), and completely define the four regimes observable in Fig. 9. Equations (132) and (130) give the expression for the crossover line in Fig. 9.

Asymptotically optimal regularization — Determining the asymptotically optimal $\ell^\star$ is a matter of finding the ℓ leading to fastest excess error decay. We focus on the far left part and the far right part of the phase diagram Fig. 9.

In the $n \gg n_2^\star \approx \sigma^{-\max\left(2,\frac{1}{\alpha\min(r,1)}\right)}$ limit where the crossover line confounds itself with its $\ell = \alpha/(1+2\alpha\min(r,1))$ asymptot, this is tantamount to solving the maximization problem

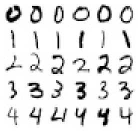
$$\ell^\star = \underset{\ell}{\operatorname{argmax}} \left(2\ell\min(r,1)\mathbb{1}_{0<\ell<\frac{\alpha}{(1+2\alpha\min(r,1))}} + \frac{\alpha-\ell}{\alpha}\mathbb{1}_{\frac{\alpha}{(1+2\alpha\min(r,1))}<\ell<\alpha} + 0\times\mathbb{1}_{\alpha<\ell}\right) \tag{133}$$

which admits as solution (124). In the $n \ll n_1^\star \approx \sigma^{-\frac{1}{\alpha\min(r,1)}}$ range, the maximization of the excess error decay reads

$$\ell^\star = \underset{\ell}{\operatorname{argmax}} \left(2\ell\min(r,1)\mathbb{1}_{0<\ell<\alpha} + 2\alpha\min(r,1)\mathbb{1}_{\alpha<\ell}\right), \tag{134}$$

and admits as solution (123).

4.4 ILLUSTRATION ON SIMPLE REAL DATA SETS



The MNIST dataset.

The Fashion MNIST dataset.

In this section we show that the derived decay rates can indeed be observed in real data sets with labels artificially corrupted by additive Gaussian noise. For real data, the decay model in eq. (120) is idealized, and in practice there is no firm reason to expect a power-law decay. However, we do find that for some of the data sets and kernels we investigated, the power law fit is reasonable and can be used to estimate the exponents α and r. For those cases, we compare the theoretically predicted exponents, eqs. (121), (122), (123) and (124) with the empirically measured learning curve, and obtain a very good agreement. We stress that the decay rates are not obtained by fitting the learning curves, but rather by fitting the exponents α and r from the data. We also observe the crossover from the noiseless (blue, green in Fig. 9) to the noisy (orange, red in Fig. 9) regime given by the theory.

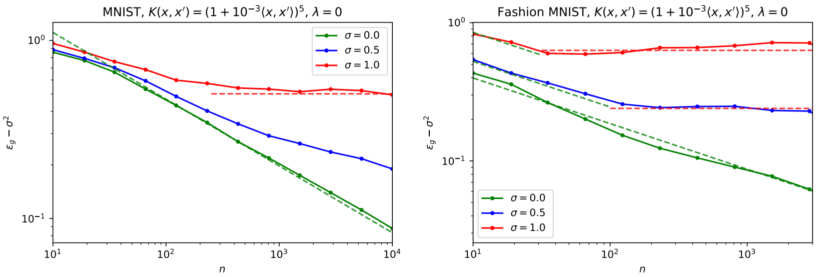


Figure 13: Excess error for MNIST odd versus even (above) and Fashion MNIST t-shirt versus coat (below) with labels corrupted by noise of variance σ^2. The kernel used is indicated in the title. Solid lines with points come from numerical experiments with zero regularization. Dashed lines are the slopes $-2\alpha r$ (as $r < 1$) or 0, predicted by the theory from the empirical values of α, r measured from the Gram matrix spectrum and the teacher for each data set, see Table 1. Colors of the dashed lines (green & red) indicate the regimes in Fig. 9.

Here we illustrate this with the learning curves for the following three data sets:

- MNIST even versus odd, a data set of 7×10^4 28×28 images of handwritten digits. Even (odd) digits were assigned label $y = 1 + \sigma\mathcal{N}(0,1)$ $(y = -1 + \sigma\mathcal{N}(0,1))$.
- Fashion MNIST t-shirts versus coats, a data set of 14702 28×28 images of clothes from an online shopping platform (Xiao et al., 2017). T-shirts (coats) were assigned label $y = 1 + \sigma\mathcal{N}(0,1)$ $(y = -1 + \sigma\mathcal{N}(0,1))$.
- Superconductivity (Hamidieh, 2018), a data set of 81 attributes of 21263 superconducting materials. The target y^μ corresponds to the critical temperature of the material, corrupted by additive Gaussian noise.

Learning curves are illustrated for a Radial Basis Function (RBF) kernel $K(x,x') = e^{-\frac{\gamma}{2}||x-x'||^2}$ with parameter $\gamma = 10^{-4}$ and a degree 5 polynomial kernel $K(x,x') = (1+\gamma\langle x,x'\rangle)^5$ with parameter $\gamma = 10^{-3}$. In Fig. 13 the regularization λ was set to 0, while in Fig. 14 λ was optimized for each sample size n using the python `scikit-learn` `GridSearchCV` package (Pedregosa et al., 2011a). KRR was carried out using the `scikit-learn` `KernelRidge` package (Pedregosa et al., 2011a). The values of α, r were independently measured for each data set, and the estimated values summarized in Table 1. From these values the theoretical decays (121), (123) and (124) were computed, and compared with the simulations with very good agreement. Since for real data the power-law form (120) does not exactly hold, the estimates for α, r slightly vary depending on how the power-law is fitted. Overall this variability does not hurt the good agreement with the simulated learning curves in Fig. 13 and 14.

When $\lambda = 0$ (Fig. 13) the characteristic plateau for large label noises is observed for both MNIST & Fashion MNIST. For polynomial kernel regression on Fashion MNIST (Fig. 13 right), the crossover between noiseless (slope

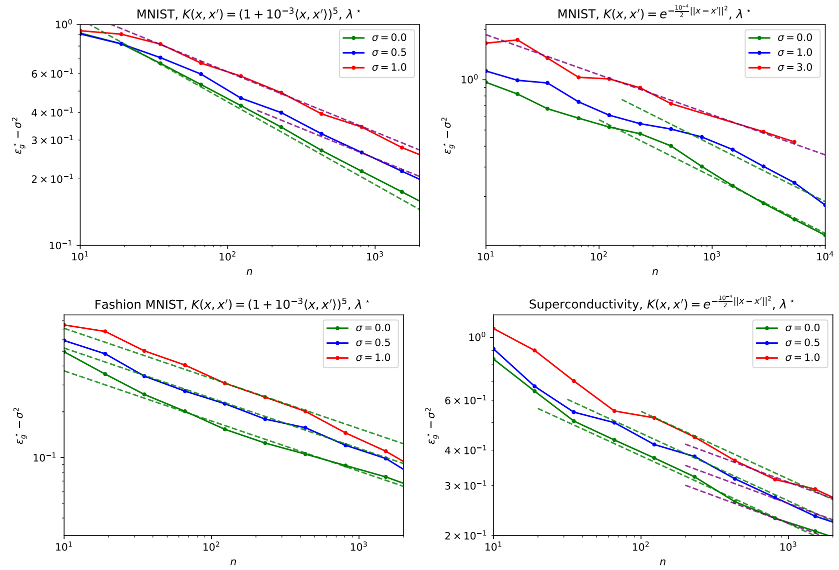


Figure 14: Excess error for MNIST odd versus even, and Fashion MNIST t-shirt versus coat, and the critical temperature regression. The kernel used is indicated in the title. Solid lines with dots come from numerical experiments with the regularization optimized using the python `scikit-learnGridSearchCV` package (Pedregosa et al., 2011a). Dashed lines are the slopes predicted by the theory, from the empirical values of α, r measured from the Gram matrix spectrum and the teacher for each data set, see Table 1. Colors of the dashed lines indicate the regime in Fig. 9.

Dataset	Kernel	α	r
Fashion MNIST	$K(x,x') = (1 + 10^{-3}\langle x,x'\rangle)^5$	1.3	0.13
MNIST	$K(x,x') = (1 + 10^{-3}\langle x,x'\rangle)^5$	1.2	0.15
MNIST	$K(x,x') = \exp(-10^{-4}\|\|x-x'\|\|^2/2)$	1.65	0.097
Superconductivity	$K(x,x') = \exp(-10^{-4}\|\|x-x'\|\|^2/2)$	2.7	0.046

Table 1: Values of the source and capacity coefficients (119) as estimated from the data sets.

$-2\alpha r$ as $r < 1$) and noisy (slope 0) regimes is apparent on the same learning curve at noise levels $\sigma = 0.5, 1$. For MNIST, the $\sigma = 0$ ($\sigma = 1$) curve is in the noiseless (noisy) regime for larger n, while at intermediary noise $\sigma = 0.5$, and small n for $\sigma = 1$, the curve is in the crossover regime between noiseless and noisy, consequently displaying in-between decay. Our results for the decays for $\sigma = 0$ agree with simulations for RBF regression on MNIST provided in (Spigler et al., 2020).

For optimal regularization $\lambda = \lambda^\star$ (Fig. 14), as the measured $r < 1$ we have exponents $-2r\alpha$ for the noiseless regime and $-2r\alpha/(1 + 2r\alpha)$ for noisy. Since the measured value of $2r\alpha$ is rather small the difference between the two rates is less prominent. Nevertheless, it seems that in our experiments the noisy regime is observed for polynomial and RBF kernels on MNIST and $\sigma = 0.5, 1$. For Superconductivity, the green and purple decay have close values and it is difficult to clearly identify the regime. For Fashion MNIST only the noiseless rate is observable in the considered noise range and sample range.

CONCLUSION

To conclude, we unify hitherto disparate lines of work, and give a comprehensive study of observable regimes, along the associated decay rates for the excess error, for kernel ridge regression with features having power-law co-variance spectrum. We show that the effect of the noise only kicks in at larger sample complexity, meaning, in particular, that the KRR transitions from a *noiseless* regime with fast error decay to a *noisy* regime with slower decay. This crossover is shown to happen for zero, decaying and optimized regularization, and is observed on a variety of real data sets corrupted with label noise.

5 SCALING LAWS FOR KERNEL CLASSIFICATION

A recent line of work (Hestness et al., 2017; Kaplan et al., 2020; Rosenfeld et al., 2019; Henighan et al., 2020) has empirically evidenced that the test error of neural networks often obey scaling laws with the number of parameters of the model, training set size, or other model parameters. Because of their implications in terms of relating performance and model size, these findings have been the object of sustained theoretical attention. Authors of (Sharma et al., 2022) relate the decay rate of the test loss with the number of parameters to the intrinsic dimension of the data. This idea is refined by (Bahri et al., 2021) for the case of regression tasks, building on the observation that in a number of settings, the covariance of the learnt features exhibits a power-law spectrum, whose rate of decay controls the scaling of the error. This investigation is actually very closely related to another large body of works. In fact, the study of a power-law features spectrum (and of a target function whose components in the corresponding eigenbasis also decay as a power-law) has a long history in the kernel literature, dating back to the seminal works of (Caponnetto et al., 2007; Caponnetto et al., 2005). The corresponding rates governing the power-laws are respectively known as the *capacity* and *source* coefficients, and the scaling of the test error with the training set size can be entirely characterized in terms of these two numbers. While the study of kernel ridge regression (Caponnetto et al., 2007; Caponnetto et al., 2005; Lin et al., 2018; Jun et al., 2019; Liu et al., 2020; Pillaud-Vivien et al., 2018; Berthier et al., 2020; Varre et al., 2021; Cui et al., 2021) therefore offers a rich viewpoint on the question of neural scaling laws with the training set size, little is so far known for kernel *classification*. Since ascertaining the test error decay under source and capacity conditions would automatically translate into neural scaling laws in *classification* tasks – similarly to (Bahri et al., 2021) for regression – this is a question of sizeable interest addressed in the present work.

RELATED WORKS

Neural scaling laws — A number of works (Hestness et al., 2017; Kaplan et al., 2020; Rosenfeld et al., 2019; Henighan et al., 2020) have provided empirical evidence of scaling laws in neural networks, with the number of parameters, training samples, compute, or other observables. These findings motivated theoretical investigations of the underlying mechanisms. Authors of (Sharma et al., 2022) show how the scaling of the test loss with the number of parameters is related to the intrinsic dimension of the data. This dimension is further tied in with the kernel spectrum by (Bahri et al., 2021), a work that leverages the kernel ridge regression viewpoint to translate, in turn, the decay of the spectrum to test error rates. Authors of (Maloney et al., 2022) similarly study a simple toy model where the power-law data is processed through a random features layer. Finally, (Hutter, 2021) investigate a toy model of scalar integer data in the context of classification, and ascertain the corresponding scaling law. Relating in classification settings the rate of decay of the kernel spectrum to the test error, like (Bahri et al., 2021) for regression, is still an open question.

Source and capacity conditions — The source and capacity conditions are standard regularity assumptions in the theoretical study of kernel methods, as they allow to subsume a large class of learning setups, c.f. (Marteau-Ferey et al., 2019; Pillaud-Vivien et al., 2018; Caponnetto et al., 2005; Caponnetto et al., 2007; Cui et al., 2021; Berthier et al., 2020). We also refer the interested reader to Chapter 4 for further discussion in the setting of KRR.

Kernel ridge regression — The error rates for kernel ridge regression have been extensively and rigorously characterized in terms of the source/capacity coefficients in the seminal work of (Caponnetto et al., 2005; Caponnetto et al., 2007), with a sizeable body of work being subsequently devoted thereto (Steinwart et al., 2009; Lin et al., 2018; Jun et al., 2019; Liu et al., 2020; Pillaud-Vivien et al., 2018; Berthier et al., 2020; Varre et al., 2021). In particular, in (Cui et al., 2021) it was shown that rates derived under worst-case assumptions (Lin et al., 2018; Jun et al., 2019; Caponnetto et al., 2005; Caponnetto et al., 2007; Bartlett et al., 2020b) are identical to the typical rates computed under the standard Gaussian design (Dobriban et al., 2018a; Dicker et al., 2016; Hsu et al., 2012) assumption. Crucially, it was observed that many real data-sets satisfy the source/capacity conditions, and display learning rates in very good agreement to the theoretical values (Cui et al., 2021).

Worst-case analyses for SVM — The worst-case bounds for SVM classification – see e.g. (Steinwart et al., 2008; Schölkopf et al., 2002) for general introductions thereto– are known from the seminal works of (Steinwart et al., 2008; Steinwart et al., 2007; Audibert et al., 2007). However, it is not known how tightly the corresponding rates hold for a given realistic data distributions, not even for synthetic Gaussian data. We show that, contrary to the case of ridge regression, for classification the worst case bounds are not tight for Gaussian data. This effectively hinders the ability to predict and understand the error rates for relevant classes of data-sets, and in particular the class of data described by source/capacity conditions, which as mentioned above includes many real data-sets (Cui et al., 2021), see Chapter 4. The key goal of this work is to fill this gap by leveraging the analysis of the learning curves for the GCM (Loureiro et al., 2021b) presented in Chapter 3 specified to data satisfying the capacity and source conditions.

MAIN CONTRIBUTION

In this work, we investigate the decay rate of the misclassification (generalization) error for noiseless kernel classification, under the Gaussian design and source/capacity regularity assumptions with capacity coefficient α and source coefficient r. Building on the analytic framework of (Loureiro et al., 2021b), we consider the two most widely used classifiers: margin-maximizing SVMs and ridge classifiers. We derive in Section 5.2 the error rate (describing the decay of the prediction error with the number of samples) for margin-maximizing SVM :

$$\epsilon_g^{\text{SVM}} \sim n^{-\frac{\alpha \min(r,\frac{1}{2})}{1+\alpha \min(r,\frac{1}{2})}}.$$

As a consequence, we conclude that the worst-case rates (Steinwart et al., 2007; Steinwart et al., 2008; Audibert et al., 2007) are indeed loose and fail to describe this class of data. This fact alone is not at all surprising. However, it becomes remarkable in the light of the fact that for ridge regression, as discussed in Chapter 4, the worst case bounds and the typical case rates do agree (Cui et al., 2021).

We contrast the SVM rate with the rate for optimally regularized ridge classification, which we establish in Section 5.3 to be

$$\epsilon_g^{\text{ridge}} \sim n^{-\frac{\alpha \min(r,1)}{1+2\alpha \min(r,1)}}.$$

We argue in the light of these findings that the SVM always displays faster rates than the ridge classifier for the classification task considered.

Finally, we observe that some real data-sets fall in the same universality class as the considered setting, in the sense that, as illustrated in Section 5.4, their error rates are in very good agreement with the ones above. This work is thus a key step for theoretically predicting the error rates of kernel classification for a broad range of real data-sets.

5.1 SETTING

5.1.1 KERNEL CLASSIFICATION

Consider a data-set $\mathscr{D} = \{(x^\mu, y^\mu)\}_{\mu=1}^n$ with n independent samples from a probability measure ν on $\mathscr{X} \times \{-1,+1\}$, with $\mathscr{X} \subset \mathbb{R}^d$. We will assume that the labels can be expressed as

$$y^\mu = \mathrm{sign}(f^\star(x^\mu)) \tag{135}$$

for some non-stochastic target function $f^\star : \mathscr{X} \to \mathbb{R}$. Note that the *noiseless* setting considered here is out of the validity domain of many worst case analyses, whose bounds become void without noise (Audibert et al., 2007), whereas a number of real learning settings are well described by a noiseless setup, see section 5.4. Learning to classify $\mathscr{D}$ in the direct space $\mathscr{X}$ for a *linear* $f^\star$ has been the object of extensive studies. In the present work, we focus on the case where $f^\star$ more generically belongs to the space of square-integrable functions $L^2(\mathscr{X})$. To classify $\mathscr{D}$, a natural method is then to perform *kernel classification* in a p- dimensional RKHS $\mathscr{H}$ associated to a kernel K, by minimizing the regularized empirical risk:

$$\hat{\mathscr{R}}_n(f) = \frac{1}{n}\sum_{\mu=1}^{n} \ell(f(x^\mu), y^\mu) + \lambda ||f||_{\mathscr{H}}^2. \tag{136}$$

The function $\ell(\cdot)$ is a loss function and λ is the strength of the ℓ_2 regularization term. In this paper we shall more specifically consider the losses $\ell(z,y) = \max(0, 1-yz)$ (hinge classification) and $\ell(z,y) = (y-z)^2$ (ridge classification), and the case of an infinite dimensional RKHS ($p = \infty$). The risk (136) admits a dual rewriting in terms of a standard parametric risk. To see this, diagonalize K in an orthogonal basis of *kernel features* $\{\psi_k(\cdot)\}_{k=1}^p$ of $L^2(\mathscr{X})$, with corresponding eigenvalues $\{\omega_k\}_{k=1}^p$:

$$\int_{\mathscr{X}} \nu(\mathrm{d}x') K(x,x') \psi_k(x') = \omega_k \psi_k(x). \tag{137}$$

It is convenient to normalize the eigenfunctions to

$$\int_{\mathscr{X}} \nu(\mathrm{d}x) \psi_k(x)^2 = \omega_k, \tag{138}$$

so that the kernel K can be rewritten in simple scalar product form $K(x,x') = \psi(x)^\top \psi(x')$, where we named $\psi(x)$ the p-dimensional vector with components $\{\psi_k(x)\}_{k=1}^p$.

Furthermore, note that the covariance Σ of the data in feature space with this choice of feature map is simply diagonal

$$\Sigma = \mathbb{E}_{x\sim\nu}(\psi(x)\psi(x)^\top) = \mathrm{diag}(\omega_1, \cdots, \omega_p). \tag{139}$$

Any function $f \in \mathscr{H}$ can then be expressed as $f(\cdot) = w^\top \psi(\cdot)$ for a vector w with square summable components. Using this parametrization, the risk (136) can be rewritten as

$$\hat{\mathscr{R}}_n(w) = \frac{1}{n} \sum_{\mu=1}^{n} \ell(w^\top \psi(x^\mu), y^\mu) + \lambda w^\top w. \tag{140}$$

Throughout this manuscript we will refer to the components of the target function in the features basis as the *teacher* $\theta^\star$, so that

$$f^\star(\cdot) = \theta^{\star\top} \psi(\cdot).$$

Note that any $f^\star \in L^2(\mathscr{X})$ can be formally written in this form with a certain $\theta^\star$ (allowing for non square-summable components if $f^\star \in L^2(\mathscr{X}) \setminus \mathscr{H}$). Similarly, the minimizer $\hat{w}$ of the parametric risk (140) is related to the argmin $\hat{f}$ of (136) by $\hat{f}(\cdot) = \hat{w}^\top \psi(\cdot)$, and will be referred to as the *estimator* in the following. We make two further assumptions : first, we work under the *Gaussian design*, and assume the features $\psi(x)$ to follow a Gaussian distribution with covariance Σ, i.e. $\psi(x) \sim \mathscr{N}(0, \Sigma)$. Note that this assumption might appear constraining , as the distribution of the data in feature space strongly depends on its distribution in the original space, and the feature map associated to the kernel. In fact, for a large class of data distributions and standard kernels, the Gaussian design assumption does not hold. However, rates derived under Gaussian design can hold more broadly. For instance, the rates established in Chapter 4 under Gaussian design were later proven by (Jin et al., 2021) under weaker conditions on the features. We will moreover discuss in Section 5.4 several settings in which our theoretical rates are in good agreement with rates observed for real data.

Second, as in Chapter 4, we assume that the *regularization strength* λ *decays as a power-law* of the number of samples n with an exponent ℓ: $\lambda = n^{-\ell}$. Note that this form of regularization is natural, since the need for regularizing is lesser for larger training sets. Furthermore, this allows to investigate the classical question of the *asymptotically optimal regularization* (Caponnetto et al., 2005; Caponnetto et al., 2007; Cui et al., 2021), i.e. the decay ℓ of the regularization yielding fastest decrease of the prediction error.

5.1.2 SOURCE AND CAPACITY CONDITIONS

Under the above assumptions of Gaussian design with features covariance Σ and existence of a teacher $\theta^\star$ that generates the labels using eq. (135) we can now study the error rates. In statistical learning theory one often uses the *source and capacity conditions*, which assume the existence of two parameters

$\alpha > 1, r \geq 0$ (hereafter referred to as the *capacity coefficient* and the *source coefficient* respectively) so that

$$\mathrm{tr}\Sigma^{\frac{1}{\alpha}} < \infty, \qquad \theta^{\star\top}\Sigma^{1-2r}\theta^{\star} < \infty. \tag{141}$$

As in (Dobriban et al., 2018a; Spigler et al., 2020; Bordelon et al., 2020; Berthier et al., 2020; Cui et al., 2021), we will consider the particular case where both the spectrum of Σ and the teacher components $\theta_k^\star$ have exactly a power-law form satisfying the limiting source/capacity conditions (141):

$$\omega_k = k^{-\alpha}, \qquad \theta_k^\star = k^{-\frac{1+\alpha(2r-1)}{2}}. \tag{142}$$

The power-law forms (142) have been empirically found in (Cui et al., 2021) in the context of kernel regression to be a reasonable approximation for a number of real data-sets including MNIST (LeCun et al., 1998a) and Fashion MNIST (Xiao et al., 2017) and a number of standard kernels such as polynomial kernels and radial basis functions. Similar observations were also made in the present work and are discussed in section 5.4.

We remind from Chapter 4 that the capacity parameter α and source parameter r capture the complexity of the data-set in feature space – i.e. after the data is transformed through the kernel feature map into $\{\psi(x^\mu), y^\mu\}_{\mu=1}^n$. A large α, for example, signals that the spectrum of the data covariance Σ displays a fast decay, implying that the data effectively lies along a small number of directions, and has a low effective dimension. Conversely, a small capacity α means that the data is effectively large dimensional, and therefore a priori harder to learn. Similarly, a large r signals a good alignment of the teacher $\theta^\star$ with the main directions of the data, and a priori an easier learning task. In terms of the target function $f^\star$, larger r correspond to smoother $f^\star$. Note that $r > 1/2$ implies that $f^\star \in \mathcal{H}$, while $r \leq 1/2$ implies $f^\star \in L^2(\mathcal{X}) \setminus \mathcal{H}$. Finally, note that while (Marteau-Ferey et al., 2019) suggested an alternative definition for the source and capacity coefficients in the case of non-square loss functions, their redefinition is not directly applicable for the hinge loss.

5.1.3 MISCLASSIFICATION ERROR

The performance of learning the data-set $\mathcal{D}$ using kernel classification (140) is quantified by the misclassification (generalization) error

$$\varepsilon_g = \frac{1}{2} - \frac{1}{2}\mathbb{E}_{\mathcal{D}}\mathbb{E}_{x,y\sim\nu}\left(y\ \mathrm{sign}(\hat{w}^\top\psi(x))\right), \tag{143}$$

where $\hat{w}$ is the minimizer of the risk (140). The error (143) corresponds to the probability for the predicted label $\mathrm{sign}(\hat{w}^\top\psi(x))$ of a test sample x to be incorrect. The rate at which the error (143) decays with the number of samples n in $\mathcal{D}$ depends on the complexity of the data-set, as captured by the source and capacity coefficients α, r eq. (142). To compute this rate, we build upon the work of (Loureiro et al., 2021b) who, following a long body of work in the statistical physics literature (Mézard et al., 1987; Dietrich et al., 1999c; Engel et al., 2001; Mézard et al., 2009; Bordelon et al., 2020; Advani et al., 2020), provided and proved a mathematically rigorous closed form asymptotic characterization of the misclassification error as

$$\epsilon_g = \frac{1}{\pi}\arccos\left(\sqrt{\eta}\right), \qquad \eta = \frac{m^2}{\rho q}, \tag{144}$$

where ρ is the squared $L^2(\mathscr{X})$ norm of the target function $f^\star$, i.e. $\rho = \int_{\mathscr{X}} \nu(\mathrm{d}x) f^\star(x)^2 = \theta^{\star\top}\Sigma\theta^\star$, and m, q are the solution of a set of self-consistent equations, which are later detailed and analyzed in Section 5.2 for margin-maximizing SVMs and section 5.3 for ridge classifiers. The order parameters m, q are known as the *magnetization* and the *self-overlap* in statistical physics and respectively correspond to the target/estimator and estimator/estimator $L^2(\mathscr{X})$ correlations:

$$\begin{aligned} m &= \mathbb{E}_{\mathscr{D}} \int_{\mathscr{X}} \nu(\mathrm{d}x) f^\star(x)\hat{f}(x) = \mathbb{E}_{\mathscr{D}}\left(\hat{w}^\top \Sigma \theta^\star\right), \\ q &= \mathbb{E}_{\mathscr{D}} \int_{\mathscr{X}} \nu(\mathrm{d}x) \hat{f}(x)^2 = \mathbb{E}_{\mathscr{D}}\left(\hat{w}^\top \Sigma \hat{w}\right). \end{aligned} \tag{145}$$

It follows from these interpretations that η has to be thought of as the *cosine-similarity* between the teacher $\theta^\star$ and the estimator $\hat{w}$, with perfect alignment ($\eta = 1$) resulting in minimal error $\epsilon_g = 0$ from (143).

Note that while this characterization has formally been proven in (Loureiro et al., 2021b) in the asymptotic proportional $n, p \to \infty, {}^n/_p = \mathscr{O}(1)$ limit, we are presently using it in the $n \ll p = \infty$ limit, thereby effectively working at ${}^n/_p = 0^+$. The non-asymptotic rate guarantees of (Loureiro et al., 2021b) are nevertheless encouraging in this respect, although a finer control of the limit would be warranted to put the present analysis on fully rigorous grounds. Further, (Cui et al., 2021) also build on (Loureiro et al., 2021b) in the ${}^n/_p = 0^+$ limit, and display solid numerics-backed results, later rigorously proven by (Jin et al., 2021). We thus conjecture that this limit can be taken as well safely in our case. Finally, we mention that a recent line of works (Li et al., 2021c; Ariosto et al., 2022a; Seroussi et al., 2023a; Cui et al., 2023a) has explored the connections between kernel regression and Bayesian learning for networks in the ${}^n/_p = \mathscr{O}(1)$ limit, where p is in this case the width of the network. While the high-dimensional limit is indeed related to the one originally discussed in (Loureiro et al., 2021b), which we relax here to ${}^n/_p = 0^+$, the main object of (Li et al., 2021c; Ariosto et al., 2022a; Seroussi et al., 2023a) was not to study kernel regression per se, but to show how observables in Bayesian regression could be expressed in terms of well-chosen kernels. In the present work, we focus on analyzing kernel classification in the ${}^n/_p = 0^+$ regime.

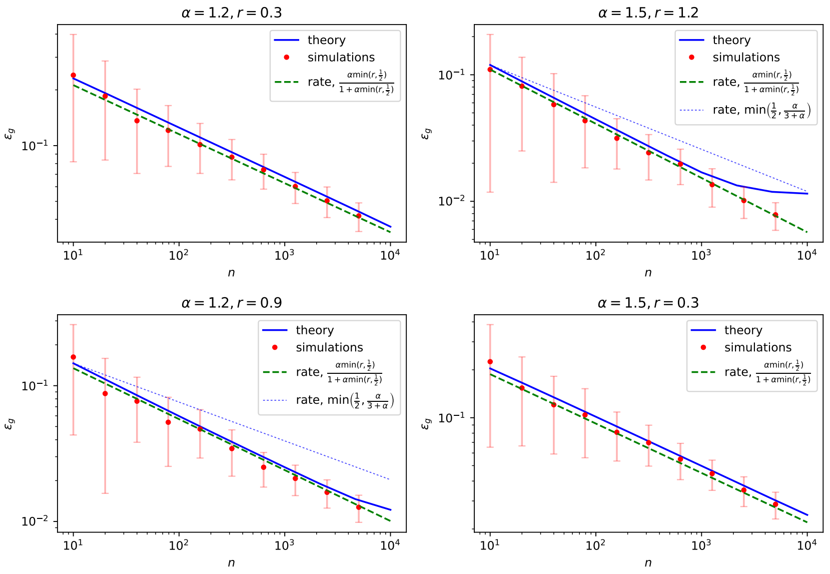


Figure 15: Misclassification error ε_g for max-margin classification on synthetic Gaussian features, as specified in (142), for different source/capacity coefficients α, r. In blue, the solution of the closed set of eqs. (147) used in the characterization (143) for the misclassification error, using the GCM package (Loureiro et al., 2021b). The dimension p was cut-off at 10^4. Red dots corresponds to simulations using the `scikit-learn` SVC(Pedregosa et al., 2011b) package run for vanishing regularization $\lambda = 10^{-4}$ and averaged over 40 instances, for $p = 10^4$. The green dashed line indicates the power-law rate (149) derived in this work. The light blue dotted line indicates the classical worst-case $\min\left(1/2, \alpha/(3+\alpha)\right)$ rate for SVM classification (Theorem 2.3 in (Steinwart et al., 2008)) in the cases where the theorem readily applies ($r > 1/2$).

5.2 MAX-MARGIN CLASSIFICATION

5.2.1 SELF-CONSISTENT EQUATIONS

In this section we study regression using Support Vector Machines. The risk (140) then reads for the hinge loss

$$\hat{\mathscr{R}}_n(w) = \frac{1}{n}\sum_{\mu=1}^{n} \max\left(0, 1 - y^\mu w^\top \psi(x^\mu)\right) + \lambda w^\top w. \tag{146}$$

In the following, we shall focus more specifically on the max-margin limit with $\lambda = 0^+$. In fact, zero regularization is asymptotically optimal for the data following eq. (142) when the target function is characterized by a source $r \leq 1/2$, i.e. $f^\star \in L^2(\mathscr{X}) \setminus \mathscr{H}$. We heuristically expect margin maximization to be *a fortiori* optimal also for easier and smoother teachers $f^\star \in \mathscr{H}$. For the risk (146) at $\lambda = 0^+$, the self-consistent equations defining m, q in (145) read

$$\begin{cases}
\rho = & \sum\limits_{k=1}^{\infty} \theta_k^{\star 2} \omega_k, \\
m = & \hat{r}_1 \frac{n}{z} \sum\limits_{k=1}^{\infty} \frac{\omega_k^2 \theta_k^{\star 2}}{1+\frac{n}{z}\omega_k}, \\
q = & \hat{r}_1^2 \frac{n^2}{z^2} \sum\limits_{k=1}^{\infty} \frac{\theta_k^{\star 2}\omega_k^3}{(1+\frac{n}{z}\omega_k)^2} \\
 & +\hat{r}_2 \frac{n}{z^2} \sum\limits_{k=1}^{\infty} \frac{\omega_k^2}{(1+\frac{n}{z}\omega_k)^2}
\end{cases},$$

$$\begin{cases}
\hat{r}_1 = & \frac{1}{2\pi\sqrt{\rho}} \frac{\left(\sqrt{2\pi}(1+\mathrm{erf}(\frac{1}{\sqrt{2q(1-\eta)}}))+2e^{-\frac{1}{2q(1-\eta)}}\sqrt{q(1-\eta)}\right)}{\int_{-\infty}^{\frac{1}{\sqrt{q}}} dx \frac{e^{-\frac{1}{2}x^2}}{\sqrt{2\pi}}\left[1+\mathrm{erf}(\sqrt{\frac{\eta}{2(1-\eta)}}x)\right]}, \\
\hat{r}_2 = & \frac{\int_{-\infty}^{\frac{1}{\sqrt{q}}} dx \frac{e^{-\frac{1}{2}x^2}}{\sqrt{2\pi}}\left[1+\mathrm{erf}(\sqrt{\frac{\eta}{2(1-\eta)}}x)\right](1-\sqrt{q}x)^2}{\left(\int_{-\infty}^{\frac{1}{\sqrt{q}}} dx \frac{e^{-\frac{1}{2}x^2}}{\sqrt{2\pi}}\left[1+\mathrm{erf}(\sqrt{\frac{\eta}{2(1-\eta)}}x)\right]\right)^2}, \\
z = & \frac{\frac{z}{n}\sum\limits_{k=1}^{\infty}\frac{\omega_k}{\frac{z}{n}+\omega_k}}{\int_{-\infty}^{\frac{1}{\sqrt{q}}} dx \frac{e^{-\frac{1}{2}x^2}}{\sqrt{2\pi}}\left[1+\mathrm{erf}(\sqrt{\frac{\eta}{2(1-\eta)}}x)\right]}.
\end{cases} \tag{147}$$

Here $\hat{r}_1$ should be thought of as the ratio between the norms of the estimator $\hat{w}$ and the teacher $\theta^\star$, while z can be loosely interpreted as an effective regularization.

5.2.2 DECAY RATES FOR MAX-MARGIN

From the investigation of the eqs. (147), the following scalings are found to hold between the order parameters:

$$m \sim \sqrt{q} \sim \hat{r}_1 \sim n\left(\frac{z}{n}\right)^{\frac{1}{\alpha}} \sim n^{\frac{\alpha\min(r,\frac{1}{2})}{1+\alpha\min(r,\frac{1}{2})}}. \tag{148}$$

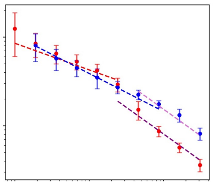
Note that the mutual scaling between m, q also follows intuitively from the interpretation of these order parameters – as the overlap of $\hat{w}$ with the ground truth and itself respectively – see the discussion around eqs. (145) and (147). Since the width of the margin is generically expected to shrink with the number of samples (as more training data are likely to be sampled close to the separating hyperplane), the increase of the norm of $\hat{w}$ (as captured by $q, \hat{r}_1$) with n is also intuitive. Finally, an analysis of the subleading corrections to m and q, leads to

$$\epsilon_g \sim n^{-\frac{\alpha\min(r,\frac{1}{2})}{1+\alpha\min(r,\frac{1}{2})}}. \tag{149}$$

The error rate (149) stands in very good agreement with numerical simulations on artificial Gaussian features generated using the model specification (142), see Fig. 15. Two observations can further be made on the decay rate (149). First, the rate is as expected an increasing function of α (low-dimensionality of the features) and r (smoothness of the target $f^\star$). Second, for a source $r > 1/2$ (corresponding to a target $f^\star \in \mathscr{H}$), the rate saturates, suggesting that all functions in $\mathscr{H}$ are all equally easy to classify, while for rougher target $f^\star \in L^2(\mathscr{X}) \setminus \mathscr{H}$ the specific roughness of the target function, as captured by its source coefficient r, matters and conditions the rate of decay of the error.

Finally, we refer to the Appendix of (Cui et al., 2023c) for a discussion the more general case where the label distribution (135) includes data noise, and show that the rates display a crossover from the noiseless value (149) to a noisy value, much like what was reported for kernel ridge regression (Cui et al., 2021).

5.2.3 COMPARISON TO CLASSICAL RATES



For SVM and ridge classification on MNIST, two rates are observed in succession with the sample complexity, like for KRR (see Chapter 4). The noisy rate seems to be $\alpha/1+\alpha$.

To the best of the authors' knowledge, there currently exists little work addressing the error rates for datasets satisfying source and capacity conditions (141). The closest result is the worst-case bound of (Steinwart et al., 2008) for SVM classification, which can be adapted to the present setting provided $f^\star \in \mathscr{H}$ ($r > 1/2$). This yields an upper bound of $\min\left(1/2, \alpha/(3+\alpha)\right)$ for the error rate for max-margin classification, which is always *slower* than (149). This rate (Steinwart et al., 2008) is plotted for comparison in Fig. 15 against numerical simulations and is visibly off, failing to capture the learning curves. It is to be expected that the worst case rates will be loose when compared to rate that assume a specific data distribution. What makes our result interesting is the comparison with the more commonly studied ridge regression where, as discussed already in the introduction, the worst case rates actually match those derived for Gaussian data, see (Cui et al., 2021) and the corresponding discussion in Chapter 4.

Importantly, the rates from (Steinwart et al., 2008) only hold for capacity $r > 1/2$, while real datasets are typically characterized by sources $r < 1/2$ (see for instance Fig. 18). The present work therefore fills an important gap in the literature in providing rates (149) which accurately capture the learning curves of datasets satisfying source and capacity conditions. Also note that while (Vecchia et al., 2021) report $\alpha/(1+\alpha)$ rates under Gaussianity assumptions, they rely on very stringent assumptions which are too strong and unfulfilled in our setting.

5.3 RIDGE CLASSIFICATION

5.3.1 SELF-CONSISTENT EQUATIONS

Another standard classification method is the ridge classifier, which corresponds to minimizing

$$\hat{\mathscr{R}}_n(w) = \frac{1}{n}\sum_{\mu=1}^{n}\left(y^\mu - w^\top \psi(x^\mu)\right)^2 + \lambda w^\top w. \tag{150}$$

As previously discussed in section 5.1, we consider a decaying regularization $\lambda = n^{-\ell}$. The self-consistent equations characterizing the quantities (q,m), read for the ridge risk (150)

$$\begin{cases} \rho = \sum_{k=1}^{\infty} \theta_k^{\star 2}\omega_k, \\ m = \sqrt{\frac{2}{\pi\rho}}\frac{n}{z}\sum_{k=1}^{\infty}\frac{\omega_k^2\theta_k^{\star 2}}{1+\frac{n}{z}\omega_k}, \\ q = \frac{n^2}{z^2}\sum_{k=1}^{\infty}\frac{\frac{2}{\pi\rho}\theta_k^{\star 2}\omega_k^3 + \frac{1+q-2m\sqrt{\frac{2}{\pi\rho}}}{n}\omega_k^2}{\left(1+\frac{n}{z}\omega_k\right)^2}, \\ z = n\lambda + \frac{z}{n}\sum_{k=1}^{\infty}\frac{\lambda_k}{\lambda_k+\frac{z}{n}}. \end{cases} \tag{151}$$

Like (147), eqs. (151) have been formally proven in the proportional $n,p \to \infty$, $n/p = \mathscr{O}(1)$ limit in (Loureiro et al., 2021b), but are expected to hold also in the present $n \ll p = \infty$ setting (Cui et al., 2021; Jin et al., 2021). Note that comparing to (147), eqs. (151) correspond to a constant student/teacher norm ratio $\hat{r}_1 = 2/(\pi\rho)$ and to a simple $\hat{r}_2 = 1 + q - 2m\sqrt{2/(\pi\rho)}$. $\hat{r}_2$ moreover admits a very intuitive interpretation as the prediction mean squared error (MSE) between the true label $y = \mathrm{sign}(\theta^{\star\top}\psi(x))$ and the pre-activation linear predictor $\hat{w}^\top\psi(x)$, i.e. $\hat{r}_2 = \mathbb{E}_{\psi(x)}\left(\mathrm{sign}\left(\theta^{\star\top}\psi(x)\right) - \hat{w}^\top\psi(x)\right)^2$.

5.3.2 DECAY RATES FOR RIDGE CLASSIFICATION

Similarly to (Bordelon et al., 2020; Cui et al., 2021), an analysis of the eqs. (151) reveals that, depending on how the rate of decay ℓ of the regularization compares to the capacity α, two regimes (called *effectively regularized* and *effectively un-regularized* in Chapter 4 in the context of KRR) can be found:

Effectively regularized regime – $\ell \leq \alpha$. In this regime, an analysis of the corrections to the self-overlap q and magnetization m shows that the misclassification error scales like

$$\epsilon_g \sim n^{-\frac{1}{2}\min\left(2\ell\min(r,1), \frac{\alpha-\ell}{\alpha}\right)}. \tag{152}$$

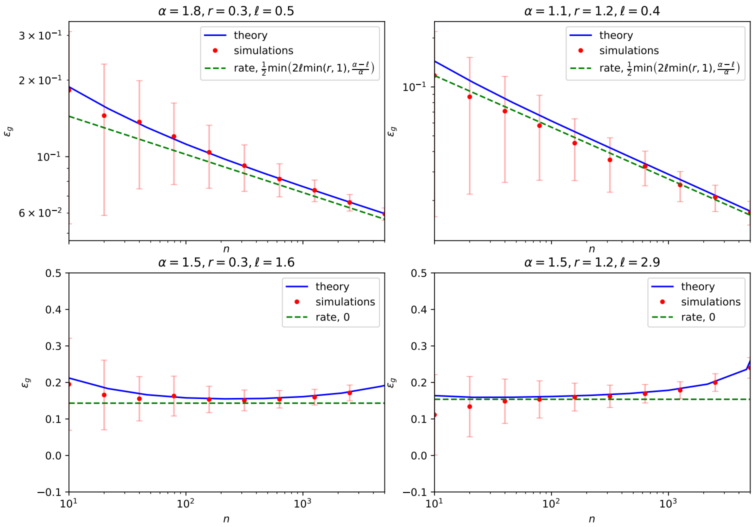


Figure 16: Misclassification error ε_g for ridge classification on synthetic Gaussian features, as specified in (142), for different source/capacity coefficients α, r, in the effectively regularized regime $\ell \leq \alpha$ (top) and unregularized regime $\ell > \alpha$ (bottom). In blue, the solution of the eqs. (147) used in the characterization (143) for the misclassification error, using the GCM package (Loureiro et al., 2021b). The dimension p was cut-off at 10^4. Red dots corresponds to simulations averaged over 40 instances, for $p = 10^4$. The green dashed lines indicate the power-laws (152) (top) and (153) (bottom) derived in this work. The slight increase of the error for larger n in the unregularized regime (bottom) is due to finite size effects of the simulations ran at $p = 10^4 < \infty$. Physically, it corresponds to the onset of the ascent preceding the second descent that is present for finite p.

The rate (152) compares very well to numerical simulations, see Fig. 16. Note that the saturation for ridge happens for $r = 1$, rather than $r = 1/2$ as for max-margin classification (see discussion in section 5.2): very smooth targets $f^\star$ characterized by a source $r \geq 1$ are all equally easily classified by ridge. For rougher teachers $f^\star$ characterized by $r \leq 1$ however, the rate of decay of the error (152) depends on the specific roughness of the target, even if, in contrast to max-margin, the latter belongs to $\mathscr{H}$ ($r > 1/2$). Two important observations should further be made on the rates (152):

- If the regularization remains small (fast decay $\alpha > \ell > \alpha/(1+2\alpha\min(r,1))$), the decay (152) is determined only by the data capacity α, while the source r plays no role. As a matter of fact, with insufficient regularization, the limiting factor to the learning is the tendency to overfit, which depends on the effective dimension of the data as captured by the capacity α.

- For larger regularizations (slow decays $\ell < \alpha/(1+2\alpha\min(r,1))$), the limiting factor becomes the complexity of the teacher $\theta^\star$, as captured by the source r.

Effectively un-regularized regime – $\ell > \alpha$. The error plateaus and stays of order 1:

$$\varepsilon_g = \mathscr{O}(1). \tag{153}$$

This pleateau actually corresponds to the first plateau in a double descent curve, with the second descent never happening since $p = \infty$. Intuitively, this phenomenon is attributable to the ridge classifier overfitting the labels using the small-variance directions of the data (142).

Interestingly, all the rates (153) and (152) correspond exactly (up to a factor $1/2$) to those reported in (Cui et al., 2021) for the MSE of ridge regression, where they are respectively called the *red*, *blue* and *orange* exponents. Notably, the plateau (153) at low regularizations and the $(\alpha-\ell)/\alpha$ exponent in (152) only appeared in (Cui et al., 2021) for noisy cases in which the labels are corrupted by an additive noise. The fact that they hold in the present *noiseless* study very temptingly suggests that model mis-specification (trying to interpolate binary labels using a linear model) effectively plays the role of a large noise.

5.3.3 OPTIMAL RATES

Optimally regularized ridge classification In practice, the strength of the regularization λ is a tunable parameter. A natural question to ask is then the one of the *asymptotically optimal* regularization, that is the regularization decay rate $\ell^\star$ leading to fastest decay rates for the misclassification error. From the expressions of (152) (which hold provided $\ell < \alpha$) and (153) (which holds provided $\ell > \alpha$), the value of ℓ maximizing the error rate is found to be

$$\ell^\star = \frac{\alpha}{1+2\alpha\min(r,1)}, \tag{154}$$

and the corresponding error rate for $\varepsilon_g^\star = \varepsilon_g(\lambda^\star = n^{-\ell^\star})$ is

$$\varepsilon_g^\star \sim n^{-\frac{\alpha\min(r,1)}{1+2\alpha\min(r,1)}}, \tag{155}$$

see the red dashed lines in Fig. 17. Coincidentally, the optimal rate (155) is up to a factor $1/2$ identical to the classical optimal rate known for the rather distinct problem of the MSE of kernel ridge regression on noisy data (Caponnetto et al., 2005; Caponnetto et al., 2007). Like the max-margin exponent (149), the optimal error rate for ridge (155) is an increasing function of both the capacity α and the source r, i.e. of the easiness of the learning task. Note that in contrast to max-margin classification which is insensitive to the specifics of the target function $f^\star$, provided it is in $\mathscr{H}$, ridge is sensitive to the source (smoothness) r of $f^\star$ up to $r = 1$.

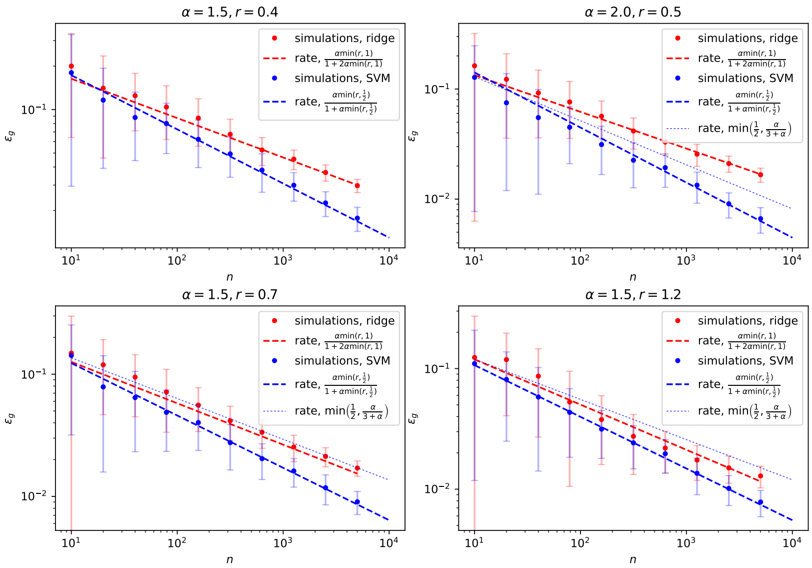


Figure 17: (red) Misclassification error ε_g for ridge classification on synthetic Gaussian features, as specified in (142), for different source/capacity coefficients α, r, for optimal regularization $\lambda^\star$. The dimension p was cut-off at 10^4 and the regularization λ numerically tuned to minimize the error ε_g for avery n. Red dots correspond to simulations averaged over 40 instances, for $p = 10^4$. Optimization over λ was performed using cross validation, with the help of the python `scikit-learn GridSearchCV` package. The red dashed line represents the power-law (155). In blue, the learning curves for max-margin for the same data-set are plotted for reference, along the corresponding power law (149) (blue) and the loose classical $\min\left(1/2, \alpha/(3+\alpha)\right)$ rate (Steinwart et al., 2008) (light blue), see Section 5.2.

Comparison to max-margin A comparison of the max-margin rate $a_{\text{SVM}} = \alpha\min(r, \frac{1}{2})/(1+\alpha\min(r, \frac{1}{2}))$ (149) and the optimal ridge exponent $a_r = \alpha\min(r,1)/(1+2\alpha\min(r,1))$ (155) reveals that for any $\alpha > 1, r \geq 0$, $a_{\text{SVM}} - a_r > 0$. In other words, the margin-maximizing SVM displays faster rates than the ridge classifier for the class of data studied (142), see Fig. 17.

We finally briefly comment on support vector proliferation. (Muthukumar et al., 2021; Hsu et al., 2020; Ardeshir et al., 2021) showed that in some settings almost *every* training sample in $\mathscr{D}$ becomes a support vector for the SVM. In such settings, the estimators $\hat{w}$ (and hence the error ε_g) consequently coincide for the ridge classifier and the margin-maximizing SVM. In the present setting however, the result $a_{\text{SVM}} - a_r > 0$ establishes that for features with a power-law decaying spectrum (142), there is *no* such support vector proliferation. Note that this result does not follow immediately from Theorem 3 in (Ardeshir et al., 2021). In fact, the spiked covariance (142), with only a small number of important (large variance) directions and a tail of unimportant (low-variance) directions does effectively not offer enough overparametrization (Bartlett et al., 2020b; Hsu et al., 2020) for support vector proliferation, and the support consists only of the subset of the training set with weakest alignment with the spike.

5.4 REMARKS FOR REAL DATA-SETS

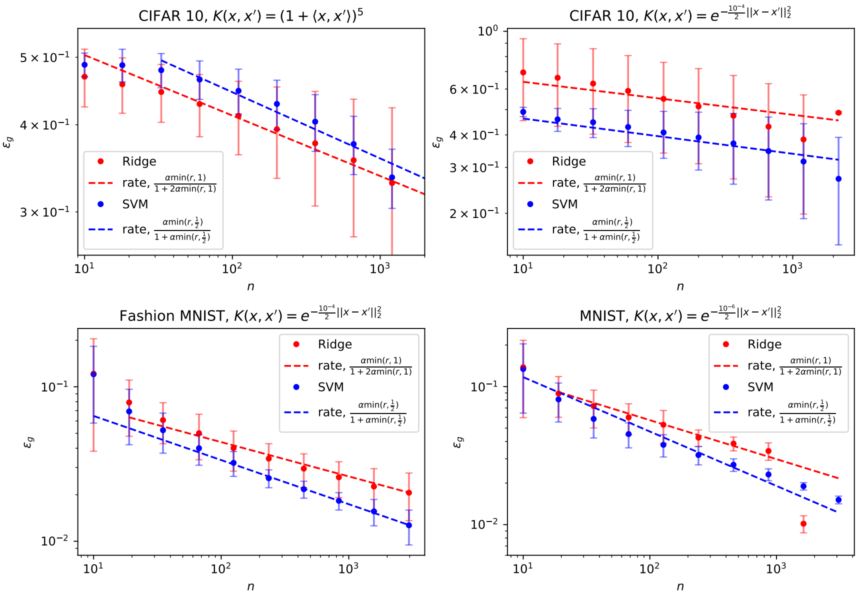
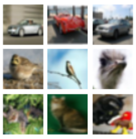


Figure 18: Dots: Misclassification error ϵ_g of kernel classification on CIFAR 10 with a polynomial kernel (top left) and an RBF kernel (top right), on Fashion MNIST with an RBF kernel (bottom left), on MNIST with an RBF kernel (bottom right), for max-margin SVM (blue) and optimally regularized ridge classification (red), using respectively the python `scikit-learn SVC` and `KernelRidge` packages. Dashed lines: Theoretical decay rates for the error ϵ_g (149) (blue) (155) (red), computed from empirically estimated capacity α and source r coefficients (see section (5.4) for details). The measure coefficients are summarized in Table 2.



The CIFAR 10 dataset.

The source and capacity condition (142) provide a simple framework to study a large class of structured data-sets. While idealized, we observe, like in Chapter 4, that many real data-sets seem to fall under this category of data-sets, and hence display learning curves which are to a good degree described by the rates (149) for SVM and (155) for ridge classification. We present here three examples of such data-sets : a data-set of 10^4 randomly sampled CIFAR 10 (Krizhevsky et al., 2009) images of animals (labelled $+1$) and means of transport (labelled -1), a data-set of 14000 FashionMNIST (Xiao et al., 2017) images of t-shirts (labelled $+1$) and coats (labelled -1), and a data-set of 14702 MNIST (LeCun et al., 1998a) images of $8s$ (labelled $+1$) and $1s$ (labelled -1). On the one hand, the learning curves for max-margin classification and optimally regularized ridge classification were obtained using the python `scikit-learn SVC, KernelRidge` packages. On the other hand, the spectrum $\{\omega_k\}_k$ of the data covariance Σ in feature space was computed, and a teacher $\theta^\star$ providing perfect classification of the data-set was fitted using margin-maximizing SVM. Then, the capacity and source coefficients α, r (142) were estimated for the data-set by fitting $\{\omega_k\}_k$ and $\{\theta^\star_k\}_k$ by power laws, and the theoretical rates (149) and (155) computed therefrom. The results of the simulations are presented in Figure 18 and compared to the theoretical

Dataset	Kernel	α	r	a_{SVM}	a_r
CIFAR 10	polynomial	1.51	0.07	0.095	0.086
CIFAR 10	RBF	1.005	0.07	0.067	0.063
Fashion MNIST	RBF	1.72	0.23	0.28	0.22
MNIST	RBF	1.65	0.39	0.39	0.28

Table 2: Values of the source and capacity coefficients (141) as estimated from the data sets, and the corresponding theoretical error rates for SVM (149) and ridge (155).

rates (149)(155) computed from the empirically evaluated source and capacity coefficients for a RBF kernel and a polynomial kernel of degree 5, with overall very good agreement. We do not compare here with the worst case bounds because the observed values of $r < 1/2$ in which case we remind the known results do not apply.

5.5 CONCLUSION

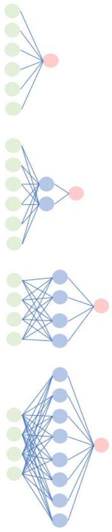
We compute the generalization error rates as a function of the source and capacity coefficients for two standard kernel classification methods, margin-maximizing SVM and ridge classification, and show that SVM classification consistently displays faster rates. Our results establish that known worst-case upper bound rates for SVM classification fail to tightly capture the rates of the class of data described by source/capacity conditions. We illustrate empirically that a number of real data-sets fall under this class, and display error rates which are to a very good degree described by the ones derived in this work.

Part III

MULTI-LAYER NETWORKS

OUTLINE AND MOTIVATIONS

Part II explored the effect of the structure of the features on learning. In daily DL practice, these features are most often learned and extracted by DNN architectures, which have proven a versatile and powerful framework to learn informative representations *of* data, *from* data. In the simplest instance of a FNN, these features are shaped by the successive propagation of the input data through the intermediate layers of the network, as it gets processed at each layer by structured weights and non-linear transformations. To be able to leverage on the insights of Part II, one thus needs to understand how the structure of the features is mathematically related to the structure of the DNN weights. On the other hand, the definition of the trained weights as the minimizers of a high-dimensional, non-convex and non-linear ERM optimization problem renders their theoretical characterization considerably challenging.



From top to bottom: The GLM, the committee machine, the extensive-width network – object of Part III–, and the infinite-width network, for depth $L = 2$.

As of the end of the past decade, a large majority of exact asymptotic studies – e.g. (Maillard et al., 2020a; Barbier et al., 2019b; Aubin et al., 2020a; Zdeborová et al., 2016; Seung et al., 1992; Sompolinsky et al., 1990; Gardner et al., 1989)– addressed GLMs with no hidden layer, or narrow two-layer networks with a small $o_d(1)$ number of hidden units (Aubin et al., 2018b; Schwarze, 1993). At the other end of the spectrum, DNNs with hidden layers of *infinite width* constitute another theoretically rather well-understood limit, thanks to their connection to kernel methods (Neal, 1996b; Williams, 1996b; Jacot et al., 2018b; Geiger et al., 2019; Chizat et al., 2018). However, due to their proximity with (generalized) linear methods and consequently rather limited expressivity, these models do not suffice to build an in-depth theoretical understanding of learning in DNNs. In particular, an investigation of *finite* – but not narrow– width DNNs is crucial. A particularly natural such limit is the *extensive-width regime*, corresponding to FNNs with widths $p, d = \Theta_n(n)$ *comparable* to the number of samples. These models should allow to probe the behaviour of overparametrized networks, while not reducing to a simple kernel limit.

Part III presents contributions to the analysis of *deep* (multilayer) FNNs in the extensive-width limit, starting from dRF networks with trainable readout and otherwise frozen random weights (Chapters 6 and 7), then networks with trainable intermediate weights (Chapters 8 and 9).

UNTRAINED NETWORKS

Recent advances have been made in the extensive-width regime for shallow RFs, corresponding to two-layer FNNs with a frozen, unstructured first layer weight matrix. The learning of such networks has been sharply characterized in theoretically controlled settings in a stream of works (Gerace et al., 2020b; Hu et al., 2022b; Mei et al., 2022d; Goldt et al., 2020c; Dhifallah et al., 2020). In particular, these analyses leverage the key insight that, as Gaussian inputs $\boldsymbol{x}$ are mixed by the random, frozen first layer weights, they propagate into features $\varphi(\boldsymbol{x})$ whose projection at readout retains asymptotically Gaussian statistics due to (a variant of) the Central Limit Theorem (CLT). This *universality* notably implies that, in terms of many learning metrics of interest, the non-linear RF is equivalent to a *noisy, linear* network with matching architecture. This equivalent view makes it clear that RFs can only implement linear methods in such regimes.

The first subpart extends this line of work to multilayer architectures. Chapter 6 addresses like these previous works the case of unstructured, isotropic weights, and evidences the presence of a similar Gaussian universality. It shows how, strikingly, a *deep, non-linear* dRF is equivalent to a *shallow, linear* network. This equivalent view offers a bridge between deep architectures and conceptually easier linear methods, and yields insights into how the architectural design of the former translates into effective biases for the latter.

Naturally, unstructured dRFs offer a model of limited realism for DNNs, whose weights display non-trivial structure after training. Chapter 7 partly palliates to this shortcoming, by considering random network ensembles with row-wise structured Gaussian weights, which were empirically found in (Guth et al., 2023) to offer a good proxy for trained networks in a number of instances. Chapter 7 provides a sharp asymptotic characterization of the learning of such colored dRF models in the extensive-width limit, and shows how this characterization can capture the learning curves of some DNNs trained with gradient-based methods, provided the weights statistics are matched. In this sense, Chapter 7 can be viewed as a refinement of Chapter 3, as it builds an effective theory of learning in DNNs taking as a starting point the weights –rather than the features– statistics.

dRFs with row-wise colored Gaussian weights were called Gaussian rainbow networks in (Guth et al., 2023)

BAYESIAN NETWORKS

The first two Chapters 6 and 7 addressed the learning of the readout weights of untrained DNNs, namely dRF models. The second subpart of Part III goes a step further to analyze DNNs with *trained* intermediary weights. Chapter 8 first considers the problem of learning a target given by a dRF function, in the framework of Bayes-optimal Bayesian learning, in the extensive-width regime. Strikingly, a form of Gaussian universality holds also when the weights are thus trained, and the Bayes-optimal test error of the dRF target

Bayes-optimal means that the priors used in the Bayesian learning correspond to the ground-truth prior associated with the ensemble the target is sampled from.

coincides with that of its equivalent single-layer GLM introduced in Chapters 6 and 7. In particular, this implies that in such cases linear methods are information-theoretically *optimal*. Chapter 8 finally provides empirical evidence that this is no longer true in more data-intensive regimes $n \gg d$, where gradient-trained DNNs manage to perfectly learn dRF targets, far outperforming linear methods.

NETWORKS AFTER ONE GRADIENT STEP

The previous Chapters evidenced that a number of settings in the extensive-width limit fall under the umbrella of Gaussian universality – where DNNs have effectively linear, thus scarce, expressivity. Informally, the presence of Gaussian universality signals that the weights are not sufficiently structured to implement informative features. While this is naturally the case for the random weights of (d)RF models (Chapters 6 and 6), Chapter 8 shows that universality can also hold when weights are learnt. When then can *feature learning* be observed in the extensive-width regime? One such setting is when the first layer weights of a two-layer FNN are trained with a single, but importantly *large* gradient step, on a GLM target. This constitutes the object of Chapter 9. After one gradient step, the weights develop a large spike aligned with the target weights, which allows the network to express *non-linear* functions in the direction of the spike, thereby breaking the curse of Gaussian universality (Ba et al., 2022a; Moniri et al., 2023; Cui et al., 2024c).

Part III A.

RANDOM FEATURES

6 DEEP RANDOM FEATURES

Despite the incredible practical progress in the applications of deep neural networks to almost all fields of knowledge, our current theoretical understanding thereof is still to a large extent incomplete. Recent progress on the theoretical front stemmed from the investigation of simplified settings, which despite their limitations are often able to capture some of the key properties of "real life" neural networks. A notable example is the recent stream of works on RF, originally introduced by (Rahimi et al., 2007a) as a computationally efficient approximation technique for kernel methods, but more recently studied as a surrogate model for two-layers neural networks in the lazy regime (Chizat et al., 2019c; Pennington et al., 2019; Mei et al., 2019c; Gerace et al., 2020a). RFs are a particular instance of random neural networks, whose statistical properties have been investigated in a sizeable body of works (Lee et al., 2018b; G. Matthews et al., 2018; Fan et al., 2020; Zavatone-Veth et al., 2021; Noci et al., 2021). The problem of training the readout layer of such networks has been addressed in the shallow (one hidden layer) case by (Mei et al., 2019c; Gerace et al., 2020a), who provide sharp asymptotic characterizations for the test error. A similar study in the generic deep case is, however, still missing. In this manuscript, we bridge this gap by considering the problem of learning the last layer of a deep, fully-connected random neural network, hereafter referred to as the dRF model. More precisely, our **main contributions** in this manuscript are:

- In 6.2, we state Theorem 6.2.4, which proves an asymptotic deterministic equivalent for the traces of the product of deterministic matrices with both conjugate kernel and sample covariance matrix of the layer-wise post-activations.

- As a consequence of Thm. 6.2.4, in 6.3 we derive a sharp asymptotic formula for the test error of the dRF model in the particular case where the target and learner networks share the same intermediate layers, and when the readout layer is trained with the squared loss. This result establishes the Gaussian equivalence of the test error for ridge regression in this setting.

- Finally, we conjecture (and provide strong numerical evidence for) the Gaussian universality of the dRF model for general convex losses, and generic target/learner network architectures. More specifically, we provide exact asymptotic formulas for the test error that leverage recent progress in high-dimensional statistics (Loureiro et al., 2021b) and a closed-form formula for the population covariance of network activations appearing in (Cui et al., 2023a). These formulas show that in terms of second-order statistics, the dRF is equivalent to a linear network with noisy layers. We discuss how this effective noise translates into a depth-induced implicit regularization in 6.4.

RELATED WORK

RF were first introduced by (Rahimi et al., 2007a). The asymptotic spectral density of the single-layer conjugate kernel was characterized in (Liao et al., 2018; Pennington et al., 2019; Benigni et al., 2021). Sharp asymptotics for the test error of the RFs model appeared in (Mei et al., 2019c; Mei et al., 2022b) for ridge regression, (Gerace et al., 2020a; Dhifallah et al., 2022) for general convex losses and (Liang et al., 2022; Bosch et al., 2022) for other penalties. The implicit regularization of RFs was discussed in (Jacot et al., 2020a). The RFs model has been studied in many different contexts as a proxy for understanding overparametrisation, e.g. in uncertainty quantification (Clarté et al., 2022), ensembling (Loureiro et al., 2022), bias-variance decomposition (D'Ascoli et al., 2020; Adlam et al., 2020b), the training dynamics (Bodin et al., 2021; Bordelon et al., 2022; Paquette et al., 2022), but also to highlight the limitations of lazy training (Ghorbani et al., 2019c; Ghorbani et al., 2021; Yehudai et al., 2019; Refinetti et al., 2021b);

Deep random networks were shown to converge to Gaussian processes in (Lee et al., 2018b; G. Matthews et al., 2018). They were also studied in the context of inference in (Manoel et al., 2017b; Gabrié et al., 2018b), and as generative priors to inverse problems in (Aubin et al., 2019; Hand et al., 2018; Aubin et al., 2020b). The distribution of outputs of deep random nets was characterized in (Zavatone-Veth et al., 2021; Noci et al., 2021). Close to our work is (Fan et al., 2020), which provide exact formulas for the asymptotic spectral density and Stieltjes transform of the NTK and conjugate kernel in the proportional limit. Our formulas for the sample and population covariance are complementary to theirs. The test error of deep networks has been recently studied in (Li et al., 2021b; Hanin et al., 2019; Pacelli et al., 2023; Zavatone-Veth et al., 2022a) through the lens of Bayesian learning;

Gaussian universality of the test error for the RFs model was shown in (Mei et al., 2019c), conjectured to hold for general losses in (Gerace et al., 2020a) and was proven in (Goldt et al., 2021a; Hu et al., 2020). Gaussian universality has also been shown to hold for other classes of features, such as two-layer NTK (Montanari et al., 2022b). (Bordelon et al., 2022; Jacot et al., 2020a; Cui et al., 2021; Cui et al., 2023c) further heuristically showed that Gaussian universality is also observed for a large class of kernel features. (Loureiro et al., 2021b) provided numerical evidence for Gaussian universality of more general feature maps, including pre-trained deep features.

