PSL UNIVERSITÉ PARIS

HABILITATION À DIRIGER DES RECHERCHES

DE L'UNIVERSITÉ PSL

Présentée à l'École normale supérieure

Statistical Physics of two-layer neural networks: from kernels to feature learning

Présentation des travaux par

Bruno LOUREIRO

Le 16 Septembre 2025

Discipline

Informatique

Composition du jury :

Gérard BEN AROUS Professeur, NYU	*Rapporteur*
Jamal NAJIM Directeur de Recherche, CNRS	*Rapporteur*
Nathan SREBRO Professeur, TTIC	*Rapporteur*
Francis BACH Directeur de Recherche, INRIA	*Examinateur*
Julie DELON Professeure, ENS Paris	*Examinatrice*
Rémi GRIBONVAL Directeur de Recherche, INRIA	*Examinateur*
Julia KEMPE Professeure, NYU & META	*Examinatrice*
Marc MÉZARD Professeur, Bocconi U	*Examinateur*
Gabriel PEYRÉ Directeur de Recherche, CNRS	*Examinateur*

ENS | PSL

Everybody knows that it's so hard to dig and get to the root.

Caetano Veloso in *Maria Bethânia*.

Acknowledgments

First and foremost, I wish to express my gratitude to my collaborators. Science is, above all, a collective endeavour, and what I value most in this profession is the opportunity to engage with colleagues. The intellectually stimulating exchanges, the chance both to challenge and to be challenged by talented minds on a daily basis, the sense of belonging to a scholarly community, and the satisfaction that comes with finally grasping an idea after hours, days, or even months of persistent thought, are all but privileges. They are, without question, a major source of motivation.

I am deeply indebted to Florent Krzakala and Lenka Zdeborová, from whom I learned much about our craft, statistical physics and mushrooms. I am also grateful to Hugo Cui, Ludovic Stephan, Cédric Gerbelot, Yue Lu, Marc Mézard, Sebastian Goldt, Federica Gerace, Antoine Maillard, Gabriele Sicuro, Marylou Gabrié, Theodor Misiakiewicz, Rodrigo Veiga, Francesca Mignacco, Zhou Fan, Elisabetta Cornacchia, Yatin Dandi, Luca Pesce, Luca Arnaboldi, Lucas Clarté, Matteo Vilucchio, Daniil Dmitriev, Emanuele Troiani, and Benjamin Aubin. Thank you for the richness of our exchanges; I hope you found the process of uncovering the results presented here as rewarding as I did.

Secondly, I thank my colleagues at the Centre for Data Science at ENS, who make this an intellectually stimulating, supportive, and enjoyable place to work. I am grateful to Gabriel Peyré, our director and indefatigable VP of all matters; Olivier Cappé, for his HDR encouragement and halloumi connaisseur advice; Kimia Nadjahi, office mate and Persian cuisine adviser; Giulio Biroli and Marylou Gabrié, companions in food and physics; Stéphane Mallat, for many enriching discussions; and Julie Delon, already a valued ally in eating out. I also thank my colleagues at the Département d'Informatique, and in particular Lise-Marie Bivard, whose guidance was invaluable in navigating French bureaucracy.

Thirdly, I am grateful to my group — Leonardo Defilippis, Arie Worstman, Alexis Aymé, Clément Loup-Forest, and Luigi Fogliani — for their trust, hard work, and understanding during my absence while writing this manuscript.

Finally, I thank my family and friends, whose encouragement sustained me throughout this journey. Above all, I am profoundly grateful to my partner in life and academia, Lavinia Maddaluno, as well as to my mother, Marcia Loureiro, and my father, Hugo Vermandel. Without their unwavering support, none of this work would have been possible.

Contents

Nomenclature

Abbreviations

(c)GET (conditional) Gaussian equivalence theorem

(C)SQ (Correlational) Statistical query

(S)GD (Stochastic) Gradient descent

(s)RF (Spiked) Random features

AMP Approximate message massing

Eq(s). equation(s)

ERM Empirical risk minimisation

GMIM Gaussian multi-index model

i.i.d. Independent and identically distributed

KRR Kernel ridge regression

ML Machine learning

p.s.d. Positive semi-definite

RFRR Random features ridge regression

RMT Random matrix theory

Symbols

1_d All-ones vector in dimension d

γ_d The d-dimensional standard normal probability density function.

$\mathbb{E}[\cdots]$ Expected value

$\mathbb{I}(A)$ Indicator function of a set A.

$\mathbb{R}^d$ d-dimensional Euclidean space

$\mathcal{N}(\mu, \Sigma)$ Multi-variate Gaussian distribution with mean μ and covariance Σ.

$\mathcal{S}_r^+$ The cone of positive semi-definite symmetric matrices.

$\|A\|_{\mathrm{F}}$ The Fröbenius norm of a matrix.

$\|A\|_{\mathrm{op}}$ The operator norm of a matrix.

$\|x\|_p$ The ℓ^p norm of a vector or a sequence.

$a \wedge b = \min(a, b)$ The minimum between $a, b \in \mathbb{R}$.

I_d $d \times d$ identity matrix.

$L^2(\mu)$ The space of square integrable functions with respect to the measure μ.

$\mathrm{Lip}_d(L)$ The space of L-Lipschitz functions over $\mathbb{R}^d$.

Preface

The non-French-speaking readers of this manuscript will sympathise with my first thoughts upon discovering that I had to write a *mémoire de recherche*. Often used in its *literary* sense in other languages, the prospect of writing a *mémoire* about one's own research can sound daunting to the scientifically minded. Behind every paper there is a story, and although stories inhabit the same mental space as the ideas from which papers are made, they do not, understandably, belong to the same written space. Nevertheless, the idea of recording the research memories behind the research ideas discussed here grew irresistibly enticing. I do not expect, however, that all readers will be interested in these memories. If you are among them, I suggest skipping directly to the science in Chapter 1.

If I had to synthesise my research path thus far in a single word, this word would be *interdisciplinary*. Trained as a high-energy theoretical physicist during my undergraduate and master's studies, I joined the Theory of Condensed Matter (TCM) group at the University of Cambridge for my doctoral thesis in 2014. At the time, a mathematical duality connecting two very different fields of physics — gravity and quantum field theory — was beginning to make its way into condensed matter theory, with the promise that hard problems on one side could be mapped to more tractable problems on the other. My supervisor, Antonio Miguel García-García, was looking for a student with a high-energy background to engage with this programme, and that's how I crossed the street from the Department of Mathematics to the Cavendish Laboratory.

Continuous exposure to condensed matter research during my TCM years grew into an interest in statistical physics, and by the end of my PhD in 2018 I was determined to move into this field. Casually googling for "*postdoctoral positions in statistical physics*" led me to the webpage of Lenka Zdeborová, who was announcing a position at the crossroads of statistical physics and machine learning. My first day of work in my new postdoc position was at the *Institut d'Études Scientifiques* in Cargèse, where Lenka was organising, together with Florent Krzakala, a conference "*Statistical Physics and Machine Learning back together*". My main recollection of this conference is the overwhelming feeling of watching a full-day programme on topics I had barely heard of, but that I was supposed to work on in the weeks to follow.

I particularly remember being struck by how many problems fit within a single posterior distribution: Stochastic Block Model, planted PCA, $\mathbb{Z}_2$ synchronisation, Restricted Boltzmann Machine, Sherrington–Kirkpatrick model, Hopfield model, low-rank matrix factorisation, etc. — making me strongly appreciate the *universality* of the statistical physics approach. But as important as the science, it was on the beach of the institute that I first met some of my future collaborators, such as Marylou Gabrié, Sebastian Goldt and Marc Mézard, and some of my current colleagues at ENS, such as Giulio Biroli, Stéphane Mallat and Guilhem Semerjian.

IPhT years (2018-2020) — My first two years of postdoc, from 2018–2020, focused on studying non-convex inverse problems such as sparse PCA and phase retrieval. From a probabilistic perspective, families of inverse problems can be classified by their factor graphs, and the statistical physics approach provides tools to characterise the asymptotic properties of the associated *free energy* (or, equivalently, the *mutual information*). There are not many factor graphs for which these quantities can be explicitly computed. For this reason, Lenka and I were interested in studying the *modularity* of this approach: given two factor graphs for which the free energy is computable, can we compute the free energy of the composed factor graph? Recent results at the time suggested a positive answer [Man+17; FRS18; Gab+18], and what I was able to show is that these particular results follow from a general *composability rule*, where the free energy of a composite graph can be systematically obtained from the free energy of the individual factor graphs. This result was particularly relevant to the emerging field of using pre-trained deep networks as priors to solve hard inverse problems, also known as *generative priors*. Leveraging this result, we were able to show that random deep generative priors were computationally more favourable than classically employed compression schemes, such as sparse priors [Aub+19; Aub+20]. We presented our results at both the main conference and the first Deep Inverse workshop, a workshop at NeurIPS entirely dedicated to this problem.

The problem of composability of graphical models naturally led us to explore the same question for the quenched disorder. Indeed, one of the major drawbacks of the statistical physics approach is that it strongly relies on the randomness in the model,[1] for instance in the form of a random graph in the context of constraint satisfaction problems, or a random data distribution in the context of learning. Understanding how to deal with more structured disorder without breaking mathematical tractability was, and still is, an important question in the field (c.f. this excellent review by Marc Mézard on the topic [Méz24]). Sebastian Goldt, who was also a postdoc in Lenka's group at the time, was exploring this question in the context of one-pass SGD. Together with Lenka, Marc and Florent, he introduced a random model for data generated from a latent space: the *Hidden Manifold Model* (HMM)

[1] Also known as the *disorder* in the statistical physics parlance.

[Gol+20]. I began working on a replica computation for this model during a one-month visit to the *Kavli Institute of Theoretical Physics* (KITP) for a programme on machine learning and physics, where I shared a kitchen with Sebastian and had many great discussions between cooking and dinner.[2] Solving this problem led us to the notion of *Gaussian universality*: the idea that since data is only seen by these models through a low-dimensional projection,[3] only lower order statistics matter in the asymptotic limit [Ger+20; Gol+22]. Exploring the extents and limitations of Gaussian universality permeated a lot of my subsequent research, which will be discussed in depth in Chapter 2.

EPFL years (2020-2022) — In September 2020, after enduring a long COVID lockdown in a small and damp Parisian apartment, Florent Krzakala proposed that I move with him to Lausanne and join the newly established *Information, Learning & Physics Laboratory* at EPFL. I have fond memories of that first autumn in Lausanne, which I spent helping Florent set up the new lab — in particular choosing a coffee machine — and finishing some ongoing work I had started in Paris on the extent to which the asymptotic formulas we had derived capture the behaviour of learning curves for real data [Gol+22; Lou+21a]. The observation that sometimes the asymptotic formulas extended well beyond Gaussian data led me to think about universality, and the extent to which it holds as a function of the task. With Gabriele Sicuro, we had derived formulas for the Gaussian mixture model [Lou+21b], which became a playground to study the limitations of universality [Ger+24; Pes+23].

This was also the period when I first learned about the source & capacity literature — or *scaling laws*, as they are referred to nowadays. At the time, there was an ongoing discussion between the groups of Matthieu Wyart at EPFL and Cengiz Pehlevan at Harvard about the rates of convergence of kernel methods they had derived in concurrent works [SGW20a; BCP20a]. Loucas-Pillaud Vivien, who was moving from Paris to EPFL after a PhD with Alessandro Rudi and Francis Bach, attended the seminars we jointly organised with Matthieu Wyart's group to present his work on this topic [PRB18]. As he introduced the classical results of Caponetto and De Vito on the optimal rates for kernel ridge regression [CD07], we realised (after some annoying notation translation) that these rates were faster than those in [SGW20a; BCP20a], apparently contradicting optimality. Florent and Lenka had just bought a pizza stone for their new Swiss barbecue grill, and I was experimenting with Neapolitan-style pizza recipes. So we made Loucas an irresistible invitation: all you can eat pizza in exchange for a summary of the kernel source & capacity literature. Together with Hugo Cui, who had just started his PhD at EPFL, we came to understand that the puzzle was due to the fact that the Pehlevan and Wyart result held in a data-limited regime where the rates are effectively noiseless, and that there exists a crossover to the rates of Caponetto and

[2] Those who have been to the KITP will probably also have fond memories of the blackboards in the kitchen of the Munger residence.

[3] Also known as the *local fields*.

De Vito as the amount of data increases beyond a certain threshold related to the noise level in the problem [Cui+21]. We also found the existence of new rates and associated crossovers, such as to plateau regimes which are of interest to the neural scaling law literature [Kap+20; Bah+24; MRS22].

The year 2021 also marked the beginning of my co-supervision of students, starting with the first cohort of Master's students at IdePHICS, Luca Pesce and Alessandro Pacco, followed by Lucas Clarté and Luca Arnaboldi (it became something of a tradition that many students were named Luca). In the same year, we welcomed Rodrigo Veiga, then a PhD student at the University of São Paulo, for a one-year research visit to IdePHICS. At the time, I had not worked on (S)GD and regarded this as an important gap in my toolbox. In brainstorming with Florent about possible projects for Rodrigo, I proposed to revisit the high-dimensional analysis of two-layer neural networks pioneered by David Saad and Sara Solla in the 1990s [SS95c; SS95b; SS95a], with the aim of extending it to the infinite-width limit and, in particular, bridging it with recent advances on the mean-field limit [CB18; MMN18; SS20; RV22]. Together with Ludovic Stéphan, who joined IdePHICS also as a postdoc later in 2021, we showed that the ODE description in terms of summary statistics could be derived in different complementary regimes, with the only structural difference in the equations being a single term proportional to the ratio between the learning rate and the network width [Vei+22]. This additional term, stemming from the variance of the SGD noise at finite learning rates, accounted for the asymptotic plateau observed at long times [SS96]. Francis Bach later pointed out to me that this plateau was the variance of the stationary distribution of SGD at finite learning rates, and had been characterised non-asymptotically in the least-squares settig [DB16]. A limitation of our first work was that the summary statistics were matrices whose size scaled with the width of the network. In a follow-up work with Luca Arnaboldi, we were able to simplify the equations further in the joint large-width and high-dimensional limit, obtaining a PDE for the evolution of the summary statistics that is consistent with the mean-field description [Arn+23]. Concurrently, Raphael Berthier, Andrea Montanari, and Kangjie Zhou reached the same conclusion from the opposite direction, by reducing the mean-field PDE in the high-dimensional limit to an ODE. Taken together, these results establish that the two limits commute [BMZ24].

Towards the end of my postdoc at EPFL in 2022, I began to take a strong interest in feature learning. It started in July 2021, when Denny Wu, then a PhD student in Toronto, wrote to me with questions about the scope of our results in [Lou+21a]. In particular, he wanted to know whether our universality results would still hold for a random features model in which the first-layer weights are correlated with the target directions — a setting motivated by gradient-step corrections to the initialisation of two-layer neural networks. In April 2022, Denny extended the Gaussian universality results and derived exact asymptotics for a random features model with trained weights in the $\Theta(1)$ learning-rate regime [Ba+22].

His analysis showed that in this regime, feature-learning corrections did not increase the expressivity of the model, which would instead require extensive learning rates. This result sparked my own interest in whether the asymptotic analysis of [Lou+21a] could be extended to the feature-learning regime. I invited Denny to EPFL in June 2022 to present his findings. One of Denny's result was that a single large gradient step is asymptotically equivalent to adding a spike correlated with the target weights. I realised that this was closely related to the Gaussian mixture model we had previously studied, which suggested that the large-step regime might also be analytically tractable. Excited by this connection, I proposed it as an open problem at the Les Houches 2022 summer school.

The ENS years (2022-to date) — I moved to my current position at the *Département d'Informatique* of ENS in October 2022. I dedicated my first months to the problem of feature learning together with Yatin Dandi, who had just started his PhD at IdePHICS. Motivated by the simpler setting of random features on a Gaussian mixture distribution, our first step was to generalise a proof scheme by Andrea Montanari and Basil Saeed [MS22] to mixture distributions, leading to a *conditional* form of Gaussian universality [Dan+23]. This advance was instrumental for addressing the one-step problem, as it established that whenever the non-Gaussian component of a high-dimensional distribution is low-dimensional (e.g. a spike), it can be factored out and treated separately. Building on this principle, which we termed *conditional Gaussian equivalence*, we progressively solved the problem of characterising large-step corrections to random features: first by deriving upper bounds [Dan+24b], then through a replica-based analysis [Cui+24], and finally with a rigorous random matrix theory proof [Dan+25]. These results precisely characterised feature learning after a large gradient step and clarified its effect on the performance. They will be discussed in detail in chapter 3.

In parallel, I began engaging with my new colleagues at ENS. In particular, Florentin Guth, then a PhD student with Stéphane Mallat, introduced me to their numerical experiments on *rainbow networks* — random networks with weights matching the statistics of trained ones — which surprisingly retained a large fraction of training performance [Gut+24]. Motivated by this observation, we extended the analysis of random features to the deep, correlated case in collaboration with Hugo Cui, Daniil Dmitriev, and Dominik Schröder, and demonstrated that in the proportional asymptotics the generalisation error of Gaussian rainbow networks is inherently limited [Sch+23; Sch+24b].

While the co-supervision of PhD students I started following during my postdoc at EPFL ensured the continuity of my collaborations with Florent and Lenka, I began to supervise my first full PhD students in Paris: Leonardo Defilippis in 2023, Arie Wortsmann in 2024, Clément Loup-Forest and Luigi Fogliani (co-supervised with Marylou Gabrié) in 2025. With Leonardo, I relaunched my collaboration with Gabriele Sicuro on heavy-tailed high-dimensional distributions [Ado+24], and initiated a project with Theodor Misiakiewicz

on a non-asymptotic analysis of random features [DLM24]. This work rigorously established the exact scaling laws of non-linear RF models, and showed that a previous result of Rudi and Rosasco on the minimal number of features required to achieve kernel minimax rates was not tight [RR17].

In the years that followed, I developed several other collaborations worth mentioning, though their results will not be discussed in detail in this manuscript. During a huddle organised by Jean Barbier, Manuel Saenz, Pragya Sur, and Subhabrata Sen in Trieste in the summer of 2023, Zhou Fan proposed using DMFT to analyse an adaptive Langevin algorithm employed in empirical Bayes methods. This led to fuitful collaboration with Zhou, Yandi Shen, Justin Ko, and Yue Lu [Fan+25a; Fan+25b], where I learned a lot about the challenges of mathematically formalising some ideas which are natural to physicists (and about Sichuanese food, too). Also in 2023, at a workshop in Tübingen, Marylou Gabrié (then at École Polytechnique) introduced me to the problem of model collapse in variational inference. Together we proposed an M2 project on the topic in 2024 and developed an analysis in a Gaussian mixture setting, showing that mode collapse corresponds to local minima of the variational inference objective that can trap the dynamics depending on the initialisation [SGL25]. With Luigi Fogliani, we are now investigating mitigation strategies for mode collapse. Finally, during their one-year visit to ENS, I worked with Julia Kempe and Nikos Tsilivis to compare their uniform convergence bounds for robust regression with exact asymptotic results I had derived with Matteo Vilucchio [Tan+25], showing that regularisation with respect to the dual norm of the attack is not always optimal in data-scarce regimes [Vil+24].

Rather than attempting to summarise all of my scientific work since 2018, I have chosen to focus on a single line of research that I believe is representative of my contributions over the past seven years, while also providing a coherent and compelling narrative. I have sought to present the discussion in an intuitive manner — retaining the mathematical detail necessary to understand the results, but without technicalities that might obscure the main message. Accordingly, most theorems are stated informally, with references to the original works for readers wishing a deeper treatment. I hope readers will find the account as engaging to read as it was to write.

Paris, 16th of September 2025.

1 | Introduction and overview

This manuscript discusses the research I have conducted since the end of my PhD. The common denominator of my different research lines are the challenges posed by the study of probability in high-dimensional spaces, particularly present in problems from the fields of computer science, statistics, signal processing and machine learning. Due to a choice for coherence and conciseness, the next chapters of this manuscript will focus mostly on one of my research lines, concerning the theory of generalisation and adaptativity in two-layer neural networks.

Although this line spans a considerable part of my research activity since the end of my PhD, it unfortunately leaves out both old and recent results which would deserve a manuscript of its own, and that showcase the diversity of my research interests. For instance, my research lines on uncertainty quantification [Cla+23c; Cla+23b; Cla+23a; Cla+24] and robustness robustness [Tan+25; Vil+24; VZL25] were left out, as well as earlier works on statistical-to-computational gaps in inference problems [Aub+19; Aub+20; Mai+20; Pes+22] and scaling limits of SGD for non-convex optimisation [Vei+22; Arn+23; Arn+24c; Arn+24a; Arn+25]. I refer the reader to the list of publications above for a complete account of my research since the end of my PhD.

The rest of this introduction sets the scientific context for what follows.

1.1 Why Statistical Physics?

A natural question for an unfamiliar reader encountering the title of this manuscript for the first time is: *what does physics have to do with neural networks* — or, more broadly — *with machine learning?* Given the centrality of this connection to what follows, it is appropriate to begin by addressing this question directly.

Statistical physics is the branch of physics concerned with understanding how the macroscopic properties of materials — such as temperature, pressure, or magnetisation — emerge from the microscopic interactions of their underlying constituents, such as particles, atoms, or molecules, often referred to as *degrees of freedom*. Motivated by the central role of thermodynamics in the first industrial revolution, statistical physics was born from the endeavour to explain how the principles of thermodynamics emerge from the fundamental laws of physics

at the microscopic scale. At its core lies the idea of regarding the physical state of a collection of particles — referred to as a *configuration* — as a sample drawn from a probability distribution defined over the set of all possible configurations, the *configuration space*.[4] From this perspective, a macroscopic property of the system is simply a *sample statistic*. Starting from the seminal works of Ludwig E. Boltzmann (1844-1906) and Josiah W. Gibbs (1839-1903) [Bol77; Gib02], physicists have developed mathematical tools to study the central question of statistical physics: *how to characterise the statistical properties of this distribution when the number of degrees of freedom is large?*.[5]

Mathematically, this problem is far from trivial. Early probability theory, as developed by Bernoulli and Laplace, focused primarily on the limiting behaviour of independent and identically distributed random variables, leading to classical results such as the law of large numbers and the central limit theorem. Statistical physics, by contrast, placed at its core the study of probability measures on systems of *interacting particles*, which mathematically corresponds to analysing limits of sequences of probability spaces rather than simple product measures. In this sense, the works of physicists such as J.W. Gibbs, L.E. Boltzmann, and A. Einstein laid the foundations for many concepts later formalised within modern probability theory, including stochastic processes [Lév40], concentration of measure [Led01], large deviations [Var66], and random matrix theory [Wig93].

Given the central role played by probability in the framing of machine learning theory, the connection between statistical physics and learning theory should thus come at no surprise to the reader. Indeed, the connection between these two fields is far from recent, and while an extensive historical account would take us way beyond the scope of this manuscript, I believe it is worth revisiting a few of the landmark works that laid ground for this connection.

Perhaps the first and most influential work to draw a precise analogy between statistical physics and optimisation was the paper by S. Kirkpatrick, C.D. Gelatti, and M.P. Vecchi introducing the *Simulated Annealing* algorithm [KGV83]. Building on the analogy between local algorithms and the relaxation dynamics of physical systems, the authors proposed an algorithm inspired by the physical process of *annealing* — the controlled cooling of a liquid to produce crystals with desired properties:

> *"There is a deep and useful connection between statistical mechanics (the behavior of systems with many degrees of freedom in thermal equilibrium at a finite temperature) and multivariate or combinatorial optimization (finding the minimum of a given function depending on many parameters). A detailed analogy with annealing in solids provides a framework for optimization of the properties of very large and complex systems. This connection to statistical mechanics exposes new information and provides an unfamiliar perspective on traditional optimization problems and methods."*

[4] In probability language, the configuration space is the sample space, and a configuration is an event.

[5] The number of particles in a typical physical system, such as the number of H_2O molecules in a gram of water, is of the order of the *Avogadro number* 10^{23}, an overwhelmingly large number.

Beyond the far-reaching impact of simulated annealing as a versatile algorithm for combinatorial optimisation, an equally important conceptual contribution of this work was to establish a link between computational complexity and the rugged energy landscapes characteristic of systems in which many variables must simultaneously satisfy competing constraints, known as *frustration* in physics [KT85].

One year earlier, John Hopfield, a physicist working in neuroscience, had drawn on a similar analogy to propose a model of associative memory in the brain [Hop82]. His central idea was to interpret stable memories as local minima of a complex energy landscape generated by the collective dynamics of neurons, which could be retrieved through a simple and biologically plausible learning rule proposed by neuroscientists, the *Hebb rule* [Heb49]. Hopfield's work was the first to establish a systematic connection between statistical physics and biological neural networks, a contribution later recognised by the award of the Nobel Prize in Physics in 2024. The *Hopfield model*, as it became known, soon attracted the attention of the statistical physics community. In 1985, Daniel Amit, Hanoch Gutfreund, and Haim Sompolinsky adapted tools developed only a few years earlier in the study of spin glasses to demonstrate the existence of a transition between a phase in which stored patterns in a Hopfield network can be successfully retrieved and a phase in which the network fails to recall them [AGS85].

The years following Hopfield's work witnessed a steady growth of interest among theoretical physicists in biological neural networks. It was soon recognised that the same set of ideas and tools could also be applied to artificial neural networks. This was pioneered by Elisabeth Gardner and Bernard Derrida, who analysed how many random points a single-layer neural network[6] could correctly classify within a given margin[7], generalising an earlier result by information theorist Thomas M. Cover, obtained using random combinatorics [Cov65]. Early contributions from Gardner, Derrida, Amit, Gutfreund, and Sompolinsky, among others, demonstrated that the connection between statistical physics and neural networks was not merely conceptual: the tools developed within statistical physics could be adapted to these problems, yielding a quantitative understanding of questions relevant to learning theory.

The period following Gardner and Derrida's work was marked by intense activity in what came to be known as the *statistical physics of learning* community. It is interesting to note that this period of increasing interest of physicists for neural networks is coincide with what is known in the deep learning folklore as the "*neural network winter*", a period during which research activity on neural networks within computer science and engineering departments was at a low point [Hin19; LeC19a; Ben22]. In his memoir, Yann Le Cun — a pioneer in the development of convolutional neural networks — recalls attending a conference on

[6] Introduced by Frank Rosenblatt in 1958, this was known at the time as the *perceptron*.
[7] Also known the the *storage capacity problem*.

neural networks at the *École de Physique des Houches* in 1985 [LeC19b]:

> *"Ma vie professionnelle bascule réellement en février 1985 lors d'un symposium aux Houches, dans les Alpes. Je rencontre là-bas la fine fleur de la recherche internationale qui s'intéresse aux réseaux de neurones : physiciens, ingénieurs, mathématiciens, neurobiologistes, psychologues, et notamment des membres d'un tout nouveau groupe de recherche en réseaux de neurones qui s'est formé aux Bell Labs, un lieu mythique pour la communauté scientifique. Grâce aux liens que je noue aux Houches, je finirai par être embauché dans ce groupe trois ans plus tard."*.

This testimony illustrates the rich exchange of ideas between the diverse communities interested in neural networks at the time, and highlights the subject's inherently multidisciplinary history. Such exchanges bore concrete fruit on both sides, as recalled by Isabelle Guyon — a pioneer in the development of the *support vector machine* (SVM) algorithm — who also attended the 1985 Les Houches school [Guy16]:

> *"I benchmarked neural networks against kernel methods with my Ph.D advisors Gerard Dreyfus and Leon Personnaz. The same year, two physicists working close-by (Marc Mézard & Werner Krauth) published a paper on an optimal margin algorithm called 'minover,' which attracted my attention... but it was not until I joined Bell Labs that I put things together and we created support vector machines."*.

The late 1980s also saw the launch of what would become the leading venue for machine learning research: the *Conference on Neural Information Processing Systems*. The first proceedings, published by the American Institute of Physics [And87], provide a clear testimony to the multidisciplinary character of the field in its early days, with contributions ranging from neuroscience and statistical physics to computer science, engineering, and applied mathematics. A direct thread connects the early developments described here and the research currently carried out at the crossroads of these fields, but a comprehensive account of the statistical physics of learning from the 1990s to the present lies well beyond the scope of this manuscript.

Taken together, the ideas discussed in this section show that the dialogue between statistical physics and machine learning is neither superficial nor recent. At their core, both fields confront the challenge of understanding the statistical properties of complex, high-dimensional systems — problems that naturally call for probabilistic methods. It is therefore no surprise that the ideas developed in the context of spin glasses found fertile ground in computer science, serving both as a conceptual framework for describing complexity and as a source of technical tools to analyse it. The fruits of this connection are far from abstract: they have inspired concrete algorithmic innovations such as *simulated annealing* and *support vector machines*. Historically, neural network theory itself emerged at the confluence of different communities, with physicists, computer scientists, and neuroscientists shaping a

shared body of ideas that deeply marked its early trajectory, and that continue to contribute to its development.

As the following sections will show, the work presented in this manuscript forms part of this ongoing shared endeavour.

1.2 The case for a *typical-case* analysis

This manuscript is primarily concerned with the question of generalisation in neural networks. Generalisation refers to how such models *learn* patterns from data and leverage them to *predict* the behaviour of data they have not previously encountered. Also known as *out-of-sample prediction*, it is a problem fundamental not only to machine learning but to statistics more broadly. Before turning to neural networks specifically, it is worth providing a precise mathematical formulation of this question and explaining *why* it is inherently challenging. Since the greater part of this manuscript is devoted to supervised learning, we frame the discussion within this setting for concreteness.

Consider a supervised learning task with training data $\mathcal{D} = \{(x_i, y_i) \in \mathbb{R}^d \times \mathcal{Y} : i \in [n]\}$, drawn i.i.d. from a joint distribution $p(x, y)$ over $\mathbb{R}^d \times \mathcal{Y}$. Throughout this manuscript, we will mainly focus on regression ($\mathcal{Y} = \mathbb{R}$) and binary classification ($\mathcal{Y} = \{-1, +1\}$) tasks. Given a loss function $\ell : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}_+$, we define the *population risk* as

$$R(f) = \mathbb{E}\left[\ell(y, f(x))\right], \tag{1.1}$$

The question of *generalisation* consists of finding an *estimator*[8] $\hat{f} : \mathbb{R}^d \to \mathcal{Y}$ such that $R(\hat{f})$ is as small as possible.[9] This can be made more precise by defining the *Bayes predictor*

$$f_\star(x) \in \arg\min_{\hat{y} \in \mathcal{Y}} \mathbb{E}[\ell(y, \hat{y}) | X = x] \tag{1.2}$$

and the associated *Bayes risk* $R_\star = \inf_{\hat{f}} R(\hat{f}) = R(f_\star)$, which is the minimally achievable risk.[10] For instance, for the square loss $\ell(y, \hat{y}) = (y - \hat{y})^2$ the Bayes predictor is the conditional expectation $f_\star(x) = \mathbb{E}[y|x]$. Therefore, the problem of generalisation can be reframed as achieving a risk close to the Bayes risk. If the statistician had access to the data distribution p, this would be a simple problem: just solve eq. (1.2). The challenge arises because p is only accessible through the training data $\mathcal{D}$. Therefore, the question of generalisation is really about how large n should be in order to achieve the Bayes risk.

Without further structure, this problem is ill-posed, as tasks of arbitrary complexity can

[8] That is, a measurable function of the training data.

[9] Note that $R(\hat{f})$ is itself a random variable, since it depends on the training data. This is therefore an inherently probabilistic statement, often phrased either *in expectation* or *with high probability* over $\mathcal{D}$.

[10] Note that even though $f_\star$ might not be unique, $R_\star$ is.

be constructed. To obtain a meaningful mathematical formulation, it is necessary to impose restrictions on the two underlying objects that define the problem:

(a) The data distribution p, or equivalently, the Bayes predictor $f_\star$ and the marginals p_x.

(b) The predictor $\hat{f}$, also known as the *hypothesis*.

It is useful to discuss in some detail the role played by these two components in learning theory.

1.2.1 The data distribution

A classical first approach in learning theory is to consider regularity conditions over the Bayes predictor. For instance, under uniform bounded covariates,[11] one can show the following lower-bound:

Theorem 1.2.1 ([Tsy08], informal)**.** Assume $x_i \sim \mathrm{Unif}([0,1]^d)$ and that $y_i = f_\star(x_i) + \varepsilon_i$ with $\mathbb{E}[\varepsilon_i] = 0$, $\mathbb{E}[\varepsilon_i^2] < \infty$ independently of x_i. Then:

$$\inf_{\hat{f}} \sup_{f_\star \in \mathrm{Lip}(1)} \mathbb{E}\left[\left(\hat{f}(x) - f_\star(x)\right)^2\right] \gtrsim n^{-\frac{2}{2+d}}, \tag{1.3}$$

where the infimum is taken over all measurable functions of the training data and the expectation is over both the training data $\mathcal{D}$ and the covariate $x \sim \mathrm{Unif}([0,1]^d)$.

Theorem 1.2.1 states that the *best predictor* requires at least $n(\epsilon) \gtrsim \epsilon^{-\frac{2+d}{2}}$ to approximate the *hardest* regular Bayes predictor $f_\star$ to precision δ in squared error. Known as the *curse of dimensionality*,[12] this exponential dependency in the data dimension implies a computational bottleneck for moderate d, as even storing the data becomes prohibitive.

While fundamental, the minimax theorem 1.2.1 is at odds with the daily practice of deep learning, where massive neural networks are trained to achieve low risk on massive data sets. This discrepancy highlights two shortcomings of this result: First, regularity might not be enough, as real data often contains stronger structural and geometrical properties. Second, this is a *worst-case* statement about learning the *hardest* possible function in the class. For most standard tasks, such as image classification, real data is not adversarial.[13]

Taken together, these considerations suggest the need for more structured, probabilistic models of data in the study of machine learning theory.

[11] The same lower-bound hold for fixed bounded covariates.

[12] This terminology was introduced by Richard Bellman in [BCC57] to describe a pervasive difficulty underlying many optimisation problems in computer science.

[13] Problems where an adversarial attacker deliberately poisons the data is a subject of its own in learning theory, but is beyond the scope of this manuscript.

1.2.2 The hypothesis class

Complementary to the discussion above is the fact that we are typically not interested in the *best predictor*. Not only would this require full knowledge of the data distribution p, but taking the infimum over all measurable functions is also computationally intractable. Instead, most learning frameworks restrict attention to a smaller class of predictors. The most common approach in machine learning practice is to consider the *empirical risk minimiser* (ERM):

$$\hat{f} \in \arg\min_{f \in \mathcal{F}} \left[\hat{R}_n(f) := \frac{1}{n} \sum_{i=1}^{n} \ell\left(y_i, f(x_i)\right) \right] \tag{1.4}$$

where the optimisation is carried out over a subclass of measurable functions $\mathcal{F}$, often parametric, $\mathcal{F} = \{f_\theta : \mathbb{R}^d \to \mathcal{Y} : \theta \in \Theta\}$, referred to as the *hypothesis class*. This procedure introduces two further challenges. First, the choice of hypothesis class typically translates a preference for certain types of functions, known as an *inductive bias*. Second, even if the risk is convex in $f \in \mathcal{F}$ (for convex losses), it need not be convex in the parameters $\theta \in \Theta$. As a result, different optimisation procedures for solving the ERM problem in eq. (1.4) may converge to different solutions with potentially different risks, a phenomenon referred to as *implicit (algorithmic) bias* [Sou+18].

To make this more concrete, consider a parametric hypothesis class together with a learning algorithm $\hat{\theta} = \mathcal{A}(\mathcal{D}) \in \Theta$. The excess risk can then be decomposed as

$$R(\hat{\theta}) - R_\star = \left[R(\hat{\theta}) - \hat{R}_n(\hat{\theta})\right] - \left[\hat{R}_n(\theta_\star) - \hat{R}_n(\hat{\theta})\right] - \left[R(\theta_\star) - \hat{R}_n(\theta_\star)\right] \tag{1.5}$$

where $\theta_\star \in \inf_{\theta \in \Theta} R(\theta)$, and we write $R(\theta) = R(f_\theta)$ by abuse of notation. The last term in this decomposition is simply the difference between the empirical mean and the expectation of i.i.d. random variables, which vanishes at rate $O(n^{-1/2})$ by the law of large numbers. The second term is an optimisation error: it quantifies how well the algorithm $\hat{\theta} = \mathcal{A}(\mathcal{D})$ succeeds in minimising the empirical risk $\hat{R}_n$. The first term is the generalisation gap. While superficially similar to the last term, there is a crucial difference: both the estimator $\hat{\theta}$ and the empirical risk $\hat{R}_n$ depend on the same training data $\mathcal{D}$, making $\hat{R}_n(\hat{\theta})$ a biased estimator of $R(\hat{\theta})$. Controlling its concentration therefore requires a more refined characterisation of $\hat{\theta}$. Instead, the standard approach in learning theory is to control the generalisation gap uniformly over the hypothesis class, in terms of its complexity. A classical result is the following:

Theorem 1.2.2 ([BM02], informal)**.** Assume $\ell(y, \cdot)$ is bounded.[14] Then, with probability

[14] This can be relaxed under assumptions on the data distribution.

$1-\delta$ over the training data,

$$\forall f \in \mathcal{F}, \qquad R(f) - \hat{R}_n(f) \leq 2\hat{\mathfrak{R}}_n(\mathcal{F}) + \sqrt{\frac{8\log 2/\delta}{n}} \tag{1.6}$$

where $\mathfrak{R}(\mathcal{F})$ is the (empirical) *Radamacher complexity* of the hypothesis class $\mathcal{F}$:

$$\hat{\mathfrak{R}}_n(\mathcal{F}) = \mathbb{E}\left[\sup_{f\in\mathcal{F}} \frac{2}{n}\sum_{i=1}^{n} \sigma_i \ell\left(y_i, f(x_i)\right) \middle| \mathcal{D}\right] \tag{1.7}$$

where $\sigma_i \sim \mathrm{Unif}(\{-1,+1\})$ i.i.d.