Deterministic equivalents of sample covariance matrices have first been established in (Marchenko et al., 1967) for separable covariances, generalizing the seminal work (Marchenko et al., 1967) on the free convolution of spectra in an anisotropic sense. More recently these results have been extended to non-separable covariances, first in tracial (Bai et al., 2008c), and then also in anisotropic sense (Louart et al., 2018a; Chouard, 2022).

Shortly after the first version of this work appeared on arXiv, we have learned about (Bosch et al., 2023a) which overlaps with some parts of our work. In particular, they show universality for strongly convex risks for the deep random features model in the well-specified setting, proving part of Conjecture 6.3.2.

6.1 SETTING & PRELIMINARIES

Let $(\boldsymbol{x}^\mu, y^\mu) \in \mathbb{R}^d \times \mathcal{Y}$, $\mu \in [n] := \{1, \cdots, n\}$, denote some training data, with $\boldsymbol{x}^\mu \sim \mathcal{N}(0_d, \Omega_0)$ independently and $y^\mu = f_\star(\boldsymbol{x}^\mu)$ a (potentially random) target function. This work is concerned with characterising the learning performance of generalised linear estimation:

$$\hat{y} = \sigma\left(\frac{\theta^\top \varphi(\boldsymbol{x})}{\sqrt{k}}\right), \tag{156}$$

with dRF:

$$\varphi(\boldsymbol{x}) := \underbrace{(\varphi_L \circ \varphi_{L-1} \circ \cdots \circ \varphi_2 \circ \varphi_1)}_{L}(\boldsymbol{x}), \tag{157}$$

where the post-activations are given by:

$$\varphi_\ell(h) = \sigma_\ell\left(\frac{1}{\sqrt{k_{\ell-1}}} W_\ell \cdot h\right), \quad \ell \in [L]. \tag{158}$$

The weights $\{W_\ell \in \mathbb{R}^{k_\ell \times k_{\ell-1}}\}_{\ell \in [L]}$ are assumed to be independently drawn Gaussian matrices with i.i.d. entries $(W_\ell)_{ij} \sim \mathcal{N}(0, \Delta_\ell)$ $\forall 1 \leq i \leq k_\ell$, $1 \leq j \leq k_{\ell-1}$. To alleviate notation, sometimes it will be convenient to denote $k_L = k$. Only the readout weights $\theta \in \mathbb{R}^k$ in (156) are trained according to the usual regularized ERM procedure:

$$\hat{\theta} = \underset{\theta \in \mathbb{R}^k}{\operatorname{argmin}} \left[\sum_{\mu=1}^{n} \ell(y^\mu, \theta^\top \varphi(\boldsymbol{x}^\mu)) + \frac{\lambda}{2} ||\theta||^2\right], \tag{159}$$

where $\ell : \mathcal{Y} \times \mathbb{R} \to \mathbb{R}_+$ is a loss function, which we assume convex, and $\lambda > 0$ sets the regularization strength.

To assess the training and test performances of the empirical risk minimizer (159), we let $g : \mathcal{Y} \times \mathbb{R} \to \mathbb{R}_+$ be any performance metric (e.g. the loss function itself or, in the case of classification, the probability of misclassifying), and define the test error:

$$\varepsilon_g(\hat{\theta}) := \mathbb{E}\left[g(y, \hat{\theta}^\top \varphi(\boldsymbol{x}))\right] \tag{160}$$

Our main goal in this work is to provide a sharp characterization of (160) in the proportional asymptotic regime $n, d, k_\ell \to \infty$ at fixed $\mathscr{O}(1)$ ratios $\alpha := {}^{n}/_{d}$ and $\gamma_\ell := {}^{k_\ell}/_{d}$, for all layer index $\ell \in [L]$. This requires a precise characterization of the sample and population covariances and the Gram matrices of the post-activations.

6.1.1 BACKGROUND ON SAMPLE COVARIANCE MATRICES

Marchenko-Pastur and free probability: We briefly introduce basic nomenclature on sample covariance matrices. For a random vector $x \in \mathbb{R}^d$ with mean zero $\mathbb{E}x = 0$ and covariance $\Sigma := \mathbb{E}xx^\top \in \mathbb{R}^{d\times d}$, we call the matrix $\Sigma := \mathscr{X}\mathscr{X}^\top/n \in \mathbb{R}^{d\times d}$ obtained from n independent copies $x_1, \ldots, x_n$ of x written in matrix form as $\mathscr{X} := (x_1, \ldots, x_n)$ the *sample covariance matrix* corresponding to the *population covariance matrix* Σ. The *Gram matrix* $\Sigma := \mathscr{X}^\top\mathscr{X}/n \in \mathbb{R}^{n\times n}$ has the same non-zero eigenvalues as the sample covariance matrix but unrelated eigenvectors. The systematic mathematical study of sample covariance and Gram matrices has a long history dating back to (Wishart, 1928). While in the "classical" statistical limit $n \to \infty$ with d being fixed the sample covariance matrix converges to the population covariance matrix $\Sigma \to \Sigma$, in the proportional regime $d \sim n \gg 1$ the non-trivial asymptotic relationship between the spectra of Σ and Σ has first been obtained in the seminal paper (Marchenko et al., 1967): the empirical spectral density $\mu(\Sigma) := d^{-1}\sum_{\lambda\in\mathrm{Spec}(\Sigma)}\delta_\lambda$ of Σ is approximately equal to the *free multiplicative convolution* of $\mu(\Sigma)$ and a Marchenko-Pastur distribution μ_{MP}^c of aspect ratio $c = d/n$,

$$\mu(\Sigma) \approx \mu(\Sigma) \boxtimes \mu_{\mathrm{MP}}^{d/n}. \tag{161}$$

Here the free multiplicative convolution $\mu \boxtimes \mu_{\mathrm{MP}}^c$ may be defined as the unique distribution ν whose Stieltjes transform $m = m_\nu(z) := \int (x-z)^{-1}\,\mathrm{d}\nu(x)$ satisfies the scalar *self-consistent equation*

$$zm = \frac{z}{1-c-czm} m_\mu\left(\frac{z}{1-c-czm}\right). \tag{162}$$

The spectral asymptotics (161) originally were obtained in the case of Gaussian $\mathscr{X}$ or, more generally, for separable correlations $\mathscr{X} = \sqrt{\Sigma}Y$ for some i.i.d. matrix $Y \in \mathbb{R}^{d\times n}$. These results were later extended (Bai et al., 2008c) to the general case under essentially optimal assumptions on concentrations of quadratic forms $x^\top Ax$ around their expectation $\mathrm{Tr}A\Sigma$.

Deterministic equivalents: It has only been recognised much later (Burda et al., 2004; Knowles et al., 2017) that the relationship (161) between the asymptotic spectra of Σ and Σ, Σ actually extends to eigenvectors as well, and that the resolvents $G(z) := (\Sigma - z)^{-1}, G(z) := (\Sigma - z)^{-1}$ are asymptotically equal to *deterministic equivalents*

$$M(z) := -\frac{(\Sigma m(z) + I_d)^{-1}}{z}, \quad M(z) := m(z) I_n, \tag{163}$$

also in an *anisotropic* rather than just a tracial sense, highlighting that despite the simple relationship between their averaged traces

$$m(z) := m_{\mu(\Sigma) \boxtimes \mu_{\mathrm{MP}}^c}(z), \quad m(z) = \frac{c-1}{z} + cm(z),$$

the sample covariance and Gram matrices carry rather different non-spectral information. The anisoptric concentration of resolvents (or in physics terminology, the self-averaging) has again first been obtained in the Gaussian or separable cases (Burda et al., 2004; Knowles et al., 2017). The extension to general sample covariance matrices was only achieved much more recently (Louart et al., 2018a; Chouard, 2022) under Lipschitz concentration assumptions. In this work we specifically use the deterministic equivalent for sample covariance matrices with general covariance from (Chouard, 2022) and extend it to cover Gram matrices.

Application to the deep random features model: In this work we apply the general theory of anisotropic deterministic equivalents to the deep random features model. As discussed in Section 6.3, to prove error universality even for the simple ridge regression case, it is not enough to only consider the spectral convergence of the matrices, and a stronger result is warranted. The application of non-linear activation functions makes the model neither Gaussian nor separable, hence our analysis relies on the deterministic equivalents from (Chouard, 2022) and our extension to Gram matrices, which appear naturally in the explicit error derivations.

6.1.2 NOTATION

We will adopt the following notation:

- For $A \in \mathbb{R}^{n \times n}$ we denote $\langle A \rangle := {}^1\!/\!{}_n \mathrm{Tr}\,(A)$.
- For matrices $A \in \mathbb{R}^{n \times m}$ we denote the operator norm (with respect to the ℓ^2-vector norm) by $\|A\|$, the max-norm by $\|A\|_{\max} := \max_{ij} |A_{ij}|$, and the Frobenius norm by $\|A\|_{\mathrm{F}}^2 := \sum_{ij} |A_{ij}|^2$.
- For any distribution μ we denote the push-forward under the map $\lambda \mapsto a\lambda + b$ by $a \otimes \mu \oplus b$ in order to avoid confusion with e.g. the convex combination $a\mu_1 + (1-a)\mu_2$ of measures μ_1, μ_2.

- We say that a sequence of random variables $(X_n)_n$ is *stochastically dominated* by another sequence $(Y_n)_n$ if for all small $\varepsilon > 0$ and large $D < \infty$ it holds that $P(X_n > n^{\varepsilon} Y_n) \leq n^{-D}$ for large enough n, and in this case write $X_n \prec Y_n$.

6.2 DETERMINISTIC EQUIVALENTS

Consider the sequence of variances defined by the recursion (recall that Δ_ℓ is the variance of the entries of W_ℓ)

$$r_{\ell+1} = \Delta_{\ell+1} \mathbb{E}_{\xi \sim \mathcal{N}(0,r_\ell)} \left[\sigma_\ell(\xi)^2\right] \tag{164}$$

with initial condition $r_1 := \Delta_1 \langle \Omega_0 \rangle$ and coefficients

$$\begin{aligned} \kappa_1^\ell &= \frac{1}{r_\ell} \mathbb{E}_{\xi \sim \mathcal{N}(0,r_\ell)} \left[\xi \sigma_\ell(\xi)\right], \\ \kappa_*^\ell &= \sqrt{\mathbb{E}_{\xi \sim \mathcal{N}(0,r_\ell)} \left[\sigma_\ell(\xi)^2\right] - r_\ell \left(\kappa_1^\ell\right)^2}. \end{aligned} \tag{165}$$

6.2.1 RIGOROUS RESULTS ON THE MULTI-LAYER SAMPLE COVARIANCE AND GRAM MATRICES

Our main result on the anisotropic deterministic equivalent of dRFs follows from iterating the following proposition. We consider a data matrix $X_0 \in \mathbb{R}^{d \times n}$ whose Gram matrix concentrates as

$$\left\| \frac{X_0^\top X_0}{d} - r_1 I \right\|_{\max} \prec \frac{1}{\sqrt{n}}, \qquad \left\| \frac{X_0}{\sqrt{d}} \right\| \prec 1 \tag{166}$$

for some positive constant r_1. The Assumption (166) for instance is satisfied if the columns $\boldsymbol{x}$ of X_0 are independent with mean $\mathbb{E}\boldsymbol{x} = 0$ and covariance $\mathbb{E}\boldsymbol{x}\boldsymbol{x}^\top = \Omega_0 \in \mathbb{R}^{d \times d}$ (together with some mild assumptions on the fourth moments), in which case $r_1 = \langle \Omega_0 \rangle$ is the normalised trace of the covariance. We then consider $X_1 := \sigma_1(W_1 X_0 / \sqrt{d})$ assuming the entries of $W_1 \in \mathbb{R}^{k_1 \times d}$ are iid. $\mathcal{N}(0,1)$ elements, and σ_1 satisfies $\mathbb{E}_{\xi \sim \mathcal{N}(0,1)} \sigma_1(\sqrt{r_1}\xi) = 0$ in the proportional $n \sim d \sim k_1$ regime. Upon changing σ_1 there is no loss in generality in assuming $\Delta_1 = 1$ which we do for notational convenience.

Proposition 6.2.1 (Deterministic equivalent for RF). *For any deterministic A and Lipschitz-continuous activation function σ_1, under the assumptions above, we have that, for any $z \in \mathbf{C} \setminus \mathbb{R}_+$*

$$\left| \left\langle A \left[\left(\frac{X_1^\top X_1}{k_1} - z \right)^{-1} - M(z) \right] \right\rangle \right| \prec \frac{\langle AA^* \rangle^{1/2}}{\delta^9 \sqrt{n}},$$

and

$$\left|\left\langle A\left(\frac{X_1X_1^\top}{k_1}-z\right)^{-1}\right\rangle-\langle A\rangle m(z)\right|\prec\frac{\langle AA^*\rangle^{1/2}}{\delta^9\sqrt{n}},$$

where $\delta:=\operatorname{dist}(z,\mathbb{R}_+)$,

$$\begin{aligned}-zM(z)&:=\Big(m(z)\Sigma_{\text{lin}}+I\Big)^{-1},\\ \Sigma_{\text{lin}}&:=(\kappa_1^1)^2\frac{X_0^\top X_0}{d}+(\kappa_*^1)^2I,\end{aligned}\tag{167}$$

and

$$m(z):=m_{\mu(\Sigma_{\text{lin}})\boxtimes\mu_{\text{MP}}^{n/k_1}}(z),\quad m(z)=\frac{n-k_1}{nz}+\frac{n}{k_1}m(z).$$

Furthermore, Assumption (166) *holds true with* X_0,r_1 *replaced by* X_1,r_2*, respectively, and we have that* $\operatorname{dist}(-1/m(z),\mathbb{R}_+)\geq\operatorname{dist}(z,\mathbb{R}_+)$.

Remark 6.2.2. *Proposition relies on the recent work of Chouard (Chouard, 2022) on deterministic equivalents of sample-covariance matrices. The main novelty here is twofold. First, we extend Chouard's result on the sample covariance matrix* $X_1^\top X_1$ *to the Gram matrix* $X_1X_1^\top$*. Second, we replace the population covariance matrix:*

$$\Sigma_{X_0}:=\mathbb{E}_{w\sim\mathcal{N}(0,I)}\sigma\left(\frac{X_0^\top w}{\sqrt{d}}\right)\sigma\left(\frac{w^\top X_0}{\sqrt{d}}\right)\approx(\kappa_1^1)^2\frac{X_0^\top X_0}{d}+(\kappa_*^1)^2I=:\Sigma_{\text{lin}}.$$

Note that both extensions are crucial for our main result on the test error since the latter naturally depends on the Gram matrix $X_lX_l^\top$ *and the iteration of* 6.2.1 *only becomes viable after linearisation.*

Remark 6.2.3. *The tracial version of* 6.2.1 *has appeared multiple times in the literature, e.g. (Bai et al., 2008c). It implies that the spectrum* μ_1 *of* $X_1^\top X_1/k_1$ *is approximately given by the free multiplicative convolution*

$$\begin{aligned}\mu_1&\approx\mu\Big((\kappa_1^1)^2\frac{X_0^\top X_0}{d}+(\kappa_*^1)^2I\Big)\boxtimes\mu_{\text{MP}}^{n/k_1}\\ &=\Big(\mu\Big((\kappa_1^1)^2\frac{X_0^\top X_0}{d}\Big)\boxplus\delta_{(\kappa_*^1)^2}\Big)\boxtimes\mu_{\text{MP}}^{n/k_1},\end{aligned}\tag{168}$$

where "≈" means that some metric between the two probability measures is small, e.g. the Kolmogorov-Smirnov distance. *Since the relation between convergence of Stieltjes transforms and and metric convergence of measures is fairly standard (see e.g. (Bai, 1993, Theorem 2.1) we refrain from elaborating on this technical point. In case* $c\leq1$*, i.e. when* μ_{MP}^c *has no atom at* 0*, it was shown in (Benaych-Georges, 2010) that*

$$\sqrt{\mu\boxtimes\mu_{\text{MP}}^c}\boxplus_c\sqrt{\mu'\boxtimes\mu_{\text{MP}}^c}=\sqrt{(\mu\boxplus\mu')\boxtimes\mu_{\text{MP}}^c}\tag{169}$$

which allows to simplify (168). *Here* $\boxplus_c$ *is the* rectangular free convolution *which models the distribution of singular values of the addition of two free rectangular random matrices, and the square-root is to be understood as the push-forward of the square-root map. Applying* (169) *to* (168) *yields*

$$\sqrt{\mu_1} \approx \left(\kappa_1^1 \otimes \sqrt{\mu_0 \boxtimes \mu_{\mathrm{MP}}^{n/k_1}}\right) \boxplus_{n/k_1} \kappa_*^1 \otimes \sqrt{\mu_{\mathrm{MP}}^{n/k_1}}, \tag{170}$$

suggesting that the non-zero singular values of $X_1/\sqrt{k}$ *can be modeled by the non-zero singular values of the* Gaussian equivalent model*:*

$$c'W'X_0 + c''W'' \tag{171}$$

for some suitably chosen constants c', c'' *and independent Gaussian matrices* W', W''.

The last assertion of 6.2.1 allows to iterate over an arbitrary (but finite) number of layers. Indeed, after one layer we have

$$\begin{aligned}\left(\frac{X_1^\top X_1}{k_1} - z_1\right)^{-1} &\approx \Big(-m(z_1)z_1\Sigma_{\mathrm{lin}} - z_1\Big)^{-1} \\ &= c_1\left(\frac{X_0^\top X_0}{k_0} - z_0\right)^{-1},\end{aligned} \tag{172}$$

using the definitions from 6.2.4 for c_1, z_0 below. Here "$\approx$" should be understood in the sense of 6.2.4.

Theorem 6.2.4. *(Deterministic equivalent for dRF) For any deterministic* A *and Lipschitz-continuous activation functions* $\sigma_1, \dots, \sigma_\ell$ *satisfying* $\mathbb{E}_\xi \sigma_m(\sqrt{r_m}\xi) = 0$ *(with* $\xi \sim \mathcal{N}(0,1)$*), under the Assumption* (166) *above, we have that for any* $z_\ell \in \mathbf{C} \setminus \mathbb{R}_+$

$$\left|\left\langle A\left(\frac{X_\ell^\top X_\ell}{k_\ell} - z_\ell\right)^{-1}\right\rangle - c_1 \cdots c_\ell m_0 \langle A\rangle\right| \prec \frac{\langle AA^*\rangle^{1/2}}{\delta_\ell^9 \sqrt{n}}$$

and that

$$\left|\left\langle A\left(\frac{X_\ell X_\ell^\top}{k_\ell} - z_\ell\right)^{-1}\right\rangle - m_\ell \langle A\rangle\right| \prec \frac{\langle AA^*\rangle^{1/2}}{\delta_\ell^9 \sqrt{n}},$$

where $\delta_\ell := \operatorname{dist}(z_\ell, \mathbb{R}_+)$*, and we recursively define*

$$\begin{aligned}\Sigma_{\mathrm{lin}}^{\ell-1} &:= (\kappa_1^\ell)^2 \frac{X_{\ell-1}^\top X_{\ell-1}}{k_{\ell-1}} + (\kappa_*^\ell)^2 I, \\ m_\ell &:= \frac{n-k_\ell}{n z_\ell} + \frac{n}{k_\ell} m_{\mu(\Sigma_{\mathrm{lin}}^{\ell-1}) \boxtimes \mu_{\mathrm{MP}}^{n/k_\ell}}(z_\ell) \\ -\frac{1}{c_\ell} &:= m_\ell z_\ell (\kappa_1^\ell)^2, \quad z_{\ell-1} := c_\ell z_\ell - \left(\frac{\kappa_*^\ell}{\kappa_1^\ell}\right)^2\end{aligned} \tag{173}$$

for $\ell \geq 1$ and finally

$$m_0 := \frac{d-n}{nz_0} + \frac{d^2}{n^2} m_{\mu(\Omega_0)\boxtimes\mu_{\mathrm{MP}}^{d/n}}\left(\frac{d}{n}z_0\right). \tag{174}$$

Remark 6.2.5. *The same iteration argument and the tracial version of 6.2.4 has appeared before in (Fan et al., 2020). The main difference to our present work is the anisotropic nature of our estimate which allows to test both sample covariance, as well as Gram resolvent against arbitrary deterministic matrices. As we will discuss in the next section, this is crucial in order to provide closed-form asymptotics for the test error of the deep random features model.*

6.2.2 CLOSED-FORMED FORMULA FOR THE POPULATION COVARIANCE

In Propostion 6.2.1 and Theorem 6.2.4 we iteratively considered $X_\ell^\top X_\ell / k_\ell$ as a sample-covariance matrix with population covariance

$$\mathbb{E}_{W_\ell} \frac{X_\ell^\top X_\ell}{k_\ell} = \mathbb{E}_w \sigma_\ell \left(\frac{X_{\ell-1}^\top w}{\sqrt{k_{\ell-1}}}\right) \sigma_\ell \left(\frac{w^\top X_{\ell-1}}{\sqrt{k_{\ell-1}}}\right) \approx \Sigma_{\mathrm{lin}}^\ell$$

and from this obtained formulas for the deterministic equivalents for both $X_\ell^\top X_\ell$ and $X_\ell X_\ell^\top$. A more natural approach would be to consider $X_\ell X_\ell^\top / n$ as a sample covariance matrix with population covariance

$$\Omega_\ell := \mathbb{E}_{X_0} \frac{X_\ell X_\ell^\top}{n}, \tag{175}$$

noting that the matrix X_ℓ conditioned on $W_1, \ldots, W_\ell$ has independent columns. A heuristic closed-form formula for the population covariance which is conjectured to be exact was recently derived in (Cui et al., 2023a), which will be the object of further discussion in Chapter 8. We now discuss this result. Consider the sequence of matrices $\{\Omega_\ell^{\mathrm{lin}}\}_\ell$ defined by the recursion

$$\Omega_{\ell+1}^{\mathrm{lin}} = \kappa_1^{(\ell+1)2} \frac{W_{\ell+1}\Omega_\ell^{\mathrm{lin}} W_{\ell+1}^\top}{k_\ell} + \kappa_*^{(\ell+1)2} I_{k_{\ell+1}}. \tag{176}$$

with $\Omega_0^{\mathrm{lin}} := \Omega_0$. Informally, $\Omega_\ell^{\mathrm{lin}}$ provides an asymptotic approximation of Ω_ℓ in the sense that the normalized distance $||\Omega_\ell^{\mathrm{lin}} - \Omega_\ell||_F / \sqrt{d}$ is of order $\mathcal{O}(1/\sqrt{d})$. Besides, the recursion (176) implies that $\Omega_\ell^{\mathrm{lin}}$ can be expressed as a sum of products of Gaussian matrices (and transposes thereof), and affords a straightforward way to derive an analytical expression its asymptotic spectral distribution.

It is an interesting question whether an approximate formula for the population covariance matrix like the one in 176 can be obtained indirectly via 6.2.4. There is extensive literature on this *inverse problem*, i.e. how to infer spectral properties of the population covariance spectrum from the sample covariance spectrum, e.g. (El Karoui, 2008a) but we leave this avenue to future work.

6.2.3 CONSISTENCY OF 6.2.4 AND THE APPROXIMATE POPULATION COVARIANCE

What we can note, however, is that 176 is *consistent* with 6.2.4. We demonstrate this in case of equal dimensions $n = d = k_1 = \cdots = k_\ell$ to avoid unnecessary technicalities due to the zero eigenvalues. We define

$$\mu_\ell := \mu\left(\frac{X_\ell^\top X_\ell}{k_\ell}\right) = \mu_\ell := \mu\left(\frac{X_\ell X_\ell^\top}{n}\right) \tag{177}$$

and recall that 6.2.1 implies that

$$\mu_\ell \approx ((\kappa_1^\ell)^2 \otimes \mu_{\ell-1} \oplus (\kappa_*^\ell)^2) \boxtimes \mu_{\mathrm{MP}}. \tag{178}$$

On the other hand (161) applied to the sample covariance matrix $X_\ell X_\ell^\top / n$ with population covariance $\Omega_\ell \approx \Omega_\ell^{\mathrm{lin}}$ implies that

$$\begin{aligned}
\mu_\ell &\approx \mu(\Omega_\ell^{\mathrm{lin}}) \boxtimes \mu_{\mathrm{MP}} \\
&= \mu\left((\kappa_1^\ell)^2 \frac{W_\ell \Omega_{\ell-1}^{\mathrm{lin}} W_\ell^\top}{k_{\ell-1}} + (\kappa_*^\ell)^2 I_{k_\ell}\right) \boxtimes \mu_{\mathrm{MP}} \\
&\approx \left((\kappa_1^\ell) \otimes \mu(\Omega_{\ell-1}^{\mathrm{lin}}) \boxtimes \mu_{\mathrm{MP}} \oplus (\kappa_*^\ell)^2\right) \boxtimes \mu_{\mathrm{MP}} \\
&\approx \left((\kappa_1^\ell) \otimes \mu_{\ell-1} \oplus (\kappa_*^\ell)^2\right) \boxtimes \mu_{\mathrm{MP}},
\end{aligned} \tag{179}$$

demonstrating that both approaches lead to the same recursion. Here in the third step we applied (161) to the sample covariance matrix $\sqrt{\Omega_{\ell-1}^{\mathrm{lin}}} W_\ell^\top$, and in the fourth step used the first approximation for ℓ replaced by $\ell - 1$.

6.3 GAUSSIAN UNIVERSALITY OF THE TEST ERROR

In the second part of this work, we discuss how the results on the asymptotic spectrum of the empirical and population covariances of the features can be used to provide sharp expressions for the test and training errors (160) when the labels are generated by a deep random neural network:

$$f_\star(\boldsymbol{x}^\mu) = \sigma^\star\left(\frac{\theta_\star^\top \varphi^\star(\boldsymbol{x}^\mu)}{\sqrt{k^\star}}\right). \tag{180}$$

The feature map $\varphi^\star$ denotes the composition $\varphi^\star_{L^\star} \circ ... \circ \varphi^\star_1$ of the $L^\star + 1$ layers:

$$\varphi^\star_\ell(\boldsymbol{x}) = \sigma^\star_\ell \left(\frac{1}{\sqrt{k^\star_{\ell-1}}} W^\star_\ell \cdot \boldsymbol{x} \right),$$

and $\theta_\star \in \mathbb{R}^{k^\star}$ is the last layer weights. To alleviate notations, we denote $k^\star := k^\star_L$. The weight matrices $\{W^\star_\ell\}_{\ell \in [L^\star]}$ have i.i.d Gaussian entries sampled from $\mathcal{N}(0, \Delta^\star_\ell)$. Note that we do not require the sequence of activations $\{\sigma^\star_\ell\}_\ell$ and widths $\{\gamma^\star_\ell := k^\star_\ell/d\}_\ell$ to match with those of the learner dRF (157). We address in succession

- The well-specified case where the target and learner networks share the same intermediate layers (i.e. same architecture, activations and weights) $\varphi^\star_\ell = \varphi_\ell$, $\ell \in [L]$ with $L^\star = L$, and the readout of the dRF is trained using ridge regression. This is equivalent to the interesting setting of ridge regression on a linear target, with features drawn from a non-Gaussian distribution, resulting from the propagation of Gaussian data through several non-linear layers.

- The general case where the target and learner possess generically distinct architectures, activations and weights, and a generic convex loss.

In both cases, we provide a sharp asymptotic characterization of the test error. Furthermore, we establish the equality of the latter with the test error of an equivalent learning problem on *Gaussian samples* with matching population covariance, thereby showing the Gaussian universality of the test error. In the well-specified case, our results are rigorous, and make use of the deterministic equivalent provided by Theorem 6.2.4. In the fully generic case, we formulate a conjecture, which we strongly support with finite-size numerical experiments.

6.3.1 WELL-SPECIFIED CASE

We first establish the Gaussian universality of the test error of dRFs in the matched setting $\varphi = \varphi^\star$, for a readout layer trained using a square loss. This corresponds to $\mathcal{Y} = \mathbb{R}$, $\ell(y, \hat{y}) = 1/2(y - \hat{y})^2$. This case is particularly simple since the empirical risk minimization problem (159) admits the following closed form solution:

$$\hat{\theta} = 1/\sqrt{k}(\lambda I_k + 1/k X_L X_L^\top)^{-1} X_L y \tag{181}$$

where we recall the reader $X_L \in \mathbb{R}^{k \times n}$ is the matrix obtained by stacking the last layer features column-wise and $y \in \mathbb{R}^n$ is the vector of labels. For a given target function, computing the test error boils down to a random matrix theory problem depending on variations of the trace of deterministic matrices times the resolvent of the features sample covariance matrices :

$$\begin{aligned}\varepsilon_g(\hat{\theta}) = \Delta\left(\left\langle \Omega_L \left(\lambda I_k + {}^1\!/\!{}_k X_L X_L^\top\right)^{-1}\right\rangle + 1\right) \\ - \lambda(\lambda-\Delta)\partial_\lambda \left\langle \Omega_L \left(\lambda I_k + {}^1\!/\!{}_k X_L X_L^\top\right)^{-1}\right\rangle \qquad (182)\end{aligned}$$

Applying Theorem 6.2.4 yields the following corollary:

Corollary 6.3.1 (Ridge universality of matched target)**.** *Let $\lambda > 0$. In the asymptotic limit $n, d, k_\ell \to \infty$ with fixed $\mathcal{O}(1)$ ratios $\alpha = {}^n\!/\!{}_d$, $\gamma_\ell := {}^{k_\ell}\!/\!{}_d$ and under the assumptions of Theorem 6.2.4, the asymptotic test error of the ridge estimator (181) on the target (180) with $L = L^*$ and $\varphi_\ell^* = \varphi_\ell$ and additive Gaussian noise with variance $\Delta > 0$ is given by:*

$$\begin{aligned}\varepsilon_g(\hat{\theta}) \xrightarrow{k\to\infty} \varepsilon_g^\star = \Delta\left(\langle \Omega_L \rangle m_L(-\lambda) + 1\right) \\ - \lambda(\lambda-\Delta)\langle \Omega_L \rangle \partial_\lambda m_L(-\lambda) \qquad (183)\end{aligned}$$

where m_L can be recursively computed from (173) respectively. In particular, this implies Gaussian universality of the asymptotic mean-squared error in this model, since (183) exactly agrees with the asymptotic test error of ridge regression on Gaussian data $\boldsymbol{x} \sim \mathcal{N}(0_d, \Omega_L)$ derived in (Dobriban et al., 2018b).

Note that, while it is not needed to establish the Gaussian equivalence of ridge dRF regression in the well-specified case, the trace of the population covariance $\langle \Omega_L \rangle$ can be explicitly computed from the closed-form formula (176).

6.3.2 GENERAL CASE

Despite the major progress stemming from the application of the random matrix theory toolbox to learning problems, the application of the latter has been mostly limited to quadratic problems where a closed-form expression of the estimators, such as (181), are available. Proving universality results akin to Corollary 6.3.1 beyond quadratic problems is a challenging task, which has recently been the subject of intense investigation. In the context of generalized linear estimation (159), universality of the test error for the $L = 1$ random features model under a generic convex loss function was heuristically studied in (Gerace et al., 2020a), where the authors have shown that the asymptotic formula for the test error obtained under the Gaussian design assumption perfectly agreed with finite-size simulations with the true features. This Gaussian universality of the test error was later proven by (Hu et al., 2020) by combining a Lindeberg interpolation scheme with a generalized central limit theorem. Our goal in the following is to provide an analogous contribution as (Gerace et al., 2020a) to the case of multi-layer random features. This result builds on a rigorous, closed-form formula for the asymptotic test error of misspecified generalized linear estimation in the high-dimensional limit considered here, which was derived in (Loureiro et al., 2021b).

We show that in the high-dimensional limit the asymptotic test error for the model introduced in 6.1 is in the *Gaussian universality class*. More precisely, the test error of this model is asymptotically equivalent to the test error of an equivalent GCM consisting of doing generalized linear estimation on a dataset $\check{\mathscr{D}} = \{v^\mu, \check{y}^\mu\}_{\mu\in[n]}$ with labels $\check{y}^\mu = f_\star({}^1\!/\!\sqrt{k^\star}\boldsymbol{\theta}_\star^\top u^\mu)$ and jointly Gaussian covariates:

$$(u, v) \sim \mathscr{N}\begin{pmatrix} \Psi_{L^\star} & \Phi_{L^\star L} \\ \Phi_{L^\star L}^\top & \Omega_L \end{pmatrix} \tag{184}$$

where we recall Ω_L is the variance of the model features (175) and $\Phi \in \mathbb{R}^{k^\star \times k}$ and $\Psi \in \mathbb{R}^{k^\star \times k^\star}$ are the covariances between the model and target features and the target variance respectively:

$$\Phi_{L^\star L} := \mathbb{E}\left[\varphi^\star(\boldsymbol{x})\varphi(\boldsymbol{x})^\top\right], \; \Psi_{L^\star} := \mathbb{E}\left[\varphi^\star(\boldsymbol{x})\varphi^\star(\boldsymbol{x})^\top\right] \tag{185}$$

This result adds to a stream of recent universality results in high-dimensional linear estimation (Loureiro et al., 2021b; Montanari et al., 2022b; Gerace et al., 2022), and generalizes the random features universality of (Mei et al., 2022b; Goldt et al., 2021a; Hu et al., 2020) to $L > 1$. It can be summarized in the following conjecture:

Conjecture 6.3.2. *In the high-dimensional limit $n, d, k_\ell \to \infty$ at fixed $\mathcal{O}(1)$ ratios $\alpha := {}^n\!/\!{}_d$ and $\gamma_\ell := {}^{k_\ell}\!/\!{}_d$, the test error of the empirical risk minimizer* (159) *trained on $\mathscr{D} = \{(\boldsymbol{x}^\mu, y^\mu)\}_{\mu\in[n]}$ with covariates $\boldsymbol{x}^\mu \sim \mathscr{N}(0_d, \Omega_0)$ and labels from* (180) *is equal to the one of a Gaussian covariate model* (184) *with matching second moments Ψ, Φ, Ω as defined in* (175) *and* (185).

We go a step further and provide a sharp asymptotic expression for the test error. Construct recursively the sequence of matrices

$$\Psi_{\ell+1}^{\mathrm{lin}} = \left(\kappa_1^{\star(\ell+1)}\right)^2 \frac{W_{\ell+1}^\star \Psi_\ell^{\mathrm{lin}} W_{\ell+1}^{\star\top}}{k_\ell^\star} + \left(\kappa_*^{\star(\ell+1)}\right)^2 I_{k_{\ell+1}^\star} \tag{186}$$

with the initial condition $\Omega_0^{\mathrm{lin}} = \Psi_0^{\mathrm{lin}} := \Omega_0$. Further define

$$\Phi_{L^\star L}^{\mathrm{lin}} = \left(\prod_{\ell=L^\star}^{1} \frac{\kappa_1^{\star\ell} W_\ell^\star}{\sqrt{k_\ell^\star}}\right) \cdot \Omega_0 \cdot \left(\prod_{\ell=1}^{L} \frac{\kappa_1^\ell W_\ell^\top}{\sqrt{k_\ell}}\right). \tag{187}$$

The sequence $\{\kappa_1^{\star\ell}, \kappa_*^{\star\ell}\}_{\ell=1}^{L^\star}$ is defined by (165) with $\sigma_\ell^\star, \Delta_\ell^\star$. In the special case $L^\star = 0$, which corresponds to a single-index target function, the first product in $\Phi_{L^\star L}^{\mathrm{lin}}$ should be replaced by I_d. This particular target architecture is also known, in the case $L = 1$, as the *hidden manifold model* (Goldt et al., 2020c; Gerace et al., 2020a) and affords a stylized model for structured data. The present paper generalizes these studies to arbitrary depths L. One is then

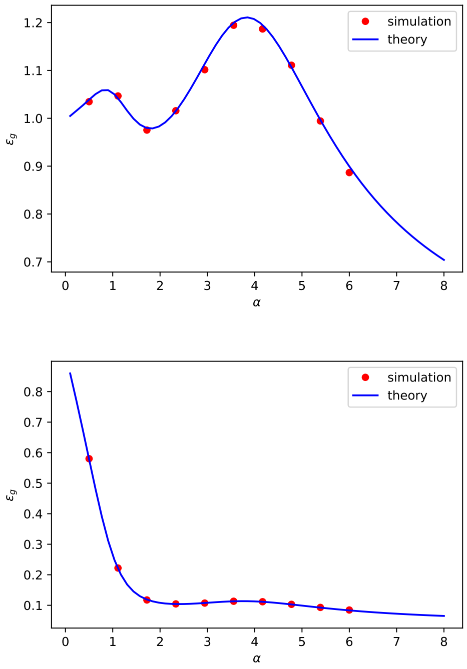


Figure 19: Learning curves $\varepsilon_g(\alpha)$, where $\alpha := {}^{n}/_{d}$ is the sample complexity, for ridge regression ($\sigma_\star = id$, $\ell(y,z) = {}^{1}/_{2}(y-z)^2$, and $g(y,\hat{y}) = (y-\hat{y})^2$). Red dots correspond to numerical simulations on the learning model (157) (180), averaged over 20 runs, in dimension $d = 500$. The solid line correspond to sharp asymptotic characterization provided by conjecture 6.3.3. (left) 2-layers target ($L^\star = 1$, $\sigma_1^\star =$ sign, $\gamma_1^\star = 1$), (right) single-layer target ($L^\star = 0$). Both are learnt with a $2-$hidden layers RF (157) with $\sigma_{1,2}(x) = \tanh(2x)$ activation, widths $\gamma_1 = 8$ and $\gamma_2 = 1$, and regularization $\lambda = 0.001$.

equipped to formulate the following, stronger, conjecture:

Conjecture 6.3.3. *In the same limit as in Conjecture 6.3.2, the test error of the empirical risk minimizer* (159) *trained on* $\mathscr{D} = \{(\boldsymbol{x}^\mu, y^\mu)\}_{\mu\in[n]}$ *with covariates* $\boldsymbol{x}^\mu \sim \mathcal{N}(0_d, \Omega_0)$ *and labels from* (180) *is equal to the one of a Gaussian covariate model* (184) *with the matrices* $\Psi_{L^\star}^{\text{lin}}, \Omega_L^{\text{lin}}, \Phi_{L^\star L}^{\text{lin}}$ (176),(187).

Conjecture 6.3.3 allows to give a fully analytical sharp asymptotic characterization of the test error. Importantly, observe that it also affords compact closed-form formulae for the population covariances $\Omega_L, \Phi_{L^\star L}, \Psi_{L^\star}$. In particular the spectrum of $\Psi_{L^\star}^{\text{lin}}, \Omega_L^{\text{lin}}$ can be analytically computed and compares excellently with empirical numerical simulations. Figs. 19 and 20 present the resulting theoretical curve and contrasts them to numerical simulations in dimensions $d = 1000$, revealing an excellent agreement.

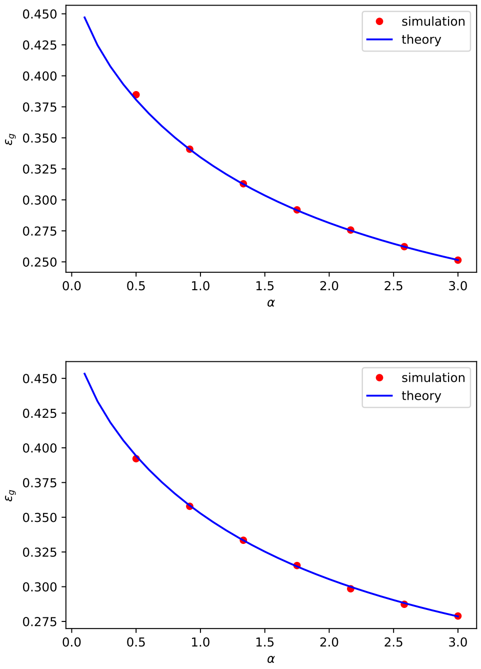


Figure 20: Learning curves $\varepsilon_g(\alpha)$, where $\alpha := n/d$ is the sample complexity, for logistic regression ($\sigma_\star = \text{sign}$, $\ell(y,z) = \ln(1 + e^{-yz})$ and metric $g(y,\hat{y}) = 1 - \Theta(y\hat{y})$). Red dots correspond to numerical simulations on the learning model (157) (180), averaged over 20 runs in dimension $d = 1200$. The solid line correspond to sharp asymptotic characterization provided by conjecture 6.3.3. (left) single-layer target ($L^\star = 0$), (right) two-layer target ($L^\star = 1$, $\sigma_1^\star = \text{erf}$, $\gamma_1^\star = 1$) (180) hidden sign layer. Both are learnt with a depth $L = 2$ dRF (157) with activation $\sigma_{1,2}(x) = \tanh(2x)$, widths $\gamma_1 = \gamma_2 = 5/3$, and regularization $\lambda = 0.05$ (top) and $\sigma_{1,2}(x) = \text{erf}(x)$ and $\lambda = 0.1$ (bottom).

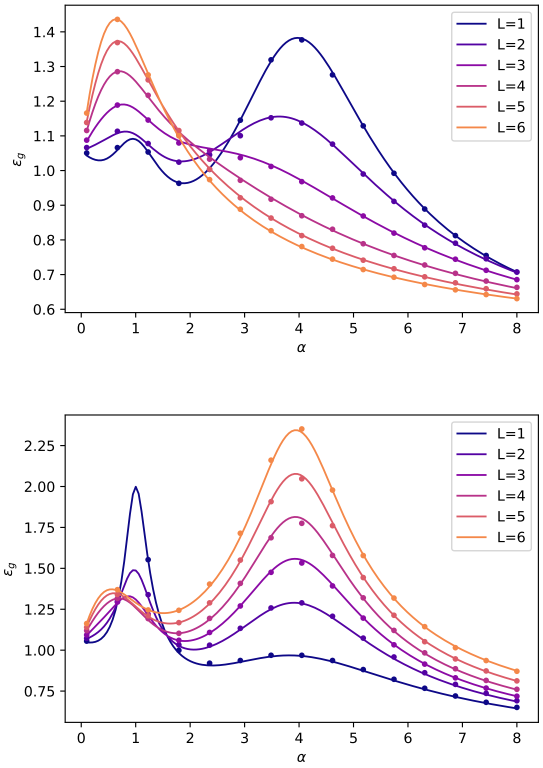


Figure 21: Learning curves for ridge regression on a 1-hidden layer target function ($\gamma_1^\star = 2$, $\sigma_1^\star = \text{sign}$) using a $L-$hidden layers learner with widths $\gamma_1 = ... = \gamma_L = 4$ and $\sigma_{1,...,L} = \tanh$ activation (left) or $\sigma_{1,...,L}(x) = 1.1 \times \text{sign}(x) \times \min(2, |x|)$ clipped linear activation (right), for depths $1 \leq L \leq 6$. The regularization is $\lambda = 0.001$. Solid lines represent theoretical curves evaluated from the sharp characterization of conjecture 6.3.3, while numerical simulations, averaged over 50 runs, are indicated by dots. The linear peak can be observed at $\alpha = 1$ (recall that $\alpha := {}^{n}/_{d}$ is the sample complexity), while the non-linear peak occurs for $\alpha = \gamma = 4$ (D'Ascoli et al., 2021b). Despite sharing the same architecture, the use of different activations induces different implicit regularizations, leading to the linear (resp. non-linear) peak being further suppressed as the depth increases for the clipped linear activation (resp. tanh activation).

6.4 DEPTH-INDUCED IMPLICIT REGULARIZATION

An informal yet extremely insightful takeaway from Conjecture 6.3.3, and in particular the closed-form expressions (176), is that the activations in a deep non-linear dRF (157) share the same population statistics as the activations in a deep *noisy* linear network, with layers

$$\varphi_\ell^{\text{lin}}(\boldsymbol{x}) = \kappa_1^\ell \frac{W_\ell^\top \boldsymbol{x}}{\sqrt{k_{\ell-1}}} + \kappa_*^\ell \xi_\ell, \tag{188}$$

where $\xi_\ell \sim \mathcal{N}(0_{k_\ell}, I_{k_\ell})$ is a Gaussian noise term. It is immediate to see that (188) lead to the same recursion as (176). This observation, which was made in the concomitant work (Cui et al., 2023a), essentially allows to equivalently think of the problem of learning using a dRF (157) as one of learning with linear noisy network. Indeed, Conjecture 6.3.3 essentially suggests that the asymptotic test error depends on the second-order statistics of the last layer acrivations, shared between the dRF and the equivalent linear network. Finally, it is worthy to stress that, while the learner dRF is deterministic conditional on the weights $\{W_\ell\}$, the equivalent linear network (188) is intrinsically stochastic in nature due to the effective noise injection ξ_ℓ at each layer. Statistical common sense dictates that this effective noise injection has a regularizing effect, by introducing some randomness in the learning, and helps mitigating overfitting. Since the effective noise is a product of the propagation through a non-linear layer, this suggest that *adding random non linear layers induces an implicit regularization.* We explore this intuition in this last section.

Observe first that the equivalent noisy linear network (188) reduces to a simple shallow noisy linear model

$$\hat{y}_\theta^{\text{lin}}(\boldsymbol{x}) = \sigma\left(\frac{1}{\sqrt{k}}\theta^\top (A_L \cdot \boldsymbol{x} + \xi_L)\right) \tag{189}$$

where the effective weight matrix A is

$$A_L := \prod_{\ell=1}^{L}\left(\kappa_1^\ell \frac{W_\ell}{\sqrt{k_{\ell-1}}}\right)$$

and the effective noise ξ_L is Gaussian with covariance C_ξ^L

$$C_\xi^L = \sum_{\ell_0=1}^{L-1} (\kappa_*^{\ell_0})^2 \left(\prod_{\ell=\ell_0+1}^{L} \frac{\kappa_1^\ell W_\ell^\top}{\sqrt{k_{\ell-1}}}\right)^\top \left(\prod_{\ell=\ell_0+1}^{L} \frac{\kappa_1^\ell W_\ell^\top}{\sqrt{k_{\ell-1}}}\right) + (\kappa_*^L)^2 I_k.$$

The signal-plus-noise structure of the equivalent linear features (189) has profound consequences on the level of the learning curves of the model (157):

- When $\alpha = 1$, there are as many training samples as the dimension of the data $d-$ dimensional submanifold $A_L\boldsymbol{x}$, resulting in a standard interpolation peak. The noise part ξ_L induces an implicit regularization which helps mitigate the overfitting.

- As $\alpha = \gamma_L$, the number of training samples matches the dimension k_L of the noise, and the *noise* part is used to interpolate the training samples, resulting in another peak. This second peak is referred to as the non-linear peak by (D'Ascoli et al., 2021b).

Therefore, there exists an interplay between the two peaks, with higher noise ξ_L both helping to mitigate the linear peak, and aggravating the non-linear peak. The depth of the network plays a role in that it modulates the amplitudes of the signal part and the noise part, depending on the activation through the recursions (165).

We give two illustrations of the regularization effect of depth in Fig. 21. Two activations are considered : $\sigma_a = \tanh$ (for which the noise level, as measure by $\mathrm{Tr}\,(C)^L_\xi$ decreases with depth) , and a very weakly non-linear activation $\sigma_b(x) = 1.1 \times \mathrm{sign}(x) \times \min(2, |x|)$, corresponding to a linear function clipped between -2.2 and 2.2 (for which $\mathrm{Tr}\,(C)^L_\xi$ increases with depth). Note σ_b is the simplest activation function for which the increase of the noise level with depth was observed. Note that, because for σ_a the effective noise decreases with depth, the linear peak is aggravated for deeper networks, while the non-linear peak is simultaneously suppressed. Conversely, for σ_b, additional layers introduce more noise and cause a higher non-linear peak, while the induced implicit regularization mitigates the linear peak.

6.5 CONCLUSION

We study the problem of learning a deep random network target function by training the readout layer of a deep network, with frozen random hidden layers (deep Random Features). We first prove an asymptotic deterministic equivalent for the conjugate kernel and sample covariance of the activations in a deep Gaussian random networks. This result is leveraged to establish a sharp asymptotic characterization of the test error in the specific case where the learner and teacher networks share the same intermediate layers, and the readout is learnt using a ridge loss. This proves the Gaussian universality of the test error of ridge regression on non-linear features corresponding to the last layer activations. In the fully generic case, we conjecture a sharp asymptotic formula for the test error, for fully general target/learner architectures and convex loss. The formulas suggest that the dRF behaves like a linear noisy network, characterized by an implicit regularization. We explore the consequences of this equivalence on the interplay between the architecture of the dRF and its generalization ability.

7 COLORED DEEP RANDOM FEATURES

Deep neural networks are the backbone of most successful machine learning algorithms in the past decade. Despite their ubiquity, a firm theoretical understanding of the very basic mechanism behind their capacity to adapt to different types of data and generalise across different tasks remains, to a large extent, elusive. For instance, what is the relationship between the inductive bias introduced by the network architecture and the representations learned from the data, and how does it correlate with generalisation? Albeit the lack of a complete picture, insights can be found in recent empirical and theoretical works.

On the theoretical side, a substantial fraction of the literature has focused on the study of deep networks at initialisation, motivated by the lazy training regime of large-width networks with standard scaling. Besides the mathematical convenience, the study of random networks at initialisation have proven to be a valuable theoretical testbed – allowing in particular to capture some empirically observed behaviour, such as the double-decent (Belkin et al., 2019) and benign overfitting (Bartlett et al., 2020b) phenomena. As such, proxys for networks at initialisation, such as the RF model (Rahimi et al., 2007a) have thus been the object of considerable theoretical attention, with their learning being asymptotically characterized in the two-layer case (Goldt et al., 2021a; Goldt et al., 2020c; Gerace et al., 2020a; Hu et al., 2020; Dhifallah et al., 2022; Mei et al., 2019c; Mei et al., 2022b) and the deep case (Zavatone-Veth et al., 2022a; Schröder et al., 2023a; Bosch et al., 2023a; Zavatone-Veth et al., 2023). With the exception of (Gerace et al., 2020a) (limited to two-layer networks) and (Zavatone-Veth et al., 2023) (limited to linear networks), all the analyses for non-linear dRFs assume unstructured random weights. In sharp contrast, the weights of trained neural networks are fundamentally structured - restricting the scope of these results to networks at initialization.

Indeed, an active research direction consists of empirically investigating how the statistics of the weights in trained neural networks encode the learned information, and how this translates to properties of the predictor, such as inductive biases (Thamm et al., 2022; Martin et al., 2021). Of particular relevance to our work is a recent observation by (Guth et al., 2023) that a random (but structured) network with the weights sampled from an ensemble with matching statistics can retain a comparable performance to the original trained neural networks. In particular, for some tasks it was shown that second order statistics suffices – defining a Gaussian *rainbow* *network* ensemble.

Our goal in this manuscript is to provide an exact asymptotic characterization of the properties of *Gaussian rainbow networks*, i.e. deep, non-linear networks with structured random weights. Our **main contributions** are:

- We derive a tight asymptotic characterization of the test error achieved by performing ridge regression with Lipschitz-continuous feature maps, in the high-dimensional limit where the dimension of the features and the number of samples grow at proportional rate. This class of feature maps encompasses as a particular case Gaussian rainbow network features.
- The asymptotic characterization is formulated in terms of the population covariance of the features. For Gaussian rainbow networks, we explicit a closed-form expression of this covariance, formulated as in the unstructured case in Chapter 6 as a simple linear recursion depending on the weight matrices of each layer. These formulae extend similar results of Chapters 6 and 8 for independent and unstructured weights to the case of structured –and potentially correlated– weights.
- We empirically find that our theoretical characterization captures well the learning curves of some networks trained by gradient descent in the lazy regime.

RELATED WORKS

Random features — RFs were introduced in (Rahimi et al., 2007a) as a computationally efficient way of approximating large kernel matrices. In the shallow case, the asymptotic spectral density of the conjugate kernel was derived in (Liao et al., 2018; Pennington et al., 2019; Benigni et al., 2021). The test error was on the other hand characterized in (Mei et al., 2019c; Mei et al., 2022b) for ridge regression, and extended to generic convex losses by (Gerace et al., 2020a; Goldt et al., 2021a; Dhifallah et al., 2022), and in (Liang et al., 2022; Loureiro et al., 2021b; Bosch et al., 2022) for other penalties. RFs have been studied as a model for networks in the lazy regime, see e.g. (Ghorbani et al., 2019c; Ghorbani et al., 2021; Yehudai et al., 2019; Refinetti et al., 2021b);

Deep RFs – Recent work have addressed the problem of extending these results to deeper architectures. In the case of linear networks, a sharp characterization of the test error is provided in (Zavatone-Veth et al., 2022a) for the case of unstructured weights and (Zavatone-Veth et al., 2023) in the case of structured weights. For non-linear RFs, (Schröder et al., 2023a) provides deterministic equivalents for the sample covariance matrices, and (Schröder et al., 2023a; Bosch et al., 2023a) provide a tight characterization of the test error. Deep random networks have been also studied in the context of Gaussian processes by (Lee et al., 2018b; G. Matthews et al., 2018), Bayesian neural networks in (Cohen et al., 2021; Naveh et al., 2021a; Li et al., 2021b; Hanin et al., 2019; Pacelli et al., 2023; Zavatone-Veth et al., 2022a) and inference in (Manoel et al., 2017b; Gabrié et al., 2018b; Aubin et al., 2019; Hand et al., 2018; Aubin et al., 2020b). The recent work of (Guth et al., 2023) provides empirical evidence that for a given trained neural network, a resampled network from an ensemble with matching statistics (*rainbow networks*) might achieve comparable generalization performance, thereby partly bridging the gap between random networks and trained networks.

7.1 SETTING

Consider a supervised learning task with training data $(\mathrm{x}_i, \mathrm{y}_i)_{i\in[n]}$. In this manuscript, we are interested in studying the statistics of linear predictors $f_{\boldsymbol{w}}(\boldsymbol{x}) = \frac{1}{\sqrt{p}}\boldsymbol{w}^\top \varphi(\boldsymbol{x})$ for a class of fixed feature maps $\varphi : \mathbb{R}^d \to \mathbb{R}^p$ and weights $\boldsymbol{w} \in \mathbb{R}^p$ trained via empirical risk minimization:

$$\hat{\boldsymbol{w}}_\lambda = \min_{\boldsymbol{w}\in\mathbb{R}^p} \sum_{i\in[n]} (\mathrm{y}_i - f_{\boldsymbol{w}}(\boldsymbol{x}_i))^2 + \lambda||\boldsymbol{w}||^2. \tag{190}$$

Of particular interest is the generalization error:

$$\mathcal{E}_{\text{gen}}(\hat{\boldsymbol{w}}_\lambda) = \mathbb{E}\left(\mathrm{y} - f_{\hat{\boldsymbol{w}}_\lambda}(\boldsymbol{x})\right)^2 \tag{191}$$

where the expectation is over a fresh sample from the same distribution as the training data. More precisely, our results will hold under the following assumptions.

Assumption 7.1.1 (Labels). *We assume that the labels y_i are generated by another feature map $\varphi_* : \mathbb{R}^d \to \mathbb{R}^k$ as*

$$\mathrm{y}_i = \frac{1}{\sqrt{k}}\boldsymbol{\theta}_*^\top \varphi_*(\boldsymbol{x}_i) + \varepsilon_i, \tag{192}$$

where $\boldsymbol{\varepsilon} \in \mathbb{R}^n$ is an additive noise vector (independent of the covariates $\boldsymbol{x}_i$) of zero mean and covariance $\Sigma := \mathbb{E}\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^\top$, and $\boldsymbol{\theta}_* \in \mathbb{R}^k$ is a deterministic weight vector.

Assumption 7.1.2 (Data & Features). *We assume that the covariates $\boldsymbol{x}_i$ are independent and come from a distribution such that*

- *the feature maps φ, φ_* are centered in the sense $\mathbb{E}\varphi(\boldsymbol{x}_i) = 0$, $\mathbb{E}\varphi_*(\boldsymbol{x}_i) = 0$,*
- *the feature covariances*

$$\begin{aligned}\Omega &:= \mathbb{E}\varphi(\boldsymbol{x}_i)\varphi(\boldsymbol{x}_i)^\top \in \mathbb{R}^{p\times p},\\ \Psi &:= \mathbb{E}\varphi_*(\boldsymbol{x}_i)\varphi_*(\boldsymbol{x}_i)^\top \in \mathbb{R}^{k\times k},\\ \Phi &:= \mathbb{E}\varphi(\boldsymbol{x}_i)\varphi_*(\boldsymbol{x}_i)^\top \in \mathbb{R}^{p\times k},\end{aligned} \tag{193}$$

have uniformly bounded spectral norm.

- *scalar Lipschitz functions of the feature matrices*

$$X := (\varphi(\boldsymbol{x}_1), \dots, \varphi(\boldsymbol{x}_n)) \in \mathbb{R}^{p\times n},$$
$$Z := (\varphi_*(\boldsymbol{x}_1), \dots, \varphi_*(\boldsymbol{x}_n)) \in \mathbb{R}^{k\times n} \tag{194}$$

are uniformly sub-Gaussian.

Assumption 7.1.3 (Proportional regime)**.** *The number of samples n and the feature dimensions p, k are all large and comparable, see 7.2.1 later.*

Remark 7.1.4. *We formulated 7.1.2 as a joint assumption on the covariates distribution and the feature maps. A conceptually simpler but less general condition would be to assume that*

(ii') the covariates $\boldsymbol{x}_i$ are Gaussian with bounded covariance $\Omega_0 := \mathbb{E}\boldsymbol{x}_i\boldsymbol{x}_i^\top$

(iii') the feature maps φ, φ_* are Lipschitz-continuous

instead of 7.1.2.

The setting above defines a quite broad class of problems, and the results that follow in Section 7.2 will hold under these generic assumptions. The main class of feature maps we are interested in are *deep structured feature models.*

Definition 7.1.5 (Deep structured feature model)**.** *For any $L \in \mathbb{N}$ and dimensions $d, p_1, \dots, p_L = p$, let $\varphi_1, \dots, \varphi_L \colon \mathbb{R} \to \mathbb{R}$ be Lipschitz-continuous* activation functions $|\varphi_l(a) - \varphi_l(b)| \lesssim |a-b|$ *applied entrywise, and let $W_1 \in \mathbb{R}^{p_1\times d}, W_2 \in \mathbb{R}^{p_2\times p_1}, \dots$ be deterministic* weight matrices *with uniformly bounded spectral norms, $\|W_l\| \lesssim 1$. We then call*

$$\varphi(\boldsymbol{x}) := \varphi_L\left(W_L\varphi_{L-1}\left(\cdots W_2\varphi_1\left(W_1\boldsymbol{x}\right)\right)\right). \tag{195}$$

a deep structured feature *model.*

Note that 195 defines a Lipschitz-continuous map[1] $\varphi \colon \mathbb{R}^d \to \mathbb{R}^p, \varphi_* \colon \mathbb{R}^d \to \mathbb{R}^k$ and therefore if both φ, φ_* are deep structured feature models (with distinct parameters in general), then 7.1.2 is satisfied whenever the feature maps φ, φ_* are centered[2] with respect to Gaussian covariates $\boldsymbol{x}_i$. As hinted in the introduction we will be particularly interested in one sub-class of 7.1.5 known as *Gaussian rainbow networks.*

Definition 7.1.6 (Gaussian rainbow ensemble)**.** *Borrowing the terminology of (Guth et al., 2023), we define a fully-connected, L-layer* Gaussian rainbow network *as a random variant of 7.1.5 where for each ℓ the hidden-layer weights $W_\ell = Z_\ell C_\ell^{1/2}$ are random matrices with $Z_\ell \in \mathbb{R}^{p_{\ell+1}\times p_\ell}$ having zero mean and i.i.d. variance $1/p_\ell$ Gaussian entries and $C_\ell \in \mathbb{R}^{p_\ell \times p_\ell}$ being uniformly bounded covariance matrices, which we allow to depend on previous layer weights $Z_1, \dots, Z_{l-1}$.*

[1] $\|\varphi(W\boldsymbol{x}) - \varphi(W\boldsymbol{x}')\|^2 = \sum_i \left|\varphi(w_i^\top\boldsymbol{x}) - \varphi(w_i^\top\boldsymbol{x}')\right|^2 \lesssim \sum_i \left|w_i^\top(\boldsymbol{x}-\boldsymbol{x}')\right|^2 = \|W(\boldsymbol{x}-\boldsymbol{x}')\|^2 \lesssim \|\boldsymbol{x}-\boldsymbol{x}'\|^2$

[2] It is sufficent that e.g. ϕ_l is odd, and $\boldsymbol{x}_i$ is centered.

Note that Gaussian rainbow networks above can be seen as a generalization of the deep random features model studied in (Schröder et al., 2023a; Bosch et al., 2023a; Fan et al., 2020) –see Chapter 6 –, with the crucial difference that the weights are structured.

NOTATIONS

For square matrices $A \in \mathbb{R}^{n\times n}$ we denote the averaged trace by $\langle A\rangle := n^{-1}\operatorname{Tr} A$, and for rectangular matrices $A \in \mathbb{R}^{n\times m}$ we denote the Frobenius norm by $\|A\|_F^2 := \sum_{ij} \left|a_{ij}\right|^2$, and the operator norm by $\|A\|$. For families of non-negative random variables $X(n), Y(n)$ we say that X is *stochastically dominated* by Y, and write $X \prec Y$, if for all ε, D it holds that $P(X(n) \geq n^{\varepsilon} Y(n)) \leq n^{-D}$ for n sufficiently large.

7.2 TEST ERROR OF LIPSCHITZ FEATURE MODELS

Under Assumptions 7.1.1 and 7.1.2 the generalization error from (191) is given by

$$\mathscr{E}_{\mathrm{gen}}(\lambda) = \frac{\theta_*^\top \Psi \theta_*}{k} + \frac{\theta_*^\top Z X^\top G \Omega G X Z^\top \theta_*}{kp^2} + \frac{n}{p}\left\langle \frac{X^\top G \Omega G X \Sigma}{p}\right\rangle - 2\frac{\theta_*^\top \Phi^\top G X Z^\top \theta_*}{kp}, \tag{196}$$

in terms of the *resolvent* $G = G(\lambda) := (XX^\top/p + \lambda)^{-1}$.

Our main result is a rigorous asymptotic expression for (196). To that end define, $m(\lambda)$ to be the unique solution to the equation

$$\frac{1}{m(\lambda)} = \lambda + \left\langle \Omega\left(I + \frac{n}{p} m(\lambda)\Omega\right)^{-1}\right\rangle, \tag{197}$$

and define

$$M(\lambda) = \left(\lambda + \frac{n}{p}\lambda m(\lambda)\Omega\right)^{-1} \tag{198}$$

which is the *deterministic equivalent* of the resolvent, $M(\lambda) \approx G(\lambda)$, see 7.2.3 later. The fact that (197) admits a unique solution $m(\lambda) > 0$ which is continuous in λ follows directly from continuity and monotonicity. Moreover, from

$$0 \leq \left\langle \Omega\left(I + \frac{n}{p} m\Omega\right)^{-1}\right\rangle \leq \min\left\{\langle\Omega\rangle, \frac{\operatorname{rank}\Omega}{n}\frac{1}{m}\right\}$$

we obtain the bounds

$$\max\left\{\frac{1}{\lambda+\langle\Omega\rangle},\frac{1-\frac{\operatorname{rank}\Omega}{n}}{\lambda}\right\}\le m(\lambda)\le\frac{1}{\lambda}. \tag{199}$$

We also remark that $m(\lambda)$ depends on Ω only through its eigenvalues $\omega_1,\ldots,\omega_p$, while $M(\lambda)$ depends on the eigenvectors. The asymptotic expression (201) for the generalization error derived below depends on the eigenvalues of Ω, the overlap of the eigenvectors of Ω with the eigenvectors of Φ, and the overlap of the eigenvectors of Ψ,Φ with θ_*.