As the name suggests, $\hat{\mathfrak{R}}_n$ measures the complexity of $\mathcal{F}$. For example, if $\mathcal{F}$ is finite then $\hat{\mathfrak{R}}_n \leq \sqrt{\log|\mathcal{F}|/n}$, while if $\mathcal{F} = \{f(x;\theta) = \langle\theta, x\rangle : ||\theta||_2 \leq R\}$ with $||x||_2 \leq 1$, then $\hat{\mathfrak{R}}_n \leq R/\sqrt{n}$. Although mathematically sound, this bound again falls short in answering questions arising in the deep learning practice, such as when and how overparametrised networks achieve small excess risk. For instance, in parametric classes the Rademacher complexity typically scales with the size of the parameter space, yielding vacuous bounds in the overparametrised regime [Zha+16]. As will be discussed in detail in Chapter 2, even simple linear models for which $\hat{\mathfrak{R}}_n = O(\sqrt{d/n})$ can have low excess risk in the $d < n$ regime while achieving $\hat{R}_n = 0$, a phenomenon known as *benign overfifting* [Bar+20]. As with the minimax lower bound in theorem 1.2.1, this limitation stems from the broad nature of the assumptions: the result is required to hold uniformly over arbitrary function classes, independently of the data distribution. This is, once again, a *worst-case* guarantee.

A similar difficulty, which we mention here only *en passant*, arises for the optimisation term in eq. (1.5). In 1988, Avrim Blum and Ronald Rivest [BR88] showed that even for a very simple architecture — a two-layer neural network with only three linear threshold nodes — deciding whether weights exist that fit training data drawn from a carefully chosen distribution is NP-complete. In other words, exactly minimising the empirical risk is computationally intractable in the worst case. This highlights a parallel with generalisation: worst-case guarantees, while mathematically rigorous, often yield results that are either too broad to be informative (in generalisation) or too demanding to be computationally feasible (in optimisation).

1.2.3 A typical-case point of view

The discussion above illustrates that broad, data-agnostic worst-case approaches to learning theory fall short of addressing the question of generalisation in the context of modern practice. Over the past decade, different mathematical approaches have emerged to tackle this challenge at a finer level. This area, loosely referred to as *deep learning theory*, is diverse, encompassing methods ranging from PAC-Bayes bounds to exact asymptotics. Despite this

heterogeneity, the field is unified by a shared recognition: genuine progress in understanding both the successes and the limitations of neural networks requires incorporating detailed information about the data distribution, the network architecture, and the descent-based optimisation algorithms used for training.

The results discussed in this manuscript will adopt a perspective known as *typical-case analysis*. Unlike worst-case analysis, which seeks guarantees valid for the most adverse instance of a problem, the typical-case approach aims to characterise the behaviour of *typical instances*. The notion of a typical instance is distribution specific, and therefore requires and therefore requires positing an explicit generative model for the data. Within this framework, the objective is to derive sharp predictions for quantities of interest — such as the risk or the convergence rate of a training algorithm — under the chosen distribution.

But what is a reasonable model for "typical data"? There is no consensual answer to this question. As motivated in Section 1.1, we approach it here as a *complex systems* problem, adopting a *bottom-up perspective* natural to statistical physics and to science more broadly. Our strategy is to model data from simple, mathematically tractable building blocks that encode the inductive biases we expect natural data to exhibit. Although necessarily simplistic, such models are intended to capture aspects of the phenomenology observed — at a certain level of granularity — when training neural networks on real data. Although inevitably simplified — much like thermodynamics is a coarse description of the inner workings of a refrigerator — these models provide a foundation from which progressively finer and more realistic descriptions can be developed.

1.3 Preliminaries

As the title suggests, our focus in this manuscript will be in studying the class of fully-connected two-layer neural networks of width p:

$$\mathcal{F}_p = \left\{ f(x; W, a) = \sum_{j=1}^{p} a_j \sigma\left(\langle w_j, x \rangle\right) : a_j \in \mathbb{R}, w_j \in \mathbb{R}^d \right\} \tag{1.8}$$

where a_j, w_j are the trainable parameters, known as *first-* and *second-layer weights*, respectively and $\sigma : \mathbb{R} \to \mathbb{R}$ denotes is a real-valued function, known as the *activation function*. Popular examples are the *rectified linear unit* (ReLU) $\sigma(t) = \max(0, t)$ and the *sigmoid* $\sigma(t) = (1+e^{-t})^{-1}$. In the context of binary classification, $f(x; \theta)$ parametrises the scores, with a *decoding function* $d : \mathbb{R} \to \{-1, +1\}$ being applied for prediction (e.g. $d(t) = \operatorname{sign}(t)$ for $\mathcal{Y} = \{-1, +1\}$). Given training data $\mathcal{D} = \{(x_i, y_i) \in \mathbb{R}^d \times \mathcal{Y} : i \in [n]\}$, the empirical risk minimisation in

eq. (1.4) therefore reads:

$$\min_{(W,a)\in\mathbb{R}^{p(d+1)}} \frac{1}{n}\sum_{i=1}^{n} \ell\left(y_i, f(x_i; W, a)\right) + r(W, a) \tag{1.9}$$

where r is a regulariser, often added to impose a constraint on the weights. A popular example is *weight decay*, where one penalises the ℓ_2 norm of the weights: $r(a, W) = \lambda_a ||a||_2^2 + \lambda_W ||W||_{\mathrm{F}}$ with $\lambda_a, \lambda_W \geq 0$. The loss function is typically a convex function of the second argument, and with the most common examples being the *squared loss* $\ell(y, t) = (y - t)^2$ for regression ($\mathcal{Y} = \mathbb{R}$) and the *logistic loss* $\ell(y, t) = \log(1 + e^{-yt})$ for binary classification ($\mathcal{Y} = \{-1, +1\}$).

1.3.1 Multi-index functions

In the spirit of the discussion in Section 1.2.3, one may ask: what constitutes a meaningful class of tasks for studying the typical properties of the learning problem in eq. (1.9)? A natural choice for typical-case analysis is a family of functions where the difficulty of learning with a given hypothesis class can be tuned. In other words, the task should be neither too simple, which would be trivially learned, nor too complex, which would bring the analysis back to a worst-case type of scenario.

A systematic way to achieve this is through the *teacher–student framework*,[15] in which the functional relationship between covariates and labels is generated, up to noise, by a model drawn from the same hypothesis class under analysis. Formally, this corresponds to a decomposition of the joint distribution $p(x, y) = p(y|x)p(x)$,[16] where the likelihood takes the form $p(y|x) = p(y|f_\star(x))$ for some $f_\star \in \mathcal{F}$, referred to as the *teacher* or *target* function. In statistical terminology, this ensures that the task is *well-specified*, i.e. that the hypothesis class is expressive enough to learn the target. More recently, the terminology has also been applied in broader, misspecified settings to denote any generative model of the form $p(x, y) = p(y|f_\star(x))p(x)$ with $f_\star \in \mathcal{F}_\star$ drawn from a different class of functions $\mathcal{F}_\star \neq \mathcal{F}$. A popular example in the context of two-layer neural networks is to take an additive noise model $y_i = f_\star(x_i) + \varepsilon_i$ with the teacher $f_\star$ itself given by two-layer neural network of smaller width $r < p$ i.e., $f_\star \in \mathcal{F}_r$. This allows, for instance, to study how overparametrisation influences learning.[17]

As we will argue, the class of multi-index functions offers a natural and effective testbed for a typical-case analysis of two-layer neural networks.

[15] Although the idea is natural and long-standing in statistics, to my knowledge the "teacher–student" terminology was first introduced by Elisabeth Gardner in [GD89].

[16] The complementary decomposition $p(x, y) = p(x|y)p(y)$ is also a useful source of models, particularly in classification tasks, where it corresponds to mixture distributions.

[17] In this case, both teacher and student may be viewed as belonging to the broader class of two-layer networks of arbitrary width, $\mathcal{F}_{2\mathrm{lnn}} = \cup_{p\geq 1}\mathcal{F}_p$.

Definition 1.3.1 (Multi-index function)**.** Let $W \in \mathbb{R}^{r\times d}$ denote a matrix with $\operatorname{rank} W = r \leq d$. A multi-index function $f : \mathbb{R}^d \to \mathbb{R}$ is defined as:

$$f(x) = g\left(Wx\right) \tag{1.10}$$

where $g : \mathbb{R}^r \to \mathbb{R}$ is a non-linear function known as the *link function*. The rows $w_k \in \mathbb{R}^d$ with $k \in [r]$ are known as the *indices*. In particular, when $r = 1$ we say f is a *single-index function*.

Remark 1.3.1 (Random link function)**.** A common variation consists of allowing g to be a (possibly) stochastic function. In this case, we denote $y \sim \mathsf{P}_y(y|Wx)$, where P_y denote the model likelihood. Note that the case of a deterministic g is a particular case given by $\mathsf{P}_y(y|Wx) = \delta\left(y - g(Wx)\right)$.

Multi-index models are classical in statistics, tracing back to the work of George Box and David Cox on how data of data can be transformed to yield a linear dependence (corresponding to g^{-1}) [BC64; BD81]. They also encompass and extend a range of semi-parametric methods, such as generalised linear models [NW72] and basis pursuit regression (corresponding to a factorised form $g(z) = \sum_{k\in[r]} g_k(z_k)$) [FS81].

Conceptually, the class of multi-index functions encode the inductive bias that the relevant directions for prediction depend only on a low-dimensional subspace of the covariates $x \in \mathbb{R}^d$. This makes it an appealing generative model for high-dimensional data. Indeed, this inductive bias is often raised as a common explanation for why models can generalise when trained with high-dimensional data despite the curse of dimensionality, also known as the *manifold hypothesis* [TSL00].

Note that a width-p two-layer neural network implements a particular index-p function with link function $g(z) = \sum_{k\in[p]} a_k\sigma(z_k)$. Therefore, it also encompasses the teacher-student setting as a particular example. On the other hand, it is a richer class, containing other popular functions studied in the literature.

Example 1.3.1. Many classical target functions studied in learning theory and signal processing can be written as multi-index functions:

- Linear functions ($r = 1$): $g(z) = z$ [Hoe59; CRT06; CD07].
- Phase retrieval ($r = 1$): $g(z) = |z|$ or $g(z) = z^2$ [Bar19; CSV13; TV23].
- Perceptron / 1-bit compressive sensing ($r = 1$): $g(z) = \operatorname{sign}(z)$ [GD88; BB08].
- Polynomials ($r > 1$): $g(z) = z_1 \dots z_r$ [CM20].
- Intersection of half-spaces ($r > 1$): $g(z) = \prod_{k\in[r]} \mathbb{I}(z_k - a_k > 0)$ [KOS04; Vem10].

- r-sparse parities ($r > 1$): $g(z) = \text{sign}(z_1 \dots z_r)$ [Kea98; BKW03; KRT17].

where we denoted $z_k = \langle w_k, x \rangle$, $k \in [r]$.

We are now ready to define the class of supervised learning tasks that will be studied in the following.

Definition 1.3.2 (Gaussian multi-index model (GMIM))**.** We say the training data $\mathcal{D} = \{(x_i, y_i) \in \mathbb{R}^d \times \mathcal{Y} : i \in [n]\}$ has been drawn from a *Gaussian r-index model* if:

$$y_i = g(W_\star x_i), \qquad x_i \sim \mathcal{N}(0, {}^1\!/_d I_d), \quad \text{i.i.d.} \tag{1.11}$$

for $g \in L^2(\gamma_r)$ and the columns of $W_\star \in \mathbb{R}^{r \times d}$ form an orthonormal family $W_\star W_\star^\top = d I_r$.

Remark 1.3.2. A few important remarks about Definition 1.3.2.

- **Orthogonality:** The orthogonality assumption on the columns of $W_\star$ is without loss of generality, as we can always go to a basis in which $W_\star$ has orthogonal rows by redefining g.
- **Scaling:** The choice of scaling $||w_k||_2^2 = d$ is merely conventional at this stage. However, it becomes meaningful in the high-dimensional limit $d \to \infty$ with fixed $r = \Theta_d(1)$ which will be of interest in the following. Indeed, together with the choice of scaling for $||x||_2^2 = \Theta_d(1)$ and $g \in L^2(\gamma_r)$, this ensures that the labels have $\Theta_d(1)$ variance, and hence a finite signal-to-noise ratio in the limit.
- **Gaussian weights:** A variation of definition 1.3.2 consists of taking $W_\star$ to be a matrix with i.i.d. Gaussian entries $\mathcal{N}(0, 1)$. Although this is not generally equivalent, almost orthonormality in the high-dimensional limit also suffice for asymptotic results stated in this manuscript to hold.

As previously motivated, definition 1.3.2 is intended to model low-dimensional structure in the data distribution. Under the isotropic Gaussian assumption on the covariates, however, this structure is encoded entirely in the functional dependence of the labels on the covariates. This is clearly a simplification of real data, where the covariates themselves typically carry meaningful information about the task. Extending the present results to such settings is an interesting research direction. Nevertheless, the isotropic case may be viewed as providing an upper bound, since additional structure in the covariates is expected to facilitate learning.

Finally, it will be useful to distinguish two notions of learnability for the Gaussian multi-index model.

Definition 1.3.3 (Learnability)**.** Let $\mathcal{D} = \{(x_i, y_i) \in \mathbb{R}^{d+1} : i \in [n]\}$ denote n samples from a GMIM. Denote by $\hat{W}(\mathcal{D})$ an estimator of $W_\star$,[18] which we assume has norm $||\hat{W}||_{\text{F}} = \Theta(d)$.

[18] In other words, any measurable function of the training data.

- We say that $\hat{W}$ has *weakly learned* or *weakly recovered* a subspace $V_\star \subset \mathrm{span}(W_\star)$ if:

$$\inf_{\substack{v \in V_\star \\ ||v||_2 = 1}} \left\| \frac{\hat{W} W_\star^\top v}{d} \right\|_2 = \Theta_d(1), \quad \text{w.h.p. as } d \to \infty \tag{1.12}$$

- Similarly, we say that $\hat{W}$ has *fully learned* or *fully recovered* a subspace $V_\star \subset \mathrm{span}(W_\star)$ if:

$$\inf_{\substack{v \in V_\star \\ \|v\|_2 = 1}} \left\| \frac{\hat{W} W_\star^\top v}{d} \right\|_2 = 1, \quad \text{w.h.p. as } d \to \infty \tag{1.13}$$

Note that full-recovery is equivalent to $\hat{W} = W_\star$ up to rotational symmetry.[19]

1.3.2 The questions

We can now revisit the question of generalisation motivated in section 1.2, now in the context of two-layer neural networks. A widely held view in the folklore about their "unreasonable effectiveness" is that neural networks succeed because they can *adapt to the data* by learning the relevant *features* during training. Making this connection between *adaptativity* and *generalisation* precise will be the main thread connecting the works in this manuscript.

More concretely, consider the empirical risk minimisation problem in eq. (1.9), where a statistician seeks to learn a two-layer neural network from a batch of training data $\mathcal{D} = (x_i, y_i) \in \mathbb{R}^d \times \mathcal{Y} : i \in [n]$ drawn from a Gaussian multi-index model (Definition 1.3.2). They key questions we would like to understand are:

(a) **Feature learning:** How does the network adapt to the low-dimensional structure in the data, and how is this related to its performance?

(b) **Approximation:** How many neurons are required to approximate the target function?

(c) **Estimation:** How much data is required to achieve low risk, and how does this compare with the optimal method for this problem?

(d) **Optimisation:** How efficient are popular training algorithms, such as SGD?

The results discussed in this manuscript will touch on each of these questions. Before turning to them, however, it is useful to discuss some expectations.

A classical approximation result is that the class of unbounded width two-layer networks $\mathcal{F}_{2\mathrm{lnn}} = \cup_{p \geq 1} \mathcal{F}_p$ with non-polynomial activation σ are *universal approximators*, [Cyb89; HSW89; Les+93]. In other words, for any smooth function $f_\star : \mathbb{R}^d \to \mathbb{R}$ and desired

[19] The *identifiability* question of exactly recovering $W_\star$ depends on the link function g. For instance, if $g(z) = \sum_{k \in [r]} g_k(z_k)$ is separable, $W_\star$ can be identified up to permutations [Yua11].

precision $\varepsilon > 0$, there exists a neural network with non-polynomial activation σ and width $p \geq 0$ such that $\sup_{x \in K} |f_\star(x) - f(x; W, a)| < \varepsilon$ on any compact subset $K \subset \mathbb{R}^d$. Those who read section 1.2 attentively will not be surprised that such a general result does not teach much about how large p needs to be. In particular, one can construct smooth targets for which $p = O(e^d)$ [ES16]. Quantitative results of the type $||f - f_\star||_{L^2(\gamma_d)} \leq O(d\gamma(g)^2/p)$ can be derived for the GMIM (definition 1.3.2), where $\gamma(g) = \Theta_d(1)$ is a complexity measure depending on the non-linearity σ — for instance the Barron norm for sigmoids [Bar93] or the TV norm for the ReLU [Bac17a].[20]

Regarding estimation, learning a multi-index target $f_\star(x) = g(Wx)$ can be seen as the composition of two problems: (i) learning a non-parametric function $g : x \in \mathbb{R}^r \mapsto g(z) \in \mathbb{R}$; (ii) inverting a linear problem $z = Wx$. As we have discussed in theorem 1.2.1, the sample complexity of (i) is $n(\epsilon) \sim \epsilon^{-\frac{2+r}{2}} = \Theta_d(1)$ in the worst-case, while inverting a linear system in (ii) has complexity $n = \Theta_d(d)$ for $d \to \infty$ at $r = \Theta_d(1)$. Therefore, we expect the bottlenack to be dominated by the latter.

Overview

This manuscript is organised as follows. Chapter 2 examines the asymptotic generalisation properties of two-layer neural networks with fixed first-layer weights, also known as the *random features approximation* to kernel methods. Although considerably simpler than the full learning problem in eq. (1.9), this setting already presents significant technical challenges due to the non-linearity of the features. A central tool in the analysis is the notion of *Gaussian equivalence*, introduced in section 2.5.1, which makes it possible to derive sharp asymptotic formulas for network performance in the high-dimensional limit. Several consequences of these formulas are then explored. Section 2.3 derives rates for the excess error under source and capacity conditions, also known as *scaling laws*. Section 2.4 highlights the limitations imposed by the lack of adaptivity in this regime, showing that the high-dimensional predictor effectively behaves as a linear function, thereby restricting expressivity. Finally, Section 2.7 extends the analysis to the multi-layer setting.

Chapter 3 turns to the impact of training the first-layer weights. It shows that even a few large updates enable the features to adapt to the low-dimensional structure in the data, thereby enhancing both expressivity and performance at fixed sample complexity. In particular, Section 3.2 characterises how the network weights correlate with the low-dimensional structure, while section 3.3 provides upper and lower bounds on the generalisation error after training. Section 3.4.2 establishes a conditional form of Gaussian equivalence for the feature matrix in the proportional asymptotics, enabling sharp asymptotic characterisations presented in section 3.4.3.

[20] See [Bac17a] for results on L^∞.

Chapter 4 investigates the fundamental computational limits of learning multi-index functions with first-order methods in the high-dimensional regime, thereby providing a benchmark for the results of the preceding sections. Section 4.2 introduces a classification of index subspaces as trivial, easy, or hard to learn. Section 4.3 explores a hierarchical learning phenomenon, showing that hard subspaces can be learned efficiently when coupled to easier ones. Section 4.4 examines spectral methods that achieve optimal sample complexity.

2 | Network at initialisation

In this chapter, we consider the empirical risk minimisation problem introduced in eq. (1.9) in the case where the first-layer weights are fixed at some $W_0 \in \mathbb{R}^{p\times d}$:

$$\min_{a\in\mathbb{R}^p} \sum_{i=1}^{n} \ell\left(y_i, f(x_i; W_0, a)\right) + \lambda r(a). \tag{2.1}$$

where for convenience we rescaled the risk by a factor n.[21] The results discussed in section 2.1 are based on [Cui+21; Def+25; Sch+23; Sch+24a], while the results discussed in section 2.5 are based on [Ger+20; Gol+22; Lou+21a].

Remark 2.0.1 (Motivation)**.** For several reasons, this problem has been widely studied in the machine learning literature.

- **Convexity:** Since $f(x; W_0, a) = \langle a, \sigma(W_0 x)$ is linear in the second-layer weights, this yields a considerably simpler optimisation problem. In particular, if both the loss and regulariser are convex functions, eq. (2.1) is a convex problem in $a \in \mathbb{R}^p$. Despite this simplification, it still retains some of the features of the generalisation curves of neural networks, as we will discuss next.

- **Random Features:** When $r(a) = \lambda||a||_2^2$ and the rows of W_0 are sampled i.i.d. as $w_{0,k} \sim \mu_w$, Equation (2.1) corresponds to a well-known problem in the learning literature: the *random features* approximation to kernel methods [BBV06; RR07]. Indeed, the representer theorem implies that in this case the predictor corresponding to the (unique) solution of eq. (2.1) can be written as:

$$f(x; W_0, \hat{a}_\lambda) = \sum_{i=1}^{n} \alpha_i \hat{K}_p(x_i, x), \tag{2.2}$$

  for some data-dependent coefficients $\alpha \in \mathbb{R}^n$, where $\hat{K}_p$ is the empirical random features

[21] This amounts to a redefinition of the regularisation term.

kernel, an approximation of a limiting kernel K:

$$\hat{K}_p(x,x') = \frac{1}{p}\sum_{k=1}^{p} \sigma\left(\langle w_{0,k}, x\rangle\right)\sigma\left(\langle w_{0,k}, x'\rangle\right) \xrightarrow[p\to\infty]{} \mathbb{E}\left[\sigma\left(\langle w_0, x\rangle\right)\sigma\left(\langle w_0, x'\rangle\right)\right] = K(x,x')$$

By carefully choosing the activation function σ, one can approximate a wide class of kernels [RR07]. Importantly, this reduces the cost of implementing kernel methods from $O(n^2)$ to $O(np)$, which can be computationally advantageous when $n \gg p$.

- **Lazy regime:** More recently, it has been shown that in the infinite-width limit $p \to \infty$, under "standard" initialisation $a_{0,j} = O(1/\sqrt{p})$,[22] the (S)GD dynamics for the second-layer weights evolve much faster than those of the first layer [COB19; Lee+19]. As a consequence, the problem in eq. (1.9) becomes equivalent to kernel regression with

$$\hat{K}(x,x') = \frac{1}{p}\sum_{k=1}^{p} \sigma\left(\langle w_{0,k}, x\rangle\right)\sigma\left(\langle w_{0,k}, x'\rangle\right) + \frac{\langle x, x'\rangle}{p}\sum_{k=1}^{p} \sigma'(\langle w_{0,k}, x\rangle)\sigma'(\langle w_{0,k}, x'\rangle) \quad (2.3)$$

where the first term is the random features kernel introduced above, and the second term is known as the *neural tangent kernel* [JGH18]. Therefore, the simpler problem in eq. (2.1) has recently gained in popularity as a proxy for studying lazy two-layer neural networks.

2.1 Random features ridge regression

As a starting point, we consider the *random features ridge regression* case, where $\ell(y, f(x)) = (y - f(x))^2$, $r(a) = ||a||_2^2$ and $w_{0,k} \sim_{i.i.d} \mu_w$. In this case, the problem in eq. (2.1) admits a closed form solution:

$$\begin{aligned}\hat{a}_\lambda(X,y) &= \arg\min_{a\in\mathbb{R}^p} \sum_{i=1}^{n} \left(y_i - \langle a, \sigma\left(W_0 x_i\right)\rangle\right)^2 + \lambda||a||_2^2 \\ &= (\Phi^\top\Phi + \lambda I_p)^{-1}\Phi^\top y \end{aligned} \quad (2.4)$$

where we define the feature matrix $\Phi_{ik} = \sigma(\langle w_{0,k}, x_i)$. Characterising the risk thus becomes a problem in random matrix theory, depending on both the feature matrix Φ and the labels y. We will work under the following assumptions:

Assumption 2.1.1 (Data)**.** Assume the training data $\mathcal{D} = \{(x_i, y_i) \in \mathbb{R}^{d+1} : i \in [n]\}$ is sampled as:

$$y_i = f_\star(x_i) + \varepsilon_i, \qquad x_i \sim \mu_x \quad \text{i.i.d.} \quad (2.5)$$

[22] This is known as the standard normalisation since it is the default normalisation on popular machine learning frameworks such as `PyTorch` and `TensorFlow`.

where μ_x is a distribution over $\mathbb{R}^d$, $f_\star \in L^2(\mu_x)$ and the noise $\varepsilon = y_i - f_\star(x)$ has zero mean $\mathbb{E}[\epsilon|x] = 0$ and finite variance $\mathbb{E}[\epsilon^2|x] = \sigma^2 < \infty$.

Remark 2.1.1. Assumption 2.1.1 are much more broader than the GMIM introduced in definition 1.3.2. Indeed, since the network features are random and independent from the low-dimensional structure of the multi-index model ($\langle w_{0,k}, w_{\star,l} = O(d^{-1/2})$ for all $k \in [p]$, $l \in [r]$), features $\varphi(x; w) = \sigma(\langle w, x)$ are uncorrelated to y. Since the predictor in assumption 2.1.1 belongs to the column space of the features, it makes no difference to assume a particular structure on $f_\star$.

2.1.1 Deterministic equivalent

The key technical idea in the analysis is to regard the feature matrix Φ as an empirical approximation of an operator on the Hilbert space $L^2(\mu_x \otimes \mu_w)$, where all relevant quantities in the problem can be diagonalised. This perspective allows us to draw on results from random operator theory to characterise the risk [KG00]. More precisely, we can associate to the features a Fredholm integral operator $\mathbb{T} : L^2(\mu_w) \to L^2(\mu_x)$:

$$(\mathbb{T}h)(x) = \int \varphi(x; w)h(w)\mu_x(\mathrm{d}w) \tag{2.6}$$

This is a compact operator, and therefore can be diagonalized:

$$\mathbb{T} = \sum_{m=1}^{\infty} \xi_m \phi_m \psi_m^\star, \tag{2.7}$$

where $(\xi_m)_{m\geq 1} \subseteq \mathbb{R}$ are the eigenvalues and $(\psi_m)_{k\geq 1}$ and $(\phi_m)_{k\geq 1}$ are orthonormal bases of $L^2(\mu_x)$ and $L^2(\mu_w)$, respectively:

$$\langle \psi_m, \psi_{m'} \rangle_{L^2(\mu_x)} = \delta_{mm'}, \qquad \langle \phi_m, \phi_{m'} \rangle_{L^2(\mu_w)} = \delta_{mm'}. \tag{2.8}$$

Without loss of generality, we assume the eigenvalues are ordered in non-increasing absolute values $|\xi_1| \geq |\xi_2| \geq \ldots$, and for simplicity of presentation we assume that all eigenvalues are non-zero, i.e., $\mathrm{Ker}(\mathbb{T}) = \{0\}$. Denote $\Sigma = \mathrm{diag}(\xi_1^2, \xi_2^2, \ldots) \in \mathbb{R}^{\infty\times\infty}$ the diagonal matrix of the squared eigenvalues. Similarly, since $f_\star \in L^2(\mu_x)$, it admits the following decomposition in $(\psi_m)_{k\geq 1}$:

$$f_\star = \sum_{k\geq 1} \theta_{\star,m} \psi_m \tag{2.9}$$

This decomposition effectively maps the non-linear problem in eq. (2.4) to a linear problem in an infinite dimensional Hilbert space. The question is then under which conditions on

ψ_m, ϕ_m and the eigenvalues ξ_m we can carry a random matrix theory analysis. A set of sufficient conditions are the following.

Assumption 2.1.2 (Tail spread)**.** There exists $\ell \in \mathbb{N}$ such that

$$p\xi_{\ell+1}^2 \leq \frac{\lambda}{n} \sum_{m=\ell+1}^{\infty} \xi_m^2. \tag{2.10}$$

Assumption 2.1.3 (Concentration of the eigenfunctions)**.** Denote the (infinite-dimensional) random vectors[23] $\psi := (\xi_m \psi_m)_{m \geq 1}$ and $\phi := (\xi_m \phi_m)_{m \geq 1}$. There exists a constant $\mathsf{C}_x > 0$ such that for any deterministic p.s.d. matrix $A \in \mathbb{R}^{\infty \times \infty}$,[24] with $\operatorname{Tr}(\Sigma A) < \infty$, we have

$$\mathbb{P}\left(|\langle \psi, A\psi \rangle - \operatorname{Tr}(\Sigma A)| \geq t \cdot \|\Sigma^{1/2} A \Sigma^{1/2}\|_F\right) \leq \mathsf{C}_x \exp\left\{-t/\mathsf{C}_x\right\}, \tag{2.11}$$

$$\mathbb{P}\left(|\langle \phi, A\phi \rangle - \operatorname{Tr}(\Sigma A)| \geq t \cdot \|\Sigma^{1/2} A \Sigma^{1/2}\|_F\right) \leq \mathsf{C}_x \exp\left\{-t/\mathsf{C}_x\right\}. \tag{2.12}$$

Assumption 2.1.2 is an assumption on the spread of the spectral tail. It essentially states that, relative to the scale $np\lambda^{-1}$, the tail of the spectrum must not be dominated by single eigenvalues. It holds, for instance, in the power law case $\xi_m^2 \propto m^{-\alpha}$ when $\alpha > 1$ as soon as $\ell \gtrsim (\alpha - 1) np\lambda^{-1}$. Instead, Assumption 2.1.3 is a Hanson-Wright type inequality on the concentration of the eigenfunctions [Ver18]. It essentially states that the eigenfunctions behave as sub-Gaussian vectors.

Under these assumptions, it can be shown that the risk can be approximated by a *deterministic equivalent*.

Definition 2.1.1 (Risk deterministic equivalent)**.** Given $\lambda > 0$, a positive-definite operator $\Sigma \in \mathbb{R}^{\infty \times \infty}$ and $\theta_\star \in \mathbb{R}^\infty$, let $\nu_2 \geq 0$ denote the unique solution of the following self-consistent equation:

$$1 + \frac{n}{p} - \sqrt{\left(1 - \frac{n}{p}\right)^2 + 4\frac{\lambda}{p\nu_2}} = \frac{2}{p} \operatorname{Tr}\left(\Sigma(\Sigma + \nu_2)^{-1}\right), \tag{2.13}$$

[23] Note that we can consider both ψ and ϕ random elements of the Hilbert space ℓ_2 with distribution induced by $x \sim \mu_x$ and $w \sim \mu_w$, where $\mathbb{E}[\psi\psi^\top] = \mathbb{E}[\phi\phi^\top] = \Sigma$ and $\operatorname{Tr}(\Sigma) < \infty$.

[24] In other words, a linear operator acting on an infinite-dimensional Hilbert space

and define the following short-hand:

$$\nu_1 = \frac{\nu_2}{2}\left[1 - \frac{n}{p} + \sqrt{\left(1 - \frac{n}{p}\right)^2 + 4\frac{\lambda}{p\nu_2}}\right]. \tag{2.14}$$

$$\Upsilon(\nu_1, \nu_2) = \frac{p}{n}\left[\left(1 - \frac{\nu_1}{\nu_2}\right)^2 + \left(\frac{\nu_1}{\nu_2}\right)^2 \frac{\mathrm{Tr}(\Sigma^2(\Sigma + \nu_2)^{-2})}{p - \mathrm{Tr}(\Sigma^2(\Sigma + \nu_2)^{-2})}\right], \tag{2.15}$$

$$\chi(\nu_2) = \frac{\mathrm{Tr}(\Sigma(\Sigma + \nu_2)^{-2})}{p - \mathrm{Tr}(\Sigma^2(\Sigma + \nu_2)^{-2})}. \tag{2.16}$$

Then, the *deterministic equivalent* for the excess risk is given by:

$$\mathsf{R}_{n,p}(\lambda, \theta_\star, \Sigma) = \mathsf{B}_{n,p}(\lambda, \theta_\star, \Sigma) + \mathsf{V}_{n,p}(\lambda, \Sigma) \tag{2.17}$$

$$\mathsf{B}_{n,p}(\lambda, \theta_\star, \Sigma) = \frac{\nu_2^2}{1 - \Upsilon(\nu_1, \nu_2)}\Big[\langle\theta_*, (\Sigma + \nu_2)^{-2}\theta_*\rangle + \chi(\nu_2)\langle\theta_*, \Sigma(\Sigma + \nu_2)^{-2}\theta_*\rangle\Big], \tag{2.18}$$

$$\mathsf{V}_{n,p}(\lambda, \Sigma) = \sigma^2 \frac{\Upsilon(\nu_1, \nu_2)}{1 - \Upsilon(\nu_1, \nu_2)}, \tag{2.19}$$

Note $\mathsf{R}_{n,p}(\lambda, \theta_\star, \Sigma)$ is a deterministic function depending only on the constants n, p, λ and on Σ and $\theta_\star$.

Theorem 2.1.1 ([DLM24], informal)**.** Let $\mathcal{R}$ denote the excess risk associated with the minimiser $\hat{a}_\lambda$ of eq. (2.4):

$$\mathcal{R}(f_\star, X, W_0, \lambda, \sigma^2) = \mathbb{E}_{\varepsilon, x\sim\mu_x}\left[(f(x; \hat{a}_\lambda, W_0) - f_\star(x))^2\right] \tag{2.20}$$

Then, for any $D > 0$, under assumption 2.1.2 and 2.1.3, with probability $1 - n^{-D} - p^{-D}$, $\mathcal{R}$ admits a *deterministic equivalent* $\mathsf{R}_{n,p} \in \mathbb{R}_+$:

$$|\mathcal{R} - \mathsf{R}_{n,p}| = \tilde{O}(n^{-1/2} + p^{-1/2}) \cdot \mathsf{R}_{n,p} \tag{2.21}$$

where $\mathsf{R}_{n,p}(\lambda, \sigma^2, \theta_\star, \Sigma)$ is given by definition 2.1.1.

This theorem tell us that, as soon as n, p are large, the risk, which is a function of the random quantities W_0, X, y can be well approximated by a function of the deterministic quantities $\theta_\star, \Sigma$. For conciseness, we omit some technical formal details in the statement of the theorem, and refer the interested reader to [DLM24].

Theorem 2.1.1 is considerably more general than previous results in the literature. First, it extends the dimension-free results of [CM24] for well-specified ridge regression and [MS24] for kernel ridge regression (see $p \to \infty$ discussion below). Moreover, the deterministic equivalent recovers as particular cases the asymptotic results derived under proportional $n, p = \Theta(d)$ [MM22; Lou+21a; Sch+23] and polynomial $n, p = \Theta(d^\kappa)$ [Xia+22; HLM24;

AFP25] scaling.

Remark 2.1.2. We draw the attention of the reader to the following important features of theorem 2.1.1.

- It provides non-asymptotic approximation bounds that hold pointwise. In particular, it does not require probabilistic assumptions over the target function coefficients $\theta_\star$.
- It does not explicitly dependent on the feature map dimension d. However, it enters implicitly through $\theta_\star, \Sigma$.
- The approximation is multiplicative, and therefore relative to the scale of the risk. In particular, they hold even if $\mathsf{R}_{n,p} \asymp n^{-\gamma}$, allowing the study of scaling laws, which will be discussed in section 2.3.
- The bound depends on λ^{-1} and $\lambda_{>\ell}^{-1}$. Following similar arguments as in [CM24; CM24], this assumption could be removed at the cost of worse rates $n^{-C} + p^{-C}$ with $C < 1/2$.

Corollary 2.1.1 (Kernel limit)**.** In the $p \to \infty$ limit both ν_1 and ν_2 converge to a single ν_K which is the unique positive solution to the following self-consistent equation

$$n - \frac{\lambda}{\nu_\mathsf{K}} = \operatorname{Tr}\left(\Sigma(\Sigma + \nu_\mathsf{K})^{-1}\right). \tag{2.22}$$

Moreover, the bias eq. (2.18) and variance eq. (2.19) terms simplify to:

$$\mathsf{B}_{\mathsf{K},n}(\theta_*, \lambda) = \frac{\nu_\mathsf{K}^2 \langle \theta_*, (\Sigma + \nu_\mathsf{K})^{-2}\theta_* \rangle}{1 - \frac{1}{n}\operatorname{Tr}(\Sigma^2(\Sigma + \nu_\mathsf{K})^{-2})}, \qquad \mathsf{V}_{\mathsf{K},n}(\lambda) = \sigma^2 \frac{\operatorname{Tr}(\Sigma^2(\Sigma + \nu_\mathsf{K})^{-2})}{n - \operatorname{Tr}(\Sigma^2(\Sigma + \nu_\mathsf{K})^{-2})}. \tag{2.23}$$

We denote the corresponding test error $\mathsf{R}_{\mathsf{K},n}(\theta_*, \lambda) = \mathsf{B}_{\mathsf{K},n}(\theta_*, \lambda) + \mathsf{V}_{\mathsf{K},n}(\lambda)$.

Note that eq. (2.19) exactly agrees with the dimension-free deterministic equivalents for kernel methods from [CM24; MS24].