Theorem 7.2.1. *Under 7.1.1, 7.1.2 and 7.1.3 for fixed $\lambda>0$ we have the asymptotics*

$$\mathscr{E}_{\text{gen}}(\lambda)=\mathscr{E}_{\text{gen}}^{\text{rmt}}(\lambda)+O\left(\frac{1}{\sqrt{n}}\right), \tag{200}$$

in the proportional $n\sim k\sim p$ regime, where

$$\begin{aligned}\mathscr{E}_{\text{gen}}^{\text{rmt}}(\lambda):=&\frac{1}{k}\theta_*^\top\frac{\Psi-\frac{n}{p}m\lambda\Phi(M+\lambda M^2)\Phi^\top}{1-\frac{n}{p}(\lambda m)^2\langle\Omega M\Omega M\rangle}\theta_*\\&+\langle\Sigma\rangle\frac{(\lambda m)^2\frac{n}{p}\langle M\Omega M\Omega\rangle}{1-\frac{n}{p}(\lambda m)^2\langle\Omega M\Omega M\rangle}.\end{aligned} \tag{201}$$

In the general case of comparable parameters we have the asymptotics with a worse error of

$$\frac{1}{\sqrt{\min\{n,p,k\}}}\left(1+\frac{\max\{n,p,k\}}{\min\{n,p,k\}}\right).$$

Remark 7.2.2 (Relation to previous results). *We focus on the misspecified case as this presents the main novelty of the present work. In the wellspecified case $Z=X$ our model essentially reduces to linear regression with data distribution $x=\varphi(\boldsymbol{x})$. There has been extensive research on the generalization error of linear regression, see e.g. in (Bach, 2023; Dobriban et al., 2018b; Bartlett et al., 2021; Cheng et al., 2022) and the references therein.*

1. *We confirm Conjecture 1 of (Loureiro et al., 2021b) in Chapter 3 under 7.1.2. The expression for the error term in 7.2.1 matches the expression presented in Chapter 3 for a GCMT-S model.*

2. *Independently and concurrently to the current work (Latourelle-Vigeant et al., 2023) (partially confirming a conjecture made in (Louart et al., 2018c)) obtained similar results under different assumptions. Most importantly (Latourelle-Vigeant et al., 2023) considers one-layer unstructured random feature models and computes the* empirical generalization error *for a deterministic data set, while we consider general Lipschitz features of random data, and compute the generalization error.*

3. *In the unstructured random feature model (Mei et al., 2022b; Adlam et al., 2020a) obtained an expression for the generalization error under the assumption that the target model is linear or rotationally invariant.*

The novelty of 7.2.1 compared to many of the previous works is, besides the level of generality, two-fold:

1. We obtain a deterministic equivalent for the generalization error involving the population covariance Φ and the sample covariance $XZ^\top$ in the general misspecified setting.
2. Our deterministic equivalent is *anisotropic*, allowing to evaluate(196) for *fixed* targets θ_* and structured noise covariance $\Sigma \neq I$.

Some of the previous rigorous results on the generalization error of ridge regression have been limited to the well-specified case, $X = Z$, since in this particular case the second term of(196) can be simplified to

$$\frac{XX^\top}{p} G\Omega G \frac{XX^\top}{p} = (1 - \lambda G)\Omega(1 - \lambda G). \tag{202}$$

When computing deterministic equivalents for terms as $G\Omega G$, some previous results have relied on the “trick” of differentiating a generalized resolvent matrix $G(\lambda, \lambda') := (XX^\top / p + \lambda'\Omega + \lambda)^{-1}$ with respect to λ'. Our approach is more robust and not limited to expressions which can be written as certain derivatives.

To illustrate 2, the conventional approach in the literature to approximating e.g. the third term on the right hand side of (196) in the case $\Sigma = I$ would be to use the cyclicity of the trace to obtain

$$\begin{aligned} \frac{1}{p^2} \operatorname{Tr} X^\top G\Omega G X &= \frac{1}{p} \operatorname{Tr} G \frac{XX^\top}{p} G\Omega \\ &= \langle G\Omega \rangle - \lambda \langle G^2 \Omega \rangle. \end{aligned} \tag{203}$$

Then upon using (197) and $\langle G\Omega \rangle \approx \langle M\Omega \rangle$, the first term of 203 can be approximated by $1/(\lambda m(\lambda)) - 1$, while for the second term it can be argued that this approximation also holds in derivative sense to obtain

$$\langle G^2 \Omega \rangle = -\frac{\mathrm{d}}{\mathrm{d}\lambda} \langle G\Omega \rangle \approx -\frac{\mathrm{d}}{\mathrm{d}\lambda} \frac{1}{\lambda m(\lambda)} = \frac{\lambda m'(\lambda) + m(\lambda)}{(\lambda m(\lambda))^2}$$

By differentiating (197), solving for m' and simplifying, it can be checked that this result agrees with the second term of (201) in the special case $\Sigma = I$. However, it is clear that any approach which only relies on *scalar* deterministic equivalents is inherently limited in the type of expressions which can be evaluated. Instead, our approach involving *anisotropic deterministic equivalents* has no inherent limitation on the structure of the expressions to be evaluated.

An alternative to evaluating rational expressions of X, Z, commonly used in similar contexts, is the technique of *linear pencils* (Adlam et al., 2020a; Latourelle-Vigeant et al., 2023). The idea here is to represent rational functions of X, Z as blocks of inverses of larger random matrices which depend linearly X,Z. The downside of linear pencils is that even for simple rational expressions the linearizations become complicated, sometimes even requiring the use of computer algebra software for the analysis[3] In comparison we believe that our approach is more direct and flexible.

7.2.1 PROOF OF 7.2.1

The main steps and ingredients for the proof of 7.2.1 consist of the following:

CONCENTRATION: As a first step we establish *concentration estimates* for Lipschitz functions of X,Z and its columns. A key aspect is the concentration of quadratic forms in the columns $x_i := \varphi(\boldsymbol{x}_i)$ of X:

$$\left|x_i^\top A x_i - \mathbb{E} x_i^\top A x_i\right| = \left|x_i^\top A x_i - \operatorname{Tr} \Omega A\right| \prec \|A\|_F$$

which follows from the Hanson-Wright inequality (Adamczak, 2015). The concentration step is very similar to analagous considerations in previous works (Chouard, 2022; Louart et al., 2018d) but we present it for completeness. The main property used extensively in the subsequent analysis is that traces of resolvents with deterministic observables concentrate as

$$|\langle A[G(\lambda) - \mathbb{E}G(\lambda)]\rangle| \prec \frac{\langle |A|^2 \rangle^{1/2}}{n\lambda^{3/2}}. \tag{204}$$

ANISOTROPIC MARCHENKO-PASTUR LAW: As a second step we prove an anisotropic Marchenko-Pastur law for the resolvent G, of the form:

Theorem 7.2.3. *For arbitrary deterministic matrices A we have the high-probability bound*

$$|\langle (G(\lambda) - M(\lambda)A\rangle| \prec \frac{\langle |A|^2 \rangle}{n\lambda^3}, \tag{205}$$

in the proportional $n \sim p$ regime.

Remark 7.2.4. *Tracial Marchenko-Pastur laws (case $A = I$ above) have a long history, going back to (Marchenko et al., 1967) in the isotropic case $\Omega = I$, (Silverstein, 1995) in the general case with separable covariance $x = \sqrt{\Omega} z$ and (Bai et al., 2008a) under quadratic form concentration assumption. Anisotropic Marchenko-Pastur laws under various conditions and with varying precision have been proven e.g. in (Rubio et al., 2011; Chouard, 2022; Louart et al., 2018c; Knowles et al., 2017).*

[3] For instance (Adlam et al., 2020a) used block matrices with up to 16 × 16 blocks in order to evaluate the asymptotic test error.

For the proof of 7.2.3 the resolvent $G := (X^\top X/p + \lambda)^{-1} \in \mathbb{R}^{n\times n}$ of the *Gram matrix* $X^\top X$ plays a key role. The main tool used in this step are the commonly used *leave-one-out identities*, e.g.

$$Gx_i = \lambda G_{ii} G_{-i} x_i, \quad G_{-i} := \Big(\sum_{j\neq i} \frac{x_j x_j^\top}{p} + \lambda\Big)^{-1} \tag{206}$$

which allow to decouple the randomness due the i-th column from the remaining randomness. Such identities are used repeatedly to derive the approximation

$$\mathbb{E}G \approx \Big(\frac{n}{p}\lambda\langle \mathbb{E}G\rangle \Omega + \lambda\Big)^{-1} \tag{207}$$

in Frobenius norm, which, together with the relation $1 - \lambda\langle G\rangle = \frac{p}{n}\big(1 - \lambda\langle G\rangle\big)$ between the traces of G and G, yields a self-consistent equation for $\langle G\rangle$. This self-consistent equation is an approximate version of (197), justifying the definition of m. The *stability* of the self-consistent equation then implies the averaged asymptotic equivalent

$$|m - \langle \mathbb{E}G\rangle| \lesssim \frac{1}{n\lambda^2}. \tag{208}$$

and therefore by 207 finally

$$\|M - \mathbb{E}G\|_F \lesssim \frac{1}{n^{1/2}\lambda^3}, \tag{209}$$

which together with 204 implies 7.2.3.

Compared to most previous anisotropic deterministic equivalents as in (Knowles et al., 2017) we measure the error of the approximation 205 with respect to the Frobenius of the observable A. As in the case of unified local laws for Wigner matrices (Cipolloni et al., 2022) this idea renders the separate handling of quadratic form bound unnecessary, considerably streamlining the proof. To illustrate the difference note that specializing A to be rank-one $A = xy^\top$ in

$$\Big|y^\top (G - M)x\Big| = |\mathrm{Tr}(G - M)A| \prec \begin{cases} \|A\| \\ \langle |A|^2\rangle^{1/2} \end{cases}$$

results in a trivial estimate $\|x\|\|y\|$ in the case of the spectral norm, and in the optimal estimate $\|x\|\|y\|/\sqrt{p}$ in the case of the Frobenius norm.

ANISOTROPIC MULTI-RESOLVENT EQUIVALENTS The main novelty of the current work lies in an asymptotic evaluationof the expressions on the right-hand-side of(196). A key property of the deterministic equivalents is that the approximation is *not* invariant under multiplication.

E.g. for the last term in(196) we have the approximations $G \approx M$ and $\frac{1}{n}XZ^\top = \frac{1}{n}\sum x_i z_i^\top \approx \Phi$, while for the product the correct deterministic equivalent is

$$G\frac{XZ^\top}{n} \approx \lambda m M \Phi, \tag{210}$$

i.e. the is an additional factor of $m\lambda$. In this case the additional factor can be obtained from a direct application of the leave-one-out identity 206 to the product $G\frac{XZ^\top}{n}$, but the derivation of the multi-resolvent equivalents requires more involved arguments. When expanding the multi-resolvent expression $\langle GAGB \rangle$ we obtain an approximative self-consistent equation of the form

$$\begin{aligned}\langle GAGB \rangle \approx &\langle MAMB \rangle \\ &+ \frac{n}{p}(m\lambda)^2 \langle MBM\Omega \rangle \langle GAG\Omega \rangle.\end{aligned}$$

Using a stability analysis this yields a deterministic equivalent for the special form $\langle GAG\Omega \rangle$ which then can be used for the general case. The second term of(196) requires the most carefuly analysis due to the interplay of the multi-resolvent expression and the dependency among Z, X.

7.3 POPULATION COVARIANCE FOR RAINBOW NETWORKS

Theorem 7.2.1 characterizes the test error for learning using Lipschitz feature maps as a function of the three features population (cross-)covariances Ω, Φ, Ψ. For the particular case where both the target and learner feature maps are drawn from the Gaussian rainbow ensemble from 7.1.6, these population covariances can be expressed in closed-form in terms of combinations of products of the weights matrices. Consider two rainbow networks

$$\begin{aligned}\varphi(\boldsymbol{x}) &= \varphi_L(W_L \varphi_{L-1}(\dots \varphi_1(W_1 \boldsymbol{x}))) \\ \varphi_*(\boldsymbol{x}) &= \psi_L(V_L \psi_{L-1}(\dots \psi_1(V_1 \boldsymbol{x})))\end{aligned} \tag{211}$$

with depths L, L. The approach we introduce here is in theory capable of obtaining linear or polynomial approximations to Ω, Φ, Ψ under very general assumptions. However, for definiteness we focus on a class of correlated rainbow networks in which, for all $k \neq j$, the k-th row of W_ℓ is independent from the j-th row of W_ℓ, V_ℓ as this allows for particularly simple expressions for the linearized covariances[4]. Note that we explicitly allow for weights to be correlated across layers.

[4] The identity matrices in 215 are a direct consequence of this assumption. In case of weight matrices with varying row-norms or covariances across rows the resulting expression would be considerably more complicated.

Assumption 7.3.1 (Correlated rainbow networks). *By symmetry we assume without loss of generality $L \leq L$. Furthermore, we assume that*

1. *for $\ell \leq L \leq L$ all the internal widths p_ℓ of W_ℓ, V_ℓ agree,*
2. *for all $\ell \leq L$, the dimensions scale proportionally, i.e. $n \sim d \sim p_\ell$,*
3. *for $\ell \leq L \leq L$ the rows w_ℓ, v_ℓ of W_ℓ, V_ℓ are mean-zero and i.i.d. with*

$$C_\ell := p_\ell \mathbb{E} w_\ell w_\ell^\top,\ C_\ell := p_\ell \mathbb{E} v_\ell v_\ell^\top,\ C_\ell := p_\ell \mathbb{E} w_\ell v_\ell^\top, \tag{212}$$

4. *for two (possibly identical) rows u, z, and for any matrix A, quadratic forms admit concentration, with high probability (w.h.p).*[5]

$$u^\top A z - \mathrm{Tr}\left(A \mathbb{E} z u^\top\right) \lesssim n^{-1/2}, \tag{213}$$

5. *for all $\ell \leq L$, operator norms of (cross-)covariance matrices admit uniform bounds*

$$\|C_\ell\| + \|C_\ell\| + \|C_\ell\| \lesssim 1. \tag{214}$$

Under 7.3.1 the *linearized population covariances* can be defined recursively as follows:

Definition 7.3.2 (Linearized population covariances). *Define the sequence of matrices $\Omega_\ell^{\mathrm{lin}}, \Phi_\ell^{\mathrm{lin}}, \Psi_\ell^{\mathrm{lin}}$ by the recursions*

$$\Omega_\ell^{\mathrm{lin}} = (\kappa_\ell^1)^2 W_\ell \Omega_{\ell-1}^{\mathrm{lin}} W_\ell^\top + (\kappa_\ell^*)^2 I \tag{215}$$

$$\Psi_\ell^{\mathrm{lin}} = (\tilde{\kappa}_\ell^1)^2 V_\ell \Psi_{\ell-1}^{\mathrm{lin}} V_\ell^\top + (\kappa_\ell^*)^2 I \tag{216}$$

$$\Phi_\ell^{\mathrm{lin}} = \kappa_\ell^1 \tilde{\kappa}_\ell^1 W_\ell \Phi_{\ell-1}^{\mathrm{lin}} V_\ell^\top + (\kappa_\ell^*)^2 I, \tag{217}$$

with $\Omega_0^{\mathrm{lin}} = \Psi_0^{\mathrm{lin}} = \Phi_0^{\mathrm{lin}} = \Omega_0$ the input covariance. The coefficients $\{\kappa_\ell^1, \tilde{\kappa}_\ell^1, \kappa_\ell^*, \tilde{\kappa}_\ell^*\}$ are defined by the recursion

$$\kappa_\ell^1 := \mathbb{E}\varphi_\ell'(N_\ell), \quad \kappa_\ell^1 := \mathbb{E}\psi_\ell'(N_\ell) \tag{218}$$

and

$$\begin{aligned} \kappa_\ell^* &= \sqrt{\mathbb{E}[\varphi_\ell(N_\ell)^2] - r_\ell(\kappa_\ell^1)^2} \\ \tilde{\kappa}_\ell^* &= \sqrt{\mathbb{E}[\psi_\ell(N_\ell)^2] - \tilde{r}_\ell(\tilde{\kappa}_\ell^1)^2} \\ \kappa_\ell^* &= \sqrt{\mathbb{E}[\varphi_\ell(N_\ell)\psi_\ell(N_\ell)] - \check{r}_\ell \kappa_\ell^1 \kappa_\ell^1}, \end{aligned} \tag{219}$$

where N_ℓ, N_ℓ are jointly mean-zero Gaussian with $\mathbb{E}N_\ell^2 = r_\ell$, $\mathbb{E}N_\ell^2 = r_\ell$, $\mathbb{E}N_\ell N_\ell = r_\ell$, with

$$r_\ell = \mathrm{Tr}\left[C_\ell \Omega_{\ell-1}^{\mathrm{lin}}\right],\ r_\ell = \mathrm{Tr}\left[C_\ell \Psi_{\ell-1}^{\mathrm{lin}}\right],\ r_\ell = \mathrm{Tr}\left[C_\ell^\top \Phi_{\ell-1}^{\mathrm{lin}}\right].$$

[5] This concentration holds in particular when rows u, z are Lipschitz concentrated with constant $O(n^{-1/2})$.

Finally, for $\tilde{L} \geq \ell \geq L+1$, define

$$\Phi_\ell^{\text{lin}} = \kappa_\ell^1 \Phi_{\ell-1}^{\text{lin}} W_\ell^\top, \tag{220}$$

with still $\kappa_\ell^1, \kappa_\ell^*$ just as before, and Ψ_ℓ^{lin} with the same recursion (215).

Conjecture 7.3.3. *The populations covariances Ω, Φ, Ψ involved in Theorem 7.2.1 can be asymptotically approximated with the last iterates of the linear recursions of Definition 7.3.2, i.e.*

$$\left\|\Omega - \Omega_L^{\text{lin}}\right\|_F + \left\|\Psi - \Psi_{\tilde{L}}^{\text{lin}}\right\|_F + \left\|\Phi - \Phi_{\tilde{L}}^{\text{lin}}\right\|_F \lesssim 1 \tag{221}$$

Note that the linearization from 7.3.2 also provides good approximation to the population covariances $\Omega_\ell, \Phi_\ell, \Psi_\ell$ of the post-activations at intermediate layers ℓ. The method we use to rigorously derive the linearizations is in theory applicable to any depths, however the estimates quickly become tedious. To keep the present work at a manageable length we provide a rigorous proof of concept only for the simplest multi-layer case.

Theorem 7.3.4. *Under 7.3.1 with $L = 1, L = 2$ we have*

$$\begin{aligned} \left\|\Omega_1 - \Omega_1^{\text{lin}}\right\|_F + \left\|\Psi_1 - \Psi_1^{\text{lin}}\right\|_F + \left\|\Phi_1 - \Phi_1^{\text{lin}}\right\|_F &\lesssim 1 \\ \left\|\Psi_2 - \Psi_2^{\text{lin}}\right\|_F + \left\|\Phi_2 - \Phi_2^{\text{lin}}\right\|_F &\lesssim 1 \end{aligned}$$

with high probability.

Remark 7.3.5 (Comparison). *The approach we take here is somewhat different from previous works (Schröder et al., 2023a; Fan et al., 2020; Chouard, 2023) on (multi-layer) RF models. In these previous results, the deterministic equivalent for the resolvent was obtained using primarily the randomness of the weights, resulting in relatively stringent assumptions (Gaussianity and independence between layers). This layer-by-layer recursive approach resulted in a deterministic equivalent for the resolvent which is* consistent *with a sample covariance matrix with linearized population covariance. Here we take the direct approach of considering feature models with arbitrary structured features, and then linearize the population covariances in a separate step for RF.*

7.3.1 PROOF OF 7.3.4

WIn the proof, we crucially rely on the theory of Wiener chaos expansion and Stein's method (see (Nourdin et al., 2012)). Gaussian Wiener chaos is a generalization of Hermite polynomial expansions, which previously have been used for approximate linearization (Fan et al., 2020; Schröder et al., 2023a) in similar contexts. The basic idea is to decompose random variables $F = F(\boldsymbol{x})$ which are functions of the Gaussian random vector $\boldsymbol{x}$, into pairwise uncorrelated components

$$F = \mathbb{E}F + \sum_{p\geq 1} I_p\left(\frac{\mathbb{E}D^p F}{p!}\right), \tag{222}$$

where I_p is a so called *multiple integral* (generalizing Hermite polynomials) and D^p is the p-th Malliavin derivative. By applying this to the one-layer quantities $\varphi_1(w^\top \boldsymbol{x}), \psi_1(u^\top \boldsymbol{x})$ we obtain, for instance

$$\begin{aligned} &\mathbb{E}\varphi_1(w^\top \boldsymbol{x})\psi_1(v^\top \boldsymbol{x}) \\ &\quad = \sum_{p\geq 1} \frac{1}{p!}\mathbb{E}\varphi_1^{(p)}(w^\top \boldsymbol{x})\mathbb{E}\psi_1^{(p)}(u^\top \boldsymbol{x})\langle w, v\rangle^p, \end{aligned} \tag{223}$$

which for independent w, v we can truncate after $p = 1$, giving rise to the linearization.

For the multi-layer case we combine the chaos expansion with Stein's method in order to prove *quantitative central limit theorems* of the type

$$d_W(F, N) \lesssim \mathbb{E}\left|\mathbb{E}F^2 - \langle DF, -DL^{-1}F\rangle\right| \tag{224}$$

for the Wasserstein distance d_W, where

$$F := w^\top \varphi_1(W\boldsymbol{x}), \quad N \sim \mathcal{N}(0, \mathbb{E}F^2), \tag{225}$$

and L^{-1} is the pseudo-inverse of the *generator of the Ornstein–Uhlenbeck semigroup*.

7.3.2 DISCUSSION OF THEOREM 7.3.4

The population covariances thus admit simple approximate closed-form expressions as linear combinations of products of relevant weight matrices. These expressions generalize similar linearizations introduced in (Cui et al., 2023a; Schröder et al., 2023a; Bosch et al., 2023a; Fan et al., 2020; Chouard, 2023) and discussed in Chapters 6 and 8 for the case of weights which are both unstructured and independent, and iteratively build upon earlier results for the two-layer case developed in (Mei et al., 2019c; Gerace et al., 2020a; Goldt et al., 2021a; Hu et al., 2020). In fact, the expressions leveraged in these works can be recovered as a special case for $C_\ell = \tilde{C}_\ell = I$ (isotropic weights) and $\check{C}_\ell = 0$ (independence). Importantly, note that possible correlation between weights across different layers do not enter in the reported expressions. In practice, we have observed in all probed settings the test error predicted by Theorem 7.2.1, in conjunction with the linearization formulae for the features covariance, to match well numerical experiments.

Figure 22 illustrates a setting where many types of weights correlations are present. It represents the learning curves of a four-layer Gaussian rainbow network with feature map $\tanh(W_3 \tanh(W_2 \tanh(W_1 \boldsymbol{x})))$, learning from a two-layer target $\theta_*^\top \tanh(W\boldsymbol{x})$. To illustrate our result, we consider both target/student correlations $C_1 = 1/2 I$, and inter-layer correlations $W_1 = W_2$. We furthermore took the covariance of the third layer to depend on the weights of the first layer, $C_3 = (W_1 W_1^\top + 1/2 I)^{-1}$. In order to have structured

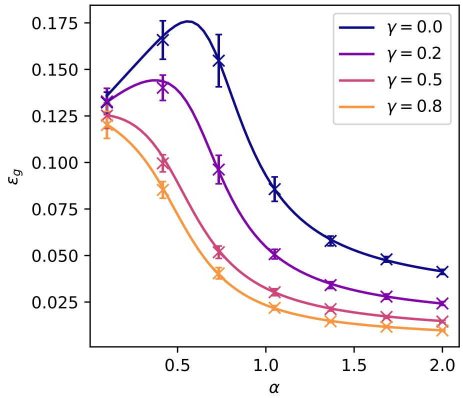


Figure 22: Test error for a target $\theta_*^\top \tanh(W_* x)$, when learning with a four-layer Gaussian rainbow network with feature map $\varphi(x) = \tanh(W_3 \tanh(W_2 \tanh(W_1 x)))$. All width were taken equal to the input dimension d, and the regularization employed is $\lambda = 10^{-4}$. The student weights are correlated across layers, with $W_1 = W_2$, and the covariance C_3 of W_3 depending on W_1 as $C_3 = (W_1 W_1^\top + 1/2 I)^{-1}$. Target/student correlations are also present, with $\check{C}_1 = 1/2 I$. The covariances $C_1, C_2, \tilde{C}_1$ were finally taken to have a spectrum with power-law decay, parametrized by γ. Solid lines: theoretical prediction of Theorem 7.2.1, in conjunction with the closed-form expression for the features population covariance of Definition 7.3.2. Crosses : numerical simulations in $d = 1000$. All experimental points were averaged over 20 instances, with error bars representing one standard deviation. Different colors represent different values for the parameter γ, with small (large) values indicating slow (fast) covariance eigenvalue decay.

weights, the covariances C_1, C_1, C_2 were chosen to have a power-law spectrum. Note that despite the presence of such non-trivial correlations, the theoretical prediction of Theorem 7.2.1 using the linearized closed-form formulae of Def. 7.3.2 for the features covariances (solid lines) captures compellingly the test error evaluated in numerical experiment (crosses).

Finally, we note that akin to (Schröder et al., 2023a), as a consequence of the simple linear recursions, it follows that the Gaussian rainbow network feature map φ shares the same second moments, and thus by Theorem 7.2.1 the same test error, as an equivalent *linear stochastic* network $\varphi^{\text{lin}} = \psi_L \circ \cdots \circ \psi_1$, with

$$\psi_\ell(x) = \kappa_\ell^1 W_\ell x + \kappa_\ell^* \xi_\ell \tag{226}$$

where $\xi_\ell \sim \mathcal{N}(0, I)$ a stochastic noise. This equivalent viewpoint has proven fruitful in yielding insights on the implicit bias of RFs (Schröder et al., 2023a; Jacot et al., 2020a) and on the fundamental limitations of DNN in the proportional regime (Cui et al., 2023a). In the 7.4 we push this perspective further, by heuristically finding that the linearization and Theorem 7.2.1 can also describe deterministic networks trained with gradient descent in the lazy regime.

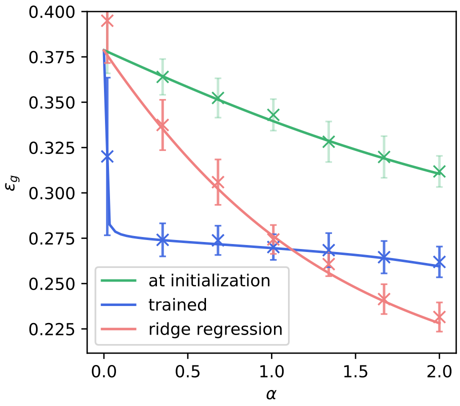


Figure 23: Crosses : Test error when training the readout layer only of a tanh-activated three-layer neural network at initialization (green) and after training (blue), using the `Pytorch` implementation of the full-batch Adam (Kingma et al., 2014a) optimizer, over 3000 epochs with leraning rate 10^{-4} and $n_0 = 1400$ samples, in dimension $d = 1000$. (red): ridge regression. The data is sampled from an isotropic Gaussian distribution. In all training procedures, an ℓ_2 optimization was employed, and the strength thereof optimized over using cross-validation. Solid lines represent the theoretical prediction of Theorem 7.2.1, using the linearized formulae of Definition 7.3.2 for the features population covariances Ω, Ψ, Φ. Crosses represent numerical experiments. Each simulation point is averaged over 10 instances, with error bars representing one standard deviation.

7.4 LINEARIZING TRAINED NEURAL NETWORKS

The previous discussion addressed feature maps associated to random Gaussian networks. However, note that the linearization itself only involves products of the weights matrices, and coefficient depending on weight covariances which can straightforwardly be estimated therefrom. The linearization 7.3.2 can thus be readily heuristically evaluated for feature maps associated to deterministic *trained* finite-width neural networks. As we discuss later in this section, the resulting prediction for the test error captures well the learning curves when re-training the readout weights of the network in a number settings. Naturally, such settings correspond to lazy learning regimes (Jacot et al., 2020a), where the network feature map is effectively *linear*, thus little expressive. However, these trained feature map, albeit linear, can still encode some inductive bias, as shown by (Ba et al., 2022b) for one gradient step in the shallow case. In this section, we briefly explore these questions for fully trained DNN, through the lens of our theoretical results.

Fig. 23 contrasts the test error achieved by linear regression (red), and regression on the feature map associated to a three-layer student at initialization (green) and after 3000 epochs of end-to-end training using full-batch Adam (Kingma et al., 2014a) at learning rate 10^{-4} and weight decay 10^{-3} over $n_0 = 1400$ training samples (blue). For all curves, the readout weights were trained using ridge regression, with regularization strength optimized over using cross-validation. Solid curves indicate the theoretical predictions of Thm. 7.2.1 leveraging the closed-form linearized formulae 7.3.2 for the features covariance. Interestingly, even for the deterministic trained network features, the formula captures the learning curve well. This observation temptingly suggests to interpret the feature map $\varphi(x)$ as the stochastic linear map

$$\varphi^g(x) = W_{\text{eff.}} x + C_{\text{eff.}}^{1/2} \xi \tag{227}$$

where $W_{\text{eff.}} \in \mathbb{R}^{p \times d}$ is proportional to the product of all the weight matrices and $\xi \sim \mathcal{N}(0, I)$ is a stochastic noise colored by the covariance

$$C_{\text{eff.}} \equiv \sum_{\ell=1}^{L-1} \left(\kappa_\ell^* \prod_{s=\ell+1}^{L} \kappa_s^1 \right)^2 \hat{W}_L \ldots \hat{W}_{\ell+1} \hat{W}_{\ell+1}^\top \ldots \hat{W}_L^\top + (\kappa_L^*)^2 I. \tag{228}$$

We denoted $\{\hat{W}_\ell\}_{1 \leq \ell \leq L}$ the trained weights. Note that the effective linear network (227) simply corresponds to the composition of the equivalent stochastic linear layers (226). A very similar expression for the covariance of the effective structured noise (228) appeared in (Schröder et al., 2023a) for the random case with unstructured and untrained random weights. The effective linear model (227) affords a concise viewpoint on a deep finite-width non-linear network trained in the lazy regime. On an intuitive level, during training, the network effectively tunes the two matrices $W_{\text{eff.}}, C_{\text{eff.}}$ which parametrize the effective model (227). The effective weights $W_{\text{eff.}}$ controls the (linear) representation of the data, while the colored noise $C_{\text{eff.}}^{1/2} \xi$ in (227) can be loosely interpreted as inducing an effective regularization.

In fact, despite the fact that all three feature maps represented in Fig. 23 are effectively just linear feature maps, they can still encode very different biases, yielding different phenomenology. In particular, remark that the trained feature map (blue) is outperformed by mere ridge regression (red) at large sample complexities, despite the former having been priorly trained on n_0 additional samples – suggesting the trained weights $W_{\text{eff.}}, C_{\text{eff.}}$ learned some form of inductive bias which is helpful at small and moderate sample complexities, but ultimately harmful for large sample complexities.

Part III b.

TRAINED FEATURES

8 BAYES-OPTIMAL LEARNING OF A RANDOM NETWORK

Learning with DNN has proven to be an extraordinarily versatile tool to approximate (learn) non-trivial functions from data. Many fundamental theoretical questions, however, remain open. For instance, the determination, for a given target function, of just *how many* training data samples are needed in order to learn the target to a given precision? This is tantamount to determining the minimal error that can be achieved from a training set of a given size.

While for a generic target function and generic training set this question is very challenging, valuable insight can be accessed by studying simplified settings with Gaussian input data and specific target functions with known functional forms. Of particular interest is the rich class of functions given by *r*andom neural networks. The lowest achievable test error is known to be obtained through Bayesian inference of the parameters of the target function, assuming (as we will) the distribution of the parameters is given. The *B*ayes-optimal test error corresponds to the information-theoretically minimal test error that any algorithm can achieve. In the context of Gaussian data, with target functions being *s*ingle-layer random neural networks the problem was studied as early as in (Opper et al., 1991a; Seung et al., 1992; Watkin et al., 1993; Schwarze, 1993). More recently, (Barbier et al., 2019a) provide a rigorous characterization of the Bayes-optimal error in the asymptotic proportional regime, where the number of samples is proportional to the input dimension and both of them are large with a fixed ratio α. These results were then extended in (Schwarze, 1993; Aubin et al., 2018a) to neural networks with one *n*arrow hidden layer, whose width remains of order one in the above limit of large dimension and a proportional number of samples.

In practice, neural networks are trained using ERM methods, and it is hence also important to know whether those methods are able to achieve the Bayes-optimal error. (Thrampoulidis et al., 2018; Montanari et al., 2019; Hastie et al., 2022; Mei et al., 2019c; Aubin et al., 2020a; Loureiro et al., 2021b), between others, addressed this question for Gaussian data, providing closed-form formulas for the ERM test error for generalized *l*inear models for target functions corresponding to single-layer neural network with random weights from a number of samples proportional to the dimension.

Here, we pursue these lines of work and study a target function given by a *n*on-linear DNN with random weights, in the limit where the layers-widths and the input dimension are comparably large, hereafter referred to as the *extensive-width* regime. We call such target function the *deep extensive-width random network*. We consider Gaussian input data. Our main question is the characterization of the test error that can be achieved information-theoretically from a given number of samples, as well as its reachability with ERM approaches. While the assumptions of Gaussian input data and the prescribed target function seem far from current machine-learning practice, from a theoretical point of view, these questions remain challenging and widely open even in such a simplified setting (even for a single hidden layer). It is hard to imagine that we could obtain a plausible theory of deep learning without being able to answer such questions first.

Main contributions For the target function corresponding to the deep extensive-width random network and random Gaussian input data we obtain the following results:

- We conjecture a closed-form characterization for the asymptotic Bayes-optimal error, for regression and classification tasks, in the proportional regime where the number of samples n scales linearly with the input dimension d.

- A fundamental step in our derivation, of independent interest, is the deep (Bayes) Gaussian Equivalence Property (GEP) , which specifies the Gaussian statistics of the output of deep networks whose weights are Gaussian, or sampled from the Bayes posterior. We show how the GEP follows from Bayes theory and the asymptotic concentration of random variables in the proportional regime.

- We contrast the Bayes-optimal test error to test errors achieved by simple ERM methods. For regression, ridge and kernel regression are found to achieve the Bayes-optimal mean-squared error, provided they are optimally regularized. An explicit formula for optimal regularization is provided. These results establish that it is impossible to learn more than a linear estimator of the target extensive-width network from linearly many samples. In the case of classification, logistic and ridge classification are found to yield test errors close (but not equal) to the Bayes error.

- We provide a numerical exploration of the regime where the number of samples n tends to infinity *faster* than linearly with the input dimension d, in which the deep (Bayes) GEPs can no longer be employed. We show that ridge and kernel methods then cease to be optimal, while gradient-trained neural networks manage to almost perfectly learn the target, evidencing the superiority of neural nets.

8.0.1 RELATED WORKS

Bayesian learning of neural networks It is well known that Bayesian learning using networks of infinite width (i.e. width much larger than the number of samples and the input dimension) is equivalent to kernel regression (Neal, 1996b; Lee et al., 2018c; Lee et al., 2019; G. Matthews et al., 2018; Hron et al., 2020). A theoretical analysis for extensive-width, however, proved for a long time a challenging endeavor. (Yaida, 2019; Roberts et al., 2021; Zavatone-Veth et al., 2022b) computed (perturbative) first-order corrections to the mean test error with respect to the infinite width limit, but only accommodate a finite number of training samples. The recent work of (Zavatone-Veth et al., 2022a; Hanin et al., 2022) respectively provide an asymptotic and non-asymptotic study of Bayesian learning, but are limited to linear activations. (Li et al., 2021a) and (Ariosto et al., 2022b) conjecture that in the proportional regime, i.e. $n \sim d$, the estimator yielded by extensive-width networks with ReLU or sign activations is still given by the associated Gaussian Process (GP) kernel, with the width only rescaling the variance term in the test error. We note that these works rely on a heuristic Gaussianity assumption and provide expressions depending explicitly on the entire dataset. Here instead we address specifically the Bayes-optimal performance for a random network target function and Gaussian inputs, which allows us to provide closed-form *scalar* formulae and leverage the principled GEP to characterize the statistics. Finally, while all the previously cited work study the case of Bayesian regression with a square log likelihood, the present work also covers classification settings.

Replica method in ML The replica method has been applied in a sizeable body of work to access asymptotic characterizations of the test error (Bayes or ERM) in a variety of setups (Seung et al., 1992; Watkin et al., 1993). While being heuristic, its predictions have been proven rigorously in many cases, e.g. (Talagrand, 2006; Barbier et al., 2019a). This toolbox has been successfully deployed to analyze architectures with one trainable layer, including generalized linear models (**cui2021large**; Advani et al., 2016; Aubin et al., 2020a; Maillard et al., 2020b; Loureiro et al., 2021b), narrow networks with frozen readout (Aubin et al., 2018a), RF (Gerace et al., 2020a) and kernel methods (Canatar et al., 2021b; Cui et al., 2021; Cui et al., 2023c). Recently (Zavatone-Veth et al., 2022a) studied the multiple layers case, in the framework of linear networks. Here we go a step further and analyze deep non-linear networks.

The proportional regime The proportional $n \sim d$ regime has been investigated for shallow networks in a sizeable body of work, leveraging tools like the convex gaussian minimax theorem (Thrampoulidis et al., 2018; Aubin et al., 2020a; Loureiro et al., 2021b; Montanari et al., 2019), random matrix theory (El Karoui, 2008b; Pennington et al., 2019; Louart et al., 2018b) or approximate message-passing (Aubin et al., 2018a; Gabrié, 2019), in addition to the replica method.

Gaussian Equivalence The equivalence between the asymptotic test error of simple ERM algorithms with that of the associated problem where the data samples are replaced by Gaussian covariates with matching population covariance has been observed in many situations, starting with the seminal work (El Karoui, 2008b) on kernel matrices. In particular, (Goldt et al., 2021a; Goldt et al., 2022b; Montanari et al., 2022b; Hu et al., 2020) have proven a *Gaussian Equivalence* principle that shows that, in the proportional regime, one can often replace projected data with Gaussian ones. Such equivalences were used, for instance, in Chapter 3 to characterize the ERM test error in a variety of setups, in terms solely of the population covariances of the target/learner networks. Concomitant works (Schröder et al., 2023b; Bosch et al., 2023b) characterize the Gaussian universality of the test error of deep learners with fixed random weights and trainable readout.

8.1 SETTING

We consider the problem of learning from a train set $\mathscr{D} = \{\boldsymbol{x}^\mu, y^\mu\}_{\mu=1}^n$, with n independently sampled Gaussian covariates $\boldsymbol{x}^\mu \in \mathbb{R}^d \sim \mathscr{N}(0, \Sigma)$. The covariance Σ is assumed to admit a well-defined limiting spectral distribution μ as $d \to \infty$ with finite non-zero first and second moments. The labels y^μ are assumed to be generated by a L-layers DNN with random weights. Denoting $\xi \sim \mathscr{N}(0, \Delta)$ a Gaussian additive output noise, we have

$$y^\mu = f_\star\left[\frac{1}{\sqrt{k_L}}\mathbf{a}_\star^\top \underbrace{(\varphi_L^\star \circ \cdots \circ \varphi_2^\star \circ \varphi_1^\star)}_{L}(\boldsymbol{x}^\mu) + \xi\right], \tag{229}$$

$$\text{with layers } \varphi_\ell^\star(\boldsymbol{x}) = \sigma_\ell\left(\frac{1}{\sqrt{k_{\ell-1}}} W_\ell^\star \cdot \boldsymbol{x}\right).$$

$(\sigma_\ell)_{\ell=1}^L$ is a sequence of activation functions, which are assumed to be odd for simplifying technical reasons. The readout function $f_\star(\cdot)$ will be taken to be the identity function for regression, and the sign function for classification. The width of the ℓ-th layer is denoted k_ℓ, and the associated weight matrix is $W_\ell^\star \in \mathbb{R}^{k_\ell \times k_{\ell-1}}$, with elements sampled i.i.d from $\mathscr{N}(0, \Delta_\ell)$. Similarly, the readout weight vector $a_\star$ is sampled from $\mathscr{N}(0, \Delta_a \mathbb{I}_{k_L})$.

We wish to characterize the Bayes-optimal test errors when learning on data produced by the target function (229). We consider that all the hyperparameters $L, k_\ell, \sigma_\ell, \Delta_a, \Delta_\ell, \Delta$ of the architecture (229) are known, but the weights $a_\star, \{W_\ell^\star\}_{\ell=1}^L$ are not known to the learner.

Throughout Sec. 8.2 to 8.3, we consider the *proportional regime*: the high-dimensional asymptotic limit where $\forall \ell,\ n, d, k_\ell \to \infty$ with fixed $\mathscr{O}(1)$ ratios

$\alpha \equiv {}^{n}/_{d}$ and $\gamma^{\star}_{\ell} \equiv {}^{k_\ell}/_{d}$. The parameters $L, \Delta_\ell, \Delta_a, \Delta$ are assumed to be $\mathcal{O}(1)$. The *quadratic regime* $n \sim d^2 \sim k_\ell^2$ is numerically explored in Sec. 8.4. It is known that learning a target of large width k (resp. using a network of large width) with a finite number of samples $n = \mathcal{O}_k(1)$ simplifies drastically to the problem of learning a Gaussian process (resp. kernel regression) (Neal, 1996b; Lee et al., 2018c; G. Matthews et al., 2018). We consider here widths $\{k_\ell\}_{\ell=1}^{L}$ at most *comparable*, and not very large compared to, the input dimension d and the number of samples n, which makes for a non-trivial, and much richer, learning problem.

8.2 BAYES-OPTIMAL ERROR

The Bayes-optimal error for data generated using the target function (229) is achieved by sampling the weights $\mathbf{a}, \{W_\ell\}_\ell$ from a posterior measure involving a neural network of matching architecture. We thus define

$$\hat{y}(\boldsymbol{x}) = \frac{1}{\sqrt{k_L}} \mathbf{a}^\top \underbrace{(\varphi_L \circ \varphi_{L-1} \circ \cdots \circ \varphi_2 \circ \varphi_1)}_{L}(\boldsymbol{x}), \tag{230}$$

$$\text{with layers } \varphi_\ell(\boldsymbol{x}) = \sigma_\ell\left(\frac{1}{\sqrt{k_{\ell-1}}} W_\ell \cdot \boldsymbol{x}\right). \tag{231}$$

The Bayes-optimal MSE is then

$$\varepsilon^{\mathrm{BO}}_{g,\mathrm{reg}} = \mathbb{E}_{\mathscr{D},\{W^\star_\ell\}_{\ell=1}^{L},\mathbf{a}_\star} \mathbb{E}_{\boldsymbol{x},y}\left[\left(y - \langle \hat{y}(\boldsymbol{x})\rangle_{\mathbf{a},\{W_\ell\}_{\ell=1}^{L}\sim\mathbb{P}}\right)^2\right] \tag{232}$$

where $\boldsymbol{x}, y$ should be understood as a test sample. The Bayes-optimal classification error (defined as the probability to wrongly classify a test sample) is given by

$$\varepsilon^{\mathrm{BO}}_{g,\mathrm{class}} = \mathbb{E}_{\mathscr{D},\{W^\star_\ell\}_{\ell=1}^{L},\mathbf{a}_\star} \mathbb{P}_{\boldsymbol{x},y}\left[y \neq \mathrm{sign}\left(\langle \mathrm{sign}(\hat{y}(\boldsymbol{x}))\rangle_{\mathbf{a},\{W_\ell\}_{\ell=1}^{L}\sim\mathbb{P}}\right)\right]. \tag{233}$$

In (232,233), the learner network is averaged over the posterior

$$\mathbb{P}\left[\mathbf{a}, \{W_\ell\}_{\ell=1}^{L} | \mathscr{D}\right] \propto e^{-\frac{||\mathbf{a}||^2}{2\Delta_a} - \sum_\ell \frac{||W_\ell||_F^2}{2\Delta_\ell}} \prod_{\mu=1}^{n} \int \frac{d\xi e^{-\frac{1}{2\Delta}\xi^2}}{\sqrt{2\pi\Delta}} \delta\left[y^\mu - f_\star(\hat{y}(\boldsymbol{x}^\mu) + \xi)\right]. \tag{234}$$

The Bayes errors (232) and (233) provide information-theoretic lower bounds on the test error for learning the target (229), in the sense that no learning algorithm can reach better performance when learning from the dataset $\mathscr{D}$.

Accessing numerically the Bayes errors (232) and (233) requires sampling an $\mathcal{O}(d^2)$-dimensional distribution, a difficult task. It is on the other hand possible to theoretically derive closed-form formulas using the replica method (Parisi, 1979b; Mézard et al., 2009) that allows characterizing the Bayes error in terms of the moments of independent instances of $\hat{y}(\boldsymbol{x})$ (the eponymous replicas) drawn from the posterior eq. (234). In the replica calculation, one averages over the randomness in the model and in order to be able to carry out such averages in a closed form, the GEP described in the next section is crucial.

8.2.1 THE BAYESIAN GAUSSIAN EQUIVALENCE PROPERTY

A seminal step in our analytical approach is the property that we can replace the statistics of the output $\hat{y}(\boldsymbol{x})$ with respect to the randomness of the input $\boldsymbol{x}$ by Gaussian, with a covariance depending linearly on the covariance of the weight matrices W_l. In fact, (Li et al., 2021a; Ariosto et al., 2022b) do rely on a related Gaussianity assumption, which (Ariosto et al., 2022b) heuristically justify for $L = 1$ using the Breuer-Major theorem, for generic datasets. Since in the present work, we consider the specific Bayes-optimal setting, we are in a position to state a more principled conjecture which follows from the GEP (Goldt et al., 2020c) and the Nishimori identities (Nishimori, 2001; Iba, 1998).

Conjecture 8.2.1. *(**Shallow Bayes GEP**) Consider $\boldsymbol{x}$ a random Gaussian vector. Then for $L=1$, in the extensive-width asymptotic limit $d, k_1 \to \infty$ with fixed $\mathcal{O}(1)$ ratio $\gamma_1 = {}^{k_1}/_{d}$, any finite number of replicas $\hat{y}^1(\boldsymbol{x}; W_1^1, \mathbf{a}^1), \ldots, \hat{y}^s(\boldsymbol{x}; W_1^s, \mathbf{a}^s)$ independently drawn from the Bayes posterior (234) are jointly Gaussian. Furthermore, their correlation reads $\mathbb{E}_{\boldsymbol{x}} \hat{y}^a(\boldsymbol{x}) \hat{y}^b(\boldsymbol{x}) = {}^{\mathbf{a}^{a\top} \Omega_1^{ab} \mathbf{a}^b}/_{k_1}$ where Ω_1^{ab} is the population covariance of the last layer post-activations $\mathbb{E}_{\boldsymbol{x}} \varphi_1^a(\boldsymbol{x}) \varphi_1^b(\boldsymbol{x})^\top$ that reads*

$$\Omega_1^{ab} = \left(\kappa_1^{(1)}\right)^2 \frac{W_1^a \Sigma W_1^{b\top}}{d} + \delta_{a,b} \left(\kappa_*^{(1)}\right)^2 \mathbb{I}_{k_1}, \tag{235}$$

where $\kappa_1^{(1)} = {}^{\mathbb{E}_z[z\sigma_1(z)]}/_{r_1}$ and $(\kappa_*^{(1)})^2 = \mathbb{E}_z\left[\sigma_1(z)^2\right] - r_1 (\kappa_1^{(1)})^2$, with $r_1 = \Delta_1 {}^{\mathrm{Tr}\Omega_1}/_{d}$ and $z \sim \mathcal{N}(0, r_1)$.

We now explain how Conjecture 8.2.1 is motivated. In the proportional regime, for $a = b$, conditional on the matrix W_1, the Gaussian Equivalence Theorem of (Goldt et al., 2021a; Hu et al., 2020; Montanari et al., 2022b) prove indeed that the model (230) for $L = 1$ shares the same second-order post-activation statistics as the noisy *linear* network $\hat{y} = \mathbf{a}^T(\kappa_1^{(1)} {}^{W_1 \boldsymbol{x}}/_{\sqrt{d}} + \kappa_*^{(1)} Z)$ (with Z a random Gaussian variable), thus leading to the covariance (235). This so-called one-dimensional Central Limit Theorem (1dCLT) (Goldt et al., 2021a; Hu et al., 2020; Montanari et al., 2022b) holds under some strict assumptions on the weight matrix W_1, that are satisfied in particular for random matrices with independent entries.

In the Bayesian setting one needs to integrate over the posterior distribution of the matrix W_1, learned from the data. For Conjecture 8.2.1 to be valid, the conditions of the 1dCLT must be satisfied w.h.p over the learned matrices, which is by no means a trivial requirement. This is where the properties of Bayes-optimal inference come in handy: indeed, a classic property of Bayesian learning (often called the Nishimori property (Nishimori, 2001; Iba, 1998; Zdeborová et al., 2015)) is that the statistics of weights drawn from the Bayes posterior is exactly the same as the one of the target network weights. This is a direct consequence of the Bayes formula (see e.g. section 1.2.3. in (Zdeborová et al., 2015)). As a consequence, the learned matrices are following Gaussian statistics as well (given this is the statistics of the target ones by definition), and thus respect the conditions of the 1dCLT.

When considering different replicas ($a \neq b$), the Nishimori conditions ensure that one of the two replicas can be taken, without loss of generality, to be the target weight $W_1^\star$. Since W_1 is learnt and therefore generically correlated with the target weights $W_1^\star$, the assessment of the covariance Ω_1^{ab} is a challenging task. However, the results of (Aubin et al., 2018a) suggest that W_1 is asymptotically uncorrelated with $W_1^\star$ for sample complexities $\alpha \lesssim k_1$. Since we consider here $\alpha = \mathcal{O}(1) \ll k_1$, this motivates the following conjecture:

Conjecture 8.2.2. *(Non-specialization) for $L = 1$, in the asymptotic limit $n, d, k_1 \to \infty$ with fixed $\mathcal{O}(1)$ ratio $\gamma_1 = {}^{k_1}/_{d}$ and $\alpha = {}^{n}/_{d}$, let W_1 be sampled from the Bayes posterior (234). Then with high probability W_1 has vanishing overlap with $W_1^\star$, i.e.*

$$\frac{1}{d} \max_{1 \leq i,j \leq k_1} \left(W_1^\star \Sigma W_1^\top \right)_{i,j} = \mathcal{O}\left({}^{1}/\sqrt{d} \right). \tag{236}$$

8.2.2 implies that the second term in the right-hand side of (235) is only present for $a = b$.

8.2.2 DEEP (BAYESIAN) GAUSSIAN EQUIVALENCE PROPERTY

We next discuss how these results generalize to deep networks ($L \geq 2$). While a sizeable body of work has been devoted to the distribution induced by the random weights for fixed inputs (Lee et al., 2018c; G. Matthews et al., 2018; Hanin et al., 2022; Hanin, 2022; Yaida, 2019), little is known, in the deep case, for the distribution induced by the input distribution, for *fixed* weights. While there is no proof of the equivalence of the 1dCLT of (Goldt et al., 2021a; Hu et al., 2022c) for $L \geq 2$, numerical evidence of the following conjecture can be found in (Cui et al., 2023a):

Conjecture 8.2.3. *The output $\hat{y}(\boldsymbol{x})$ of a deep random network, conditional on its Gaussian weights $\{W_\ell\}_{\ell=1}^{L}, \mathbf{a}$, in the extensive-width limit $d \to \infty$ and $\forall \ell, k_\ell \to \infty$ with fixed ratios $\gamma_\ell = {}^{k_\ell}/_{d}$, is asymptotically Gaussian with respect to $\boldsymbol{x}$.*

Conjecture 8.2.3 thus extends the first part of 8.2.1 to the deep setting. This intuitively follows from the fact that higher order cumulants of the post-activations at intermediary layers are asymptotically suppressed (as shown in (Fischer et al., 2022)) and thus approximately Gaussian – allowing one to iterate 8.2.1. A closed-form expression for the variance of $\hat{y}(\boldsymbol{x})$, which like the shallow case 8.2.1 is amenable to being interpreted in terms of an equivalent noisy network, can be reached using linearization arguments like in Chapter 7. We defer the discussion of the latter to the subsection 8.2.3. Finally, the Nishimori property again ensures that conjecture 8.2.3 transfers to weights sampled from the Bayes posterior (234). Defining the following recursion on $\{r_\ell\}_{\ell=1}^L$, $\{\kappa_1^{(\ell)}\}_{\ell=1}^L$ and $\{\kappa_*^{(\ell)}\}_{\ell=1}^L$:

$$
\begin{aligned}
r_{\ell+1} &= \Delta_{\ell+1}\mathbb{E}_{z\sim\mathcal{N}(0,r_\ell)}\left[\sigma_\ell(z)^2\right],\\
\kappa_1^{(\ell)} &= \frac{1}{r_\ell}\mathbb{E}_{z\sim\mathcal{N}(0,r_\ell)}\left[z\sigma_\ell(z)\right],\\
\kappa_*^{(\ell)} &= \sqrt{\mathbb{E}_{z\sim\mathcal{N}(0,r_\ell)}\left[\sigma_\ell(z)^2\right] - r_\ell\left(\kappa_1^{(\ell)}\right)^2},
\end{aligned} \tag{237}
$$

with $r_1 \equiv \Delta_1 {}^{\mathrm{Tr}(\Sigma)}/{}_d$, the deep version of (8.2.1) and 8.2.2 reads:

Conjecture 8.2.4. ***(Deep Bayes GEP)*** *in the extensive width asymptotic limit $d\to\infty$ and $\forall\ell, k_\ell\to\infty$ with fixed ratios $\gamma_\ell = {}^{k_\ell}/{}_d$, let $\hat{y}^1(\boldsymbol{x}),...,\hat{y}^s(\boldsymbol{x})$ be any finite number of replicas independently drawn from the Bayes posterior eq. (234). Then $\hat{y}^1(\boldsymbol{x}),...,\hat{y}^s(\boldsymbol{x})$ are jointly Gaussian with correlation $\mathbb{E}_{\boldsymbol{x}}\hat{y}^a(\boldsymbol{x})\hat{y}^b(\boldsymbol{x}) = {}^{\mathbf{a}^{a\top}\Omega_L^{ab}\mathbf{a}^b}/{}_{k_1}$, where the population covariance $\Omega_L^{ab} \equiv \mathbb{E}_{\boldsymbol{x}}(\varphi_L^a\circ...\varphi_1^a(\boldsymbol{x}))(\varphi_L^b\circ...\varphi_1^b(\boldsymbol{x}))^\top$ is given by*

$$
\Omega_\ell^{ab} = \left(\kappa_1^{(\ell)}\right)^2\frac{W_\ell^a\Omega_{\ell-1}^{ab}W_\ell^{b\top}}{k_{\ell-1}} + \delta_{ab}\left(\kappa_*^{(\ell)}\right)^2\mathbb{I}_{k_\ell}. \tag{238}
$$

Finally, defining $\Omega_\ell^{a\star} \equiv \mathbb{E}_{\boldsymbol{x}}(\varphi_\ell^a\circ...\varphi_1^a(\boldsymbol{x}))(\varphi_\ell^\star\circ...\varphi_1^\star(\boldsymbol{x}))^\top$ for any a, there is no specialization, i.e. with high probability

$$
\frac{1}{d}\max_{1\le i,j\le k_\ell}\left(W_\ell^a\Omega_{\ell-1}^{a\star}(W_\ell^\star)^\top\right)_{i,j} = \mathcal{O}\left({}^1/\sqrt{d}\right). \tag{239}
$$

In (238), $\Omega_0^{ab} \equiv \Sigma$. We precise that 8.2.4 holds for any sequence of activations $\{\sigma_\ell\}_{\ell=1}^L$ satisfying $\forall\ell$, $\mathbb{E}_{z\sim\mathcal{N}(0,r_\ell)}\left[\sigma_\ell(z)\right] = 0$. This is in particular always true for odd activations. We adopt in this work the latter (stronger) assumption for the sake of definiteness and clarity. An important note is that the population covariance between post-activations at *any* two layers $1 \le \ell, \ell' \le L$ (not just $\ell = \ell' = L$) can be generically computed. Because the post-activations result from the propagation of the Gaussian variable $\boldsymbol{x}$ through several non-linear layers, this computation is non-trivial for $L \ge 2$ and of independent interest.

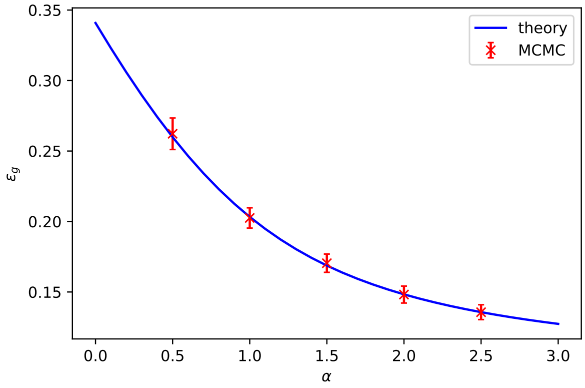


Figure 24: (solid lines) Theoretical predition for the Bayes MSE (240), for a one-hidden layer rectangular neural network ($\gamma_1 = 1$) with shifted ReLU activation $\sigma_1(\cdot) = (\cdot)_+ - 1/\sqrt{2\pi}$. (red crosses) Monte Carlo simulations using the Gibbs sampling algorithm of (Piccioli et al., 2023). The MSE was estimated over the last 15000 iterations of the algorithm, after 15000 initial thermalization steps. Error bars represent a single standard deviation over $N = 20$ instances of the target network. The simulations were performed in dimension $d = 1000$.

8.2.3 BAYES-OPTIMAL ERRORS

Conjectures 8.2.1 and 8.2.4 allow to characterize the Bayes error in terms of the sole second-order statistic $q = \mathbb{E}_{\mathscr{D},W_1^\star,\mathbf{a}_\star}\langle \mathbb{E}_{\boldsymbol{x}}\hat{y}^a(\boldsymbol{x})\hat{y}^b(\boldsymbol{x})\rangle_{\mathbb{P}}$. q is known as the *self overlap* in the statistical physics literature. The replica computation then proceeds in a rather standard way, provided one employs the so-called *RS ansatz*, which is always correct in Bayes-optimal settings (see e.g. (Zdeborová et al., 2015)). One finally reaches the following characterizations :

For regression, the Bayes-optimal MSE reads

$$\varepsilon_{g,\mathrm{reg}}^{\mathrm{BO}} = \prod_{\ell=1}^{L}\left(\kappa_1^{(\ell)}\right)^2\left(\Delta_a\left(\int z\mathrm{d}\mu(z)\right)\prod_{\ell=1}^{L}\Delta_\ell - q\right) + \varepsilon_r \tag{240}$$

with the self-overlap q satisfying the equation

$$q = \frac{1}{2}\int \frac{\alpha \prod\limits_{\ell=1}^{L}\left(\kappa_1^{(\ell)}\right)^2 z^2\Delta_a^2 \prod\limits_{\ell=1}^{L}\Delta_\ell^2}{\varepsilon_{g,\mathrm{reg}}^{\mathrm{BO}} + \alpha\prod\limits_{\ell=1}^{L}\left(\kappa_1^{(\ell)}\right)^2 z\Delta_a \prod\limits_{\ell=1}^{L}\Delta_\ell}\mathrm{d}\mu(z). \tag{241}$$

We have denoted the residual error

$$\varepsilon_r \equiv \sum_{\ell_0=1}^{L-1}\left(\kappa_*^{(\ell_0)}\right)^2\Delta_a\prod_{\ell=\ell_0+1}^{L}\left(\kappa_1^{(\ell)}\right)^2\Delta_\ell + \left(\kappa_*^{(L)}\right)^2\Delta_a + \Delta. \tag{242}$$

For classification, the Bayes-optimal error reads

$$\varepsilon_{g,\text{class}}^{\text{BO}} = \frac{1}{\pi}\arccos\left[\frac{\sqrt{\prod_{\ell=1}^{L}\left(\kappa_1^{(\ell)}\right)^2 q}}{\sqrt{\Delta_a \int z \mathrm{d}\mu(z)\prod_{\ell=1}^{L}\left(\kappa_1^{(\ell)}\right)^2 \Delta_\ell + \varepsilon_r}}\right], \tag{243}$$

where the self-overlap q satisfies the system of equations

$$\begin{cases} q = \int \frac{\hat{q}\Delta_a^2 \prod_{\ell=1}^{L}\Delta_\ell^2 z^2}{\hat{q} z \Delta_a \prod_{\ell=1}^{L}\Delta_\ell + 1}\mathrm{d}\mu(z) \\ \hat{q} = \frac{2\alpha\prod_{\ell=1}^{L}\left(\kappa_1^{(\ell)}\right)^2}{\Delta_a\int z\mathrm{d}\mu(z)\prod_{\ell=1}^{L}\left(\kappa_1^{(\ell)}\right)^2\Delta_\ell + \varepsilon_r - \prod_{\ell=1}^{L}\left(\kappa_1^{(\ell)}\right)^2 q} \\ \int \frac{d\xi}{(2\pi)^{\frac{3}{2}}}\frac{2e^{-\frac{1}{2}\frac{\Delta_a\int z\mathrm{d}\mu(z)\prod_{\ell=1}^{L}\left(\kappa_1^{(\ell)}\right)^2\Delta_\ell+\varepsilon_r+\prod_{\ell=1}^{L}\left(\kappa_1^{(\ell)}\right)^2 q}{\Delta_a\int z\mathrm{d}\mu(z)\prod_{\ell=1}^{L}\left(\kappa_1^{(\ell)}\right)^2\Delta_\ell+\varepsilon_r-\prod_{\ell=1}^{L}\left(\kappa_1^{(\ell)}\right)^2 q}\xi^2}}{1-\mathrm{erf}\left(\frac{\prod_{\ell=1}^{L}\kappa_1^{(\ell)}\sqrt{q}\xi}{\sqrt{2\left(\Delta_a\int z\mathrm{d}\mu(z)\prod_{\ell=1}^{L}\left(\kappa_1^{(\ell)}\right)^2\Delta_\ell+\varepsilon_r-\prod_{\ell=1}^{L}\left(\kappa_1^{(\ell)}\right)^2 q\right)}}\right)} \end{cases}.$$

As previously discussed, numerically sampling the Bayes posterior (234) is a hard task. However, for regression in the shallow $L=1$ case and a shifted ReLU activation $\sigma_1(\cdot) = (\cdot)_+ - 1/\sqrt{2\pi}$ (with the shift ensuring the condition $\mathbb{E}_{z\sim\mathcal{N}(0,r_1)}[\sigma_1(z)] = 0$ discussed below Conjecture 8.2.4 is satisfied), the recent work of (Piccioli et al., 2023) provides an efficient implementation of a Gibbs sampler. We use this algorithm to run simulations for this particular target network, which can be observed in Fig. 24 to agree well with the theoretical prediction (240).

Equivalent shallow network Remarkably, the Bayes errors (240) and (243) are equal to the Bayes errors of a simple single-layer target function with random weights

$$y^{\text{eq}}(\boldsymbol{x}) = f_\star\left(\frac{\sqrt{\rho}\boldsymbol{\theta}^\top\boldsymbol{x}}{\sqrt{d}} + \sqrt{\varepsilon_r}\xi\right), \tag{244}$$

where $\xi \sim \mathcal{N}(0,1)$ and $\boldsymbol{\theta}$ is a Gaussian weight vector with independent Gaussian entries of unit variance. We have defined the effective signal strength

$$\rho \equiv \Delta_a \prod_{\ell=1}^{L}\left(\kappa_1^{(\ell)}\right)^2\Delta_\ell.$$

To gain intuition on this equivalence, observe that the deep Bayes GEP 8.2.4 applied to a single replica implies that the deep non-linear network (229) is characterized by the same second-order activations statistics as a network with noisy *linear* layers

$$\varphi_\ell^{\text{eq.}}(\boldsymbol{x}) = \kappa_1^{(\ell)} \frac{1}{\sqrt{k_{\ell-1}}} W_\ell^\star \cdot \boldsymbol{x} + \kappa_*^{(\ell)} \mathcal{N}(0, \mathbb{I}_{k_\ell}). \tag{245}$$

In turn, this deep noisy network reduces equivalently to the shallow network (244). Interestingly, note that while the multilayer target (229) is deterministic for a given instance of the weights, the equivalent target (244) displays a stochastic output noise ξ. This noise subsumes the effect of the higher order terms introduced by the non-linearities, which are not learnt in the proportional regime (Mei et al., 2022b).

Fig. 25 shows the Bayes MSE, eq. (240), for networks with tanh activation, with $L = 1, 2$ hidden layers. This is contrasted to the MSE achieved by an expressive ANN with twice the target width, optimized end-to-end with full batch gradient descent. Fig. 27 presents the same experiment in the classification setting. As expected, even this expressive learning algorithm cannot achieve a lower error than the information-theoretic lower bounds (240), (243).

8.3 ERM WITH LINEAR METHODS

Eqs. (240) and (243) provide the information-theoretic minimal error for deep extensive-width targets (229). However, (Barbier et al., 2019a; Aubin et al., 2018a) evidenced that the Bayes error is not always attainable, in practice, by known polynomial time algorithms. In this section, we investigate the performance of some standard ERM methods. We provide a tight asymptotic characterization of the test error of each algorithm and show that for regression the Bayes error is, in fact, also achievable algorithmically. We address in succession: ridge regression, RF regression, kernel regression, logistic regression and ridge classification.

We give, for each of the considered ERM algorithms, a sharp asymptotic characterization of the associated test error. The fact that the deep non-linear target (229) shares the same Bayes error as the equivalent shallow model (244) suggests that the test error of ERM methods should also be identical. Applying Theorem 1 of Chapter 8 on the equivalent shallow target (244) thus leads to the formulas provided here. This heuristic line of reasoning is put on a firm basis in some settings in the concomitant work of (Schröder et al., 2023b), where the formulas characterizing the performance of the considered ERM methods are derived.

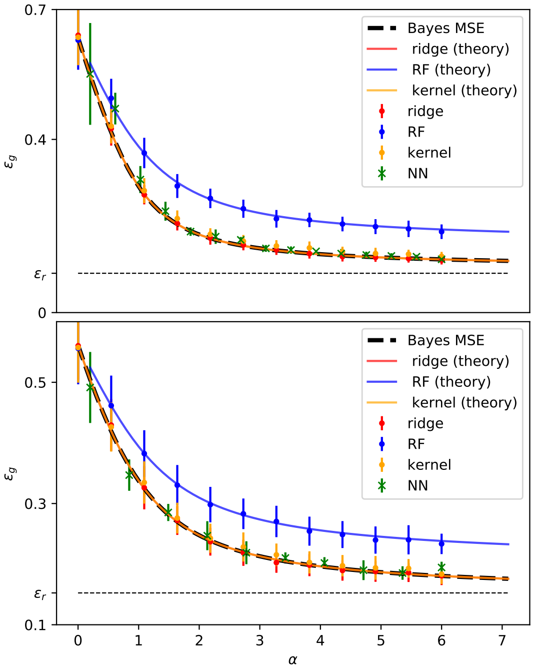


Figure 25: Targets (229) with $L = 1$ (top) and $L = 2$ (bottom) hidden layers , with $\sigma_{1,2}(x) = \tanh(2x)$ activation, widths $k_{1,2} = 700$, and $\Delta_a = \Delta_{1,2} = 1$, $\Delta = 0$, in dimension $d = 500$. The Bayes-optimal MSE (240) (dashed black) is contrasted to the replica predictions and simulations for optimally regularized ridge regression (247) (red), optimally regularized random features (251) with $\sigma(x) = \tanh(2x)$ non-linearity (blue), and optimally regularized kernel regression (252) with $\sigma(x) = \tanh(2x)$ non-linearity (orange). Green dots represent simulations for a one (top) and two (bottom) hidden layers neural network of width 1500, optimized with full-batch GD, learning rate $\eta = 8.10^{-3}$ and weight decay $\lambda = 0.1$. Dashed grey lines represent the residual error ε_r (242). Error bars represent one standard deviation over 30 trials.

8.3.1 RIDGE REGRESSION

We first consider ridge regression, corresponding to the minimization of the risk

$$\mathscr{R}(w) = \sum_{\mu=1}^{n} \left(y^\mu - \frac{w \cdot \boldsymbol{x}^\mu}{\sqrt{d}} \right)^2 + \frac{\lambda}{2} ||w||^2, \tag{246}$$

with a ℓ_2 regularization term of strength λ. The associated test error can be computed by combining the deep GEP (237) and the theorem of Chapter 3, as

$$\varepsilon_g = \rho \int z \mathrm{d}\mu(z) + q - 2 \prod_{\ell=1}^{L} \kappa_1^{(\ell)} m + \varepsilon_r, \tag{247}$$

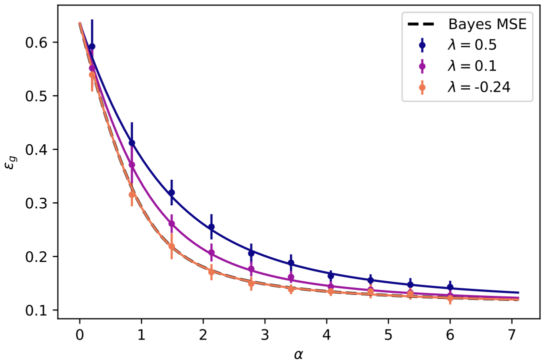


Figure 26: Arccosine kernel regression on a two-layers target of width $k_1 = 700$, $\Delta_a = \Delta_1 = 1$, $\Delta = 0$, with $\sigma_1(x) = \tanh(2x)$ activation, in dimension $d = 500$. Different curves correspond to different regularizations $\lambda = 0.5, 0.1$ and the optimal $\lambda^\star \approx -0.24$ (253). Solid lines correspond to the replica predictions (252). The learning curve for the optimal *negative* regularizer $\lambda^\star$ (red) superimposes with the Bayes-optimal MSE (240). Error bars represent one standard deviation over 30 trials.

where m, q, V are the solutions of the system of equations

$$\begin{cases} \hat{V} = \frac{\alpha}{1+V} \\ \hat{q} = \alpha \frac{\epsilon_g}{(1+V)^2} \\ \hat{m} = \frac{\prod_{\ell=1}^{L} \kappa_1^{(\ell)} \alpha}{1+V} \end{cases} \qquad \begin{cases} V = \int \frac{z}{\lambda + \hat{V} z} \mathrm{d}\mu(z) \\ q = \int \frac{\Delta_a \prod_{\ell=1}^{L} \Delta_\ell \hat{m}^2 z^3 + \hat{q} z^2}{(\lambda + \hat{V} z)^2} \mathrm{d}\mu(z) \\ m = \Delta_a \prod_{\ell=1}^{L} \Delta_\ell \hat{m} \int \frac{z^2}{\lambda + \hat{V} z} \mathrm{d}\mu(z) \end{cases} \tag{248}$$

Ridge regression is Bayes-optimal The optimal regularization $\lambda^\star$ leading to minimal test error can be found to admit the compact expression

$$\lambda^\star = \frac{\epsilon_r}{\rho}. \tag{249}$$

The expression (249) mirrors the result of (Sahraee-Ardakan et al., 2022) for GP targets. Eq. (249) intuitively corresponds to requiring that the regularization λ used in the ERM (246) should be equal to the true noise-to-signal ratio of the equivalent target (244).

For $\lambda = \lambda^\star$, the equation (248) reduces to (240), implying that optimally regularized ridge regression achieves the Bayes-optimal MSE. The solution of (247), (248), (249), with the corresponding numerical simulations, is plotted in Fig. 25, and can indeed be seen to exactly fall on the Bayes-optimal baseline (240). This implies that in the proportional high-dimensional limit $n \sim d$, no algorithm can learn more accurately the non-linear function (229) than optimally regularized linear regression. This echoes the claims of (Sahraee-Ardakan et al., 2022) when the target is a GP.

8.3.2 RANDOM FEATURES

RF learning (Rahimi et al., 2007a) was initially introduced as a means to speed up kernel methods. They correspond to the ERM

$$\mathscr{R}(w) = \sum_{\mu=1}^{n} \left(y^{\mu} - \frac{1}{\sqrt{k}} w \cdot \sigma\left(\frac{F \cdot \boldsymbol{x}^{\mu}}{\sqrt{d}}\right) \right)^2 + \frac{\lambda}{2} ||w||^2, \tag{250}$$

where σ is a nonlinearity with associated GEP coefficients (237) κ_1, κ_*, and $F \in \mathbb{R}^{k \times d}$ is the random feature matrix, assumed in the following to possess i.i.d Gaussian entries. RF learning corresponds formally to ridge regression in a $k-$dimensional space, often taken larger than the $d-$ dimensional input space to allow for overparametrization. Again, we consider the proportional regime $n, d, k \to \infty$ and introduce the width/dimension ratio $\gamma \equiv k/d$. In the following, for simplicity, we restrict ourselves to the case of isotropic covariates $\Sigma = \mathbb{I}_d$. Sharp asymptotics for the test error of such models have been provided in (Mei et al., 2019c; Gerace et al., 2020a) for single-layer targets. We tie those results in with (237) in the case of the deep random target (229). The asymptotic test error is given again by (247), where q, m satisfy

$$\begin{cases} \hat{V} = \frac{\frac{\alpha}{\gamma}}{1+V} \\ \hat{q} = \frac{\alpha}{\gamma} \frac{\epsilon_g}{(1+V)^2} \\ \hat{m} = \sqrt{\Delta_a \prod\limits_{\ell=1}^{L} \Delta_\ell} \sqrt{\gamma} \frac{\prod\limits_{\ell=1}^{L} \kappa_1^{(\ell)} \frac{\alpha}{\gamma}}{1+V} \end{cases} \tag{251}$$

$$\begin{cases} V = \frac{1}{\hat{V}} - \frac{\lambda}{\hat{V}^2 \kappa_1^2} g\left(-\frac{\lambda + \hat{V}\kappa_*^2}{\hat{V}\kappa_1^2}\right) \\ q = \frac{\hat{m}^2 + \hat{q}}{\hat{V}^2} - \frac{1}{\kappa_1^2 \hat{V}^2} \left(\frac{2\lambda(\hat{m}^2 + \hat{q})}{\hat{V}} + \hat{m}^2 \kappa_*^2\right) g\left(-\frac{\lambda + \hat{V}\kappa_*^2}{\hat{V}\kappa_1^2}\right) \\ \qquad + \frac{\lambda}{\kappa_1^4 \hat{V}^3} \left(\frac{\lambda(\hat{m}^2 + \hat{q})}{\hat{V}} + \hat{m}^2 \kappa_*^2\right) g'\left(-\frac{\lambda + \hat{V}\kappa_*^2}{\hat{V}\kappa_1^2}\right) \\ m = \sqrt{\gamma} \frac{\hat{m}}{\hat{V}} \left[1 - \frac{1}{\kappa_1^2}\left(\frac{\lambda}{\hat{V}} + \kappa_*^2\right) g\left(-\frac{\lambda + \hat{V}\kappa_*^2}{\hat{V}\kappa_1^2}\right)\right]. \end{cases}$$

$g(z)$ is the Stieljes transform of the Marchenko-Pastur distribution with aspect ratio γ. The solution of (251) is plotted in Fig. 25, with the regularization λ being numerically optimized over, and is observed to yield higher MSE than the optimal (240). Intuitively, this is because from (244) the target function (229) effectively acts as a linear function on the original space, while the RF transformation $\sigma(F \cdot)$ (250) introduces a mismatch between the original basis where the target acts and the features basis in which the linear regression readout is carried out. This mismatch can be shown to be benign only in the infinite overparametrization (infinite width) $\gamma \to \infty$ limit, has discussed in the next subsection.

8.3.3 KERNELS

In the $\gamma \to \infty$ limit, RF learning (250) becomes equivalent to kernel regression (Neal, 1996b; Lee et al., 2018c). The saddle-point equations (251) characterizing the generalization error (247) then reduce to

$$\begin{cases} \hat{V} = \frac{\alpha}{1+V} \\ \hat{q} = \alpha \frac{\varepsilon_g}{(1+V)^2} \\ \hat{m} = \alpha \frac{\prod_{\ell=1}^{L} \kappa_1^{(\ell)}}{1+V} \end{cases} \qquad \begin{cases} V & = \frac{\kappa_*^2}{\lambda} + \frac{\kappa_1^2}{\lambda+\hat{V}\kappa_1^2} \\ q & = \frac{\Delta_a \prod_{\ell=1}^{L} \Delta_\ell \hat{m}^2 \kappa_1^4 + \hat{q}\kappa_1^4}{(\lambda+\hat{V}\kappa_1^2)^2} \\ m & = \Delta_a \prod_{\ell=1}^{L^\star} \Delta_\ell \hat{m} \frac{\kappa_1^2}{\lambda+\hat{V}\kappa_1^2} \end{cases} . \tag{252}$$

Kernel regression is Bayes-optimal The regularization λ minimizing the test error (247) for kernel regression (252) can be shown as in the ridge case to admit the simple expression

$$\lambda^\star = \kappa_1^2 \left(\frac{\varepsilon_r}{\rho} - \frac{\kappa_*^2}{\kappa_1^2} \right). \tag{253}$$

Again, the expression (253) also admits a very intuitive interpretation. An informal takeaway from (Mei et al., 2022b; Hu et al., 2022c) is indeed that in the linear $n \sim d$ regime, kernel regression can be seen as effectively implementing ridge regression with an implicit ℓ_2 regularization equal to $\lambda + \kappa_*^2/\kappa_1^2$. Requiring this implicit regularization to match the true target noise-to-signal ratio (244) naturally leads to (253). The test error (247) evaluated at the solution of (252), at the optimal regularization (253), is plotted in Fig. 25, and can be seen to exactly superimpose with the Bayes-optimal baseline (240).

When the activation σ of the kernel is very non-linear (as quantified by the ratio κ_*^2/κ_1^2) and implements too strong an implicit regularization (Wu et al., 2020b; Hastie et al., 2022) compared to the actual target noise, the optimal regularization (253) can be *negative*. This negative ridge phenomenon can for example be observed when learning a target 2−layer network with tanh activation with a kernel using the more non-linear sign activation, see Fig. 26. Note that even in such cases where the optimal explicit regularization λ is negative, the risk remains convex due to the implicit ℓ_2 regularization.

8.3.4 LOGISTIC AND RIDGE CLASSIFICATION

This last subsection addresses ERM in the classification setting. We consider two standard classification methods, namely logistic regression and ridge classification. They correspond to ERM on the risk

$$\mathscr{R}(w) = \sum_{\mu=1}^{n} \ell\left(y^\mu, \frac{w \cdot \boldsymbol{x}^\mu}{\sqrt{d}} \right) + \frac{\lambda}{2} ||w||^2 \tag{254}$$

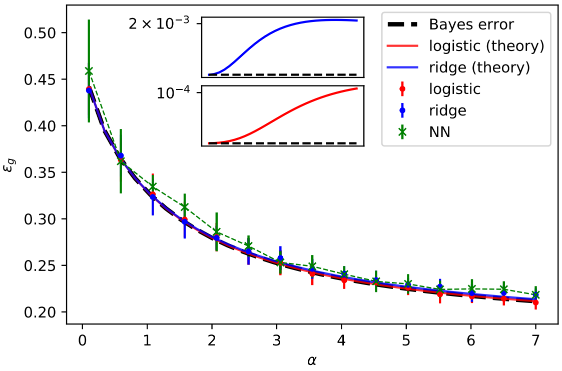


Figure 27: Learning curves for classification, with a three-layers target (229) with $\tanh(2\cdot)$ activation and width $k_{1,2} = 700$, $\Delta_a = \Delta_{1,2} = 1$, $\Delta = 0$, in dimension $d = 500$. The black dashed line represents the Bayes-optimal error (243). The theoretical learning curves for optimally regularized logistic regression (ridge classification) are shown in red (blue), alongside the corresponding numerical simulations. Green dots show the test error of a three layers fully connected network trained end-to-end with full-batch Adam, learning rate 0.003 and weight decay 0.01, after 2000 epochs. Error bars represent one standard deviation over 30 trials. (inset) Zoom in on the theoretical learning curves for logistic regression (red) and ridge classification (blue), with the Bayes-optimal baseline (243) subtracted. Logistic regression and ridge classification can be seen to be very close, but not equal, to the Bayes-optimal baseline (up to 10^{-4} and 10^{-3} respectively).

with $\ell(y,z) = \ln(1+e^{-yz})$ and $\ell(y,z) = 1/2(y-z)^2$ respectively. For simplicity, we assume the noiseless $\Delta = 0$ setting. An asymptotic expression for the test error of logistic regression can again be reached. These theoretical predictions are plotted in Fig. 27, and contrasted to the Bayes-optimal baseline (243). At odds with the regression case, the learning curves of logistic regression and ridge classification do not exactly superimpose with the Bayes-optimal baseline but lie extremely close. Fig. 27 shows that for a $\sigma(x) = \tanh(2x)$ activation the difference is of order of 10^{-4} (for logistic regression) and 10^{-3} (for ridge classification). Such a gap has also been observed by (Aubin et al., 2020a) for targets without hidden layer.

8.4 BEYOND THE PROPORTIONAL REGIME

Section 8.3 establishes that ridge regression and kernel regression are Bayes-optimal in the linear $n \sim d$ regime for the target (229). These conclusions and the ones of (Sahraee-Ardakan et al., 2022) in the context of GP targets, and are reminiscent on a high level of the empirical observations of (Arora et al., 2020) that DNN can only marginally outperform kernel methods for small train sets on several benchmark real datasets. Note that the study (Lee et al., 2020) also employs networks with comparable width to the dataset size and

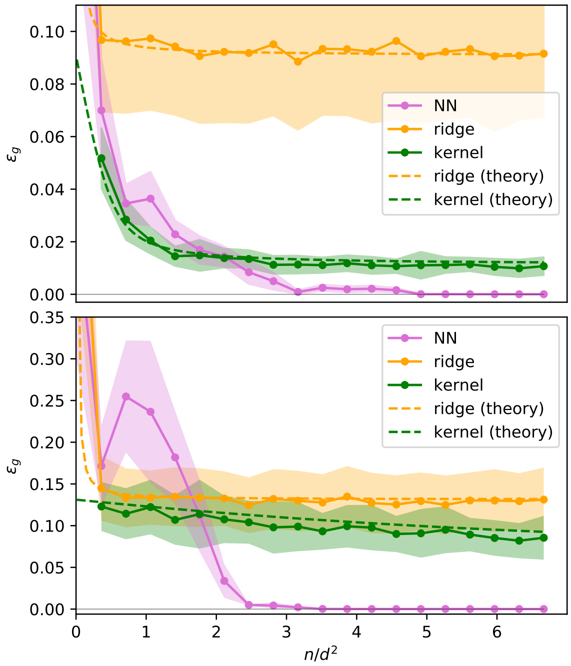


Figure 28: MSE of regression on a ReLU (top) $\text{erf}(2x)$ (bottom) 2−layers target of width $k_1 = 20$, $\Delta_a = \Delta_1 = 1$, $\Delta = 0$, in dimension $d = 30$. Dots represent simulations for optimally regularized ridge regression (orange), optimally regularized arccosine (top), and arcsine (bottom) kernel (green). Dashed lines represent the theoretical predictions for the MSE of the kernel in polynomial regimes of (Hu et al., 2020). Purple dots indicate the MSE of a 2−layers fully connected neural network of width $k = 30$ trained end-to-end using Adam (purple), batch size $n/3$ and learning rate $\eta = 3.10^{-3}$, over 2000 epochs. Error bars represent one standard deviation over 10 trials.For a quadratic number of samples $n \sim d^2$ the network learns the target perfectly (up to errors of order 10^{-5}) while kernel regression only learns a quadratic approximation of the target and yields MSEs larger by an order of 10^3. Ridge regression can only learn the best linear approximation and leads to even higher MSE.

dimension, and reaches similar conclusions.

In fact, the information-theoretic optimality of linear/kernel ERM methods is essentially due to the fact that a proportional number of samples $n \sim d$ is not enough to learn features beyond linear approximations. The conclusions drawn in other scaling regimes are anticipated to be very different. In particular, beyond the linear $n \sim d$ regime, the test error should pick up, and depend on, non-Gaussian asymptotic corrections to the GEP (Goldt et al., 2021a), Bayes GEP 8.2.1, and deep Bayes GEP 8.2.4. Besides, information-theoretic intuition suggests that since the target (229) is parametrized by $\mathcal{O}(d^2)$ real numbers, a quadratic number of samples $n \sim d^2$ is needed, and sufficient, to learn it perfectly. In other words, one expects the Bayes-optimal error to vanish in the quadratic regime. This means in particular that kernel methods (which are known to learn at best a quadratic approximation of the target in this scaling regime (Misiakiewicz, 2022; Xiao et al., 2022; Hu et al., 2022c)), and ridge regression (which learns a linear approximation) will cease to be optimal. In this section, this intuition is further explored numerically.

Fig. 28 contrasts the MSE of an Adam-optimized neural network, optimally regularized ridge regression, and optimally regularized arcosine kernel regression, in the $n \sim d^2$ regime, for a ReLU network target with one hidden layer. For completeness, we also plot the theoretical predictions for the test error for kernel regression derived in (Hu et al., 2022c). While ridge (kernel) regression can only learn the best linear (quadratic) approximation of the non-linear target, the neural network manages to learn the target almost perfectly (up to an MSE of $\mathcal{O}(10^{-5})$). These results show the superiority of neural networks over kernel methods in the quadratic regime for the multi-layer target (229), and complement similar superiority results (see e.g. (Ghorbani et al., 2019a; Ghorbani et al., 2021)) for single-layer targets.

CONCLUSION

We investigate the problem of learning a deep, non-linear, extensive-width random neural network. We propose asymptotic expressions for the Bayes-optimal error and the test error of linear / kernel ERM methods, for classification and regression. The technical backbone of the derivations is the deep Bayes GEP conjecture 8.2.4, and novel closed-form formulae for second-order population statistics of network post-activations. The conclusion is that kernel methods are optimal in this regime. We showed, however, that the situation is drastically different in the quadratic sample regime, and evidence the onset of feature learning leading to a vanishing test error for neural nets, while kernel methods can learn only quadratic approximation and thus become suboptimal. This marks a clear separation in the power of neural nets with respect to kernels as soon as $n \sim d^2$.

From a theoretical standpoint, a quantitative analysis of the learning curve for the quadratic regime is challenging. In particular, Gaussian universality results such as (Goldt et al., 2021a; Hu et al., 2020; Mei et al., 2019c), cease to hold outside of the proportional regime and would need to be extended. Building a theory for extensive-width networks in these superlinear sample regimes could unveil rich behavior and properties, and constitutes in the authors's point of view a crucial challenge in machine learning theory.

9 FEATURE LEARNING AFTER ONE GRADIENT-STEP

A common deep learning intuition behind the unreasonable effectiveness of neural networks is their capacity to effectively adapt to the training data which makes them superior to kernel methods. While kernel methods and their finite width approximations are known to be data hungry (see e.g. (Adlam et al., 2023)), neural networks have proven themselves to be flexible and efficient in practice. Their adaptivity and capacity to learn features from data is behind the success in efficiently solving problems from image classification to text generation. A large part of our current theoretical understanding of neural networks stems from the investigation of their lazy regime where features are *n*ot learned during training. This includes a set of works that falls under the umbrella of GP (Neal, 1996a; Lee et al., 2018d), the Neural Tangent Kernel (NTK) (Jacot et al., 2018a) and the Lazy regime (Chizat et al., 2019a). A crucial question in the theoretical machine learning community is thus to characterize the advantages of two-layer neural networks beyond these convex optimization approaches.

Different theoretical works have offered sharp separation results between kernel and feature learning regimes (see e.g. (Ghorbani et al., 2019c; Refinetti et al., 2021a; Damian et al., 2022; Abbe et al., 2023)). In particular, (Ba et al., 2022a; Damian et al., 2022), and later (Dandi et al., 2023a), discussed the advantage of neural networks when training with only *o*ne single step of large-batch gradient descent with a large learning rate. Specifically, (Ba et al., 2022a) highlighted that the weight matrix after the first training step can be decomposed in a bulk plus a rank-one spike, effectively mapping the learned features to a spiked Random Features (sRF) model, defined in eq. (260). This observation has fueled many further studies on the effect of spiked structure, see e.g. (Dandi et al., 2023a; Ba et al., 2023; Mousavi-Hosseini et al., 2023; Moniri et al., 2023).

In this paper, we follow this line of work, and provide an exact high-dimensional description of the test error achieved by a two-layer network for after a *s*ingle, large, gradient step. To our knowledge, our work provides the first sharp asymptotic treatment of a model where feature learning is accounted for in the high-dimensional, non-perturbative large learning rate regime $\eta = \Theta_d(d)$, quantitatively illustrating the benefits of feature learning over the lazy regime.

Our **main contributions** in this paper are the following:

- **Exact asymptotics for two-layer networks after one gradient step** – We provide a sharp asymptotic characterization of the test error, alongside a set of summary statistics, for two-layers neural networks with first layer weights trained with a large learning rate gradient step. The derivation leverages the replica method from statistical physics (Parisi, 1979a; Parisi, 1983b), and provides a set of scalar self-consistent equations for the generalization error.
- **Conditional Gaussian Equivalence** – Building upon (Dandi et al., 2023a), we show (and provide strong numerical evidence to support) that the learning properties of the sRF model are asymptotically equivalent to a simple *conditional* Gaussian model in the high-dimensional proportional regime. The conditional Gaussian distribution is characterized by the projections of the input data on the spike in the weight matrix. This mapping constitutes the extension of related theoretical results that unveiled a similar Gaussian equivalence property for the training and generalization error for non-spiked vanilla RFs (Goldt et al., 2022a; Goldt et al., 2020c; Gerace et al., 2020a; Hu et al., 2022b; Dhifallah et al., 2022; Mei et al., 2022c; Cui et al., 2023a; Schröder et al., 2023a; Bosch et al., 2023a; Dandi et al., 2023b).
- **Feature learning** – We provide an extensive discussion on how feature learning leads to a drastic improvement on the generalization performance over random features in a data limited regime, demonstrating a clear and quantitative separation with respect to kernel method and random feature models. In particular, we derive both upper and lower bounds on the generalization error and discuss under which conditions they are tight.

9.1 SETTING, MOTIVATION AND RELATED WORK

We study fully-connected two-layer networks

$$f_{W,\boldsymbol{a}}(\boldsymbol{x}) = \frac{1}{\sqrt{p}} \sum_{i=1}^{p} a_i \sigma(\boldsymbol{w}_i^\top \boldsymbol{x}), \tag{255}$$

and their capacity to learn a single-index target function of isotropic Gaussian covariates:

$$f_\star(x) = \sigma_\star(\boldsymbol{\theta}^\top \boldsymbol{x}/\sqrt{d}), \quad \boldsymbol{x} \sim \mathcal{N}(0, I_d) \tag{256}$$

from finite batch $\mathscr{D}=\{(\boldsymbol{x}^\mu,y^\mu)_{\mu=1}^n\}$ of n independently drawn training samples. We consider a layer-wise training procedure where the first layer weights $W\in\mathbb{R}^{p\times d}$ are trained for a single gradient step:

$$\begin{aligned}\boldsymbol{w}_i^{(1)} &= \boldsymbol{w}_i^{(0)}-\eta\boldsymbol{g}_i \\ \boldsymbol{g}_i &= \frac{1}{\sqrt{p}}\sum_{\mu=1}^{n_0}(y^\mu-f_{W^{(0)},\boldsymbol{a}^{(0)}}(\boldsymbol{x}^\mu))a_i^{(0)}\sigma'(\boldsymbol{w}_i^{(0)}\cdot\boldsymbol{x}^\mu)\boldsymbol{x}^\mu\end{aligned} \tag{257}$$

on a subset $\mathscr{D}_0\subset\mathscr{D}$ of size n_0, where $(W^{(0)},\boldsymbol{a}^{(0)})$ denote the initial weights and $\eta>0$ the learning rate. For simplicity, we assume $\boldsymbol{a}^{(0)}=1_p/\sqrt{p}$ (uniform initialization) and $\boldsymbol{w}_i^{(0)}$ with unit norm $\|\boldsymbol{w}_i^{(0)}\|=1$ and weak correlation $\boldsymbol{w}_i^{(0)}\cdot\boldsymbol{w}_j^{(0)}=O_d(\text{polylog}(d)/\sqrt{d})$ for $i\neq j$ (e.g. uniformly drawn from the unit sphere $\mathbb{S}^{d-1}$). Given the updated weights $W^{(1)}$, we train the read-out layer on the remaining data $\mathscr{D}_1=\mathscr{D}\setminus\mathscr{D}_0$:

$$\hat{\boldsymbol{a}}_\lambda=\underset{\boldsymbol{a}\in\mathbb{R}^p}{\operatorname{argmin}}\frac{1}{2}\sum_{\mu=1}^{n_1}\left(y^\mu-f_{W^{(1)},\boldsymbol{a}}(\boldsymbol{x}^\mu)\right)^2+\frac{\lambda}{2}\|\boldsymbol{a}\|^2. \tag{258}$$

with $\lambda\in\mathbb{R}_+$ being a regularization parameter. Note that the layer-wise training procedure considered here is commonly studied in the theoretical machine learning literature (Ba et al., 2022a; Damian et al., 2022; Abbe et al., 2023; Berthier et al., 2023; Dandi et al., 2023a; Moniri et al., 2023) due to its mathematical tractability.

Our main goal in the following is to provide a tight asymptotic characterization of the generalization error:

$$\varepsilon_g=\mathbb{E}_{\mathscr{D},\boldsymbol{x}}\left(f_\star(\boldsymbol{x})-f_{W^{(1)},\hat{\boldsymbol{a}}_\lambda}(\boldsymbol{x})\right)^2. \tag{259}$$

in the high-dimensional proportional limit where $n_0,n_1,d,p,\eta\to\infty$ at fixed ratios $\alpha_0={n_0}/{d}$, $\alpha={n_1}/{d},\beta={p}/{d},\tilde{\eta}={\eta}/{d}$.

MOTIVATION

Driven by the lazy-training regime of learning of large-width networks (Chizat et al., 2019a), a large body of literature has been dedicated to the particular case where the first layer weights are fixed at initialization $W^{(0)}$ ($\eta=0$), also known as the RF model (Rahimi et al., 2007b). In particular, (Ghorbani et al., 2019c; Ghorbani et al., 2020a; Mei et al., 2022a) have shown that with $n_1=\Theta_d(d)$ samples $f_{W^{(0)},\hat{\boldsymbol{a}}_\lambda}$ can only approximate, at best, a linear function of $\boldsymbol{x}^\mu$, with the non-linear part playing a role akin to additive label noise. This is a strong limitation: RF network requires $p$olynomials number of data and neurons to fit a simple polynomial (Mei et al., 2022a; Xiao et al., 2022). It is one of our motivation here to discuss how, with a single gradient steps, these limitations are lifted.

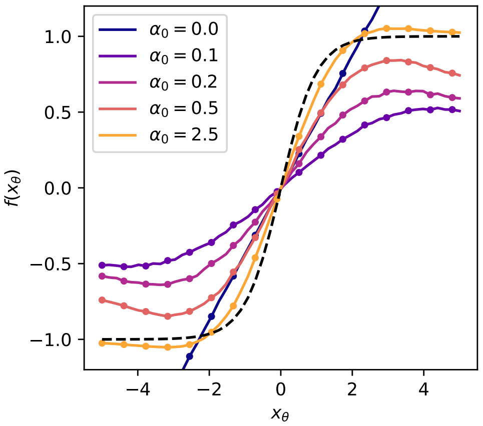


Figure 29: Numerical estimation of the function $f(x_\theta) = \mathbb{E}_x[f_{W^{(1)},\hat{\boldsymbol{a}}_\lambda} | \boldsymbol{\theta}^\top \boldsymbol{x}/\sqrt{d} = x_\theta]$ implemented by the trained network (255) in the direction spanned by the weights $\boldsymbol{\theta}$ of the target function. The activations are $\sigma = \sigma_\star = \tanh$, and simulations were run in dimensions $d = p = 2000$, for a learning rate $\eta = 2.5p$, and a readout regularization $\lambda = 0.01$. The readout was trained with $n_1 = 2d$ samples. Different colors corresponds to different sample complexities $\alpha_0 \equiv n_0/d$ used to implement the gradient step on the first layer weights, with $\alpha_0 = 0$ corresponding to not implementing the step.

Behind the scenes in this effective linearity of RF is a GEP (Goldt et al., 2022a; Mei et al., 2022c; Hu et al., 2022b; Montanari et al., 2022a; Dandi et al., 2023b), which states that in this regime the random feature map $\varphi = \sigma(W^{(0)}\boldsymbol{x})$ is statistically equivalent to a rescaled stochastic linear map $\varphi^g(\boldsymbol{x}) \asymp \mu_0 \mathbf{1} + \mu_1 W^{(0)}\boldsymbol{x} + \mu_\star \boldsymbol{z}$, with $\boldsymbol{z} \sim \mathcal{N}(0, I_p)$. Related linearizations are the object of further in-depth discussion in the previous Chapters of Part III. This surprising universality result allows to go beyond lower bounds for the generalization performance, making the problem amenable to a tight high-dimensional characterization of all relevant statistics in these models (Mei et al., 2022c; Gerace et al., 2020a; Dhifallah et al., 2022).

Figure 29 illustrates this fundamental limitation of RF models contrasted with the function $f(x_\theta) = \mathbb{E}_{\boldsymbol{x}}[f_{W^{(1)},\hat{a})\lambda}(\boldsymbol{x}) | \boldsymbol{\theta}^\top \boldsymbol{x}/\sqrt{d} = x_\theta]$, along the direction of the target $\boldsymbol{\theta}$, implemented by the network $f_{W^{(1)},\boldsymbol{a}}$ trained with a single gradient step (eq. (257)). Varying the amounts of data $n_0 = \alpha_0 d$ used in the first gradient step, the function $f(x_\theta)$ moves from a linear approximation of $\sigma_\star$ in the RF limit ($\alpha_0 = 0$), to an accurate non-linear one ($\alpha_0 = 2.5$).

Learning a non-linear approximation of $\sigma_\star$ in the high-dimensional proportional regime therefore requires learning features. (Ba et al., 2022a) have proven that with $\eta = \Theta_d(1)$, the GEP holds even after a few gradient steps, corroborating a fact empirically observed by (Loureiro et al., 2021b). Indeed, they have shown that $\eta = \Theta_d(d)$ is *sufficient* to go beyond a linear approximation of $f_\star$ in this regime. (Moniri et al., 2023) considered intermediate scalings of step-size $\eta = \Theta_d(d^s)$ for $1/2 < s < 1$, which allows the network to fit target functions along $\boldsymbol{\theta}$ having finite degree. In this intermediate regime, the feature matrix can be approximated through a finite-number of spikes corresponding to increasing degree of functions along $\boldsymbol{\theta}$. Instead, we consider the full-scaling of $\eta = \Theta_d(d)$, where such a finite-spike approximation is insufficient and the network can fit arbitrary functions along $\boldsymbol{\theta}$. (Dandi et al., 2023a) proved that even if the target depends on multiple directions (multi-index model), only a (non-linear) function of a *s*ingle direction $\boldsymbol{\theta}$ can be learned with a single gradient step and $\eta = \Theta_d(d)$. This observation justifies the focus on single-index functions (256) on the regime of interest.

FURTHER RELATED WORKS

Random features — Random Features (RFs) were first introduced as a computationally efficient approximation to kernel methods (Rahimi et al., 2007b). Recently, they have enjoyed renewed interest also as models of two-layer neural networks in the lazy regime. Tight asymptotics for the random features model have been derived by (Goldt et al., 2020c; Goldt et al., 2022a; Gerace et al., 2020a; Mei et al., 2022c; Hu et al., 2022b; Dhifallah et al., 2022) in the two-layer case, and were extended to deep networks in (Schröder et al., 2023a; Bosch et al., 2023a) in the deep case. Importantly, with the exception of (Gerace et al., 2020a) who considered rotationally invariant weights and (Zavatone-Veth et al., 2023) for the case of deep *linear* random features, all these work assumed unstructured weights. In sharp contrast, gradient-trained neural networks have fundamentally structured weights. In the present manuscript, we consider such a case, when the weights are given by a bulk random matrix plus a rank-one spike emerging after a single large gradient step.

Feature learning regime – Perturbative feature learning corrections to the large-width lazy regime have been extensively studied in the literature (Yaida, 2020; Hanin et al., 2020; Dyer et al., 2020; Seroussi et al., 2023b; Naveh et al., 2021b; Bordelon et al., 2023). Our work radically contrasts with this line, since we exactly account for feature learning in the first step, *non-perturbatively* (note the gradient in (257) has a norm comparable with the initial weights). Beyond the lazy regime, a major recent development has been the understanding that the training dynamics of two-layer neural networks with small learning rate can be mapped to a Wasserstein gradient flow, known as the *mean-field regime* (Mei et al., 2018; Chizat et al., 2018; Rotskoff et al., 2022; Sirignano et al., 2020). Over the past few years, this flow was investigated under different classes of generative data models, such as staircase functions (Abbe et al., 2021; Abbe et al., 2022; Abbe et al., 2023), single-index (Berthier et al., 2023; Arnaboldi et al., 2023a) and multi-index models (Arnaboldi et al., 2023b), symmetric targets (Hajjar et al., 2023) and Gaussian mixture models (Refinetti et al., 2021b; Ben Arous et al., 2022).

9.2 MAIN TECHNICAL RESULTS

Our main technical results are a tight asymptotic characterization of the test error achieved by two-layer networks trained with a single large gradient step followed by a ridge regression on the readout weights. These results are enabled through the mapping to first an equivalent sRF model in subsection 9.2.1, which can in turn be mapped to a equivalent Gaussian model in subsection 9.2.2. Subsection 9.2.3 finally states the tight asymptotic characterization of the test error.

9.2.1 ASYMPTOTICS OF THE FIRST LAYER WEIGHTS AFTER ONE (LARGE) GRADIENT STEP

The first step is to derive an explicit asymptotic expression for the hidden-layer weights $W^{(1)}$ after one (large) gradient step. In the following, we show that the learning problem introduced in Section 9.1 is equivalent to a sRF model, which we first define.

Definition 9.2.1 (sRF model). *We define a sRF model with bulk variance c, spike strength r as the two-layer neural network*

$$g_{F,\boldsymbol{a}}(x) = \frac{1}{\sqrt{p}}\boldsymbol{a}^{\top}\sigma\left(Fx\right) \tag{260}$$

with trainable readout $\boldsymbol{a}$ and frozen random first layer weights:

$$F = W + r\frac{\boldsymbol{u}\boldsymbol{v}^{\top}}{\sqrt{d}}. \tag{261}$$

where W is a random matrix with rows independently sampled from $\mathbb{S}^{d-1}(\sqrt{c})$, and $\boldsymbol{u} \in \mathbb{S}^{d-1}(\sqrt{p}), \boldsymbol{v} \in \mathbb{S}^{d-1}(\sqrt{d})$. We further say that a sRF has alignment γ with $\boldsymbol{\theta}$ when $\boldsymbol{v}$ is uniformly sampled uniformly sampled among vectors with norm $\sqrt{d}$ satisfying $\boldsymbol{v}^{\top}\boldsymbol{\theta}/d = \gamma$.

The next result shows that after a large gradient step, our problem is asymptotically equivalent to a particular sRF model with spikes $\boldsymbol{u} = \mathbf{1}_p$ and $\boldsymbol{v}$ that correlated to the target weights $\boldsymbol{\theta}$. We give a tight characterization of the parameters c, r, γ.

Result 9.2.2 (Equivalence to a sRF model). *Consider two-layer networks with first-layer weights trained with a single gradient step of learning rate η from initial conditions $\boldsymbol{a}^{(0)} = \mathbf{1}_p/\sqrt{p}$ and $\boldsymbol{w}_i^{(0)}$ with unit norm $\|\boldsymbol{w}_i^{(0)}\| = 1$ and weak correlation $\boldsymbol{w}_i^{(0)} \cdot \boldsymbol{w}_j^{(0)} = O_d(1/\sqrt{d})$ for $i \neq j$ (eq. (257)). In the asymptotic limit $n_0, n, d, p \to \infty$, with $\alpha_0 = n_0/d, \alpha = n/d, \beta = p/d, \tilde{\eta} = \eta/d = \Theta_d(1)$, the test error achieved by performing ridge regression on the readout weights $\boldsymbol{a}$*

of the network after the gradient step (258) is identical to that achieved by an equivalent sRF model with parameters

$$c = 1 + \frac{\tilde{\eta}^2 h_1^2 \check{h}_1^2 h_2^\star}{\alpha_0 \beta^2} \tag{262}$$

$$r = \frac{\tilde{\eta} h_1}{\beta}\left(\frac{h_2^\star}{\alpha_0} + h_1^{\star 2}\right)^{1/2} \tag{263}$$

$$\gamma = \frac{h_1^\star}{\left(\frac{h_2^\star}{\alpha_0} + h_1^{\star 2}\right)^{1/2}} \tag{264}$$

where:

$$\begin{aligned} h_1 &= \mathbb{E}_z[z\sigma(z)], & h_1^\star &= \mathbb{E}_z[z\sigma_\star(z)], \\ \check{h}_1^2 &= \mathbb{E}_z[(\sigma'(z) - h_1)^2], & h_2^\star &= \mathbb{E}_z[\sigma_\star(z)^2], \end{aligned} \tag{265}$$

with $z \sim \mathcal{N}(0,1)$.

The derivation of Result (9.2.2) leverages the decomposition from (Ba et al., 2022a) of the gradient $\boldsymbol{g}_i = {}^{ru_i\boldsymbol{v}}/_{d} + \Delta$ into a rank-one term and a correction term. Note that above we have assumed a uniform initialization for the readout layer. This can be relaxed in the equivalence above, for instance for $\boldsymbol{a}^{(0)}$ taking a finite number of values, leading to finite-rank term instead of a single spike.

9.2.2 CONDITIONAL GAUSSIAN EQUIVALENCE

The sharp characterization of the test performance, which we state in Result 9.2.5 in the following subsection, is enabled by further mapping the sRF model to an exactly solvable (conditional) Gaussian model. We adapt the rigorous result in (Dandi et al., 2023a) (Theorem 4) by constructing explicitly the equivalent stochastic feature map, that we believe is of independent interest.

Result 9.2.3 (Conditional Gaussian Equivalence)**.** *Consider the sRF model with weights $F = W + r\boldsymbol{u}\boldsymbol{v}^\top/\sqrt{d}$, with $\boldsymbol{u} = \mathbf{1}_p$ and parameters c, r, γ, and the corresponding feature map given by*

$$\boldsymbol{\varphi}(\boldsymbol{x}) = \sigma(F\boldsymbol{x}) \tag{266}$$

Define the equivalent stochastic feature map

$$\boldsymbol{\varphi}^g(\boldsymbol{x}) \overset{d}{=} \mu_0(\kappa)\mathbf{1}_p + \mu_1(\kappa)W\boldsymbol{x} + \mu_2(\kappa)\mathcal{N}(0, \mathbb{I}_p), \tag{267}$$

where $\kappa \equiv \boldsymbol{v}^\top \boldsymbol{x}/\sqrt{d}$. We introduced the coefficients $\mu_0(\kappa), \mu_1(\kappa), \mu_2(\kappa)$ defined as

$$\begin{aligned}
\mu_0(\kappa) &= \mathbb{E}_z \sigma(z + r\kappa) \\
\mu_1(\kappa) &= \frac{1}{c}\mathbb{E}_z z\sigma(z + r\kappa) \\
\mu_2(\kappa) &= \sqrt{\mathbb{E}_z \sigma^2(z + r\kappa) - c(\mu_1(\kappa))^2 - (\mu_0(\kappa))^2},
\end{aligned} \tag{268}$$

with expectations bearing over $z \sim \mathcal{N}(0, c)$. The test error ε_g achieved by ridge regression

$$\hat{\boldsymbol{a}}_\lambda = \underset{\boldsymbol{a}\in\mathbb{R}^p}{\mathrm{argmin}} \frac{1}{2}\sum_{\mu=1}^{n_1}\left(y^\mu - \frac{1}{\sqrt{p}}\boldsymbol{a}^\top \phi(\boldsymbol{x})\right)^2 + \frac{\lambda}{2}\|\boldsymbol{a}\|^2. \tag{269}$$

is asymptotically identical for $\phi = \varphi$ and $\phi = \varphi^g$.

Result 9.2.3 extends the similar linearizations provided e.g. in (Goldt et al., 2020c; Hu et al., 2022b; Cui et al., 2023a) for unstructured RFs to sRFs (261), and we refer to Chapters 6, 7 and 8 for further discussions. Informally, the quantity κ in the stochastic feature map (267) represents the projection of the input on the spike defining the sRF $\kappa = \boldsymbol{x}^\top \boldsymbol{v}/\sqrt{d}$. The equivalent network $1/\sqrt{d}\boldsymbol{a}^\top \boldsymbol{\varphi}^g(x)$ obtained by replacing $\boldsymbol{\varphi}$ by the equivalent $\boldsymbol{\varphi}^g$ is a linear combination of terms such as $\mu_0(\kappa)$, $\mu_1(\kappa)\kappa$, $\mu_1(\kappa)\boldsymbol{a}^\top W(\Pi^\perp \boldsymbol{x})$, plus noise. On an intuitive level, this makes it apparent that sRFs can thus express *non-linear* functions of the component κ along the spike $\boldsymbol{v}$, but only linear functions of the component $\Pi^\perp \boldsymbol{x}$ orthogonal thereto. This feature learning correction to the linear regime is visually exemplified in Fig. 29. In the next subsection, we make this discussion more quantitative by providing a tight asymptotic characterization of the test error achieved by two-layer networks trained with a single large gradient step followed by a ridge regression on the readout weights.

9.2.3 TIGHT ASYMPTOTIC CHARACTERIZATION OF THE TEST ERROR

Finally, we leverage on the sequential mappings of Result 9.2.2 and Result 9.2.3 to offer sharp asymptotic guarantees on the test error achieved after training the readout weights using eq. (258).

Assumption 9.2.4. *Denote by $\{\boldsymbol{e}_i\}_{i=1}^p$ ($\{\boldsymbol{f}_i\}_{i=1}^d$) the left (resp. right) singular vectors of W. We further note $\{\lambda_i^\ell\}_{i=1}^p$ the squared singular values of W. The squared singular values $\{\lambda_i^\ell\}_{i=1}^p$ and the projection of the teacher vector $\boldsymbol{\theta}$ and* *the spike v on the eigenvectors $\{\boldsymbol{f}_i^\top \boldsymbol{v}\}_{i,\ell}$,$\{\boldsymbol{f}_i^\top \Pi^\perp \boldsymbol{\theta}\}_{i,\ell}$ are assumed to admit a well-defined joint distribution ν as $d \to \infty$.*

$$\frac{1}{p}\sum_{i=1}^{\min(p,d)} \delta\left(\lambda_i - \rho\right)\delta\left(\boldsymbol{f}_i^\top v - \tau\right)\delta\left(\boldsymbol{f}_i^\top \Pi^\perp \boldsymbol{\theta} - \pi\right) \xrightarrow{d\to\infty} \nu(\rho,\tau,\pi).$$

Result 9.2.5 (Test error asymptotics). *Consider the ERM problem (258) associated with the training of the readout weights $\boldsymbol{a}$, and assume 9.2.4 to hold. Define $\Pi^\perp \equiv \mathbb{I}_d - {}^{vv^\top}\!/_d$ the projection to the subspace orthogonal to the spike $\boldsymbol{v}$. In the asymptotic limit $d,p,n \to \infty$ with $\alpha = {}^n\!/_d, \beta = {}^p\!/_d = \Theta_d(1)$, the summary statistics*

$$\begin{aligned} q_1 &= \frac{\hat{\boldsymbol{a}}^\top W\Pi^\perp W^\top \hat{\boldsymbol{a}}}{p}, & q_2 &= \frac{\hat{\boldsymbol{a}}^\top \hat{\boldsymbol{a}}}{p}, & m &= \frac{\boldsymbol{1}_p^\top \hat{\boldsymbol{a}}}{\sqrt{p}}, \\ \zeta &= \frac{\hat{\boldsymbol{a}}^\top W\boldsymbol{v}}{\sqrt{dp}}, & \psi &= \frac{\hat{\boldsymbol{a}}^\top W\Pi^\perp \boldsymbol{\theta}}{\sqrt{dp}}, & \rho^2 &= \frac{\boldsymbol{\theta}^\top \Pi^\perp \boldsymbol{\theta}}{d} \end{aligned} \tag{270}$$

concentrate in probability to the solutions of the system of equations

$$\begin{cases} q_1 = \int d\nu(\rho,\tau,\pi)\rho\frac{\left(\hat{q}_1\rho+\hat{q}_2+\hat{\zeta}^2\rho\tau^2+\hat{\psi}^2\rho\pi^2\right)}{\left(\lambda+\hat{V}_1\rho+\hat{V}_2\right)^2} \\ \qquad -\beta\hat{\zeta}^2\frac{I(\hat{V}_1,\hat{V}_2)^2}{\left(1-\beta\hat{V}_1 I(\hat{V}_1,\hat{V}_2)\right)^2} \\ \qquad -\hat{\zeta}^2\frac{\int d\nu(\rho,\tau,\pi)\frac{\tau^2\rho^2}{(\lambda+\hat{V}_1\rho+\hat{V}_2)^2}\left[\left(1-\beta\hat{V}_1 I(\hat{V}_1,\hat{V}_2)\right)^2-1\right]}{\left(1-\beta\hat{V}_1 I(\hat{V}_1,\hat{V}_2)\right)^2} \\ q_2 = \int d\nu(\rho,\tau,\pi)\frac{\left(\hat{q}_1\rho+\hat{q}_2+\hat{\zeta}^2\rho\tau^2+\hat{\psi}^2\rho\pi^2\right)}{\left(\lambda+\hat{V}_1\rho+\hat{V}_2\right)^2} \\ \qquad -\hat{\zeta}^2\int d\nu(\rho,\tau,\pi)\frac{\tau^2\rho}{(\lambda+\hat{V}_1\rho+\hat{V}_2)^2}\left[1-\frac{1}{\left(1-\beta\hat{V}_1 I(\hat{V}_1,\hat{V}_2)\right)^2}\right] \\ V_1 = \int d\nu(\rho,\tau,\pi)\rho\frac{1}{\lambda+\hat{V}_1\rho+\hat{V}_2} \\ V_2 = \int d\nu(\rho,\tau,\pi)\frac{1}{\lambda+\hat{V}_1\rho+\hat{V}_2} \\ m = \frac{1}{\mathbb{E}_\kappa\left[\frac{\mu_0(\kappa)^2}{1+V(\kappa)}\right]}\mathbb{E}_{\kappa,y}\left[\frac{\mu_0(\kappa)(\sigma_\star(\kappa,y)-\mu_1(\kappa)\kappa\zeta)}{1+V(\kappa)}\right] \\ \zeta = \hat{\zeta}\sqrt{\beta}\int d\nu(\rho,\tau,\pi)\rho\tau^2\frac{1}{\lambda+\hat{V}_1\rho+\hat{V}_2} \\ \qquad +\beta^{3/2}\hat{\zeta}\hat{V}_1\frac{I(\hat{V}_1,\hat{V}_2)^2}{1-\beta\hat{V}_1 I(\hat{V}_1,\hat{V}_2)} \\ \psi = \hat{\psi}\sqrt{\beta}\int d\nu(\rho,\tau,\pi)\rho\pi^2\frac{1}{\lambda+\hat{V}_1\rho+\hat{V}_2} \end{cases}, \tag{271}$$

$$\begin{cases} \hat{V}_1 = \frac{\alpha}{\beta}\mathbb{E}_{\kappa}\frac{\rho\mu_1(\kappa)^2}{1+V(\kappa)} \\ \hat{q}_1 = \frac{\alpha}{\beta}\mathbb{E}_{\kappa,y}\mu_1(\kappa)^2\frac{b(\kappa,y)^2+\rho q(\kappa)-\mu_1(\kappa)^2\psi^2}{(1+V(\kappa))^2} \\ \hat{V}_2 = \frac{\alpha}{\beta}\mathbb{E}_{\kappa}\frac{\rho\mu_2(\kappa)^2}{1+V(\kappa)} \\ \hat{q}_2 = \frac{\alpha}{\beta}\mathbb{E}_{\kappa,y}\mu_2(\kappa)^2\frac{b(\kappa,y)^2+\rho q(\kappa)-\mu_1(\kappa)^2\psi^2}{(1+V(\kappa))^2} \\ \hat{\zeta} = \frac{\alpha}{\sqrt{\beta}}\mathbb{E}_{\kappa,y}\kappa\mu_1(\kappa)\frac{b(\kappa,y)}{1+V(\kappa)} \\ \hat{\psi} = \frac{\alpha}{\sqrt{\beta}}\mathbb{E}_{\kappa,y}\frac{y\mu_1(\kappa)b(\kappa,y)+\psi\mu_1(\kappa)^2}{1+V(\kappa)} \end{cases} \tag{272}$$

and

$$\rho^2 = 1-\gamma^2. \tag{273}$$

We introduced the shorthands

$$b(y,\kappa) \equiv \sigma_\star(\gamma\kappa+\sqrt{1-\gamma^2}y)-\mu_0(\kappa)m-\kappa\mu_1(\kappa)\zeta-\mu_1(\kappa)\psi y \tag{274}$$

$$V(\kappa) \equiv \mu_1(\kappa)^2V_1+\mu_2(\kappa)^2V_2 \tag{275}$$

$$q(\kappa) \equiv \mu_1(\kappa)^2q_1+\mu_2(\kappa)^2q_2, \tag{276}$$

$$I(\hat{V}_1,\hat{V}_2) = \int d\nu(\rho,\tau,\pi)\frac{\tau^2\rho}{\hat{V}_1\rho+\hat{V}_2+\lambda} \tag{277}$$

In (271)*, the expectations over* κ, y *bear over standard Gaussian variables. We remind that the quantities* r, c, γ *are characterized in Result* 9.2.2.
Finally, the test error (259) *admits the sharp characterization*

$$\epsilon_g = \mathbb{E}_{\kappa,y}\left[\left(\sigma_\star(\gamma\kappa+\sqrt{1-\gamma^2}y)-\mu_0(\kappa)m-\mu_1(\kappa)\kappa\zeta-\frac{\mu_1(\kappa)\psi}{\sqrt{\rho}}y\right)^2+q(\kappa)-\frac{\mu_1(\kappa)^2\psi^2}{\rho}\right] \tag{278}$$

where expectations bear over standard Gaussian variables.

Result 9.2.5 thus rephrases the high-dimensional learning problem (258) in terms of a finite set of scalar summary statistics which can be efficiently evaluated, yielding excellent agreement with finite d numerical simulations, see Figs. 30-32. While we state Result 9.2.5 for the square loss and ℓ_2 regularization for clarity, a sharp characterization can be derived for any generic convex loss ℓ. While we have used the (non-rigorous) replica method to derive these equations, an interesting avenue of future research is to provide a rigorous probabilistic proof. A possible way to address the problem is to apply Gordon's Gaussian min-max inequalities (Gordon, 1988; Thrampoulidis et al., 2014; Stojnic, 2013b; Stojnic, 2013c)), and generalized the approach used for pure random features (Hu et al., 2022b; Loureiro et al., 2021b). This is, however, beyond the scope of this manuscript.

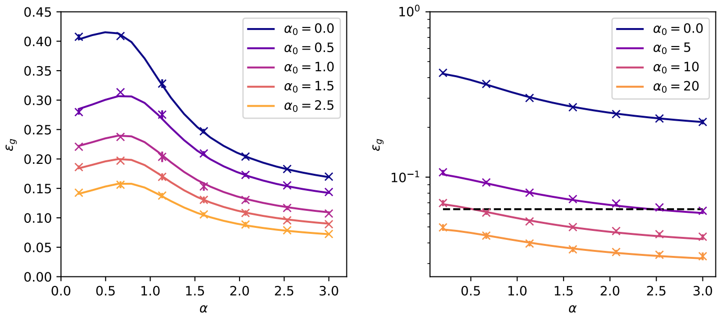


Figure 30: Test error achieved by a two-layer network with activation σ whose first layer has been trained following the protocol detailed in section 9.1 on a single-index target with activation $\sigma_\star$. (**Left**) Activations $\sigma = \sigma_\star = \tanh$, learning rate $\tilde{\eta} = 1$, readout regularization $\lambda = 0.01$. (**Right**) Activations $\sigma = \tanh, \sigma_\star = \sin$, learning rate $\tilde{\eta} = 3$, readout regularization $\lambda = 0.1$. The dashed black line represents the lowest achievable MSE for kernel/-linear methods, namely $h_2^\star - (h_1^\star)^2$ (Ba et al., 2022a). In both plots, solid lines correspond to the theoretical Result 9.2.5, and crosses correspond numerical experiments in $d = 3000$ (left) and $d = 2000$ (right). All points were averaged over 5 instances. Different colours represent different initial sample complexities $\alpha_0 = {}^{n_0}/_{d}$ used for the first gradient step.

9.3 DISCUSSION OF MAIN RESULTS

While the self-consistent equations in Result 9.2.5 might appear cumbersome, they offer valuable insight into the mechanism behind feature learning in two-layer neural networks trained with gradient descent. In this section, we discuss and highlight some of these insights.

9.3.1 SPIKED RANDOM FEATURES VS RANDOM FEATURES

The asymptotic characterization 9.2.5 encompasses, as a special case, usual RFs (when setting the spike strength r to zero). More precisely, for zero spike strength $r = 0$ in (261), sRFs coincide with RFs, as the coefficients (268) lose their κ dependence, reducing to usual Hermite coefficients. The equivalent feature map $\boldsymbol{\varphi}^g$ then reduces to the Gaussian equivalent feature map employed in e.g. (Goldt et al., 2020c; Hu et al., 2022b; Schröder et al., 2023a). Importantly, while the equivalent feature map for unspiked RFs is *linear* in the input $\boldsymbol{x}$, the equivalent feature map $\boldsymbol{\varphi}^g$ (267) of Result 9.2.3 is *non-linear* in the component κ of the input $\boldsymbol{x}$.

sRF models, which are equivalent to two-layer networks after a single large gradient step through Result 9.2.2, offer an ideal playground to test the intertwined influence of the spike/target correlation and the test performance of the model. Fig. 32 (left) presents the learning curves of (s)RFs for varying spike strengths r. As is intuitive, larger spikes allow the sRF to more easily express non-linear function in the direction of the spike, and thus lead to relatively smaller test errors. Furthermore, note that even rather small spike strengths $r = 0.2$ already yield test errors which are sizably lower than vanilla (unspiked) RFs ($r = 0$), hinting at the qualitative difference between sRF and RF models discussed in section 9.1 and Fig. 29. This is in qualitative agreement with the result of (Moniri et al., 2023), who studied the intermediate scaling regime where the learning rate $\eta = \Theta_d(d^s)$ with $s \in (1/2, 1)$, and have shown that in this regime a polynomial approximation of $\sigma_\star$ is learned.

The plot shows a compelling agreement between the theoretical predictions (continuous line) and numerical simulations. Fig. 32 (right) represents the theoretical closed-form expression for the test error (278) for different values of the spike/target alignment γ, for a single-index target sign($\boldsymbol{\theta}^\top \boldsymbol{x}$), with good agreement with numerical experiments. Higher alignments γ lead to overall lower test errors escaping the linear curse of Gaussian models.

9.3.2 BEATING KERNELS IN A SINGLE STEP

A key parameter in our formulas is $\gamma = \boldsymbol{v}^\top \boldsymbol{\theta}/d \in [0, 1]$, the correlation between the effective gradient spike and the target weights. From its asymptotic expression eq. 262 in Result 9.2.2, this is an increasing function of $\alpha_0 = n_0/d$, the number of samples in the first step. As shown in Fig. 30 (left) for $\sigma = \sigma_\star = \tanh$, larger sample complexities in the representation learning step α_0 allow for better feature learning when implementing a gradient step on the first layer, enabling a lower test error after the readout layer is retrained. As expected, the lowest error is achieved when $\alpha_0 \to \infty$, in which case the spike $\boldsymbol{v}$ is perfectly aligned with the target weights $\boldsymbol{\theta}$ and $\gamma = 1$, as can be seen from (262).

Figure 30 (right) presents similar curves for another target activation $\sigma^\star = \sin$, including the test error achieved by the network at initialization ($\alpha_0 = 0$), which corresponds to a usual RF model. The latter is well above the lowest MSE achievable by a linear estimator (plotted as a dashed black line), namely the projection of the teacher function on Hermite polynomials $||\mathbb{P}_{>1} f_\star||_2^2$. This performance corresponds to the kernel one when the number of samples scales proportionally with the input dimension (Ghorbani et al., 2020a); using the notations of Result 9.2.2 the best linear MSE is readily written as $h_2^\star - (h_1^\star)^2$. In sharp contrast, networks with trained first layers ($\alpha_0 > 0$) can learn non-linear functions of the inputs and outperform this baseline.

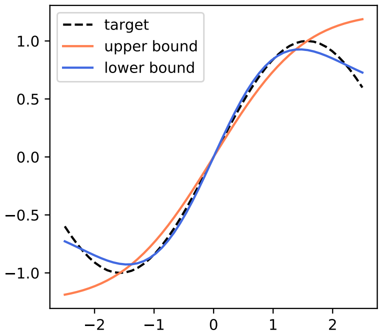


Figure 31: ($c = 1, r = 0.9$ and $\gamma = 1$) Illustration of the functions realizing the upper bound (279) (orange) and lower bound (280) (blue), for $\sigma = \tanh$, for a target $\sigma_\star = \sin$ (dashed black).

9.3.3 WHAT CAN BE LEARNED WITH A SINGLE STEP?

While training a two-layer network with large gradient steps allows it to escape the linear limitations of, e.g. RF models, it generically only learns the target up to small test error, yet not perfectly. In fact, a closer examination of the sharp asymptotic expression (278) for ε_g in Result 9.2.5 reveals that even for large initial batches $\alpha_0 \to \infty$ (and thus perfect spike/target alignments $\gamma = 1$), at fixed learning rate strength $\tilde{\eta}$ and sample complexity $\alpha > 0$, the test error optimized over the regularization λ is upper bounded as

$$\inf_{\lambda \geq 0} \varepsilon_g \leq \inf_{b_1} \mathbb{E}_\kappa \left[\sigma_\star(\kappa) - b_1 \mu_0(\kappa)\right]^2, \tag{279}$$

and lower bounded by:

$$\inf_{\lambda \geq 0} \varepsilon_g \geq \inf_{b_1, b_2} \mathbb{E}_\kappa \left[\sigma_\star(\kappa) - b_1 \mu_0(\kappa) - b_2 \mu_1(\kappa)\kappa\right]^2. \tag{280}$$

The upper bound (279) is the equivalent of Lemma 6 of (Ba et al., 2022a) in our case of uniform readout initialization $\boldsymbol{a}^{(0)} = 1_p/\sqrt{p}$, and is achieved for $\lambda \to \infty$. The lower bound (280), however, shows that the test error cannot be lower than the Gaussian-weighted L^2 distance between the target function $\sigma_\star$ and $\mathrm{span}(\mu_0, \tilde{\mu}_1)$, where $\tilde{\mu}_1(\kappa) \equiv \kappa\mu_1(\kappa)$, and the best approximation would be reached for the projection of the target thereon. Fig. 31 illustrates what functions realize the upper (279) and lower bounds (280), and how they compare with the target $\sigma_\star$. Finally, note that in the vanilla RF limit $r = 0$, the functions $\mu_0(\kappa), \mu_1(\kappa)$ reduce to constants independent of κ, constraining the class of functions that can be learned to that of linear functions.

Finally, let us mention that while this discussion was made at fixed learning rate $\tilde{\eta}$ for clarity, the latter is in practice a tunable hyper-parameter, and the functions μ_0, μ_1 depend thereupon via the spike strength r (268). One can thus refine – and lower– the bounds (279) and (280) over the functions

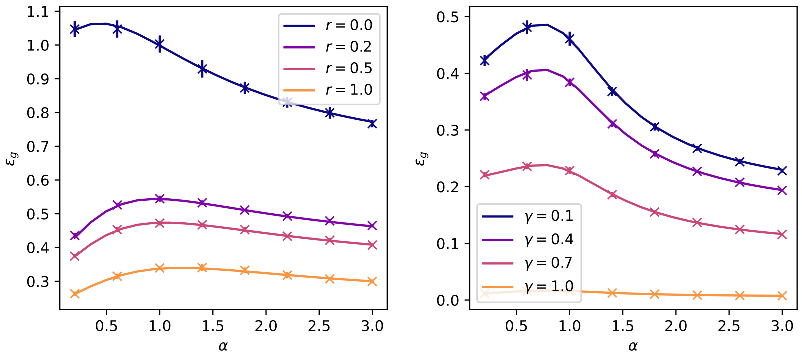


Figure 32: Test error for a sRF with activation σ learning from a single-index model $\sigma_\star(\boldsymbol{\theta}^\top \boldsymbol{x}/\sqrt{d})$, with regularization $\lambda = 0.1$. **(Left)** $\sigma = \sin, \sigma_\star = sign$. Different colours corrwespond to different spike strengths r (261), with $r = 0$ corresponding to the vanilla RF model. **(Right)** $\sigma = \sigma_\star = \tanh$. Different colours correspond to different overlaps $\gamma \equiv \boldsymbol{\theta}^\top \boldsymbol{v}/d$ between the target weights $\boldsymbol{\theta}$ and the spike $\boldsymbol{v}$. Solid lines: theoretical characterization of Result 9.2.5. Crosses : numerical simulations in dimensions $d = p = 2000$. Each point is averaged over 10 instances of the problem.

μ_0, μ_1, which can take values in the realizable set $\mathscr{M} = \{(\mu_0(r), \mu_1(r)\}_{r \geq 0}$ (emphasizing the dependence on r via equation (268)) as $\tilde{\eta}$ is varied:

$$\inf_{\lambda \geq 0, \tilde{\eta} \geq 0} \epsilon_g \leq \inf_{b_1 \in \mathbb{R}, \nu_0 \in \mathscr{M}} \mathbb{E}_\kappa \left[\sigma_\star(\kappa) - b_1 \nu_0(\kappa)\right]^2, \tag{281}$$

$$\inf_{\lambda \geq 0, \tilde{\eta} \geq 0} \epsilon_g \geq \inf_{\substack{b_1, b_2 \in \mathbb{R}, \\ (\nu_0, \nu_1) \in \mathscr{M}}} \mathbb{E}_\kappa \left[\sigma_\star(\kappa) - b_1 \nu_0(\kappa) - b_2 \nu_1(\kappa)\kappa\right]^2.$$

In other words, by tuning the learning rate $\tilde{\eta}$ – and thus the spike strength r–, one gains the freedom to choose the "best" subspace $\mathrm{span}(\mu_0(r), \mu_1(r))$, i.e. the one allowing to approximate the target $\sigma_\star$ best.

Finally, as discussed in (Ba et al., 2022a), observe that for $\sigma = \sigma_\star = \mathrm{erf}$, the upper bound in (281) is zero provided one tunes the learning rate to $\tilde{\eta} = \sqrt{3\beta}/h_1$, and perfect learning is therefore achievable.