Remark 2.1.3 (Degrees-of-freedom)**.** The quantities appearing the the expressions of eq. (2.25) are known as *degrees-of-freedom* [CD07]:

$$\mathsf{df}_1(\nu) = \operatorname{Tr}\left\{\Sigma(\Sigma + \nu)^{-1}\right\}, \qquad \mathsf{df}_2(\nu) = \operatorname{Tr}\left\{\Sigma^2(\Sigma + \nu)^{-2}\right\}. \tag{2.24}$$

The degrees-of-freedom can be seen as a "soft count" of how many eigenvalues are larger than the parameter ν, since eigenvalues $\xi_m^2 \ll \nu$ contribute to the trace, while eigenvalues $\xi_m^2 \gg \nu$ are shrank. With this notation, we can rewrite:

$$\mathsf{B}_{\mathsf{K},n}(\theta_*, \lambda) = \frac{\nu_\mathsf{K}^2 \langle \theta_*, \Sigma(\Sigma + \nu_\mathsf{K})^{-2}\theta_* \rangle}{1 - \frac{1}{n}\mathsf{df}_2(\nu_\mathsf{K})}, \qquad \mathsf{V}_{\mathsf{K},n}(\lambda) = \sigma^2 \frac{\mathsf{df}_2(\nu_\mathsf{K})}{n - \mathsf{df}_2(\nu_\mathsf{K})}. \tag{2.25}$$

where ν_{K} is the unique solution of $n - \lambda/\nu = \mathsf{df}_1(\nu)$. Comparing these expressions to eqs. (2.18) and (2.19) give an interesting interpretation of quantities appearing in the RF deterministic equivalent. Indeed, the variance in the RF case is controlled by Υ, which contains not only df_2 (as in the kernel case) but an additional term due to the additional randomness of W_0. Similarly, the bias term also contains an additional term due to the variance of W_0, given by $\chi(\nu_2)$. We refer the reader to [Lou+22] for a detailed discussion how to decompose the variance contribution of W_0 with respect to the other sources of randomness in the problem.

Finally, the second limit of interest is the $n \to \infty$ where data is abundant. In this case, the empirical risk eq. (1.4) converge to the population risk, and therefore the bottleneck in the risk is given by the capacity of the random feature class to approximate the target $f_\star$.

Corollary 2.1.2 (Approximation limit)**.** In the $n \to \infty$ limit, we have $\nu_1 \to 0$ and $\nu_2 \to \nu_{\mathsf{A}}$ satisfying the following simplified self-consistent equation:

$$p = \mathrm{Tr}\left(\Sigma(\Sigma + \nu_{\mathsf{A}})^{-1}\right). \tag{2.26}$$

Moreover, the bias eq. (2.18) and variance eq. (2.19) terms simplify to:

$$\mathsf{B}_{\mathsf{A},p}(\theta_*) = \nu_{\mathsf{A}}\langle\theta_*, (\Sigma + \nu_{\mathsf{A}})^{-1}\theta_*\rangle, \qquad \mathsf{V}_{\mathsf{A},n} = 0. \tag{2.27}$$

We denote the risk in this case $\mathsf{R}_{\mathsf{A},p}(\theta_*) = \mathsf{B}_{\mathsf{A},p}(\theta_*)$, which as expected does not depend on λ.

2.1.2 Intuition

As previously hinted, the key intuition behind theorem 2.1.1 is to regard the feature matrix as the empirical version of an infinite dimensional Fredholm operator. To see this more precisely, consider again the diagonalisation of the feature map in a basis of $L^2(\mu_x \otimes \mu_w)$:

$$\varphi(x; w) = \sum_{m=1}^{\infty} \xi_m \phi_m(w)\psi_m(x)^*, \tag{2.28}$$

Defining the "matrices" $U \in \mathbb{R}^{n\times\infty}$ and $V \in \mathbb{R}^{p\times\infty}$ with components:

$$U_{im} = \psi_m(x_i), \qquad V_{km} = \phi_m(w_k) \tag{2.29}$$

The feature matrix $\Phi \in \mathbb{R}^{n\times p}$ can be re-written as $\Phi = U\Lambda V^\top$, where $\Lambda = \mathrm{diag}(\xi_m)$. Under the assumptions in the problem, the matrices U, V behave similar to a Gaussian matrices with rank n and p, respectively. For instance, since $x_i \sim \mu_x$ are independently sampled, U is close

to left-orthogonal due to the law of large numbers:

$$(U^\top U)_{mm'} = \sum_{i=1}^{n} \psi_m(x_i)\psi_{m'}(x_i) \sim n\left(\delta_{mm'} + O(n^{-1/2})\right) \tag{2.30}$$

On the other hand, assumption 2.1.3 on concentration of quadratic forms implies it behaves similarly to a Gaussian matrix on the right-side. Although heuristic, this intuition is exact: the expression in theorem 2.1.1 can be derived by pretending U, V are Gaussian matrices and following standard random matrix theory arguments for Wishart matrices.

This is an instance of *Gaussian universality*, which will be discussed in more detail in section 2.5.1.

2.1.3 Comparison with high-dimensional ridge regression

The formulas in definition 2.1.1 bear close resemblance with a the formulas for the classical random design analysis of ridge regression in the proportional regime. Indeed, this connection can provide an useful intuition for this result.

Consider the standard well-specified ridge regression problem, where we are interested in studying the performance of the linear predictor $f(x;\beta) = \langle \hat{\beta}_\lambda, x\rangle$ with $\hat{\beta}_\lambda \in \mathbb{R}^d$:

$$\hat{\beta}_\lambda(X, y) = \left(X^\top X + \lambda I_d\right)^{-1} X^\top y \tag{2.31}$$

for data $\mathcal{D} = \{(x_i, y_i) \in \mathbb{R}^{d+1} : i \in [n]\}$ drawn according to:

$$y_i = \langle \beta_\star, x_i\rangle + \varepsilon_i, \qquad x_i \sim \mu_x, \tag{2.32}$$

where $\mathbb{E}[\varepsilon_i|x] = 0$ and $\mathbb{E}[\varepsilon_i^2] = \sigma^2 < \infty$. Opening up the expressions, it is easy to show that the excess risk admits the following bias-variance decomposition:

$$\mathcal{R}(\beta_\star, X, \lambda, \sigma^2) = \mathbb{E}_{\varepsilon, x\sim\mu_x}\left[\left(\langle \hat{\beta}_\lambda, x\rangle - \langle \beta_\star, x\rangle\right)^2\right] = \mathcal{B}(\beta_\star, X, \lambda) + \mathcal{V}(X, \lambda, \sigma^2) \tag{2.33}$$

where:

$$\mathcal{B}(\beta_\star, X, \lambda) = \lambda^2 \langle \beta_\star, \left(X^\top X + \lambda I_d\right)^{-1} \Sigma \left(X^\top X + \lambda I_d\right)^{-1} \beta_\star\rangle \tag{2.34}$$

$$\mathcal{V}(X, \lambda, \sigma^2) = \sigma^2 \operatorname{Tr}\left\{X^\top X \left(X^\top X + \lambda I_d\right)^{-2} \Sigma\right\} \tag{2.35}$$

where $\Sigma = \mathbb{E}[xx^\top]$. The high-dimensional asymptotics of eq. (2.34) has been studied by several works in the literature, under various assumptions on the covariate distribution μ_x, target weights $\beta_\star \in \mathbb{R}^d$ and covariance $\Sigma \in \mathbb{R}^{d\times d}$, see for example [DW18; WX20; Lou+21a; Has+22; Bac24]. For concreteness, consider Proposition 4.1 from [Bac24].

Proposition 2.1.1 ([Bac24]). Assume the following assumptions hold in the high-dimensional limit $d \to \infty$:

- $x_i = \Sigma^{1/2} z_i$ with z_i i.i.d., zero mean, unit variance sub-Gaussian random variables and $\Sigma = \sum_{j=1}^{d} \tau_j v_j v_j^\top$ with non-increasing $\tau_j > 0$ and $\sigma_1 = ||\Sigma||_{\text{op}} < \infty$
- The empirical spectral measure $1/d \sum_{j=1}^{d} \delta_{\tau_j}$ converge to a compactly supported probability density ρ on $\mathbb{R}_+$.
- $||\beta_\star||_2 < \infty$ and $\sum_{j=1}^{d} \langle v_j, \beta_\star \rangle \delta_{\tau_j}$ converge to a measure with bounded mass.

Then, in the proportional limit where $n, d \to \infty$ with $n = \Theta(d)$:

$$\mathcal{R}(\beta_\star, X, \lambda, \sigma^2) \xrightarrow{a.s.} \mathsf{R}_{\text{prop}} = \mathsf{B}_{\text{prop}} + \mathsf{V}_{\text{prop}} \tag{2.36}$$

where:

$$\mathsf{B}_{\text{prop}}(\beta_\star, \Sigma, \lambda, \gamma) = \frac{\nu(\lambda)^2 \langle \beta_\star, \Sigma \left(\Sigma + \nu(\lambda) I_d\right)^{-2} \beta_\star \rangle}{1 - \frac{1}{n} \operatorname{Tr}\left\{\Sigma^2 (\Sigma + \nu(\lambda) I_d)^{-2}\right\}} \tag{2.37}$$

$$\mathsf{V}_{\text{prop}}(\Sigma, \lambda, \sigma^2, \gamma) = \sigma^2 \frac{\operatorname{Tr}\left\{\Sigma^2 \left(\Sigma + \nu(\lambda) I_d\right)^{-2}\right\}}{n - \operatorname{Tr}\left\{\Sigma^2 (\Sigma + \nu(\lambda) I_d)^{-2}\right\}} \tag{2.38}$$

with $\nu(\lambda) \in \mathbb{R}_+$ the unique solution of the following self-consistent equation:

$$n - \frac{\lambda}{\nu} = \operatorname{Tr}\left\{\Sigma (\Sigma + \nu I_d)^{-1}\right\} \tag{2.39}$$

Note that even though we write the bias and variance terms as a function of the diverging dimensions n, d and the high-dimensional objects $\beta_\star$ and Σ, all the quantities involved in the above expression are finite. For example,

$$\frac{1}{n} \operatorname{Tr}\left\{\Sigma^2 (\Sigma + \nu I_d)^{-2}\right\} \to \gamma \int \rho(\mathrm{d}\tau) \frac{\tau^2}{(\tau + \nu)^2}. \tag{2.40}$$

Therefore, the expressions in eqs. (2.37) and (2.38) could be written only as a function of scalar quantities. However, for the sake of comparison, it is useful to write them in this way.

Remark 2.1.4 (Comparison with KRR). Note that the equations are almost identical to the kernel ridge regression deterministic equivalent from corollary 2.1.1. The only difference is the presence of an additional Σ in enumerator of the bias:

$$\langle \beta_\star, \Sigma (\Sigma + \nu)^{-2} \beta_\star \rangle \qquad \text{vs.} \qquad \langle \theta_\star, (\Sigma + \nu)^{-2} \theta_\star \rangle \tag{2.41}$$

This can be understood from the fact that the covariate x_i plays the role of the kernel features in KRR. Indeed, in KRR we decomposed the target function as a linear combination in the

basis $(\psi_m)_{m\geq 0}$ of $L^2(\mu_x)$, and not in the feature basis $f_\star(x) = \langle \theta_\star, x_i \rangle$ as we did in the ridge regression case. Therefore, a one-to-one identification is given by $\theta_\star = \Sigma^{1/2}\beta_\star$.

The fact that the high-dimensional deterministic equivalents agree exactly with the non-asymptotic is remarkable but intuitive. Indeed, ridge regression is equivalent to a linear kernel, and the sub-Gaussian assumption on the covariates implies assumption 2.1.3.

2.2 The double descent phenomenon

As motivated in Section 1.2, an important observation in modern deep learning practice is that overparametrised neural networks can achieve small excess risk even while interpolating the training data.

This behaviour contrasts with the traditional view in statistics, according to which increasing the number of parameters beyond a certain point leads to overfitting and consequently a degradation of risk. This intuition is formalised in the *bias–variance trade-off*, commonly taught in introductory statistics: a model should be expressive enough to capture the underlying signal, but not so flexible that it fits the noise [Jam+13]. The empirical observation that neural networks can generalise while overfitting the training data defies this picture, while also challenging complexity based statistical learning bounds such as theorem 1.2.2 [GBD92; Zha+16].

Instead, the deterministic equivalent from theorem 2.1.1 give us access to an exact characterisation of the excess risk for the two-layer neural network at initialisation. As a first application of this result, we can study the exact dependence of the bias and variance terms on the network width p, as well as how it depends on the geometry of the data distribution.

To study the empirical risk interpolator, consider the ridgeless limit where $\lambda \to 0^+$. In this limit, the ridge estimator reduces to the ordinary least squares estimator:

$$\hat{a}_{\text{ols}}(X, y) = \Phi^\dagger y \tag{2.42}$$

where $\Phi^\dagger \in \mathbb{R}^{p\times n}$ denotes the Moore-Penrose inverse of $\Phi \in \mathbb{R}^{n\times p}$.

There are two regimes of interest:

- **Underparametrised regime ($n > p$):** In this case, $\nu_1 = 0$ and ν_2 satisfies $\operatorname{Tr}\{\Sigma(\Sigma + \nu_2)^{-1}\} = p$. It follows that $\Upsilon = {}^{n}/_{p}$, and we can simplify the deterministic equivalent for the bias and variance:

$$\mathsf{B}_{n>p} = \frac{n}{n-p}\nu_2\langle \theta_\star, (\Sigma + \nu_2)^{-1}\theta_\star \rangle, \qquad \mathsf{V}_{n>p} = \sigma^2\frac{p}{n-p} \tag{2.43}$$

  Interestingly, the variance coincides exactly with that of well-specified ordinary least squares in the classical regime $n > p$. It is a monotonically increasing function of p

for $p < n$, diverging precisely at $p = n$. By contrast, the bias is not monotonic in p: while the factor $\frac{n}{n-p}$ increases with p, the term $\nu_2 \langle \theta_\star, (\Sigma + \nu_2)^{-1} \theta_\star \rangle$ decreases, with the rate of decay governed by $\theta_{\star,m}$ and ξ_m^2 (the faster these decay, the faster this term decreases).

- **Overparametrised regime ($n < p$):** In this regime, we have $\nu_1/\nu_2 = 1 - n/p > 0$, with ν_2 determined by the condition $\mathrm{Tr}\{\Sigma(\Sigma + \nu_2)^{-1}\} = n$. Notably, ν_2 does not depend on the width p. The expressions for the bias and variance do not simplify further, but one can show that the risk starts from the peak $\mathsf{R}_{n,p} \sim (p - n)^{-1}$ as $p \to n^-$ and then decreases monotonically with p, eventually reaching the plateau value given by the kernel limit described in corollary 2.1.1.

The increasing behaviour of the variance and the decreasing behaviour of the bias as the width p grows are consistent with the classical bias–variance trade-off. What is more unusual, however, is that the bias itself begins to increase through the variance-like factor $\frac{n}{n-p}$ once p approaches n. This unexpected behaviour of the bias can be traced back to an additional source of variance induced by the randomness of the weights W_0. A finer bias–variance decomposition, which separates these contributions, makes this point explicit; see [dAs+20; AP20; LD21; Lou+22] for detailed discussions.

The divergence of the excess risk at $n = p$ is referred to as the *interpolation peak*. It occurs exactly at the transition where the linear system $y = \Phi a$ changes from being overdetermined (no exact solution) to underdetermined (infinitely many solutions), with a unique solution $\hat{a}_{\mathrm{ols}} = \Phi^{-1} y$ at the critical point $n = p$. This is the point with largest variance, and indeed can be mitigated by properly regularising the problem [KH91; Nak+21b].

Finally, the second descent in the overparametrised region ($p > d$) is known as the *double descent phenomenon* [Bel+19]. Early works on exact asymptotics for linear regression already observed a second descent [Opp+90; KH91], but in that setting the minimum error occurs before interpolation ($p < n$), consistent with the classical bias–variance trade-off. By contrast, empirical studies have shown that neural networks can continue to improve their performance even beyond the interpolation peak [GBD92; Spi+19; Nak+21a], a behaviour referred to as *benign overfitting*.

Interestingly, both behaviours are captured by the two-layer network with fixed first layer weights, and depend on the interaction between the data distribution and the network architecture, here parametrised by $\theta_\star$, σ^2, and Σ. In particular, the location of the first minimum is controlled by the rate of decay of Σ and $\theta_\star$, which control how easy is the task. Figure 2.1 illustrates these two scenarios for a task with $\theta_\star = e_1$ and $\sigma^2 = 0.01$. With a fast-decaying spectrum $\xi_m^2 = 2^{-m}$, the risk is minimised before interpolation ($p < n$), and performance deteriorates afterwards, corresponding to *malign* overfitting. By contrast, with a slowly decaying spectrum $\xi_m^2 = m^{-1.2}$, we observe benign overfitting, where the best risk

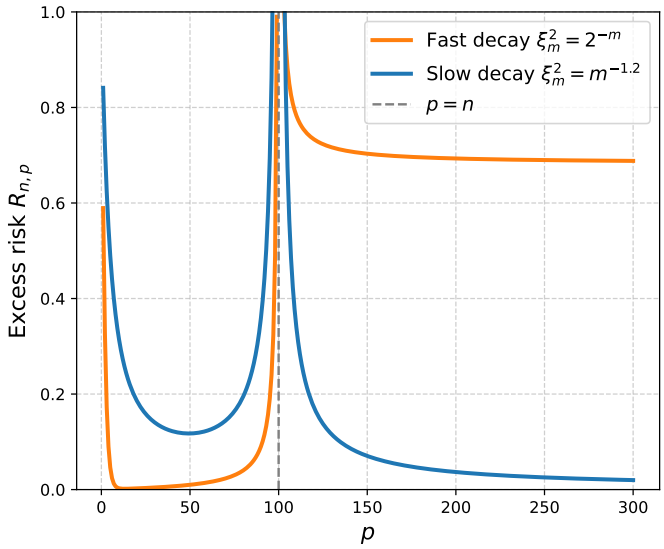

Figure 2.1: Excess risk $\mathsf{R}_{n,p}$ as a function of the network width p for $n = 100$, $\sigma^2 = 0.01$ and $\theta_\star = e_1$. The orange curve shows a *malign overfitting* case with fast decaying spectrum $\xi_m^2 = 2^{-m}$, while the blue curve shows a *benign overfitting case* with slow decaying spectrum $\xi_m^2 = m^{-1.2}$.

is achieved in the kernel regime $p \to \infty$. For a detailed discussion on the criteria for benign overfitting in linear models, see [Bar+20; CM24].

2.3 Scaling laws

A classical question in learning theory is to determine how quickly the excess risk converges to its asymptotic value as $n \to \infty$. The convergence rate depends critically on both the hypothesis class and the structure of the data. In the setting of random features regression, this rate is governed by a trade-off between the regularity of the target function $f_\star$ and the expressivity of the random features kernel. Concretely, it is determined by the relative decay of the spectrum of the operator in eq. (2.7) and of the target coefficients $\theta_\star$ when expanded in its eigenfunction basis, eq. (2.9).

Recall that as long as the Fredholm operator $\mathbb{T}$ is full rank ($\xi_m^2 > 0$ for all $m \geq 0$), the RKHS $\mathcal{H} = \mathrm{Im}(\mathbb{T})$ spanned by the $p \to \infty$ RF kernel is dense in $L^2(\mu_x)$ (i.e. it is a *universal approximator*). From a functional-analytic perspective, the relative decay therefore quantifies the "size" of $\mathcal{H}$ within $L^2(\mu_x)$. See [Bac17b] for further discussion of this viewpoint.

Instead of studying specific target functions and kernel, it is common to study a family of problems parametrised by a given relative decay of these quantities. A classical setup in

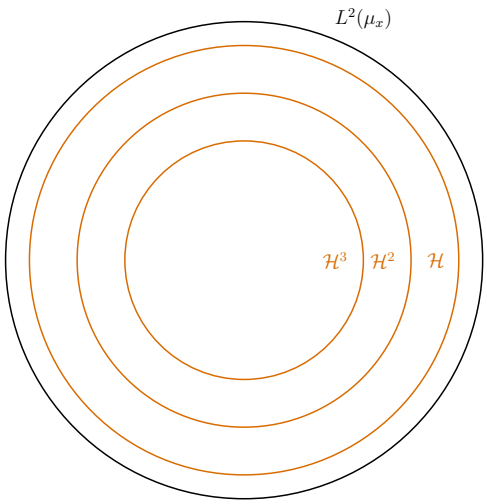

Figure 2.2: (**Left**) Illustration of the source condition, where the spaces $\mathcal{H}^{2r}$ corresponding to functions with finite $||\Sigma^r \theta_\star|| < \infty$ are nested in an increasing order $\mathcal{H}^3 \subset \mathcal{H}^2 \subset \mathcal{H} \subset L^2(\mu_x)$.

this direction are the *source and capacity* conditions [CD07]:

$$\operatorname{Tr} \Sigma^{1/\alpha} < \infty \qquad \text{(Capacity)}$$

$$||\Sigma^r \theta_\star||_{L^2(\mu_x)} < \infty \qquad \text{(Source)}$$

where $\alpha > 1$ and $r \geq 0$. Note these conditions are equivalent to a power-law assumption on the decay of the eigenvalues and target coefficients, i.e. there exists constants C_1, C_2 such that:

$$\xi_m^2 = C_1 m^{-\alpha}, \qquad \theta_{*,m} = C_2 m^{-\frac{1+2\alpha r}{2}}. \tag{2.44}$$

Note that the larger r, the faster de target coefficients decay, meaning that the easier it is to express it with the kernel. In particular, for $r < 1/2$ we have $f_\star \notin \mathcal{H}$ while for $r \geq 1/2$ we have $f_\star \in \mathcal{H}$, see fig. 2.2 for an illustration. A similar role is played by the capacity exponent $\alpha > 1$: the faster is the decay of the kernel eigenvalues, the larger is the kernel capacity. An alternative and useful picture is to think of α as controlling the effective dimension of the kernel feature space [Zha05]. This can be made quantitative through the lens of the degree-of-freedom $\mathsf{df}_1(\nu)$ introduced in eq. (2.24), which can be seen as a measure of the effective dimension of the feature space. Then, the source condition is equivalent to $\mathsf{df}_1(\nu) \leq C\nu^{-1/\alpha}$, i.e. the larger $\alpha > 0$, the smaller the effective dimension of the RKHS.

Recent empirical studies in deep learning have shown that the large-scale performance of neural networks often follows a scaling-law relationship with the number of samples, model size, and computational budget [Kap+20; Hes+17]. These *neural scaling laws* have had a major impact on practice, as they provide a principled way to scale up small models (for instance, those used in fine-tuning) while avoiding performance bottlenecks — such as oversizing a

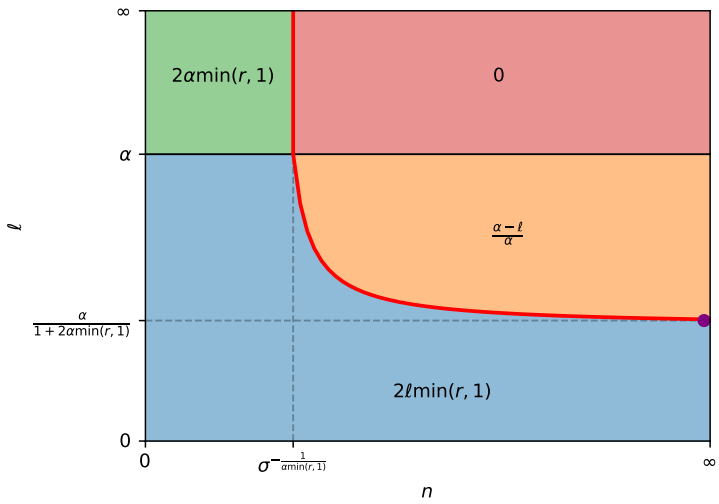

Figure 2.3: Different decays for the excess risk for different values of n and different regularization $\lambda = \Theta(n^{-\ell-1})$ decays, at given noise variance $\sigma \geq 0$. The red solid line represents the noise-induced crossover line, separating the effectively noiseless regime (green and 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 $\ell > \alpha/(1 + 2\alpha(r \wedge 1))$. The corresponding learning curve will accordingly exhibit a crossover from a fast decay (noiseless regime) to a slow decay (noisy regime). Plot from [Cui+21].

model when limitations actually stem from data or compute.

Neural scaling laws have sparked renewed interest in the deep learning theory community in the classical literature on kernel source and capacity conditions. Building on the connection between kernels and neural networks discussed in chapter 2, several recent works have examined these laws through a kernel lens [MRS22; Bah+24; AZP24; BAP24; Paq+24; Lin+24]. As we will see, key features of the observed neural scaling laws — such as the trade-offs between model and sample complexity — are captured by random features regression.

2.3.1 Kernel ridge regression rates

As a starting point, consider the kernel limit where $p \to \infty$. The excess risk rates under the source and capacity conditions in eq. (2.44) can be obtained from corollary 2.1.1.

Theorem 2.3.1 (KRR excess risk rates, [Cui+21])**.** Under source and capacity conditions eq. (2.44) and regularisation scaling $\lambda = \Theta(n^{-\ell-1})$, the deterministic equivalent eq. (2.17) rate is given by:

$$\mathsf{R}_{\mathsf{K},n}(\sigma^2, \alpha, r, \ell) = \Theta\left(n^{-2\alpha\left(\frac{\ell}{\alpha}\wedge 1\right)(r\wedge 1)} + \sigma^2 n^{-1+\left(\frac{\ell}{\alpha}\wedge 1\right)}\right) \tag{2.45}$$

We refer the interested reader to [Cui+21] for a derivation of these rates from corollary 2.1.1. Theorem 2.3.1 is best summarised in a diagram, see fig. 2.3. We can distinguish

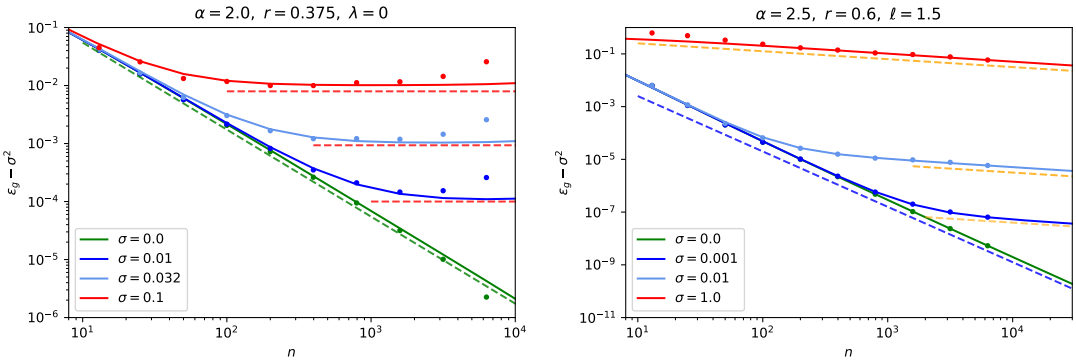

Figure 2.4: Excess risk for KRR under source and capacity conditions as a function of the number of samples. **(Left)** Cross-over between the green and red (plateau) regions in the weakly regularisation regime $\ell > \alpha$. **(Right)** Cross-over between the blue and orange (plateau) regions in the strongly regularisation regime $\ell < \alpha$. Solid lines correspond to the theoretical prediction of theorem 2.3.1 and points are simulations conducted using the python `scikit-learn KernelRidge` package [Ped+11]. Dashed lines represent the slopes predicted by eq. (2.46), with the colours in correspondence to the regime from fig. 2.3. Plot from [Cui+21]

four different regimes, with cross-overs between bias-dominated and variance-dominated rates.

- **Weak regularisation** $\ell \geq \alpha$ (green and red regions),

$$\mathsf{R}_{\mathsf{K},n} = \Theta\left(\max\left(\sigma^2, n^{-2\alpha(r\wedge 1)}\right)\right) . \tag{2.46}$$

  The excess error transitions from a fast decay $2\alpha(r \wedge 1)$ (green region in fig. 2.3 and green dashed line in fig. 2.4 (left) to a plateau (red region in fig. 2.3 and red dashed line in fig. 2.4 (left) with no decay as n increases. This corresponds to a crossover from the green region to the red region in the phase diagram fig. 2.3.

- **Strong regularisation** $\ell \leq \alpha$ (blue and orange regions),

$$\mathsf{R}_{\mathsf{K},n} = \Theta\left(\max\left(\sigma^2, n^{1-2\ell\min(\mathrm{r},1)-\frac{\ell}{\alpha}}\right) n^{\frac{\ell-\alpha}{\alpha}}\right) . \tag{2.47}$$

  The excess error transitions from a fast decay $2\ell(r \wedge 1)$ (blue region in fig. 2.3) to a slower decay $(\alpha - \ell)/\alpha$ (orange region in fig. 2.3) as n is increased and the effect of the additive noise kicks in, see fig. 2.4 (right). The crossover disappears for too slow decays $\ell \leq \alpha/(1 + 2\alpha(r \wedge 1))$, as the regularization λ is always sufficiently large to completely mitigate the variance. This corresponds to the max in (2.47) being realized by its second argument for all n.

An important curve in the diagram of fig. 2.3 is given by the optimal choice of regularisation, i.e. the value $\ell_\star$ leading to fastest decreasing of the excess risk.

Corollary 2.3.1 (Optimal KRR rates)**.** The optimal excess risk rates achieved by the KRR under source and capacity conditions eq. (2.44) and regularisation $\lambda = \Theta(n^{-\ell-1})$ are given by:

- **Noiseless rate:** If $n \ll \sigma^{-\frac{1}{\alpha(r\wedge 1)}}$, any $\ell_\star \in (\alpha, \infty)$ yields excess error decay

$$\mathsf{R}_{\mathsf{K},n} = \Theta\left(n^{-2\alpha(r\wedge 1)}\right) \tag{2.48}$$

  This corresponds to the vertical red line (—) in fig. 2.3.

- **Noisy rate:** If $n \gg \sigma^{-\max\left(2, \frac{1}{\alpha(r\wedge 1)}\right)}$, then there exists an unique optimal regularisation scale $\ell^\star = \frac{\alpha}{1+2\alpha(r\wedge 1)}$ yielding the minimax excess risk rates:

$$\mathsf{R}_{\mathsf{K},n} = \Theta\left(n^{-\frac{2\alpha(r\wedge 1)}{1+2\alpha(r\wedge 1)}}\right). \tag{2.49}$$

  This corresponds to the red curve in fig. 2.3.

Remark 2.3.1 (Relationship with literature)**.** Theorem 2.3.1 and corollary 2.3.1 encompasses some known rates in the kernel ridge regression literature.

- The optimally regularised noisy rate in eq. (2.49) corresponding to the $n \to \infty$ asymptotic (purple dot in fig. 2.4) is the well-known rate from Caponnetto and De Vito [CD07]. This is also the minimax rate under linear hypothesis and source condition $r > 1/2$. Curiously, corollary 2.3.1 shows the existence of a faster, *noiseless rate* at low sample complexity. Therefore, the optimal decay for the excess risk exhibits a cross-over from a fast decay $2\alpha(r \wedge 1)$ — corresponding to, effectively, the optimal rates expected in a "noiseless" situation — to a slower decay $2\alpha(r \wedge 1)/(1 + 2\alpha(r \wedge 1))$ corresponding to the classical "noisy" optimal rate. This is illustrated in fig. 2.5 where the two rates are observed in succession for the same data as the number of points is increased.

- The nomenclature "noiseless" is motivated by [SGW20b; BCP20b], who obtained these rates in a noiseless data setting.

Note that in all scenarios (whether with fixed or optimal regularisation) one observes a *cross-over* from an effectively *noiseless* regime (green or blue in fig. 2.4) to an effectively *noisy* regime (red or orange in fig. 2.4), depending on the amount of data available. Importantly, although noise is present in the green and blue "noiseless" regimes, its effect is not felt, and the learning curves follow noiseless rates. In fact, if the noise level is small, the classical noisy rates may only appear once an astronomical amount of data is available. Intuitively, for small sample sizes n low-variance directions of the feature space are used to overfit the noise, while the high-variance directions are accurately captured. In these noiseless regions, the excess

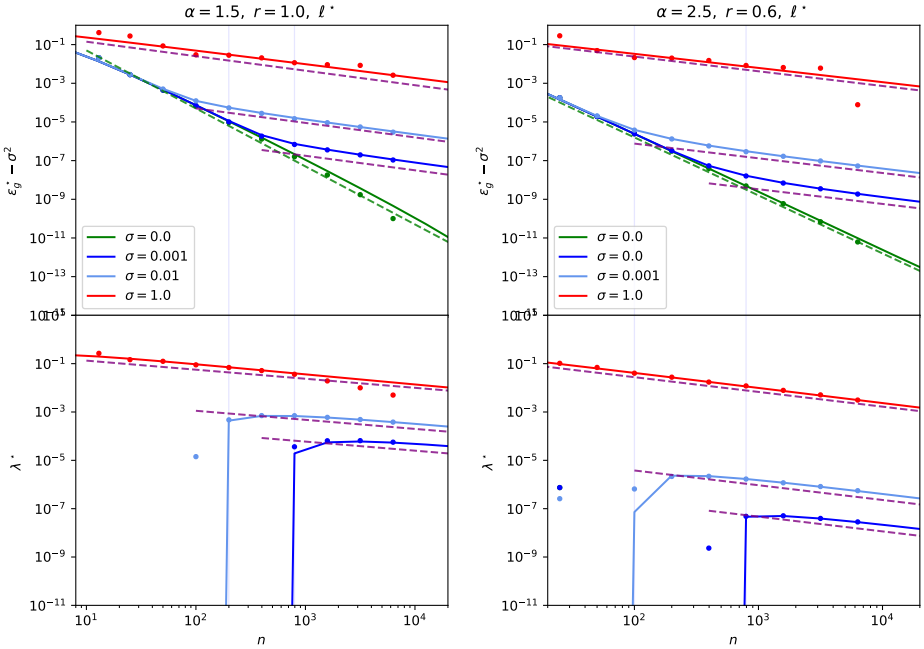

Figure 2.5: Excess risk as a function of the number of samples, under optimal regularisation $\lambda_\star = \Theta(n^{-\ell_\star})$. Solid lines correspond to the predictions from corollary 2.1.1. Points are simulations conducted with the python `scikit-learn KernelRidge` package [Ped+11]. In simulations, the best $\lambda^\star$ was determined using python `scikit-learn GridSearchCV` cross validation package [Ped+11]. 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 corollary 2.3.1, with the colours in correspondence to the regimes in fig. 2.4. (**Top**) excess error. (**Bottom**) Optimal $\lambda_\star$. Note the noiseless case has $\lambda_\star = 0$. Plot from [Cui+21]

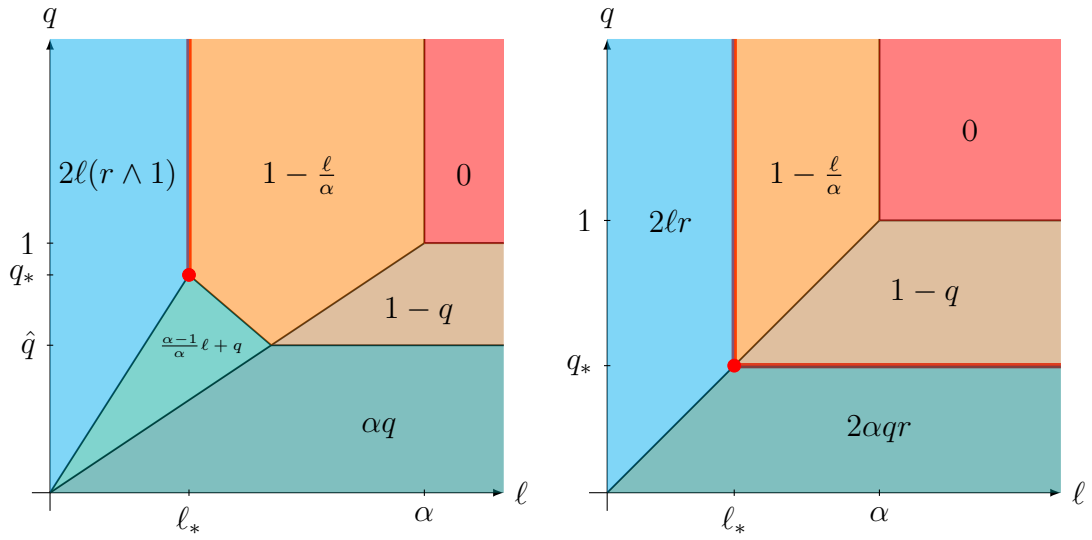

Figure 2.6: Excess error rate γ in the noisy regime $n \gg \sigma^{-1/(\gamma_B(\ell,q)-\gamma_V(\ell,q))}$ as a function of (ℓ, q) given by theorem 2.3.2. for $r \geq 1/2$ **(Left)** and $r \in [0, 1/2)$ **(Right)**. The explicit crossover points $\ell_\star, q_\star, \hat{q}$ are defined in eq. (2.53) as a function of the source r and capacity α exponents. Plot from [DLM24].

error is therefore dominated by a fast decay. This phenomenon, where the noise variance is diluted over the dimensions of lesser importance, is connected to *benign overfitting* [Bar+20; TB23].

2.3.2 Random features ridge regression rates

Although more involved, a similar analysis can be carried over on the deterministic equivalent of the random features excess risk from theorem 2.1.1.