9.3.4 MORE VARIABILITY MEANS BETTER FEATURE LEARNING

The discussion of subsection 9.3.3 thus affords an insightful perspective on the learning of two-layer neural networks in terms of approximating the target activation $\sigma_\star$ in a two-dimensional functional space. Interestingly, introducing variability in the readout layer at initialization leads to an even richer functional basis, and hence greater expressivity of the network. When $\boldsymbol{a}^{(0)}$ is no longer proportional to $\mathbf{1}_p$, but is rather initialized from a distribution over a finite vocabulary V of size $|V| > 1$–e.g. $V = \{-1, 0, +1\}$– the equivalent feature map takes the form:

$$\boldsymbol{\varphi}^g(\boldsymbol{x}) = \begin{pmatrix} \mu_0(u_1\kappa) \\ \mu_0(u_2\kappa) \\ \vdots \\ \mu_0(u_p\kappa) \end{pmatrix} + \begin{pmatrix} \mu_1(u_1\kappa) \\ \mu_1(u_2\kappa) \\ \vdots \\ \mu_1(u_p\kappa) \end{pmatrix} \odot W\boldsymbol{x} + \begin{pmatrix} \mu_2(u_1\kappa) \\ \mu_2(u_2\kappa) \\ \vdots \\ \mu_2(u_p\kappa) \end{pmatrix} \odot \boldsymbol{\xi} \tag{282}$$

where $\odot$ denotes element-wise multiplication, $\boldsymbol{\xi} \sim \mathcal{N}(0, \mathbb{I}_p)$, and $\boldsymbol{u} \not\propto \mathbf{1}_p$ has entries which can now take a finite number of different values. The coefficients of the equivalent map (282) are thus neuron-dependent and thus afford a richer functional basis $\{\mu_0(\omega\cdot), \tilde{\mu}_1(\omega\cdot)\}_{\omega\in V}$, thereby allowing the network to express a larger class of functions. As a matter of fact, the functional space spanned by these functions is generically of dimension $2|V|$ for non-uniform readout initializations $\boldsymbol{a}^{(0)}$, compared to just 2 in the uniform readout case. A sharp asymptotic characterization of the test error for the case of non-uniform readout initialization can also be reached along similar lines as Result 9.2.5.

We briefly discuss the limiting case of interest $\lambda, \alpha_0, \tilde{\eta} \to \infty$, for which the equivalent feature map $\boldsymbol{\varphi}^g(\boldsymbol{x})$ (282) reduces to its first term $(\mu_0(u_i\kappa))_{i=1}^p$. Further observe that $\mu_0(u_i\kappa)$ can be viewed as a one-dimensional neuron acting on the one-dimensional input variable κ with a random weight u_i. Standard results on approximation errors for random feature mappings of finite-dimensional inputs (Rahimi et al., 2007b) imply that a large class of functions can be approximated from the network features $\boldsymbol{\varphi}^g(\boldsymbol{x})$, provided the vocabulary size $|V|$ is large enough. Similar random features approximations have been leveraged in (Ba et al., 2022a; Ba et al., 2023; Damian et al., 2022). The equivalent feature map (282) provides an intuitive picture on how such random feature mappings of low-dimensional inputs can naturally emerge in the setting of the learning of a two-layer network.

Finally, note that the bounds (279) and (280) can be readily generalised to non-uniform readout initializations, provided one replaces in the discussion of subsection 9.3.3 the two-dimensional functional basis $\mathrm{span}(\mu_0, \tilde{\mu}_1)$ in the case of uniform initialization by the richer functional space $\mathrm{span}(\{\mu_0(\omega\cdot), \tilde{\mu}_1(\omega\cdot)\}_{\omega\in V})$ for non-uniform initialization. For instance, the lower bound (280) thus involves in the non-uniform case the distance between the target $\sigma_\star$ and $\mathrm{span}(\{\mu_0(\omega\cdot), \tilde{\mu}_1(\omega\cdot)\}_{\omega\in V})$.

CONCLUSION

We provided a tight asymptotic description of the learning of a two-layer neural network after training its first layer with a large, single, gradient step, in the limit where the number of samples, the hidden layer width, and the input dimension are proportionally large. Our results sharply characterize the feature maps learned from the data, and how it achieves a test error which non-perturbatively improves over the kernel regime. Crucially, the trained network can efficiently approximate non-linear functions in the direction of the gradient – sizeably improving upon the network at initialization, which can only express linear functions. We further discuss bounds for the test error and which functions are learnable after a single gradient step. Finally, extensive numerical support is provided to illustrate our findings. To our knowledge, our work provides the first exact asymptotic result in the non-perturbative $\eta = \Theta_d(d)$ regime for feature learning.

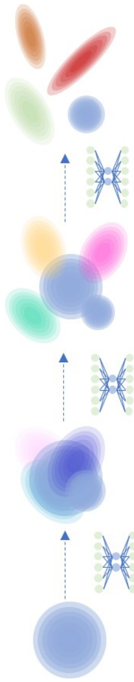
We believe the present work opens exciting research avenues, paving the way towards a tight theoretical understanding of feature learning in gradient-trained networks. Prominent among these research directions is the extension of our results to readout initialization with generic (not necessarily finite) support. to a finite number of gradient steps, and ultimately fully trained networks.

Part IV

ASPECTS OF MODERN ML

OUTLINE AND MOTIVATIONS

Parts II and III focused on the study of supervised tasks with FNN architectures. While such settings have historically been the principal focal point of theoretical effort in ML research, they represent but a small part of the modern zoology of ML tasks and architectures. In fact, partly prompted by the need to match the fast pace of empirical breakthroughs, the emphasis in ML theory is increasingly shifting towards other learning paradigms. On the one hand, diffusion and flow-based models (Ho et al., 2020a; Sohl-Dickstein et al., 2015) introduced a new framework for the sampling of complex distributions, offering efficient alternatives to GAN-based methods. The training and downstream deployment of such models follow a slightly different pattern from vanilla supervised learning, as we discuss further in Chapter 11, and thus represent unique technical challenges for theoretical analyses. Secondly, recent years have witnessed quick-paced progress in the learning of language data. Transformer architectures (Vaswani et al., 2017) have in particular taken center stage as powerful and computationally efficient feature extractors for sequential data. The rapid diversification of ML practice warrants matching developments at the level of ML theory. Part IV presents tight asymptotic analyses – to our awareness the first– of the learning of a flow-based generative model and a non-linear attention layer from limited data.



In diffusion models, a ANN- parametrized velocity field transports an simple base distribution (bottom) to a complex target density (top).

FLOW-BASED GENERATIVE MODELS

The first two Chapters of Part IV investigate the learning of flow-based generative models. These models allow to sample from realistic data distributions, or approximations thereof, by transporting an easy-to-sample base density into an approximation of the target density. The velocity field associated to this transport map is typically parametrized by an ANN, trained to denoise a series of increasingly noisier versions of the training data.

Chapter 10 first focuses on analyzing this building block in isolation, and discusses self-supervised denoising tasks with denoiser ANNs. It more precisely considers the learning of the simplest instance of such networks, the DAE, when trained to denoise data sampled from a high-dimensional Gaussian mixture density. Chapter 11 builds upon these insights to analyze the performance of a DAE-parametrized flow-based generative model in generating samples from a bimodal Gaussian mixture, when learning from n training points. It provides a tight description of the learning and transport processes, and shows how the model learns to generate an approximation of the target distribution with a $\Theta_n(1/n)$ error.

DOT-PRODUCT ATTENTION

The information encoded in sequential data such as language is embedded in the tokens themselves (a.k.a their *semantics*) and their respective ordering within the sequence (a.k.a their *positions*). Attention layers (Vaswani et al., 2017) are a priori able to implement both semantic and positional mechanisms, i.e. to extract either type of information from a sequence. Under what condition do transformers learn to implement one mechanism or the other, and to which relative extents? Chapter 12 investigates this question in a theoretically controlled setting, namely that of supervised learning with a single attention head parametrized by low-rank matrices. Two local minima are found in the ERM loss landscape, respectively corresponding to positional and semantic mechanisms, with the former (resp. latter) being global at small (resp. large) sample complexities. The analysis reveals a *first-order phase transition* between the learning of the two mechanisms.

10

SELF-SUPERVISED LEARNING WITH AUTO-ENCODERS

Machine learning techniques have a long history of success in denoising tasks. The recent breakthrough of diffusion-based generation (Song et al., 2021; Ho et al., 2020b) has further revived the interest in denoising networks, demonstrating how they can also be leveraged, beyond denoising, for generative tasks. However, this rapidly expanding range of applications stands in sharp contrast to the relatively scarce theoretical understanding of denoising neural networks, even for the simplest instance thereof – namely DAEs (Vincent et al., 2010).

Theoretical studies of autoencoders have hitherto almost exclusively focused on data compression tasks using Reconstruction Auto-Encoder (RAE)s, where the goal is to learn a concise latent representation of the data. A majority of this body of work addresses *linear* autoencoders (Oftadeh et al., 2020; Kunin et al., 2019; Bao et al., 2020; Gidel et al., 2019). The authors of (Refinetti et al., 2022; Shevchenko et al., 2022) analyze the gradient-based training of non-linear autoencoders with online stochastic gradient descent or in population, thus implicitly assuming the availability of an infinite number of training samples. Furthermore, two-layer RAEs were shown to learn to essentially perform Principal Component Analysis (PCA) (Eckart et al., 1936; Bourlard et al., 1988; Baldi et al., 1989), i.e. to learn a linear model. Ref. (Nguyen, 2021) shows that this is also true for infinite-width architectures. Learning in DAEs has been the object of theoretical investigations only in the linear case (Pretorius et al., 2018), while the case of non-linear DAEs remains theoretically largely unexplored.

MAIN CONTRIBUTIONS

The present work considers the problem of denoising data sampled from a Gaussian mixture by learning a two-layer DAE with a skip connection and tied weights via empirical risk minimization. Throughout the manuscript, we consider the high-dimensional limit where the number of training samples n and the dimension d are large ($n, d \to \infty$) while remaining comparable, i.e. $\alpha \equiv {}^{n}/_{d} = \Theta(1)$. Our main contributions are:

- Leveraging the replica method, we provide sharp, closed-form formulae for the MSE for DAEs, as a function of the sample complexity α and the problem parameters. We also provide a sharp characterization for other learning metrics including the weights norms, skip connection strength, and cosine similarity between the weights and the cluster means. These formulae encompass as a corollary the case of RAEs. We show that these formulae also describe quantitatively rather well the denoising MSE for *real* data sets, including MNIST (LeCun et al., 1998b) and FashionMNIST (Xiao et al., 2017).

- We find that PCA denoising (namely denoising by projecting the noisy data along the principal component of the training samples) is widely sub-optimal compared to the DAE, leading to a MSE superior by a difference of $\Theta(d)$, thereby establishing that DAEs do *not* simply learn to perform PCA.

- Building on the formulae, we quantify the role of each component of the DAE architecture (skip connection and the bottleneck network) in its overall performance. We find that the two components have complementary effects in the denoising process –namely preserving the data nuances and removing the noise– and discuss how the training of the DAE results from a tradeoff between these effects.

RELATED WORKS

Theory of autoencoders – Various aspects of RAEs have been studied, for example, memorization (Radhakrishnan et al., 2019), or latent space alignment (Jain et al., 2021). However, the largest body of work has been dedicated to the analysis of gradient-based algorithms when training RAEs. Ref. (Kunin et al., 2019) established that minimizing the training loss leads to learning the principal components of the data. Authors of (Bourlard et al., 1988; Baldi et al., 1989) have analyzed how a linear RAE learns these components during training. These studies were later extended to non-linear networks by (Nguyen et al., 2019b; Refinetti et al., 2022; Shevchenko et al., 2022), at the sacrifice of further assuming an infinite number of training samples to be available –either by considering online stochastic gradient descent, or the population loss. Refs. (Nguyen et al., 2019a; Nguyen, 2021) are able to address a finite sample complexity, but in exchange, have to consider infinite-width architectures, which (Nguyen, 2021) further shows, also tend to a large extent to learn to perform PCA.

Exact asymptotics from the replica method – The replica method (Parisi, 1979b; Parisi, 1983a; Zdeborová et al., 2015; Gabrié, 2019) has proven a very valuable gateway to access sharp asymptotic characterizations of learning metrics for high-dimensional machine learning problems. Past works have addressed –among others– single-(Gardner et al., 1988; Opper et al., 1991b; Barbier et al., 2019a; Aubin et al., 2020a) and multi-index models (Aubin et al., 2018a), or kernel methods (Dietrich et al., 1999b; Bordelon et al., 2020; Gerace et al., 2020a; Cui et al., 2023c). While the approach has traditionally addressed convex problems, for which its prediction can be proven e.g. using the convex Gordon minimax theorem (Thrampoulidis et al., 2018), the replica method allows to average over *all* the global minimizers of the loss, and therefore also accommodates non-convex settings. Refs. (Zavatone-Veth et al., 2022a; Cui et al., 2023b) are two recent examples of its application to non-convex losses. In the present manuscript, we leverage this versatility to study the minimization of the empirical risk of DAEs, whose non-convexity represents a considerable hurdle to many other types of analyses.

10.1 SETTING

Data model We consider the problem of denoising data $\boldsymbol{x} \in \mathbb{R}^d$ corrupted by Gaussian white noise of variance Δ,

$$\tilde{\boldsymbol{x}} = \sqrt{1-\Delta}\boldsymbol{x} + \sqrt{\Delta}\boldsymbol{\xi},$$

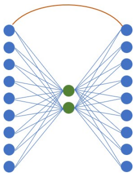
where we denoted $\tilde{\boldsymbol{x}}$ the noisy data point, and $\boldsymbol{\xi} \sim \mathcal{N}(0, \mathbb{I}_d)$ the additive noise. The rescaling of the clean data point by a factor $\sqrt{1-\Delta}$ is a practical choice that entails no loss of generality, and allows to easily interpolate between the noiseless case ($\Delta = 0$) and the case where the signal-to-noise ratio vanishes ($\Delta = 1$). Furthermore, it allows us to seamlessly connect with works on diffusion-based generative models, where the rescaling naturally follows from the way the data is corrupted by an Ornstein-Uhlenbeck process (Song et al., 2021; Ho et al., 2020b). In the present work, we assume the clean data $\boldsymbol{x}$ to be drawn from a Gaussian mixture distribution $\mathbb{P}$ with K clusters

$$\boldsymbol{x} \sim \sum_{k=1}^{K} \rho_k \mathcal{N}(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k). \tag{283}$$

The $k-$th cluster is thus centered around $\boldsymbol{\mu}_k \in \mathbb{R}^d$, has covariance $\boldsymbol{\Sigma}_k \succeq 0$, and relative weight ρ_k.

DAE model An algorithmic way to retrieve the clean data $\boldsymbol{x}$ from the noisy data $\tilde{\boldsymbol{x}}$ is to build a neural network taking the latter as an input and yielding the former as an output. A particularly natural choice for such a network is an autoencoder architecture (Vincent et al., 2010). The intuition is that the narrow hidden layer of an autoencoder forces the network to learn a succinct latent representation of the data, which is robust against noise corruption of the input. In this work, we analyze a two-layer DAE. We further assume that the weights are tied. Additionally, mirroring modern denoising architectures like U-nets (Ronneberger et al., 2015b) or (Mao et al., 2016; Tong et al., 2017; Kim et al., 2016a; Kim et al., 2016b), we also allow for a (trainable) skip-connection:

$$f_{b,\boldsymbol{w}}(\tilde{\boldsymbol{x}}) = b \times \tilde{\boldsymbol{x}} + \frac{\boldsymbol{w}^\top}{\sqrt{d}} \sigma\left(\frac{\boldsymbol{w}\tilde{\boldsymbol{x}}}{\sqrt{d}}\right). \tag{284}$$



DAE architecture, for $p = 2$.

The DAE (284) is therefore parametrized by the scalar skip connection strength $b \in \mathbb{R}$ and the weights $\boldsymbol{w} \in \mathbb{R}^{p\times d}$, with p the width of the DAE hidden layer. The normalization of the weight $\boldsymbol{w}$ by $\sqrt{d}$ in (284) is the natural choice which ensures for high dimensional settings $d \gg 1$ that the argument of the non-linearity $\sigma(\cdot)$ stays $\Theta(1)$. Like (Refinetti et al., 2022), we focus on the case with $p \ll d$. The assumption of weight-tying affords a more concise theoretical characterization and thus clearer discussions. Note that it is also a strategy with substantial practical history, dating back to (Vincent et al., 2010), as it prevents the DAE from functioning in the linear region of its non-linearity $\sigma(\cdot)$. This choice of architecture is also motivated by a particular case of Tweedie's formula (Efron, 2011), which will be the object of further discussion in Section 10.3.

We also consider two other simple architectures

$$u_{\boldsymbol{v}}(\tilde{\boldsymbol{x}}) = \frac{\boldsymbol{v}^\top}{\sqrt{d}} \sigma\left(\frac{\boldsymbol{v}\tilde{\boldsymbol{x}}}{\sqrt{d}}\right), \qquad\qquad r_c(\tilde{\boldsymbol{x}}) = c \times \tilde{\boldsymbol{x}}, \tag{285}$$

which correspond to the building blocks of the complete DAE architecture $f_{b,\boldsymbol{w}}$ (284) (hereafter referred to as the *full* DAE). Note that indeed $f_{b,\boldsymbol{w}} = r_b + u_{\boldsymbol{w}}$. The part $u_{\boldsymbol{v}}(\cdot)$ is a DAE without skip connection (hereafter called the *bottleneck network* component), while $r_c(\cdot)$ correspond to a simple single-parameter trainable rescaling of the input (hereafter called the *rescaling* component).

To train the DAE (284), we assume the availability of a training set $\mathscr{D} = \{\tilde{\boldsymbol{x}}^\mu, \boldsymbol{x}^\mu\}_{\mu=1}^n$, with n clean samples $\boldsymbol{x}^\mu$ drawn i.i.d from $\mathbb{P}$ (283) and the corresponding noisy samples $\tilde{\boldsymbol{x}}^\mu = \boldsymbol{x}^\mu + \boldsymbol{\xi}^\mu$ (with the noises $\boldsymbol{\xi}^\mu$ assumed mutually independent). The DAE is trained to recover the clean samples $\boldsymbol{x}^\mu$ from the noisy samples $\tilde{\boldsymbol{x}}^\mu$ by minimizing the empirical risk [1]

$$\hat{\mathscr{R}}(b,\boldsymbol{w}) = \sum_{\mu=1}^{n} \|\boldsymbol{x}^\mu - f_{b,\boldsymbol{w}}(\tilde{\boldsymbol{x}}^\mu)\|^2 + g(\boldsymbol{w}), \tag{286}$$

where $g : \mathbb{R}^{p\times d} \to \mathbb{R}_+$ is an arbitrary convex regularizing function. We denote by $\hat{b}, \hat{\boldsymbol{w}}$ the minimizers of the empirical risk (286) and by $\hat{f} \equiv f_{\hat{b},\hat{\boldsymbol{w}}}$ the corresponding trained DAE (284). For future discussion, we also consider training independently the components (285) via ERM, by which we mean

[1] Observe that if $K = 1$, $\left\|\boldsymbol{x}^\mu - f_{b,\boldsymbol{w}}(\tilde{\boldsymbol{x}}^\mu)\right\|^2 = d(1 - b\sqrt{1-\Delta})^2\mathrm{Tr}[\Sigma_1] + d\Delta + {}^1\!/_d\|\boldsymbol{w}\|^2\sigma({}^{\boldsymbol{w}\tilde{\boldsymbol{x}}}\!/\!\sqrt{d})^2 - 2{}^{\boldsymbol{w}\boldsymbol{x}}\!/\!\sqrt{d})\sigma({}^{\boldsymbol{w}\tilde{\boldsymbol{x}}}\!/\!\sqrt{d}) \equiv \ell({}^{\|\boldsymbol{w}\|^2}\!/_d, {}^{X\boldsymbol{w}}\!/\!\sqrt{d})$, which is a slight generalization of the seq-GLM analyzed in the introductory Chapter 2, where the loss can further depend on the weight norm $\|\boldsymbol{w}\|$. We noted the sequence $X \in \mathbb{R}^{2\times d}$ as the concatenation of the rows $\boldsymbol{x}, \boldsymbol{\xi}$.

replacing $f_{b,\boldsymbol{w}}$ by $u_{\boldsymbol{v}}$ or r_c in (286). We similarly denote $\hat{\boldsymbol{v}}$ (resp. $\hat{c}$) the learnt weight of the bottleneck network (resp. rescaling) component and $\hat{u} \equiv u_{\hat{\boldsymbol{v}}}$ (resp. $\hat{r} \equiv r_{\hat{c}}$). Note that generically, $\hat{\boldsymbol{v}} \neq \hat{\boldsymbol{w}}$ and $\hat{c} \neq \hat{b}$, and therefore $\hat{f} \neq \hat{u} + \hat{r}$, since $\hat{b}, \hat{\boldsymbol{w}}$ result from their joint optimization as parts of the full DAE $f_{b,\boldsymbol{w}}$, while $\hat{c}$ (or $\hat{\boldsymbol{v}}$) are optimized independently. As we discuss in Section 10.3, training the sole rescaling r_c does not afford an expressive enough denoiser, while an independently learnt bottleneck network component $u_{\boldsymbol{v}}$ essentially only learns to implement PCA. However, when *jointly* trained as components of the full DAE $f_{b,\boldsymbol{w}}$ (284), the resulting denoiser $\hat{f}$ is a genuinely non-linear model which yields a much lower test error than PCA, and learns to leverage flexibly its two components to balance the preservation of the data nuances and the removal of the noise.

Learning metrics The performance of the DAE (284) trained with the loss (286) is quantified by its reconstruction (denoising) test MSE, defined as

$$\mathrm{mse}_{\hat{f}} \equiv \mathbb{E}_{\mathscr{D}} \mathbb{E}_{\boldsymbol{x} \sim \mathbb{P}} \mathbb{E}_{\boldsymbol{\xi} \sim \mathcal{N}(0, \mathbb{I}_d)} \left\| \boldsymbol{x} - f_{\hat{b}, \hat{\boldsymbol{w}}} \left(\sqrt{1-\Delta} \boldsymbol{x} + \sqrt{\Delta} \boldsymbol{\xi} \right) \right\|^2 . \tag{287}$$

The expectations run over a fresh test sample $\boldsymbol{x}$ sampled from the Gaussian mixture $\mathbb{P}$ (283), and a new additive noise $\boldsymbol{\xi}$ corrupting it. Note that an expectation over the train set $\mathscr{D}$ is also included to make $\mathrm{mse}_{\hat{f}}$ a metric that does not depend on the particular realization of the train set. The denoising test MSEs $\mathrm{mse}_{\hat{u}}, \mathrm{mse}_{\hat{r}}$ are defined similarly as the denoising test errors of the independently learnt components (285). Aside from the denoising MSE (287), another question of interest is how much the DAE manages to learn the structure of the data distribution, as described by the cluster means $\boldsymbol{\mu}_k$. This is measured by the cosine similarity matrix $\boldsymbol{\theta} \in \mathbb{R}^{p \times K}$, where for $i \in [\![1, p]\!]$ and $k \in [\![1, K]\!]$,

$$\theta_{ik} \equiv \mathbb{E}_{\mathscr{D}} \left[\frac{\hat{\boldsymbol{w}}_i^\top \boldsymbol{\mu}_k}{\|\hat{\boldsymbol{w}}_i\| \|\boldsymbol{\mu}_k\|} \right] . \tag{288}$$

In other words, θ_{ik} measures the alignment of the $i-$th row $\hat{\boldsymbol{w}}_i$ of the trained weight matrix $\hat{\boldsymbol{w}}$ with the mean of the $k-$th cluster $\boldsymbol{\mu}_k$.

High-dimensional limit We analyze the optimization problem (286) in the high-dimensional limit where the input dimension d and number of training samples n jointly tend to infinity, while their ratio $\alpha = {}^{n}/_{d}$ stays $\Theta(1)$. The hidden layer width p, the noise level Δ, the number of clusters K and the norm of the cluster means $\|\boldsymbol{\mu}_k\|$ are also assumed to remain $\Theta(1)$. This corresponds to a rich limit, where the number of parameters of the DAE is not large compared to the number of samples like in (Nguyen et al., 2019a; Nguyen, 2021), and therefore cannot trivially fit the train set, or simply memorize it (Radhakrishnan et al., 2019). Conversely, the number of samples n is not infinite like in (Refinetti et al., 2022; Shevchenko et al., 2022; Nguyen et al., 2019b), and therefore importantly allows to study the effect of a finite train set on the representation learnt by the DAE.

10.2 ASYMPTOTIC FORMULAE FOR DAES

We now state the main result of the present work, namely the closed-form asymptotic formulae for the learning metrics $\mathrm{mse}_{\hat{f}}$ (287) and θ (288) for a DAE (284) learnt with the empirical loss (286), derived using the replica method in its RS formulation.

Assumption 10.2.1. *The covariances $\{\boldsymbol{\Sigma}\}_{k=1}^{K}$ admit a common set of eigenvectors $\{\boldsymbol{e}_i\}_{i=1}^{d}$. We further note $\{\lambda_i^k\}_{i=1}^{d}$ the eigenvalues of $\boldsymbol{\Sigma}_k$. The eigenvalues $\{\lambda_i^k\}_{i=1}^{d}$ and the projection of the cluster means on the eigenvectors $\{\boldsymbol{e}_i^\top \boldsymbol{\mu}_k\}_{i,k}$ are assumed to admit a well-defined joint distribution ν as $d \to \infty$ – namely, for $\gamma = (\gamma_1, ..., \gamma_K) \in \mathbb{R}^K$ and $\tau = (\tau_1, ..., \tau_K) \in \mathbb{R}^K$:*

$$\frac{1}{d}\sum_{i=1}^{d}\prod_{k=1}^{K}\delta\left(\lambda_i^k-\gamma_k\right)\delta\left(\sqrt{d}\boldsymbol{e}_i^\top\boldsymbol{\mu}_k-\tau_k\right)\xrightarrow{d\to\infty}\nu\left(\gamma,\tau\right). \tag{289}$$

Moreover, the marginals ν_γ (resp. ν_τ) are assumed to have a well-defined first (resp. second) moment.

Assumption 10.2.2. *$g(\cdot)$ is a ℓ_2 regularizer with strength λ, i.e. $g(\cdot) = \lambda/2\|\cdot\|_F^2$.*

Result 10.2.3. ***(Closed-form asymptotics for DAEs trained with empirical risk minimization)*** *Under Assumptions 10.2.1 and 10.2.2, in the high-dimensional limit $n, d \to \infty$ with fixed ratio α, the denoising test MSE $\mathrm{mse}_{\hat{f}}$ (287) admits the expression*

$$\begin{aligned}&\mathrm{mse}_{\hat{f}}-\mathrm{mse}_\circ=\\&\sum_{k=1}^{K}\rho_k\mathbb{E}_z\,\mathrm{Tr}\left[q\sigma\left(\sqrt{1-\Delta}m_k+\sqrt{\Delta q(1-\Delta)}q_k z\right)^{\otimes 2}\right]\\&-2\sum_{k=1}^{K}\rho_k\mathbb{E}_{u,v}\left[\sigma\left(\sqrt{1-\Delta}m_k+\sqrt{q_k(1-\Delta)}u+\sqrt{\Delta q}v\right)^\top\left((1-\hat{b}\sqrt{1-\Delta})(m_k+\sqrt{q_k}u)-\hat{b}\sqrt{\Delta q}v\right)\right]\\&+o(1),\end{aligned} \tag{290}$$

where the averages bear over independent Gaussian variables $z, u, v \sim \mathcal{N}(0, \mathbb{I}_p)$. We denoted

$$\mathrm{mse}_\circ = d\Delta\hat{b}^2+\left(1-\sqrt{1-\Delta}\hat{b}\right)^2\left[\sum_{k=1}^{K}\rho_k\left(\int d\nu_\tau(\tau)\tau_k^2+d\int d\nu_\gamma(\gamma)\gamma_k\right)\right]. \tag{291}$$

The learnt skip connection strength $\hat{b}$ is

$$\hat{b}=\frac{\left(\sum_{k=1}^{K}\rho_k\int d\nu_\gamma(\gamma)\gamma_k\right)\sqrt{1-\Delta}}{\left(\sum_{k=1}^{K}\rho_k\int d\nu_\gamma(\gamma)\gamma_k\right)(1-\Delta)}+\Delta+o(1). \tag{292}$$

The cosine similarity θ (288) admits the compact formula for $i \in [\![1,p]\!]$ and $k \in [\![1,K]\!]$

$$\theta_{ik} = \frac{(m_k)_i}{\sqrt{q_{ii} \int d\nu_\tau(\tau)\tau_k^2}}, \tag{293}$$

where we have introduced the summary statistics

$$q = \lim_{d\to\infty} \mathbb{E}_{\mathscr{D}}\left[\frac{\hat{\boldsymbol{w}}\hat{\boldsymbol{w}}^\top}{d}\right], \quad q_k = \lim_{d\to\infty} \mathbb{E}_{\mathscr{D}}\left[\frac{\hat{\boldsymbol{w}}\boldsymbol{\Sigma}_k\hat{\boldsymbol{w}}^\top}{d}\right], \quad m_k = \lim_{d\to\infty} \mathbb{E}_{\mathscr{D}}\left[\frac{\hat{\boldsymbol{w}}\boldsymbol{\mu}_k}{\sqrt{d}}\right]. \tag{294}$$

Thus $q, q_k \in \mathbb{R}^{p\times p}$, $m_k \in \mathbb{R}^p$. The summary statistics q, q_k, m_k can be determined as solutions of the system of equations

$$\begin{cases}
\hat{q}_k = \alpha\rho_k \mathbb{E}_{\xi,\eta} V_k^{-1}\left(prox_y^k - q_k^{\frac{1}{2}}\eta - m_k\right)^{\otimes 2} V_k^{-1} \\
\hat{V}_k = -\alpha\rho_k q_k^{-\frac{1}{2}} \mathbb{E}_{\xi,\eta} V_k^{-1}\left(prox_y^k - q_k^{\frac{1}{2}}\eta - m_k\right)\eta^\top \\
\hat{m}_k = \alpha\rho_k \mathbb{E}_{\xi,\eta} V_k^{-1}\left(prox_y^k - q_k^{\frac{1}{2}}\eta - m_k\right) \\
\hat{q} = \frac{\alpha}{\Delta}\sum_{k=1}^K \rho_k \mathbb{E}_{\xi,\eta} V^{-1}\left(prox_x^k - \sqrt{\Delta}q^{\frac{1}{2}}\xi\right)^{\otimes 2} V^{-1} \\
\hat{V} = -\alpha\sum_{k=1}^K \rho_k \mathbb{E}_{\xi,\eta}\left[\frac{1}{\sqrt{\Delta}}V^{-1}\left(prox_x^k - \sqrt{\Delta}q^{\frac{1}{2}}\xi\right)\xi^\top q^{-\frac{1}{2}} - \sigma\left(\sqrt{1-\Delta}prox_y^k + prox_x^k\right)^{\otimes 2}\right]
\end{cases}$$

$$\begin{cases}
q_k = \int d\nu(\gamma,\tau)\gamma_k\left(\lambda\mathbb{I}_p + \hat{V} + \sum_{j=1}^K \gamma_j\hat{V}_j\right)^{-1} \\
\qquad \left(\hat{q} + \sum_{j=1}^K \gamma_j\hat{q}_j + \sum_{1\le j,l\le K}\tau_j\tau_l\hat{m}_j\hat{m}_l^\top\right)\left(\lambda\mathbb{I}_p + \hat{V} + \sum_{j=1}^K \gamma_j\hat{V}_j\right)^{-1} \\
V_k = \int d\nu(\gamma,\tau)\gamma_k\left(\lambda\mathbb{I}_p + \hat{V} + \sum_{j=1}^K \gamma_j\hat{V}_j\right)^{-1} \\
m_k = \int d\nu(\gamma,\tau)\tau_k\left(\lambda\mathbb{I}_p + \hat{V} + \sum_{j=1}^K \gamma_j\hat{V}_j\right)^{-1}\sum_{j=1}^K \tau_j\hat{m}_j \\
q = \int d\nu(\gamma,\tau)\left(\lambda\mathbb{I}_p + \hat{V} + \sum_{j=1}^K \gamma_j\hat{V}_j\right)^{-2} \\
\qquad \left(\hat{q} + \sum_{j=1}^K \gamma_j\hat{q}_j + \sum_{1\le j,l\le K}\tau_j\tau_l\hat{m}_j\hat{m}_l^\top\right) \\
V = \int d\nu(\gamma,\tau)\left(\lambda\mathbb{I}_p + \hat{V} + \sum_{j=1}^K \gamma_j\hat{V}_j\right)^{-1}
\end{cases} \tag{295}$$

In (295)*,* $\hat{q}_k, \hat{V}_k, \hat{V}, V \in \mathbb{R}^{p\times p}$ *and* $\hat{m}_k \in \mathbb{R}^p$*, and the averages bear over finite-dimensional i.i.d Gaussians* $\xi, \eta \sim \mathscr{N}(0, \mathbb{I}_p)$*. Finally,* $prox_x^k, prox_y^k$ *are given as the solutions of the finite-dimensional optimization*

$$\begin{aligned}
&prox_x^k,\ prox_y^k = \\
&\underset{x,y\in\mathbb{R}^p}{\mathrm{arginf}}\left\{ \mathrm{Tr}\left[V_k^{-1}\left(y - q_k^{\frac{1}{2}}\eta - m_k\right)^{\otimes 2}\right] + \frac{1}{\Delta}\mathrm{Tr}\left[V^{-1}\left(x - \sqrt{\Delta}q^{\frac{1}{2}}\xi\right)^{\otimes 2}\right]\right. \\
&\left. + \mathrm{Tr}\left[q\sigma(\sqrt{1-\Delta}y + x)^{\otimes 2}\right] - 2\sigma(\sqrt{1-\Delta}y + x)^\top((1-\sqrt{1-\Delta}\hat{b})y - \hat{b}x)\right\}.
\end{aligned} \tag{296}$$

In fact, Assumptions 10.2.1 and 10.2.2 are not strictly necessary, and can be simultaneously relaxed to address arbitrary convex regularizer $g(\cdot)$ and generically non-commuting $\{\boldsymbol{\Sigma}_k\}_{k=1}^K$ – but at the price of more intricate formulae. For this reason, we choose to discuss here Result 10.2.3, following the same lines as the computation detailed in the introductory Part I. Result 10.2.3 encompasses as special cases the asymptotic characterization of the components $\hat{r}, \hat{u}$ (285):

Corollary 10.2.4. *(**MSE of components**) The test MSE of* $\hat{r}$ (285) *is given by* $\mathrm{mse}_{\hat{r}} = \mathrm{mse}_\circ$ (291)*. Furthermore, the learnt value of its single parameter* $\hat{c}$ *is given by* (292)*. The test MSE, cosine similarity and summary statistics of the bottleneck network* $\hat{u}$ (285) *follow from Result 10.2.3 by setting* $\hat{b} = 0$*.*

The implications of Corollary 10.2.4 shall be further discussed in Section 10.3. Finally, remark that in the noiseless limit $\Delta = 0$, the denoising task reduces to a reconstruction task, with the autoencoder being tasked with reproducing the clean data as an output when taking the same clean sample as an input. Therefore Result 10.2.3 also includes RAEs (by definition, without skip connection) as a special case.

Corollary 10.2.5. *(**RAEs**) In the* $n, d \to \infty$ *limit, the MSE, cosine similarity and summary statistics for an RAE follow from Result 10.2.3 by setting* $x = 0$ *in* (296)*, removing the first term in the brackets in the equation of* $\hat{V}$ (295) *and taking the limit* $\Delta, \hat{q}, \hat{b} \to 0$*.*

Corollary 10.2.5 will be the object of further discussion in Section 10.3. Note that Corollary 10.2.5 provides a characterization of RAEs as a function of the sample complexity α, where previous studies on non-linear RAEs rely on the assumption of an infinite number of available training samples (Nguyen, 2021; Refinetti et al., 2022; Shevchenko et al., 2022).

Equations (292) and (294) of Result 10.2.3 thus characterize the statistics of the learnt parameters $\hat{b}, \hat{\boldsymbol{w}}$ of the trained DAE (284). These summary statistics are, in turn, sufficient to fully characterize the learning metrics (287) and (288) via equations (290) and (293). We thus have reduced the high-dimensional optimization (286) and the high-dimensional average over the train set $\mathscr{D}$

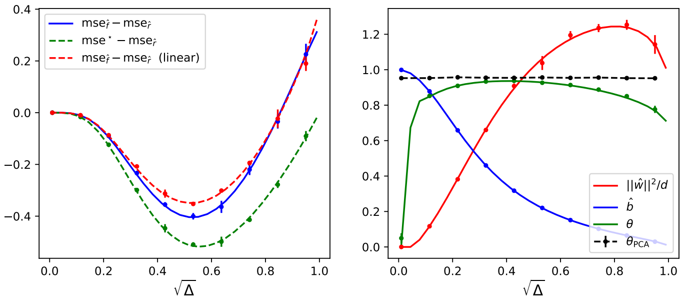


Figure 33: $\alpha = 1, K = 2, \rho_{1,2} = 1/2, \mathbf{\Sigma}_{1,2} = 0.09 \times \mathbb{I}_d, p = 1, \lambda = 0.1, \sigma(\cdot) = \tanh(\cdot)$; the cluster mean $\boldsymbol{\mu}_1 = -\boldsymbol{\mu}_2$ was taken as a random Gaussian vector of norm 1. (left) In blue, the difference in MSE between the full DAE $\hat f$ (284) and the rescaling component $\hat r$ (285). Solid lines correspond to the sharp asymptotic characterization of Result 10.2.3. Dots represent numerical simulations for $d = 700$, training the DAE using the `Pytorch` implementation of full-batch Adam, with learning rate $\eta = 0.05$ over 2000 epochs, averaged over $N = 10$ instances. Error bars represent one standard deviation. For completeness, the MSE of the oracle denoiser is given as a baseline in green, see Section 10.3. The performance of a linear DAE ($\sigma(x) = x$) is represented in dashed red. (right) Cosine similarity θ (288) (green), squared weight norm $\|\hat{\boldsymbol{w}}\|_F^2/d$ (red) and skip connection strength $\hat b$ (blue). Solid lines correspond to the formulae (293)(294) and (292) of Result 10.2.3; dots are numerical simulations. For completeness, the cosine similarity of the first principal component of the clean train data $\{\boldsymbol{x}^\mu\}_{\mu=1}^n$ is plotted in dashed black.

involved in the definition of the metrics (287) and (288) to a simpler system of equations over $4 + 6K$ variables (295) which can be solved numerically. It is important to note that all the summary statistics involved in (295) are *finite-dimensional* as $d \to \infty$, and therefore Result 10.2.3 is a fully asymptotic characterization, in the sense that it does not involve any high-dimensional object. In the next paragraphs, we give two examples of applications of Result 10.2.3, to a simple binary isotropic mixture, and to real data sets.

Example 1: Isotropic homoscedastic mixture We give as a first example the case of a synthetic binary Gaussian mixture with $K = 2, \boldsymbol{\mu}_1 = -\boldsymbol{\mu}_2, \mathbf{\Sigma}_{1,2} = 0.09 \times \mathbb{I}_d, \rho_{1,2} = 1/2$, using a DAE with $\sigma = \tanh$ and $p = 1$. Since this simple case exhibits the key phenomenology discussed in the present work, we refer to it in future discussions. The MSE $\mathrm{mse}_{\hat f}$ (290) evaluated from the solutions of the self-consistent equations (295) is plotted as the solid blue line in Fig. 33 (left) and compared to numerical simulations corresponding to training the DAE (284) with the `Pytorch` implementation of the Adam optimizer (Kingma et al., 2014b) (blue dots), for sample complexity $\alpha = 1$ and ℓ_2 regularization (weight decay) $\lambda = 0.1$. The agreement between the theory and simulation is compelling. The green solid line and corresponding green dots in Fig. 33 (right) correspond to the replica prediction (293) and simula-

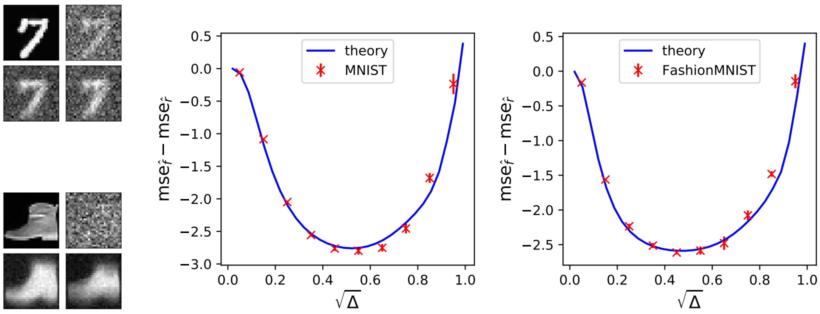


Figure 34: Difference in MSE between the full DAE (284) and the rescaling component (285) for the MNIST data set (middle), of which for simplicity only 1s and 7s were kept, and FashionMNIST (right), of which only boots and shoes were kept. In blue, the theoretical predictions resulting from using Result 10.2.3 with the empirically estimated covariances and means. In red, numerical simulations of a DAE ($p = 1$, $\sigma = \tanh$) trained with $n = 784$ training points, using the `Pytorch` implementation of full-batch Adam, with learning rate $\eta = 0.05$ and weight decay $\lambda = 0.1$ over 2000 epochs, averaged over $N = 10$ instances. Error bars represent one standard deviation. (left) illustration of the denoised images: (top left) original image, (top right) noisy image, (bottom left) DAE $\hat{f}$ (284), (bottom right) rescaling $\hat{r}$ (285).

tions for the cosine similarity θ (288), and again display very good agreement.

A particularly striking observation is that due to the non-convexity of the loss (286), there is a priori no guarantee that an Adam-optimized DAE should find a global minimum, as described by the Result 10.2.3, rather than a local minimum. The compelling agreement between theory and simulations in Fig. 33 temptingly suggests that the loss landscape of DAEs (284) trained with the loss (286) for the data model (283) should in some way be benign. Authors of (Baldi et al., 1989) have shown, for *linear RAEs*, that there exists a unique global and local minimum for the square loss and no regularizer. Ref. (Pretorius et al., 2018) offers further insight for a linear DAE in dimension $d = 1$, and shows that, aside from the global minima, the loss landscape only includes an unstable saddle point from which the dynamics easily escapes. Extending these works and intuitions to non-linear DAEs is an exciting research topic for future work.

Example 2: MNIST, FashionMNIST It is reasonable to ask whether Result 10.2.3 is restricted to Gaussian mixtures (283). The answer is negative – in fact, Result 10.2.3 also describes well a number of real data distributions. We provide such an example for FashionMNIST (Xiao et al., 2017) (from which, for simplicity, we only kept boots and shoes) and MNIST (LeCun et al., 1998b) (1s and 7s), in Fig. 34. For each data set, samples sharing the same label were considered to belong to the same cluster. The mean and covariance thereof were estimated numerically, and combined with Result 10.2.3. The resulting denoising MSE predictions $\mathrm{mse}_{\hat{f}}$ are plotted as solid lines in Fig. 34, and agree very well with numerical simulations of DAEs optimized over the real data sets using the `Pytorch` implementation of Adam (Kingma et al., 2014b).

The observation that the MSEs of real data sets are to such degree of accuracy captured by the equivalent Gaussian mixture strongly hints at the presence of Gaussian universality (Goldt et al., 2020b). This opens a gateway to future research, as Gaussian universality has hitherto been exclusively addressed in classification and regression (rather than denoising) settings, see e.g. (Goldt et al., 2020b; Hu et al., 2020; Montanari et al., 2022b). Denoising tasks further constitute a particularly intriguing setting for universality results, as Gaussian universality would signify that only second-order statistics of the data can be reconstructed using a shallow autoencoder.

10.3 THE ROLE AND IMPORTANCE OF THE SKIP CONNECTION.

Result 10.2.3 for the full DAE $\hat{f}$ (284) and Corollary 10.2.4 for its components $\hat{r}$, $\hat{u}$ (285) allow to disentangle the contribution of each part, and thus to pinpoint their respective roles in the DAE architecture. We sequentially present a comparison of $\hat{f}$ with $\hat{r}$, and $\hat{f}$ with $\hat{u}$. We remind that $\hat{f}$, $\hat{r}$ and $\hat{u}$ result from *independent* optimizations over the same train set $\mathscr{D}$, and that while $f_{b,\mathbf{w}} = u_{\mathbf{w}} + h_b$, $\hat{f} \neq \hat{u} + \hat{r}$.

Full DAE and the rescaling component We start this section by observing that for noise levels Δ below a certain threshold, the full DAE $\hat{f}$ yields better MSE than the learnt rescaling $\hat{r}$, as can be seen by the negative value of $\mathrm{mse}_{\hat{f}} - \mathrm{mse}_{\hat{r}}$ in Fig. 33 and Fig. 34. The improvement is more sizeable at intermediate noise levels Δ, and is observed for a growing region of Δ as the sample complexity α increases, see Fig. 35 (a). This lower MSE further translates into visible qualitative changes in the result of denoising. As can be seen from Fig. 34 (left), the full DAE $\hat{f}$ (284) (bottom left) yields denoised images with sensibly higher definition and overall contrast, while a simple rescaling $\hat{r}$ (bottom right) leads to a still largely blurred image.

In its simplest formulation, Tweedie's *formula states that when $\tilde{\mathbf{x}} \sim \mathcal{N}(\mathbf{x}, \mathbb{I}_d)$, $\mathbb{E}[\mathbf{x}|\tilde{\mathbf{x}}] = \tilde{\mathbf{x}} + s(\tilde{\mathbf{x}})$, where the score corresponds to $s(\cdot) = {}^{d}\!/_{dz} \ln P$, with P the probability density of $\tilde{x}$. Similar results are at the heart of denoising-driven generative models, see also Chapter 11.*

We provide one more comparison: for the isotropic binary mixture (see Fig. 33), the DAE test error $\mathrm{mse}_{\hat{f}}$ in fact approaches the information-theoretic lowest achievable MSE $\mathrm{mse}^{\star}$ as the sample complexity α increases. To see this, note that $\mathrm{mse}^{\star}$ is given by the application of Tweedie's formula (Efron, 2011), that requires perfect knowledge of the cluster means $\boldsymbol{\mu}_k$ and covariances $\boldsymbol{\Sigma}_k$ – it is, therefore, an *oracle* denoiser.

As can be observed from Fig. 35 (a), the DAE MSE (284) approaches the oracle test error $\mathrm{mse}^{\star}$ as the number of available training samples n grows,

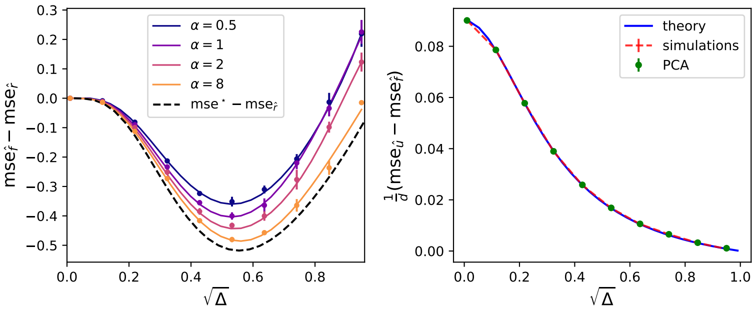
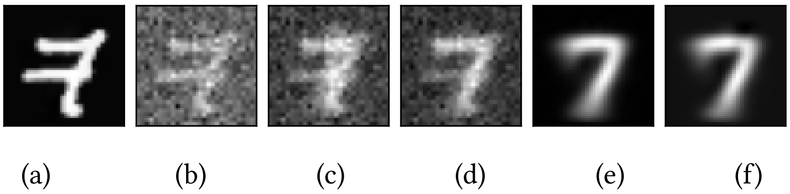


Figure 35: (left) Solid lines: difference in MSE between the full DAE $\hat{f}$ (284), with $\sigma = \tanh$, $p = 1$, and the rescaling $\hat{r}$ (285). Dashed: the same curve for the oracle denoiser. Different colours represent different sample complexities α (solid lines). (right) Difference in MSE between the bottleneck network $\hat{u}$ (285) and the complete DAE $\hat{f}$ (284). In blue, the theoretical prediction (297); in red, numerical simulations for the bottleneck network (285) ($\sigma = \tanh$, $p = 1$) trained with the `Pytorch` implementation of full-batch Adam, with learning rate $\eta = 0.05$ and weight decay $\lambda = 0.1$ over 2000 epochs, averaged over $N = 5$ instances, for $d = 700$. In green, the MSE (minus the MSE of the complete DAE (284)) achieved by PCA. Error bars represent one standard deviation. The model and parameters are the same as in Fig. 33.



(a) (b) (c) (d) (e) (f)

Figure 36: Illustration of the denoised image for the various networks and algorithms. (a) original image (b) noisy image, for $\sqrt{\Delta} = 0.2$ (c) trained rescaling $\hat{r}$ (285) (d) full DAE $\hat{f}$ (284) (e) bottleneck network $\hat{u}$ (285) (f) PCA. The DAE and training parameters are the same as Fig. 34.

and is already sensibly close to the optimal value for $\alpha = 8$.

DAEs with(out) skip connection We now turn our attention to comparing the full DAE $\hat{f}$ (284) to the bottleneck network component $\hat{u}$ (285). It follows from Result 10.2.3 and Corollary 10.2.4 that $\hat{u}$ (285) leads to a higher MSE than the full DAE $\hat{f}$ (284), with the gap being $\Theta(d)$. More precisely,

$$\frac{1}{d}\left(\mathrm{mse}_{\hat{u}} - \mathrm{mse}_{\hat{f}}\right) = \frac{\left(\int d\nu_\gamma(\gamma) \sum_{k=1}^{K} \rho_k \gamma_k\right)^2 (1-\Delta)}{\left(\int d\nu_\gamma(\gamma) \sum_{k=1}^{K} \rho_k \gamma_k\right)(1-\Delta)} + \Delta. \qquad (297)$$

The theoretical prediction (297) compares excellently with numerical simulations; see Fig. 35 (right). Strikingly, we find that PCA denoising yields an MSE almost indistinguishable from $\hat{u}$, see Fig. 35, strongly suggesting that $\hat{u}$

essentially learns, also in the denoising setting, to project the noisy data $\tilde{\boldsymbol{x}}$ along the principal components of the training set. The last two images of Fig. 36 respectively correspond to $\hat{u}$ and PCA, which can indeed be observed to lead to visually near-identical results. This echoes the findings of (Eckart et al., 1936; Bourlard et al., 1988; Refinetti et al., 2022; Shevchenko et al., 2022; Nguyen, 2021) in the case of RAEs that bottleneck networks are limited by the PCA reconstruction performance – a conclusion that we can also recover from Corollary 10.2.5. Crucially however, it *also* means that compared to the *full* DAE $\hat{f}$ (284), *PCA is sizeably suboptimal*, since $\mathrm{mse}_{\mathrm{PCA}} \approx \mathrm{mse}_{\hat{u}} = \mathrm{mse}_{\hat{f}} + \Theta(d)$.

This last observation has an important consequence: in contrast to previously studied RAEs (Eckart et al., 1936; Baldi et al., 1989; Bourlard et al., 1988; Shevchenko et al., 2022; Nguyen, 2021), the full DAE $\hat{f}$ does *not* simply learn to perform PCA. In contrast to bottleneck RAE networks (Refinetti et al., 2022; Shevchenko et al., 2022; Nguyen, 2021), the non-linear DAE hence does not reduce to a linear model after training. The non-linearity is important to improve the denoising MSE, see Fig. 33. We stress this finding: trained alone, the bottleneck network $\hat{u}$ only learns to perform PCA; trained jointly with the rescaling component as part of the full DAE $f_{b,\boldsymbol{w}}$ (284), it learns a richer, non-linear representation. The full DAE (284) thus offers a genuinely non-linear learning model and opens exciting research avenues for the theory of autoencoders, beyond linear (or effectively linear) cases. In the next paragraph, we explore further the interaction between the rescaling component and the bottleneck network.

A tradeoff between the rescaling and the bottleneck network Result (10.2.3), alongside Corollary (10.2.4) and the discussion in Section 10.3 provide a firm theoretical basis for the well-known empirical intuition (discussed e.g. in (Mao et al., 2016)) that skip-connections allow to better propagate information from the input to the output of the DAE, thereby contributing to preserving intrinsic characteristics of the input. This effect is clearly illustrated in Fig. 36, where the resulting denoised image of an MNIST 7 by $\hat{r}$, $\hat{f}$, $\hat{u}$ and PCA are presented. While the bottleneck network $\hat{u}$ perfectly eliminates the background noise and produces an image with a very good resolution, it essentially collapses the image to the cluster mean, and yields, like PCA, the average MNIST 7. As a consequence, the denoised image bears little resemblance with the original image – in particular, the horizontal bar of the 7 is lost in the process. Conversely, the rescaling $\hat{r}$ preserves the nuances of the original image, but the result is still largely blurred and displays overall poor contrast. Finally, the complete DAE (284) manages to preserve the characteristic features of the original data, while enhancing the image resolution by slightly overlaying the average 7 thereupon.

The optimization of the DAE (286) is therefore described by a tradeoff between two competing effects – namely the preservation of the input nuances by the skip connection, and the enhancement of the resolution/noise removal by the bottleneck network. This allows us to discuss the curious non-monotonicity of the cosine similarity θ as a function of the noise level Δ, see Fig. 33 (left). While it may at first seem curious that the DAE seemingly does not manage to learn the data structure better for low Δ than for intermediate Δ (where the cosine similarity θ is observed to be higher), this is actually due to the afore-dicussed tradeoff. Indeed, for small Δ, the data is still substantially clean, and there is therefore no incentive to enhance the contrast by using the cluster means –which are consequently not learnt. This phase is thus characterized by a large skip connection strength $\hat{b}$, and small cosine similarity θ and weight norm $||\hat{\boldsymbol{w}}||_F$. Conversely, at high noise levels Δ, the nuances of the data are already lost because of the noise. Hence the DAE does not rely on the skip connection component (whence the small values of $\hat{b}$), and the only way to produce reasonably denoised data is to collapse to the cluster mean using the network component (whence a large $||\hat{\boldsymbol{w}}||_F$).

CONCLUSION

We consider the problem of denoising a high-dimensional Gaussian mixture, by training a DAE via empirical risk minimization, in the limit where the number of training samples and the dimension are proportionally large. We provide a sharp asymptotic characterization of a number of summary statistics of the trained DAE weight, average MSE, and cosine similarity with the cluster means. These results contain as a corollary the case of RAEs. Building on these findings, we isolate the role of the skip connection and the bottleneck network in the DAE architecture and characterize the tradeoff between those two components in terms of preservation of the data nuances and noise removal – thereby providing some theoretical insight into a longstanding practical intuition in machine learning.

We believe the present work also opens exciting research avenues. First, our real data experiments hint at the presence of Gaussian universality. While this topic has gathered considerable attention in recent years, only classification/regression supervised learning tasks have been hitherto addressed. Which aspects of universality carry over to denoising tasks, and how they differ from the current understanding of supervised regression/classification, is an important question. Second, the DAE with skip connection (284) provides an autoencoder model which does not just simply learn the principal components of the training set. It, therefore, affords a genuinely non-linear network model where richer learning settings can be investigated.

11 FLOW-BASED GENERATIVE MODELS

Flow and diffusion-based generative models have introduced a shift in paradigm for density estimation and sampling problems, leading to state-of-the art algorithms e.g. in image generation (Rombach et al., 2022; Ramesh et al., 2022; Saharia et al., 2022). Instrumental in these advances was the realization that the sampling problem could be recast as a transport process from a simple –typically Gaussian– base distribution to the target density. Furthermore, the velocity field governing the flow can be characterized as the minimizer of a quadratic loss function, which can be estimated from data by (a) approximating the loss by its empirical estimate using available training data and (b) parametrizing the velocity field using a denoiser neural network. These ideas have been fruitfully implemented as part of a number of frameworks, including score-based diffusion models (Song et al., 2019; Song et al., 2020; Karras et al., 2022; Ho et al., 2020b), and stochastic interpolation (Albergo et al., 2022; Albergo et al., 2023; Lipman et al., 2022; Liu et al., 2022). A tight analytical understanding of the learning of generative models from limited data, and the resulting generative process, is however still largely missing. This constitutes the research question addressed in the present manuscript.

A line of recent analytical works (Benton et al., 2023; Chen et al., 2022a; Chen et al., 2023a; Chen et al., 2023c; Chen et al., 2023d; Wibisono et al., 2022; Lee et al., 2022; Lee et al., 2023; Li et al., 2023a; De Bortoli et al., 2021; De Bortoli, 2022; Pidstrigach, 2022; Block et al., 2020) have mainly focused on the study of the transport problem, and provide rigorous convergence guarantees, taking as a starting point the assumption of an L^2-accurate estimate of the velocity or score. They hence bypass the investigation of the learning problem –and in particular the question of ascertaining the sample complexity needed to obtain such an accurate estimate. More importantly, the study of the effect of learning from a *limited* sample complexity (and thus e.g. of possible network overfitting and memorization) on the generated density, is furthermore left unaddressed. On the other hand, very recent works (Cui et al., 2024b; Shah et al., 2023) have characterized the learning of DAEs (Vincent et al., 2010; Vincent, 2011) in high dimensions on Gaussian mixture densities, see Chapter 10. Neither work however studies the consequences on the generative process. Bridging that gap, recent works have offered a *joint* analysis of the learning and generative processes. (Oko et al., 2023; Chen et al., 2023b; Yuan et al., 2023) derive rigorous bounds at finite sample complexity, under the assumption of data with a *low-dimensional* structure. Closer to our manuscript, a concurrent work (Mei et al., 2023) bounds the Kullback-Leibler distance between the generated and target densities, when parametrizing the flow using a ResNet, for high-dimensional graphical models. On the other hand, these bounds do not go to zero as the sample complexity increases, and are a priori not tight.

The present manuscript aims at complementing and furthering this last body of works, by providing a tight end-to-end analysis of a flow-based generative model – starting from the study of the high-dimensional learning problem with a finite number of samples, and subsequently elucidating the implications thereof on the generative process.

Main contributions– We study the problem of estimating and sampling a Gaussian mixture using a flow-based generative model, in the framework of stochastic interpolation (Albergo et al., 2022; Albergo et al., 2023; Lipman et al., 2022; Liu et al., 2022). We consider the case where a non-linear two-layer DAE with one hidden unit is used to parametrize the velocity field of the associated flow, and is trained with a finite training set. In the high-dimensional limit,

- We provide a sharp asymptotic closed-form characterization of the learnt velocity field, as a function of the target Gaussian mixture parameters, the stochastic interpolation schedule, and the number of training samples n.
- We characterize the associated flow by providing a tight characterization of a small number of summary statistics, tracking the dynamics of a sample from the Gaussian base distribution as it is transported by the learnt velocity field.
- We show that even with a finite number of training samples, the learnt generative model allows to sample from a mixture whose mean asymptotically approaches the mean of the target mixture as $\Theta_n(1/n)$ in squared distance, with this rate being tight.
- Finally, we show that this rate is in fact Bayes-optimal.

RELATED WORKS

Diffusion and flow-based generative models Score-based diffusion models (Song et al., 2019; Song et al., 2020; Karras et al., 2022; Ho et al., 2020b) build on the idea that any density can be mapped to a Gaussian density by degrading samples through an Ornstein-Uhlenbeck process. Sampling from the original density can then be carried out by time-reversing the corresponding stochastic transport, provided the score is known – or estimated. These ideas were subsequently refined in (Albergo et al., 2022; Albergo et al., 2023; Lipman et al., 2022; Liu et al., 2022), which provide a flexible framework to bridge between two arbitrary densities in finite time.

Convergence bounds In the wake of the practical successes of flow and diffusion-based generative models, significant theoretical effort has been devoted to studying the convergence of such methods, by bounding appropriate distances between the generated and the target densities. A common assumption of (Benton et al., 2023; Chen et al., 2022a; Chen et al., 2023a; Chen et al., 2023c; Chen et al., 2023d; Wibisono et al., 2022; Lee et al., 2022; Lee et al., 2023; Li et al., 2023a; De Bortoli et al., 2021; De Bortoli, 2022; Pidstrigach, 2022; Block et al., 2020) is the availability of a good estimate for the score, i.e. an estimate whose average (population) squared distance with the true score is bounded by a small constant ε. Under this assumption, Chen et al. (2022a) and Lee et al. (2022) obtain rigorous control on the Wasserstein and total variation distances with very mild assumptions on the target density. (Ghio et al., 2023) explore the connections between algorithmic hardness of the score/flow approximation and the hardness of sampling in a number of graphical models.

Asymptotics for DAE learning The backbone of flow and diffusion-based generative models is the parametrization of the score or velocity by a denoiser-type network, whose most standard realization is arguably the DAE (Vincent et al., 2010; Vincent, 2011). Very recent works have provided a detailed analysis of its learning on denoising tasks, for data sampled from Gaussian mixtures. (Cui et al., 2024b) sharply characterize how a DAE can learn the mixture parameters with $n = \Theta_d(d)$ training samples when the cluster separation is $\Theta_d(1)$. Closer to our work, for arbitrary cluster separation, Shah et al. (2023) rigorously show that a DAE trained with gradient descent on the denoising diffusion probabilistic model loss (Ho et al., 2020b) can recover the cluster means with a polynomial number of samples. While these works complement the aforediscussed convergence studies in that they analyze the effect of a finite number of samples, neither explores the flow associated to the learnt score.

Network-parametrized models Tying together these two body of works, a very recent line of research has addressed the problem of bounding, at finite sample complexity, appropriate distances between the generated and target densities, assuming a network-based parametrization. (Oko et al., 2023) provide such bounds when parametrizing the score using a class of ReLU networks. These bounds however suffer from the curse of dimensionality. (Oko et al., 2023; Yuan et al., 2023; Chen et al., 2023b) surmount this hurdle by assuming a target density with low-dimensional structure. On a heuristic level, (Biroli et al., 2023) estimate the order of magnitude of the sample complexity needed to sample from a high-dimensional Curie-Weiss model. Finally, a work concurrent to ours (Mei et al., 2023) derives rigorous bounds for a number of high-dimensional graphical models. On the other hand, these bounds are a priori not tight, and do not go to zero as the sample complexity becomes large. The present manuscript aims at furthering this line of work, and provides a *sharp* analysis of a high-dimensional flow-based generative model.

11.1 SETTING

We start by giving a concise overview of the problem of sampling from a target density ρ_1 over $\mathbb{R}^d$ in the framework of stochastic interpolation (Albergo et al., 2022; Albergo et al., 2023).

Recasting sampling as an optimization problem Samples from ρ_1 can be generated by drawing a sample from an easy-to-sample base density ρ_0 –henceforth taken to be a standard Gaussian density $\rho_0 = \mathcal{N}(0, \mathbb{I}_d)$–, and evolving it according to the flow described by the Ordinary Differential Equation (ODE)

$$\frac{d}{dt}\boldsymbol{X}_t = \boldsymbol{b}(\boldsymbol{X}_t, t), \tag{298}$$

for $t \in [0,1]$. Specifically, as shown in (Albergo et al., 2023), if $\boldsymbol{X}_{t=0} \sim \rho_0$, then the final sample $\boldsymbol{X}_{t=1}$ has probability density ρ_1, if the velocity field $\boldsymbol{b}(\boldsymbol{x}, t)$ governing the flow (298) is given by

$$\boldsymbol{b}(\boldsymbol{x}, t) = \mathbb{E}[\dot{\alpha}(t)\boldsymbol{x}_0 + \dot{\beta}(t)\boldsymbol{x}_1 | \boldsymbol{x}_t = \boldsymbol{x}], \tag{299}$$

where we denoted $\boldsymbol{x}_t \equiv \alpha(t)\boldsymbol{x}_0 + \beta(t)\boldsymbol{x}_1$ and the conditional expectation bears over $\boldsymbol{x}_1 \sim \rho_1$, $\boldsymbol{x}_0 \sim \rho_0$, with $\boldsymbol{x}_0 \perp \boldsymbol{x}_1$. The result holds for any fixed choice of schedule functions $\alpha, \beta \in \mathscr{C}^2([0,1])$ satisfying $\alpha(0) = \beta(1) = 1, \alpha(1) = \beta(0) = 0$, and $\alpha(t)^2 + \beta(t)^2 > 0$ for all $t \in [0,1]$. In addition to the velocity field $\boldsymbol{b}(\boldsymbol{x}, t)$, it is convenient to consider the field $\boldsymbol{f}(\boldsymbol{x}, t)$, related to $\boldsymbol{b}(\boldsymbol{x}, t)$ by the simple relation

$$\boldsymbol{b}(\boldsymbol{x}, t) = \left(\dot{\beta}(t) - \frac{\dot{\alpha}(t)}{\alpha(t)}\beta(t)\right)\boldsymbol{f}(\boldsymbol{x}, t) + \frac{\dot{\alpha}(t)}{\alpha(t)}\boldsymbol{x}. \tag{300}$$

Note that $\boldsymbol{f}(\boldsymbol{x}, t)$ can be alternatively expressed as $\mathbb{E}[\boldsymbol{x}_1 | \boldsymbol{x}_t = \boldsymbol{x}]$, and thus admits a natural interpretation as a *denoising* function, tasked with recovering the target value $\boldsymbol{x}_1$ from the interpolated (noisy) sample $\boldsymbol{x}_t$. The denoiser $\boldsymbol{f}(\boldsymbol{x}, t)$ can furthermore characterized as the minimizer of the objective

$$\mathscr{R}[\boldsymbol{f}] = \int_0^1 \mathbb{E}\, \|\boldsymbol{f}(\boldsymbol{x}_t, t) - \boldsymbol{x}_1\|^2 \, dt. \tag{301}$$

The loss (301) is a simple sequence of quadractic *denoising* objectives.

Learning the velocity from data There are several technical hurdles in carrying out the minimization (301). First, since the analytical form of ρ_1 is generically unknown, the population risk has to be approximated by its empirical version, provided a dataset $\mathscr{D} = \{\boldsymbol{x}_1^\mu, \boldsymbol{x}_0^\mu\}_{\mu=1}^n$ of n training samples $\boldsymbol{x}_1^\mu$ ($\boldsymbol{x}_0^\mu$) independently drawn from ρ_1 (ρ_0) is available. Second, the minimization in (301) bears over a time-dependent vector field $\boldsymbol{f}$. To make the optimization tractable, the latter can be parametrized at each time step t by a separate neural network $\boldsymbol{f}_{\theta_t}(\cdot)$ with trainable parameters θ_t. Under those approximations, the population risk (301) thus becomes

$$\hat{\mathscr{R}}(\{\theta_t\}_{t\in[0,1]}) = \int_0^1 \sum_{\mu=1}^{n} \left\| \boldsymbol{f}_{\theta_t}(\boldsymbol{x}_t^\mu) - \boldsymbol{x}_1^\mu \right\|^2 dt. \tag{302}$$

Remark that in practice, the time t can enter as an input of the neural network, and only one network then needs to be trained. In the present manuscript however, for technical reasons, we instead consider the case where a *separate* network is trained *for each time step* t. Besides, note that since the base density ρ_0 is a priori easy to sample from, one could in theory augment the dataset $\mathscr{D}$ with several samples from ρ_0 for each available $\boldsymbol{x}_1^\mu$. For conciseness, we do not examine such an augmentation technique in the present manuscript, and leave a precise investigation thereof to future work. Denoting by $\{\hat{\theta}_t\}_{t\in[0,1]}$ the minimizer of (302), the learnt velocity field $\hat{\boldsymbol{b}}$ is related to the trained denoiser $\boldsymbol{f}_{\hat{\theta}_t}$ by (301) as

$$\hat{\boldsymbol{b}}(\boldsymbol{x},t) = \left(\dot{\beta}(t) - \frac{\dot{\alpha}(t)}{\alpha(t)}\beta(t)\right) \boldsymbol{f}_{\hat{\theta}_t}(\boldsymbol{x}) + \frac{\dot{\alpha}(t)}{\alpha(t)}\boldsymbol{x}. \tag{303}$$

The sampling can finally be carried out by using $\hat{\boldsymbol{b}}$ as a proxy for the unknown $\boldsymbol{b}$ in (298):

$$\frac{d}{dt}\boldsymbol{X}_t = \hat{\boldsymbol{b}}(\boldsymbol{X}_t,t) \tag{304}$$

Note that the solution $\boldsymbol{X}_1$ at time $t=1$ of the ODE (304) has a law $\hat{\rho}_1 \neq \rho_1$ due to the two approximations in going from the population function-space objective (301) to the empirical parametric proxy (302). The present manuscript presents a sharp analysis of the learning problem (302) and the resulting flow (304) for a solvable model, which we detail below.

Data model We consider the case of a target density ρ_1 given by a binary isotropic and homoscedastic Gaussian mixture

$$\rho_1 = \frac{1}{2}\mathscr{N}(\boldsymbol{\mu}, \sigma^2\mathbb{I}_d) + \frac{1}{2}\mathscr{N}(-\boldsymbol{\mu}, \sigma^2\mathbb{I}_d). \tag{305}$$

Each cluster is thus centered around its mean $\pm\boldsymbol{\mu}$ and has variance σ^2. For definiteness, we consider here a balanced mixture, where the two clusters have equal relative probabilities. Note that a sample $\boldsymbol{x}_1^\mu$ can then be decomposed as $\boldsymbol{x}_1^\mu = s^\mu\boldsymbol{\mu} + \boldsymbol{z}^\mu$, with $s^\mu \sim \mathscr{U}(\{-1,+1\})$ and $\boldsymbol{z}^\mu \sim \mathscr{N}(0,\sigma^2\mathbb{I}_d)$. Finally, note that the closed-form expression for the exact velocity field $\boldsymbol{b}$ (298) associated to the density ρ_1 is actually known (see e.g. (Efron, 2011; Albergo et al., 2023)). This manuscript explores the question whether a neural network can learn a good approximate $\hat{\boldsymbol{b}}$ thereof *without* any knowledge of the density ρ_1, and only from a finite number of samples drawn therefrom.

Network architecture We consider the case where the denoising function $\boldsymbol{f}$ (301) is parametrized with a two-layer non-linear DAE with one hidden neuron, and –taking inspiration from modern practical architectures such as U-nets (Ronneberger et al., 2015a)– a trainable skip connection:

$$\boldsymbol{f}_{\boldsymbol{w}_t,c_t}(\boldsymbol{x}) = c_t \times \boldsymbol{x} + \boldsymbol{w}_t \times \varphi(\boldsymbol{w}_t^\top \boldsymbol{x}), \tag{306}$$

where φ is assumed to tend to 1 (resp. -1) as its argument tends to $+\infty$ (resp $-\infty$). Sign, tanh and erf are simple examples of such an activation function. The trainable parameters are therefore $c_t \in \mathbb{R}, \boldsymbol{w}_t \in \mathbb{R}^d$. Note that (306) is a special case of the architecture studied in Chapter 10. It differs from the very similar network considered in Shah et al. (2023) in that it covers a slightly broader range of activation functions (Shah et al. (2023) address the case $\varphi = \tanh$), and in that the skip connection istrainable –rather than fixed–. Since we consider the case where a separate network is trained at every time step, the empirical risk (302) decouples over the time index t. The parameters $\boldsymbol{w}_t, c_t$ of the DAE (306) should therefore minimize

$$\hat{\mathscr{R}}_t(\boldsymbol{w}_t, c_t) = \sum_{\mu=1}^{n} \|\boldsymbol{f}_{c_t,\boldsymbol{w}_t}(\boldsymbol{x}_t^\mu) - \boldsymbol{x}_1^\mu\|^2 + \frac{\lambda}{2}\|\boldsymbol{w}_t\|^2, \tag{307}$$

where for generality we also allowed for the presence of a ℓ_2 regularization of strength λ. We remind that $\boldsymbol{x}_t^\mu = \alpha(t)\boldsymbol{x}_0^\mu + \beta(t)\boldsymbol{x}_1^\mu$, with $\{\boldsymbol{x}_1^\mu\}_{\mu=1}^n$ (resp. $\{\boldsymbol{x}_0^\mu\}_{\mu=1}^n$) n training samples independently drawn from the target density ρ_1 (305) (resp. the base density $\rho_0 = \mathcal{N}(0, \mathbb{I}_d)$), collected in the training set $\mathscr{D}$.

Asymptotic limit We consider in this manuscript the asymptotic limit $d \to \infty$, with $n, \|\boldsymbol{\mu}\|^2/d, \sigma = \Theta_d(1)$. For definiteness, in the following, we set $\|\boldsymbol{\mu}\|^2/d = 1$. Note that Chapter 10 considered the different limit $\|\boldsymbol{\mu}\| = \Theta_d(1)$. Shah et al. (2023) on the other hand address a larger range of asymptotic limits, including the present one, but does not provide tight characterizations, nor an analysis of the generative process.

11.2 LEARNING

In this section, we first provide sharp closed-form characterizations of the minimizers $\hat{c}_t, \hat{\boldsymbol{w}}_t$ of the objective $\hat{\mathscr{R}}_t$ (307). The next section discusses how these formulae can be leveraged to access a tight characterization of the associated flow.

Result 11.2.1. *(**Sharp characterization of minimizers of** (307)) For any given activation φ satisfying $\varphi(x) \xrightarrow{x\to\pm\infty} \pm 1$ and any $t \in [0,1]$, in the limit $d \to \infty$, $n, \|\boldsymbol{\mu}\|^2/d, \sigma = \Theta_d(1)$, the skip connection strength $\hat{c}_t$ minimizing (307) is given by*

$$\hat{c}_t = \frac{\beta(t)(\lambda(1+\sigma^2) + (n-1)\sigma^2)}{\alpha(t)^2(\lambda + n - 1) + \beta(t)^2(\lambda(1+\sigma^2) + (n-1)\sigma^2)}. \tag{308}$$

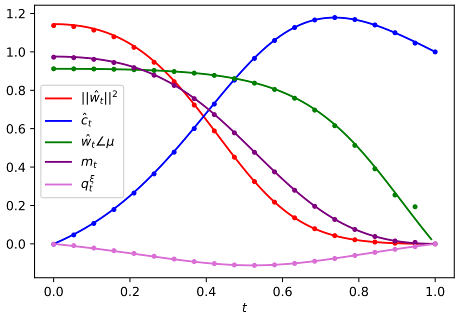


Figure 37: $n=4, \sigma=0.9, \lambda=0.1, \alpha(t)=1-t, \beta(t)=t, \varphi=\tanh$. Solid lines: theoretical predictions of Result 11.2.1: squared norm of the DAE weight vector $\|\hat{\boldsymbol{w}}_t\|^2$ (red), skip connection strength $\hat{c}_t$ (blue) cosine similarity between the weight vector $\hat{\boldsymbol{w}}_t$ and the target cluster mean $\boldsymbol{\mu}$, $\hat{\boldsymbol{w}}_t\angle\boldsymbol{\mu} \equiv \hat{\boldsymbol{w}}_t^\top\boldsymbol{\mu}/\|\boldsymbol{\mu}\|\|\hat{\boldsymbol{w}}_t\|$ (green), components m_t, q_t^ξ of $\hat{\boldsymbol{w}}_t$ along the vectors $\boldsymbol{\mu}_{\text{emp.}}, \boldsymbol{\xi}$ (purple, pink, orange). Dots: numerical simulations in dimension $d=5\times10^4$, corresponding to training the DAE (306) on the risk (307) using the `Pytorch` implementation of full-batch Adam, with learning rate 0.0001 over 4×10^4 epochs and weight decay $\lambda=0.1$. The experimental points correspond to a single instance of the model.

Furthermore, the learnt weight vector $\hat{\boldsymbol{w}}_t$ *is asymptotically contained in* $\text{span}(\boldsymbol{\mu}_{\text{emp.}}, \boldsymbol{\xi})$ *(in the sense that its projection on the orthogonal space* $\text{span}(\boldsymbol{\mu}_{\text{emp.}}, \boldsymbol{\xi})$ *has asymptotically vanishing norm), where*

$$\boldsymbol{\xi} \equiv \sum_{\mu=1}^{n} s^\mu \boldsymbol{x}_0^\mu, \qquad\qquad \boldsymbol{\mu}_{\text{emp.}} = \frac{1}{n}\sum_{\mu=1}^{n} s^\mu x_1^\mu. \tag{309}$$

In other words, $\boldsymbol{\mu}_{\text{emp.}}$ *is the empirical mean of the training samples. We remind that* $s^\mu = \pm 1$ *was defined below* (305) *and indicates the cluster the* $\mu-th$ *sample* $\boldsymbol{x}_1^\mu$ *belongs to. The components of* $\hat{\boldsymbol{w}}_t$ *along each of these three vectors is described by the summary statistics*

$$m_t = \frac{\boldsymbol{\mu}_{\text{emp.}}^\top \hat{\boldsymbol{w}}_t}{d(1+\sigma^2/n)}, \qquad\qquad q_t^\xi = \frac{\hat{\boldsymbol{w}}_t^\top \boldsymbol{\xi}}{nd}, \tag{310}$$

which concentrate as $d\to\infty$ *to the quantities characterized by the closed-form formulae*

$$\begin{cases} m_t = \frac{n}{\lambda+n}\frac{\alpha(t)^2(\lambda+n-1)}{\alpha(t)^2(\lambda+n-1)+\beta(t)^2(\lambda(1+\sigma^2)+(n-1)\sigma^2)} \\ q_t^\xi = \frac{-\alpha(t)}{\lambda+n}\frac{\beta(t)(\lambda(1+\sigma^2)+(n-1)\sigma^2)}{\alpha(t)^2(\lambda+n-1)+\beta(t)^2(\lambda(1+\sigma^2)+(n-1)\sigma^2)} \end{cases}. \tag{311}$$

The derivation of Result 11.2.1 involves a heuristic partition function computation, borrowing ideas from statistical physics. The theoretical predictions for the skip connection strength $\hat{c}_t$ and the component m_t, q_t^ξ of the weight vector $\hat{\boldsymbol{w}}_t$ are plotted as solid lines in Fig. 37, and display good agreement with numerical simulations, corresponding to training the DAE (306) on the risk (307) using the `Pytorch` (Paszke et al., 2019b) implementation of the Adam optimizer (Kingma et al., 2014a).

A notable consequence of (310) is that the weight vector $\hat{\boldsymbol{w}}_t$ is contained at all times t in the two-dimensional subspace spanned by the empirical cluster mean $\boldsymbol{\mu}_{\text{emp.}}$ and the vectors $\boldsymbol{\xi}$ (309) – in other words, the learnt weights align to some extent with the empirical mean, but still possess a non-zero component along $\boldsymbol{\xi}$, which is orthogonal thereto. $\boldsymbol{\xi}$ subsumes the aggregated effect of the base vectors $\{\boldsymbol{x}_0^\mu\}_{\mu=1}^n$ used in the train set. Rather remarkably, the training samples thus only enter in the characterization of $\hat{\boldsymbol{w}}_t$ through the form of simple sums (309). Since the vector $\boldsymbol{\xi}$ is associated to the training samples, the fact that the learnt vector $\hat{\boldsymbol{w}}_t$ has non-zero components along $\boldsymbol{\xi}$ hence signals a form of overfitting and memorization. Interestingly, Fig. 37 shows that the extent of this overfitting is non-monotonic in time, as $|q_t^\xi|$ first increases then decreases. Finally, note that this effect is as expected mitigated as the number of training samples n increases. From (311), for large n, $m_t = \Theta_n(1)$ while the components q_t^ξ is suppressed as $\Theta_n(1/n)$. Finally, Result 11.2.1 and equation (303) can be straightforwardly combined to yield a sharp characterization of the learnt estimate $\hat{\boldsymbol{b}}$ of the velocity field $\boldsymbol{b}$ (298). This characterization can be in turn leveraged to build a tight description of the generative flow (304). This is the object of the following section.

11.3 GENERATIVE PROCESS

While Corollary 11.2.1, together with the definition (303), provides a concise characterization of the velocity field $\hat{\boldsymbol{b}}$, the sampling problem (304) remains formulated as a high-dimensional, and therefore hard to analyze, transport process. The following result shows that the dynamics of a sample $\boldsymbol{X}_t$ following the differential equation (304) can nevertheless be succinctly tracked using a finite number of scalar summary statistics.

Result 11.3.1. *(**Summary statistics**) Let $\boldsymbol{X}_t$ be a solution of the ordinary differential equation (304) with initial condition $\boldsymbol{X}_0$. For a given t, the projection of $\boldsymbol{X}_t$ on* span$(\boldsymbol{\mu}_{\text{emp.}}, \boldsymbol{\xi}$ *is characterized by the summary statistics*

$$M_t \equiv \frac{\boldsymbol{X}_t^\top \boldsymbol{\mu}_{\text{emp.}}}{d(1+\sigma^2/n)}, \qquad\qquad Q_t^\xi \equiv \frac{\boldsymbol{X}_t^\top \boldsymbol{\xi}}{nd}. \tag{312}$$

With probability asymptotically $1/2$ *the summary statistics M_t, Q_t^ξ (312) concentrate for all t to the solution of the ordinary differential equations*

$$\begin{cases} \frac{d}{dt}M_t = \left(\dot{\beta}(t)\hat{c}_t + \frac{\dot{\alpha}(t)}{\alpha(t)}(1-\hat{c}_t\beta(t))\right)M_t + \left(\dot{\beta}(t) - \frac{\dot{\alpha}(t)}{\alpha(t)}\beta(t)\right)m_t \\ \frac{d}{dt}Q_t^\xi = \left(\dot{\beta}(t)\hat{c}_t + \frac{\dot{\alpha}(t)}{\alpha(t)}(1-\hat{c}_t\beta(t))\right)Q_t^\xi + \left(\dot{\beta}(t) - \frac{\dot{\alpha}(t)}{\alpha(t)}\beta(t)\right)q_t^\xi \end{cases}, \tag{313}$$

with initial condition $M_0 = Q_0^\xi = 0$, and with probability asymptotically $1/2$ *they concentrate to minus the solution of* (313)*. Furthermore, the orthogonal component $\boldsymbol{X}_t^\perp \in \mathrm{span}(\boldsymbol{\mu}_{\mathrm{emp.}}, \boldsymbol{\xi})^\perp$ obeys the simple linear differential equation*

$$\frac{d}{dt}\boldsymbol{X}_t^\perp = \left(\dot{\beta}(t)\hat{c}_t + \frac{\dot{\alpha}(t)}{\alpha(t)}(1-\hat{c}_t\beta(t))\right)\boldsymbol{X}_t^\perp. \tag{314}$$

Finally, the statistic $Q_t \equiv \|\boldsymbol{X}_t\|^2/d$ is given with high probability by

$$Q_t = M_t^2(1+\sigma^2/n) + n(Q_t^\xi)^2 + e^{2\int_0^t \left(\dot{\beta}(t)\hat{c}_t + \frac{\dot{\alpha}(t)}{\alpha(t)}(1-\hat{c}_t\beta(t))\right)dt}. \tag{315}$$

Taking a closer look at (313), it might seem at first from equations (313) that there is a singularity for $t=1$ since $\alpha(1)=0$ in the denominator. Remark however that both $1-\beta(t)\hat{c}_t$ (308) and m_t (311) are actually proportional to $\alpha(t)^2$, and therefore (313) is in fact also well defined for $t=1$. In practice, the numerical implementation of a generative flow like (304) often involves a discretization thereof, given a discretization scheme $\{t_k\}_{k=0}^N$ of $[0,1]$, where $t_0=0$ and $t_N=1$:

$$\boldsymbol{X}_{t_{k+1}} = \boldsymbol{X}_{t_k} + \hat{\boldsymbol{b}}(\boldsymbol{X}_{t_k}, t_k)(t_{k+1}-t_k). \tag{316}$$

The evolution of the summary statistics introduced in Result 11.3.1 can be rephrased in more actionable form to track the discretized flow (316).

Remark 11.3.2. ***(Summary statistics for the discrete flow)*** *Let $\{\boldsymbol{X}_{t_k}\}_{k=0}^N$ be a solution of the discretized learnt flow* (304)*, for an arbitrary discretization scheme $\{t_k\}_{k=0}^N$ of $[0,1]$, where $t_0=0$ and $t_N=1$, with initial condition $\boldsymbol{X}_{t_0}\sim\rho_0$. The summary statistics introduced in Result* 11.3.1 *are then equal to the solutions of the recursions*

$$\begin{cases} M_{t_{k+1}} = M_{t_k} + \delta t_k\left(\dot{\beta}(t_k)\hat{c}_{t_k} + \frac{\dot{\alpha}(t_k)}{\alpha(t_k)}(1-\hat{c}_{t_k}\beta(t_k))\right)M_{t_k} \\ \qquad +\delta t_k\left(\dot{\beta}(t_k) - \frac{\dot{\alpha}(t_k)}{\alpha(t_k)}\beta(t_k)\right)m_{t_k} \\ Q^\xi_{t_{k+1}} = Q^\xi_{t_k} + \delta t_k\left(\dot{\beta}(t_k)\hat{c}_{t_k} + \frac{\dot{\alpha}(t_k)}{\alpha(t_k)}(1-\hat{c}_{t_k}\beta(t_k))\right)Q^\xi_{t_k} \\ \qquad +\delta t_k\left(\dot{\beta}(t_k) - \frac{\dot{\alpha}(t_k)}{\alpha(t_k)}\beta(t_k)\right)q^\xi_{t_k} \end{cases}, \tag{317}$$

with probability $1/2$*, and to the opposite theoreof with probability* $1/2$*. In* (317)*, the initial conditions are understood as $M_{t_0} = Q^\xi_{t_0} = 0$, and we have denoted $\delta t_k \equiv t_{k+1}-t_k$ for clarity. Furthermore, the orthogonal component $\boldsymbol{X}_{t_k}^\perp \in \mathrm{span}(\boldsymbol{\mu}_{\mathrm{emp.}}, \boldsymbol{\xi})^\perp$ obeys the simple linear recursion*

$$\boldsymbol{X}^\perp_{t_{k+1}} = \left[1 + \delta t_k\left(\dot{\beta}(t_k)\hat{c}_{t_k} + \frac{\dot{\alpha}(t_k)}{\alpha(t_k)}(1-\hat{c}_{t_k}\beta(t_k))\right)\right]\boldsymbol{X}^\perp_{t_k}. \tag{318}$$

Finally, the statistic $Q_{t_k} \equiv \|\boldsymbol{x}_{t_k}\|^2/d$ is given with high probability by

$$Q_{t_k} = M_{t_k}^2(1+\sigma^2/n) + n(Q_{t_k}^{\xi})^2 + \prod_{\ell=0}^{k}\left[1+\left(\dot{\beta}(t_\ell)\hat{c}_{t_\ell}+\frac{\dot{\alpha}(t_\ell)}{\alpha(t_\ell)}(1-\hat{c}_{t_\ell}\beta(t_\ell))\right)\delta t_\ell\right]^2. \tag{319}$$

Equations (317),(318) and (319) of Remark 11.3.2 are consistent discretizations of the continuous flows (313),(314) and (315) of Result 11.3.1 respectively, and converge thereto in the limit of small discretization steps $\max_k \delta t_k \to 0$. An important consequence of Result 11.3.1 is that the transport of a sample $\boldsymbol{X}_0 \sim \rho_0$ by (304) factorizes into the low-dimensional deterministic evolution of its projection on the low-rank subspace $\mathrm{span}(\boldsymbol{\mu}_{\mathrm{emp.}}, \boldsymbol{\xi})$, as tracked by the two summary statistics M_t, Q_t^{ξ}, and the simple linear dynamics of its projection on the orthogonal space $\mathrm{span}(\boldsymbol{\mu}_{\mathrm{emp.}}, \boldsymbol{\xi})^{\perp}$. Result 11.3.1 thus reduces the high-dimensional flow (304) into a set of two scalar ordinary differential equations (313) and a simple homogeneous linear differential equation (314). The theoretical predictions of Result (11.3.1) and Remark 11.3.2 for the summary statistics M_t, Q_t^{ξ}, Q_t are plotted in Fig. 38, and display convincing agreement with numerical simulations, corresponding to discretizing the flow (304) in $N = 100$ time steps, and training a separate network for each step as described in Section 11.1. A PCA visualization of the flow is further provided in Fig. 38 (middle).

Leveraging the simple characterization of Result 11.3.1, one is now in a position to characterize the generated distribution $\hat{\rho}_1$, which is the density effectively sampled by the generative model. In particular, Result 11.3.1 establishes that the distribution $\hat{\rho}_1$ is Gaussian over $\mathrm{span}(\boldsymbol{\mu}_{\mathrm{emp.}}, \boldsymbol{\xi})^{\perp}$ – since $\boldsymbol{X}_0^{\perp}$ is Gaussian and the flow is linear–, while the density in $\mathrm{span}(\boldsymbol{\mu}_{\mathrm{emp.}}, \boldsymbol{\xi})$ concentrates along the vector $\hat{\boldsymbol{\mu}}$ described by the components (313). The density $\hat{\rho}_1$ is thus described by a mixture of two clusters, Gaussian along $d-2$ directions, centered around $\pm\hat{\boldsymbol{\mu}}$. The following corollary provides a sharp characterization of the squared distance between the mean $\hat{\boldsymbol{\mu}}$ of the generated density $\hat{\rho}_1$ and the true mean $\boldsymbol{\mu}$ of the target density ρ_1.

Corollary 11.3.3. ***(Mean squared error of the mean estimate)*** *Let $\hat{\boldsymbol{\mu}}$ be the cluster mean of the density $\hat{\rho}_1$ generated by the (continuous) learnt flow (304). In the asymptotic limit described by Result 11.2.1, the squared distance between $\hat{\boldsymbol{\mu}}$ and the true mean $\boldsymbol{\mu}$ is given by*

$$\frac{1}{d}\|\hat{\boldsymbol{\mu}} - \boldsymbol{\mu}\|^2 = M_1^2 + n(Q_1^{\xi})^2 + n\sigma^2(Q_1^{\eta})^2 + 1 - 2M_1, \tag{320}$$

with $M_1, Q_1^{\xi}, Q_1^{\eta}$ being the solutions of the ordinary differential equations (313) evaluated at time $t = 1$. Furthermore, the cosine similarity between $\hat{\boldsymbol{\mu}}$ and the true mean $\boldsymbol{\mu}$ is given by

$$\hat{\boldsymbol{\mu}}\angle\boldsymbol{\mu} = \frac{M_1}{\sqrt{Q_1}}. \tag{321}$$

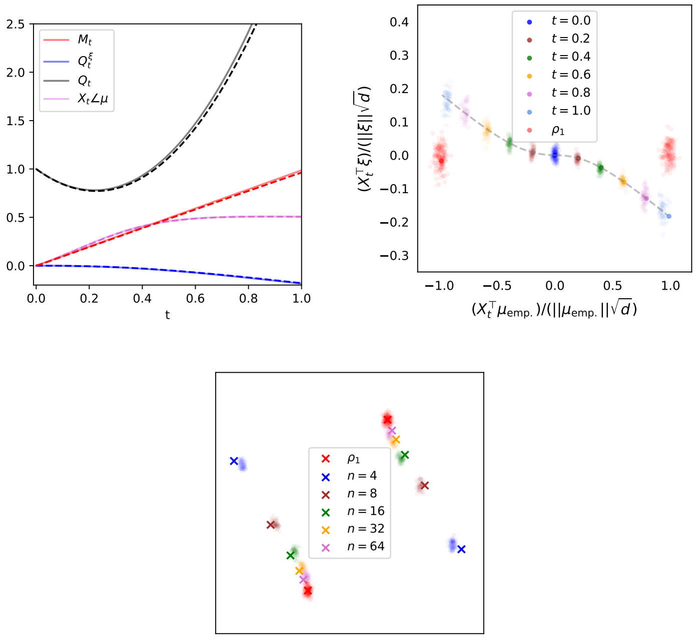


Figure 38: In all three plots, $\lambda = 0.1, \alpha(t) = 1-t, \beta(t) = t, \varphi = \text{sign}$. **(left)** $\sigma = 1.5, n = 8$. Temporal evolution of the summary statistics $M_t, Q_t^\xi, Q_t, \boldsymbol{X}_t\angle\boldsymbol{\mu}$ (312). Solid lines correspond to the theoretical prediction of (312) in Result 11.3.1, while dashed lines correspond to numerical simulations of the generative model, by discretizing the differential equation (304) with step size $\delta t = 0.01$, and training a separate DAE for each time step using Adam with learning rate 0.01 for 2000 epochs. All experiments were conducted in dimension $d = 5000$, and a single run is represented. **(middle)** $\sigma = 2, n = 16$. Projection of the distribution of $\boldsymbol{X}_t$ (304) in $\text{span}(\boldsymbol{\mu}_{\text{emp.}}, \boldsymbol{\xi})$, transported by the velocity field $\hat{\boldsymbol{b}}$ (303) learnt from data. The point clouds correspond to numerical simulations. The dashed line corresponds to the theoretical prediction of the means of the cluster, as given by equation (313) of Result 11.3.1. The target Gaussian mixture ρ_1 is represented in red. The base zero-mean Gaussian density ρ_0 (dark blue) is split by the flow (304) into two clusters, which approach the target clusters (red) as time accrues . **(right)** $\sigma = 2$. PCA visualization of the generated density $\hat{\rho}_1$, by training the generative model on n samples, for $n \in \{4, 8, 16, 32, 64\}$. Point clouds represent numerical simulations of the generative model. Crosses represent the theoretical predictions of Result 11.3.1 for the means of the clusters of $\hat{\rho}_1$, as given by equation (313) of Result 11.3.1 for $t = 1$. As the number of training samples n increases, the generated clusters of $\hat{\rho}_1$ approach the target clusters of ρ_1, represented in red.

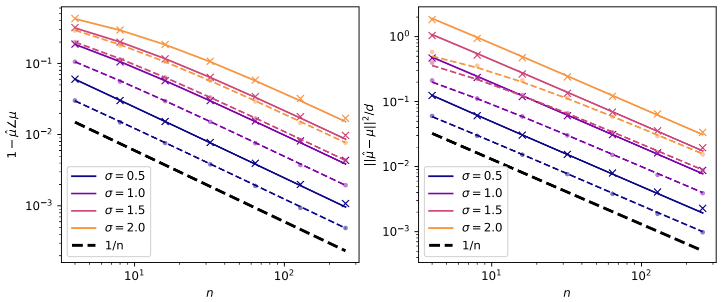


Figure 39: $\alpha(t) = 1-t, \beta(t) = t, \varphi =$ sign. Cosine asimilarity (left) and mean squared distance (right) between the mean $\hat{\boldsymbol{\mu}}$ of the generated mixture $\hat{\rho}_1$ and the mean $\boldsymbol{\mu}$ of the target density ρ_1, as a function of the number of training samples n, for various variances σ of ρ_1. Solid lines represent the theoretical characterization of Corollary 11.3.3. Crosses represent numerical simulations of the generative model, by discretizing the differential equation (304) with step size $\delta t = 0.01$, and training a separate DAE for each time step using the `Pytorch` implementation of the full-batch Adam optimizer, with learning rate 0.04 and weight decay $\lambda = 0.1$ for 6000 epochs. All experiments were conducted in dimension $d = 5 \times 10^4$, and a single run is represented. Dashed lines indicate the performance of the Bayes-optimal estimator $\hat{\boldsymbol{\mu}}^\star$, as theoretically characterized in Remark 11.4.1. Dots indicate the performance of the PCA estimator, which is found as in Cui et al. (2024b) to yield performances nearly identical to the Bayes-optimal estimator.

Finally, both the MSE $1/d\|\hat{\boldsymbol{\mu}} - \boldsymbol{\mu}\|^2$ (320) *and the cosine asimilarity* $1 - \hat{\boldsymbol{\mu}}\angle\boldsymbol{\mu}$ (321) *decay as* $\Theta_n(1/n)$ *for large number of samples* n.

The theoretical predictions of the learning metrics (320) and (321) are plotted in Fig. 39 as a function of the number of samples, along with the corresponding numerical simulations, and display a clear $\Theta_n(1/n)$ decay, signalling the convergence of the generated density $\hat{\rho}_1$ to the true target density ρ_1 as the sample complexity accrues. A PCA visualization of this convergence is further presented in Fig.38 (right). Intuitively, this is because the DAE learns the empirical means up to a $\Theta_n(1/n)$ component along $\boldsymbol{\xi}$, and that the empirical means itself converges to the true mean with rate $\Theta_n(1/n)$.

11.4 BAYES-OPTIMAL BASELINE

Corollary 11.3.3 completes the study of the performance of the DAE-parametrized generative model. It is natural to wonder whether one can improve on the $\Theta_n(1/n)$ rate that it achieves. A useful baseline to compare with is the Bayes-optimal estimator $\hat{\boldsymbol{\mu}}^\star$, yielded by Bayesian inference when in addition to the dataset $\mathscr{D} = \{\boldsymbol{x}_1^\mu\}_{\mu=1}^n$, the form of the distribution (305) and the variance σ are known, but *not* the mean $\boldsymbol{\mu}$ –which for definiteness and without loss of generality will be assumed in this section to be have been drawn at random from $\mathcal{N}(0, \mathbb{I}_d)$. The following remark provides a tight characterization of the MSE achieved by this estimator.

Remark 11.4.1. *(**Bayes-optimal estimator of the cluster mean**) The Bayes-optimal estimator $\hat{\boldsymbol{\mu}}^\star$ of $\boldsymbol{\mu}$ assuming knowledge of the functional form of the target density (305), the cluster variance σ, and the training set $\mathscr{D}$, is defined as the minimizer of the average squared error*

$$\hat{\boldsymbol{\mu}}^\star = \underset{\nu}{\operatorname{arginf}} \mathbb{E}_{\boldsymbol{\mu}\sim\mathcal{N}(0,\mathbb{I}_d),\mathscr{D}\sim\rho_1^{\otimes n}} \|\nu(\mathscr{D}) - \boldsymbol{\mu}\|^2. \tag{322}$$

In the asymptotic limit of Result 11.2.1, the Bayes-optimal estimator $\hat{\boldsymbol{\mu}}^\star(\mathscr{D})$ is parallel to the empirical mean $\boldsymbol{\mu}_{\text{emp.}}$. Its component $m^\star \equiv \boldsymbol{\mu}_{\text{emp.}}^\top \hat{\boldsymbol{\mu}}^\star(\mathscr{D})/d(1+\sigma^2/n)$ concentrate asymptotically to

$$m^\star = \frac{n}{n+\sigma^2}, \tag{323}$$

Finally, with high probability, the Bayes-optimal MSE reads

$$\frac{1}{d}\|\hat{\boldsymbol{\mu}}^\star(\mathscr{D}) - \boldsymbol{\mu}\|^2 = \frac{\sigma^2}{n+\sigma^2}. \tag{324}$$

In particular, (324) implies that the optimal MSE decays as $\Theta_n(1/n)$.

Remark 11.4.1 thus establishes that the Bayes-optimal MSE decays as $\Theta_n(1/n)$ with the number of available training samples. Note that while the Bayes-optimal estimator is colinear to the empirical mean, it is differs therefrom by a non-trivial multiplicative factor. On the other hand, the $\Theta_n(1/n)$ rate is intuitively due to the $\Theta_n(1/n)$ convergence of the empirical mean to the true mean. Contrasting to Corollary 11.3.3 for the MSE associated to the mean $\hat{\boldsymbol{\mu}}$ of the density $\hat{\rho}_1$ learnt by the generative model, it follows that *the latter achieves the Bayes-optimal learning rate*. The Bayes-optimal MSE (324) predicted by Remark 11.4.1 is plotted in dashed lines in Fig. 39, alongside the MSE achieved by the generative model (see Corollary 11.3.3). The common $1/n$ decay rate is also plotted in dashed black for comparison. Finally, we observe that the estimate of $\boldsymbol{\mu}$ inferred by PCA, plotted as dots in Fig. 39, leads to a cosine similarity which is very close to the Bayes-optimal one, echoing the findings of Chapter 10 in another asymptotic limit. We however stress an important distinction between the generative model analyzed in previous sections and the Bayes and PCA estimators dicussed in the present section. The generative model is tasked with estimating the full distribution ρ_1 only from data, while being completely agnostic thereof. In contrast, PCA and Bayesian inference only offer an estimate of the cluster mean, and require an exact oracle knowledge of its functional form (305) and the cluster variance σ. They do *not*, therefore, constitute generative models and are only discussed in the present section as insightful baselines.

It is a rather striking finding that the DAE (306) succeeds in approximately sampling from ρ_1(305) when trained on but $n = \Theta_d(1)$ samples –instead of simply generating back memorized training samples–, and further displays information-theoretically optimal learning rates. The answer to this puzzle, lies in the fact that the architecture (306) is very close to the functional form of the exact velocity field b (298), and is therefore implicitly biased towards learning the latter – while also not being expressive enough to too detrimentally overfit. A thorough exploration of the inductive bias for more complex architectures is an important and fascinating entreprise, which falls out of the scope of the present manuscript and is left for future work.

CONCLUSION

We conduct a tight end-to-end asymptotic analysis of estimating and sampling a binary Gaussian mixture using a flow-based generative model, when the flow is parametrized by a shallow auto-encoder. We provide sharp closed-form characterizations for the trained weights of the network, the learnt velocity field, a number of summary statistics tracking the generative flow, and the distance between the mean of the generated mixture and the mean of the target mixture. The latter is found to display a $\Theta_n(1/n)$ decay rate, where n is the number of samples, which is further shown to be the Bayes-optimal rate. In contrast to most studies of flow-based generative models in high dimensions, the learning and sampling processes are jointly and sharply analyzed in the present manuscript, which affords the possibility to explicitly investigate the effect of a limited sample complexity at the level of the generated density.

12

DOT-PRODUCT ATTENTION

Recent years have seen an upheaval in our ability to learn and implement complex tasks from textual data. Central in these advances is the use of self-attention layers (Vaswani et al., 2017), which provide an efficient method of extracting information from sentences – both the information encoded in the ordering (i.e. *positions*) of the words, and that encoded in the meaning (i.e. *semantics*) of the words. In theory, attention mechanisms are able to leverage both types of information, by having tokens attend to each other based on their respective positions (called positional attention in (Jelassi et al., 2022)) and/or respective meanings (henceforth referred to as semantic attention).

We aim to expand the up-to-now rather scarce theoretical understanding of learning with attention layers. Seminal open questions include: To what extent do transformers learn semantic or positional attention matrices? How does this depend on the amount of available data, the task, or the type of embedding? The present manuscript explores these questions, by proposing and analysing a solvable model of dot-product attention that can learn to implement both positional and semantic attention mechanisms from data. In particular our contributions are:

- We first illustrate how, for the histogram task, two qualitatively different solutions exist in the same loss landscape of a simple transformer, respectively corresponding to positional and semantic attention.
- We then move on to describe a model with a single self-attention layer with tied, low-rank query and key matrices. On Gaussian input data and realizable outputs, we show that this model exhibits a phase transition in terms of sample complexity between a semantic and positional mechanism.
- For this model, in the asymptotic limit where the embedding dimension d of the tokens and the number n of training samples are proportionally large, we provide a tight closed-form characterization of the test error and training loss achieved at the minima of the non-convex empirical loss. Using this high-dimensional characterization, we locate the positional-semantic phase transition, thus providing the first theoretical result about the emergence of sharp phase transitions in a model of dot-product attention.
- We contrast the performance of the dot-product attention layer with that of a linear attention layer, which can only implement positional

mechanisms, and how the former outperforms the latter once it learns the semantic mechanism.

12.0.1 RELATED WORK

Theory of attention Attention models have been the object of sizeable theoretical scrutiny in recent years, with a growing body of work investigating various aspects such as their expressivity (Fu et al., 2023; Edelman et al., 2022; Hahn, 2020), inductive bias (Sahiner et al., 2022; Tarzanagh et al., 2023b; Tarzanagh et al., 2023a), training dynamics (Jelassi et al., 2022; Boix-Adsera et al., 2023; Li et al., 2023b; Tian et al., 2023), and in-context learning (Bai et al., 2023; Guo et al., 2023; Li et al., 2023c; Zhang et al., 2023). Geshkovski et al. (2023) and Fu et al. (2023) analyze models with frozen non-trainable queries and keys, while Jelassi et al. (2022) similarly studies the learning of the value matrix and positional encodings only, fixing keys and queries to identity. The works of (Sahiner et al., 2022; Zhang et al., 2023) address trainable queries and keys for linear or ReLu-activated attention mechanisms. Li et al. (2023b) and Edelman et al. (2022) provide error bounds for non-linear models, with trainable queries and keys. Because these studies are not tight, they do not allow to capture sharp changes in the behaviour of attention mechanisms such as phase transitions. A first tight analysis was provided in (Rende et al., 2023), in the context of learning a high-dimensional graphical model with factored attention, leveraging its formal equivalence to a linear and convex learning problem. The present manuscript conducts a tight analysis of the non-convex learning of a non-linear attention model with trainable tied queries and keys, thereby allowing the description of sharp phase transitions in the behaviour and performance of the model.

Positional encodings To combine the positional and semantic information in textual or general sequential data, a plethora of models and input encodings have been explored. Many approaches are based on autoregressive models, e.g. recurrent architectures (Elman, 1990), where the positional information is provided implicitly by the order in which the input is processed. While some transformers can leverage implicit positional information through causal masks in training (Haviv et al., 2022; Sinha et al., 2022; Kazemnejad et al., 2023), in principle a dot product attention layer requires an explicit encoding of positional information as it views the input sequence in parallel, as a bag of words (Vaswani et al., 2017). Several works experimentally explore different types of positional encodings with the goal of improving the downstream task performance (Shaw et al., 2018; Ruoss et al., 2023). In this work, we provide a tractable model to quantify the generalization error of a single layer of attention in the presence of positional encodings.

Mechanistic interpretability Recently, there has been an effort to reverse engineer the algorithms learned by a neural network for a specific task (Weiss et al., 2021; Olsson et al., 2022; Von Oswald et al., 2023) to uncover their

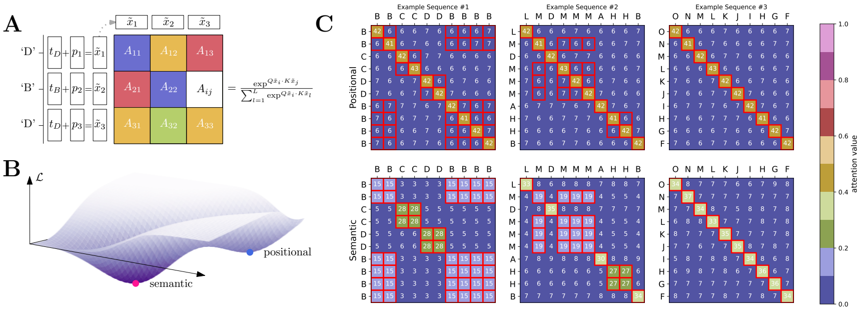


Figure 40: **Several solutions exist for the histogram task.** *(A)* The sequence $[D, B, D]$ is processed by a single layer of dot-product attention. After embedding each token into learned vectors $[t_D, t_B, t_D] \in \mathbb{R}^{3\times d}$, the absolute positional encodings $[p_1, p_2, p_3]$ are added to give the inputs to the attention layer. The colored elements A_{ij} represent the values of the attention matrix, as generated using the key and query matrices Q and K and after applying softmax. *(B)* A schematic loss landscape containing two stable solutions. *(C)* Elements of attention matrices for the histogram task for local minima in the loss landscape. We generated a dataset of sequences by sampling each token of the sequence i.i.d. from the uniform distribution over all tokens. Models were trained with their respective frozen initialization using $n = 35{,}000$ samples and the Adam optimizer. *Top Row:* The attention matrix of the positional solution is largely independent of the specific input sequence. *Bottom Row:* The attention matrices from the semantic solution vary based on the input token. Red squares highlight the elements of A_{ij} where $x_i = x_j$.

limitations and generalization abilities. It has been shown that a transformer may implement two qualitatively different algorithms for modular addition (Zhong et al., 2023). In a similar spirit, our work provides examples of two tasks for which an attention layer can implement two qualitatively different solutions, based on either the positions or semantics of the inputs. While many works in mechanistic interpretability rely on a careful introspection and interpretation of the learned model to come to that conclusion, our second example allows for a theoretical analysis.

12.1 TWO SOLUTIONS FOR THE HISTOGRAM TASK

In this section, we demonstrate that for a simple counting task two qualitatively different solutions exist in the loss landscape of a simple transformer using a dot-product attention layer with positional encodings. One solution corresponds to a dot-product attention matrix which is largely independent of the tokens making up the input sequence, and another strongly varies based on the tokens (and thus the semantic content of the input). Both solutions achieve a close to 100% test accuracy.

The training task is a sequence-to-sequence counting task, referred to as the *histogram task* in (Weiss et al., 2021). Given an input sequence $\boldsymbol{x} =$

$[x_1, x_2, \cdots, x_L]$ of length L of tokens from a fixed alphabet, the goal is to return a sequence $\boldsymbol{y} = [y_1, y_2, \cdots, y_L]$, where each token y_i is the number of occurrences of the token x_i in $\boldsymbol{x}$. In Fig. 40 (A), we show an example where we consider sequences where the tokens are from the fixed alphabet $\mathscr{X} = \{A, B, C, \cdots\}$ of size $|\mathscr{X}| = 15$. When the input data is limited to length L, the output elements y_i thus take values up to the maximum count L.

In this setting, for instance, the sequence $\boldsymbol{x} = [A, B, B, C, A, B]$ should be mapped to its histogram sequence $\boldsymbol{y} = [2, 3, 3, 1, 2, 3]$.

We encode the input using token embeddings and absolute positional encodings which are trained jointly with the model weights. As an architecture we consider a small transformer made up of a single layer of dot-product attention, followed by a fully connected hidden layer and with learned embeddings for both tokens and positions. For each output position, it generates logits for the L possible classes of the output alphabet; training is done using the cross-entropy loss.

We conduct experiments where we set two different sections of the models' weights to zero at the initialization of training –removing either the model's access to positional or semantic information and keeping the weights frozen throughout training with the Adam optimizer. After convergence, we check that the resulting configurations of weights are stable in the unconstrained loss landscape, i.e. without frozen weights. More precisely, we ascertain that these weights only change marginally when further trained with SGD on the unconstrained loss, and that the qualitative behaviour of the attention layer is retained. Our experiments demonstrate that the loss landscape of the transformer has at least two qualitatively different local minimizers (or close to minimizers), subsequently referred to as the semantic and positional solution.

We inspect the learnt attention matrix for different input sequences in Fig. 40 (C). The positional solution corresponds to a learnt attention matrix whose $i, j-$th component only depend on the positions i, j, and little on the tokens occupying these positions. The attention matrix is thus almost independent of the input sequence. In fact, the attention matrix is similar to the identity. In this case, the attention layer simply serves to aggregate the other tokens uniformly, and the fully connected layer learns the counting.

In contrast, the attention matrix learnt at the semantic solution displays larger $i, j-$th component if the tokens at position i and j are identical. In other words, identical tokens attend more to each other. This mechanism hence does not rely on the positions, but rather on the semantic content of the tokens. Both solutions and associated attention matrices thus correspond to feasible algorithms which ultimately allow the transformer to solve the downstream task.

Our experimental exploration gives compelling evidence that different stable solutions exist in the empirical loss landscape of simple transformers, which correspond to different algorithmic solutions to a given task. However, it remains an interpretation of an experiment and does not allow for a precise characterization of their behaviour or of the conditions under which they are established. In the remainder of this work, we turn to a simpler model of the attention layer, which presents similar phenomenology yet can be analyzed theoretically. More precisely, in Section 12.2, we provide a tight characterization of the global minimum of the empirical loss, and show in Section 12.4 that it corresponds to a semantic or positional mechanism, depending on the amount of training data and the task with a phase transition between them.

12.2 TIED LOW-RANK ATTENTION MODEL

This section introduces a simple model of supervised learning with an attention layer parametrized by learnable, tied and low-rank query and key matrices.

Input data model We consider a model of embedded sentences with uncorrelated (1-gram) words. More precisely, a sentence $\boldsymbol{x} \in \mathbb{R}^{L\times d}$, where L is the sentence length and d represents the embedding dimension, consists of L tokens $\{\boldsymbol{x}_\ell\}_{1\leq\ell\leq L}$ independently drawn from a Gaussian distribution $\boldsymbol{x}_\ell \sim \mathcal{N}(0,\Sigma_\ell)$ with covariance $\Sigma_\ell \in \mathbb{R}^{d\times d}$. In the following, we denote the probability distribution of $\boldsymbol{x}$ as p_x. Note that while this sentence model does not involve in itself statistical correlations between tokens, the task (target function) will entail interactions between different tokens.

Target function The target (teacher) is assumed to be of the form

$$y(x) = \mathtt{T}\left[\frac{1}{\sqrt{d}}\boldsymbol{x}\boldsymbol{Q}_\star\right]\boldsymbol{x} \tag{325}$$

for $\mathtt{T} : \mathbb{R}^{L\times t} \to \mathbb{R}^{L\times L}$, and $\boldsymbol{x}_\ell \in \mathbb{R}^d$ is the $\ell-$the word, i.e. the $\ell-$th row of the sentence $\boldsymbol{x} \in \mathbb{R}^{L\times d}$. The term $\mathtt{T}\left[{}^1/\sqrt{d}\boldsymbol{x}\boldsymbol{Q}_\star\right] \in \mathbb{R}^{L\times L}$ in (325) should be interpreted as the target attention matrix, which mixes the tokens of the input $\boldsymbol{x}$, with and is parametrized by the matrix $\boldsymbol{Q}_\star \in \mathbb{R}^{d\times r_t}$,

Tied attention We consider the learning of the target (325) using a parametric family of attention matrices

$$f_Q(x) = \mathtt{S}\left[\frac{1}{\sqrt{d}}(\boldsymbol{x}+\boldsymbol{p})\boldsymbol{Q}\right](\boldsymbol{x}+\boldsymbol{p}). \tag{326}$$

In (326), $\boldsymbol{p} \in \mathbb{R}^{L\times d}$ is a *fixed* matrix, corresponding to positional encodings, and $\boldsymbol{Q} \in \mathbb{R}^{d\times r_s}$ is a trainable matrix. We denote subsequently $\boldsymbol{p}_\ell \in \mathbb{R}^d$ the $\ell-$th row of $\boldsymbol{p}$. Like the target (325), the parametric function (326) takes the form of a data-dependent attention matrix $\mathtt{S}\left[{}^1/\sqrt{d}(\boldsymbol{x}+\boldsymbol{p})\boldsymbol{Q}\right] \in \mathbb{R}^{L\times L}$ mixing the tokens of the input $\boldsymbol{x}$. Note that, compared to the usual attention mechanism (Vaswani et al., 2017), (326) corresponds to setting the value matrix to identity, and – since (326) is parametrized by a single matrix $\boldsymbol{Q}$- tying the key and query matrices.

Empirical risk minimization We study the learning of the attention layer (326), when a training set is $\mathscr{D} = \{\boldsymbol{x}^\mu, y(\boldsymbol{x}^\mu)\}_{\mu=1}^n$ with n independently sampled sentences $\{\boldsymbol{x}^\mu\}_{\mu=1}^n$ is available. The target (325) can be learnt by carrying out an empirical risk minimization:

$$\hat{\boldsymbol{Q}} = \underset{\boldsymbol{Q}\in\mathbb{R}^{d\times r}}{\mathrm{argmin}} \left[\sum_{\mu=1}^{n} \frac{1}{2d} \|y(\boldsymbol{x}^\mu) - f_Q(\boldsymbol{x}^\mu)\|^2 + \frac{\lambda}{2}\|\boldsymbol{Q}\|^2\right]. \tag{327}$$

The performance of the resulting trained model $f_{\hat{\boldsymbol{Q}}}$ is measured at test time by the MSE

$$\varepsilon_g \equiv \frac{1}{dL}\mathbb{E}_{\boldsymbol{x}\sim p_x}\left\|y(\boldsymbol{x}) - f_{\hat{\boldsymbol{Q}}}(\boldsymbol{x})\right\|^2. \tag{328}$$

12.3 CLOSED-FORM CHARACTERIZATION OF THE TRAINING

High-dimensional limit We analyze the learning problem (327) in the limit where the embedding dimension d and the number of training samples n jointly tend to infinity, while their ratio $\alpha = {}^{n}/_{d}$ (henceforth referred to as the sample complexity) stays of order $\Theta_d(1)$. We further assume the sentence length L, the ranks r_s, r_t of the weights $\boldsymbol{Q}, \boldsymbol{Q}_\star$, and the norm of the positional embeddings $\|\boldsymbol{p}\|$, to be $\Theta_d(1)$. We consider this limit as it permits a closed-form characterization presented in the next section. At the same time, this asymptotic limit exhibits rich learning phenomenology closely related to the experimental observations reported in Section 12.1, and which we further explore in Section (12.4).

The main technical result of the present work is a closed-formed characterization of the test MSE (328) and training loss (327) achieved in the high-dimensional limit when training the model (326) via the empirical risk minimization of (327).

Assumption 12.3.1. *The covariances $\{\boldsymbol{\Sigma}_\ell\}_{\ell=1}^L$ admit a common set of eigenvectors $\{\boldsymbol{e}_i\}_{i=1}^d$. We further note $\{\lambda_i^\ell\}_{i=1}^d$ the eigenvalues of Σ_ℓ. The eigenvalues $\{\lambda_i^\ell\}_{i=1}^d$ and the projection of the positional embedding $\boldsymbol{p}_\ell$ and the teacher rows $\boldsymbol{Q}_j^\star$ on the eigenvectors $\{\boldsymbol{e}_i^\top \boldsymbol{p}_\ell\}_{i,\ell}$, $\{\boldsymbol{e}_i^\top \boldsymbol{Q}_j^\star\}_{i,j}$ are assumed to admit a well-defined joint distribution ν as $d \to \infty$ – namely, for $\gamma = (\gamma_1, ..., \gamma_L) \in \mathbb{R}^L$,$\pi = (\pi_1, ..., \pi_t) \in \mathbb{R}^t$ and $\tau = (\tau_1, ..., \tau_L) \in \mathbb{R}^L$:*

$$\frac{1}{d}\sum_{i=1}^{d}\prod_{k=1}^{K}\prod_{j=1}^{t}\delta\left(\lambda_i^\ell - \gamma_\ell\right)\delta\left(\sqrt{d}\boldsymbol{e}_i^\top \boldsymbol{p}_\ell - \tau_\ell\right)\delta\left(\boldsymbol{e}_i^\top \boldsymbol{Q}_j^\star - \pi_j\right) \xrightarrow{d\to\infty} \nu\left(\gamma, \tau\right). \tag{329}$$

Moreover, the marginals ν_γ (resp. ν_τ) are assumed to have a well-defined first (resp. second) moment.

Result 12.3.2. *Under Assumption 12.3.1, in the limit $n, d \to \infty$, $\|\boldsymbol{p}_\ell\|, {}^n/_d, L, r_s, r_t = \Theta_d(1)$, the summary statistics*

$$\rho_\ell \equiv \frac{\boldsymbol{Q}_\star^\top \Sigma_\ell \boldsymbol{Q}_\star}{d} \in \mathbb{R}^{r_t \times r_t}, \qquad q_\ell \equiv \frac{\hat{\boldsymbol{Q}}^\top \Sigma_\ell \hat{\boldsymbol{Q}}}{d} \in \mathbb{R}^{r_s \times r_s},$$
$$m_\ell \equiv \frac{\hat{\boldsymbol{Q}}^\top \boldsymbol{p}_\ell}{d} \in \mathbb{R}^{r_s}, \qquad \theta_\ell \equiv \frac{\hat{\boldsymbol{Q}}^\top \Sigma_\ell \boldsymbol{Q}_\star}{d} \in \mathbb{R}^{r_s \times r_t} \tag{330}$$

concentrate in probability, and are solutions of the set of finite-dimensional self-consistent equations

$$\begin{cases} q_\ell = \int d\nu(\gamma,\tau,\pi)\gamma_\ell \left(\lambda \mathbb{I}_r + \sum\limits_{\kappa=1}^{L} \gamma_\kappa \hat{V}_\kappa\right)^{-1} \\ \qquad \left(\sum\limits_{\kappa=1}^{L} \gamma_\kappa \hat{q}_\kappa + \left(\sum\limits_{\kappa=1}^{L} \hat{m}_\kappa \tau_\kappa + \gamma_\kappa \hat{\theta}_\kappa \cdot \pi\right)^{\otimes 2}\right) \left(\lambda \mathbb{I}_r + \sum\limits_{\kappa=1}^{L} \gamma_\kappa \hat{V}_\kappa\right)^{-1} \\ V_\ell = \int d\nu(\gamma,\tau,\pi)\gamma_\ell \left(\lambda \mathbb{I}_r + \sum\limits_{\kappa=1}^{L} \gamma_\kappa \hat{V}_\kappa\right)^{-1} \\ m_\ell = \int d\nu(\gamma,\tau,\pi)\tau_\ell \left(\lambda \mathbb{I}_r + \sum\limits_{\kappa=1}^{L} \gamma_\kappa \hat{V}_\kappa\right)^{-1} \\ \qquad \left(\sum\limits_{\kappa=1}^{L} \hat{m}_\kappa \tau_\kappa + \gamma_\kappa \hat{\theta}_\kappa \cdot \pi\right) \\ \theta_\ell = \int d\nu(\gamma,\tau,\pi)\gamma_\ell \left(\lambda \mathbb{I}_r + \sum\limits_{\kappa=1}^{L} \gamma_\kappa \hat{V}_\kappa\right)^{-1} \\ \qquad \left(\sum\limits_{\kappa=1}^{L} \hat{m}_\kappa \tau_\kappa + \gamma_\kappa \hat{\theta}_\kappa \cdot \pi\right) \pi^\top. \end{cases} \tag{331}$$

$$\begin{cases} \hat{q}_\ell = \alpha \mathbb{E}_{\Xi,U} V_\ell^{-1} \left(prox(\Xi,U)_\ell - q_\ell^{\frac{1}{2}} \xi_\ell - m_\ell\right)^{\otimes 2} V_\ell^{-1} \\ \hat{V}_\ell = \hat{\theta}_\ell \theta_\ell^\top q_\ell^{-1} - \alpha \mathbb{E}_{\Xi,U} V_\ell^{-1} \left(prox(\Xi,U)_\ell - q_\ell^{\frac{1}{2}} \xi_\ell - m_\ell\right) \xi_\ell^\top q_\ell^{-\frac{1}{2}} \\ \hat{m}_\ell = \alpha \mathbb{E}_{\Xi,U} V_\ell^{-1} \left(prox(\Xi,U)_\ell - q_\ell^{\frac{1}{2}} \xi_\ell - m_\ell\right) \\ \hat{\theta}_\ell = \alpha \mathbb{E}_{\Xi,U} V_\ell^{-1} \left(prox(\Xi,U)_\ell - q_\ell^{\frac{1}{2}} \xi_\ell - m_\ell\right) \\ \qquad \left(u_\ell - \xi_\ell^\top q_\ell^{-1/2} \theta_\ell\right)^\top \left(\rho_\ell - \theta_\ell^\top q_\ell^{-1} \theta_\ell\right)^{-1} \end{cases} \tag{332}$$

In (331), $U = \{u_\ell\}_{\ell=1}^L$ and $\Xi = \{\xi_\ell\}_{\ell=1}^L$, with $u_\ell \sim \mathcal{N}(\xi_\ell^\top q_\ell^{-1/2} \theta_\ell, \rho_\ell - \theta_\ell^\top q_\ell^{-1} \theta_\ell)$ and $\xi_\ell \sim \mathcal{N}(0, \mathbb{I}_r)$, and $\cdot^{\otimes 2}$ denotes the outer product of a vector with itself.

Finally, the resolvents $\{prox(\Xi, U)_\ell\}_{\ell=1}^L$ are defined as the minimizers of the Moreau envelope

$$\begin{aligned}
&\mathscr{M}(\Xi, U) = \\
&\inf_{z_1,\dots,z_L} \frac{1}{2}\left\{ \sum_{\ell=1}^{L} \mathrm{Tr}\left[V_\ell^{-1}\left(x_\ell - q_\ell^{1/2}\xi_\ell - m_\ell \right)^{\otimes 2} \right] + \mathrm{Tr}\left[\mathtt{S}(Z)\rho_\Sigma \mathtt{S}(Z)^\top \right] \right. \\
&\left. - 2\,\mathrm{Tr}\left[\mathtt{T}(U)\rho_\Sigma \mathtt{S}(Z)^\top \right] \right\}.
\end{aligned} \tag{333}$$

We noted $Z \in \mathbb{R}^{L\times r_s}$ (resp. $U \in \mathbb{R}^{L\times r_t}$ the matrix whose rows are z_ℓ (resp. u_ℓ). In (333),

$$\rho_\Sigma \equiv \mathrm{diag}\left[\left(\int d\nu(\gamma,\tau)\gamma_\ell \right)_{\ell=1}^{L} \right] \in \mathbb{R}^{L\times L}. \tag{334}$$

In the same limit, the test error (328) converges in probability to

$$\begin{aligned}
\epsilon_g =& \mathbb{E}_h \mathrm{Tr}\left[\mathtt{S}[h]\rho_\Sigma \mathtt{S}[h]^\top \right] + \mathbb{E}_{h^\star} \mathrm{Tr}\left[\mathtt{T}[h^\star]\rho_\Sigma \mathtt{T}[h^\star]^\top \right] \\
&- 2\mathbb{E}_{h,h^\star} \mathrm{Tr}\left[\mathtt{S}[h]\rho_\Sigma \mathtt{T}[h^\star]^\top \right].
\end{aligned} \tag{335}$$

where the average bears on $h \in \mathbb{R}^{L\times r_s}, h^\star \in \mathbb{R}^{L\times r_t}$ with independent rows with statistics

$$(h_\ell, h_\ell^\star) \sim \mathcal{N}\left[\begin{pmatrix} m_\ell \\ 0 \end{pmatrix}, \left(\begin{array}{c|c} q_\ell & \theta_\ell \\ \hline \theta_\ell^\top & \rho_\ell \end{array} \right) \right] \tag{336}$$

Finally, the training loss ϵ_t converges in probability to

$$\begin{aligned}
\epsilon_t =& \alpha \mathbb{E}_{Y,\Xi} \mathscr{M} - \frac{1}{2}\sum_{\ell=1}^{L} \mathrm{Tr}[\hat{q}_\ell V_\ell] \\
&+ \frac{\lambda}{2} \int d\nu(\gamma,\tau)\, \mathrm{Tr}\left[\left(\lambda + \sum_{\ell=1}^{L} \gamma_\ell \hat{V}_\ell \right)^{-1} \left(\sum_{\ell=1}^{L} \gamma_\ell \hat{q}_\ell + \left(\sum_{\ell=1}^{L} \tau_\ell \hat{m}_\ell + \hat{\theta}_\ell \cdot \pi \right)^{\otimes 2} \right) \right].
\end{aligned} \tag{337}$$

The derivation of Result 12.3.2 is exploiting a mapping of the model (326) to a (variant of) a GLM (Nelder et al., 1972; McCullagh, 2019). The summary statistics characterized by the equations (331) (often called SE (Javanmard et al., 2013) in this context) asymptotically describe the fixed points of a GAMP algorithm (Rangan et al., 2016). The stable fixed points of GAMP in turn correspond to critical (zero-gradient) points of the non-convex empirical loss landscape (327). Therefore, while Result (12.3.2) is stated as a characterization of the global minimum of (327), which is the main concern of the present work, solutions of (331) also describe local minima.

This strategy to study asymptotics of high-dimensional problem has been used in many recent work, see e.g. (Bayati et al., 2011b; Donoho et al., 2016;

Emami et al., 2020; Loureiro et al., 2021a; Gerbelot et al., 2022). We note, however, that we importantly assume the point-wise convergence of GAMP. While we believe that this point can be rigorously justified, it would require a considerable amount of work —in particular, the usual rigorous tools used in recent works fall short because of the non-convexity of the loss— and we leave this point for further studies. Again, we mention that while Result 12.3.2 is presented for an ℓ_2 regularization of the empirical loss (327) for clarity, similar results can be reached for generic convex regularizers, following the lines of the analysis presented in the introductory Part I. In the following section, we explore the phenomenology uncovered from the study of the equations (331) of Result 12.3.2, for the special case of dot-product attention.

12.4 POSITIONAL-TO-SEMANTIC PHASE TRANSITION

12.4.1 RANK ONE DOT-PRODUCT ATTENTION

In the following, we turn to a special case of tied low-rank attention (326), which exhibits a similar phenomenology as the histogram task empirically probed in Section 12.1 – namely a dot-product attention layer:

$$\mathtt{S}\left[\frac{1}{\sqrt{d}}(\boldsymbol{x}+\boldsymbol{p})\boldsymbol{Q}\right] = \mathrm{softmax}\left(\frac{1}{d}(\boldsymbol{x}+\boldsymbol{p})\boldsymbol{Q}\boldsymbol{Q}^\top(\boldsymbol{x}+\boldsymbol{p})^\top\right). \tag{338}$$

As in (326), we allow for positional encodings $\boldsymbol{p}$ in the dot-product attention parametrization (338). We further consider a specific case of target attention matrix (325) of the form

$$\mathtt{T}\left[\frac{1}{\sqrt{d}}\boldsymbol{x}\boldsymbol{Q}_\star\right] = (1-\omega)\mathrm{softmax}\left(\frac{1}{d}\boldsymbol{x}\boldsymbol{Q}_\star\boldsymbol{Q}_\star^\top\boldsymbol{x}^\top\right) + \omega A\cdot \tag{339}$$

with $A \in \mathbb{R}^{L\times L}$ a fixed matrix. In (339), the parameter $\omega \in [0,1]$ tunes the relative strength of the dot-product term and the fixed matrix term, and interpolates between a fully positional and a fully semantic task:

- For $\omega = 0$, the target reduces to the first dot-product term, and is purely semantic, in that the i, j—th element of the score matrix $\mathrm{softmax}(1/d\boldsymbol{x}\boldsymbol{Q}_\star\boldsymbol{Q}_\star^\top\boldsymbol{x}^\top)$ only depends on the tokens $\boldsymbol{x}_i,\boldsymbol{x}_j$ and not explicitely on their respective placements i, j inside the sentence. To learn satisfyingly the target, the learning model thus has to learn a *semantic* attention matrix.

- For $\omega = 1$, the target reduces to the second fixed term A in (339). The attention matrix A associated thereto is purely positional, in the sense that A_{ij} is a function of i, j but not of $\boldsymbol{x}_i,\boldsymbol{x}_j$. To complete the learning task, a *positional* mechanism thus needs to be learnt.