Theorem 2.3.2 (RFRR excess risk rates)**.** Under source and capacity conditions eq. (2.44), regularisation $\lambda = \Theta(n^{-\ell-1})$ and width $p = \Theta(n^q)$ scaling, the deterministic equivalent rate from eq. (2.17) is given by:

$$\mathsf{R}_{n,p}(\sigma^2, \alpha, r, \ell) = \Theta\left(n^{-\gamma_B(\ell,q)} + \sigma^2 n^{-\gamma_V(\ell,q)}\right) = \Theta\left(n^{-\gamma(\ell,q)}\right), \tag{2.50}$$

where $\gamma(\ell, q) = \gamma_B(\ell, q) \wedge \gamma_V(\ell, q)$ for non-zero noise variance $\sigma^2 \neq 0$, otherwise $\gamma(\ell, q) = \gamma_B(\ell, q)$. The exponents γ_B and γ_V are respectively the decay rates of the bias and variance terms eqs. (2.18) and (2.19), and are explicitly given by

$$\gamma_B = \left[2\alpha\left(\frac{\ell}{\alpha} \wedge q \wedge 1\right)(r \wedge 1)\right] \wedge \left[\left(2\alpha\left(r \wedge \frac{1}{2}\right) - 1\right)\left(\frac{\ell}{\alpha} \wedge q \wedge 1\right) + q\right], \tag{2.51}$$

$$\gamma_V = 1 - \left(\frac{\ell}{\alpha} \wedge q \wedge 1\right). \tag{2.52}$$

In particular, note these reduce to the kernel rates from theorem 2.3.1 when $q \to \infty$.

We refer the interested reader to [DLM24] for a derivation of these rates from theorem 2.1.1. Again, the expressions in eq. (2.51) are easier to visualise in a diagram. Figure 2.6 shows the excess risk exponent $\gamma(\ell, q)$ as a function of the parameters ℓ and q, in the case where $\sigma^2 \neq 0$ for $r \geq 1/2$ (left) and $r < 1/2$ (right). Note that the key difference between the diagrams is the presence of an additional region for $r \geq 1/2$.[25] Defining the following shorthand:

$$\ell_\star = \frac{\alpha}{2\alpha(r \wedge 1) + 1}, \qquad q_\star = 1 - \ell_\star(2r \wedge 1), \qquad \hat{q} = \frac{1}{\alpha(2r \wedge 1) + 1} = q_\star \vee \frac{1}{\alpha + 1} \tag{2.53}$$

we can identify two main regions in the (ℓ, q) plane, corresponding to a trade-off between the bias γ_B and variance γ_V terms:

(a) **Variance dominated region** ($\gamma_V < \gamma_B$): if $\ell > \ell_\star$, $q > \hat{q}$ and $p > \lambda$, the excess risk is dominated by the variance term, provided the number of samples is large enough $n \gg \sigma^{-1/(\gamma_B(\ell,q) - \gamma_V(\ell,q))}$. Inside this region it is possible to further distinguish between two regimes:

- **Slow decay regime** (orange and brown): for $\ell < \alpha$ and $q < 1$ ($p \ll n$), $\gamma_V = 1 - (\ell/\alpha \wedge q)$, hence the decay depends on the interplay between regularization strength and number of random features and it is slower as $(\ell/\alpha \wedge q)$ increases;
- **Plateau regime** (red): for $\ell \geq \alpha$ and $q \geq 1$ ($p \geq n$) the excess risk converges to a constant value and does not decay as n increases.

(b) **Bias dominated region** ($\gamma_V > \gamma_B$): if $\ell < \ell_\star$, $q < \hat{q}$ and $p < \lambda$, the excess risk is dominated by the bias term, whose decay is faster as $(\ell/\alpha \wedge q)$ increases (cyan, emerald and teal).

Analogously to the discussion in section 2.3.1 for the KRR case, one can also identify a noiseless regime, valid when $n < \sigma^{-1/(\gamma_B(\ell,q) - \gamma_V(\ell,q))}$, with the corresponding cross-overs to the noisy regime as n increases. For conciseness, we omit the details here and refer the interested reader to [DLM24] for a complete account.

As in the KRR setting, one may further ask how to optimally tune the regularisation and width scales in order to achieve the fastest decay of the excess risk. This is obtained by optimising over ℓ and q in the formula above.

Corollary 2.3.2 (Optimal rates)**.** The optimal excess risk rate achieved by RFRR under source and capacity conditions eq. (2.44), regularisation $\lambda = \Theta(n^{-\ell-1})$ and width $p = \Theta(n^q)$

[25] Recall that these two cases correspond to the target function $f_\star$ belonging ($r \geq 1/2$) or not ($r < 1/2$) to the RKHS spanned by the asymptotic kernel

is given by:

$$\gamma_\star = \max_{\ell,q} \gamma(\ell, q) = \frac{2\alpha(r \wedge 1)}{2\alpha(r \wedge 1) + 1}, \tag{2.54}$$

and it is attained for:

$$\begin{cases} \lambda = \lambda_\star = n^{-(\ell_\star - 1)} \\ p \geq p_\star = n^{q_*} = \lambda_\star \end{cases} \qquad \text{for } r \geq 1/2, \tag{2.55}$$

$$\begin{cases} \lambda = \lambda_\star \\ p \geq p_\star = (\lambda_\star^{-1} n)^{1/\alpha} \end{cases} \quad \text{or} \quad \begin{cases} \lambda \leq \lambda_\star \\ p = p_\star = (\lambda_\star^{-1} n)^{1/\alpha} \end{cases} \qquad \text{for } r < 1/2 \tag{2.56}$$

corresponding to the bold red line (—) in fig. 2.6. In particular, the minimal number of random features $p_\star = n^{q_\star}$ required to achieve the optimal rate $\gamma_\star$ is given by:

$$q_\star = 1 - \frac{\alpha(2r \wedge 1)}{2\alpha(r \wedge 1) + 1} \tag{2.57}$$

and corresponds to the bold red dot (•) in fig. 2.6.

Remark 2.3.2 (Relationship with the literature)**.** Theorem 2.3.2 and corollary 2.3.2 relate to different results in the source & capacity and neural scaling law literature.

- **Minimal width:** As expected, the optimal excess risk for RFRR in the noisy regime are consistent with corollary 2.3.1 and the minimax optimal rates from [CD07]. An important question in the context of the random features approximation of kernels is: *what is the minimal number of random features $p_\star = n^{q_\star}$ in order to achieve the minimax optimal rate in eq.* (2.54)? This question has been studied in several works in the literature [CMT10; Yan+12; RR08; Bac17b; RR17]. In particular, the most refined results showed that for a target in the RKHS $r \geq 1/2$, it suffices to take $p \geq p_0 = O(n^{q_0})$ with:

  $$q_0 = \frac{\alpha + 2r - 1}{2\alpha r + 1} \tag{2.58}$$

  to achieve the minimax rate $\mathsf{R}_{n,p} = \Theta(n^{-\frac{2\alpha r}{2\alpha r+1}})$. The result in corollary 2.3.2 shows that this rate can indeed be achieved with a smaller number of features when $r > 1/2$, since

  $$q_0 - q_\star = \frac{2(1-r)(\alpha - 1)}{2\alpha r + 1} > 0, \qquad \text{for all } \alpha > 1. \tag{2.59}$$

- **Neural scaling laws:** Motivated by the neural scaling laws literature, different recent works have turned to the study of linear models under source and capacity conditions

as a playground to understand the emergence of different bottlenecks in the excess error rates [Bah+24; MRS22].

The model studied in these works is given by ridge regression on data $y_i = \langle \theta_\star, x_i \rangle$ with $x \sim \mathcal{N}(0, \mathrm{diag}((d/k)^\alpha))$ and $\theta_\star \sim \mathcal{N}(0, 1/d I_d)$ with a linear projection model $f(x; a, W_0) = \langle a, W_0 x \rangle$, where W_0 is an i.i.d. Gaussian matrix. Note this model is a particular case of the model discussed here, corresponding to a linear feature map and random target function. Moreover, since the variance of the target is constant, the source is entirely determined by the capacity α of the asymptotic kernel, here controlled by the decay of the covariance of the input data.

The approximation limit from Corollary 2.1.2 and the kernel limit from Corollary 2.1.1 are known in this literature as *Variance* and *Resolution limited regimes*, respectively [Bah+24]. They correspond precisely to the bottlenecks in the excess risk arising from the limited approximation capacity of the random feature model or the limited availability of training data. The rates in the variance limited regime can also be obtained from Theorem 2.3.2, and correspond to particular cases in fig. 2.6. We refer the reader to the Appendix E of [DLM24] for a detailed discussion of the relationship between these two literatures.

2.4 High-dimensional bottlenecks

So far, our results were agnostic to the covariate distribution μ_x, as they only appear indirectly in the definition of the Friedholm operator $\mathbb{T}$ in eq. (2.6). For instance, the covariate dimension d does not appear in the deterministic equivalents of theorem 2.1.1. To under better the limitations of learning with high-dimensional random features, we need to explicitly connect Σ to μ_x.

Consider the Gaussian i.i.d. case $\mu_x = \gamma_d$. The space $L^2(\gamma_d)$ admits the following orthogonal decompositions:

$$L^2(\gamma_d) = \bigoplus_{m \geq 0} V_m \tag{2.60}$$

where V_m are the linear subspaces spanned by the Hermite tensors $H_\alpha(x) = \prod_{j=1}^d h_{\alpha_j}(x_j)$ of degree $m = |\alpha| = \sum_{j=1}^d \alpha_j$, where $(h_m)_{m \geq 0}$ are the normalised probabilist Hermite polynomials. The dimension of these linear subspaces are given by:

$$\dim(V_m) = \binom{d+m-1}{m}. \tag{2.61}$$

Therefore, in this basis the decomposition of the target function in eq. (2.9) can be equiva-

lently written in this basis as:

$$f_\star(x) = \sum_{m=0}^{\infty} \sum_{\substack{\alpha \in \mathbb{Z}_+^d \\ |\alpha|=m}} c_\alpha(f_\star) H_\alpha(x). \tag{2.62}$$

with the coefficients c_α quantifying how much of the total energy of the target $||f_\star||^2_{\gamma_d} = \sum_\alpha c_\alpha^2$ lies in each subspace V_m. The random features ridge predictor in eq. (2.4) is a linear operator $y \in \mathbb{R}^n \mapsto \hat{a}\lambda(X, y) \in \mathbb{R}^p$ mapping the training data into the column space of Φ, which—assuming Φ is full rank[26]—is a linear subspace of dimension $\mathrm{rank}(\Phi) = \min(n, p)$. A simple power-counting argument then suggests that learning the component of the target in subspace V_m requires $\min(n, p) = O(m)$, with the smaller of n and p acting as the bottleneck for approximating V_m. Consequently, fitting all components up to $m \leq \kappa$ requires:

$$\min(n, p) \simeq \sum_{m=0}^{\kappa} \binom{d+m-1}{m} = O(d^\kappa) \tag{2.63}$$

This intuition was made precise in the following result in [MMM22].

Theorem 2.4.1 ([MMM22], informal)**.** Let $\delta > 0$. With $\min(n, p) = O(d^{\kappa+\delta})$, the random features predictor in eq. (2.4) can learn at best a degree κ approximation of the target function $f_\star$. In other words:

$$\mathbb{E}\left[||f(x; \hat{a}_\lambda, W_0) - f_\star(x)||_2^2\right] = ||P_{>\kappa} f_\star||_{L^2(\gamma_d)} + o_d(1) \tag{2.64}$$

where $P_{>\kappa}$ is the orthogonal projector into the space $\bigoplus_{m>\kappa} V_m$ of polynomials of degree larger than κ, i.e. $||P_{>\kappa} f_\star||_{L^2(\gamma_d)} = \sum_{|\alpha|>\kappa} c_\alpha^2$.

In other words: for high-dimensional isotropic data, RFRR is simply performing a polynomial regression of the target function.

Remark 2.4.1. The key ingredient in the argument above is that the data distribution is isotropic. This implies that the target will have energy uniformly spread over the different frequencies in the orthogonal decomposition. Similarly, since the features are not adapted to the structure of the data, to fit a given frequency it requires spanning the full space to be fitted.

2.4.1 Gaussian equivalence for RFRR

The $\min(n, p) = O(d^{\kappa+\delta})$ in theorem 2.4.1 ensures that the number of samples and parameters are large enough such that the full subspace V_κ is learned. A more quantitative descriptions of

[26] This holds almost surely for Gaussian covariates and full-rank W_0.

the transition can be obtained with a finer characterisation of the feature covariance matrix at the scale $n, d = \Theta(d^\kappa)$. For simplicity of exposition, consider the case $\kappa = 1$, known as the *proportional regime*.

Conditionally on W_0, the pre-activations $z_k = \langle w_{0,k}, x_i \rangle$ are Gaussian random variables with zero mean and covariance $1/d \langle w_{k,0}, w_{k',0} \rangle$. Therefore, it is also natural to expand the feature matrix $\Phi_{ik} = \sigma(\langle w_{0,k}, x_i \rangle)$ in the basis of Hermite polynomials:

$$\Phi_{ik} = \sum_{m=0}^{\infty} \mu_m h_m(\langle w_{0,k}, x_i \rangle) \tag{2.65}$$

Therefore, the (finite p) feature covariance matrix $\hat{\Sigma} \in \mathbb{R}^{p \times p}$ is given by:

$$\hat{\Sigma}_{kk'} = \mathbb{E}_{x\sim}[\sigma(\langle w_{0,k}, x \rangle)\sigma(\langle w_{0,k'}, x \rangle)] = \sum_{m=0}^{\infty} \mu_m^2 \left(\frac{\langle w_{0,k}, w_{0,k'} \rangle}{d} \right)^m \tag{2.66}$$

Note that since the neurons are independent, the scalar product above is $\Theta_d(1)$ for $j = k$ and $O(d^{-1/2})$ for $j \neq k$, meaning that the frequency m has weight $\Theta(d^{-m/2})$ in the decomposition of the off-diagonal components. In other words, the high-frequency components of the population covariance decay faster as $d \to \infty$. This implies that in order to compute the excess risk from theorem 2.1.1 in the proportional high-dimensional limit it suffices to keep the leading order terms in this expansion:

$$\hat{\Sigma} \simeq \mu_0^2 1_p 1_p^\top + \mu_1^2 W_0 W_0^\top + \mu_\star^2 I_p \tag{2.67}$$

where $\mu_\star^2 = \sum_{m \geq 2} \mu_m^2$. Higher-order terms, as well as corrections to the diagonal term, are negligible in the limit.[27]

Note that the features covariance in eq. (2.67) is equivalent to the one of a model with Gaussian features:

$$G = \mu_0 1_p + \mu_1 X W_0^\top + \mu_\star Z \in \mathbb{R}^{n \times p} \tag{2.68}$$

where Z is a Gaussian matrix with i.i.d. Gaussian entries $\mathcal{N}(0, 1)$. This linearisation of the features covariance is known as *Gaussian equivalence* [Gol+22].

2.5 Beyond ridge: convex ERM

The discussion thus far has focused on the square loss and quadratic penalty. This choice considerably simplifies the analysis, since the empirical risk minimisation problem in eq. (2.1)

[27] This approximation should be understood in operator norm, since what matters for the risk in theorem 2.1.1 is the spectrum.

admits a closed-form solution, effectively reducing it to the study of random matrices.

In this section, we extend these ideas to the setting of convex ERM, i.e. eq. (2.1) with convex loss ℓ and regulariser r. This includes several important cases, such as the logistic / cross-entropy loss $\ell(y,z) = \log(1+e^{-yz})$, the hinge loss $\ell(y,z) = \max(0, 1-yz)$, and ℓ_q penalties $r(a) = ||a||_q$ for $q \geq 1$.

As discussed in section 2.1.2, the central technical tool in the ridge analysis was the *Gaussian universality* of the features — the idea that, for the purpose of analysing excess risk, the non-linear features can be asymptotically replaced by correlated Gaussian variables. In the case of RFRR, this can be viewed as an instance of universality in random matrix theory [TV11]. The starting point for extending the analysis to the non-quadratic case is to adapt the idea of Gaussian equivalence.

From now on, we restrict the discussion to the proportional high-dimensional regime, where $d \to \infty$ with $n, p = \Theta(d)$.[28]

2.5.1 Gaussian equivalence

As a starting point, let's formalise what we mean by Gaussian equivalence in the context of more general empirical risk minimisation problems. Consider the empirical risk minimiser

$$\hat{a}_\lambda(\Phi, y) = \min_{a \in \mathbb{R}^p} \sum_{i=1}^{n} \ell\left(y_i, \langle a, \varphi(x_i)\rangle\right) + \lambda r(a). \tag{2.69}$$

under a general convex loss function ℓ and penalty r, where $\varphi(x) = \sigma(W_0 x_i)$. For simplicity, assume that the labels were drawn from a Gaussian single-index model (definition 1.3.2):

$$y_i = g(\langle \theta_\star, x_i\rangle), \qquad x_i \sim \mathcal{N}(0, {}^1\!/_d I_d), \quad i.i.d. \tag{2.70}$$

Our goal is to compute the population risk

$$R(\hat{a}_\lambda) = \mathbb{E}_{x\sim\gamma_d}\left[\ell\left(g(\langle\theta_\star, x\rangle), \langle \hat{a}_\lambda, \varphi(x)\rangle\right)\right] \tag{2.71}$$

for the minimiser $\hat{a}_\lambda$ in the high-dimensional asymptotic limit. The challenge, as before, is that the features $\varphi(x) = \sigma(W_0 x)$ are a non-linear functions of the covariates. Inspired by our discussion from section 2.4.1, consider the *Gaussian equivalent model* with linearised features:

$$g_i = \mu_0 1_p + \mu_1 W_0 x_i + \mu_\star z_i \tag{2.72}$$

[28] Extending the ideas that follow to polynomial regimes is an open research question.

where $z_i \sim \mathcal{N}(0, I_p)$ independently from x_i and $\mu_0, \mu_1, \mu_\star$ are related to the Hermite decomposition of the non-linearity $\sigma(t)$:

$$\mu_0 = \mathbb{E}[\sigma(\xi)], \qquad \mu_1 = \mathbb{E}[\sigma(\xi)\xi], \qquad \mu_\star = \sqrt{\mathbb{E}[\sigma(\xi)^2] - \mu_0^2 - \mu_1^2} \tag{2.73}$$

for $\xi \sim \mathcal{N}(0, 1)$. In other words, we have that, to leading order the equivalent Gaussian model has the same second-order statistics than the full non-linear model: $g_i \sim \mathcal{N}(0, \hat{\Sigma})$ with $\hat{\Sigma} = \mathbb{E}[\varphi(x)\varphi(x)^\top]$. Then, Gaussian equivalence states that in the high-dimensional limit, the risks of these two models are exactly the same.

Theorem 2.5.1 (Gaussian equivalence for RFRR, informal)**.** In the proportional high-dimensional limit $d \to \infty$ with $n, p = \Theta(d)$, the asymptotic population and empirical risks of these two problems coincide:

$$\begin{aligned} |R(\hat{a}_\lambda(\Phi, y)) - R(\hat{a}_\lambda(G, y))| &\xrightarrow{\mathbb{P}} 0 \\ |\hat{R}_n(\hat{a}_\lambda(\Phi, y)) - \hat{R}_n(\hat{a}_\lambda(G, y))| &\xrightarrow{\mathbb{P}} 0 \end{aligned} \tag{2.74}$$

This result was first conjectured in [Ger+20], where the equivalent model was analysed using the replica method from statistical physics to characterise the limiting risk, and later proved in [HL22; MS22].

The idea behind the proof of theorem 2.5.1 is that the risk in eq. (2.71) depends on the covariates only through the joint distribution of the pre-activations $z = \langle \theta_\star, x \rangle$ and $s = \langle \hat{a}_\lambda, \varphi(x) \rangle$. It therefore suffices to establish their asymptotic joint Gaussianity. A naive approach would be to invoke the central limit theorem for s, since it is expressed as a sum of random variables. The difficulty, however, is that s involves not an arbitrary $a \in \mathbb{R}^p$ but the optimiser $\hat{a}_\lambda(X, y)$ of the empirical risk, making it a correlated rather than independent sum. Handling this correlation is the main challenging in the proof.

The proof in [HL22; MS22] follows this scheme, breaking the argument in two steps. First, one argues that for a fixed predictor $a \in \mathcal{S}_p$ in some constraint set $\mathcal{S}_p \subset \mathbb{R}^p$ (e.g. the set of $||a||_\infty < \infty$), the pre-activation $s = \langle a, \varphi(x) \rangle$ satisfies a central limit theorem type of result:

$$\lim_{d \to \infty} \sup_{a \in \mathbb{R}^p} \Big| \mathbb{E}\left[h\left(\langle a, \varphi(x) \rangle\right)\right] - \mathbb{E}\left[h\left(\langle a, g \rangle\right)\right] \Big| = 0 \tag{2.75}$$

for any bounded Lipshchitz function $h : \mathbb{R} \to \mathbb{R}$, where g is the Gaussian equivalent. This type of pointwise CLT was first established for random features in [Gol+22; HL22].

The second step is to show that the CLT implies the universality of the risk in eq. (2.74). In [HL22], this was achieved using a Lindeberg interpolation argument, in which each feature vector $\varphi(x_i)$ in the training set is progressively replaced by a Gaussian vector g_i. This

substitution allows the CLT to be applied to the modified predictor. The cumulative error introduced in both the predictor and the loss can then be controlled, and shown to remain negligible asymptotically, even after all features in the training set have been swapped.

The strength of this two-step proof scheme is that it allows to establish universality of the risk independently of the particular feature map, as proposed in [MS22]. In other words, to prove the universality of the risk under the general assumption that the features satisfy a point-wise CLT. This allows to reduce the proof of universality of the empirical risk minimiser to proving the point-wise CLT for the features of interest, independently of the optimisation problem.

2.5.2 Gaussian covariate model

Under Gaussian equivalence, investigating the risk of the ERM problem in eq. (2.1) is equivalent to studying a model with correlated Gaussian features. This motivate us to introduce the following *Gaussian covariate model* [Lou+21a]:

$$\hat{a}_\lambda = \arg\min_{a\in\mathbb{R}^p} \sum_{i=1}^{n} \ell(g(\langle\theta_\star, u_i\rangle), \langle a, v_i\rangle) + \lambda r(a) \tag{2.76}$$

where the pairs $u_i \in \mathbb{R}^d$ and $v \in \mathbb{R}^p$ are jointly Gaussian vectors:

$$(u_i, v_i) \sim \mathcal{N}\left(\begin{bmatrix}0_d\\0_p\end{bmatrix}, \begin{bmatrix}\Psi & \Phi\\ \Phi^\top & \Omega\end{bmatrix}\right), \qquad i.i.d. \tag{2.77}$$

with $\Psi \in \mathbb{R}^{d\times d}$ and $\Omega \in \mathbb{R}^{p\times p}$ positive-definite symmetric matrices and $\Phi \in \mathbb{R}^{p\times d}$. The general covariance structure is motivated by the fact that both the data and model features could be drawn from different feature maps $u = \varphi_\star(x)$ and $v = \varphi(x)$, e.g. two random feature maps with different widths.

Remark 2.5.1 (Equivalent model)**.** By Gaussian conditioning, we can always rewrite (u, v) in terms of independent Gaussian vectors $z \sim \mathcal{N}(0, I_d)$ and $x \sim \mathcal{N}(0, I_p)$:

$$u = \Phi^\top\Omega^{-1/2}x + (\Psi - \Phi^\top\Omega^{-1}\Phi)^{1/2}z \tag{2.78}$$

$$v = \Omega^{1/2}x. \tag{2.79}$$

This give us a decomposition of the data features u in terms of a piece which is observed by the statistician (proportional to x) and a part which is unobserved (proportional to z), playing the role of an effective structured noise. Since the label distribution only depends on the pre-activation, we can rewrite: $\langle\theta_\star, u\rangle = \langle\beta_\star, v\rangle + \xi$ where $\beta_\star = \Omega^{-1}\Phi\theta_\star$ is the signal component and $\xi = \langle\theta_\star, (\Psi - \Phi^\top\Omega^{-1}\Phi)^{1/2}z\rangle$ is the effective noise induced by model

mispecification. This decomposition can be useful in some contexts, e.g. in the study of the best achievable (Bayes-optimal) error in this model [Cla+23b].

Since the features are Gaussian, one can leverage techniques from high-dimensional probability to derive a sharp characterisation of the risk in the high-dimensional limit.

Theorem 2.5.2 ([Lou+21a], informal)**.** In the proportional high-dimensional limit where $d \to \infty$ with $n/p \to \alpha > 0$ and $p/d \to \gamma > 0$,

$$R(\hat{a}_\lambda) \xrightarrow{\mathbb{P}} \mathsf{R}(\alpha, \gamma, \lambda) = \mathbb{E}[\ell(g(z), s)] \tag{2.80}$$

where $(z, s) \in \mathbb{R}^2$ are jointly Gaussian variables:

$$(z, s) \sim \mathcal{N}\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \rho & m_\star \\ m_\star & q_\star \end{bmatrix} \right) \tag{2.81}$$

with $\rho = \lim_{d\to\infty} \langle \theta_\star, \Psi\theta_\star \rangle \in \mathbb{R}_+$ and $(m_\star, q_\star) \in \mathbb{R}_+^2$ the unique solutions of the following system of self-consistent equations:

$$\begin{cases} \hat{v} &= -\alpha \mathbb{E}_\eta \int \mathrm{d}y \; Z_\star\left(y, \frac{m}{\sqrt{q}}\eta, \rho - \frac{m^2}{q}\right) \partial_\omega f_y(y, \sqrt{q}\eta, v) \\ \hat{q} &= \alpha \mathbb{E}_\eta \int \mathrm{d}y \; Z_\star\left(y, \frac{m}{\sqrt{q}}\eta, \rho - \frac{m^2}{q}\right) f_y(y, \sqrt{q}\eta, v)^2 \\ \hat{m} &= \alpha \mathbb{E}_\eta \int \mathrm{d}y \; Z_\star\left(y, \frac{m}{\sqrt{q}}\eta, \rho - \frac{m^2}{q}\right) f_y(y, \sqrt{q}\eta, v) \end{cases} \tag{2.82}$$

$$\begin{cases} v &= \mathbb{E}_{\xi,\theta_\star}\left[\nabla_b \cdot f_a\left(\hat{m}\Omega^{-1/2}\Phi\theta_\star + (\hat{q}\Omega)^{1/2}\xi, \hat{v}\Omega\right)\right] \\ q &= \mathbb{E}_{\xi,\theta_\star}\left[||f_a\left(\hat{m}\Omega^{-1/2}\Phi\theta_\star + (\hat{q}\Omega)^{1/2}\xi, \hat{v}\Omega\right)||_2^2\right] \\ m &= \mathbb{E}_{\xi,\theta_\star}\left[\langle f_a\left(\hat{m}\Omega^{-1/2}\Phi\theta_\star + (\hat{q}\Omega)^{1/2}\xi, \hat{v}\Omega\right), \theta_\star\rangle\right] \end{cases} \tag{2.83}$$

In these equations, $\eta \sim \mathcal{N}(0, 1)$, $\xi \sim \mathcal{N}(0, I_p)$ independently, and we have defined the short-hand:

$$f_y(y, \omega, v) = \frac{1}{v}\left(\omega - \mathrm{prox}_{v\ell(y,\cdot)}(\omega)\right) \tag{2.84}$$

$$f_a(b, A) = \mathrm{prox}_{r\left(A^{-1/2}\cdot\right)}(A^{-1/2}b) \tag{2.85}$$

$$Z_\star(y, \omega, v) = \mathbb{E}_{z\sim\mathcal{N}(\omega,V)}\left[\mathsf{P}_y(y|z)\right] \tag{2.86}$$

where $\mathrm{prox}_{\tau f}(x) = \arg\min_{z\in\mathbb{R}^p}\left\{\frac{1}{2\tau}||z - x||_2^2 + f(z)\right\}$ is the proximal operator.

Remark 2.5.2 (Intuition)**.** Despite cumbersome, this result simply states that in the high-dimensional limit the joint statistics of the pre-activations $z = \langle \theta_\star, u \rangle$ and $s = \langle \hat{a}_\lambda, v \rangle$ can be fully characterised by a set of scalar equations. Given a specific choice of $\ell, r, \theta_\star, \mathsf{P}_y$ (or g) and Ψ, Ω, Φ, these can be efficiently integrated numerically. Since the ERM problem for the

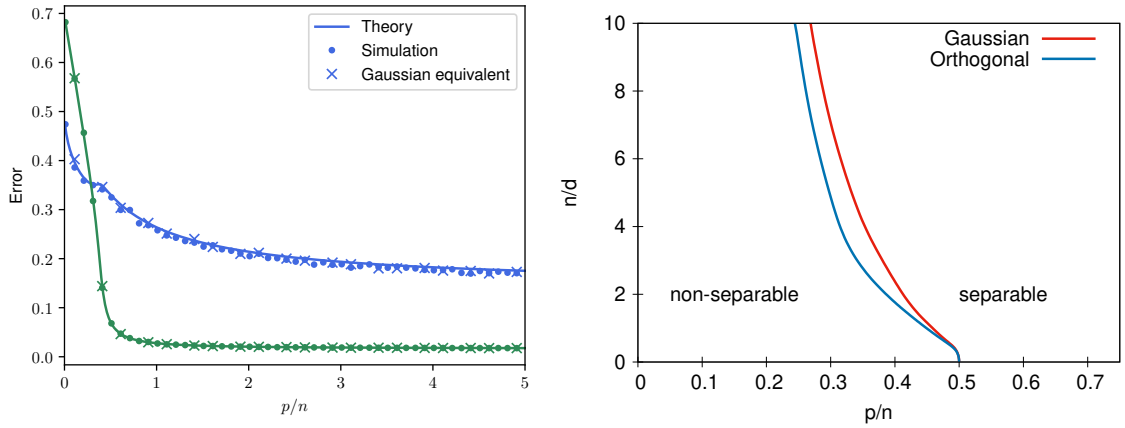

Figure 2.7: (**Left**) Population (blue) and empirical (green) misclassification risk for logistic regression as a function of the normalised width p/n for logistic regression $\ell(y,z) = \log(1+e^{-yz})$ with random features with $\sigma = \text{sign}$ activation at fixed sample complexity $n/d = 3$ and vanishing ridge penalty $\lambda \to 0^+$. Solid lines denote theoretical results, as derived from theorem 2.5.2, dots denote simulations of the full problem, while crosses denote simulations of the Gaussian equivalent model, with $d = 200$. (**Right**) Interpolation threshold in the sample complexity n/d vs. normalised width p/n plane, for two choices of random features weight ensembles, Gaussian (red line) and orthogonal (blue line). Figures from [Ger+20]

Gaussian covariate model only depends on the high-dimensional covariates (u, v) through the pre-activations (z, s), this suffices to characterise the asymptotic risk.

Corollary 2.5.1 (ℓ_2 penalty)**.** Consider the ℓ_2 penalty $r(a) = ||a||_2^2$. Then, the last three equations of theorem 2.5.2 considerably simplify:

$$\begin{cases} v &= \text{Tr}\left\{\Omega(\lambda I_p + \hat{v}\Omega)^2\right\} \\ q &= \text{Tr}\left\{\left[\hat{q}\Omega + \hat{m}^2\Phi\theta_\star\theta_\star^\top\Phi^\top\right]\Omega\left(\lambda I_p + \hat{v}\Omega\right)^{-2}\right\} \\ m &= \hat{m}\langle\Phi\theta_\star, (\lambda I_p + \hat{v}\Omega)^{-1}\Phi\theta_\star\rangle. \end{cases} \tag{2.87}$$

In particular, letting $\Omega = \sum_{j=1}^p \xi_j^2 v_j v_j^\top$, this depends only on the asymptotic joint statistics of ξ_j^2 and $\langle v_j, \Phi\theta_\star\rangle$.

A similar statement to theorem 2.5.2 holds for the asymptotic empirical risk, but we refer the interested reader to [Lou+21a] for the details.

The asymptotic characterisation in theorem 2.5.2 was proven in [Lou+21a] using Gordon's Gaussian min-max inequalities (CGMT) [Gor85; Sto13; OTH13], and can alternatively be derived using the replica method from statistical physics. It encompasses several settings of interest in the literature, such as logistic regression, M-estimators and the LASSO.

As an illustration of how theorem 2.5.2 can be combined with Gaussian equivalence, fig. 2.7 (left) displays the empirical and population misclassification error of logistic regression $\ell(y,z) = \log(1+e^{-zy})$ with a vanishing ridge penalty $\lambda \to 0^+$, $r(a) = ||a||_2^2$. The curves are

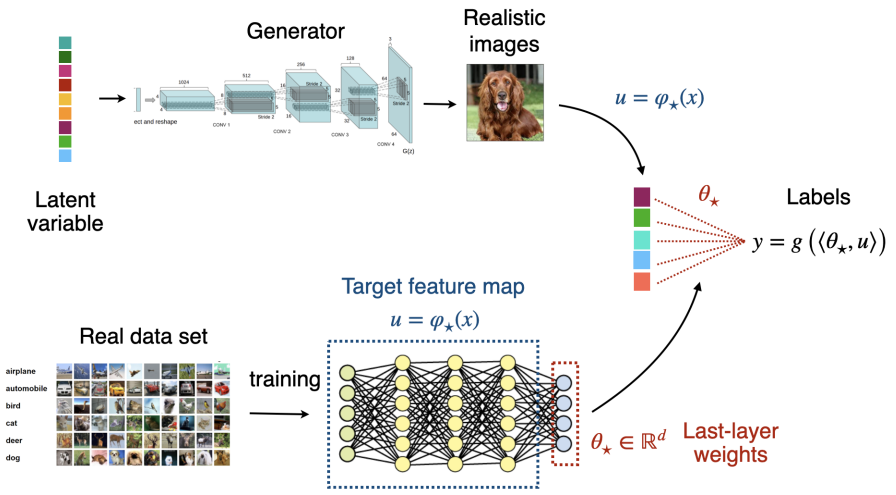

Figure 2.8: Sketch of a possible pipeline to generate realistic synthetic data. Here, a generative model is trained on a real data set, which can be used to generate realistic images from Gaussian data (upper part of the diagram). To generate labels, a second network (lower-part) is trained on the same data set to discriminate different classes. This defines a map from the images to the label, which can be used to generate fake labels for the GAN generated synthetic images. Figure from [Lou+21a].

obtained by solving the self-consistent equations of theorem 2.5.2 for the Gaussian equivalent model with $\Psi = {}^1\!/_d I_d$, $\Phi = \mu_1 W_0$, and $\Omega = \mu_1^2 W_0 W_0^\top + \mu_\star^2 I_p$, where W_0 is a Gaussian i.i.d. matrix. The results show excellent agreement between theoretical predictions and numerical simulations, both for the full non-linear model ($\bullet$) and for its Gaussian equivalent ($\times$). In this setting, the interpolator corresponds to the maximum-margin solution [RZH03], which is also the limit reached by gradient descent on the unregularised problem [Sou+18]. The interpolation peak thus does not occur at $p = n$, but instead at the width where the network becomes able to linearly separate the data. This separability threshold can be computed from the theory and is shown in fig. 2.7 (right) for two choices of W_0 ensemble: Gaussian and orthogonal. Interestingly, although both ensembles yield the same asymptotic kernel, at finite width and sample size orthogonal random features consistently outperform Gaussian ones — a fact well documented in practice [CRW17].

2.6 Towards realistic data

Motivated by the discussion in section 2.5.1, one can ask to what extent Gaussian universality holds beyond random feature maps on Gaussian data. Theorem 2.5.2 allow us to numerically investigate this question in two scenarios.

Realistic synthetic data — The first scenario is the one of realistic feature maps on synthetic data. In this case, we assume that the labels are drawn from a generative model on

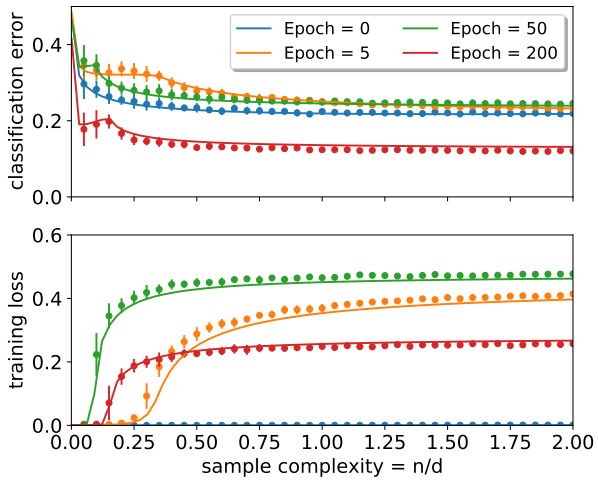

Figure 2.9: Misclassification population risk (top) and empirical training loss (bottom) as a function of the sample complexity $\alpha = n/p$ for logistic regression on a learned feature map trained on dcGAN-generated CIFAR10-like images labelled by a trained fully-connected neural network (see Appendix in [Lou+21a] for the details on the architecture), 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). Figure from [Lou+21a].

random Gaussian data:

$$y_i = f_\star(z_i; \Theta_\star) = g\left(\langle a_\star, \varphi_\star(x_i; W_\star)\rangle\right), \qquad x_i \sim \mathcal{N}(0, {}^1\!/_d I_d) \tag{2.88}$$

Here, $f_\star(x_i; \Theta_\star)$ may denote any parametric generative model, but for concreteness one can think of a deep neural network generative model, such as a Generative Adversarial Network (GAN) [Goo+14]. In the second equality, the model is further decomposed into the last-layer weights $a_\star$, the network features $\varphi_\star$ with trained weights $W_\star$, and a decoding map $g : \mathbb{R} \to \mathcal{Y}$. Therefore, the map $\varphi_\star : \mathbb{R}^D \to \mathbb{R}^d$ can be thought as mapping a Gaussian latent variable to a realistic image, see fig. 2.8 for a possible pipeline for this process.

This process can be used to generate a realistic dataset $\mathcal{D} = (u_i, y_i) \in \mathbb{R}^d \times \mathcal{Y} : i \in [n]$ with $u_i = \varphi_\star(x_i)$, which in turn serves as input for an ERM problem of interest. Since arbitrarily large datasets can be generated, the population covariances Ψ, Ω, Φ appearing in theorem 2.5.2 can be estimated to any desired precision, making it possible to directly compare the theoretical predictions with finite-size numerical experiments based on this pipeline. Note that there is no reason for Gaussian universality to hold in this setting. Nevertheless, this pipeline was extensively studied in [Gol+22; Lou+21a], showing that theorem 2.5.2 can surprisingly capture the behaviour of the learning curves produced by this procedure.

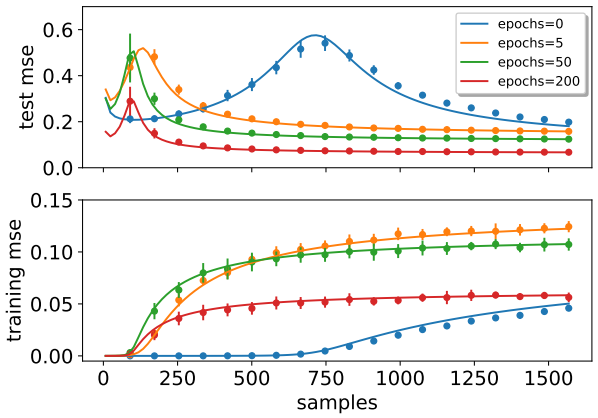

Figure 2.10: Population (top) and empirical (bottom) risks as a function of the number of samples n for ridge regression on the Fashion-MNIST data set, with vanishing regularisation $\lambda \to 0^+$. In this plot, the model feature map φ is a 3-layer fully-connected neural network with $p = 2352$ hidden 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. Figure from [Lou+21a].

Real data — The challenge of applying theorem 2.5.2 to real data is that the equations depend on both the target weights $\theta_\star$ and the target-model feature covariance Φ — which are not accessible for a real data set. However, in the particular case of ridge regression, these quantities only appear in the equations of theorem 2.5.2 in the particular combination $\Phi\theta_\star$, see corollary 2.5.1. This implies the following corollary.

Corollary 2.6.1 (Universality of linear target)**.** Consider ridge regression on the the Gaussian covariate model $\ell(y, z) = (y - z)^2$ and $r(a) = ||a||_2^2$ with linear target function $g(z) = z$. Then, the asymptotic performance given by theorem 2.5.2 is the same for any target feature u_i and target weights $\theta_\star$ that exactly interpolates the data $y_i = \langle \theta_\star, u_i \rangle$, $\forall i \in [n]$.

Although this result might seen surprising at first sight, it is quite intuitive. Indeed, the information about the target model only enters the Gaussian covariate model eq. (2.76) through the statistics of $\langle \theta_\star, u \rangle$. For a linear target $g(z) = z$, this is precisely given by the labels. Under this assumption, this result allow us to estimate empirically the relevant quantities from the data. For that, let $\mathcal{D} = \{(x_i, y_i) \in \mathbb{R}^{d+1} : i \in [n_{\text{tot}}]\}$ denote some real data, e.g. MNIST or CIFAR10, which we assume has been centred. Then, given a feature map of interest $\varphi : \mathbb{R}^d \to \mathbb{R}^p$ and denoting $v_i = \varphi(x_i)$, we can empirically estimate

$$\Omega = \sum_{i=1}^{n_{\text{tot}}} \frac{v_i v_i^\top}{n_{\text{tot}}}, \qquad \rho = \frac{1}{n_{\text{tot}}} \sum_{i=1}^{n_{\text{tot}}} (y_i)^2 , \qquad \Phi\theta_\star = \frac{1}{n_{\text{tot}}} \sum_{i=1}^{n_{\text{tot}}} y_i v_i . \tag{2.89}$$

which allow us to evaluate the equations in theorem 2.5.2. This was extensively studied in [Lou+21a] for different data sets. As an illustration of these results, consider a trained a 3-layer fully connected neural network with ReLU activations on the full Fashion-MNIST

dataset to distinguish clothing worn above vs. below the waist [XRV17]. The model feature map $\varphi : \mathbb{R}^{784} \to \mathbb{R}^p$ is obtained by removing the last layer. Figure 2.10 reports the test and training errors of the ridge estimator on a subsampled batch $n < n_{\text{tot}}$ of Fashion-MNIST images. The learning curve obtained from simulations shows remarkable agreement with the theoretical prediction given by the corresponding Gaussian covariate model.

Recall from section 2.2 that for the square loss and $\lambda \to 0^+$, the interpolation peak is located at the point where the linear system becomes invertible. Interestingly, fig. 2.10 shows that the fully connected network progressively learns a low-rank representation of the data during training. This can be directly verified by examining the spectrum of Ω, whose number of zero eigenvalues increases over time: from a full-rank matrix at initialization to rank 380 after 200 training epochs.

2.6.1 Limitations of Gaussian equivalence

The empirical results discussed in section 2.6 naturally raise the question of how far Gaussian universality extends in high-dimensional asymptotics. A difficulty in investigating this is that there are relatively few settings where exact high-dimensional limits for the risk can be computed outside the Gaussian case. An exception is provided by *Gaussian mixture models* (GMM):

$$\mu_x = \sum_{c \in \mathcal{C}} p_c \mathcal{N}(\mu_c, \Sigma_c) \tag{2.90}$$

with mixture weights $p_c \in [0, 1]$ satisfying $\sum_{c \in \mathcal{C}} p_c = 1$. The proportional asymptotic limit of the risk for linear classifiers trained on GMM data was established in [Lou+21b], in a result analogous to theorem 2.5.2 for GMM covariates. This provides a useful testbed for universality, as it permits a direct comparison of the two asymptotic formulas under the same loss and regularisation. This yields a well-defined mathematical question: *under what conditions do the asymptotic risks predicted by these two formulas agree*?

This question was investigated in [Ger+24; Pes+23]. The first work considered the case of *random labels*, where the problem depends only on the geometry of the covariates. In this setting, the asymptotic training loss of linear classifiers admits a closed-form characterisation that coincides exactly with the Gaussian covariate model for a broad class of input distributions, including arbitrary mixtures of Gaussian components. In the interpolation limit $\lambda \to 0^+$, universality becomes even stronger: the asymptotic loss is independent of the data covariance and reduces to the prediction of the isotropic Gaussian model.

When the task is structured and the labels are correlated with the inputs, universality is more fragile. This case was studied in [Pes+23] in a model where labels are generated by a single-index function of the covariates, $y = g(\langle \theta_\star, x \rangle)$. Here, the conditions under which Gaussian universality holds are restrictive: in particular, if the mixture is homoscedastic and

the target weights $\theta_\star$ are uncorrelated with the cluster means μ_c and covariances Σ_c, the asymptotic training and generalisation errors coincide with those of the Gaussian covariate model with matching second order statistics. Conversely, universality breaks down when the task correlates with the mixture structure or under strong heteroscedasticity, in which case the limiting risk departs systematically from the Gaussian prediction. We refer the reader interested in the details to [Ger+24; Pes+23].

2.7 Going deeper

In Section 2.6 we have discussed empirical evidence showing that Gaussian universality holds beyond random features maps. In these cases, the features statistics were computed numerically. In this section, we discuss another example where we can exactly characterise the Gaussian equivalent model, which is the extension of the random features model to the multi-layer case:

$$\varphi\left(x; \{W_\ell\}_{\ell\in[L]}\right) = \sigma_L\left(W_L\sigma_{L-1}\left(\cdots W_2\sigma_1\left(W_1 x\right)\right)\right). \tag{2.91}$$

where $\sigma_1, \cdots, \sigma_L$ are non-linear scalar activation functions acting entry-wise and $W_1 \in \mathbb{R}^{p_1\times d}$ and $W_\ell \in \mathbb{R}^{p_\ell\times p_{\ell-1}}$ for $\ell \in [L]$ are deterministic weight matrices.

Consider two such deep feature maps $\varphi\left(x; \{W_\ell\}_{\ell\in[L]}\right)$ and $\varphi_\star\left(x; \{V_\ell\}_{\ell\in[L_\star]}\right)$, with potentially different weights and activation functions $\sigma_\ell, \sigma_\ell^\star$, and without loss of generality $L \le L_\star$. Then, one can derive a similar linearisation the one discussed in section 2.4.1 layer-wise, under the following simplifying assumptions.

Assumption 2.7.1. Consider the following simplifying assumptions.

- Gaussian covariates: $x \sim \mathcal{N}(0, \Omega_0)$ with $||\Omega_0||_{\text{op}} < \infty$.
- The activation functions $\sigma_\ell, \sigma_\ell^\star$ are Lipschitz functions and such that the feature map is centred $\mathbb{E}\left[\varphi(x)\right] = 0$ and $\mathbb{E}\left[\varphi_\star(x)\right] = 0$.
- All the internal widths p_ℓ of W_ℓ, V_ℓ are equal.
- The rows w_ℓ, v_ℓ of W_ℓ, V_ℓ are i.i.d. sub-Gaussian random vectors with mean zero and covariances

$$C_\ell := p_\ell\mathbb{E}\left[w_\ell w_\ell^\top\right], \ \tilde{C}_\ell = p_\ell\mathbb{E}\left[v_\ell v_\ell^\top\right], \ \check{C}_\ell := p_\ell\mathbb{E}\left[w_\ell v_\ell^\top\right],$$

  with $||C_\ell||_{\text{op}} + ||\tilde{C}_\ell||_{\text{op}} + ||\check{C}_\ell||_{\text{op}} \lesssim 1$. Moreover, they are asymptotically orthogonal: let $w, w' \in \mathbb{R}^{\ell-1}$ be two independent copies of a row of W_ℓ. Then, $\langle w, w'\rangle = O(d^{-1/2})$ (similarly for V_ℓ).

Conjecture 2.7.1 ([Sch+23; Sch+24b], Gaussian equivalence for deep RFs)**.** In the proportional high-dimensional limit where $p \to \infty$ at fixed $p_\ell = \Theta_d(d)$ for all $\ell \in [L]$, the features populations covariances Ω, Φ, Ψ:

$$\Psi = \mathbb{E}[\varphi_\star(x)\varphi_\star(x)^\top], \qquad \Phi = \mathbb{E}[\varphi(x)\varphi_\star(x)^\top], \qquad \Omega = \mathbb{E}[\varphi(x)\varphi(x)^\top] \tag{2.92}$$

can be asymptotically approximated by the linearised covariances:

$$||\Omega - \Omega_{\text{lin}}||_{\text{F}} + ||\Psi - \Psi_{\text{lin}}||_{\text{F}} + ||\Phi - \Phi_{\text{lin}}||_{\text{F}} \lesssim 1 \tag{2.93}$$

where $\Psi_{\text{lin}}, \Phi_{\text{lin}}, \Omega_{\text{lin}}$ are defined as the last iterate $\ell = L$ of the following recursion:

$$\begin{aligned}
\Omega_\ell^{\text{lin}} &= (\kappa_\ell^1)^2 W_\ell \Omega_{\ell-1}^{\text{lin}} W_\ell^\top + (\kappa_\ell^*)^2 I_{p_\ell} \\
\Psi_\ell^{\text{lin}} &= (\tilde{\kappa}_\ell^1)^2 V_\ell \Psi_{\ell-1}^{\text{lin}} V_\ell^\top + (\tilde{\kappa}_\ell^*)^2 I_{p_\ell} \\
\Phi_\ell^{\text{lin}} &= \kappa_\ell^1 \tilde{\kappa}_\ell^1 W_\ell \Phi_{\ell-1}^{\text{lin}} V_\ell^\top + (\breve{\kappa}_\ell^*)^2 I_{p_\ell},
\end{aligned} \tag{2.94}$$

with $\Omega_0^{\text{lin}} = \Psi_0^{\text{lin}} = \Phi_0^{\text{lin}} = \Omega_0$ the input covariance. The coefficients $\{\kappa_\ell^1, \tilde{\kappa}_\ell^1, \kappa_\ell^*, \tilde{\kappa}_\ell^*, \breve{\kappa}_\ell^*\}$ are defined by the recursion

$$\kappa_\ell^1 := \mathbb{E}\left[\sigma_\ell'(N_\ell)\right], \quad \tilde{\kappa}_\ell^1 := \mathbb{E}\left[\tilde{\sigma}_\ell'(\tilde{N}_\ell)\right] \tag{2.95}$$

and

$$\begin{aligned}
\kappa_\ell^* &= \sqrt{\mathbb{E}[\sigma_\ell(N_\ell)^2] - r_\ell (\kappa_\ell^1)^2} \\
\tilde{\kappa}_\ell^* &= \sqrt{\mathbb{E}[\sigma_\ell^\star(\tilde{N}_\ell)^2] - \tilde{r}_\ell (\tilde{\kappa}_\ell^1)^2} \\
\breve{\kappa}_\ell^* &= \sqrt{\mathbb{E}[\sigma_\ell(N_\ell)\sigma_\ell^\star(\tilde{N}_\ell)] - \breve{r}_\ell \kappa_\ell^1 \tilde{\kappa}_\ell^1},
\end{aligned} \tag{2.96}$$

where $N_\ell, \tilde{N}_\ell$ are jointly mean-zero Gaussian with $\mathbb{E}\left[N_\ell^2\right] = r_\ell$, $\mathbb{E}\left[\tilde{N}_\ell^2\right] = \tilde{r}_\ell$, $\mathbb{E}\left[N_\ell \tilde{N}_\ell\right] = \breve{r}_\ell$, with

$$r_\ell = \text{Tr}\left[C_\ell \Omega_{\ell-1}^{\text{lin}}\right], \ \tilde{r}_\ell = \text{Tr}\left[\tilde{C}_\ell \Psi_{\ell-1}^{\text{lin}}\right], \ \breve{r}_\ell = \text{Tr}\left[\breve{C}_\ell^\top \Phi_{\ell-1}^{\text{lin}}\right].$$

Finally, for $\tilde{L} \geq \ell \geq L+1$, define

$$\Phi_\ell^{\text{lin}} = \tilde{\kappa}_\ell^1 \Phi_{\ell-1}^{\text{lin}} V_\ell^\top, \tag{2.97}$$

with still $\tilde{\kappa}_\ell^1, \tilde{\kappa}_\ell^*$ just as before, and Ψ_ℓ^{lin} with the same recursion (2.94).

This conjecture generalises the prior results on Gaussian universality for ridge regression discussed in Section 2.4.1. Although it is challenging to prove it in full generality, it can be shown in the particular case of three-layer random features.

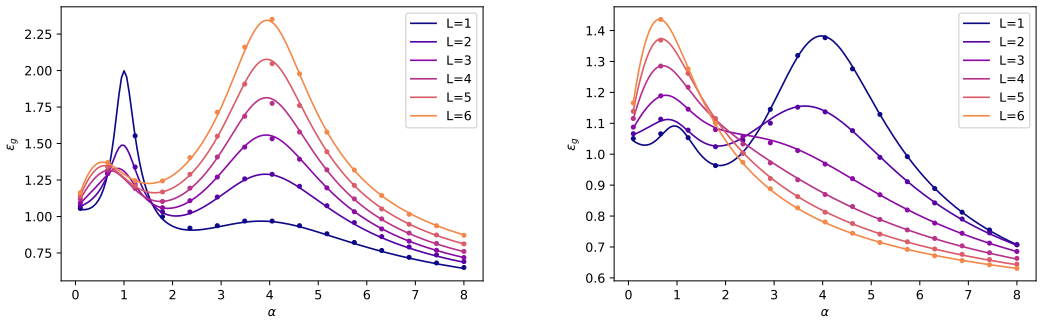

Figure 2.11: Risk for ridge regression on a 1-hidden layer target function ($p_\star/d = 2$, $\sigma_1^\star = \mathrm{sign}$) using a $L-$hidden layers model with widths $p_\ell = 4d$ and $\sigma_\ell = \tanh$ activation **(left)** or $\sigma_\ell(x) = 1.1 \times \mathrm{sign}(x) \times \min(2, |x|)$ clipped linear activation **(right)**, for depths $1 \leq L \leq 6$. The regularization is $\lambda = 0.001$. Solid lines represent theoretical curves evaluated from theorem 2.5.2 and conjecture 2.7.1, while numerical simulations, averaged over 50 runs, are indicated by dots. Despite sharing the same architecture, the use of different activations induces different implicit regularizations. Figure from [Sch+23].

Theorem 2.7.1 ([Sch+24b])**.** Conjecture 2.7.1 holds under Assumption 2.7.1 for $L = L_\star = 2$.

Combined with theorem 2.5.2, the linearisation in 2.7.1 can be used to study the asymptotic performance of models where both the target and hypothesis were drawn from a deep random features model with structured weights. Recently, it has been empirically observed that resampling the weights of trained neural networks from an ensemble that partially preserves their statistics can retain the generalisation performance [Gut+24]. Conjecture 2.7.1 allows to study these *rainbow networks*, as they are known, in the case where only second order statistics on the trained weights are retained. We refer the interested reader to [Sch+24b], where the inductive bias of the Gaussian rainbow network ensemble, as well as the high-dimensional bottlenecks implied by the linearisation of the features were studied. In Figure 2.11, we illustrate the performance of deep random features in the case of i.i.d. Gaussian weights.

3 | Networks away from initialisation

In Chapter 2, we analysed in detail the generalisation properties of two-layer neural networks with fixed features. A key outcome of that discussion was that, due to the lack of adaptivity, the network's performance is insensitive to any structure present in the training data. For instance, in Section 2.4 we showed that, for isotropic random data, random features ridge regression is asymptotically equivalent to polynomial regression when both the width and the sample size are fixed.

Improving on this requires the network to *learn features*, i.e. to adapt its basis to the target task. Understanding this process at the same level of generality as for fixed-feature networks remains challenging. Unlike the second-layer optimisation, the optimisation over the first-layer weights in eq. (1.9) is non-convex. In practice, training is often carried out using gradient-based methods with early stopping, which introduces implicit, algorithm-dependent regularisation.

In this chapter, we discuss this problem in a simplified setting where the first-layer weights are trained only for a few but large steps of gradient descent. Despite its simplicity, this setting captures several key aspects of adaptivity and allows us to make precise the connection between feature learning and the generalisation capacity of the model.

The results discussed in sections 3.2, 3.3 and 3.4.2 are based on [Dan+24a], while the results discussed in section 3.4.3 are based on [Cui+24; Dan+25].

3.1 Learning features, one step at a time

Throughout this chapter, we focus on a regression problem with square loss and ℓ_2 penalty, for which eq. (1.9) reads:

$$\min_{(W,a)\in\mathbb{R}^{p(d+1)}} \frac{1}{n}\sum_{i=1}^{n} \left(y_i - f(x_i; W, a)\right)^2 + \lambda||a||_2^2 \tag{3.1}$$

where we recall the reader we are interested in the class of width-p two-layer neural networks:

$$f(x; W, a) = \frac{1}{\sqrt{p}} \sum_{k=1}^{p} a_k \sigma\left(\langle w_k, x \rangle\right). \tag{3.2}$$

Note that we adopt a different normalisation from eq. (1.8), which, without loss of generality, amounts to a redefinition of the readout weights. We also include weight decay on the second-layer weights, but not on the first-layer.

In what follows, we assume that the data are sampled from a Gaussian multi-index model as defined in definition 1.3.2. As discussed in section 1.3.1, multi-index functions are a rich hypothesis class that captures the inductive bias that high-dimensional tasks often possess an underlying low-dimensional structure. The isotropic Gaussian assumption on the covariates further enables us to quantify precisely the benefits of feature learning, in comparison with the untrained network whose bottlenecks were discussed in section 2.4.

Finally, consider the following step-wise training procedure for the ERM problem in eq. (3.1).

Definition 3.1.1 (Step-wise training)**.** Let $\eta \in \mathbb{R}_+$, $T \in \mathbb{Z}_+$, $(W_0, a_0) \in \mathbb{R}^{p(d+1)}$, $\mathcal{D} = \left\{(x_i, y_i) \in \mathbb{R}^{d+1} : i \in [N]\right\}$ denote the learning rate, training horizon, initial weights and training data, respectively. Let $N = \sum_{t=0}^{T} n_t$ denote a partition of the training samples into $T+1$ batches of n_t samples. Define $\mathcal{A}_{\eta,T,\lambda} : \mathcal{D} \mapsto (\hat{W}, \hat{a}_\lambda) \in \mathbb{R}^{p(d+1)}$ to be the following step-wise training algorithm:

1. Train the first-layer weights for T steps with (one-pass) stochastic gradient descent on the first T batches:

$$w_{t+1,k} = w_t - \eta \nabla_w \frac{1}{n_t} \sum_{i=1}^{n_t} \left(y_i - f(x_i; W_t, a_0)\right)^2, \qquad 0 \le t \le T-1 \tag{3.3}$$

2. Train the second-layer weights on the last batch $n := n_T$ by ridge regression:

$$\hat{a}_\lambda = \arg\min_{a \in \mathbb{R}^p} \frac{1}{n} \sum_{i=1}^{n} \left(y_i - f(x_i; W_T, a)\right)^2 + \lambda ||a||_2^2 \tag{3.4}$$

3. Return $\hat{W} = W_T$ and $\hat{a}_\lambda$.

In particular, our analysis will assume the following balanced initialisation for the network weights.

Assumption 3.1.1 (Balanced initialisation)**.** We assume that the initial weights (W_0, a_0) are drawn i.i.d. as $a_{0,k} \sim \mathrm{Unif}([-1/\sqrt{p}, 1/\sqrt{p}])$ and $w_{0,k} \sim \mathrm{Unif}(\mathbb{S}^{d-1}(\sqrt{d}))$ with

$$a_{0,k} = a_{0,p-k-1}, \qquad w_{0,k} = w_{0,p-k-1}, \quad k \in [p/2]. \tag{3.5}$$

Note this enforces that the initialisation is *balanced* $f(x; W_0, a_0) = 0$ and that for large width p the initial SGD steps eq. (3.3) are in the mean-field regime [COB19].

Remark 3.1.1. A few comments on the setting and assumptions are in place.

- The step-wise training procedure in eq. (3.1) is a common simplification in the theoretical literature, e.g. [DLS22; Ba+22; Bie+22].
- The partition assumption $\mathcal{D} = \bigcup_{t\in[T+1]} \mathcal{D}_t$ in the training data means that each batch of data is only seen by the gradient once. In this algorithm, known as *one-pass SGD*, the stochastic gradients are unbiased estimates of the population ($n_t \to \infty$) gradient.
- Although the exact balanced assumption is useful, it can be generalised to an approximate balanced conditions in the limits we will consider in the following, i.e. $f(x; W_0, a_0) = o_d(1)$ with $a_{0,k} = O(1/\sqrt{p})$ and $||w_k||_2^2 = O(d)$, $\langle w_{0,k}, w_{0,k'}\rangle = \delta_{kk'}$.

3.2 Weak learnability

As discussed in remark 2.1.1, a sign of the lack of adaptivity of the network at initialisation is that the (random) weights of the network are uncorrelated with the underlying structure of the data. Therefore, a crucial question is whether this changes after training, i.e. does $\hat{W} \in \mathbb{R}^{p\times d}$ meaningfully correlate with the indices $W_\star \in \mathbb{R}^{r\times d}$? This notion was precisely defined in definition 1.3.3, and is known as *weak learnability*.

3.2.1 One gradient step

In order to build intuition, we start our discussion of what can be (weakly) learned in a single step $T = 1$. The first step update in eq. (3.3) can be writen as $w_{1,k} = w_{0,k} + \eta g_k$, where $g_k \in \mathbb{R}^d$ is the gradient at initialisation:

$$g_k = -\frac{2a_{0,k}}{\sqrt{p}n_0} \sum_{i=1}^{n_0} y_i \sigma'(\langle w_{0,k}, x_i\rangle) x_i \tag{3.6}$$

where we used that $f(x_i; W_0, a_0) = 0$. For simplicity, consider the case where $y_i = g(\langle w_\star, x_i\rangle)$ is a single-index function. The expected correlation between the gradient and the index can be obtained with an argument almost identical to the one discussed in section 2.4.1:

$$\mathbb{E}[\langle g, w_\star\rangle] = -\frac{a_{0,k}}{\sqrt{p}} \sum_{m=0}^{\infty} c_{m+1}\mu_{m+1} \left(\frac{\langle w_{0,k}, w_\star\rangle}{d}\right)^m \tag{3.7}$$

where, as before $\mu_m = \mathbb{E}[\sigma(z)h_m(z)]$ and $c_m = \mathbb{E}[g(z)h_m(z)]$ are the coefficients of the decomposition of σ, g in the basis of Hermite polynomials, respectively. Since $1/\sqrt{d}\langle w_{0,k}, w_\star\rangle =$

$O(d^{-1/2})$, each term in this sum is of order $O_d(d^{-m/2})$, meaning that for large $d \to \infty$ the sum is dominated by the lowest order term in the sum. Assuming that $\mu_m \neq 0$ for all $m \in \mathbb{Z}_+$, this is given by $\ell_\star = \min\{m \in \mathbb{Z}_+ : \mathbb{E}[g(z)h_m(z) \neq 0]\}$. This quantity, known as the *information exponent*, was first introduced in the study of SGD on single-index models in [AGJ21]. Finally, in order to ensure weak recovery according to definition 1.3.3, we must ensure (a) the gradient term in $w_{1,k} = w_{0,k} - \eta g_k$ is strong enough; (b) concentration of $\langle g_k, w_\star \rangle$. The former is obtained by taking a large learning rate $\eta = \Theta_d(pd^{\frac{\ell_\star - 1}{2}})$ and the latter a batch size $n = \Omega(d^{\ell_\star})$, such that the variance is vanishing in the limit.

This argument can be generalised to a general target function $f_\star$. As in Section 2.4, consider the decomposition of the target in the orthonormal basis of Hermite tensors:

$$f_\star(x) = \sum_{m=0}^{\infty} \langle c_m(f_\star), H_m(x) \rangle \tag{3.8}$$

where $c_m(f_\star) \in \mathrm{Sym}^m(\mathbb{R}^d)$ is a symmetric rank-m tensor in $\mathbb{R}^d$.[29] For the multi-index target $f_\star(x) = g(W_\star x)$, it can be shown that $c_m(f_\star) = d^{-m/2} c_m(f) \cdot (W_\star, \cdots, W_\star)$, where $\cdot$ denote the multilinear multiplication operator — we refer the unfamiliar reader to [Gre12]. This implies that, $c_m(f_\star)$ is at most a rank$-r$ tensor, i.e. its singular vectors (in the sense of (in the sense of [DDV00]) span a subspace of $\mathrm{span}(W_\star)$. Since $c_m(f_\star) = O(d^{-m/2})$, as in the rank-one case the gradient at initialisation is dominated by the lowest order frequency the the Hermite tensor basis.

Theorem 3.2.1 ([Dan+24a], informal)**.** Let $\ell_\star = \min\{m \in \mathbb{Z}_+ : \mathbb{E}[g(z)H_m(z)] \neq 0\}$ denote the *leap exponent* of the target link function, where H_m are the Hermite tensors and $z \sim \mathcal{N}(0, I_r)$, and define $V_{\ell_\star} \subset \mathrm{span}(W_\star)$ to be the space spanned by the singular vectors of $c_{\ell_\star}(f_\star)$. Then, after a single step of gradient descent:

- If $n_0 = \Theta(d^{\ell_\star - \delta})$ for $\delta > 0$, only a vanishing fraction of weights W_1 correlated with the target indices $W_\star$. In other words: if the batch size is not large enough, no subspace is weakly recovered after the gradient step.
- Otherwise, if $n_0 = \Omega(d^{\ell_\star})$ and $\eta = O(pd^{\frac{\ell_\star - 1}{2}})$, with high probability as $d \to \infty$ the weights W_1 weakly learns the subspace $V_{\ell_\star}$.

We refer the reader to [Dan+24a] for a formal statement.

Remark 3.2.1 (Relation with literature)**.** Theorem 3.2.1 extends two previous results in the literature. The rank-one property of the gradient for $n = \Theta(d)$ proven in [Ba+22] implies weak recovery in this regime, while [DLS22] showed the positive part the theorem for $n = \Theta(d^2)$, under the assumption that $V_2 = \mathrm{span}(W_\star)$ (their Assumption 2). Theorem 3.2.1

[29] Note the equivalence with the decomposition in eq. (2.62) is given by grouping $c_m = \{c_\alpha\}_{\alpha \in \mathcal{Z}_+^d : |\alpha| = m}$.

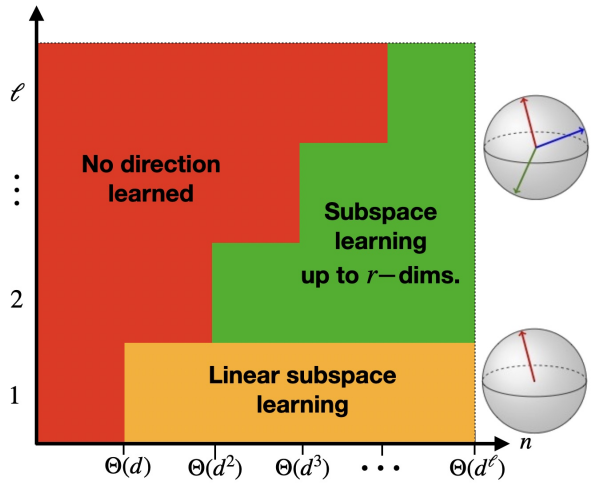

Figure 3.1: Illustration of theorem 3.2.1 of the weak recovery after a single gradient step. The y-axis shows the leap exponent of the target function and the x-axis the batch size n_0. In this figure, we assume the learning rate is chosen at the critical scaling $\eta = O(pd^{\frac{\ell_\star - 1}{2}})$. Figure from [Dan+24a].

complements these works by providing the matching lower bounds, demonstrating their tightness and establishing a general picture for *all higher* powers of d. In particular, it proves a clear separation between the class of functions that can be learned in the $\Theta(d)$ batch-size regime of [Ba+22] and those accessible in the $\Theta(d^2)$ regime of [DLS22], and it further establishes a hierarchy of functions requiring progressively larger batch sizes to be learned with a single gradient step.

Although cumbersome to state due to the technicalities of the tensor notation, Theorem 3.2.1 is quite intuitive: it provides a tight characterisation of the batch and learning rate sizes required to weakly recover the "easiest" (lowest frequency) directions of the target, as weighted by the link function. A visual summary of this result is given in fig. 3.1. An immediate corollary for the discussion that will follow in section 3.4.3 is the following:

Corollary 3.2.1 (Proportional regime)**.** In the proportional high-dimensional regime where $d \to \infty$ with $n, p = \Theta(d)$, W_1 correlates with at most a one-dimensional subspace of $\mathrm{span}(W_\star)$. More precisely, if $\ell_\star > 1$ or $n_0 = o_d(d)$, no subspace is (weakly) learned. Otherwise, if $\ell_\star \in \{0, 1\}$, $n_0, \eta = \Omega(d)$, a one dimensional subspace is (weakly) learned.

3.2.2 Few gradient steps

Consider now taking a few steps of stochastic gradient descent, $T = \Theta_d(1)$, according to eq. (3.3). From a technical perspective, this amounts to determining which subspaces become accessible given what has already been learned, thereby establishing a hierarchy of index subspaces of increasing difficulty.

On a high-level, this can be done by iteratively applying theorem 3.2.1 conditionally on the subspace which has been learned in the previous steps. However, this is technically intricate for two reasons. First, if $\ell_\star \in \{0, 1\}$, then after a single step we have $f(x; W_1, a_0) \neq 0$, which breaks the balancedness condition in assumption 3.1.1 and complicates the analysis. Second, the hierarchy must be defined in a basis-independent manner. To illustrate this point, consider the polynomial two-index function $g(z) = z_1 + z_2 + z_1^2 - z_2^2$ with $z_k = \langle w_{\star,k}, x\rangle$. This is a function with $\ell_\star = 1$, and according to theorem 3.2.1, with $n_0 = O(d)$ one can learn the one-dimensional subspace spanned by $w_{\star,1} + w_{\star,2}$.

Suppose we follow this intuition, and in a second step we condition on the directions already learned and apply theorem 3.2.1 again. The remaining part, $z_1^2 - z_2^2$, is a quadratic polynomial, and a naive application of theorem 3.2.1 would suggest that it cannot be learned at linear sample complexity $n = \Theta(d)$. However, this reasoning is misleading: note that $z_1^2 - z_2^2 = (z_1 + z_2)(z_1 - z_2)$, and hence, conditionally on $(z_1 + z_2) = a \in \mathbb{R}$, the expression reduces to $a(z_1 - z_2)$, which is simply a linear function in disguise and can indeed be learned with $n_0 = \Theta(d)$.

In summary, the quadratic polynomial $g(z) = z_1 + z_2 + z_1^2 - z_2^2$ can be learned with $n_0 = \Theta(d)$ in two steps: the first step identifies the one-dimensional subspace $V_1 = \mathrm{span}(w_{\star,1}+w_{\star,2})$, and the second step identifies $V_2 = \mathrm{span}(w_{\star,1} - w_{\star,2})$.

This example belongs to a broader class of non-linear functions known as *staircase functions*, introduced in [Abb+21; AAM22]. Staircase functions admit a decomposition into a sequence of subspaces of increasing complexity, with the key property that, conditional on the first T subspaces, the $T + 1$ subspace is linearly coupled to them.

To formalise this idea, we first need to make precise what we mean by conditioning a functions on a subspace.

Definition 3.2.1 (Subspace conditioning)**.** Let V be a vector space, and $U \subseteq V$ a subspace. For any function $f : V \to \mathbb{R}$, and $x \in U$, we define the *conditional* function $f_{U,\mathbf{x}} : U^\perp \to \mathbb{R}$ as

$$f_{U,x}(x^\perp) = f(x + x^\perp) \tag{3.9}$$

To get some intuition for this definition in our context of the GMIM where $g \in L^2(\gamma_r)$, consider the case in which $U = \mathrm{span}(v) \subset \mathbb{R}^p$ is a one-dimensional subspace. Then, any $z \sim \mathcal{N}(0, I_r)$ can be decomposed as $z = z^\perp + v$ with $\langle z^\perp, v\rangle = 0$. Thanks to Gaussian conditioning, $z^\perp$ is itself Gaussian, and therefore:

$$g(z) = g(z^\perp + v) = \sum_{m=0}^{\infty} \sum_{\alpha \in \mathbb{Z}_+^d} C_\alpha(v) H_\alpha(z^\perp) \tag{3.10}$$

with $C_\alpha(v) = \mathbb{E}[g(z^\perp + v)H_\alpha(z^\perp)]$. In other words, $H_\alpha(z^\perp)$ has no components along v. A

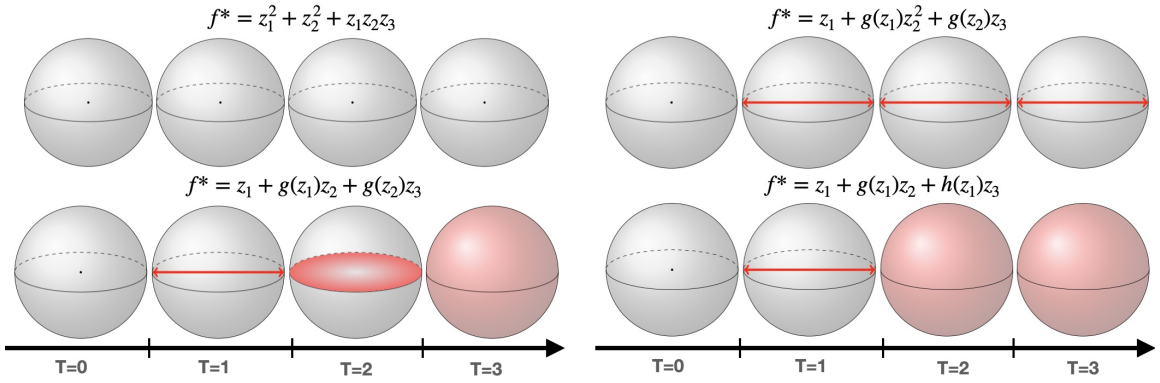

Figure 3.2: Illustration of theorem 3.2.2, showing the index space $\mathrm{span}(W_\star)$ for different multi-index targets $g(z)$ and the directions W_t learned after t one-pass SGD steps. In the early stages of training, the network first identifies a single direction associated with the linear component of the target, and subsequently uncovers additional directions that become linear once conditioned on those already discovered. Let $e_{kk\in[r]}$ denote the standard basis of $\mathbb{R}^r$. The four panels illustrate distinct scenarios: (**Top left**) no direction is learned; (**Top right**) only a single direction e_1 is recovered (single-index regime); (**Bottom left**) directions are learned sequentially, with e_1 followed by e_2 and then e_3; (**Bottom right**) e_1 is learned in the first step, and both e_2 and e_3 are learned in the second step. Figure from [Dan+24a].

very explicit example is $v = e_1$ where we can explicit write:

$$H_\alpha(z^\perp) = \prod_{j=2}^{d} h_{\alpha_j}(z_j) \tag{3.11}$$

This allow us to show the following result.

Theorem 3.2.2 ([Dan+24a], informal)**.** Consider a r-index GMIM target $f_\star(x) = g(W_\star x)$. Let $T \in \mathbb{Z}_+$ and assume that $n_0, \cdots, n_{T-1} = \Theta(d)$, and $\eta > 0$, $p \in \mathbb{Z}_+$ are fixed. Define a sequence of nested subspaces $U_0^\star \subseteq U_1^\star \subseteq U_{T-1}^\star \subseteq \mathrm{span}(W_\star)$ as

- $U_0^\star = \{0\}$,
- for any $t \in [T]$, $U_{t+1}^\star = U_t^\star \oplus \big(\{c_{U_t^\star, z}(g) : z \in U_t^\star\}\big)$, where $c_{U,z}(g) = \mathbb{E}_{z^\perp \sim \mathcal{N}(0, I_r)}\left[\nabla_{z^\perp} g_{U,z}\right]$ is the first Hermite tensor coefficient of $g_{U,z}$.