The parameter ω thus allows to tune the extent to which the task requires the model to implement semantic attention (small ωs) or rather positional

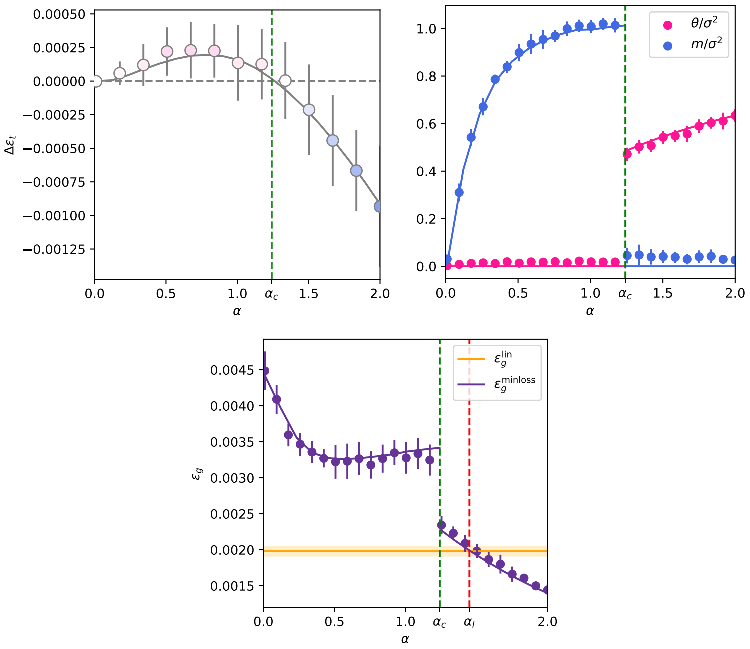


Figure 41: Mixed positional/semantic teacher for $\omega = 0.3$. Setting is $r_s = r_t = 1, L = 2, A = ((0.6, 0.4), (0.4, 0.6)), \Sigma_1 = \Sigma_2 = 0.25\mathbb{I}_d$, $\boldsymbol{p}_1 = \mathbf{1}_d = -\boldsymbol{p}_2$ and $\boldsymbol{Q}_\star \sim \mathcal{N}(0, \mathbb{I}_d)$. **(left)** Solid lines: difference in training loss $\Delta\epsilon_t$ between the semantic and positional solutions of (331) in Result 12.3.2. Markers: difference in training loss at convergence achieved by training the model (326) using gradient descent initialized resp. at $\boldsymbol{Q}_\star$ and at $\boldsymbol{p}_1$. Marker color according to phase diagram in Fig. 42. **(middle)** (blue) overlap θ between the learnt weights $\hat{\boldsymbol{Q}}$ and the target weights $\boldsymbol{Q}_\star$ (red) overlap m between the learnt weights $\hat{\boldsymbol{Q}}$ and the positional embedding p_1. Solid lines represent the theoretical characterization of these two summary statistics provided by Result 12.3.2. Only the solution of (331) corresponding to the lowest found training loss is represented (respectively the positional solution for $\alpha < \alpha_c$ and the semantic solution for $\alpha > \alpha_c$). Markers represent experimental measures of these quantities, for gradient descent at convergence. Gradient descent was initialized at p_1 for $\alpha < \alpha_c$ and at $\boldsymbol{Q}_\star$ for $\alpha > \alpha_c$. **(right)** Test MSE. Solid lines represent the theoretical characterization of Result 12.3.2. Only the solution corresponding to minimal training loss is represented. Markers indicate the MSE experimentally reached by the model (326) trained using gradient descent, initialized at p_1 for $\alpha < \alpha_c$ and at $\boldsymbol{Q}_\star$ for $\alpha > \alpha_c$. The yellow line represents the MSE achieved by the dense linear baseline (340), as analytically characterized by Result 12.4.1. All experiments were performed in $d = 1{,}000$ with the `Pytorch` implementation of full-batch gradient descent, for $T = 5{,}000$ epochs and learning rate $\eta = 0.15$. All points are averaged over 24 instances of the problem each.

attention (large ωs).

In the following, for definiteness, we further assume $r_s = r_t = 1$ and set $\boldsymbol{Q}_\star$ to be a fixed random Gaussian vector drawn from $\mathcal{N}(0, \mathbb{I}_d)$, and choose the positional encodings $\boldsymbol{p}_1 = -\boldsymbol{p}_2 = \mathbf{1}_d$. Finally, for simplicity, we consider sentences with two tokens $L = 2$ and isotropic token covariances $\Sigma_1 = \Sigma_2 = \sigma^2 \mathbb{1}_d$.

12.4.2 SEMANTIC AND POSITIONAL MECHANISMS

The summary statistics θ_ℓ, m_ℓ describing the global minimizer of the empirical loss minimization (327) of the dot-product attention (338) on the target (339) are captured alongside the corresponding test error (328) and training loss (327), by Result 12.3.2. The solution of the system of equations (331) is not unique, and different stable fixed points describe different corresponding critical points of the non-convex empirical loss landscape (327). In practice, we notably find two solutions of (331), corresponding to two different mechanisms implemented by the dot-product attention (338) when approximating the target (339):

- *Positional solution*– One solution of (331) correspond to vanishing overlap $\theta = 0$ between the trained weights $\hat{\boldsymbol{Q}}$ and the semantic target weights $\boldsymbol{Q}_\star$, and non-zero $m > 0$ between the trained weights $\hat{\boldsymbol{Q}}$ and the positional embedding $\boldsymbol{p}_1 = -\boldsymbol{p}_2$. Consequently, the argument of the dot-product attention $\hat{\boldsymbol{Q}}(\boldsymbol{x} + \boldsymbol{p})$ has a sizeable token-independent –thus positional– contribution $\hat{\boldsymbol{Q}}\boldsymbol{p}$, alongside a token-dependent semantic part $\hat{\boldsymbol{Q}}\boldsymbol{x}$. Because of the positional term, the elements of the resulting learnt attention attention matrix $\mathrm{softmax}({}^1\!/\!_d(\boldsymbol{x} + \boldsymbol{p})\hat{\boldsymbol{Q}}\hat{\boldsymbol{Q}}^\top(\boldsymbol{x} + \boldsymbol{p})^\top)$ implement a partly positional mechanism.

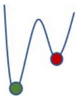
- *Semantic solution*– Another solution of the system of equations (331) is associated to vanishing overlap $m = 0$ between the learnt weights $\hat{\boldsymbol{Q}}$ and the positional embeddings, and finite overlap $\theta > 0$ with the target weights $\boldsymbol{Q}_\star$. Therefore the resulting learnt attention matrix $\mathrm{softmax}({}^1\!/\!_d(\boldsymbol{x} + \boldsymbol{p})\hat{\boldsymbol{Q}}\hat{\boldsymbol{Q}}^\top(\boldsymbol{x} + \boldsymbol{p})^\top) \approx \mathrm{softmax}({}^1\!/\!_d\boldsymbol{x}\hat{\boldsymbol{Q}}\hat{\boldsymbol{Q}}^\top\boldsymbol{x}^\top)$ is largely semantic.

While the system of self-consistent equations (331) may admit other solutions, we did not find solutions with lower training loss than the two aforedescribed fixed points. Which of these solution corresponds to the global minimum – and thus the solution of the optimization (327)– depends on the sample complexity α and the positional/semantic parameter ω (339), as we describe in the following subsection.

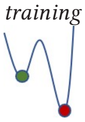
12.4.3 POSITIONAL-TO-SEMANTIC PHASE TRANSITION

For a fixed parameter ω in (339), an analysis of equations (331) reveals that for a sizeable range of ω, in the probed setups, there exists a threshold α_c for the sample complexity so that

- For $\alpha < \alpha_c$, the global minimum of (327) corresponds to a positional mechanism, and is described by the positional solution of (331) of Result 12.3.2 with $\theta = 0, m > 0$.
- For $\alpha > \alpha_c$, the global minimum of (327) corresponds to a semantic mechanism, and is described by the semantic solution of (331) of Result 12.3.2 with $\theta > 0, m = 0$.



For $\alpha < \alpha_c$, the positional minimum (green) is lower in training loss.



For $\alpha > \alpha_c$, the semantic minimum (red) becomes global.

The dot-product attention thus displays a phase transition in sample complexity from a positional mechanism to a semantic mechanism, implementing the simpler positional mechanism when having access to small amounts of data, and only learning the semantic content of the target (339) when presented sufficient data. The critical sample complexity α_c generically grows with the positionality ω of the target function (339), as the semantic content – i.e. the first term of (339)– is less apparent for larger ω, and thus require larger amounts of data to be identified and approximated by the dot-product attention (338). An example for $\omega = 0.3$ is given in Fig. 41.

Algorithmically, the positional minimum can be reached for $\alpha < \alpha_c$ by gradient descent by initializing the weights $\boldsymbol{Q}$ of the attention (338) close to the positional embedding $\boldsymbol{p}_1$. By the same token, the semantic minimum can be reached by gradient descent from an initialization at the teacher weights $\boldsymbol{Q}_\star$ (339). Henceforth, we refer with a slight abuse to the minimum experimentally reached from a positional (resp. semantic) initialization as the positional (resp. semantic) minimum, even when it is not global. Finally, note importantly that the semantic initialization is informed by nature, in that it necessitates the knowledge of the target parameters $\boldsymbol{Q}_\star$. A precise analysis of the dynamics of gradient descent from an agnostic (random) initialization, and ascertaining whether the optimizer reaches the global minimum, is an interesting question which falls out of the scope of the present manuscript – whose aim is rather to provide a characterization of the global minimum alone.

The difference in training loss $\Delta\epsilon_t$ between the positional and semantic solutions of (331) is represented in Fig. 42, alongside the difference in training loss at convergence experimentally reached by gradient descent from a positional ($\boldsymbol{Q} = \boldsymbol{p}_1$) and semantic ($\boldsymbol{Q} = \boldsymbol{Q}_\star$) initializations. For small (resp. large) sample complexity $\alpha < \alpha_c$ (resp. $\alpha > \alpha_c$), the training loss of the positional (resp. semantic) minimum is lower, and thus corresponds to the global minimum.

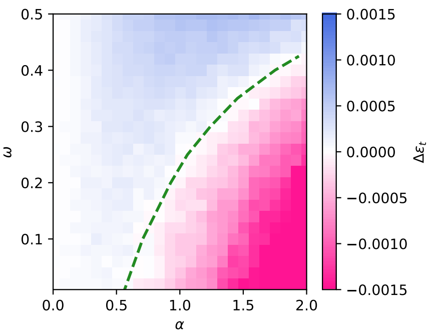


Figure 42: Phase transition between semantic and positional training loss. and$r_s = r_t = 1, L = 2, A = ((0.6, 0.4), (0.4, 0.6)), \Sigma_1 = \Sigma_2 = 0.25\mathbb{I}_d$, $\boldsymbol{p}_1 = \mathbf{1} = -\boldsymbol{p}_2$ and $\boldsymbol{Q}_\star \sim \mathcal{N}(0, \mathbb{I}_d)$. The color map represents the difference in training loss at convergence when training the model (326) using the Pytorch implementation of full-batch gradient descent, respectively from an initialization at p_1 and an initialization at $\boldsymbol{Q}_\star$. The green dashed lines represents the theoretical prediction for the threshold $\alpha_c(\omega)$ above which the semantic solution of (331) in Result 12.3.2 has lower loss than the positional solution. Experiments were performed as in Fig. 41.

The analytical equations (331) are observed to capture well the difference in training loss between both minima (global and local) across the whole range of probed sample complexities, see Fig. 41. Finally, Fig. 41 (middle, right) presents the theoretical predictions of Result 12.3.2 for the weights/target and weights/embedding overlaps θ, m and the generalization MSE ε_g achieved at the global minimum of the loss landscape (327). These analytical characterizations are compared with experimental estimates of the same metrics obtained by optimizing (327) with the Pytorch (Paszke et al., 2019a) implementation of gradient descent, from a positional (resp. semantic) initialization for $\alpha < \alpha_c$ (resp. $\alpha > \alpha_c$), displaying overall good agreement.

The dot-product attention (338) thus implements a semantic mechanism when learning from sufficient amounts of data. The learning of the semantic mechanism by the dot-product attention at sample complexities $\alpha > \alpha_c$ corresponds to a noticeable drop in the generalization MSE as can be observed in Fig. 41, right. But just how essential is the learning of semantic mechanism in the ability of the dot-product attention to generalize well? We explore this question in the following subsection, by comparing the dot-product attention (338) to a purely positional attention model.

12.4.4 PURELY POSITIONAL BASELINE

In this subsection, for the same target (339), we contrast the dot-product attention model (338), analyzed in the previous subsections, to the baseline given by a linear attention layer:

$$f_W(\boldsymbol{x}) = W \cdot \boldsymbol{x}, \tag{340}$$

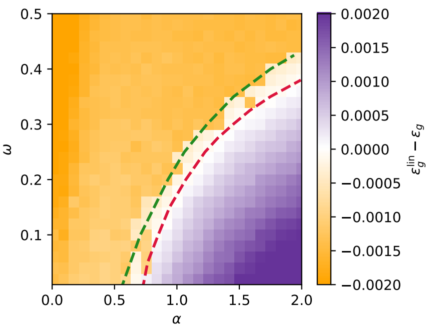


Figure 43: Phase transition between semantic and positional training loss. $r_s = r_t = 1, L = 2, A = ((0.6, 0.4), (0.4, 0.6)), \Sigma_1 = \Sigma_2 = 0.25\mathbb{I}_d, \boldsymbol{p}_1 = \boldsymbol{1} = -\boldsymbol{p}_2$ and $\boldsymbol{Q}_\star \sim \mathcal{N}(0, \mathbb{I}_d)$. The color map represents the difference in test MSE at convergence when training the model (338) using the Pytorch implementation of full-batch gradient descent initialized at $\boldsymbol{Q}_\star$, and the dense linear baseline (340). The red dashed lines indicate the theoretical prediction –following from Result 12.3.2 and Result 340– for the threshold sample complexity α_l above which the dot-product attention (326) outperforms the baseline (340). For comparison, the positional-to-semantic threshold $\alpha_c(\omega)$ is reminded in green. Experiments were performed as in Fig. 41.

with a trainable weights matrix $W \in \mathbb{R}^{L \times L}$. As for the dot-product attention (338), we consider the case where the weights $\hat{W}$ are learnt by minimizing the empirical risk

$$\hat{W} = \underset{W \in \mathbb{R}^{L \times L}}{\operatorname{argmin}} \sum_{\mu=1}^{n} \| y(\boldsymbol{x}^\mu) - f_W(\boldsymbol{x}^\mu) \|^2 \tag{341}$$

The model (340) is a natural counterpart to the dot-product architecture (338). In (340), the attention matrix is parametrized by a single fully-trainable matrix W, instead of being parametrized as a dot-product attention as in (338). A seminal difference in the two parametrizations is that while the elements of softmax$({}^1/_d \boldsymbol{x}\boldsymbol{Q}\boldsymbol{Q}^\top \boldsymbol{x}^\top)$ can depend on the input tokens $\boldsymbol{x}$, and therefore express semantic information, the elements W_{ij} of W can only depend on the positions i, j. The model (340) can thus only implement *positional mechanisms*, while the dot-product attention (338) can implement both linear and semantic mechanisms, as discussed above. Finally, observe that the model (340) is closely related to the one analyzed by (Rende et al., 2023) in another asymptotic limit. The following result characterizes the test error achieved by the purely positional model (340):

Result 12.4.1. *In the same asymptotic limit as Result* (12.3.2)*, the learnt weights* $\hat{W}$ *trained by minimizing the empirical risk* (341) *coincide with the minimizer of the population square risk, and thus admit the compact expression*

$$\hat{W} = \mathbb{E}_{\boldsymbol{x}} \mathrm{T}\left[\frac{1}{\sqrt{d}} \boldsymbol{x} \boldsymbol{Q}_\star \right] = \mathbb{E}_h \mathrm{T}[h] \tag{342}$$

where the average bears over a finite-dimensional matrix $h \in \mathbb{R}^{L\times t}$ *with independent rows* h_ℓ *with statistics* $h_\ell \sim \mathcal{N}(0, \rho_\ell)$, *where* ρ_ℓ *was defined in* (330) *in Result* (12.3.2). *We remind that* $\mathtt{T}\left[{}^1\!/\!\sqrt{d}\boldsymbol{x}\boldsymbol{Q}_\star\right]$ *corresponds to the target score matrix* (325). *Finally, the test MSE* ${}^1\!/\!{}_{dL}\mathbb{E}_{\boldsymbol{x}}\|y(\boldsymbol{x}) - f_{\hat{W}}(\boldsymbol{x})\|^2$ *achieved by the trained dense linear model* $f_{\hat{W}}$ (340) *admits the asymptotic characterization*

$$\varepsilon_g^{\text{lin}} = \frac{1}{L}\operatorname{Tr}\left[\hat{W}\rho_\Sigma\hat{W}^\top\right] + \frac{1}{L}\mathbb{E}_h \operatorname{Tr}\left[\mathtt{T}[h]\rho_\Sigma\mathtt{T}[h]^\top\right] - \frac{2}{L}\mathbb{E}_h \operatorname{Tr}\left[\hat{W}\rho_\Sigma\mathtt{T}[h]^\top\right]. \tag{343}$$

The MSE achieved by the baseline (340) when learning the target (339) is plotted in Fig. 41 (right) as the orange solid line, alongside the MSE achieved by the dot-product attention (338) discussed in previous subsections. Remarkably, in the setup of Fig. 41, in the positional regime $\alpha < \alpha_c$ when the dot-product attention relies on a positional mechanism $\theta = 0, m > 0$ to approximate the target, the dot-product attention (338) is outperformed by the purely positional attention (340) $\varepsilon_g > \varepsilon_g^{\text{lin}}$. In contrast, in the semantic regime $\alpha > \alpha_c$ where the dot-product attention learns the semantic mechanism, there exists a sample complexity $\alpha_l \geq \alpha_c$ above which $\varepsilon_g < \varepsilon_g^{\text{lin}}$, i.e. the dot-product attention (338) outperforms the dense linear baseline (340). This threshold value α_l is plotted for various positionality strengths ω in Fig. 43, alongside the positional-to-semantic threshold α_c. Interestingly, we observe $\alpha_l \geq \alpha_c$ in all probed settings, temptingly suggesting the natural interpretation that the dot-product attention needs to learn the semantic mechanism first (at $\alpha = \alpha_c$) in order to then be able to outperform the best positional approximation $f_{\hat{W}}$ (at $\alpha = \alpha_l$). This highlights the importance of the semantic mechanism, enabled by the dot-product parametrization (338), in learning targets with semantic content such as (339).

CONCLUSION

We explored the interplay between positional and semantic attention, both through an empirical example and the prism of tied low-rank self-attention in high dimensions. In the empirical setting we showed how a simple algorithmic counting task can be solved using a positional or semantic mechanism in the attention layer. For a different task, in a theoretically controlled setting, we characterized the global optimum of the empirical loss, when learning a target attention layer. This global optimum was found to correspond to either a positional or a semantic mechanism, with a phase transition between the two mechanisms occurring as the sample complexity increases. We believe the present asymptotic analysis of the inner workings of attention mechanisms opens up exciting research directions. Considering untied query and key matrices, appending a readout network after the attention layer, or addressing more practical training procedures such as masked language modelling, are some possible extensions which will hopefully pave the way towards a satisfactory theoretical comprehension of attention mechanisms.

Part V

BIBLIOGRAPHY

BIBLIOGRAPHY

[1] Emmanuel Abbe, Enric Boix Adsera, and Theodor Misiakiewicz. '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. 2022, pp. 4782–4887 (cit. on p. 168).

[2] Emmanuel Abbe, Enric Boix-Adsera, Matthew S Brennan, Guy Bresler, and Dheeraj Nagaraj. 'The staircase property: How hierarchical structure can guide deep learning.' In: *Advances in Neural Information Processing Systems* 34 (2021), pp. 26989–27002 (cit. on p. 168).

[3] Emmanuel Abbe, Enric Boix-Adsera, and Theodor Misiakiewicz. *SGD learning on neural networks: leap complexity and saddle-to-saddle dynamics.* 2023. arXiv: *2302.11055* [cs.LG] (cit. on pp. 36, 164, 166, 168).

[4] Radosław Adamczak. 'A note on the hanson-wright inequality for random vectors with dependencies.' In: *Electron. Commun. Prob.* 20 (2015). arXiv: *1409.8457* (cit. on p. 131).

[5] Ben Adlam, Jaehoon Lee, Shreyas Padhy, Zachary Nado, and Jasper Snoek. 'Kernel Regression with Infinite-Width Neural Networks on Millions of Examples.' In: *arXiv preprint arXiv:2303.05420* (2023) (cit. on p. 164).

[6] Ben Adlam and Jeffrey Pennington. 'The Neural Tangent Kernel in High Dimensions: Triple Descent and a Multi-Scale Theory of Generalization.' Preprint. 2020. arXiv: *2008.06786* (cit. on pp. 129–131).

[7] Ben Adlam and Jeffrey Pennington. 'Understanding double descent requires a fine-grained bias-variance decomposition.' In: *Advances in neural information processing systems* 33 (2020), pp. 11022–11032 (cit. on p. 106).

[8] Madhu Advani and Surya Ganguli. 'Statistical mechanics of optimal convex inference in high dimensions.' In: *Physical Review X* 6.3 (2016), p. 031034 (cit. on p. 147).

[9] Madhu S Advani, Andrew M Saxe, and Haim Sompolinsky. 'High-dimensional dynamics of generalization error in neural networks.' In: *Neural Networks* 132 (2020), pp. 428–446 (cit. on pp. 66, 86).

[10] Michael Albergo, Nicolas Boffi, and Eric Vanden-Eijnden. 'Stochastic Interpolants: A Unifying Framework for Flows and Diffusions.' In: *ArXiv* abs/2303.08797 (2023) (cit. on pp. 199, 200, 202, 203).

[11] Michael Albergo and Eric Vanden-Eijnden. 'Building normalizing flows with stochastic interpolants.' In: *arXiv preprint arXiv:2209.15571* (2022) (cit. on pp. 199, 200, 202).

[12] Mathieu Andreux et al. 'Kymatio: Scattering Transforms in Python.' In: *Journal of Machine Learning Research* 21.60 (2020), pp. 1–6 (cit. on pp. 45, 47).

[13] Navid Ardeshir, Clayton Sanford, and Daniel Hsu. 'Support vector machines and linear regression coincide with very high-dimensional features.' In: *ArXiv* abs/2105.14084 (2021) (cit. on p. 93).

[14] S Ariosto, R Pacelli, M Pastore, F Ginelli, M Gherardi, and P Rotondo. 'Statistical mechanics of deep learning beyond the infinite-width limit.' In: *arXiv preprint arXiv:2209.04882* (2022) (cit. on pp. 36, 86).

[15] S. Ariosto, R. Pacelli, M. Pastore, F. Ginelli, M. Gherardi, and P. Rotondo. 'Statistical mechanics of deep learning beyond the infinite-width limit.' In: *arXiv:2209.04882* (2022) (cit. on pp. 147, 150).

[16] Luca Arnaboldi, Florent Krzakala, Bruno Loureiro, and Ludovic Stephan. *Escaping mediocrity: how two-layer networks learn hard single-index models with SGD.* 2023. arXiv: *2305.18502* [stat.ML] (cit. on p. 168).

[17] Luca Arnaboldi, Ludovic Stephan, Florent Krzakala, and Bruno Loureiro. *From high-dimensional & mean-field dynamics to dimensionless ODEs: A unifying approach to SGD in two-layers networks.* arXiv:2302.05882 [cond-mat, stat] type: article. Feb. 2023 (cit. on p. 168).

[18] Sanjeev Arora, Simon Shaolei Du, Zhiyuan Li, Ruslan Salakhutdinov, Ruosong Wang, and Dingli Yu. 'Harnessing the Power of Infinitely Wide Deep Nets on Small-data Tasks.' In: *Proc. Int. Conf. Learning Rep. (ICLR)* (2020) (cit. on p. 160).

[19] Benjamin Aubin, Florent Krzakala, Yue Lu, and Lenka Zdeborová. 'Generalization error in high-dimensional perceptrons: Approaching bayes error

with convex optimization.' In: *Advances in Neural Information Processing Systems* 33 (2020), pp. 12199–12210 (cit. on pp. 11, 16, 34, 37, 40, 45, 48, 97, 145, 147, 160, 186).

[20] Benjamin Aubin, Bruno Loureiro, Antoine Baker, Florent Krzakala, and Lenka Zdeborová. 'Exact asymptotics for phase retrieval and compressed sensing with random generative priors.' In: *Proceedings of The First Mathematical and Scientific Machine Learning Conference*. Ed. by Jianfeng Lu and Rachel Ward. Vol. 107. Proceedings of Machine Learning Research. PMLR, 20–24 Jul 2020, pp. 55–73 (cit. on pp. 106, 126).

[21] Benjamin Aubin, Bruno Loureiro, Antoine Maillard, Florent Krzakala, and Lenka Zdeborová. 'The spiked matrix model with generative priors.' In: *Advances in Neural Information Processing Systems*. Vol. 32. Curran Associates, Inc., 2019 (cit. on pp. 106, 126).

[22] Benjamin Aubin, Antoine Maillard, Jean Barbier, Florent Krzakala, Nicolas Macris, and Lenka Zdeborová. 'The committee machine: computational to statistical gaps in learning a two-layers neural network.' In: *Journal of Statistical Mechanics: Theory and Experiment* 2019 (2018) (cit. on pp. 37, 145, 147, 148, 151, 155, 186).

[23] Benjamin Aubin, Antoine Maillard, Florent Krzakala, Nicolas Macris, Lenka Zdeborová, et al. 'The committee machine: Computational to statistical gaps in learning a two-layers neural network.' In: *Advances in Neural Information Processing Systems* 31 (2018) (cit. on pp. 11, 12, 34, 40, 97).

[24] Jean-Yves Audibert and Alexandre B. Tsybakov. 'Fast learning rates for plug-in classifiers.' In: *Annals of Statistics* 35 (2007), pp. 608–633 (cit. on pp. 81–83).

[25] Jimmy Ba, Murat A Erdogdu, Taiji Suzuki, Zhichao Wang, and Denny Wu. 'Learning in the presence of low-dimensional structure: a spiked random matrix perspective.' In: *Conference on Parsimony and Learning (Recent Spotlight Track)*. 2023 (cit. on pp. 164, 178).

[26] Jimmy Ba, Murat A Erdogdu, Taiji Suzuki, Zhichao Wang, Denny Wu, and Greg Yang. 'High-dimensional Asymptotics of Feature Learning: How One Gradient Step Improves the Representation.' In: *Advances in Neural Information Processing Systems*. Ed. by S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh. Vol. 35. Curran Associates, Inc., 2022, pp. 37932–37946 (cit. on pp. 36, 99, 164, 166, 167, 170, 174, 176–178).

[27] Jimmy Ba, Murat A. Erdogdu, Taiji Suzuki, Zhichao Wang, Denny Wu, and Greg Yang. *High-dimensional Asymptotics of Feature Learning: How One Gradient Step Improves the Representation*. 2022 (cit. on p. 138).

[28] Francis Bach. 'High-dimensional analysis of double descent for linear regression with random projections.' Preprint. 2023. arXiv: *2303.01372* (cit. on p. 129).

[29] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 'Neural machine translation by jointly learning to align and translate.' In: *arXiv preprint arXiv:1409.0473* (2014) (cit. on p. 8).

[30] Yasaman Bahri, Ethan Dyer, Jared Kaplan, Jaehoon Lee, and Utkarsh Sharma. 'Explaining neural scaling laws.' In: *arXiv preprint arXiv:2102.06701* (2021) (cit. on pp. 80, 81).

[31] Yu Bai, Fan Chen, Huan Wang, Caiming Xiong, and Song Mei. 'Transformers as Statisticians: Provable In-Context Learning with In-Context Algorithm Selection.' In: *arXiv preprint arXiv:2306.04637* (2023) (cit. on p. 215).

[32] Z. D. Bai. 'Convergence Rate of Expected Spectral Distributions of Large Random Matrices. Part I. Wigner Matrices.' In: *The Annals of Probability* 21.2 (1993) (cit. on p. 111).

[33] Zhidong Bai and Wang Zhou. 'LARGE SAMPLE COVARIANCE MATRICES WITHOUT INDEPENDENCE STRUCTURES IN COLUMNS.' In: *Statistica Sinica* 18.2 (2008), pp. 425–442 (cit. on p. 131).

[34] Zhidong Bai and Wang Zhou. 'Large sample covariance matrices without independence structures in columns.' In: *Statistica Sinica* (2008), pp. 425–442 (cit. on p. 59).

[35] Zhidong Bai and Wang Zhou. 'Large sample covariance matrices without independence structures in columns.' In: *Statist. Sinica* 18.2 (2008), pp. 425–442 (cit. on pp. 107, 108, 111).

[36] Pierre Baldi and Kurt Hornik. 'Neural networks and principal component analysis: Learning from examples without local minima.' In: *Neural Networks* 2 (1989), pp. 53–58 (cit. on pp. 184, 185, 193, 196).

[37] Xuchan Bao, James Lucas, Sushant Sachdeva, and Roger B Grosse. 'Regularized linear autoencoders recover the principal components, eventually.' In:

Advances in Neural Information Processing Systems 33 (2020), pp. 6971–6981 (cit. on p. 184).

[38] Jean Barbier, Jean Barbier, Florent Krzakala, Nicolas Macris, Léo Miolane, and Lenka Zdeborová. 'Optimal errors and phase transitions in high-dimensional generalized linear models.' In: *Proc. Natl. Acad. Sci. U. S. A.* 116.12 (2019), pp. 5451–5460 (cit. on pp. 145, 147, 155, 186).

[39] Jean Barbier, Florent Krzakala, Nicolas Macris, Léo Miolane, and Lenka Zdeborová. 'Optimal errors and phase transitions in high-dimensional generalized linear models.' In: *Proceedings of the National Academy of Sciences* 116.12 (2019), pp. 5451–5460 (cit. on pp. 11, 34, 37, 40, 97).

[40] Peter L. Bartlett, Philip M. Long, Gábor Lugosi, and Alexander Tsigler. 'Benign overfitting in linear regression.' In: *Proceedings of the National Academy of Sciences* 117.48 (2020), pp. 30063–30070. eprint: *https://www.pnas.org/content/117/48/30063.full.pdf* (cit. on p. 49).

[41] Peter L. Bartlett, Philip M. Long, Gábor Lugosi, and Alexander Tsigler. 'Benign overfitting in linear regression.' In: *Proceedings of the National Academy of Sciences* 117.48 (2020), pp. 30063–30070 (cit. on pp. 65, 66, 70, 81, 93, 124).

[42] Peter L. Bartlett, Andrea Montanari, and Alexander Rakhlin. 'Deep learning: a statistical viewpoint.' Preprint. 2021. arXiv: *2103.09177* (cit. on p. 129).

[43] Mohsen Bayati and Andrea Montanari. 'The dynamics of message passing on dense graphs, with applications to compressed sensing.' In: *IEEE Transactions on Information Theory* 57.2 (2011), pp. 764–785 (cit. on p. 20).

[44] Mohsen Bayati and Andrea Montanari. 'The LASSO risk for Gaussian matrices.' In: *IEEE Transactions on Information Theory* 58.4 (2011), pp. 1997–2017 (cit. on p. 221).

[45] Mikhail Belkin, Daniel Hsu, Siyuan Ma, and Soumik Mandal. 'Reconciling modern machine-learning practice and the classical bias–variance trade-off.' In: *Proceedings of the National Academy of Sciences* 116.32 (2019), pp. 15849–15854 (cit. on pp. 9, 45, 58, 124).

[46] Mikhail Belkin, Daniel Hsu, and Ji Xu. 'Two models of double descent for weak features.' In: *SIAM Journal on Mathematics of Data Science* 2.4 (2020), pp. 1167–1180 (cit. on pp. 45, 48, 66).

[47] Gerard Ben Arous, Reza Gheissari, and Aukosh Jagannath. 'High-dimensional limit theorems for sgd: Effective dynamics and critical scaling.' In: *Advances in Neural Information Processing Systems* 35 (2022), pp. 25349–25362 (cit. on p. 168).

[48] Florent Benaych-Georges. 'On a surprising relation between the Marchenko-Pastur law, rectangular and square free convolutions.' In: *Ann. Inst. Henri Poincaré Probab. Stat.* 46.3 (2010), pp. 644–652 (cit. on p. 111).

[49] Lucas Benigni and Sandrine Péché. 'Eigenvalue distribution of some nonlinear models of random matrices.' In: *Electron. J. Probab.* 26 (2021), Paper No. 150, 37 (cit. on pp. 106, 125).

[50] Joe Benton, George Deligiannidis, and Arnaud Doucet. 'Error Bounds for Flow Matching Methods.' In: *arXiv preprint arXiv:2305.16860* (2023) (cit. on pp. 199, 201).

[51] Raphael Berthier, F. Bach, and P. Gaillard. 'Tight Nonparametric Convergence Rates for Stochastic Gradient Descent under the Noiseless Linear Model.' In: *ArXiv* abs/2006.08212 (2020) (cit. on pp. 64, 65, 67, 70, 80, 81, 85).

[52] Raphaël Berthier, Andrea Montanari, and Kangjie Zhou. *Learning time-scales in two-layers neural networks.* 2023. arXiv: *2303.00055* [cs.LG] (cit. on pp. 36, 166, 168).

[53] Giulio Biroli and Marc Mézard. 'Generative diffusion in very large dimensions.' In: *arXiv preprint arXiv:2306.03518* (2023) (cit. on p. 201).

[54] Adam Block, Youssef Mroueh, and Alexander Rakhlin. 'Generative modeling with denoising auto-encoders and Langevin sampling.' In: *arXiv preprint arXiv:2002.00107* (2020) (cit. on pp. 199, 201).

[55] Avrim Blum and Ronald Rivest. 'Training a 3-node neural network is NP-complete.' In: *Advances in neural information processing systems* 1 (1988) (cit. on pp. 9, 40).

[56] Antoine Bodin and Nicolas Macris. 'Model, sample, and epoch-wise descents: exact solution of gradient flow in the random feature model.' In: *Advances in Neural Information Processing Systems.* Ed. by M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan. Vol. 34. Curran Associates, Inc., 2021, pp. 21605–21617 (cit. on p. 106).

[57] Enric Boix-Adsera, Etai Littwin, Emmanuel Abbe, Samy Bengio, and Joshua Susskind. 'Transformers learn through gradual rank increase.' In: *arXiv preprint arXiv:2306.07042* (2023) (cit. on p. 215).

[58] Blake Bordelon, Abdulkadir Canatar, and Cengiz Pehlevan. 'Spectrum dependent learning curves in kernel regression and wide neural networks.' In: *International Conference on Machine Learning*. PMLR. 2020, pp. 1024–1034 (cit. on pp. 49, 55, 64, 65, 67, 70, 85, 86, 90, 186).

[59] Blake Bordelon and Cengiz Pehlevan. *Dynamics of Finite Width Kernel and Prediction Fluctuations in Mean Field Neural Networks*. 2023. arXiv: *2304.03408* [stat.ML] (cit. on p. 168).

[60] Blake Bordelon and Cengiz Pehlevan. 'Learning Curves for SGD on Structured Features.' In: *International Conference on Learning Representations*. 2022 (cit. on p. 106).

[61] David Bosch, Ashkan Panahi, and Babak Hassibi. 'Precise Asymptotic Analysis of Deep Random Feature Models.' In: *Proceedings of Thirty Sixth Conference on Learning Theory*. Ed. by Gergely Neu and Lorenzo Rosasco. Vol. 195. Proceedings of Machine Learning Research. PMLR, Dec. 2023, pp. 4132–4179 (cit. on pp. 107, 124, 125, 128, 136, 165, 168).

[62] David Bosch, Ashkan Panahi, and Babak Hassibi. 'Precise Asymptotic Analysis of Deep Random Feature Models.' In: *ArXiv* abs/2302.06210 (2023) (cit. on p. 148).

[63] David Bosch, Ashkan Panahi, Ayca Ozcelikkale, and Devdatt Dubhash. 'Double Descent in Random Feature Models: Precise Asymptotic Analysis for General Convex Regularization.' In: *arXiv:2204.02678* (2022) (cit. on pp. 106, 125).

[64] Hervé Bourlard and Yves Kamp. 'Auto-association by multilayer perceptrons and singular value decomposition.' In: *Biological Cybernetics* 59 (1988), pp. 291–294 (cit. on pp. 184, 185, 196).

[65] Leo Breiman. 'Reflections after refereeing papers for nips.' In: *The Mathematics of Generalization*. CRC Press, 1995, pp. 11–15 (cit. on pp. 4, 9).

[66] J. Bruna and S. Mallat. 'Invariant Scattering Convolution Networks.' In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 35.8 (2013), pp. 1872–1886 (cit. on p. 58).

[67] Z. Burda, A. Görlich, A. Jarosz, and J. Jurkiewicz. 'Signal and noise in correlation matrix.' In: *Phys. A* 343.1-4 (2004), pp. 295–310 (cit. on p. 109).

[68] Francesco Camilli, Daria Tieplova, and Jean Barbier. 'Fundamental limits of overparametrized shallow neural networks for supervised learning.' In: *arXiv preprint arXiv:2307.05635* (2023) (cit. on p. 34).

[69] Abdulkadir Canatar, B. Bordelon, and C. Pehlevan. 'Spectral bias and task-model alignment explain generalization in kernel regression and infinitely wide neural networks.' In: *Nature Communications* 12 (2021), pp. 1–12 (cit. on p. 74).

[70] Abdulkadir Canatar, Abdulkadir Canatar, Blake Bordelon, Cengiz Pehlevan, Blake Bordelon, and Cengiz Pehlevan. 'Spectral bias and task-model alignment explain generalization in kernel regression and infinitely wide neural networks.' In: *Nat. Commun.* 12.1 (2021) (cit. on p. 147).

[71] Emmanuel J Candès, Pragya Sur, et al. 'The phase transition for the existence of the maximum likelihood estimate in high-dimensional logistic regression.' In: *The Annals of Statistics* 48.1 (2020), pp. 27–42 (cit. on pp. 45, 48).

[72] Andrea Caponnetto and Ernesto De Vito. 'Fast Rates for Regularized Least-squares Algorithm.' In: . 2005 (cit. on pp. 41, 64, 65, 68, 70, 80, 81, 84, 92).

[73] Andrea Caponnetto and Ernesto De Vito. 'Optimal rates for the regularized least-squares algorithm.' In: *Foundations of Computational Mathematics* 7.3 (2007), pp. 331–368 (cit. on pp. 41, 55, 65, 67, 68, 80, 81, 84, 92).

[74] Michael Celentano, Andrea Montanari, and Yuting Wei. 'The Lasso with general Gaussian designs with applications to hypothesis testing.' In: *arXiv preprint arXiv:2007.13716* (2020) (cit. on pp. 48, 52).

[75] Hongrui Chen, Holden Lee, and Jianfeng Lu. 'Improved analysis of score-based generative modeling: User-friendly bounds under minimal smoothness assumptions.' In: *International Conference on Machine Learning*. PMLR. 2023, pp. 4735–4763 (cit. on pp. 199, 201).

[76] Minshuo Chen, Kaixuan Huang, Tuo Zhao, and Mengdi Wang. 'Score approximation, estimation and distribution recovery of diffusion models on low-dimensional data.' In: *arXiv preprint arXiv:2302.07194* (2023) (cit. on pp. 199, 201).

[77] Sitan Chen, Sinho Chewi, Holden Lee, Yuanzhi Li, Jianfeng Lu, and Adil Salim. 'The probability flow ODE is provably fast.' In: *arXiv preprint arXiv:2305.11798* (2023) (cit. on pp. 199, 201).

[78] Sitan Chen, Sinho Chewi, Jerry Li, Yuanzhi Li, Adil Salim, and Anru R Zhang. 'Sampling is as easy as learning the score: theory for diffusion models with minimal data assumptions.' In: *arXiv preprint arXiv:2209.11215* (2022) (cit. on pp. 199, 201).

[79] Sitan Chen, Giannis Daras, and Alex Dimakis. 'Restoration-degradation beyond linear diffusions: A non-asymptotic analysis for DDIM-type samplers.' In: *International Conference on Machine Learning*. PMLR. 2023, pp. 4462–4484 (cit. on pp. 199, 201).

[80] Zhengdao Chen, Eric Vanden-Eijnden, and Joan Bruna. 'A Functional-Space Mean-Field Theory of Partially-Trained Three-Layer Neural Networks.' In: *arXiv preprint arXiv:2210.16286* (2022) (cit. on p. 35).

[81] Chen Cheng and Andrea Montanari. 'Dimension free ridge regression.' Preprint. 2022. arXiv: *2210.08571* (cit. on p. 129).

[82] Xiuyuan Cheng and Amit Singer. 'The spectrum of random inner-product kernel matrices.' In: *Random Matrices: Theory and Applications* 2.04 (2013), p. 1350010 (cit. on p. 59).

[83] Lenaic Chizat and Francis Bach. 'On the global convergence of gradient descent for over-parameterized models using optimal transport.' In: *Advances in neural information processing systems* 31 (2018) (cit. on pp. 35, 41, 97, 168).

[84] Lénaïc Chizat, Edouard Oyallon, and Francis Bach. 'On Lazy Training in Differentiable Programming.' In: *Advances in Neural Information Processing Systems*. Vol. 32. Curran Associates, Inc., 2019 (cit. on pp. 34, 164, 166).

[85] Lénaïc Chizat, Edouard Oyallon, and Francis Bach. 'On Lazy Training in Differentiable Programming.' In: *Advances in Neural Information Processing Systems*. Vol. 32. Curran Associates, Inc., 2019 (cit. on p. 64).

[86] Lénaı c Chizat, Edouard Oyallon, and Francis Bach. 'On Lazy Training in Differentiable Programming.' In: *Proceedings of the 33rd International Conference on Neural Information Processing Systems*. Red Hook, NY, USA: Curran Associates Inc., 2019 (cit. on p. 105).

[87] Clément Chouard. 'Deterministic equivalent of the Conjugate Kernel matrix associated to Artificial Neural Networks.' Preprint. 2023. arXiv: *2306.05850* (cit. on pp. 135, 136).

[88] Clément Chouard. 'Quantitative deterministic equivalent of sample covariance matrices with a general dependence structure.' In: *arXiv:2211.13044* (2022) (cit. on pp. 107, 109, 111, 131).

[89] SueYeon Chung, Daniel D Lee, and Haim Sompolinsky. 'Classification and geometry of general perceptual manifolds.' In: *Physical Review X* 8.3 (2018), p. 031003 (cit. on p. 37).

[90] Giorgio Cipolloni, László Erdős, and Dominik Schröder. 'Rank-uniform local law for Wigner matrices.' In: *Forum Math., Sigma* 10 (2022). arXiv: *2203.01861* (cit. on p. 132).

[91] Lucas Clarté, Bruno Loureiro, Florent Krzakala, and Lenka Zdeborová. 'A study of uncertainty quantification in overparametrized high-dimensional models.' In: *arXiv:2210.12760* (2022) (cit. on p. 106).

[92] Omry Cohen, Or Malka, and Zohar Ringel. 'Learning curves for overparametrized deep neural networks: A field theory perspective.' In: *Phys. Rev. Res.* 3 (2 Apr. 2021), p. 023034 (cit. on p. 126).

[93] Hugo Cui, Freya Behrens, Florent Krzakala, and Lenka Zdeborová. 'A phase transition between positional and semantic learning in a solvable model of dot-product attention.' In: *arXiv preprint arXiv:2402.03902* (2024) (cit. on p. 11).

[94] Hugo Cui, Florent Krzakala, and Lenka Zdeborová. 'Bayes-optimal learning of deep random networks of extensive-width.' In: *International Conference on Machine Learning* 40 (2023), 6468–6521, (Oral) (cit. on pp. 36, 86, 105, 113, 121, 136, 137, 151, 165, 171).

[95] Hugo Cui, Florent Krzakala, and Lenka Zdeborová. 'Optimal Learning of Deep Random Networks of Extensive-width.' In: *ArXiv* abs/2302.00375 (2023) (cit. on p. 186).

[96] Hugo Cui, Bruno Loureiro, Florent Krzakala, and Lenka Zdeborová. 'Error scaling laws for kernel classification under source and capacity conditions.' In: *Machine Learning: Science and Technology* 4.3 (2023), p. 035033 (cit. on pp. 89, 106, 147, 186).

[97] Hugo Cui, Bruno Loureiro, Florent Krzakala, and Lenka Zdeborová. 'Generalization error rates in kernel regression: The crossover from the noiseless to noisy regime.' In: *Advances in Neural Information Processing Systems* **34** pp. 10131–10143*;* invited to the special Machine Learning issue of *Journal*

of Statistical Mechanics: Theory and Experiment, **11** pp. 114004 (2021) (cit. on pp. 80–82, 84–86, 89, 90, 92, 106, 147).

[98] Hugo Cui and Lenka Zdeborová. 'High-dimensional asymptotics of denoising autoencoders.' In: *Advances in Neural Information Processing Systems* 36 (2024), (Spotlight) (cit. on pp. 199, 201, 210).

[99] Hugo Cui et al. 'Asymptotics of feature learning in two-layer networks after one gradient-step.' In: *International Conference on Machine Learning* 41 (2024) (cit. on p. 99).

[100] Stéphane D'Ascoli, Marylou Gabrié, Levent Sagun, and Giulio Biroli. 'On the interplay between data structure and loss function in classification problems.' In: *Advances in Neural Information Processing Systems*. Ed. by M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan. Vol. 34. Curran Associates, Inc., 2021, pp. 8506–8517 (cit. on pp. 37, 40).

[101] Stéphane D'Ascoli, Maria Refinetti, Giulio Biroli, and Florent Krzakala. 'Double Trouble in Double Descent: Bias and Variance(s) in the Lazy Regime.' In: *Proceedings of the 37th International Conference on Machine Learning*. Ed. by Hal Daumé III and Aarti Singh. Vol. 119. Proceedings of Machine Learning Research. PMLR, 13–18 Jul 2020, pp. 2280–2290 (cit. on p. 106).

[102] Stéphane D'Ascoli, Levent Sagun, and Giulio Biroli. 'Triple descent and the two kinds of overfitting: where and why do they appear?' In: *J. Stat. Mech. Theory Exp.* 2021.12 (2021), Paper No. 124002, 21 (cit. on pp. 120, 121).

[103] Alexandru Damian, Jason Lee, and Mahdi Soltanolkotabi. 'Neural Networks can Learn Representations with Gradient Descent.' In: *Proceedings of Thirty Fifth Conference on Learning Theory*. Ed. by Po-Ling Loh and Maxim Raginsky. Vol. 178. Proceedings of Machine Learning Research. PMLR, Feb. 2022, pp. 5413–5452 (cit. on pp. 36, 164, 166, 178).

[104] Yatin Dandi, Florent Krzakala, Bruno Loureiro, Luca Pesce, and Ludovic Stephan. *How Two-Layer Neural Networks Learn, One (Giant) Step at a Time*. 2023. arXiv: *2305.18270* [stat.ML] (cit. on pp. 164–166, 168, 170).

[105] Yatin Dandi, Ludovic Stephan, Florent Krzakala, Bruno Loureiro, and Lenka Zdeborová. *Universality laws for Gaussian mixtures in generalized linear models*. 2023. arXiv: *2302.08933* [math.ST] (cit. on pp. 165, 167).

[106] Valentin De Bortoli. 'Convergence of denoising diffusion models under the manifold hypothesis.' In: *arXiv preprint arXiv:2208.05314* (2022) (cit. on pp. 199, 201).

[107] Valentin De Bortoli, James Thornton, Jeremy Heng, and Arnaud Doucet. 'Diffusion Schrödinger bridge with applications to score-based generative modeling.' In: *Advances in Neural Information Processing Systems* 34 (2021), pp. 17695–17709 (cit. on pp. 199, 201).

[108] Oussama Dhifallah and Yue M Lu. 'A precise performance analysis of learning with random features.' In: *arXiv preprint arXiv:2008.11904* (2020) (cit. on pp. 49, 98).

[109] Oussama Dhifallah and Yue M. Lu. *A Precise Performance Analysis of Learning with Random Features*. 2022 (cit. on pp. 106, 124, 125, 165, 167, 168).

[110] Lee H Dicker et al. 'Ridge regression and asymptotic minimax estimation over spheres of growing dimension.' In: *Bernoulli* 22.1 (2016), pp. 1–37 (cit. on pp. 65, 81).

[111] Rainer Dietrich, Manfred Opper, and Haim Sompolinsky. 'Statistical Mechanics of Support Vector Networks.' In: *Phys. Rev. Lett.* 82 (14 Apr. 1999), pp. 2975–2978 (cit. on p. 55).

[112] Rainer Dietrich, Manfred Opper, and Haim Sompolinsky. 'Statistical mechanics of support vector networks.' In: *Physical review letters* 82.14 (1999), p. 2975 (cit. on pp. 65, 186).

[113] Rainer Dietrich, Manfred Opper, and Haim Sompolinsky. 'Statistical mechanics of Support Vector networks.' In: *Physical Review Letters* 82 (1999), pp. 2975–2978 (cit. on p. 86).

[114] Edgar Dobriban and Stefan Wager. 'High-dimensional asymptotics of prediction: Ridge regression and classification.' In: *The Annals of Statistics* 46.1 (2018), pp. 247–279 (cit. on pp. 49, 65, 67, 73, 81, 85).

[115] Edgar Dobriban and Stefan Wager. 'High-dimensional asymptotics of prediction: ridge regression and classification.' In: *Ann. Statist.* 46.1 (2018), pp. 247–279 (cit. on pp. 116, 129).

[116] David Donoho and Andrea Montanari. 'High dimensional robust m-estimation: Asymptotic variance via approximate message passing.' In: *Probability Theory and Related Fields* 166.3-4 (2016), pp. 935–969 (cit. on pp. 45, 221).

[117] David L Donoho, Arian Maleki, and Andrea Montanari. 'Message-passing algorithms for compressed sensing.' In: *Proceedings of the National Academy of Sciences* 106.45 (2009), pp. 18914–18919 (cit. on pp. 12, 45).

[118] David L Donoho, Arian Maleki, and Andrea Montanari. 'The noise-sensitivity phase transition in compressed sensing.' In: *IEEE Transactions on Information Theory* 57.10 (2011), pp. 6920–6941 (cit. on p. 40).

[119] Ethan Dyer and Guy Gur-Ari. 'Asymptotics of Wide Networks from Feynman Diagrams.' In: *International Conference on Learning Representations*. 2020 (cit. on p. 168).

[120] Carl Eckart and G. Marion Young. 'The approximation of one matrix by another of lower rank.' In: *Psychometrika* 1 (1936), pp. 211–218 (cit. on pp. 184, 196).

[121] Benjamin L Edelman, Surbhi Goel, Sham Kakade, and Cyril Zhang. 'Inductive biases and variable creation in self-attention mechanisms.' In: *International Conference on Machine Learning*. PMLR. 2022, pp. 5793–5831 (cit. on p. 215).

[122] Bradley Efron. 'Tweedie's Formula and Selection Bias.' In: *Journal of the American Statistical Association* 106 (2011), pp. 1602–1614 (cit. on pp. 187, 194, 203).

[123] Noureddine El Karoui et al. 'Concentration of measure and spectra of random matrices: Applications to correlation matrices, elliptical distributions and beyond.' In: *Annals of Applied Probability* 19.6 (2009), pp. 2362–2405 (cit. on p. 59).

[124] Noureddine El Karoui. 'Spectrum estimation for large dimensional covariance matrices using random matrix theory.' In: *Ann. Statist.* 36.6 (2008), pp. 2757–2790 (cit. on p. 114).

[125] Noureddine El Karoui. 'Spectrum estimation for large dimensional covariance matrices using random matrix theory.' In: *The Annals of Statistics* 36.6 (2008), pp. 2757–2790 (cit. on pp. 147, 148).

[126] Noureddine El Karoui et al. 'The spectrum of kernel random matrices.' In: *Annals of statistics* 38.1 (2010), pp. 1–50 (cit. on pp. 12, 49, 59).

[127] Noureddine El Karoui, Derek Bean, Peter J Bickel, Chinghway Lim, and Bin Yu. 'On robust regression with high-dimensional predictors.' In: *Proceedings of the National Academy of Sciences* 110.36 (2013), pp. 14557–14562 (cit. on p. 45).

[128] Jeffrey L. Elman. 'Finding structure in time.' In: *Cognitive Science* 14.2 (1990), pp. 179–211 (cit. on p. 215).

[129] Melikasadat Emami, Mojtaba Sahraee-Ardakan, Parthe Pandit, Sundeep Rangan, and Alyson Fletcher. 'Generalization error of generalized linear models in high dimensions.' In: *International Conference on Machine Learning*. PMLR. 2020, pp. 2892–2901 (cit. on p. 222).

[130] Andreas Engel and Christian Van den Broeck. *Statistical mechanics of learning*. Cambridge University Press, 2001 (cit. on pp. 45, 48, 86).

[131] Zhou Fan and Andrea Montanari. 'The spectral norm of random inner-product kernel matrices.' In: *Probability Theory and Related Fields* 173.1 (2019), pp. 27–85 (cit. on p. 59).

[132] Zhou Fan and Zhichao Wang. 'Spectra of the Conjugate Kernel and Neural Tangent Kernel for Linear-Width Neural Networks.' In: *Proceedings of the 34th International Conference on Neural Information Processing Systems*. NIPS'20. Vancouver, BC, Canada: Curran Associates Inc., 2020 (cit. on pp. 105, 106, 113, 128, 135, 136).

[133] Kirsten Fischer, Alexandre Ren'e, Christian Keup, Moritz Layer, David Dahmen, and Moritz Helias. 'Decomposing neural networks as mappings of correlation functions.' In: *Physical Review Research* (2022) (cit. on p. 152).

[134] Simon Fischer and Ingo Steinwart. 'Sobolev Norm Learning Rates for Regularized Least-Squares Algorithms.' In: *Journal of Machine Learning Research* 21.205 (2020), pp. 1–38 (cit. on p. 65).

[135] Hengyu Fu, Tianyu Guo, Yu Bai, and Song Mei. 'What can a single attention layer learn? a study through the random features lens.' In: *arXiv preprint arXiv:2307.11353* (2023) (cit. on p. 215).

[136] Kunihiko Fukushima. 'Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position.' In: *Biological cybernetics* 36.4 (1980), pp. 193–202 (cit. on p. 8).

[137] Alexander G. de G. Matthews, Mark Rowland, Jiri Hron, Richard E. Turner, and Zoubin Ghahramani. 'Gaussian Process Behaviour in Wide Deep Neural Networks.' In: *NeurIPS Workshop on Advances in Approximate Bayesian Inference* (2018) (cit. on pp. 105, 106, 125, 147, 149, 151).

[138] Marylou Gabrié. 'Mean-field inference methods for neural networks.' In: *Journal of Physics A: Mathematical and Theoretical* 53 (2019) (cit. on pp. 21, 34, 148, 185).

[139] Marylou Gabrié. 'Mean-field inference methods for neural networks.' In: *Journal of Physics A: Mathematical and Theoretical* 53.22 (2020), p. 223002 (cit. on pp. xviii, 10, 18, 40).

[140] Marylou Gabrié, Andre Manoel, Clément Luneau, Nicolas Macris, Florent Krzakala, Lenka Zdeborová, et al. 'Entropy and mutual information in models of deep neural networks.' In: *Advances in neural information processing systems* 31 (2018) (cit. on p. 35).

[141] Marylou Gabrié et al. 'Entropy and mutual information in models of deep neural networks.' In: *Advances in Neural Information Processing Systems*. Vol. 31. Curran Associates, Inc., 2018 (cit. on pp. 106, 126).

[142] Elizabeth Gardner and Bernard Derrida. 'Optimal storage properties of neural network models.' In: *Journal of Physics A: Mathematical and general* 21.1 (1988), p. 271 (cit. on pp. 10, 12, 34, 185).

[143] Elizabeth Gardner and Bernard Derrida. 'Three unfinished works on the optimal storage capacity of networks.' In: *Journal of Physics A: Mathematical and General* 22.12 (1989), p. 1983 (cit. on pp. 10, 12, 34, 40, 48, 97).

[144] Mario Geiger, Stefano Spigler, Arthur Jacot, and Matthieu Wyart. 'Disentangling feature and lazy learning in deep neural networks: an empirical study.' In: *arXiv preprint arXiv:1906.08034* (2019) (cit. on pp. 34, 41, 97).

[145] Mario Geiger et al. 'Scaling description of generalization with number of parameters in deep learning.' In: *Journal of Statistical Mechanics: Theory and Experiment* 2020.2 (2020), p. 023401 (cit. on p. 9).

[146] Stuart Geman, Elie Bienenstock, and René Doursat. 'Neural networks and the bias-variance dilemma.' In: *Neural computation* 4.1 (1992), pp. 1–58 (cit. on p. 9).

[147] Federica Gerace, Florent Krzakala, Bruno Loureiro, Ludovic Stephan, and Lenka Zdeborová. *Gaussian Universality of Linear Classifiers with Random Labels in High-Dimension*. 2022 (cit. on p. 117).

[148] Federica Gerace, Bruno Loureiro, Florent Krzakala, Marc Mezard, and Lenka Zdeborova. 'Generalisation error in learning with random features and the hidden manifold model.' In: *Proceedings of the 37th International Conference on Machine Learning*. Ed. by Hal Daumé III and Aarti Singh. Vol. 119. Proceedings of Machine Learning Research. PMLR, 13–18 Jul 2020, pp. 3452–3462 (cit. on pp. 105, 106, 116, 117, 124, 125, 136, 147, 158, 165, 167, 168, 186).

[149] Federica Gerace, Bruno Loureiro, Florent Krzakala, Marc Mézard, and Lenka Zdeborová. 'Generalisation error in learning with random features and the hidden manifold model.' In: *International Conference on Machine Learning*. PMLR. 2020, pp. 3452–3462 (cit. on pp. 49, 53, 66, 98).

[150] Cedric Gerbelot, Alia Abbara, and Florent Krzakala. 'Asymptotic errors for teacher-student convex generalized linear models (or: How to prove Kabashima's replica formula).' In: *IEEE Transactions on Information Theory* 69.3 (2022), pp. 1824–1852 (cit. on p. 222).

[151] Cédric Gerbelot, Alia Abbara, and Florent Krzakala. 'Asymptotic Errors for High-Dimensional Convex Penalized Linear Regression beyond Gaussian Matrices.' In: *Conference on Learning Theory*. PMLR. 2020, pp. 1682–1713 (cit. on p. 49).

[152] Borjan Geshkovski, Cyril Letrouit, Yury Polyanskiy, and Philippe Rigollet. 'A mathematical perspective on Transformers.' In: *arXiv preprint arXiv:2312.10794* (2023) (cit. on p. 215).

[153] Davide Ghio, Yatin Dandi, Florent Krzakala, and Lenka Zdeborová. 'Sampling with flows, diffusion and autoregressive neural networks: A spin-glass perspective.' In: *arXiv preprint arXiv:2308.14085* (2023) (cit. on p. 201).

[154] B. Ghorbani, Song Mei, Theodor Misiakiewicz, and Andrea Montanari. 'Linearized two-layers neural networks in high dimension.' In: *ArXiv* abs/1904.12191 (2019) (cit. on p. 162).

[155] Behrooz Ghorbani, Song Mei, Theodor Misiakiewicz, and Andrea Montanari. 'Limitations of Lazy Training of Two-layers Neural Network.' In: *Advances in Neural Information Processing Systems*. Vol. 32. 2019, pp. 9111–9121 (cit. on p. 66).

[156] Behrooz Ghorbani, Song Mei, Theodor Misiakiewicz, and Andrea Montanari. 'Limitations of Lazy Training of Two-layers Neural Network.' In: *Advances in Neural Information Processing Systems*. Vol. 32. Curran Associates, Inc., 2019 (cit. on pp. 106, 125, 164, 166).

[157] Behrooz Ghorbani, Song Mei, Theodor Misiakiewicz, and Andrea Montanari. 'When Do Neural Networks Outperform Kernel Methods?' In: *Advances in Neural Information Processing Systems*. Vol. 33. Curran Associates, Inc., 2020, pp. 14820–14830 (cit. on pp. 166, 175).

[158] Behrooz Ghorbani, Song Mei, Theodor Misiakiewicz, and Andrea Montanari. 'When do neural networks outperform kernel methods?' In: *Advances in Neural Information Processing Systems*. Vol. 33. 2020 (cit. on pp. 48, 66).

[159] Behrooz Ghorbani, Song Mei, Theodor Misiakiewicz, and Andrea Montanari. 'When do neural networks outperform kernel methods?' In: *J. Stat. Mech. Theory Exp.* 2021.12 (2021), Paper No. 124009, 110 (cit. on pp. 106, 125, 162).

[160] Gauthier Gidel, Francis R. Bach, and Simon Lacoste-Julien. 'Implicit Regularization of Discrete Gradient Dynamics in Deep Linear Neural Networks.' In: *Neural Information Processing Systems*. 2019 (cit. on p. 184).

[161] S. Goldt, M. Mézard, F. Krzakala, and L. Zdeborová. 'Modeling the influence of data structure on learning in neural networks: The hidden manifold model.' In: *Phys. Rev. X* 10.4 (2020), p. 041044 (cit. on pp. 37, 40, 49).

[162] Sebastian Goldt, Bruno Loureiro, Galen Reeves, Florent Krzakala, Marc Mezard, and Lenka Zdeborova. 'The Gaussian equivalence of generative models for learning with shallow neural networks.' In: *Proceedings of the 2nd Mathematical and Scientific Machine Learning Conference*. Ed. by Joan Bruna, Jan Hesthaven, and Lenka Zdeborova. Vol. 145. Proceedings of Machine Learning Research. PMLR, 16–19 Aug 2022, pp. 426–471 (cit. on pp. 165, 167, 168).

[163] Sebastian Goldt, Bruno Loureiro, Galen Reeves, Florent Krzakala, Marc Mezard, and Lenka Zdeborová. 'The Gaussian equivalence of generative models for learning with shallow neural networks.' In: *Proceedings of the 2nd Mathematical and Scientific Machine Learning Conference*. Proceedings of Machine Learning Research. 145. 2021, pp. 426–471 (cit. on pp. 106, 117, 124, 125, 136, 148, 150, 151, 161, 162).

[164] Sebastian Goldt, Bruno Loureiro, Galen Reeves, Florent Krzakala, Marc Mézard, and Lenka Zdeborová. 'The Gaussian equivalence of generative models for learning with shallow neural networks.' In: *Mathematical and Scientific Machine Learning*. 2020 (cit. on pp. 37, 194).

[165] Sebastian Goldt, Bruno Loureiro, Galen Reeves, Florent Krzakala, Marc Mézard, and Lenka Zdeborová. 'The Gaussian equivalence of generative models for learning with shallow neural networks.' In: *Mathematical and Scientific Machine Learning*. PMLR. 2022, pp. 426–471 (cit. on p. 148).

[166] Sebastian Goldt, Bruno Loureiro, Galen Reeves, Marc Mézard, Florent Krzakala, and Lenka Zdeborová. 'The Gaussian equivalence of generative models for learning with two-layer neural networks.' In: *Mathematical and Scientific Machine Learning*. 2021 (cit. on p. 49).

[167] Sebastian Goldt, Marc Mézard, Florent Krzakala, and Lenka Zdeborová. 'Modeling the influence of data structure on learning in neural networks: The hidden manifold model.' In: *Physical Review X* 10.4 (2020), p. 041044 (cit. on pp. 12, 98, 117, 124, 150, 165, 168, 171, 174).

[168] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep learning*. MIT press, 2016 (cit. on p. 4).

[169] Ian Goodfellow et al. *Generative Adversarial Networks*. 2014. arXiv: *1406.2661* [stat.ML] (cit. on p. 45).

[170] Yehoram Gordon. 'On Milman's inequality and random subspaces which escape through a mesh in Rn.' In: *Geometric Aspects of Functional Analysis: Israel Seminar (GAFA) 1986–87*. Springer. 1988, pp. 84–106 (cit. on p. 173).

[171] Yehoram Gordon. 'Some inequalities for Gaussian processes and applications.' In: *Israel Journal of Mathematics* 50.4 (1985), pp. 265–289 (cit. on pp. 46, 49).

[172] Tianyu Guo et al. 'How Do Transformers Learn In-Context Beyond Simple Functions? A Case Study on Learning with Representations.' In: *arXiv preprint arXiv:2310.10616* (2023) (cit. on p. 215).

[173] Florentin Guth, Brice Ménard, Gaspar Rochette, and Stéphane Mallat. 'A Rainbow in Deep Network Black Boxes.' In: *arXiv preprint arXiv:2305.18512* (2023) (cit. on pp. 98, 124, 126, 127).

[174] G Györgyi and N Tishby. 'Neural networks and spin glasses.' In: (1990) (cit. on pp. 10, 34).

[175] Michael Hahn. 'Theoretical limitations of self-attention in neural sequence models.' In: *Transactions of the Association for Computational Linguistics* 8 (2020), pp. 156–171 (cit. on p. 215).

[176] Karl Hajjar and Lenaic Chizat. *On the symmetries in the dynamics of wide two-layer neural networks*. 2023. arXiv: *2211.08771* [cs.LG] (cit. on p. 168).

[177] Kam Hamidieh. 'A data-driven statistical model for predicting the critical temperature of a superconductor.' In: *Computational Materials Science* 154 (2018), pp. 346–354 (cit. on p. 76).

[178] Paul Hand, Oscar Leong, and Vlad Voroninski. 'Phase Retrieval Under a Generative Prior.' In: *Advances in Neural Information Processing Systems*. Vol. 31. Curran Associates, Inc., 2018 (cit. on pp. 106, 126).

[179] Boris Hanin. 'Correlation Functions in Random Fully Connected Neural Networks at Finite Width.' In: *ArXiv* abs/2204.01058 (2022) (cit. on p. 151).

[180] Boris Hanin and Mihai Nica. 'Finite Depth and Width Corrections to the Neural Tangent Kernel.' In: *ArXiv* abs/1909.05989 (2019) (cit. on pp. 106, 126).

[181] Boris Hanin and Mihai Nica. 'Finite Depth and Width Corrections to the Neural Tangent Kernel.' In: *International Conference on Learning Representations*. 2020 (cit. on p. 168).