Then, after $t \in [T]$ gradient steps of one-pass SGD in eq. (3.3) with a balanced initialisation assumption 3.1.1, W_t weakly recover the index subspace $U_t^\star$ almost surely over a_0.

Informally, theorem 3.2.2 states that with linear sample complexity one-pass SGD weakly learns staircase components of $f_\star$. In particular, if the link function is a staircase function, it weakly recovers the full $\mathrm{span}(W_\star)$ in $T \in [r]$ steps with linear sample complexity. A few examples are discussed in fig. 3.2.

3.2.3 Batch reusing and CSQ vs. SQ classes

As discussed in section 1.3.1, many efficient algorithms for learning multi-index functions exist in the literature. In particular, as we will see in chapter 4, weak recovery can be achieved computationally with $n = \Theta(d)$ samples for most multi-index functions, including monomials of the form $g(z) = z_1 \cdots z_r$, which have leap exponent $\ell_\star = r$ [CM20]. It follows that the one-pass SGD sample complexities derived in theorems 3.2.1 and 3.2.2 are suboptimal.

This limitation is closely tied to the fact that in one-pass SGD the gradient updates in eq. (3.3) are independent. Each step has the form $\mathbb{E}[y\sigma'(\langle w_{t,k}, x\rangle)x]$, where $w_{t,k}$ is independent of (x, y) (c.f. eq. (3.6)), and hence carries vanishing information about the indices as $d \to \infty$ when $\ell_\star \geq 1$. This corresponds to a particular instance of the so-called *correlational statistical query* (CSQ) model, where the signal is accessed only through expectations of the form $\mathbb{E}[y\phi(x)]$ for some transformation $\phi : \mathbb{R}^d \to \mathbb{R}$. In [Dam+23], it was shown that CSQ algorithms require at least $n = \Omega(d^{\ell_\star/2})$ samples to weakly learn a single-index model ($r = 1$). Moreover, they demonstrated that this lower bound can be attained by one-pass SGD if the loss function is modified via a smoothing procedure inspired by statistical physics landscape analysis [BCR20]. In any case, since the training procedure in definition 3.1.1 effectively implements a CSQ algorithm, it follows that one-pass SGD is fundamentally constrained in this setting.

The situation changes when the batches of data are reused across the iterations. As shown in [Dan+24c], full-batch gradient descent can weakly recover high-frequency subspaces, such as $g(z) = h_3(z)$, with the information-theoretic sample complexity $n = \Theta(d)$ in only two steps. This improvement arises from the *memory effect* induced by data reuse, which introduces bias into the gradients. Importantly, this effect is not unique to full-batch gradient descent. As demonstrated in [Arn+24b; Lee+24], it already appears in the simplest setting of reusing a single data pair (x, y) twice. To see this, consider the single-index target $y = h_{\ell_\star}(\langle w_\star, x\rangle)$ with leap $\ell_\star > 1$. As shown in eq. (3.7), for one-pass SGD the correlation $\mathbb{E}[\langle w_\star, g_k\rangle] = \Theta(d^{-(\ell_\star-1)/2})$, and thanks to the independence of batches this remains true at every step. By contrast, when two steps of SGD are performed on the same data point, the first step yields the same correlation as in one-pass SGD, but the second step leads to

$$\begin{aligned}\mathbb{E}[\langle g_{2,k}, w_\star\rangle] &\propto -\mathbb{E}[g(\langle w_\star, x\rangle)\sigma'(\langle w_{0,k} - \eta g_{0,k}, x\rangle)\langle w_\star, x\rangle] \\ &\propto -\mathbb{E}[g(z_\star)\sigma'(z_k - \eta g(z_\star)\sigma'(z_k)z_\star]. \end{aligned} \tag{3.12}$$

for $(z_\star, z_k)$ jointly Gaussian variables with unit variance and correlation $1/d\langle w_\star, w_{0,k}\rangle = O(d^{-1/2})$. Expanding in the limit $d \to \infty$, one finds that the resulting expectation is of order $\Theta_d(1)$ — see [Arn+24b] for the derivation.

The key difference between eq. (3.12) and eq. (3.7) is that the former takes the form

$\mathbb{E}[\psi(x, y)]$, where $\psi : \mathbb{R}^{d+1} \to \mathbb{R}$ is a joint transformation of both covariates and labels. This corresponds to a *statistical query* (SQ) in the sense of [Kea98], and defines a richer class of algorithms than correlational statistical queries (CSQ). For Gaussian single-index models, a SQ lower bound of $n = \Omega(d^{k_\star/2})$ was established in [Dam+24], where the sample complexity is governed by the so-called *generative exponent* $k_\star \in \mathbb{Z}_+$, which is strictly smaller than the information-theoretic exponent. We will return to this point in chapter 4, where it will be discussed in the context of optimal algorithms for the GMIM.

3.3 From weak recovery to generalisation

So far, our discussion has focused only on weak recovery, i.e. the first step in the step-wise training in definition 3.1.1. In this section, we discuss the implications to generalisation once the second layer is trained.

Intuitively, once a subspace $U \subset \mathrm{span}(W_\star)$ of the indices has been weakly learned during training, the contribution of the target lying in this subspace should no longer contribute to the risk. In the most favourable case, all the energy of the target function supported on U is effectively removed from the error, leaving only the orthogonal component to be fitted. This intuition can be formalised in the following conjecture, which provides a lower bound on the risk.

Conjecture 3.3.1 ([Dan+24a], informal)**.** Assume that $\min(n, p) = \Theta(d^\kappa)$ and that the learned first layer weights $\hat{W}$ span a subspace $U \subseteq \mathrm{span}(W_\star)$. Then, the risk of ridge regression on the second-layer weights is lower-bounded as

$$\mathbb{E}\left[\left(f_\star(x) - f(x; \hat{a}_\lambda, \hat{W})\right)^2\right] \geq ||P_{U,>\kappa} f_\star||^2_{L^2(\gamma_d)} - o(1), \tag{3.13}$$

where $P_{U,>\kappa}$ is the projector of the target $f_\star$ into the polynomials of degree $> \kappa$ which are orthogonal to the learned subspace

$$P_{U,>\kappa} f_\star(x) = \sum_{m>\kappa} \sum_{\substack{\alpha \in \mathbb{Z}_+^d \\ |\alpha|=m}} C_\alpha(P_U x) H_\alpha(x^\perp) \tag{3.14}$$

Proving conjecture 3.3.1 in full generality is a challenging open problem, and numerical evidence can be found in [Dan+24a]. Nevertheless, we can show it holds in the particular case of the proportional asymptotics.

Theorem 3.3.1 ([Dan+24a])**.** Conjecture 3.3.1 holds in the proportional regime $d \to \infty$ with $n, p = \Theta(d)$.

Together with theorem 3.2.1, this results show that if $f_\star$ has leap $\ell_\star \in \{0, 1\}$, in the

proportional regime the network can learn at least a single-index approximation of the target function $f_\star(x) \approx g(\langle \theta_\star, x\rangle) + \text{noise}$.

3.4 Sharp asymptotics

Theorem 3.3.1 establishes a lower bound on what can be learned in the proportional regime. We now examine this result more closely, aiming at an exact characterisation of feature learning after a single gradient step.

For the sake of clarity, let's repeat the setting we will be focusing in this section. Let $\mathcal{D} = \{(x_i, y_i) \in \mathbb{R}^{d+1} : i \in [N]\}$ denote the training data, which we assume has been drawn from a Gaussian single-index model:

$$y_i = g(\langle w_\star, x_i\rangle), \qquad w_\star \in \mathbb{S}^{d-1}(\sqrt{d}), \quad x_i \sim \mathcal{N}(0, {}^1\!/_d I_d) \tag{3.15}$$

Since we will be interested in the proportional regime, thanks to theorem 3.3.1 this is without loss of generality. Let $N = n_0 + n$ denote a split of the data in two disjoin batches, and consider one step of SGD:

$$w_{1,k} = w_{0,k} - \eta \nabla_w \frac{1}{n_0} \sum_{i=1}^{n_0} \left(g(\langle w_\star, x_i\rangle) - f(x_i; W_0, a_0)\right)^2 . \tag{3.16}$$

from initial condition $w_{k,0} \sim \text{Uni}(\mathbb{S}^{d-1}(\sqrt{d}))$ and $a_{k,0}$. We will be interested in the following ERM problem

$$\hat{a}_\lambda = \arg\min_{a \in \mathbb{R}^p} \sum_{i=1}^{n} \left(g(\langle w_\star, x_i\rangle) - f(x_i; W_1, a)\right)^2 + \lambda ||a||_2^2 \tag{3.17}$$

for a two-layer neural network $f(x; W, a) = {}^1\!/_{\sqrt{p}} \langle a, \sigma(Wx)\rangle$ in the feature rich proportional asymptotics where $d \to \infty$ with:

$$\frac{n}{d} \to \alpha, \qquad \frac{n_0}{d} \to \alpha_0, \qquad \frac{p}{d} \to \gamma, \qquad \frac{\eta}{d} \to \tilde{\eta}, \qquad \sqrt{p}\, a_{0,k} \to \tilde{a}_{0,k} \tag{3.18}$$

where the right-hand side quantities are $\Theta_d(1)$.

Remark 3.4.1 (Unbalanced initialisation)**.** Note that the above is slightly more general than the assumptions we made in section 3.1. In particular, we do not require initialisation to be balanced.

As in the random features case discussed in Chapter 2, deriving sharp asymptotics requires analysing the limiting behaviour of the feature matrix $\Phi_{ik} = \sigma(\langle w_{1,k}, x_i\rangle)$. At initialisation, this was achieved via Gaussian equivalence; here, an analogous result is needed.

3.4.1 Equivalent spiked random features model

The starting point is to have a sharper control of the gradient update. By separating the activation in the linear and non-linear parts on the Hermite basis $\sigma(z) = \mu_1 z + \sigma_{>1}(z)$ in eq. (3.16), one can show that the update can be re-writen as:

$$W_1 = W_0 + uv^\top + \Delta \tag{3.19}$$

where:

$$u = \frac{\mu_1 \eta a_0}{\sqrt{p}}, \qquad v = \frac{X^\top y}{n_0} \tag{3.20}$$

and Δ account for the remaining, high-frequency terms. Note that $u \in \mathbb{R}^p$ is proportional to the second-layer weights at initialisation, while $v \in \mathbb{R}^d$ is the signal part of the gradient. Indeed, v can be seen as implementing an average over the gradient of $f_\star$:

$$v = \frac{1}{n_0} \sum_{i=1}^{n_0} x_i g(\langle w_\star, x_i \rangle) \xrightarrow{n_0 \to \infty} \mathbb{E}[g'(\langle w_\star, x \rangle) w_\star] = c_1 w_\star \tag{3.21}$$

where $c_1 = \mathbb{E}[g'(z)]$. Indeed, this approximation is good as soon as $n_0 = \Theta(d^{1+\delta})$, i.e. $\alpha_0 \to \infty$. It is easy to see that the rows of the high-frequency part of the gradient $\Delta_k \in \mathbb{R}^d$ are independent, and one can show they satisfy the following properties.

Lemma 3.4.1 ([Dan+24a], informal)**.** With high-probability on as $d \to \infty$ with $n_0, p, \eta = \Theta(d)$:

- **Vanishing correlation with the signal:**

$$\langle \Delta_k, w_\star \rangle = O\left(\frac{\text{polylog}(d)}{p\sqrt{d}}\right) \tag{3.22}$$

- **Operator norm**:

$$||\Delta||_{\text{op}} = O\left(\frac{\text{polylog}(d)}{\sqrt{d}}\right) \tag{3.23}$$

- **Orthogonality**:

$$\langle \Delta_k, \Delta_l \rangle = O\left(\frac{\text{polylog}(d)}{p^2\sqrt{d}}\right), \quad k \neq l \tag{3.24}$$

Intuitively, we would like to treat W_1 as a spiked matrix model, where $W_0 + \Delta$ plays the role of the bulk and $uv^\top$ is a rank-one spike. However, this is not straightforward: while

both Δ and W_0 have independent and almost orthogonal rows, differently from W_0, the rows of Δ are anisotropic, making the analysis challenging. It can be shown, however, that the anisotropic components of Δ are of $O(\alpha_0)$, meaning that in the limit $\alpha_0 \to \infty$ we can approximate W_1 by a isotropic spiked model. Together with eq. (3.21), this yields the following result.

Lemma 3.4.2 ([Dan+25]). For any $\delta > 0$, in the limit $d \to \infty$ with $n, p, \eta = \Theta(d)$ and $n_0 = \Theta(d^{1+\delta})$ the first gradient step can be approximated by a isotropic spiked matrix model:

$$||W_1 - \left(F + c_\star u w_\star^\top\right)||_{\mathrm{F}} \to 0, \qquad \text{a.s.} \quad \mathrm{d} \to \infty. \tag{3.25}$$

where $w_\star \in \mathbb{S}^{d-1}(\sqrt{d})$, $u = \gamma^{-1}\mu_1\tilde{\eta}\tilde{a}_0$ and $F \in \mathbb{R}^{p\times}$ is a matrix i.i.d. rows satisfying $\langle f_k, f_l\rangle = d\delta_{kl}\left(1 + \Theta(d^{-1/2})\right)$.

This result effectively means that studying the two-step training procedure in the proportional asymptotics is equivalent to studying the ERM problem in eq. (3.17) with the following *spiked random features model* (sRF) [Cui+24]:

$$\min_{a\in\mathbb{R}^p} \sum_{i=1}^{n} \left(g(\langle w_\star, x_i\rangle) - \langle a, \sigma\left(Fx_i + u\langle v, x_i\rangle\right)\rangle\right)^2 + \lambda||a||_2^2 \tag{3.26}$$

with $\lim_{d\to\infty} {}^1\!/_d\langle v, w_\star\rangle = c_1$.

3.4.2 Conditional Gaussian equivalence

The final step in characterising the risk is to approximate the spiked feature matrix $\Phi = \sigma(XF^\top + Xvu^\top)$ in the high-dimensional limit. As in the random features model, this relies on an equivalence result showing that only the low-frequency components of the features contribute meaningfully to the error, while the high-frequency components act as effective noise. Unlike the random features case, however, this approximation retains information about the component of the target function already captured during training. In other words, the learned features are crucially correlated with the index $w_\star$.

Theorem 3.4.1 ([Dan+24a], informal). Consider the spiked random features model with features:

$$\Phi_{ik} = \sigma\left(\langle f_k, x_i\rangle + u_k\langle v, x_i\rangle\right) \tag{3.27}$$

Define the linearised feature map:

$$G_{ik} = \mu_0\left(u_k\langle v, x\rangle\right) + \mu_1\left(u_k\langle v, x\rangle\right)\langle f_k, x_i\rangle + \mu_\star\left(u_k\langle v, x\rangle\right)Z_{ik}, \tag{3.28}$$

where $Z_{ik} \sim \mathcal{N}(0,1)$, $\kappa = \langle v, x \rangle$ and:

$$\begin{aligned}
\mu_0(\kappa) &= \mathbb{E}\left[\sigma(z+\kappa)\right], \\
\mu_1(\kappa) &= \mathbb{E}\left[z\sigma(z+\kappa)\right], \\
\mu_\star(\kappa) &= \sqrt{\mathbb{E}\left[\sigma^2(z+\kappa)\right] - \mu_1(\kappa)^2 - \mu_0(\kappa)^2},
\end{aligned} \tag{3.29}$$

where $z \sim \mathcal{N}(0,1)$. Then, in the proportional high-dimensional limit the test error of the minimiser in eq. (3.26) is asymptotic to the test error of the conditional equivalent problem with linearised features eq. (3.28).

Remark 3.4.2 (cGET beyond sRF)**.** The intuition behind the derivation of theorem 3.4.1 is that the non-Gaussian components of the features span only a low-dimensional subspace, while the remaining directions behave as isotropic Gaussian. This idea was first introduced in [Dan+23], where it was used to establish conditional universality in mixture models.[30] We conjecture that this principle extends beyond these two settings, to any problem in which the non-Gaussian components of the features influence the risk only through a limited number of directions.

3.4.3 High-dimensional asymptotics

Combining the characterisation of the gradient and the conditional Gaussian equivalence for the features allow us to derive a sharp characterisation of the risk in the proportional high-dimensional regime, akin to the results discussed of chapter 2.

Theorem 3.4.2 ([Cui+24; Dan+25], informal)**.** Assume that $\tilde{a}_{0,k}$ is drawn i.i.d. from a finite vocabulary of size $k = \Theta_d(1)$ with probabilities $(\pi_q)_{q\in[k]}$. Then, in the limit $d \to \infty$ with $n, p, \eta = \Theta_d(d)$ and $n_0 = \Theta_d(d^{1+\delta})$ with $\delta > 0$, the excess risk of the minimiser in eq. (3.17) admits a deterministic equivalent:

$$|R(\hat{a}_\lambda) - \mathsf{R}(\alpha, \gamma, \lambda, \tilde{\eta}, \pi)| \to 0, \qquad \text{a.s.} \quad d \to \infty \tag{3.30}$$

The explicit formulas for R are rather cumbersome and are omitted here for brevity; we refer the interested reader to the original works [Cui+24; Dan+25].

Remark 3.4.3 (One step vs. sRF)**.** A deterministic equivalent for the risk, analogous to theorem 3.4.2, can be derived for the spiked random features model in the regime $n_0 = \Theta_d(d)$ (i.e. $\alpha_0 = \Theta_d(1)$) under isotropic F. However, as discussed in section 3.4.1, this provides only an approximation to the full bulk, which in this regime contains anisotropic corrections.

[30] A Gaussian mixture model can also be expressed as a spiked matrix model $X = y\mu^\top + Z$, but with the spike appearing in the covariates rather than in the weights.

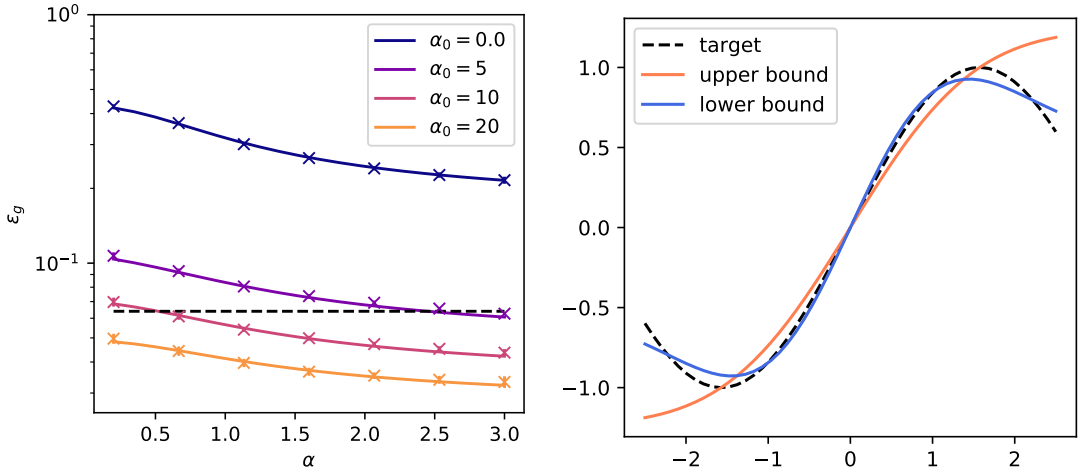

Figure 3.3: **(Left)** Generalisation error of learning of the isotropic sRF model as a function of $\alpha = {}^{n}/_{p}$ for different values of batch size $\alpha_0 = {}^{n_0}/_{d}$, $g = \sin$, $\sigma = \tanh$, $\tilde{\eta} = 1$, $\lambda = 0.1$ and uniform initialisation $\tilde{a} = 1_p$. Solid lines denote theoretical results from a generalisation of theorem 3.4.2 to $\alpha_0 = \Theta_d(1)$, and crosses denote finite size simulations with $d = 2000$. The dashed black line represents the lowest achievable MSE for kernel/linear methods. **(Right)** Upper (orange) and lower (blue) bounds for the approximation of a sine function $g = \sin$ with $\sigma = \tanh$ as predicted from corollary 3.4.1. Figures from [Cui+24].

Obtaining a sharp characterisation of the risk for the anisotropic spiked random features model remains an open problem.

Theorem 3.4.2 provides a sharp characterisation of feature learning after a single gradient step, yielding not only the asymptotic risk but also an effective description of how the network adapts its features to the low-dimensional structure of the target, as a function of $(\alpha, \gamma, \tilde{\eta}, \tilde{a})$. In Figure 3.3 (left), we compare the best achievable risk of kernel methods at $n = \Theta(d)$, as characterised in section 2.4, with the risk attained after one step of training for different batch sizes $\alpha = {}^{n_0}/_{d}$. While the performance at initialisation is always lower bounded by that of the best linear method (theorem 2.4.1), a single gradient step with moderate sample complexity surpasses this high-dimensional bottleneck thanks to the adaptivity of the features to the target.

The explicit formulas in theorem 3.4.2 can be considerably simplified in the case where the readout layer weights are initialised homogeneously $a_0 = {}^{1}/_{\sqrt{p}} 1_p$. From the exact formulas, one can derive interpretable lower and upper bounds from the risk.

Corollary 3.4.1 ([Cui+24], informal)**.** Under the same setting of section 3.4.3 with the additional assumption that $\tilde{a}_0 = 1_p$, the excess risk of ridge regression satisfies the following

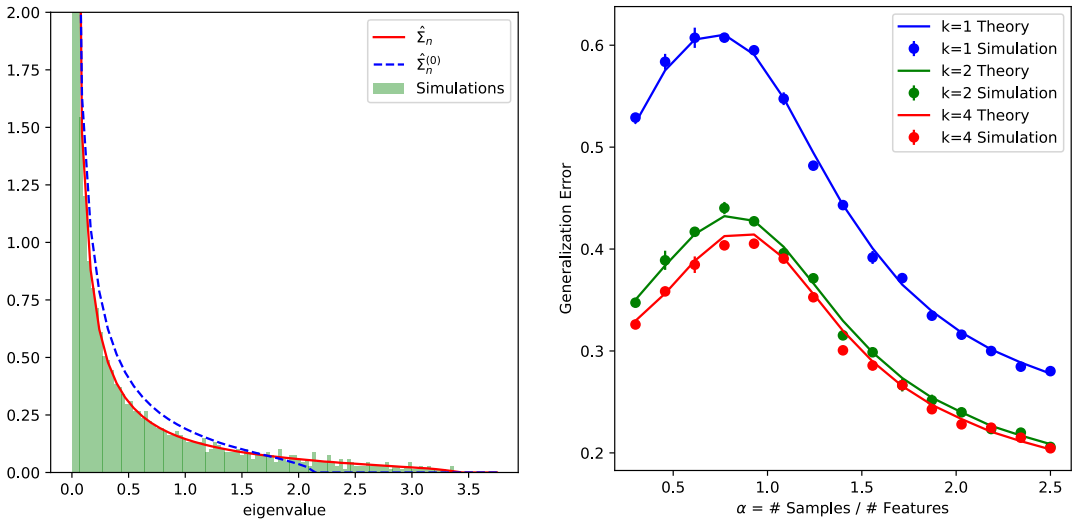

Figure 3.4: **(Left)** Asymptotic spectral density of the features empirical covariance matrix before (dashed blue) and after (solid red) training with one-step of SGD. The lines denote the theoretical predictions, while the green histogram denotes finite size simulations. **(Right)** Risk as a function of $\alpha = {}^{n}/_{p}$ for different vocabulary sizes $k \in [4]$ and $\lambda = 0.01$, $\tilde{\eta} = 0.5$, $\gamma = 1.5$. Solid curves denote the theoretical predictions from theorem 3.4.2, and dots denote simulations with $p = 2048$. Figure from [Dan+25].

upper and lower bounds under optimal choice of regularisation $\lambda \geq 0$:

$$\inf_{\lambda \geq 0} \mathsf{R}(\alpha, \gamma, \lambda, \tilde{\eta}) \leq \inf_{\nu_1} \mathbb{E}_\kappa \left[\left(g(\kappa) - \nu_1 \mu_0(\kappa) \right)^2 \right], \tag{3.31}$$

$$\inf_{\lambda \geq 0} \mathsf{R}(\alpha, \gamma, \lambda, \tilde{\eta}) \geq \inf_{\nu_1, \nu_2} \mathbb{E}_\kappa \left[\left(g(\kappa) - \nu_1 \mu_0(\kappa) - \nu_2 \mu_1(\kappa) \kappa \right)^2 \right]. \tag{3.32}$$

where $\kappa \sim \mathcal{N}(0, 1)$.

The upper bound (3.31) corresponds to Lemma 6 of [Ba+22] in the case of uniform readout initialisation $a_0 = {}^{1_p}/_{\sqrt{p}}$, and is attained in the limit $\lambda \to \infty$. The lower bound (3.32), by contrast, shows that the risk cannot fall below the $L^2(\gamma_1)$ distance between the target link function g and the span of $\{\mu_0, \tilde{\mu}_1\}$, where $\tilde{\mu}_1(\kappa) = \kappa \mu_1(\kappa)$, with the best approximation given by the orthogonal projection of g onto this span. Figure 3.3 (right) illustrates the functions that realise the upper and lower bounds, and how these compare to the target g. Finally, in the random features limit the functions $\mu_0(\kappa)$ and $\mu_1(\kappa)$ collapse to constants independent of κ, restricting the class of learnable functions to linear ones — consistent with the discussion in Section 2.4. This result therefore give us a low-dimensional summary of how the high-dimensional features adapt after one step of SGD.

One of the steps in the characterisation of the asymptotic risk in theorem 3.4.2 involves controlling the empirical covariance of the feature matrix, $\hat{\Sigma}_n = {}^{1}/_{n} \Phi^\top \Phi$. A corollary of this result is a deterministic equivalent for $\hat{\Sigma}_n$, from which the asymptotic spectral density can be derived; see [Dan+25] for explicit formulas. Figure 3.4 (left) compares the bulk of the

asymptotic spectral density of the empirical feature matrix before and after training. Relative to the initialisation spectrum[31], training produces a modified bulk with broader support and heavier tails. This theoretical prediction resonates with a range of empirical findings [MM21; MPM21; Wan+24], which have reported the emergence of heavier-tailed spectra following feature learning, often correlating with improved generalisation. Notably, this phenomenon persists even when networks are trained with multiple large stochastic gradient steps or adaptive optimisers such as Adam [Kin14], as observed empirically in [Wan+24].

Remark 3.4.4 (Readout initialisation)**.** A consequence of the theory in [Cui+24; Dan+25] is that the effective number of parameters in the high-dimensional limit is proportional to the number of elements of the alphabet in which $a_{0,j}$ is initialised. This implies that adding diversity in the initial weights $a_{0,j}$ increases the expressivity of the network. Indeed, from the conditional Gaussian equivalent characterisation in theorem 3.4.1, for a vocabulary V, the functional basis $\{\mu_0(\omega\cdot), \tilde{\mu}_1(\omega\cdot)\}_{\omega\in V}$, thereby allowing the network to span a larger class of functions. This implies that the functional space spanned by these functions is generically of dimension $2|V|$ for non-uniform readout initializations a_0, compared to just 2 in the uniform readout case. Figure 3.4 (right) illustrates how a larger vocabulary size for a given task leads to an improvement in the risk.

[31] Given by a shifted Marchenko–Pastur distribution due to Gaussian universality, see eq. (2.67)

4 | Fundamental limitations

In this chapter, we examine the fundamental computational limits of learning Gaussian multi-index models with limited data in the high-dimensional regime. Our focus is on classifying which multi-index functions are computationally tractable and which are intrinsically hard to learn, depending on how the link function couples the different indices. This analysis serves as a benchmark for the results developed in chapters 2 and 3 on two-layer neural networks, providing a reference point against which the benefits and limitations of feature learning can be assessed.

The results discussed in sections 4.2 and 4.3 are based on [Tro+25], while the results discussed in section 4.4 are based on [Def+25]. The discussion is partially inspired from a lecture taught on these results at the *Statistical Physics & Machine Learning: moving forward* summer school in August 2025 [Lou25].

4.1 Approximate message passing

Let $\mathcal{D} = \{(x_i, y_i) \in \mathbb{R}^{d+1} : i \in [n]\}$ denote a batch of training data drawn from the Gaussian multi-index model introduced in definition 1.3.2, which we recall here for convenience:

$$y_i \sim \mathsf{P}_y(\cdot|W_\star x_i), \qquad x_i \sim \mathcal{N}(0, {}^1\!/_d I_d) \quad \text{i.i.d.} \tag{4.1}$$

where P_y is the likelihood parametrising the link function and $W_\star \in \mathbb{R}^{r\times d}$ are the indices, which in this chapter we will assume are Gaussian i.i.d. matrices.

Consider the most favourable scenario where the link function (or likelihood) P_y and the distribution of the indices P_W are known to the statistician, who seeks to estimate the specific realisation of the indices $W_\star \sim p_W$ that generated the training data $\mathcal{D}$, also known as the *Bayes-optimal* scenario. Under the square loss, the estimator achieving the best possible risk is given by the posterior mean:

$$\text{mmse} = \min_{\hat{W}\in\mathbb{R}^{r\times d}} \mathbb{E}\left[||W_\star - \hat{W}(\mathcal{D})||_{\text{F}}^2\right] = \mathbb{E}\left[||W_\star - \mathbb{E}[W|\mathcal{D}]||_{\text{F}}^2\right] \tag{4.2}$$

where the posterior distribution is explicitly written as

$$p(W|\mathcal{D}) = \frac{p(W)}{Z_d(\mathcal{D})} \prod_{i=1}^{n} \mathsf{P}_y(y_i|Wx_i). \tag{4.3}$$

Hence, the marginals of this posterior distribution characterise the information-theoretic limits of recovery. Analysing them when d is large, however, is notoriously challenging: sampling from $p(W|\mathcal{D})$ is, in the worst case, computationally intractable, requiring time exponential in the dimension.

Our main tool in the following will be the *approximate message passing* algorithm. This is an iterative algorithm that seeks to approximate the marginals of the posterior eq. (4.3) from an initial guess $\hat{W}_0$:

$$\Omega_t = X f_{\text{in}}(B_t, A_t) - g_{\text{out}}(y, \Omega_{t-1}, V_t) V_t^\top \tag{4.4}$$

$$B_{t+1} = X^\top g_{\text{out}}(y, \Omega_t, V_t) - f_{\text{in}}(B_t, A_t) A_t^\top \tag{4.5}$$

where $\Omega_t \in \mathbb{R}^{n\times r}$ and $B_t \in \mathbb{R}^{d\times r}$ are matrices with rows $\omega_i, b_j \in \mathbb{R}^r$, respectively, and $f_{\text{in}}(\cdot; A) : \mathbb{R}^r \to \mathbb{R}^r$ and $g_{\text{out}}(y, \cdot; V) : \mathbb{R} \times \mathbb{R}^r \to \mathbb{R}$ are two vector-valued functions acting row-wise on the matrices B_t, Ω_t:

$$g_{\text{out}}(y, \omega, V) = \mathbb{E}_{z\sim\mathcal{N}(0,I_r)} \left[V^{-1}(z - \omega) \mathsf{P}_y(y|z) \right], \qquad f_{\text{in}}(b, A) = (I_r - A)^{-1} b \tag{4.6}$$

and A_t, V_t are given by:

$$A_t = \frac{1}{d} \sum_{i=1}^{n} \nabla_{\omega_i} g_{\text{out}}(y_i, \omega_i, V_i), \qquad V_t = \frac{1}{d} \sum_{j=1}^{d} \nabla_{b_j} f_{\text{in}}(b_j, A_j) \tag{4.7}$$

Finally, the estimate of $W_\star$ after T steps is obtained by $\hat{W}_{\text{amp}} = f_{\text{in}}(B_T, A_T)^\top$.

Intuitively, the AMP iterates in eq. (4.4) can be viewed as a two-step procedure: first estimating the pre-activations $z \in \mathbb{R}^r$ from the observations $y = g(z)$, and then recovering the indices $W \in \mathbb{R}^{r\times d}$ from the relation $z = Wx$. The functions f_{in} and g_{out} defined in eq. (4.6) are precisely the Bayes-optimal denoisers for these two subproblems. Moreover, AMP is a first-order method: it requires only the evaluation of r-dimensional functions and matrix–vector multiplications by $X \in \mathbb{R}^{n\times d}$ and its transpose. For $r = O(1)$, the computational cost is therefore dominated by these multiplications, scaling linearly with the size of the data matrix, i.e. $\Theta(nd)$.

What distinguishes AMP from other first-order algorithms is that it is provably optimal in the proportional high-dimensional regime $d \to \infty$ with $n = \Theta(d)$. This has been proven in [CMW20; MW24], showing that AMP achieves the best possible performance among all first-order methods in this limit. As such, AMP provides a fundamental computational

benchmark for this class of algorithms.

The key fact that makes AMP a powerful theoretical tool is that its asymptotic performance can be tracked by a set of state evolution equations, which was first derived and proven for the Gaussian-multi index model in [Aub+18].

Lemma 4.1.1 (State evolution [Aub+18; GB23])**.** Let $\mathcal{D} = \{(x_i, y_i) \in \mathbb{R}^{d+1} : i \in [n]\}$ denote i.i.d. samples from the Gaussian multi-index model lemma 4.1.1. Run AMP from random initialisation $\hat{W}_0 \in \mathbb{R}^{r\times d}$ with $\hat{w}_{0,k} \sim \mathcal{N}(0, I_d)$ i.i.d. Denote by $\hat{W}_t$ the resulting estimator at time $t \in [T]$. Then, in the high-dimensional limit $n, d \to \infty$ with fixed ratio ${}^{n}/_{d} \to \alpha = \Theta_d(1)$, constant $r, T = \Theta_d(1)$, the limiting overlaps satisfy:

$$\frac{1}{d}\hat{W}_t\hat{W}_t^\top \xrightarrow{P} M_t, \qquad \frac{1}{d}\hat{W}_t W_\star^\top \xrightarrow{P} M_t, \tag{4.8}$$

with M_t satisfying the *state evolution equations* from initial condition M_0 iterated with:

$$M_{t+1} = F(M_t) \tag{4.9}$$

where:

$$F(M_t) = G\left(\alpha \mathbb{E} I_r g[g_{\text{out}}\left(Y_t, \sqrt{M_t}\xi, I_r - M_t\right)^{\otimes 2} I_r g]\right). \tag{4.10}$$

where $G(M) = (I_r + M)^{-1}M \in \mathbb{R}^{r\times r}$ and the expectation is taken over the following effective process

$$Y_t \sim \mathsf{P}_y\left(\cdot|\sqrt{I_r - M_t}Z + \sqrt{M_t}\xi\right), \tag{4.11}$$

with $Z, \xi \sim \mathcal{N}(0, I_r)$ independently. The asymptotic mean-squared error on the label prediction is then given by:

$$\mathbb{E}\left[\left(y - g\left(\hat{W}_t(X, y)x\right)\right)^2\right] \xrightarrow{P} \mathbb{E}[(Y_t - g(Z))^2],$$

where the expectation is taken over the effective estimation process eq.(4.11) and $\xrightarrow{P}$ denotes convergence in probability w.r.t the training data as $n, d \to \infty$.

Remark 4.1.1. It can be shown that F preserves the symmetry and positive semi-definite properties. Therefore, if $M_0 \succeq 0$ is a symmetric p.s.d. matrix, so is $M_{t+1} = F(M_t)$ for every $t \in [T]$.

The state evolution equations reduce the task of analysing the fundamental computational bottlenecks of learning GMIM in the high-dimensional limit to studying a deterministic dynamical system on the cone of positive semi-definite matrices $M \in \mathbb{R}^{r\times r}$. Next, we can leverage lemma 4.1.1 to classify which problems are easy or hard for AMP in the high-dimensional limit.

4.2 Trivial, easy and hard subspaces

Since in the Bayes-optimal scenario the prior distribution is known to the statistician, a natural choice of initial condition is $W_0 \sim P_W$. Consequently, at initialisation the overlap matrix is element-wise small:

$$\frac{\langle \hat{w}_{k,0}, w_{\star,k'} \rangle}{d} = \tilde{O}_{\mathbb{P}}(d^{-1/2}) \tag{4.12}$$

In the high-dimensional $d \to \infty$ limit, this implies that state evolution from a so-called *uninformed initialisation* reduces to the zero matrix, $M_0 = 0 \in \mathbb{R}^{r \times r}$. It is therefore necessary to examine the trajectory of state evolution eq. (4.9) starting from 0. In particular, the first question is whether the zero matrix constitutes a fixed point.

Lemma 4.2.1 (Existence of uninformed fixed point)**.** $M = 0 \in \mathbb{R}^{r \times r}$ is a fixed point of the state evolution eq. (4.9) if and only if the following condition holds almost surely over the effective process $Y = g(Z)$ for $Z \sim \mathcal{N}(0, I_r)$:

$$g_{\text{out}}(Y, 0, I_r) = \mathbb{E}[Z|Y] = \int_{\mathbb{R}^r} \frac{\mathrm{d}z}{(2\pi)^{d/2}} e^{-\frac{1}{2}||z||_2^2} \mathsf{P}_y(Y|z) = 0 \tag{4.13}$$

This is an intrinsic property of the likelihood $\mathsf{P}_y(Y|\cdot)$, and it is satisfied for instance if P_y is an odd function of z.

4.2.1 Trivial subspace

If $M = 0$ is not a fixed point of eq. (4.9), then $M_1 = F(M_0) \succ 0$. In other words, a single iteration of AMP from the initialisation suffices — *regardless of how small the sample complexity* $\alpha > 0$ may be — to weakly recover a subspace $T_\star$ of dimension $\text{rank}(M_1) > 0$, in the sense of eq. (1.12). We refer to this subspace $T_\star \subset \text{span}(W_\star)$ as a *trivial subspace*. More formally, this can be define as follows.

Definition 4.2.1 (Trivial subspace)**.** Let $H_\star \subset \text{span}(W_\star)$ denote the subspace spanned by the vectors $v \in \mathbb{R}^r$ satisfying:

$$\langle g_{\text{out}}(Y, 0, I_r), v \rangle = \lim_{d \to \infty} \mathbb{E}[\langle W_\star^\top v, x \rangle | Y = y] = 0 \tag{4.14}$$

where equality holds almost surely over $Y = g(Z)$ with $Z \sim \mathcal{N}(0, I_r)$. The trivial subspace $T_\star$ is define as the orthogonal complement of $H_\star$, i.e. $\text{span}(W_\star) = T_\star \oplus H_\star$

As previously said, trivial subspaces can be learned at any sample complexity by a single step of AMP.

Theorem 4.2.1. For any $\alpha > 0$, with high-probability as $d \to \infty$, the AMP algorithm in eq. (4.4) weakly recovers $T_\star$ as per eq. (1.12) in a single iteration.

Example 4.2.1. To get some intuition, we can have a look at a few examples of trivial subspaces.

- For single-index models ($r = 1$), T_* is one dimensional if and only if g is non-even, e.g. $g(z) = \mathrm{He}_3(z)$. This follows from requiring that $g_{\mathrm{out}}(y, 0, 1) \neq 0$ for at least one value of y. In particular, on any open interval where g_{out} is invertible we have $g_{\mathrm{out}} = g^{-1}$.
- For a linear multi-index model, $g(z) = \sum_{k=1}^{r} z_i$, $T_\star$ is spanned by $1_r \in \mathbb{R}^r$ (all-one vector).
- For a committee $g(z) = \sum_{k=1}^{r} \mathrm{sign}(z_k)$, the trivial subspace $T_\star$ is again 1d, spanned by $1_r \in \mathbb{R}^r$.
- For monomials $g(z) = z_1 \dots z_r$, the trivial subspace $T_\star$ is non-empty if and only if $p = 1$.
- For leap one staircase functions [AAM23]:

$$g(z) = z_1 + z_1 z_2 + z_1 z_2 z_3 + \dots \tag{4.15}$$

  The trivial subspace is $T_\star = \mathbb{R}^r$ and is spanned by the canonical basis. In other words, AMP learns all the directions with a *single step* for any $\alpha > 0$.

Theorem 4.2.1 already highlights a stark contrast between one-pass SGD and AMP. Indeed, certain functions that can be learned with arbitrarily small sample complexity by a single step of AMP require diverging sample complexity under one-pass SGD. This is exemplified by the cubic single-index function $g(z) = h_3(z)$, which is trivial for AMP but, as shown in theorem 3.2.1, requires $n = \Theta(d^2)$ samples for weak recovery under one-pass SGD. As discussed in section 3.2.3, this inefficiency of one-pass SGD can be understood within the statistical query framework as a separation between CSQ and SQ queries. A similar analogy holds for AMP.

Remark 4.2.1 (Optimal label pre-processing)**.** The condition in lemma 4.2.1 admits a natural interpretation in the framework of statistical queries [Kea98]. Specifically, the denoiser g_{out} can be viewed as a non-linear transformation of the labels, $y \mapsto g_{\mathrm{out}}(y, 0, I_r)$. From this perspective, the condition for the existence of a non-empty trivial subspace translates into the requirement

$$\mathbb{E}[g_{\mathrm{out}}(y, 0, I_r)^\top v \langle W_\star^\top v, x \rangle] = \mathbb{E}\left[\mathbb{E}[\langle W_\star^\top v, x \rangle | Y = y]^2\right] \neq 0 \tag{4.16}$$

where $v \in T_\star$. The left-hand side can be seen as a statistical query of the type $\mathbb{E}[\psi(y)\phi(x)]$ with label pre-processing $\varphi = g_{\text{out}}$. In fact the denoiser g_{out} is the optimal such transformation in the sense that when g_{out} fails to obtain a linear correlation along v, i.e when $v \in T_\star$, then no transformation can.

4.2.2 Easy subspace

On the other hand, if $M = 0$ is a fixed point of eq. (4.9), two qualitatively distinct behaviours may arise: it may be an *unstable* fixed point (repeller) or a *stable* fixed point (attractor) of the dynamics. These scenarios have markedly different implications. If we initialise $\hat{W}0$ such that $||M_0||_{\text{F}} \approx \epsilon > 0$ with ϵ arbitrarily small, then in the case of a repeller we have $M1$ moving away from 0, whereas in the case of an attractor the dynamics drive M_1 back towards 0. In particular, if $M = 0$ is an attractor, a random initialisation does not suffice for AMP to develop a meaningful correlation with $W_\star$ at later times. To determine the stability of a fixed point, one must examine the Jacobian of F around $M = 0$:

$$F(M) \approx \alpha \mathcal{F}(\delta M) + O(\|\delta M\|^2) \tag{4.17}$$

where $\mathcal{F}(\delta M)$ is a linear operator on the cone $\mathcal{S}_r^+$ of p.s.d. matrices of dimension r:

$$\mathcal{F}(M) = \mathbb{E}\left[G(Y)MG(Y)^\top\right]. \tag{4.18}$$

where the expectation is with respect to $Y = g(Z)$ with $Z \sim \mathcal{N}(0, I_r)$, and the operator $\hat{G}$ is given by:

$$\hat{G}(y) = \nabla_\omega g_{\text{out}}(y, 0, I_r) = \mathbb{E}[ZZ^\top - I_r | y] \in \mathbb{R}^{r\times r} \tag{4.19}$$

The stability of the $M = 0$ fixed point is then closely related to the operator norm of $\mathcal{F}$ on the psd cone:

Lemma 4.2.2 ([Tro+25])**.** If $M = 0 \in \mathbb{R}^{r\times r}$ is a fixed point of the state evolution equations. Then, it is an unstable fixed point if and only if $\|\mathcal{F}(M)\|_{\text{F}} > 0$ and $n > \alpha_c d$, where the critical sample complexity α_c, known as the *weak recovery threshold*, is given by:

$$\frac{1}{\alpha_c} = \sup_{\substack{M\in\mathbb{R}^{r\times r}\\ \|M\|_F=1}} \|\mathcal{F}(M)\|_{\text{F}}, \tag{4.20}$$

Moreover, if $\mathcal{F}(M) \neq 0$, there exists at least one $M_\star \neq 0 \in \mathcal{S}_r^+$ achieving the above supremum. While if $\mathcal{F}(M) = 0$, then $M = 0$ is a stable fixed point for any $n = \Theta(d)$.

This implies that for $\alpha > \alpha_c$, iterating the state evolution equations eq. (4.9) will eventually move us away from initialisation, provided we have $M_0 \neq 0$.

We can further characterise the subspace learned in this case, which we denote the *easy subspace*:

Definition 4.2.2 (Easy subspace $E_\star$)**.** Let $H_* \subset \mathrm{span}(W_\star)$ denote a subspace of directions $v \in \mathbb{R}^r$ satisfying

$$\langle v, \hat{G}(Y)v \rangle = 0, \tag{4.21}$$

almost surely over $Y = g(Z)$, $Z \sim \mathcal{N}(0, I_r)$. We define the easy subspace $E_\star$ as the orthogonal complement of $H_\star$, i.e. $\mathrm{span}(W_\star) = E_\star \oplus H_\star$

The main difficulty in proving that AMP successfully learns the easy subspace for $\alpha > \alpha_c$ from the state evolution equations with random initialisation lies in the fact that $M = 0$ is, by construction, a fixed point of the dynamics. A standard way to circumvent this issue is to assume an arbitrarily small (yet $\Theta_d(1)$) initial overlap, and to show that AMP converges to the easy subspace within a time that does not diverge too rapidly as this initial correlation tends to zero. Formally, this is equivalent to assuming access to an arbitrarily noisy version of $W_\star$, referred to as a *side information* channel:

$$S = \sqrt{\lambda} W_\star + \sqrt{1-\lambda} Z \tag{4.22}$$

where $Z \in \mathbb{R}^{r \times d}$ is a random Gaussian matrix with $\mathcal{N}(0,1)$ entries, and $\lambda > 0$ quantifies the amount of side information.

Theorem 4.2.2 ([Tro+25])**.** Let $M_{d,t} = {}^1\!/_d \hat{W}_t W_\star{}^\top$ denote the model-target overlap matrix at any finite time t. Suppose that $T_\star = 0$ and consider the AMP algorithm eq. (4.4). Then, with high probability as $d \to \infty$:

(i) For $\alpha \geq \alpha_c$, $\exists \delta > 0$ such that for sufficiently small λ ,$M_d^t \succ \delta M_\star$ for $t = \Theta(\log {}^1\!/\lambda)$, where $M_\star$ is any of the extremisers defined in eq. (4.20). Furthermore, there exists an $\alpha \geq \alpha_c$ and a $\delta > 0$ such that $M_d^t \succ \delta M_{E_*}$ in $t = \Theta(\log {}^1\!/\lambda)$ iterations, where $M_{E^\star} \in \mathcal{S}_r^+$ spans $E_\star$.

(ii) For $\alpha < \alpha_c$ however, $M_{d,t} = 0$ is asymptotically stable i.e. there exist constants $\lambda' < 1$ and $C > 0$ such that for $\lambda < \lambda'$, $\sup_{t \geq 0} \|M_d^t\| \leq C\sqrt{\lambda}$.

Remark 4.2.2 (Relationship with the literature)**.** The computational weak learnability threshold for single-index models were first established in [MM19; LAL19; Bar+19b]. In this simpler case, the computational and information theoretical thresholds for full-recovery were also studied in [Bar+19b; Mai+20].

Theorem 4.2.2 formalises our intuition about the stability of the uninformed fixed point. It implies that for $\alpha < \alpha_c$, not only does AMP fail to find any pertinent directions, but it also fails to improve on the small side-information. For $\alpha > \alpha_c$, however, AMP will develop a

growing overlap M_t along a non-empty subspace starting with *arbitrarily small* (but $\Theta_d(1)$) side information.

Heuristically, we would like to identify $\lambda = \Theta(1/d)$. Of course, this is not justified, since in the state evolution equations we have already taken the high-dimensional limit, and would imply $\lambda = 0$ in this limit. To make sense of this in the AMP framework requires a non-asymptotic control of the state evolution equations. This is mathematically challenging, and only available for a few simpler problems, see [RV18; LW22; LFW23]. Nevertheless, numerical evidence suggests a similar result holds here.

Conjecture 4.2.1 ([Tro+25])**.** AMP initialized randomly will find a finite overlap with the easy directions for $\alpha > \alpha_c$ in $\mathcal{O}(\log d)$ steps, without side information.

Example 4.2.2. Two examples of easy multi-index functions.

- The monomials $g(z) = \prod_{k=1}^{r} z_k$ with $r > 1$ can always be learned with $\alpha > \alpha_c(r)$ [CM20]. For instance, we have $\alpha_c(2) \approx 0.5937$, $\alpha_c(3) \approx 3.725$, $\alpha_c(4) \approx 4.912$ and $\alpha_c(r) \sim r^{1.2}$ for large r.

- The 2-sparse parity $g(z) = \mathrm{sign}(z_1 z_2)$ is easy, and can be learned with $\alpha_c = \pi^2/4$.

4.2.3 Hard subspace

Finally, there are functions for which $M = 0$ is a stable fixed point for all $\alpha = \Theta_d(1)$, meaning $\alpha_c \to \infty$, i.e. the supremum in the right-hand side of eq. (4.20) is zero. This is what we call a *hard function*. A canonical example of a hard function is the r-sparse parity for $r > 2$ [BKW03]:

$$g(z) = \prod_{k=1}^{r} \mathrm{sign}(z_k) \tag{4.23}$$

These are functions that are intrinsically hard for AMP, and for which a more refined analysis — beyond proportional asymptotics — is required. Two alternative notions of typical-case complexity, extensively studied in the theoretical computer science literature, are the statistical query (SQ) framework [Kea98], previously discussed in section 3.2.3, and the low-degree polynomial (LDP) framework [Bar+19a; HS17]. Gaussian single-index models have been analysed under both frameworks in [Dam+24], which established in each case a sample complexity lower bound of $n = \Omega(d^{k_\star/2})$, where $k_\star$ is the so-called *generative exponent*:

$$k_\star = \min\{m \in \mathbb{Z}_+ : \mathbb{E}[\mathbb{E}[h_m(Z) \mid Y = y]] \neq 0\}. \tag{4.24}$$

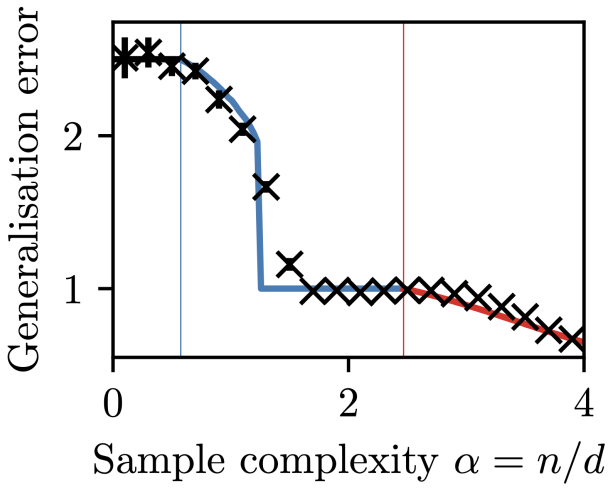

Figure 4.1: Risk as a function of the sample complexity $\alpha = {}^{n}/_{d}$ for Bayes-optimal AMP on the 3-index function $g(z_1, z_2, z_3) = z_1^2 + \text{sign}(z_1 z_2 z_3)$. The solid lines denote the theoretical prediction from state evolution lemma 4.1.1, while crosses denote finite size runs of the AMP algorithm in eq. (4.4) with $d = 500$. The vertical lines denote the weak-recovery thresholds in which the different components are learned. Figure from [Tro+25].

For $k_\star = 2$, the inner conditional expectation coincides with the definition of the function $\hat{G}(y) = \nabla_\omega g_{\text{out}}(y, 0, I_r)$ given in eq. (4.19) with $r = 1$. In this case, the condition is equivalent to the existence of an easy subspace as discussed in section 4.2.2. In [DLB25], this notion was extended to multi-index models, where an analogous lower bound $n = \Omega(d^{k_\star/2})$ was proven in terms of a *leap generative exponent* $\ell_\star$, which again matches the condition for AMP easy subspaces when $\ell_\star = 2$. Taken together, these results demonstrate that LDP, SQ, and AMP provide consistent notions of typical-case hardness in the proportional high-dimensional regime for multi-index models. An interesting open problem is to determine whether AMP-based analyses can be generalised to capture non-linear scaling regimes.

4.3 The grand staircase

The discussion so far has focused on the the analysis of the initial condition $M_0 = 0$. The classification of trivial, hard and easy subspaces do not account for what happens after AMP escapes initialisation. Similarly to this case, this will be governed by the other fixed points of the state evolution equations. As in the analysis of multiple SGD steps discussed in section 3.2.2, the key idea here will be to hierarchically study what subspaces are accessible once the algorithm has learned a given space, emplying a similar idea of conditioning. This requires generalising the notion of trivial, easy and hard subspaces conditionally on a learned space.

Suppose that the estimator $\hat{W}_t$ has developed an overlap along a subspace contained in $\text{span}(W_\star)$, yielding a non-zero overlap $M_t \succ 0$. From this point onwards, the main difference

relative to the previous discussion is that the variable ω in the linear operator $\mathcal{F}$ defined in eq. (4.18) is no longer zero, since it is distributed as $\omega = \sqrt{M}\xi$. This modification alters the span of $\mathcal{F}(M)$ and consequently the stability condition in theorem 4.2.2. In particular, as we show below, learning certain directions may in turn facilitate the learning of larger subspaces. We refer to this phenomenon as the *grand staircase*, in analogy with the staircase behaviour for SGD discussed in Section 3.2.2.

Definition 4.3.1 (Conditionally trivial and easy subspaces)**.** Let $U \subset \mathbb{R}^r$. We define $H_T^\star(U)$ to be the subspace spanned by $v \in U^\perp$ such that

$$\langle v, g_{\text{out}}(Y, \sqrt{M_U}\xi, I_r - \sqrt{M_U})\rangle = 0 \tag{4.25}$$

almost surely over $\xi \sim \mathcal{N}(0, I_r)$ and Y for any $M_U \in \mathcal{S}_r^+$ such that $\text{span}(M_U) = U$. We define the *trivially-coupled subspace* $T_U^\star$ for U as the orthogonal complement of $H_T^\star(U)$.

Analogously, let $H_E^\star(U)$ be the subspace spanned by directions $v \in U^\perp$ such that

$$\langle v, \partial_\omega g_{\text{out}}(Y, \sqrt{M_U}\xi, I_r - \sqrt{M_U})v\rangle = 0 \tag{4.26}$$

almost surely over ξ and Y for any $M_U \in \mathcal{S}_r^+$ such that $\text{span}(M_U) = U$.

When M_U is additionally a fixed point of $\mathcal{F}_M$, one can linearise $\mathcal{F}_M$ along the orthogonal complement of U. We define the *easy-coupled subspace* $E_U^\star$ for U as the orthogonal complement of $H_E^\star(U)$. Next, suppose that $M_U \in \mathcal{S}_r^+$ with $\text{span}(M_U) = U$ is a fixed-point of $\mathcal{F}_M$. Let $\mathcal{F}_{M_U}$ denote the linearization of $F(M)$ along the orthogonal complement $U^\perp$ at $M = M_U$.

We define the *grand staircase threshold* $\alpha_{\text{gst}}(M_U)$ at $M = M_U$ as the conditional weak recovery threshold.

$$\frac{1}{\alpha_{\text{gst}}(M_U)} = \sup_{M^\perp \in U^\perp} \left\|\mathcal{F}_{M_U}(M^\perp)\right\|_{\text{F}} \tag{4.27}$$

The above definitions extend the notions of *trivial* and *easy* subspaces in definitions 4.2.1 and 4.2.2 to the setting where recovery is conditioned on a previously learned subspace U. They identify the directions whose recovery becomes possible once overlap along U has been established. Concretely, after developing an initial overlap on U, the directions in $T_U^\star$ and $E_U^\star$ can be recovered in direct analogy with the recovery of $T^\star$ and $E^\star$ in theorems 4.2.1 and 4.2.2.

Proposition 4.3.1 ([Tro+25],informal)**.** Let $U \subseteq \mathbb{R}^r$ be a subspace such that $E_U^\star$ is non-empty. Consider AMP iterates with the Bayes-optimal choice of denoisers $f_{\text{in}}, g_{\text{out}}$ (c.f. eq. (4.6)) for sufficiently small $\lambda > 0$. Suppose that $M_{d,t} = {}^1\!/_d \hat{W}_t W_\star^\top$ is an approximate fixed point of $F(M)$ in eq. (4.10) such that $M_{d,t} \approx M_U$ where M_U spans U. Then:

- For $\alpha > \alpha_{\text{gst}}(M_U)$, AMP recovers $M_{M_U}^*$ in additional $\Theta(\log {}^1\!/_\lambda)$ steps for arbitrarily

small λ, where $M^*_{M_U}$ denotes any matrix in $\mathcal{S}_r^+$ achieving the supremum in eq. (4.27).

- For $\alpha < \alpha_{\text{gst}}(M_U)$ and sufficiently small λ, AMP remains at the approximate fixed point M_U and fails to gain weak-recovery along $U^\perp$.

A concrete example of a function displaying this phenomenon is a linear combination between *hard* parity function and an *easy* polynomial:

$$g(z_1, z_2, z_3) = z_1^2 + \text{sign}(z_1 z_2 z_3) \tag{4.28}$$

The sign component corresponds to a sparse parity with $r = 3$, which cannot be learned with $n = \Theta(d)$ samples. However, the quadratic term z_1^2 in the function enables weak recovery of the first component, i.e. $U = (1, 0, 0)$, provided that $\alpha > 1/2$. Conditioned on U, the effective multi-index model reduces to $\text{sign}(z_2 z_3)$, which — as discussed in example 4.2.2 — is an *easy* function. This behaviour is illustrated in Figure 4.1: the component z_1 is first recovered at $\alpha_1 \approx 0.575$, and for larger values $\alpha > \alpha_2$, all directions are learned. In other words, knowing z_1 transforms the *hard* three-parity problem into an *easy* two-parity one.

4.4 Spectral methods

An unsatisfactory aspect of the AMP-based results discussed above is the assumption of a warm start in establishing the weak learnability of easy subspaces in theorem 4.2.2. Although such assumptions are standard in the study of phase transitions in mathematical physics, they stand in fundamental contradiction with the notion of weak learnability in computer science: assuming a $\Theta(1)$ correlation with $W_\star$ at initialisation already entails weak recovery of a subspace. Overcoming this assumptions in the context of AMP is a technically challenging problem.

To certify that the weak-recovery threshold in theorem 4.2.2 can be attained without a warm start, we seek an alternative algorithm that both achieves the same threshold from random initialisation and is simpler to analyse. A natural candidate is a *spectral method* whose success or failure follows a BBP-type transition [BAP05].

This perspective — rooted in the Bethe Hessian and non-backtracking operators for sparse graphs [Krz+13; SKZ14] and pioneered in the context of AMP for single-index models in [MM19; LAL19; Mai+22] — relies on a simple construction: weak recovery corresponds to a linear instability of the uninformed fixed point. At the threshold, the spectral radius of the Jacobian crosses 1; beyond it, an expanding mode emerges. Consequently, locating the transition reduces to analysing the first-order linearisation of the dynamics, which is equivalent to power iteration of a data-dependent operator. The spectral properties of this operator can then be characterised via random matrix theory, thereby providing a proof of achievability of the computational thresholds without the need for a warm start.

Linearising the AMP iterations around eq. (4.4) $(\Omega, W) = (0, 0)$ and keeping only the first order terms yield the following linear system

$$\delta\Omega_t = X\delta\hat{W}_t - \mathrm{mat}\Big(\hat{G}\mathrm{vec}(\delta\Omega_{t-1})\Big) \tag{4.29}$$

$$\delta\hat{W}_{t+1} = X^\top \mathrm{mat}\Big(\hat{G}\mathrm{vec}(\delta\Omega_t)\Big). \tag{4.30}$$

where $\mathrm{vec} : \mathbb{R}^{n\times r} \to \mathbb{R}^{nr}$ is the vectorisation operation, $\mathrm{mat} : \mathbb{R}^{nr} \to \mathbb{R}^{n\times r}$ its inverse and $\hat{G} \in \mathbb{R}^{(nr)\times(nr)}$ is a block-diagonal matrix with elements $\hat{G}_{ik,jl} = \delta_{ij}\hat{G}(y_i)_{kl}$, with $\hat{G}(y)$ the Jacobian of the g_{out} denoiser, defined in eq. (4.19):

$$\hat{G}(y) = \nabla_\omega g_{\mathrm{out}}(y, 0, I_r) = \mathbb{E}[ZZ^\top - I_r|y] \in \mathbb{R}^{r\times r} \tag{4.31}$$

From eqs. (4.29) and (4.30), one can proceed in two ways: either solve for $\delta\Omega$ or for $\delta\hat{W}$. Each of these choices yield a different spectral method. For instance, solving for $\delta\Omega_t$ yields a *power iteration* on a data-dependent operator

$$\mathrm{vec}(\delta\Omega_{t+1}) = L\mathrm{vec}(\delta\Omega_t), \tag{4.32}$$

where $L \in \mathbb{R}^{(nr)\times(nr)}$ is the linear operator with entries:

$$L_{(ik),(jl)} \;=\; \big((XX^\top)_{ij} - \delta_{ij}\big)\, G(y_j)_{kl}, \qquad i,j \in [n], \quad k,l \in [r]. \tag{4.33}$$

Note that this operator is not symmetric, and therefore its eigenvalues are complex. For $r = 1$, this is the LAMP method from [Mai+22]. On the other hand, solving for $\delta\hat{W}_t$ yield a second spectral method:

$$\mathrm{vec}(\delta\Omega_{t+1}) = T\mathrm{vec}(\delta\Omega_t) \tag{4.34}$$

where $T \in \mathbb{R}^{(rd)\times(rd)}$ is the linear operator with entries

$$T_{(k\mu),(l\nu)} = \sum_{i=1}^{n} X_{i\mu}X_{i\nu}\left[G(y_i)\left(G(y_i) + I_r\right)\right]^{-1}_{kl}, \qquad \mu,\nu \in [d], \quad k,l \in [r]. \tag{4.35}$$

This is a symmetric operator, with real eigenvalues. For $r = 1$, this corresponds to the Bethe Hessian method from [Mai+22].

These two methods were analysed in detail in [Def+25]. A first set of result is obtained by translating the state evolution of AMP into an analogous state evolution for these methods, yielding a precise characterisation of the spectral edge and of the asymptotic correlation between the leading eigenvectors and the target indices. However, as with AMP, these state-evolution-based results rely on the same warm start assumption.

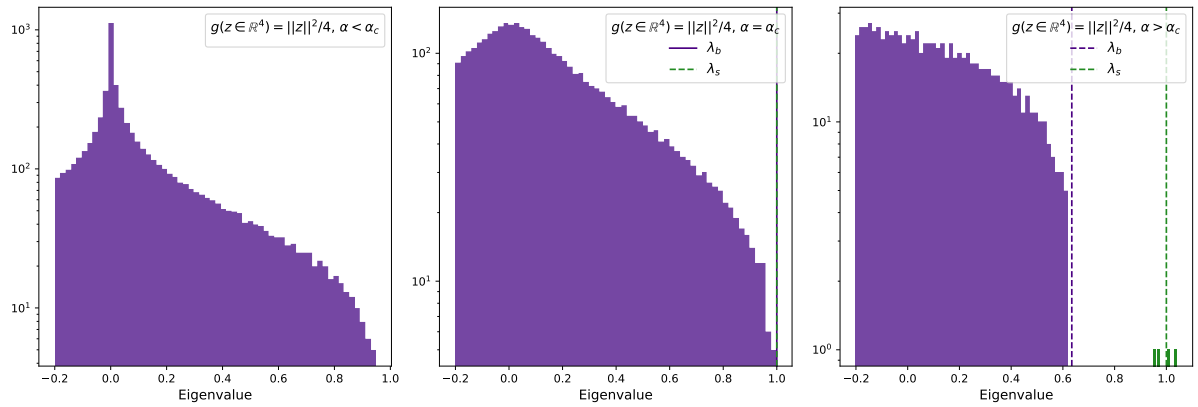

Figure 4.2: Distribution of the eigenvalues of T, $d = 10^4$, for the link function $g(z) = r^{-1}\|z\|^2$, $p = 4$. The critical threshold is $\alpha_c = 2$. The distribution is truncated on the left. **(Left)** $\alpha = 1 < \alpha_c$. **(Center)** $\alpha = \alpha_c$. **(Right)** $\alpha = 6 > \alpha_c$. As predicted by the state evolution, we observe four eigenvalues (in green) separated from the main bulk, centred around $\lambda_s = 1$ (green vertical line). The vertical purple line correspond to the value λ_b provided in Theorem 4.4.1 as a bound for the bulk. Figure from [Def+25].

An alternative analysis which does not rely on side information is possible using tools from random matrix theory. In particular, under an additional technical assumption, one can establish the existence of a BBP transition for the top eigenvalue, occurring precisely at the weak-recovery threshold identified in theorem 4.2.2.

Theorem 4.4.1 ([Def+25])**.** Assume that the matrix $G(y_i) \in \mathbb{R}^{r\times r}$ admits a basis of orthonormal eigenvectors independent of y. Then, in the high-dimensional limit $n, d \to \infty$, $n/d \to \alpha$, above the AMP weak recovery threshold $\alpha > \alpha_c$ defined in theorem 4.2.2, the largest eigenvalue of $T \in \mathbb{R}^{rd\times rd}$ converges to $\lambda_s = 1$. Moreover, the empirical spectral distribution of the pd eigenvalues of T converges weakly almost surely to a density upper bounded by $\lambda_b < 1$.

Remark 4.4.1 (Relationship to the literature)**.** A concurrent proof of theorem 4.4.1 appeared in [KZM25], under similar technical assumptions.

The joint diagonalisability assumption is satisfied by all the examples discussed in this chapter, including the monomials $g(z) = \prod_{k\in[r]} z_k$ and the 2-sparse parity $g(z) = \text{sign}(z_1 z_2)$. However, it fails for certain natural functions, such as $g(z) = z_1/z_2$. Extending the proof to the general case presents a substantially more difficult problem in random matrix theory. We believe, nevertheless, that this is only a technical limitation: both state evolution and numerical evidence strongly support the validity of theorem 4.4.1 without this assumption. Figure 4.2 illustrates the BBP transition on the spectrum of the symmetric method T, as predicted from theorem 4.4.1.

5 | Conclusion

This manuscript summarises a major strand of my research activity over the past seven years, since the completion of my PhD in August 2018. The unifying theme of the results presented here is the question of *adaptivity* in two-layer neural networks, namely how feature learning — adjusting to structure in the data during training — enables efficient generalisation in regimes where data is scarce. The guiding perspective has been that of typical-case analysis, whereby the investigation of synthetic generative models for high-dimensional structured data allows for a precise description of the training procedure and the resulting risk. Our main thread has been the study of multi-index models, a class of functions capturing the inductive bias that many tasks depend on a small number of linear projections combined through a non-linear transformation. This class, which includes both simple functions such as linear models and notably hard ones such as sparse parities, offers a tractable yet flexible laboratory in which the role of adaptativity in two-layer neural networks can be made precise.

The starting point of our discussion was the theory of generalisation for non-adaptive networks, where the features are frozen and only the readout layer is trained. Also known as the random features model, this hypothesis implements a finite width approximation of a kernel method. Chapter 2 presented a detailed analysis of the excess risk for empirical risk minimisation in this problem. Despite its simplicity, this analysis yields important insights into the interplay (or lack thereof) between overparametrisation and generalisation, helping to demystify some of the non-intuitive statistical phenomena first observed in the learning curves of overparametrised neural networks (section 2.2). Building on these formulas, section 2.3 investigated the scaling laws of the risk, uncovering cross-overs between fast and slow scaling regimes, as well as width- and data-driven bottlenecks reminiscent of the neural scaling laws observed in large-scale models. This analysis also provided a sharp lower bound on the width required to achieve the optimal kernel rates in the random features model. We further discussed extensions to other losses and penalties (section 2.5), to deep random features (section 2.7), and explored the extent to which these asymptotic formulas describe real data (section 2.6). Nevertheless, as emphasised in section 2.4, the absence of adaptivity to the underlying data structure in fixed-feature models implies a fundamental limitation of their generalisation capacity under finite samples.

Overcoming this high-dimensional bottleneck requires adaptivity. In Chapter 3, we studied how the network features evolve during the first few steps of training the first-layer weights. Section 3.2 provided a sharp characterisation of the amount of data required for the weights to develop meaningful correlation with the target low-dimensional space, identifying precisely which subspaces are learned in the early stages of training. Section 3.3 analysed how weak-recovery of a subspace translates into generalisation when the second layer is trained in a step-wise procedure. A key outcome is that, in the proportional asymptotic regime, feature learning enables access to a one-dimensional subspace of the target, allowing the network to potentially learn any non-linear function in this direction. Finally, Section 3.4.3 provided an in-depth investigation of this regime, showing that feature learning after a single step can be understood as a BBP transition in the weights, and established via random matrix theory a sharp characterisation of the excess risk, together with upper and lower bounds that provide a precise summary-statistics picture of adaptivity in two-layer networks after one training step. These results show that even limited training suffices to unlock a qualitatively richer generalisation regime, unattainable at initialisation.

Finally, Chapter 4 discussed the fundamental computational limits of learning Gaussian multi-index functions in the proportional regime. Leveraging an optimal message passing scheme, we provided a classification of what what link functions are trivial, easy or hard to weakly learn from initialisation in the high-dimensional regime, and discuss its relationship to the SQ framework and low-degree polynomials methods. section 4.3 discussed how the way subspaces are coupled through the link function can lead to a *grand staircase phenomenon*, a hierarchical learning phenomena where hard directions become accessible once they are coupled to easy directions. To conclude, in section 4.4 we discussed how to remove a warm start assumptions by deriving optimal spectral methods that allow to study weak recovery using random matrix theory techniques. These results serve as a benchmark, allowing us to situate the performance of two-layer networks trained by empirical risk minimisation relative to the fundamental computational limits.

Taken together, the typical-case analysis of high-dimensional two-layer neural networks discussed in this manuscript provides a clear mathematical picture of the mechanisms underlying feature learning, and shows how adaptivity to structure in the data enables efficient generalisation. Yet this is just the tip of the iceberg. The fast-paced evolution of deep learning practice only widens the already considerable gap between our understanding of learning and generalisation in simple, shallow networks and the methods currently deployed in the field. Nevertheless, and despite the view that theory risks obsolescence in an increasingly practice-driven discipline, a solid mathematical foundation remains essential for the sustainable and reliable development of this area. I hope this manuscript makes a strong case that the continued cross-pollination between statistical physics and learning theory, which began in the 1980s, can play a central role in advancing that understanding.

Bibliography

[AAM22] 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.

[AAM23] Emmanuel Abbe, Enric Boix Adsera, and Theodor Misiakiewicz. "Sgd learning on neural networks: leap complexity and saddle-to-saddle dynamics". In: *The Thirty Sixth Annual Conference on Learning Theory*. PMLR. 2023, pp. 2552–2623.

[Abb+21] 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.

[Ado+24] Urte Adomaityte, **Defilippis, Leonardo**, Bruno Loureiro, and Gabriele Sicuro. "High-dimensional robust regression under heavy-tailed data: Asymptotics and universality". In: *Journal of Statistical Mechanics: Theory and Experiment* 2024.11 (2024), p. 114002.

[AFP25] Fabián Aguirre-López, Silvio Franz, and Mauro Pastore. "Random features and polynomial rules". In: *SciPost Physics* 18.1 (2025), p. 039.

[AGJ21] Gerard Ben Arous, Reza Gheissari, and Aukosh Jagannath. "Online stochastic gradient descent on non-convex losses from high-dimensional inference". In: *Journal of Machine Learning Research* 22.106 (2021), pp. 1–51.

[AGS85] Daniel J Amit, Hanoch Gutfreund, and Haim Sompolinsky. "Storing infinite numbers of patterns in a spin-glass model of neural networks". In: *Physical Review Letters* 55.14 (1985), p. 1530.

[And87] Dana Z Anderson. *Neural Information Processing Systems: Proceedings of a conference held in Denver, Colorado, November 1987*. American Institute of Physics, 1987.

[AP20] 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.

[Arn+23] 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". In: *The Thirty Sixth Annual Conference on Learning Theory*. PMLR. 2023, pp. 1199–1227.

[Arn+24a] Luca Arnaboldi, Yatin Dandi, Florent Krzakala, Bruno Loureiro, Luca Pesce, and Ludovic Stephan. "Online Learning and Information Exponents: The Importance of Batch size & Time/Complexity Tradeoffs". In: *International Conference on Machine Learning*. PMLR. 21–27 Jul 2024, pp. 1730–1762.

[Arn+24b] Luca Arnaboldi, Yatin Dandi, Florent Krzakala, Luca Pesce, and Ludovic Stephan. "Repetita iuvant: Data repetition allows sgd to learn high-dimensional multi-index functions". In: *arXiv preprint arXiv:2405.15459* (2024).

[Arn+24c] Luca Arnaboldi, Florent Krzakala, Bruno Loureiro, and Ludovic Stephan. "Escaping mediocrity: how two-layer networks learn hard generalized linear models with SGD". In: *arXiv preprint arXiv:2305.18502* (2024).

[Arn+25] Luca Arnaboldi, Bruno Loureiro, Ludovic Stephan, Florent Krzakala, and Lenka Zdeborova. "Asymptotics of SGD in Sequence-Single Index Models and Single-Layer Attention Networks". In: *arXiv preprint arXiv:2506.02651* (2025).

[Aub+18] 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).

[Aub+19] 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* 32 (2019).

[Aub+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: *Mathematical and Scientific Machine Learning*. PMLR. 2020, pp. 55–73.

[AZP24] Alexander Atanasov, Jacob A Zavatone-Veth, and Cengiz Pehlevan. "Scaling and renormalization in high-dimensional regression". In: *arXiv preprint arXiv:2405.00592* (2024).

[Ba+22] 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* 35 (2022), pp. 37932–37946.

[Bac17a] Francis Bach. “Breaking the curse of dimensionality with convex neural networks”. In: *Journal of Machine Learning Research* 18.19 (2017), pp. 1–53.

[Bac17b] Francis Bach. “On the equivalence between kernel quadrature rules and random feature expansions”. In: *Journal of machine learning research* 18.21 (2017), pp. 1–38.

[Bac24] Francis Bach. “High-dimensional analysis of double descent for linear regression with random projections”. In: *SIAM Journal on Mathematics of Data Science* 6.1 (2024), pp. 26–50.

[Bah+24] Yasaman Bahri, Ethan Dyer, Jared Kaplan, Jaehoon Lee, and Utkarsh Sharma. “Explaining neural scaling laws”. In: *Proceedings of the National Academy of Sciences* 121.27 (2024), e2311878121.