[182] Boris Hanin and Alexander Zlokapa. 'Bayesian Interpolation with Deep Linear Networks.' In: *ArXiv* abs/2212.14457 (2022) (cit. on pp. 147, 151).

[183] Trevor Hastie, Andrea Montanari, Saharon Rosset, and Ryan J. Tibshirani. 'Surprises in High-Dimensional Ridgeless Least Squares Interpolation.' In: *Ann. Stat.* 50.2 (2022), pp. 949–986 (cit. on pp. 45, 48, 49, 53, 66, 145, 159).

[184] Adi Haviv, Ori Ram, Ofir Press, Peter Izsak, and Omer Levy. 'Transformer Language Models without Positional Encodings Still Learn Positional Information.' In: *Findings of the Association for Computational Linguistics: EMNLP 2022*. Ed. by Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang. Abu Dhabi, United Arab Emirates: Association for Computational Linguistics, Dec. 2022, pp. 1382–1390 (cit. on p. 215).

[185] Matthias Hein and Jean-Yves Audibert. 'Intrinsic dimensionality estimation of submanifolds in Rd.' In: *Proceedings of the 22nd international conference on Machine learning*. 2005, pp. 289–296 (cit. on pp. 9, 40).

[186] Tom Henighan et al. 'Scaling laws for autoregressive generative modeling.' In: *arXiv preprint arXiv:2010.14701* (2020) (cit. on pp. 41, 80).

[187] Joel Hestness et al. 'Deep learning scaling is predictable, empirically.' In: *arXiv preprint arXiv:1712.00409* (2017) (cit. on pp. 41, 80).

[188] Jonathan Ho, Ajay Jain, and Pieter Abbeel. 'Denoising diffusion probabilistic models.' In: *Advances in neural information processing systems* 33 (2020), pp. 6840–6851 (cit. on pp. 8, 181).

[189] Jonathan Ho, Ajay Jain, and Pieter Abbeel. 'Denoising diffusion probabilistic models.' In: *Advances in Neural Information Processing Systems* 33 (2020), pp. 6840–6851 (cit. on pp. 184, 186, 199–201).

[190] Tadaaki Hosaka, Yoshiyuki Kabashima, and Hidetoshi Nishimori. 'Statistical mechanics of lossy data compression using a nonmonotonic perceptron.' In: *Physical Review E* 66.6 (2002), p. 066126 (cit. on p. 40).

[191] Jiri Hron, Yasaman Bahri, Roman Novak, Jeffrey Pennington, and Jascha Sohl-Dickstein. 'Exact posterior distributions of wide Bayesian neural networks.' In: *ArXiv* abs/2006.10541 (2020) (cit. on p. 147).

[192] Daniel Hsu, Sham M. Kakade, and Tong Zhang. 'Random Design Analysis of Ridge Regression.' In: *Proceedings of the 25th Annual Conference on Learning Theory*. Ed. by Shie Mannor, Nathan Srebro, and Robert C. Williamson. Vol. 23. Proceedings of Machine Learning Research. Edinburgh, Scotland: JMLR Workshop and Conference Proceedings, 25–27 Jun 2012, pp. 9.1–9.24 (cit. on pp. 65, 81).

[193] Daniel Hsu, Vidya Muthukumar, and Ji Xu. *On the proliferation of support vectors in high dimensions*. 2020. arXiv: *2009.10670* [math.ST] (cit. on p. 93).

[194] Hong Hu and Yue M Lu. 'Sharp asymptotics of kernel ridge regression beyond the linear regime.' In: *arXiv preprint arXiv:2205.06798* (2022) (cit. on p. 37).

[195] Hong Hu and Yue M Lu. 'Universality laws for high-dimensional learning with random features.' In: *IEEE Transactions on Information Theory* 69.3 (2022), pp. 1932–1964 (cit. on pp. 49, 53, 98, 165, 167, 168, 171, 173, 174).

[196] Hong Hu and Yue M. Lu. 'Sharp Asymptotics of Kernel Ridge Regression Beyond the Linear Regime.' In: *arXiv:2205.06798* (2022) (cit. on pp. 151, 159, 162).

[197] Hong Hu and Yue M. Lu. 'Universality Laws for High-Dimensional Learning With Random Features.' In: *IEEE Transactions on Information Theory* 69

(2020), pp. 1932–1964 (cit. on pp. 106, 116, 117, 124, 136, 148, 150, 161, 162, 194).

[198] Hanwen Huang and Qinglong Yang. 'Large scale analysis of generalization error in learning using margin based classification methods.' In: *Journal of Statistical Mechanics: Theory and Experiment* 2020.10 (2020), p. 103407 (cit. on p. 49).

[199] Marcus Hutter. 'Learning curve theory.' In: *arXiv preprint arXiv:2102.04074* (2021) (cit. on p. 81).

[200] Yukito Iba. 'The Nishimori line and Bayesian statistics.' In: *Journal of Physics A* 32 (1998), pp. 3875–3888 (cit. on pp. 150, 151).

[201] Arthur Jacot, Franck Gabriel, and Clément Hongler. 'Neural tangent kernel: Convergence and generalization in neural networks.' In: *Advances in neural information processing systems*. 2018, pp. 8571–8580 (cit. on pp. 34, 41, 45, 64, 164).

[202] Arthur Jacot, Franck Gabriel, and Clement Hongler. 'Neural Tangent Kernel: Convergence and Generalization in Neural Networks.' In: *Advances in Neural Information Processing Systems*. Vol. 31. Curran Associates, Inc., 2018 (cit. on p. 97).

[203] Arthur Jacot, Berfin Simsek, Francesco Spadaro, Clément Hongler, and Franck Gabriel. 'Implicit regularization of random feature models.' In: *International Conference on Machine Learning*. PMLR. 2020, pp. 4631–4640 (cit. on pp. 106, 137, 138).

[204] Arthur Jacot, Berfin Şimşek, Francesco Spadaro, Clément Hongler, and Franck Gabriel. 'Kernel Alignment Risk Estimator: Risk Prediction from Training Data.' In: *arXiv preprint arXiv:2006.09796* (2020) (cit. on pp. 49, 59, 66).

[205] Saachi Jain, Adityanarayanan Radhakrishnan, and Caroline Uhler. 'A Mechanism for Producing Aligned Latent Spaces with Autoencoders.' In: *ArXiv* abs/2106.15456 (2021) (cit. on p. 185).

[206] Adel Javanmard and Andrea Montanari. 'State evolution for general approximate message passing algorithms, with applications to spatial coupling.' In: *Information and Inference: A Journal of the IMA* 2.2 (2013), pp. 115–144 (cit. on p. 221).

[207] Samy Jelassi, Michael Sander, and Yuanzhi Li. 'Vision transformers provably learn spatial structure.' In: *Advances in Neural Information Processing Systems* 35 (2022), pp. 37822–37836 (cit. on pp. 214, 215).

[208] Hui Jin, Pradeep Kr. Banerjee, and Guido Montúfar. *Learning curves for Gaussian process regression with power-law priors and targets*. 2021. arXiv: *2110.12231* [cs.LG] (cit. on pp. 84, 86, 90).

[209] Kwang-Sung Jun, Ashok Cutkosky, and Francesco Orabona. 'Kernel Truncated Randomized Ridge Regression: Optimal Rates and Low Noise Acceleration.' In: *NeurIPS*. 2019 (cit. on pp. 65, 80, 81).

[210] Yoshiyuki Kabashima. 'Inference from correlated patterns: a unified theory for perceptron learning and linear vector channels.' In: *Journal of Physics: Conference Series*. Vol. 95. 1. IOP Publishing. 2008, p. 012001 (cit. on pp. 37, 65).

[211] M. Kanagawa, P. Hennig, D. Sejdinovic, and B. K. Sriperumbudur. 'Gaussian Processes and Kernel Methods: A Review on Connections and Equivalences.' In: *Arxiv e-prints* arXiv:1805.08845v1 [stat.ML] (2018) (cit. on p. 65).

[212] Jared Kaplan et al. 'Scaling laws for neural language models.' In: *arXiv preprint arXiv:2001.08361* (2020) (cit. on pp. 41, 80).

[213] Tero Karras, Miika Aittala, Timo Aila, and Samuli Laine. 'Elucidating the design space of diffusion-based generative models.' In: *Advances in Neural Information Processing Systems* 35 (2022), pp. 26565–26577 (cit. on pp. 199, 200).

[214] Amirhossein Kazemnejad, Inkit Padhi, Karthikeyan Natesan, Payel Das, and Siva Reddy. 'The Impact of Positional Encoding on Length Generalization in Transformers.' In: *Thirty-seventh Conference on Neural Information Processing Systems*. 2023 (cit. on p. 215).

[215] Jiwon Kim, Jung Kwon Lee, and Kyoung Mu Lee. 'Accurate image super-resolution using very deep convolutional networks.' In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016, pp. 1646–1654 (cit. on p. 187).

[216] Jiwon Kim, Jung Kwon Lee, and Kyoung Mu Lee. 'Deeply-recursive convolutional network for image super-resolution.' In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016, pp. 1637–1645 (cit. on p. 187).

[217] George Kimeldorf and Grace Wahba. 'Some results on Tchebycheffian spline functions.' In: *Journal of mathematical analysis and applications* 33.1 (1971), pp. 82–95 (cit. on p. 7).

[218] Diederik P Kingma and Jimmy Ba. 'Adam: A method for stochastic optimization.' In: *arXiv preprint arXiv:1412.6980* (2014) (cit. on pp. 138, 206).

[219] Diederik P. Kingma and Jimmy Ba. 'Adam: A Method for Stochastic Optimization.' In: *CoRR* abs/1412.6980 (2014) (cit. on pp. 192, 194).

[220] Antti Knowles and Jun Yin. 'Anisotropic local laws for random matrices.' In: *Probab. Theory Related Fields* 169.1-2 (2017), pp. 257–352 (cit. on pp. 109, 131, 132).

[221] Dmitry Kobak, Jonathan Lomond, and Benoit Sanchez. 'The optimal ridge penalty for real-world high-dimensional data can be zero or negative due to the implicit ridge regularization.' In: *Journal of Machine Learning Research* 21.169 (2020), pp. 1–16 (cit. on p. 66).

[222] Alex Krizhevsky, Geoffrey Hinton, et al. 'Learning multiple layers of features from tiny images.' In: - (2009) (cit. on p. 94).

[223] Daniel Kunin, Jonathan M. Bloom, Aleksandrina Goeva, and Cotton Seed. 'Loss Landscapes of Regularized Linear Autoencoders.' In: *International Conference on Machine Learning*. 2019 (cit. on pp. 184, 185).

[224] Hugo Latourelle-Vigeant and Elliot Paquette. 'Matrix Dyson equation for correlated linearizations and test error of random features regression.' Preprint. 2023. arXiv: *2312.09194* (cit. on pp. 129, 130).

[225] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. 'Deep learning.' In: *nature* 521.7553 (2015), pp. 436–444 (cit. on p. 8).

[226] Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. 'Gradient-based learning applied to document recognition.' In: *Proceedings of the IEEE* 86.11 (1998), pp. 2278–2324 (cit. on pp. 36, 85, 94).

[227] Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. 'Gradient-based learning applied to document recognition.' In: *Proc. IEEE* 86 (1998), pp. 2278–2324 (cit. on pp. 185, 193).

[228] Olivier Ledoit and Sandrine Péché. 'Eigenvectors of some large sample covariance matrix ensembles.' In: *Probability Theory and Related Fields* 151.1 (2011), pp. 233–264 (cit. on pp. 59, 65).

[229] Holden Lee, Jianfeng Lu, and Yixin Tan. 'Convergence for score-based generative modeling with polynomial complexity.' In: *Advances in Neural Information Processing Systems* 35 (2022), pp. 22870–22882 (cit. on pp. 199, 201).

[230] Holden Lee, Jianfeng Lu, and Yixin Tan. 'Convergence of score-based generative modeling for general data distributions.' In: *International Conference on Algorithmic Learning Theory*. PMLR. 2023, pp. 946–985 (cit. on pp. 199, 201).

[231] Jaehoon Lee, Yasaman Bahri, Roman Novak, Samuel S. Schoenholz, Jeffrey Pennington, and Jascha Sohl-Dickstein. *Deep Neural Networks as Gaussian Processes*. ICLR 2018, arXiv:1711.00165. 2018 (cit. on p. 64).

[232] Jaehoon Lee, Yasaman Bahri, Roman Novak, Samuel S. Schoenholz, Jeffrey Pennington, and Jascha Sohl-Dickstein. 'Deep Neural Networks as Gaussian Processes.' In: *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net, 2018 (cit. on pp. 105, 106, 125).

[233] Jaehoon Lee, Yasaman Bahri, Roman Novak, Samuel S. Schoenholz, Jeffrey Pennington, and Jascha Sohl-Dickstein. 'Deep Neural Networks as Gaussian Processes.' In: *International Conference on Learning Representations* (2018) (cit. on pp. 147, 149, 151, 159).

[234] Jaehoon Lee, Jascha Sohl-dickstein, Jeffrey Pennington, Roman Novak, Sam Schoenholz, and Yasaman Bahri. 'Deep Neural Networks as Gaussian Processes.' In: *International Conference on Learning Representations*. 2018 (cit. on p. 164).

[235] Jaehoon Lee et al. 'Finite Versus Infinite Neural Networks: an Empirical Study.' In: *NeurIPS* (2020) (cit. on p. 160).

[236] Jaehoon Lee et al. 'Wide neural networks of any depth evolve as linear models under gradient descent.' In: *Journal of Statistical Mechanics: Theory and Experiment* 2020 (2019) (cit. on p. 147).

[237] Gen Li, Yuting Wei, Yuxin Chen, and Yuejie Chi. 'Towards Faster Non-Asymptotic Convergence for Diffusion-Based Generative Models.' In: *arXiv preprint arXiv:2306.09251* (2023) (cit. on pp. 199, 201).

[238] Hongkang Li, Meng Wang, Sijia Liu, and Pin-Yu Chen. 'A theoretical understanding of shallow vision transformers: Learning, generalization, and sample complexity.' In: *arXiv preprint arXiv:2302.06015* (2023) (cit. on p. 215).

[239] Qianyi Li, Qianyi Li, Haim Sompolinsky, Haim Sompolinsky, and Haim Sompolinsky. 'Statistical Mechanics of Deep Linear Neural Networks: The Backpropagating Kernel Renormalization.' In: *Phys. Rev. X* 11.3 (2021) (cit. on pp. 147, 150).

[240] Qianyi Li and Haim Sompolinsky. 'Statistical Mechanics of Deep Linear Neural Networks: The Backpropagating Kernel Renormalization.' In: *Phys. Rev. X* 11 (3 2021), p. 031059 (cit. on pp. 106, 126).

[241] Qianyi Li and Haim Sompolinsky. 'Statistical mechanics of deep linear neural networks: The backpropagating kernel renormalization.' In: *Physical Review X* 11.3 (2021), p. 031059 (cit. on pp. 36, 86).

[242] Yingcong Li, Muhammed Emrullah Ildiz, Dimitris Papailiopoulos, and Samet Oymak. 'Transformers as algorithms: Generalization and stability in in-context learning.' In: *International Conference on Machine Learning*. PMLR. 2023, pp. 19565–19594 (cit. on p. 215).

[243] Tengyuan Liang and Pragya Sur. 'A precise high-dimensional asymptotic theory for boosting and minimum-$\ell 1$-norm interpolated classifiers.' In: *Ann. Statist.* 50.3 (2022), pp. 1669–1695 (cit. on pp. 106, 125).

[244] Zhenyu Liao and Romain Couillet. 'On the Spectrum of Random Features Maps of High Dimensional Data.' In: *Proceedings of the 35th International Conference on Machine Learning*. Ed. by Jennifer Dy and Andreas Krause. Vol. 80. Proceedings of Machine Learning Research. PMLR, Oct. 2018, pp. 3063–3071 (cit. on pp. 106, 125).

[245] Zhenyu Liao, Romain Couillet, and Michael W Mahoney. 'A random matrix analysis of random Fourier features: beyond the Gaussian kernel, a precise phase transition, and the corresponding double descent.' In: *Advances in Neural Information Processing Systems*. Vol. 33. 2020 (cit. on pp. 11, 49, 66).

[246] Junhong Lin, Alessandro Rudi, L. Rosasco, and V. Cevher. 'Optimal Rates for Spectral Algorithms with Least-Squares Regression over Hilbert Spaces.' In: *Applied and Computational Harmonic Analysis* 48 (2018), pp. 868–890 (cit. on pp. 65, 80, 81).

[247] Yaron Lipman, Ricky TQ Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. 'Flow matching for generative modeling.' In: *arXiv preprint arXiv:2210.02747* (2022) (cit. on pp. 199, 200).

[248] Fanghui Liu, Zhenyu Liao, and Johan AK Suykens. 'Kernel regression in high dimension: Refined analysis beyond double descent.' In: *arXiv preprint arXiv:2010.02681* (2020) (cit. on pp. 49, 66, 80, 81).

[249] Xingchao Liu, Chengyue Gong, and Qiang Liu. 'Flow straight and fast: Learning to generate and transfer data with rectified flow.' In: *arXiv preprint arXiv:2209.03003* (2022) (cit. on pp. 199, 200).

[250] Cosme Louart and Romain Couillet. 'Concentration of measure and large random matrices with an application to sample covariance matrices.' In: *arXiv preprint arXiv:1805.08295* (2018) (cit. on pp. 49, 59, 107, 109).

[251] Cosme Louart, Zhenyu Liao, and Romain Couillet. 'A random matrix approach to neural networks.' In: *Ann. Appl. Probab.* 28.2 (2018), pp. 1190–1248 (cit. on pp. 11, 67, 147).

[252] Cosme Louart, Zhenyu Liao, and Romain Couillet. 'A random matrix approach to neural networks.' In: *Ann. Appl. Probab.* 28.2 (2018), pp. 1190–1248 (cit. on pp. 129, 131).

[253] Cosme Louart, Zhenyu Liao, and Romain Couillet. 'A random matrix approach to neural networks.' In: *Ann. Appl. Probab.* 28.2 (2018), pp. 1190–1248. arXiv: *1702.05419* (cit. on p. 131).

[254] Bruno Loureiro, Cedric Gerbelot, Maria Refinetti, Gabriele Sicuro, and Florent Krzakala. 'Fluctuations, Bias, Variance & Ensemble of Learners: Exact Asymptotics for Convex Losses in High-Dimension.' In: *Proceedings of the 39th International Conference on Machine Learning*. Ed. by Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato. Vol. 162. Proceedings of Machine Learning Research. PMLR, 17–23 Jul 2022, pp. 14283–14314 (cit. on p. 106).

[255] Bruno Loureiro, Gabriele Sicuro, Cedric Gerbelot, Alessandro Pacco, Florent Krzakala, and Lenka Zdeborová. 'Learning Gaussian Mixtures with Generalized Linear Models: Precise Asymptotics in High-dimensions.' In: *Advances in Neural Information Processing Systems*. Ed. by M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan. Vol. 34. Curran Associates, Inc., 2021, pp. 10144–10157 (cit. on p. 222).

[256] Bruno Loureiro et al. 'Learning curves of generic features maps for realistic datasets with a teacher-student model.' In: *Advances in Neural Information Processing Systems* **34** pp. 18137–18151*;* invited to the special Machine Learning issue of *Journal of Statistical Mechanics: Theory and Experiment,* **11** pp. 114001 (2021) (cit. on pp. 37, 82, 86, 87, 90, 91, 105, 106, 117, 125, 129, 145, 147, 167, 173).

[257] Antoine Maillard, Florent Krzakala, Marc Mézard, and Lenka Zdeborová. 'Perturbative construction of mean-field equations in extensive-rank matrix factorization and denoising.' In: *Journal of Statistical Mechanics: Theory and Experiment* 2022.8 (2022), p. 083301 (cit. on p. 35).

[258] Antoine Maillard, Bruno Loureiro, Florent Krzakala, and Lenka Zdeborová. 'Phase retrieval in high dimensions: Statistical and computational phase transitions.' In: *Advances in Neural Information Processing Systems* 33 (2020), pp. 11071–11082 (cit. on pp. 12, 40, 97).

[259] Antoine Maillard, Bruno Loureiro, Florent Krzakala, and Lenka Zdeborová. 'Phase retrieval in high dimensions: Statistical and computational phase transitions.' In: *Advances in Neural Information Processing Systems* 33 (2020), pp. 11071–11082 (cit. on p. 147).

[260] Stéphane Mallat. 'Understanding deep convolutional networks.' In: *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences* 374.2065 (2016), p. 20150203 (cit. on pp. 5, 9).

[261] Alexander Maloney, Daniel A Roberts, and James Sully. 'A Solvable Model of Neural Scaling Laws.' In: *arXiv preprint arXiv:2210.16859* (2022) (cit. on p. 81).

[262] Andre Manoel, Florent Krzakala, Marc Mézard, and Lenka Zdeborová. 'Multilayer generalized linear estimation.' In: *2017 IEEE International Symposium on Information Theory (ISIT)*. IEEE. 2017, pp. 2098–2102 (cit. on p. 35).

[263] Andre Manoel, Florent Krzakala, Marc Mézard, and Lenka Zdeborová. 'Multilayer generalized linear estimation.' In: *2017 IEEE International Symposium on Information Theory (ISIT)*. 2017, pp. 2098–2102 (cit. on pp. 106, 126).

[264] Xiao-Jiao Mao, Chunhua Shen, and Yubin Yang. 'Image Restoration Using Convolutional Auto-encoders with Symmetric Skip Connections.' In: *ArXiv* abs/1606.08921 (2016) (cit. on pp. 186, 196).

[265] Vladimir Alexandrovich Marchenko and Leonid Andreevich Pastur. 'Distribution of eigenvalues for some sets of random matrices.' In: *Matematicheskii Sbornik* 114.4 (1967), pp. 507–536 (cit. on pp. 106–108, 131).

[266] Ulysse Marteau-Ferey, Dmitrii Ostrovskii, Francis R. Bach, and Alessandro Rudi. 'Beyond Least-Squares: Fast Rates for Regularized Empirical Risk Minimization through Self-Concordance.' In: *COLT*. 2019 (cit. on pp. 81, 85).

[267] Charles H. Martin and Michael W. Mahoney. 'Implicit Self-Regularization in Deep Neural Networks: Evidence from Random Matrix Theory and Implications for Learning.' In: *Journal of Machine Learning Research* 22.165 (2021), pp. 1–73 (cit. on p. 124).

[268] Peter McCullagh. *Generalized linear models*. Routledge, 2019 (cit. on p. 221).

[269] Song Mei, Theodor Misiakiewicz, and Andrea Montanari. 'Generalization error of random feature and kernel methods: Hypercontractivity and kernel matrix concentration.' In: *Applied and Computational Harmonic Analysis* 59 (2022). Special Issue on Harmonic Analysis and Machine Learning, pp. 3–84 (cit. on p. 166).

[270] Song Mei, Theodor Misiakiewicz, and Andrea Montanari. 'Generalization error of random feature and kernel methods: hypercontractivity and kernel matrix concentration.' In: *Appl. Comput. Harmon. Anal.* 59 (2022), pp. 3–84 (cit. on pp. 106, 117, 124, 125, 129, 155, 159).

[271] Song Mei, Theodor Misiakiewicz, and Andrea Montanari. 'Mean-field theory of two-layers neural networks: dimension-free bounds and kernel limit.' In: (2019), pp. 2388–2464 (cit. on p. 34).

[272] Song Mei and Andrea Montanari. 'The generalization error of random features regression: Precise asymptotics and double descent curve.' In: *arXiv preprint arXiv:1908.05355* (2019) (cit. on pp. 45, 48, 49, 53, 59, 66).

[273] Song Mei and Andrea Montanari. 'The Generalization Error of Random Features Regression: Precise Asymptotics and the Double Descent Curve.' In: *Communications on Pure and Applied Mathematics* 75 (2019) (cit. on pp. 105, 106, 124, 125, 136, 145, 158, 162).

[274] Song Mei and Andrea Montanari. 'The Generalization Error of Random Features Regression: Precise Asymptotics and the Double Descent Curve.' In: *Communications on Pure and Applied Mathematics* 75.4 (2022), pp. 667–

766. eprint: *https://onlinelibrary.wiley.com/doi/pdf/10.1002/cpa.22008* (cit. on pp. 37, 165, 167, 168).

[275] Song Mei and Andrea Montanari. 'The generalization error of random features regression: Precise asymptotics and the double descent curve.' In: *Communications on Pure and Applied Mathematics* 75.4 (2022), pp. 667–766 (cit. on p. 98).

[276] Song Mei, Andrea Montanari, and Phan-Minh Nguyen. 'A mean field view of the landscape of two-layer neural networks.' In: *Proceedings of the National Academy of Sciences* 115.33 (2018), E7665–E7671 (cit. on pp. 35, 168).

[277] Song Mei and Yuchen Wu. 'Deep Networks as Denoising Algorithms: Sample-Efficient Learning of Diffusion Models in High-Dimensional Graphical Models.' In: *arXiv preprint arXiv:2309.11420* (2023) (cit. on pp. 200, 201).

[278] Marc Mézard and Andrea Montanari. *Information, physics, and computation.* Oxford Graduate Texts. Oxford University Press, Oxford, 2009, pp. xiv+569 (cit. on pp. xviii, 10, 12, 21, 47, 86, 149).

[279] Marc Mézard, Giorgio Parisi, and Miguel Virasoro. *Spin glass theory and beyond: An Introduction to the Replica Method and Its Applications.* Vol. 9. World Scientific Publishing Company, 1987 (cit. on pp. 46, 86).

[280] Francesca Mignacco, Florent Krzakala, Yue Lu, Pierfrancesco Urbani, and Lenka Zdeborova. 'The role of regularization in classification of high-dimensional noisy gaussian mixture.' In: *International conference on machine learning*. PMLR. 2020, pp. 6874–6883 (cit. on pp. 37, 40).

[281] Francesca Mignacco, Florent Krzakala, Yue M. Lu, and Lenka Zdeborová. 'The role of regularization in classification of high-dimensional noisy Gaussian mixture.' In: *International Conference on Machine Learning*. 2020 (cit. on p. 40).

[282] Léo Miolane and Andrea Montanari. 'The distribution of the lasso: Uniform control over sparse balls and adaptive parameter tuning.' In: *arXiv preprint arXiv:1811.01212* (2018) (cit. on p. 52).

[283] Theodor Misiakiewicz. 'Spectrum of inner-product kernel matrices in the polynomial regime and multiple descent phenomenon in kernel ridge regression.' In: *ArXiv* abs/2204.10425 (2022) (cit. on p. 162).

[284] Partha P Mitra. 'Understanding overfitting peaks in generalization error: Analytical risk curves for l_2 and l_1 penalized interpolation.' In: *arXiv preprint arXiv:1906.03667* (2019) (cit. on p. 49).

[285] Mehryar Mohri, Afshin Rostamizadeh, and Ameet Talwalkar. *Foundations of machine learning*. MIT press, 2018 (cit. on p. 4).

[286] Behrad Moniri, Donghwan Lee, Hamed Hassani, and Edgar Dobriban. *A Theory of Non-Linear Feature Learning with One Gradient Step in Two-Layer Neural Networks.* 2023. arXiv: *2310.07891* [stat.ML] (cit. on pp. 99, 164, 166, 167, 175).

[287] Andrea Montanari, Feng Ruan, Youngtak Sohn, and Jun Yan. 'The generalization error of max-margin linear classifiers: High-dimensional asymptotics in the overparametrized regime.' In: *arXiv preprint arXiv:1911.01544* (2019) (cit. on pp. 48, 49, 52, 145, 147).

[288] Andrea Montanari and Basil N Saeed. 'Universality of empirical risk minimization.' In: *Conference on Learning Theory*. PMLR. 2022, pp. 4310–4312 (cit. on p. 167).

[289] Andrea Montanari and Basil N. Saeed. 'Universality of empirical risk minimization.' In: *Proceedings of Thirty Fifth Conference on Learning Theory.* Ed. by Po-Ling Loh and Maxim Raginsky. Vol. 178. Proceedings of Machine Learning Research. PMLR, Feb. 2022, pp. 4310–4312 (cit. on pp. 106, 117, 148, 150, 194).

[290] Alireza Mousavi-Hosseini, Denny Wu, Taiji Suzuki, and Murat A Erdogdu. 'Gradient-based feature learning under structured data.' In: *arXiv preprint arXiv:2309.03843* (2023) (cit. on p. 164).

[291] Vidya Muthukumar, Adhyyan Narang, Vignesh Subramanian, Mikhail Belkin, Daniel Hsu, and Anant Sahai. 'Classification vs regression in overparameterized regimes: Does the loss function matter?' In: *Journal of Machine Learning Research* 22.222 (2021), pp. 1–69 (cit. on p. 93).

[292] EA Nadaraya. 'On a regression estimate.' In: *Teor. Verojatnost. i Primenen.* 9 (1964), pp. 157–159 (cit. on p. 64).

[293] Gadi Naveh, Oded Ben David, Haim Sompolinsky, and Zohar Ringel. 'Predicting the outputs of finite deep neural networks trained with noisy gradients.' In: *Phys. Rev. E* 104 (6 Dec. 2021), p. 064301 (cit. on p. 126).

[294] Gadi Naveh and Zohar Ringel. 'A self consistent theory of Gaussian Processes captures feature learning effects in finite CNNs.' In: *Advances in*

Neural Information Processing Systems. Ed. by M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan. Vol. 34. Curran Associates, Inc., 2021, pp. 21352–21364 (cit. on p. 168).

[295] R.M. Neal. *Bayesian Learning for Neural Networks*. Lecture Notes in Statistics. Springer New York, 1996 (cit. on p. 164).

[296] Radford M. Neal. 'Priors for Infinite Networks.' In: *Bayesian Learning for Neural Networks*. New York, NY: Springer New York, 1996, pp. 29–53 (cit. on pp. 41, 64, 97, 147, 149, 159).

[297] John Ashworth Nelder and Robert WM Wedderburn. 'Generalized linear models.' In: *Journal of the Royal Statistical Society Series A: Statistics in Society* 135.3 (1972), pp. 370–384 (cit. on p. 221).

[298] Arkadij Semenovic Nemirovskij and David Borisovich Yudin. *Problem complexity and method efficiency in optimization*. Wiley-Interscience, 1983 (cit. on p. 65).

[299] Phan-Minh Nguyen. 'Analysis of feature learning in weight-tied autoencoders via the mean field lens.' In: *ArXiv* abs/2102.08373 (2021) (cit. on pp. 34, 184, 185, 188, 191, 196).

[300] Thanh Van Nguyen, Raymond K. W. Wong, and Chinmay Hegde. 'Benefits of Jointly Training Autoencoders: An Improved Neural Tangent Kernel Analysis.' In: *IEEE Transactions on Information Theory* 67 (2019), pp. 4669–4692 (cit. on pp. 185, 188).

[301] Thanh Van Nguyen, Raymond K. W. Wong, and Chinmay Hegde. 'On the Dynamics of Gradient Descent for Autoencoders.' In: *International Conference on Artificial Intelligence and Statistics*. 2019 (cit. on pp. 185, 188).

[302] Hidetoshi Nishimori. *Statistical Physics of Spin Glasses and Information Processing*. Oxford:Clarendon, 2001 (cit. on pp. 150, 151).

[303] Lorenzo Noci, Gregor Bachmann, Kevin Roth, Sebastian Nowozin, and Thomas Hofmann. 'Precise characterization of the prior predictive distribution of deep ReLU networks.' In: *Advances in Neural Information Processing Systems*. Ed. by M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan. Vol. 34. Curran Associates, Inc., 2021, pp. 20851–20862 (cit. on pp. 105, 106).

[304] Ivan Nourdin and Giovanni Peccati. *Normal approximations with Malliavin calculus: from Stein's method to universality*. Vol. 192. Cambridge University Press, 2012 (cit. on p. 135).

[305] Reza Oftadeh, Jiayi Shen, Zhangyang Wang, and Dylan A. Shell. 'Eliminating the Invariance on the Loss Landscape of Linear Autoencoders.' In: *International Conference on Machine Learning*. 2020 (cit. on p. 184).

[306] Kazusato Oko, Shunta Akiyama, and Taiji Suzuki. 'Diffusion models are minimax optimal distribution estimators.' In: *arXiv preprint arXiv:2303.01861* (2023) (cit. on pp. 199, 201).

[307] Catherine Olsson et al. 'In-context Learning and Induction Heads.' In: *Transformer Circuits Thread* (2022). https://transformer-circuits.pub/2022/in-context-learning-and-induction-heads/index.html (cit. on p. 215).

[308] Opper and Haussler. 'Generalization performance of Bayes optimal classification algorithm for learning a perceptron.' In: *Physical review letters* 66 20 (1991), pp. 2677–2680 (cit. on p. 145).

[309] M. Opper and R. Urbanczik. 'Universal Learning Curves of Support Vector Machines.' In: *Phys. Rev. Lett.* 86 (19 May 2001), pp. 4410–4413 (cit. on p. 55).

[310] M. Opper and R. Urbanczik. 'Universal Learning Curves of Support Vector Machines.' In: *Phys. Rev. Lett.* 86 (19 May 2001), pp. 4410–4413 (cit. on p. 65).

[311] Manfred Opper and David Haussler. 'Generalization performance of Bayes optimal classification algorithm for learning a perceptron.' In: *Physical Review Letters* 66.20 (1991), p. 2677 (cit. on p. 185).

[312] Manfred Opper and Wolfgang Kinzel. 'Statistical mechanics of generalization.' In: *Models of neural networks III*. Springer, 1996, pp. 151–209 (cit. on pp. 48, 58, 65).

[313] Samet Oymak, Christos Thrampoulidis, and Babak Hassibi. 'The squared-error of generalized lasso: A precise analysis.' In: *2013 51st Annual Allerton Conference on Communication, Control, and Computing (Allerton)*. IEEE. 2013, pp. 1002–1009 (cit. on pp. 11, 49).

[314] R. Pacelli, S. Ariosto, M. Pastore, F. Ginelli, M. Gherardi, and P. Rotondo. 'A statistical mechanics framework for Bayesian deep neural networks beyond the infinite-width limit.' In: *Nature Machine Intelligence* 5.12 (Dec. 2023), pp. 1497–1507 (cit. on pp. 106, 126).

[315] Courtney Paquette, Elliot Paquette, Ben Adlam, and Jeffrey Pennington. 'Homogenization of SGD in high-dimensions: Exact dynamics and gen-

eralization properties.' In: *arXiv preprint arXiv:2205.07069* (2022) (cit. on p. 106).

[316] Giorgio Parisi. 'Order Parameter for spin glasses.' In: *Phys. Rev. Lett* 50 (1983), pp. 1946–1948 (cit. on p. 185).

[317] Giorgio Parisi. 'Order parameter for spin-glasses.' In: *Physical Review Letters* 50.24 (1983), p. 1946 (cit. on pp. 12, 14, 165).

[318] Giorgio Parisi. 'Toward a mean field theory for spin glasses.' In: *Physics Letters A* 73.3 (1979), pp. 203–205 (cit. on pp. 12, 14, 165).

[319] Giorgio Parisi. 'Towards a mean field theory for spin glasses.' In: *Phys. Lett* 73.A (1979), pp. 203–205 (cit. on pp. 149, 185).

[320] Adam Paszke et al. 'PyTorch: An Imperative Style, High-Performance Deep Learning Library.' In: *Advances in Neural Information Processing Systems 32*. Curran Associates, Inc., 2019, pp. 8024–8035 (cit. on p. 226).

[321] Adam Paszke et al. 'Pytorch: An imperative style, high-performance deep learning library.' In: *Advances in neural information processing systems* 32 (2019) (cit. on p. 206).

[322] F. Pedregosa et al. 'Scikit-learn: Machine Learning in Python.' In: *Journal of Machine Learning Research* 12 (2011), pp. 2825–2830 (cit. on pp. 71, 72, 76, 77).

[323] Fabian Pedregosa et al. 'Scikit-learn: Machine learning in Python.' In: *the Journal of machine Learning research* 12 (2011), pp. 2825–2830 (cit. on p. 87).

[324] Jeffrey Pennington and Pratik Worah. 'Nonlinear random matrix theory for deep learning.' In: *Advances in Neural Information Processing Systems*. Vol. 30. 2017, pp. 2637–2646 (cit. on pp. 49, 59).

[325] Jeffrey Pennington and Pratik Worah. 'Nonlinear random matrix theory for deep learning.' In: *J. Stat. Mech. Theory Exp.* 12 (2019), pp. 124005, 14 (cit. on pp. 105, 106, 125, 147).

[326] Giovanni Piccioli, Emanuele Troiani, and Lenka Zdeborová. 'Gibbs Sampling the Posterior of Neural Networks.' In: *arXiv preprint arXiv:2306.02729* (2023) (cit. on pp. 153, 154).

[327] Jakiw Pidstrigach. 'Score-based generative models detect manifolds.' In: *Advances in Neural Information Processing Systems* 35 (2022), pp. 35852–35865 (cit. on pp. 199, 201).

[328] Loucas Pillaud-Vivien, Alessandro Rudi, and Francis Bach. 'Statistical Optimality of Stochastic Gradient Descent on Hard Learning Problems through Multiple Passes.' In: *Advances in Neural Information Processing Systems*. Vol. 31. 2018, pp. 8114–8124 (cit. on pp. 55, 64, 65, 80, 81).

[329] Boris T Polyak and Anatoli B Juditsky. 'Acceleration of stochastic approximation by averaging.' In: *SIAM journal on control and optimization* 30.4 (1992), pp. 838–855 (cit. on p. 65).

[330] Arnu Pretorius, Steve Kroon, and Herman Kamper. 'Learning Dynamics of Linear Denoising Autoencoders.' In: *International Conference on Machine Learning*. 2018 (cit. on pp. 184, 193).

[331] A. Radford, L. Metz, and S. Chintala. 'Unsupervised representation learning with deep convolutional generative adversarial networks.' In: *ICLR*. 2016 (cit. on p. 56).

[332] Adityanarayanan Radhakrishnan, Mikhail Belkin, and Caroline Uhler. 'Overparameterized neural networks implement associative memory.' In: *Proceedings of the National Academy of Sciences of the United States of America* 117 (2019), pp. 27162–27170 (cit. on pp. 185, 188).

[333] Ali Rahimi and Benjamin Recht. 'Random Features for Large-Scale Kernel Machines.' In: *Advances in Neural Information Processing Systems*. Ed. by J. Platt, D. Koller, Y. Singer, and S. Roweis. Vol. 20. Curran Associates, Inc., 2007 (cit. on pp. 105, 106, 124, 125, 158).

[334] Ali Rahimi and Benjamin Recht. 'Random features for large-scale kernel machines.' In: *Advances in neural information processing systems* 20 (2007) (cit. on pp. 7, 36, 166, 168, 178).

[335] Ali Rahimi and Benjamin Recht. 'Random features for large-scale kernel machines.' In: *Advances in neural information processing systems*. 2008, pp. 1177–1184 (cit. on pp. 45, 53).

[336] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. 'Hierarchical text-conditional image generation with clip latents.' In: *arXiv preprint arXiv:2204.06125* (2022) (cit. on p. 199).

[337] Sundeep Rangan, Philip Schniter, Erwin Riegler, Alyson K Fletcher, and Volkan Cevher. 'Fixed points of generalized approximate message passing

with arbitrary matrices.' In: *IEEE Transactions on Information Theory* 62.12 (2016), pp. 7464–7474 (cit. on pp. 20, 221).

[338] Maria Refinetti and Sebastian Goldt. 'The dynamics of representation learning in shallow, non-linear autoencoders.' In: *International Conference on Machine Learning*. PMLR. 2022, pp. 18499–18519 (cit. on pp. 184, 185, 187, 188, 191, 196).

[339] Maria Refinetti, Sebastian Goldt, Florent Krzakala, and Lenka Zdeborová. 'Classifying high-dimensional Gaussian mixtures: Where kernel methods fail and neural networks succeed.' In: *International Conference on Machine Learning*. PMLR. 2021, pp. 8936–8947 (cit. on p. 164).

[340] Maria Refinetti, Sebastian Goldt, Florent Krzakala, and Lenka Zdeborová. 'Classifying high-dimensional Gaussian mixtures: Where kernel methods fail and neural networks succeed.' In: *Proceedings of the 38th International Conference on Machine Learning*. Ed. by Marina Meila and Tong Zhang. Vol. 139. Proceedings of Machine Learning Research. PMLR, 18–24 Jul 2021, pp. 8936–8947 (cit. on pp. 106, 125, 168).

[341] Riccardo Rende, Federica Gerace, Alessandro Laio, and Sebastian Goldt. 'Optimal inference of a generalised Potts model by single-layer transformers with factored attention.' In: *arXiv preprint arXiv:2304.07235* (2023) (cit. on pp. 215, 227).

[342] Dominic Richards, Jaouad Mourtada, and Lorenzo Rosasco. 'Asymptotics of ridge (less) regression under general source condition.' In: *International Conference on Artificial Intelligence and Statistics*. PMLR. 2021, pp. 3889–3897 (cit. on pp. 66, 73).

[343] Daniel A. Roberts, Sho Yaida, and Boris Hanin. 'The Principles of Deep Learning Theory.' In: *ArXiv* abs/2106.10165 (2021) (cit. on p. 147).

[344] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. 'High-resolution image synthesis with latent diffusion models.' In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2022, pp. 10684–10695 (cit. on p. 199).

[345] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. 'U-net: Convolutional networks for biomedical image segmentation.' In: *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18*. Springer. 2015, pp. 234–241 (cit. on pp. 8, 204).

[346] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. 'U-net: Convolutional networks for biomedical image segmentation.' In: *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18*. Springer. 2015, pp. 234–241 (cit. on p. 186).

[347] Frank Rosenblatt. 'The perceptron: a probabilistic model for information storage and organization in the brain.' In: *Psychological review* 65.6 (1958), p. 386 (cit. on p. 7).

[348] Jonathan S Rosenfeld, Amir Rosenfeld, Yonatan Belinkov, and Nir Shavit. 'A constructive prediction of the generalization error across scales.' In: *arXiv preprint arXiv:1909.12673* (2019) (cit. on pp. 41, 80).

[349] Saharon Rosset, Ji Zhu, and Trevor Hastie. 'Margin Maximizing Loss Functions.' In: *NIPS*. 2003, pp. 1237–1244 (cit. on p. 54).

[350] Grant Rotskoff and Eric Vanden-Eijnden. 'Trainability and Accuracy of Artificial Neural Networks: An Interacting Particle System Approach.' In: *Communications on Pure and Applied Mathematics* 75.9 (2022), pp. 1889–1935 (cit. on pp. 35, 168).

[351] Francisco Rubio and Xavier Mestre. 'Spectral convergence for a general class of random matrices.' In: *Statistics & Probability Letters* 81.5 (2011), pp. 592–602 (cit. on p. 131).

[352] Anian Ruoss et al. 'Randomized Positional Encodings Boost Length Generalization of Transformers.' In: *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. Ed. by Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki. Toronto, Canada: Association for Computational Linguistics, July 2023, pp. 1889–1903 (cit. on p. 215).

[353] David Saad and Sara A. Solla. 'On-line learning in soft committee machines.' In: *Physical Review E* 52.4 (Oct. 1995), pp. 4225–4243 (cit. on p. 34).

[354] Chitwan Saharia et al. 'Photorealistic text-to-image diffusion models with deep language understanding.' In: *Advances in Neural Information Processing Systems* 35 (2022), pp. 36479–36494 (cit. on p. 199).

[355] Arda Sahiner, Tolga Ergen, Batu Ozturkler, John Pauly, Morteza Mardani, and Mert Pilanci. 'Unraveling attention via convex duality: Analysis and interpretations of vision transformers.' In: *International Conference on Machine Learning*. PMLR. 2022, pp. 19050–19088 (cit. on p. 215).

[356] Mojtaba Sahraee-Ardakan, Melikasadat Emami, Parthe Pandit, Sundeep Rangan, and Alyson K. Fletcher. 'Kernel Methods and Multi-layer Perceptrons Learn Linear Models in High Dimensions.' In: *ArXiv* abs/2201.08082 (2022) (cit. on pp. 157, 160).

[357] Fariborz Salehi, Ehsan Abbasi, and Babak Hassibi. 'The Performance Analysis of Generalized Margin Maximizers on Separable Data.' In: *International Conference on Machine Learning*. PMLR. 2020, pp. 8417–8426 (cit. on pp. 45, 48).

[358] B. Scholkopf and A.J. Smola. *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond*. Adaptive Computation and Machine Learning. MIT Press, 2018 (cit. on p. 54).

[359] Bernhard Schölkopf and Alexander J Smola. *Learning with kernels: support vector machines, regularization, optimization, and beyond*. MIT press, 2002 (cit. on p. 81).

[360] Dominik Schröder, Hugo Cui, Daniil Dmitriev, and Bruno Loureiro. 'Deterministic equivalent and error universality of deep random features learning.' In: *International Conference on Machine Learning* 40 (2023), pp. 30285–30320 (cit. on pp. 124, 125, 128, 135–137, 139, 165, 168, 174).

[361] Dominik Schröder, Hugo Cui, Daniil Dmitriev, and Bruno Loureiro. 'Deterministic equivalent and error universality of deep random features learning.' In: *ArXiv* abs/2302.00401 (2023) (cit. on pp. 148, 155).

[362] Henry Schwarze. 'Learning a rule in a multilayer neural network.' In: *Journal of Physics A: Mathematical and General* 26.21 (1993), p. 5781 (cit. on pp. 10, 34, 97, 145).

[363] Mohamed El Amine Seddik, Cosme Louart, Mohamed Tamaazousti, and Romain Couillet. 'Random matrix theory proves that deep learning representations of gan-data behave as gaussian mixtures.' In: *International Conference on Machine Learning*. PMLR. 2020, pp. 8573–8582 (cit. on pp. 49, 59).

[364] Inbar Seroussi, Gadi Naveh, and Zohar Ringel. 'Separation of scales and a thermodynamic description of feature learning in some CNNs.' In: *Nature Communications* 14.1 (2023), p. 908 (cit. on p. 86).

[365] Inbar Seroussi, Gadi Naveh, and Zohar Ringel. 'Separation of scales and a thermodynamic description of feature learning in some CNNs.' In: *Nature Communications* 14.1 (Feb. 2023), p. 908 (cit. on p. 168).

[366] Hyunjune Sebastian Seung, Haim Sompolinsky, and Naftali Tishby. 'Statistical mechanics of learning from examples.' In: *Physical review A* 45.8 (1992), p. 6056 (cit. on pp. 10–12, 34, 40, 45, 48, 97, 145, 147).

[367] Kulin Shah, Sitan Chen, and Adam Klivans. 'Learning Mixtures of Gaussians Using the DDPM Objective.' In: *arXiv preprint arXiv:2307.01178* (2023) (cit. on pp. 199, 201, 204).

[368] Utkarsh Sharma and Jared Kaplan. 'Scaling Laws from the Data Manifold Dimension.' In: *J. Mach. Learn. Res.* 23 (2022), pp. 9–1 (cit. on p. 80).

[369] Peter Shaw, Jakob Uszkoreit, and Ashish Vaswani. 'Self-Attention with Relative Position Representations.' In: *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*. NAACL-HLT 2018. New Orleans, Louisiana: Association for Computational Linguistics, June 2018, pp. 464–468 (cit. on p. 215).

[370] A. E. Shevchenko, Kevin Kögler, Hamed Hassani, and Marco Mondelli. 'Fundamental Limits of Two-layer Autoencoders, and Achieving Them with Gradient Methods.' In: *ArXiv* abs/2212.13468 (2022) (cit. on pp. 184, 185, 188, 191, 196).

[371] Takashi Shinzato and Yoshiyuki Kabashima. 'Learning from correlated patterns by simple perceptrons.' In: *Journal of Physics A: Mathematical and Theoretical* 42.1 (2008), p. 015005 (cit. on p. 37).

[372] Takashi Shinzato and Yoshiyuki Kabashima. 'Perceptron capacity revisited: classification ability for correlated patterns.' In: *Journal of Physics A: Mathematical and Theoretical* 41.32 (2008), p. 324013 (cit. on p. 37).

[373] J.W. Silverstein. 'Strong Convergence of the Empirical Distribution of Eigenvalues of Large Dimensional Random Matrices.' In: *Journal of Multivariate Analysis* 55.2 (1995), pp. 331–339 (cit. on p. 131).

[374] Koustuv Sinha, Amirhossein Kazemnejad, Siva Reddy, Joelle Pineau, Dieuwke Hupkes, and Adina Williams. 'The Curious Case of Absolute Position Embed-

dings.' In: *Findings of the Association for Computational Linguistics: EMNLP 2022*. Ed. by Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang. Abu Dhabi, United Arab Emirates: Association for Computational Linguistics, Dec. 2022, pp. 4449–4472 (cit. on p. 215).

[375] Justin Sirignano and Konstantinos Spiliopoulos. 'Mean field analysis of neural networks: A central limit theorem.' In: *Stochastic Processes and their Applications* 130.3 (2020), pp. 1820–1852 (cit. on p. 168).

[376] Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. 'Deep unsupervised learning using nonequilibrium thermodynamics.' In: *International conference on machine learning*. PMLR. 2015, pp. 2256–2265 (cit. on pp. 8, 181).

[377] Haim Sompolinsky, Naftali Tishby, and H Sebastian Seung. 'Learning from examples in large neural networks.' In: *Physical Review Letters* 65.13 (1990), p. 1683 (cit. on pp. 10–12, 34, 40, 97).

[378] Yang Song and Stefano Ermon. 'Generative modeling by estimating gradients of the data distribution.' In: *Advances in neural information processing systems* 32 (2019) (cit. on pp. 199, 200).

[379] Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. 'Score-based generative modeling through stochastic differential equations.' In: *arXiv preprint arXiv:2011.13456* (2020) (cit. on pp. 199, 200).

[380] Yang Song, Jascha Sohl-Dickstein, Diederik P. Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. 'Score-Based Generative Modeling through Stochastic Differential Equations.' In: *International Conference on Learning Representations* (2021) (cit. on pp. 184, 186).

[381] Stefano Spigler, Mario Geiger, Stéphane d'Ascoli, Levent Sagun, Giulio Biroli, and Matthieu Wyart. 'A jamming transition from under-to over-parametrization affects generalization in deep learning.' In: *Journal of Physics A: Mathematical and Theoretical* 52.47 (2019), p. 474001 (cit. on p. 58).

[382] Stefano Spigler, Mario Geiger, and Matthieu Wyart. 'Asymptotic learning curves of kernel methods: empirical data versus teacher-student paradigm.' In: *Journal of Statistical Mechanics: Theory and Experiment* 2020.12 (2020), p. 124001 (cit. on pp. 64, 65, 67, 70, 78, 85).

[383] Ingo Steinwart and Andreas Christmann. 'Support Vector Machines.' In: *Information science and statistics*. 2008 (cit. on pp. 81, 82, 87, 89, 93).

[384] Ingo Steinwart, Don R Hush, Clint Scovel, et al. 'Optimal Rates for Regularized Least Squares Regression.' In: *COLT*. 2009, pp. 79–93 (cit. on pp. 55, 64, 65, 81).

[385] Ingo Steinwart and Clint Scovel. 'Fast rates for support vector machines using Gaussian kernels.' In: *Annals of Statistics* 35 (2007), pp. 575–607 (cit. on pp. 81, 82).

[386] Mihailo Stojnic. 'A framework to characterize performance of lasso algorithms.' In: *arXiv preprint arXiv:1303.7291* (2013) (cit. on p. 49).

[387] Mihailo Stojnic. 'Meshes that trap random subspaces.' In: *arXiv preprint arXiv:1304.0003* (2013) (cit. on p. 173).

[388] Mihailo Stojnic. 'Upper-bounding L1-optimization weak thresholds.' In: *arXiv preprint arXiv:1303.7289* (2013) (cit. on p. 173).

[389] Michel Talagrand. 'Concentration of measure and isoperimetric inequalities in product spaces.' In: *Publications Mathématiques de l'Institut des Hautes Études Scientifiques* 81 (1994), pp. 73–205 (cit. on p. 67).

[390] Michel Talagrand. 'The parisi formula.' In: *Annals of mathematics* (2006), pp. 221–263 (cit. on p. 147).

[391] Davoud Ataee Tarzanagh, Yingcong Li, Christos Thrampoulidis, and Samet Oymak. 'Transformers as support vector machines.' In: *arXiv preprint arXiv:2308.16898* (2023) (cit. on p. 215).

[392] Davoud Ataee Tarzanagh, Yingcong Li, Xuechen Zhang, and Samet Oymak. 'Max-margin token selection in attention mechanism.' In: *Thirty-seventh Conference on Neural Information Processing Systems*. 2023 (cit. on p. 215).

[393] Matthias Thamm, Max Staats, and Bernd Rosenow. 'Random matrix analysis of deep neural network weight matrices.' In: *Phys. Rev. E* 106.5 (2022), Paper No. 054124, 15 (cit. on p. 124).

[394] Christos Thrampoulidis, Ehsan Abbasi, and Babak Hassibi. 'Precise Error Analysis of Regularized M-Estimators in High Dimensions.' In: *IEEE Transactions on Information Theory* 64.8 (2018), pp. 5592–5628 (cit. on pp. 11, 48, 145, 147, 186).

[395] Christos Thrampoulidis, Samet Oymak, and Babak Hassibi. 'The Gaussian min-max theorem in the presence of convexity.' In: *arXiv preprint arXiv:1408.4837* (2014) (cit. on pp. 11, 173).

[396] Yuandong Tian, Yiping Wang, Beidi Chen, and Simon Du. 'Scan and Snap: Understanding Training Dynamics and Token Composition in 1-layer Transformer.' In: *arXiv preprint arXiv:2305.16380* (2023) (cit. on p. 215).

[397] Tong Tong, Gen Li, Xiejie Liu, and Qinquan Gao. 'Image super-resolution using dense skip connections.' In: *Proceedings of the IEEE international conference on computer vision*. 2017, pp. 4799–4807 (cit. on p. 186).

[398] Alexander Tsigler and P. Bartlett. 'Benign overfitting in ridge regression.' In: *arXiv preprint arXiv:2009.14286* (2020) (cit. on pp. 67, 70).

[399] Leslie G Valiant. 'A theory of the learnable.' In: *Communications of the ACM* 27.11 (1984), pp. 1134–1142 (cit. on p. 10).

[400] Laurens Van der Maaten and Geoffrey Hinton. 'Visualizing data using t-SNE.' In: *Journal of machine learning research* 9.11 (2008) (cit. on p. 36).

[401] Aditya Varre, Loucas Pillaud-Vivien, and Nicolas Flammarion. 'Last iterate convergence of SGD for Least-Squares in the Interpolation regime.' In: *ArXiv* abs/2102.03183 (2021) (cit. on pp. 65, 70, 80, 81).

[402] Ashish Vaswani et al. 'Attention is all you need.' In: *Advances in neural information processing systems* 30 (2017) (cit. on pp. 8, 181, 182, 214, 215, 218).

[403] Andrea Della Vecchia, Jaouad Mourtada, Ernesto de Vito, and Lorenzo Rosasco. 'Regularized ERM on random subspaces.' In: *ArXiv* abs/2006.10016 (2021) (cit. on p. 89).

[404] Pascal Vincent. 'A connection between score matching and denoising autoencoders.' In: *Neural computation* 23.7 (2011), pp. 1661–1674 (cit. on pp. 199, 201).

[405] Pascal Vincent, Hugo Larochelle, Isabelle Lajoie, Yoshua Bengio, and Pierre-Antoine Manzagol. 'Stacked Denoising Autoencoders: Learning Useful Representations in a Deep Network with a Local Denoising Criterion.' In: *J. Mach. Learn. Res.* 11 (2010), pp. 3371–3408 (cit. on pp. 8, 184, 186, 187, 199, 201).

[406] Johannes Von Oswald et al. 'Transformers learn in-context by gradient descent.' In: *Proceedings of the 40th International Conference on Machine Learning*. ICML'23. Honolulu, Hawaii, USA: JMLR.org, 2023 (cit. on p. 215).

[407] Timothy LH Watkin, Albrecht Rau, and Michael Biehl. 'The statistical mechanics of learning a rule.' In: *Reviews of Modern Physics* 65.2 (1993), p. 499 (cit. on pp. 40, 45, 48, 145, 147).

[408] Geoffrey S. Watson. 'Smooth Regression Analysis.' In: *Sankhyā: The Indian Journal of Statistics, Series A (1961-2002)* 26.4 (1964), pp. 359–372 (cit. on p. 64).

[409] Gail Weiss, Yoav Goldberg, and Eran Yahav. 'Thinking Like Transformers.' In: *Proceedings of the 38th International Conference on Machine Learning*. Ed. by Marina Meila and Tong Zhang. Vol. 139. Proceedings of Machine Learning Research. PMLR, 18–24 Jul 2021, pp. 11080–11090 (cit. on pp. 215, 216).

[410] Andre Wibisono and Kaylee Yingxi Yang. 'Convergence in KL divergence of the inexact Langevin algorithm with application to score-based generative models.' In: *arXiv preprint arXiv:2211.01512* (2022) (cit. on pp. 199, 201).

[411] Christopher K. I. Williams. 'Computing with Infinite Networks.' In: *Proceedings of the 9th International Conference on Neural Information Processing Systems*. NIPS'96. Denver, Colorado: MIT Press, 1996, pp. 295–301 (cit. on pp. 41, 64).

[412] Christopher K. I. Williams. 'Computing with Infinite Networks.' In: *Proceedings of the 9th International Conference on Neural Information Processing Systems*. NIPS'96. Denver, Colorado: MIT Press, 1996, pp. 295–301 (cit. on pp. 47, 97).

[413] John Wishart. 'The generalised product moment distribution in samples from a normal multivariate population.' In: *Biometrika* (1928), pp. 32–52 (cit. on p. 108).

[414] Denny Wu and Ji Xu. 'On the Optimal Weighted ℓ_2 Regularization in Overparameterized Linear Regression.' In: *Advances in Neural Information Processing Systems*. Vol. 33. 2020 (cit. on pp. 49, 66, 68, 73).

[415] Denny Wu and Ji Xu. 'On the Optimal Weighted ℓ_2 Regularization in Overparameterized Linear Regression.' In: *Advances in Neural Information Processing Systems*. Vol. 33. Curran Associates, Inc., 2020, pp. 10112–10123 (cit. on p. 159).

[416] H. Xiao, K. Rasul, and Roland Vollgraf. 'Fashion-MNIST: a Novel Image Dataset for Benchmarking Machine Learning Algorithms.' In: *ArXiv* abs/1708.07747 (2017) (cit. on pp. 58, 76, 85, 94, 185, 193).

[417] Lechao Xiao, Hong Hu, Theodor Misiakiewicz, Yue Lu, and Jeffrey Pennington. 'Precise Learning Curves and Higher-Order Scalings for Dot-product Kernel Regression.' In: *Advances in Neural Information Processing Systems* 35 (2022), pp. 4558–4570 (cit. on pp. 11, 34, 162, 166).

[418] Sho Yaida. 'Non-Gaussian processes and neural networks at finite widths.' In: *Mathematical and Scientific Machine Learning*. 2019 (cit. on pp. 147, 151).

[419] Sho Yaida. 'Non-Gaussian processes and neural networks at finite widths.' In: *Proceedings of The First Mathematical and Scientific Machine Learning Conference*. Ed. by Jianfeng Lu and Rachel Ward. Vol. 107. Proceedings of Machine Learning Research. PMLR, 20–24 Jul 2020, pp. 165–192 (cit. on p. 168).

[420] Gilad Yehudai and Ohad Shamir. 'On the Power and Limitations of Random Features for Understanding Neural Networks.' In: *Advances in Neural Information Processing Systems*. Vol. 32. Curran Associates, Inc., 2019 (cit. on pp. 106, 125).

[421] Hui Yuan, Kaixuan Huang, Chengzhuo Ni, Minshuo Chen, and Mengdi Wang. 'Reward-Directed Conditional Diffusion: Provable Distribution Estimation and Reward Improvement.' In: *arXiv preprint arXiv:2307.07055* (2023) (cit. on pp. 199, 201).

[422] Jacob Zavatone-Veth and Cengiz Pehlevan. 'Exact marginal prior distributions of finite Bayesian neural networks.' In: *Neural Information Processing Systems*. 2021 (cit. on pp. 105, 106).

[423] Jacob Zavatone-Veth and Cengiz Pehlevan. 'Learning curves for deep structured Gaussian feature models.' In: *arXiv preprint arXiv:2303.00564* (2023) (cit. on pp. 124, 125, 168).

[424] Jacob Zavatone-Veth, Cengiz Pehlevan, and William L. Tong. 'Contrasting random and learned features in deep Bayesian linear regression.' In: *Phys. Rev. E* 105.6 (2022) (cit. on pp. 36, 106, 124–126, 147, 186).

[425] Jacob Zavatone-Veth et al. 'Asymptotics of representation learning in finite Bayesian neural networks.' In: *J. Stat. Mech. Theory Exp.* 2022.11 (2022) (cit. on p. 147).

[426] Lenka Zdeborová and Florent Krzakala. 'Statistical physics of inference: Thresholds and algorithms.' In: *Advances in Physics* 65.5 (2016), pp. 453–552 (cit. on pp. xviii, 10, 14, 18, 21–23, 34, 37, 45, 97).

[427] Lenka Zdeborová and Florent Krzakala. 'Statistical physics of inference: thresholds and algorithms.' In: *Advances in Physics* 65 (2015), pp. 453–552 (cit. on pp. 151, 153, 185).

[428] C. Zhang, S. Bengio, M. Hardt, B. Recht, and O. Vinyals. 'Understanding deep learning requires rethinking generalization.' In: *ICLR*. 2017 (cit. on p. 45).

[429] Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. 'Understanding deep learning (still) requires rethinking generalization.' In: *Communications of the ACM* 64.3 (2021), pp. 107–115 (cit. on p. 9).

[430] Ruiqi Zhang, Spencer Frei, and Peter L Bartlett. 'Trained Transformers Learn Linear Models In-Context.' In: *arXiv preprint arXiv:2306.09927* (2023) (cit. on p. 215).

[431] Ziqian Zhong, Ziming Liu, Max Tegmark, and Jacob Andreas. *The Clock and the Pizza: Two Stories in Mechanistic Explanation of Neural Networks*. 2023. arXiv: *2306.17844* [cs.LG] (cit. on p. 216).

RESUME

For a full up-to-date resume please visit *https://hugocui.github.io/*.

EDUCATION

2020-2024 *EPFL, Lausanne, Switzerland*
PhD student, Statistical Physics of Computation laboratory (SPOC), advised by Lenka Zdeborová.

2019 *IPhT, Paris, France*
Master thesis, advised by Lenka Zdeborová.

2017-2019 *ENS, Paris, France*
Master's degree, Theoretical Physics (International Center for Fundamental Physics)

2016-2017 *ENS, Paris, France*
Bachelor's degree, Physics

TALKS

- LemanTh 2024, *invited speaker*
- EurOPT 2024, *invited speaker*
- ICLR 2021, *poster*
- Youth in high dimensions 2024 ICTP, *invited speaker*
- NeurIPS 2023 workshop on diffusion models, *poster*
- NeurIPS 2023, *spotlight poster*
- 5th International Workshop on neural scaling laws, *invited speaker*
- ICML 2023, *oral*
- Cargèse summer school 2023, Machine learning and statisical physics back together, *contributed talk*
- ITS seminar, CUNY, *invited speaker*
- les Houches 2023 workshop, Towards an understanding of artificial and biological networks, *poster*
- Youth in high dimensions 2023 ICTP, *poster*
- EPFL-RIKEN young rising stars joint workshop, *invited speaker*
- Learning: Optimization and Stochastics Summer Research Institute, *invited speaker*
- Learning and Optimization, CIRM conference, *contributed talk*
- 4th IMA Conference on the mathematical challenges of big Data, Oxford, *poster*

- Workshop on the theory of overparameterized machine learning (TOPML 2022), *contributed talk*
- Advanced Course on Data and Learning Summer School (ACDL 2022), *poster and talk, best presentation award*
- Fundamental problems in statistical physics summer school, *poster*
- Cargèse 2021 summer school, Glassy systems and interdisciplinary applications, *poster*
- NeurIPS 2021, *poster*
- Fundamentals of Learning and AI Research (FLAIR), EPFL, *invited speaker*
- MSML 2020, *contributed talk*

TEACHING

2023	Machine learning for physicists, TA, *Masters, EPFL*
2021-2022	Statistical physics of computation, TA, *Masters, EPFL*
2022	Statistical physics II, TA, *Bachelors, EPFL*
2021	Physics for earth sciences, TA, *Bachelors, Unil*
2021	Physics I, TA, *Bachelors, EPFL*
2023	Supervision of Nolan Sandgathe (Master thesis)
2021	Supervision of Oscar Bouverot-Dupuis (Master internship)

DISTINCTIONS

- G-research PhD prize in Mathematics and Data Science 2024, EPFL,3rd *prize*
- Famelab 2021 science communication competition, *Switzerland national winner and international finalist*
- ENS entrance national competitive exam 2016, *ranked* 1st.
- Laureate of two thematic awards (bronze medal at the International Chemistry Olympiads, 2nd prize at French Chemistry Olympiads) from the French Académie des Sciences, 2014.

Ph.D dissertation of Hugo Cui
Topics in Statistical Physics of High-Dimensional Machine Learning © May 6
2024

SUPERVISORS:
Lenka Zdeborová

LOCATION:
Institute of Physics
EPFL, Lausanne, Switzerland

TIME FRAME:
September 1, 2020 — May 6, 2024