[BAP05] Jinho Baik, Gérard Ben Arous, and Sandrine Péché. “Phase transition of the largest eigenvalue for nonnull complex sample covariance matrices”. In: *The Annals of Probability* 33.5 (2005), pp. 1643–1697. DOI: 10.1214/009117905000000233. URL: https://doi.org/10.1214/009117905000000233.

[BAP24] Blake Bordelon, Alexander Atanasov, and Cengiz Pehlevan. “A dynamical model of neural scaling laws”. In: *arXiv preprint arXiv:2402.01092* (2024).

[Bar+19a] Boaz Barak, Samuel Hopkins, Jonathan Kelner, Pravesh K Kothari, Ankur Moitra, and Aaron Potechin. “A nearly tight sum-of-squares lower bound for the planted clique problem”. In: *SIAM Journal on Computing* 48.2 (2019), pp. 687–735.

[Bar+19b] 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.

[Bar+20] 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.

[Bar19] David Aaron Barmherzig. *The Phase Retrieval Problem: Theory, Algorithms, and Applications*. Stanford University, 2019.

[Bar93] Andrew R Barron. “Universal approximation bounds for superpositions of a sigmoidal function”. In: *IEEE Transactions on Information theory* 39.3 (1993), pp. 930–945.

[BB08] Petros T Boufounos and Richard G Baraniuk. “1-bit compressive sensing”. In: *2008 42nd Annual Conference on Information Sciences and Systems*. IEEE. 2008, pp. 16–21.

[BBV06] Maria-Florina Balcan, Avrim Blum, and Santosh Vempala. "Kernels as features: On kernels, margins, and low-dimensional mappings". In: *Machine Learning* 65.1 (2006), pp. 79–94.

[BC64] George EP Box and David R Cox. "An analysis of transformations". In: *Journal of the Royal Statistical Society Series B: Statistical Methodology* 26.2 (1964), pp. 211–243.

[BCC57] R. Bellman, Rand Corporation, and Karreman Mathematics Research Collection. *Dynamic Programming*. Rand Corporation research study. Princeton University Press, 1957. ISBN: 9780691079516. URL: https://books.google.fr/books?id=wdtoPwAACAAJ.

[BCP20a] 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.

[BCP20b] 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.

[BCR20] Giulio Biroli, Chiara Cammarota, and Federico Ricci-Tersenghi. "How to iron out rough landscapes and get optimal performances: averaged gradient descent and its application to tensor PCA". In: *Journal of Physics A: Mathematical and Theoretical* 53.17 (2020), p. 174003.

[BD81] Peter J Bickel and Kjell A Doksum. "An analysis of transformations revisited". In: *Journal of the american statistical association* 76.374 (1981), pp. 296–311.

[Bel+19] 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.

[Ben22] Yoshua Bengio. *A Deep Learning Journey*. New in ML NeurIPS workshop. 2022. URL: https://nehzux.github.io/NewInML2022NeurIPS/assets/YoshuaBengio-NewInML-NeurIPS-28nov2022.pdf.

[Bie+22] Alberto Bietti, Joan Bruna, Clayton Sanford, and Min Jae Song. "Learning single-index models with shallow neural networks". In: *Advances in neural information processing systems* 35 (2022), pp. 9768–9783.

[BKW03] Avrim Blum, Adam Kalai, and Hal Wasserman. "Noise-tolerant learning, the parity problem, and the statistical query model". In: *J. ACM* 50.4 (July 2003), pp. 506–519. ISSN: 0004-5411. DOI: 10.1145/792538.792543. URL: https://doi.org/10.1145/792538.792543.

[BM02] Peter L Bartlett and Shahar Mendelson. "Rademacher and gaussian complexities: Risk bounds and structural results". In: *Journal of machine learning research* 3.Nov (2002), pp. 463–482.

[BMZ24] Raphaël Berthier, Andrea Montanari, and Kangjie Zhou. "Learning time-scales in two-layers neural networks". In: *Foundations of Computational Mathematics* (2024), pp. 1–84.

[Bol77] L. Boltzmann. *Über die Beziehung zwischen dem zweiten Hauptsatze des mechanischen Wärmetheorie und der Wahrscheinlichkeitsrechnung, respective den Sätzen über das Wärmegleichgewicht*. K.k. Hof- und Staatsdruckerei, 1877.

[BR88] Avrim Blum and Ronald Rivest. "Training a 3-node neural network is NP-complete". In: *Advances in neural information processing systems* 1 (1988).

[CB18] 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).

[CD07] 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.

[Cla+23a] Lucas Clarté, Bruno Loureiro, Florent Krzakala, and Lenka Zdeborová. "Expectation consistency for calibration of neural networks". In: *Uncertainty in Artificial Intelligence*. PMLR. 2023, pp. 443–453.

[Cla+23b] Lucas Clarté, Bruno Loureiro, Florent Krzakala, and Lenka Zdeborová. "On double-descent in uncertainty quantification in overparametrized models". In: *International Conference on Artificial Intelligence and Statistics*. PMLR. 2023, pp. 7089–7125.

[Cla+23c] Lucas Clarté, Bruno Loureiro, Florent Krzakala, and Lenka Zdeborová. "Theoretical characterization of uncertainty in high-dimensional linear classification". In: *Machine Learning: Science and Technology* 4.2 (2023), p. 025029.

[Cla+24] Lucas Clarté, Adrien Vandenbroucque, Guillaume Dalle, Bruno Loureiro, Florent Krzakala, and Lenka Zdeborová. "Analysis of Bootstrap and Subsampling in High-dimensional Regularized Regression". In: *Uncertainty in Artificial Intelligence*. PMLR. 2024, pp. 787–819.

[CM20] Sitan Chen and Raghu Meka. "Learning Polynomials in Few Relevant Dimensions". In: *Proceedings of Thirty Third Conference on Learning Theory*. Ed. by Jacob Abernethy and Shivani Agarwal. Vol. 125. Proceedings of Machine Learning Research. PMLR, Sept. 2020, pp. 1161–1227. URL: https://proceedings.mlr.press/v125/chen20a.html.

[CM24] Chen Cheng and Andrea Montanari. "Dimension free ridge regression". In: *The Annals of Statistics* 52.6 (2024), pp. 2879–2912.

[CMT10] Corinna Cortes, Mehryar Mohri, and Ameet Talwalkar. "On the impact of kernel approximation on learning accuracy". In: *Proceedings of the thirteenth international conference on artificial intelligence and statistics*. JMLR Workshop and Conference Proceedings. 2010, pp. 113–120.

[CMW20] Michael Celentano, Andrea Montanari, and Yuchen Wu. "The estimation error of general first order methods". In: *Conference on Learning Theory*. PMLR. 2020, pp. 1078–1141.

[COB19] Lenaic Chizat, Edouard Oyallon, and Francis Bach. "On lazy training in differentiable programming". In: *Advances in neural information processing systems* 32 (2019).

[Cov65] Thomas M Cover. "Geometrical and statistical properties of systems of linear inequalities with applications in pattern recognition". In: *IEEE transactions on electronic computers* 3 (1965), pp. 326–334.

[CRT06] Emmanuel J Candes, Justin K Romberg, and Terence Tao. "Stable signal recovery from incomplete and inaccurate measurements". In: *Communications on Pure and Applied Mathematics: A Journal Issued by the Courant Institute of Mathematical Sciences* 59.8 (2006), pp. 1207–1223.

[CRW17] Krzysztof M Choromanski, Mark Rowland, and Adrian Weller. "The unreasonable effectiveness of structured random orthogonal embeddings". In: *Advances in neural information processing systems* 30 (2017).

[CSV13] Emmanuel J Candes, Thomas Strohmer, and Vladislav Voroninski. "Phaselift: Exact and stable signal recovery from magnitude measurements via convex programming". In: *Communications on Pure and Applied Mathematics* 66.8 (2013), pp. 1241–1274.

[Cui+21] 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 (2021), pp. 10131–10143.

[Cui+24] Hugo Cui, Luca Pesce, Yatin Dandi, Florent Krzakala, Yue Lu, Lenka Zdeborova, and Bruno Loureiro. "Asymptotics of feature learning in two-layer networks after one gradient-step". In: *International Conference on Machine Learning*. PMLR. 21–27 Jul 2024, pp. 9662–9695.

[Cyb89] George Cybenko. "Approximation by superpositions of a sigmoidal function". In: *Mathematics of control, signals and systems* 2.4 (1989), pp. 303–314.

[Dam+23] Alex Damian, Eshaan Nichani, Rong Ge, and Jason D Lee. "Smoothing the landscape boosts the signal for sgd: Optimal sample complexity for learning single index models". In: *Advances in Neural Information Processing Systems* 36 (2023), pp. 752–784.

[Dam+24] Alex Damian, Loucas Pillaud-Vivien, Jason Lee, and Joan Bruna. "Computational-Statistical Gaps in Gaussian Single-Index Models". In: *Conference on Learning Theory*. PMLR. 2024, pp. 1262–1262.

[Dan+23] Yatin Dandi, Ludovic Stephan, Florent Krzakala, Bruno Loureiro, and Lenka Zdeborová. "Universality laws for gaussian mixtures in generalized linear models". In: *Advances in Neural Information Processing Systems* 36 (2023), pp. 54754–54768.

[Dan+24a] Yatin Dandi, Florent Krzakala, Bruno Loureiro, Luca Pesce, and Ludovic Stephan. "How two-layer neural networks learn, one (giant) step at a time". In: *Journal of Machine Learning Research* 25.349 (2024), pp. 1–65.

[Dan+24b] Yatin Dandi, Ludovic Stephan, Florent Krzakala, Bruno Loureiro, and Lenka Zdeborová. "Universality laws for Gaussian mixtures in generalized linear models*". In: *Journal of Statistical Mechanics: Theory and Experiment* 2024.10 (Oct. 2024), p. 104015. DOI: 10.1088/1742-5468/ad65e7.

[Dan+24c] Yatin Dandi, Emanuele Troiani, Luca Arnaboldi, Luca Pesce, Lenka Zdeborova, and Florent Krzakala. "The Benefits of Reusing Batches for Gradient Descent in Two-Layer Networks: Breaking the Curse of Information and Leap Exponents". In: *International Conference on Machine Learning*. PMLR. 2024, pp. 9991–10016.

[Dan+25] Yatin Dandi, Luca Pesce, Hugo Cui, Florent Krzakala, Yue Lu, and Bruno Loureiro. "A Random Matrix Theory Perspective on the Spectrum of Learned Features and Asymptotic Generalization Capabilities". In: *International Conference on Artificial Intelligence and Statistics*. PMLR. 2025, pp. 2224–2232.

[dAs+20] Stéphane d'Ascoli, Maria Refinetti, Giulio Biroli, and Florent Krzakala. "Double trouble in double descent: Bias and variance (s) in the lazy regime". In: *International Conference on Machine Learning*. PMLR. 2020, pp. 2280–2290.

[DB16] Aymeric Dieuleveut and Francis Bach. "Nonparametric stochastic approximation with large step-sizes". In: *The Annals of Statistics* 44.4 (2016), pp. 1363–1399. DOI: 10.1214/15-AOS1391. URL: https://doi.org/10.1214/15-AOS1391.

[DDV00] Lieven De Lathauwer, Bart De Moor, and Joos Vandewalle. "A multilinear singular value decomposition". In: *SIAM journal on Matrix Analysis and Applications* 21.4 (2000), pp. 1253–1278.

[Def+25] **Defilippis, Leonardo**, Yatin Dandi, Pierre Mergny, Florent Krzakala, and Bruno Loureiro. "Optimal Spectral Transitions in High-Dimensional Multi-Index Models". In: *arXiv preprint arXiv:2502.02545* (2025).

[DLB25] Alex Damian, Jason D Lee, and Joan Bruna. "The Generative Leap: Sharp Sample Complexity for Efficiently Learning Gaussian Multi-Index Models". In: *arXiv preprint arXiv:2506.05500* (2025).

[DLM24] **Defilippis, Leonardo**, Bruno Loureiro, and Theodor Misiakiewicz. "Dimension-free deterministic equivalents and scaling laws for random feature regression". In: *Advances in Neural Information Processing Systems*. Ed. by A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang. Vol. 37. Curran Associates, Inc., 2024, pp. 104630–104693.

[DLS22] Alexandru Damian, Jason Lee, and Mahdi Soltanolkotabi. "Neural networks can learn representations with gradient descent". In: *Conference on Learning Theory*. PMLR. 2022, pp. 5413–5452.

[DW18] 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.

[ES16] Ronen Eldan and Ohad Shamir. "The power of depth for feedforward neural networks". In: *Conference on learning theory*. PMLR. 2016, pp. 907–940.

[Fan+25a] Zhou Fan, Justin Ko, Bruno Loureiro, Yue M Lu, and Yandi Shen. "Dynamical mean-field analysis of adaptive Langevin diffusions: Propagation-of-chaos and convergence of the linear response". In: *arXiv preprint arXiv:2504.15556* (2025).

[Fan+25b] Zhou Fan, Justin Ko, Bruno Loureiro, Yue M Lu, and Yandi Shen. "Dynamical mean-field analysis of adaptive Langevin diffusions: Replica-symmetric fixed point and empirical Bayes". In: *arXiv preprint arXiv:2504.15558* (2025).

[FRS18] Alyson K Fletcher, Sundeep Rangan, and Philip Schniter. "Inference in deep networks in high dimensions". In: *2018 IEEE International Symposium on Information Theory (ISIT)*. IEEE. 2018, pp. 1884–1888.

[FS81] Jerome H Friedman and Werner Stuetzle. "Projection pursuit regression". In: *Journal of the American statistical Association* 76.376 (1981), pp. 817–823.

[Gab+18] 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).

[GB23] Cédric Gerbelot and Raphaël Berthier. “Graph-based approximate message passing iterations”. In: *Information and Inference: A Journal of the IMA* 12.4 (2023), pp. 2562–2628.

[GBD92] Stuart Geman, Elie Bienenstock, and René Doursat. “Neural networks and the bias/variance dilemma”. In: *Neural computation* 4.1 (1992), pp. 1–58.

[GD88] 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.

[GD89] 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.

[Ger+20] 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.

[Ger+24] Federica Gerace, Florent Krzakala, Bruno Loureiro, Ludovic Stephan, and Lenka Zdeborová. “Gaussian universality of perceptrons with random labels”. In: *Physical Review E* 109.3 (2024), p. 034305.

[Gib02] Josiah Willard Gibbs. *Elementary principles in statistical mechanics: developed with especial reference to the rational foundations of thermodynamics*. C. Scribner’s sons, 1902.

[Gol+20] 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.

[Gol+22] 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.

[Goo+14] Ian J Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. “Generative adversarial nets”. In: *Advances in neural information processing systems* 27 (2014).

[Gor85] Yehoram Gordon. “Some inequalities for Gaussian processes and applications”. In: *Israel Journal of Mathematics* 50.4 (1985), pp. 265–289.

[Gre12] W.H. Greub. *Multilinear Algebra*. Grundlehren der mathematischen Wissenschaften. Springer Berlin Heidelberg, 2012. ISBN: 9783662007952.

[Gut+24] Florentin Guth, Brice Ménard, Gaspar Rochette, and Stéphane Mallat. "A rainbow in deep network black boxes". In: *Journal of Machine Learning Research* 25.350 (2024), pp. 1–59.

[Guy16] Isabelle Guyon. *Data Mining History: The Invention of Support Vector Machines*. 2016. URL: https://www.kdnuggets.com/2016/07/guyon-data-mining-history-svm-support-vector-machines.html.

[Has+22] Trevor Hastie, Andrea Montanari, Saharon Rosset, and Ryan J Tibshirani. "Surprises in high-dimensional ridgeless least squares interpolation". In: *Annals of statistics* 50.2 (2022), p. 949.

[Heb49] D.O. Hebb. *The Organization of Behavior: A Neuropsychological Theory*. Wiley and Sons, 1949. ISBN: 978-0-471-36727-7.

[Hes+17] Joel Hestness, Sharan Narang, Newsha Ardalani, Gregory Diamos, Heewoo Jun, Hassan Kianinejad, Md Mostofa Ali Patwary, Yang Yang, and Yanqi Zhou. "Deep learning scaling is predictable, empirically". In: *arXiv preprint arXiv:1712.00409* (2017).

[Hin19] Geoffrey Hinton. *The Deep Learning Revolution*. Royal Society of Edinburgh. 2019. URL: https://rse.org.uk/wp-content/uploads/2021/08/Hintons-Presentation_20190718.pdf.

[HL22] 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.

[HLM24] Hong Hu, Yue M. Lu, and Theodor Misiakiewicz. "Asymptotics of Random Feature Regression Beyond the Linear Scaling Regime". In: *arXiv preprint arXiv:2403.08160* (2024).

[Hoe59] Arthur E Hoerl. "Optimum solution of many variables equations". In: *Chemical Engineering Progress* 55.11 (1959), pp. 69–78.

[Hop82] John J Hopfield. "Neural networks and physical systems with emergent collective computational abilities." In: *Proceedings of the national academy of sciences* 79.8 (1982), pp. 2554–2558.

[HS17] Samuel B Hopkins and David Steurer. "Efficient bayesian estimation from few samples: community detection and related problems". In: *2017 IEEE 58th Annual Symposium on Foundations of Computer Science (FOCS)*. IEEE. 2017, pp. 379–390.

[HSW89] Kurt Hornik, Maxwell Stinchcombe, and Halbert White. "Multilayer feedforward networks are universal approximators". In: *Neural networks* 2.5 (1989), pp. 359–366.

[Jam+13] G. James, D. Witten, T. Hastie, and R. Tibshirani. *An Introduction to Statistical Learning: with Applications in R*. Springer Texts in Statistics. Springer New York, 2013. ISBN: 9781461471387.

[JGH18] Arthur Jacot, Franck Gabriel, and Clément Hongler. "Neural tangent kernel: Convergence and generalization in neural networks". In: *Advances in neural information processing systems* 31 (2018).

[Kap+20] Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. "Scaling laws for neural language models". In: *arXiv preprint arXiv:2001.08361* (2020).

[Kea98] Michael Kearns. "Efficient noise-tolerant learning from statistical queries". In: *Journal of the ACM (JACM)* 45.6 (1998), pp. 983–1006.

[KG00] Vladimir Koltchinskii and Evarist Giné. "Random Matrix Approximation of Spectra of Integral Operators". In: *Bernoulli* 6.1 (2000), pp. 113–167. ISSN: 13507265. (Visited on 09/03/2025).

[KGV83] Scott Kirkpatrick, C Daniel Gelatt Jr, and Mario P Vecchi. "Optimization by simulated annealing". In: *science* 220.4598 (1983), pp. 671–680.

[KH91] Anders Krogh and John Hertz. "A simple weight decay can improve generalization". In: *Advances in neural information processing systems* 4 (1991).

[Kin14] Diederik P Kingma. "Adam: A method for stochastic optimization". In: *arXiv preprint arXiv:1412.6980* (2014).

[KOS04] Adam R Klivans, Ryan O'Donnell, and Rocco A Servedio. "Learning intersections and thresholds of halfspaces". In: *Journal of Computer and System Sciences* 68.4 (2004), pp. 808–840.

[KRT17] Gillat Kol, Ran Raz, and Avishay Tal. "Time-space hardness of learning sparse parities". In: *Proceedings of the 49th Annual ACM SIGACT Symposium on Theory of Computing*. 2017, pp. 1067–1080.

[Krz+13] Florent Krzakala, Cristopher Moore, Elchanan Mossel, Joe Neeman, Allan Sly, Lenka Zdeborová, and Pan Zhang. "Spectral redemption in clustering sparse networks". In: *Proceedings of the National Academy of Sciences* 110.52 (2013), pp. 20935–20940.

[KT85] Scott Kirkpatrick and Gérard Toulouse. "Configuration space analysis of travelling salesman problems". In: *Journal de physique* 46.8 (1985), pp. 1277–1292.

[KZM25] Filip Kovačević, Yihan Zhang, and Marco Mondelli. "Spectral estimators for multi-index models: Precise asymptotics and optimal weak recovery". In: *arXiv preprint arXiv:2502.01583* (2025).

[LAL19] Wangyu Luo, Wael Alghamdi, and Yue M Lu. "Optimal spectral initialization for signal recovery with applications to phase retrieval". In: *IEEE Transactions on Signal Processing* 67.9 (2019), pp. 2347–2356.

[LD21] Licong Lin and Edgar Dobriban. "What causes the test error? going beyond bias-variance via anova". In: *Journal of Machine Learning Research* 22.155 (2021), pp. 1–82.

[LeC19a] Yann LeCun. "Deep learning hardware: Past, present, and future". In: *2019 IEEE International Solid-State Circuits Conference-(ISSCC)*. IEEE. 2019, pp. 12–19.

[LeC19b] Yann LeCun. *Quand la machine apprend: la révolution des neurones artificiels et de l'apprentissage profond*. Odile Jacob, 2019.

[Led01] M. Ledoux. *The Concentration of Measure Phenomenon*. Mathematical surveys and monographs. American Mathematical Society, 2001. ISBN: 9780821837924. URL: https://books.google.fr/books?id=mCX_cWL6rqwC.

[Lee+19] Jaehoon Lee, Lechao Xiao, Samuel Schoenholz, Yasaman Bahri, Roman Novak, Jascha Sohl-Dickstein, and Jeffrey Pennington. "Wide neural networks of any depth evolve as linear models under gradient descent". In: *Advances in neural information processing systems* 32 (2019).

[Lee+24] Jason D Lee, Kazusato Oko, Taiji Suzuki, and Denny Wu. "Neural network learns low-dimensional polynomials with sgd near the information-theoretic limit". In: *Advances in Neural Information Processing Systems* 37 (2024), pp. 58716–58756.

[Les+93] Moshe Leshno, Vladimir Ya Lin, Allan Pinkus, and Shimon Schocken. "Multilayer feedforward networks with a nonpolynomial activation function can approximate any function". In: *Neural networks* 6.6 (1993), pp. 861–867.

[Lév40] M Paul Lévy. "Le mouvement brownien plan". In: *American Journal of Mathematics* 62.1 (1940), pp. 487–550.

[LFW23] Gen Li, Wei Fan, and Yuting Wei. "Approximate message passing from random initialization with applications to Z 2 synchronization". In: *Proceedings of the National Academy of Sciences* 120.31 (2023), e2302930120.

[Lin+24] Licong Lin, Jingfeng Wu, Sham M Kakade, Peter L Bartlett, and Jason D Lee. "Scaling laws in linear regression: Compute, parameters, and data". In: *Advances in Neural Information Processing Systems* 37 (2024), pp. 60556–60606.

[Lou+21a] Bruno Loureiro, Cedric Gerbelot, Hugo Cui, Sebastian Goldt, Florent Krzakala, Marc Mezard, and Lenka Zdeborová. "Learning curves of generic features maps for realistic datasets with a teacher-student model". In: *Advances in Neural Information Processing Systems* 34 (2021), pp. 18137–18151.

[Lou+21b] Bruno Loureiro, Gabriele Sicuro, Cédric 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* 34 (2021), pp. 10144–10157.

[Lou+22] 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: *International conference on machine learning*. PMLR. 2022, pp. 14283–14314.

[Lou25] Bruno Loureiro. "A (very) biased overview of Gaussian multi-index models". In: (2025). URL: https://brloureiro.github.io/assets/pdf/cargese2025.pdf.

[LW22] Gen Li and Yuting Wei. "A non-asymptotic framework for approximate message passing in spiked models". In: *arXiv preprint arXiv:2208.03313* (2022).

[Mai+20] 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.

[Mai+22] Antoine Maillard, Florent Krzakala, Yue M Lu, and Lenka Zdeborová. "Construction of optimal spectral methods in phase retrieval". In: *Mathematical and Scientific Machine Learning*. PMLR. 2022, pp. 693–720.

[Man+17] Andre Manoel, Florent Krzakala, Marc Mézard, and Lenka Zdeborová. "Multi-layer generalized linear estimation". In: *2017 IEEE International Symposium on Information Theory (ISIT)*. IEEE. 2017, pp. 2098–2102.

[Méz24] Marc Mézard. "Spin glass theory and its new challenge: structured disorder". In: *Indian Journal of Physics* 98.11 (2024), pp. 3757–3768.

[MM19] Marco Mondelli and Andrea Montanari. "Fundamental Limits of Weak Recovery with Applications to Phase Retrieval". In: *Foundations of Computational Mathematics* 19.3 (June 2019), pp. 703–773. ISSN: 1615-3383. DOI: 10.1007/s10208-018-9395-y.

[MM21] 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.

[MM22] 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. DOI: https://doi.org/10.1002/cpa.22008.

[MMM22] 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), pp. 3–84.

[MMN18] 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.

[MPM21] Charles H Martin, Tongsu Peng, and Michael W Mahoney. "Predicting trends in the quality of state-of-the-art neural networks without access to training or testing data". In: *Nature Communications* 12.1 (2021), p. 4122.

[MRS22] Alexander Maloney, Daniel A Roberts, and James Sully. "A solvable model of neural scaling laws". In: *arXiv preprint arXiv:2210.16859* (2022).

[MS22] Andrea Montanari and Basil N Saeed. "Universality of empirical risk minimization". In: *Conference on Learning Theory*. PMLR. 2022, pp. 4310–4312.

[MS24] Theodor Misiakiewicz and Basil Saeed. "A non-asymptotic theory of Kernel Ridge Regression: deterministic equivalents, test error, and GCV estimator". In: *arXiv preprint arXiv:2403.08938* (2024).

[MW24] Andrea Montanari and Yuchen Wu. "Statistically optimal firstorder algorithms: a proof via orthogonalization". In: *Information and Inference: A Journal of the IMA* 13.4 (2024), iaae027.

[Nak+21a] Preetum Nakkiran, Gal Kaplun, Yamini Bansal, Tristan Yang, Boaz Barak, and Ilya Sutskever. "Deep double descent: Where bigger models and more data hurt". In: *Journal of Statistical Mechanics: Theory and Experiment* 2021.12 (2021), p. 124003.

[Nak+21b] Preetum Nakkiran, Prayaag Venkat, Sham M. Kakade, and Tengyu Ma. "Optimal Regularization can Mitigate Double Descent". In: *International Conference on Learning Representations*. 2021.

[NW72] 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.

[Opp+90] Manfred Opper, W Kinzel, J Kleinz, and R Nehl. "On the ability of the optimal perceptron to generalise". In: *Journal of Physics A: Mathematical and General* 23.11 (1990), p. L581.

[OTH13] 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.

[Paq+24] Elliot Paquette, Courtney Paquette, Lechao Xiao, and Jeffrey Pennington. "4+ 3 phases of compute-optimal neural scaling laws". In: *Advances in Neural Information Processing Systems* 37 (2024), pp. 16459–16537.

[Ped+11] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. "Scikit-learn: Machine Learning in Python". In: *Journal of Machine Learning Research* 12 (2011), pp. 2825–2830.

[Pes+22] Luca Pesce, Bruno Loureiro, Florent Krzakala, and Lenka Zdeborová. "Subspace clustering in high-dimensions: Phase transitions & Statistical-to-Computational gap". In: *Advances in Neural Information Processing Systems* 35 (2022), pp. 27087–27099.

[Pes+23] Luca Pesce, Florent Krzakala, Bruno Loureiro, and Ludovic Stephan. "Are Gaussian data all you need? The extents and limits of universality in high-dimensional generalized linear estimation". In: *International Conference on Machine Learning*. PMLR. 2023, pp. 27680–27708.

[PRB18] 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* 31 (2018).

[RR07] Ali Rahimi and Benjamin Recht. "Random features for large-scale kernel machines". In: *Advances in neural information processing systems* 20 (2007).

[RR08] Ali Rahimi and Benjamin Recht. "Weighted sums of random kitchen sinks: Replacing minimization with randomization in learning". In: *Advances in neural information processing systems* 21 (2008).

[RR17] Alessandro Rudi and Lorenzo Rosasco. "Generalization properties of learning with random features". In: *Advances in neural information processing systems* 30 (2017).

[RV18] Cynthia Rush and Ramji Venkataramanan. "Finite sample analysis of approximate message passing algorithms". In: *IEEE Transactions on Information Theory* 64.11 (2018), pp. 7264–7286.

[RV22] 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.

[RZH03] Saharon Rosset, Ji Zhu, and Trevor Hastie. "Margin maximizing loss functions". In: *Advances in neural information processing systems* 16 (2003).

[Sch+23] 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*. PMLR. 2023, pp. 30285–30320.

[Sch+24a] Dominik Schröder, Hugo Cui, Daniil Dmitriev, and Bruno Loureiro. "Deterministic equivalent and error universality of deep random features learning*". In: *Journal of Statistical Mechanics: Theory and Experiment* 2024.10 (Oct. 2024), p. 104017. DOI: 10.1088/1742-5468/ad65e2.

[Sch+24b] Dominik Schröder, Daniil Dmitriev, Hugo Cui, and Bruno Loureiro. "Asymptotics of Learning with Deep Structured (Random) Features". In: *International Conference on Machine Learning*. Vol. 235. PMLR, 21–27 Jul 2024, pp. 43862–43894.

[SGL25] Soletskyi, Roman, Marylou Gabrié, and Bruno Loureiro. "A theoretical perspective on mode collapse in variational inference". In: *Machine Learning: Science and Technology* 6.2 (2025), p. 025056.

[SGW20a] 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.

[SGW20b] 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.

[SKZ14] Alaa Saade, Florent Krzakala, and Lenka Zdeborová. "Spectral clustering of graphs with the bethe hessian". In: *Advances in neural information processing systems* 27 (2014).

[Sou+18] Daniel Soudry, Elad Hoffer, Mor Shpigel Nacson, Suriya Gunasekar, and Nathan Srebro. "The implicit bias of gradient descent on separable data". In: *Journal of Machine Learning Research* 19.70 (2018), pp. 1–57.

[Spi+19] 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.

[SS20] 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.

[SS95a] David Saad and Sara Solla. "Dynamics of on-line gradient descent learning for multilayer neural networks". In: *Advances in neural information processing systems* 8 (1995).

[SS95b] David Saad and Sara A Solla. "Exact solution for on-line learning in multilayer neural networks". In: *Physical Review Letters* 74.21 (1995), p. 4337.

[SS95c] David Saad and Sara A Solla. "On-line learning in soft committee machines". In: *Physical Review E* 52.4 (1995), p. 4225.

[SS96] David Saad and Sara Solla. "Learning with noise and regularizers in multilayer neural networks". In: *Advances in Neural Information Processing Systems* 9 (1996).

[Sto13] Mihailo Stojnic. "A framework to characterize performance of lasso algorithms". In: *arXiv preprint arXiv:1303.7291* (2013).

[Tan+25] Kasimir Tanner, Matteo Vilucchio, Bruno Loureiro, and Florent Krzakala. "A High Dimensional Statistical Model for Adversarial Training: Geometry and Trade-Offs". In: *International Conference on Artificial Intelligence and Statistics*. PMLR. 2025, pp. 2530–2538.

[TB23] Alexander Tsigler and Peter L Bartlett. "Benign overfitting in ridge regression". In: *Journal of Machine Learning Research* 24.123 (2023), pp. 1–76.

[Tro+25] Emanuele Troiani, Yatin Dandi, **Defilippis, Leonardo**, Lenka Zdeborova, Bruno Loureiro, and Florent Krzakala. "Fundamental computational limits of weak learnability in high-dimensional multi-index models". In: *International Conference on Artificial Intelligence and Statistics*. PMLR. 2025, pp. 2467–2475.

[TSL00] Joshua B Tenenbaum, Vin de Silva, and John C Langford. "A global geometric framework for nonlinear dimensionality reduction". In: *science* 290.5500 (2000), pp. 2319–2323.

[Tsy08] A.B. Tsybakov. *Introduction to Nonparametric Estimation*. Springer Series in Statistics. Springer New York, 2008. ISBN: 9780387790527.

[TV11] Terence Tao and Van Vu. "Random matrices: Universality of local eigenvalue statistics". In: *Acta Mathematica* 206.1 (2011), pp. 127–204. DOI: 10.1007/s11511-011-0061-3.

[TV23] Yan Shuo Tan and Roman Vershynin. "Online stochastic gradient descent with arbitrary initialization solves non-smooth, non-convex phase retrieval". In: *Journal of Machine Learning Research* 24.58 (2023), pp. 1–47.

[Var66] SR Srinivasa Varadhan. "Asymptotic probabilities and differential equations". In: *Communications on Pure and Applied Mathematics* 19.3 (1966), pp. 261–286.

[Vei+22] Rodrigo Veiga, Ludovic Stephan, Bruno Loureiro, Florent Krzakala, and Lenka Zdeborová. "Phase diagram of stochastic gradient descent in high-dimensional two-layer neural networks". In: *Advances in Neural Information Processing Systems* 35 (2022), pp. 23244–23255.

[Vem10] Santosh S Vempala. "A random-sampling-based algorithm for learning intersections of halfspaces". In: *Journal of the ACM (JACM)* 57.6 (2010), pp. 1–14.

[Ver18] R. Vershynin. *High-Dimensional Probability: An Introduction with Applications in Data Science*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 2018. ISBN: 9781108415194.

[Vil+24] Matteo Vilucchio, Nikolaos Tsilivis, Bruno Loureiro, and Julia Kempe. "On the Geometry of Regularization in Adversarial Training: High-Dimensional Asymptotics and Generalization Bounds". In: *arXiv preprint arXiv:2410.16073* (2024).

[VZL25] Matteo Vilucchio, Lenka Zdeborová, and Bruno Loureiro. "On the existence of consistent adversarial attacks in high-dimensional linear classification". In: *arXiv preprint arXiv:2506.12454* (2025).

[Wan+24] Zhichao Wang, Andrew Engel, Anand D Sarwate, Ioana Dumitriu, and Tony Chiang. "Spectral evolution and invariance in linear-width neural networks". In: *Advances in Neural Information Processing Systems* 36 (2024).

[Wig93] Eugene P Wigner. "Characteristic vectors of bordered matrices with infinite dimensions i". In: *The Collected Works of Eugene Paul Wigner: Part A: The Scientific Papers*. Springer, 1993, pp. 524–540.

[WX20] Denny Wu and Ji Xu. "On the optimal weighted ℓ_2 regularization in overparameterized linear regression". In: *Advances in Neural Information Processing Systems* 33 (2020), pp. 10112–10123.

[Xia+22] Lechao Xiao, Hong Hu, Theodor Misiakiewicz, Yue M Lu, and Jeffrey Pennington. "Precise learning curves and higher-order scaling limits for dot product kernel regression". In: *Thirty-sixth Conference on Neural Information Processing Systems (NeurIPS)*. 2022.

[XRV17] Han Xiao, Kashif Rasul, and Roland Vollgraf. “Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms”. In: *arXiv preprint arXiv:1708.07747* (2017).

[Yan+12] Tianbao Yang, Yu-Feng Li, Mehrdad Mahdavi, Rong Jin, and Zhi-Hua Zhou. “Nyström method vs random fourier features: A theoretical and empirical comparison”. In: *Advances in neural information processing systems* 25 (2012).

[Yua11] Ming Yuan. “On the identifiability of additive index models”. In: *Statist. Sinica* 21.4 (2011), pp. 1901–1911. ISSN: 1017-0405,1996-8507. DOI: 10.5705/ss.2008.117. URL: https://doi.org/10.5705/ss.2008.117.

[Zha+16] Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. “Understanding deep learning requires rethinking generalization”. In: *arXiv preprint arXiv:1611.03530* (2016).

[Zha05] Tong Zhang. “Learning bounds for kernel regression using effective data dimensionality”. In: *Neural computation* 17.9 (2005), pp. 2077–2098.

RÉSUMÉ

Ce manuscrit résume une partie de mon activité de recherche menée au cours des sept dernières années. Le fil conducteur des résultats présentés est la question de *l'adaptativité* dans les réseaux de neurones à deux couches, c'est-à-dire la manière dont l'apprentissage des représentations — l'ajustement à la structure des données au cours de l'entraînement — permet une généralisation efficace dans des régimes où les données sont limitées. L'accent principal est mis sur l'analyse asymptotique de la performance de généralisation typique des réseaux à deux couches entraînés sur des tâches présentant une structure latente de basse dimension, le modèle multi-indices. Le Chapitre 2 présente une caractérisation précise du risque dans un cadre non adaptatif où les représentations du réseau sont figées, aussi connu sous le nom de *random features model*. Le Chapitre 3 étudie comment l'apprentissage des représentations au cours des premières étapes de l'entraînement améliore la généralisation par rapport à ce modèle de référence à caractéristiques figées. Enfin, le Chapitre 4 aborde les limites computationnelles fondamentales de l'apprentissage des fonctions multi-indices dans le régime proportionnel. Pris dans leur ensemble, ces résultats offrent une image mathématique des mécanismes qui sous-tendent l'apprentissage des représentations et montrent comment l'adaptativité à la structure des données permet une généralisation efficace.

ABSTRACT

This manuscript summarises part of my research activity over the past seven years. The unifying theme of the results presented here is the question of *adaptivity* in two-layer neural networks - namely, how feature learning, i.e. adjusting to structure in the data during training, enables efficient generalisation in regimes where data is scarce. The main focus is the typical-case asymptotic analysis of the generalisation performance of two-layer networks trained on tasks with underlying low-dimensional structure, the multi-index model. Chapter 2 discusses a sharp characterisation of the risk in a non-adaptive setting where the network features are fixed, known as the random features model. Chapter 3 investigates how representational learning during the first few steps of training improves generalisation compared with this fixed-feature benchmark. Finally, Chapter 4 addresses the fundamental computational limits of learning multi-index functions in the proportional regime. Taken together, these results provide a clear mathematical picture of the mechanisms underlying feature learning, and demonstrate how adaptivity to structure in the data enables efficient generalisation.