Thesis for The Degree of Doctor of Philosophy

Asymptotic Analysis of Machine Learning Models

Comparison Theorems and Universality

David Bosch

Department of Computer Science and Engineering
Chalmers University of Technology | University of Gothenburg
Gothenburg, Sweden, 2025

Asymptotic Analysis of Machine Learning Models
Comparison Theorems and Universality

David Bosch

ISBN 978-91-8103-287-1
Doktorsavhandlingar vid Chalmers tekniska högskola, Ny serie nr 5745.
ISSN 0346-718X

Department of Computer Science and Engineering
Division of Data Science and AI
Chalmers University of Technology | University of Gothenburg
SE-412 96 Göteborg,
Sweden
Phone: +46(0)31 772 1000

Printed by Chalmers Digitaltryck,
Gothenburg, Sweden 2025.

To my family.

Asymptotic Analysis of Machine Learning Models
Comparison Theorems and Universality

David Bosch

Department of Computer Science and Engineering
Chalmers University of Technology | University of Gothenburg

Abstract

This thesis investigates the asymptotic regime of machine learning models - a regime in which both the number of trainable parameters (model size) and the number of data points grow infinitely at a fixed ratio. Understanding model behavior in this limit provides valuable theoretical insights into model statistics such as training error and generalization error, particularly in high-dimensional settings relevant to contemporary machine learning practice.

The core methodological tools used throughout this work are Gaussian comparison theorems, with a special emphasis on the Convex Gaussian Min-max Theorem (CGMT). These theorems enable the rigorous analysis of complex learning algorithms by comparing them to alternative surrogate problems, which are simpler to analyze. By constructing such asymptotically equivalent optimization problems, we are able to derive characterizations of the models of interest by proxy.

A secondary but significant theme in this thesis is the concept of universality in the asymptotic regime. Universality results demonstrate that many statistical properties of machine learning models are asymptotically governed only by low-order moments (e.g., means and variances) of the data distribution, rather than its full structure. This insight justifies the use of Gaussian surrogate models that match these moments, making them amenable to analysis via Gaussian comparison tools.

Keywords

Asymptotic Analysis, Learning Curves, Convex Gaussian Min-max Theorem, CGMT, Universality, Comparison Theorem

List of Publications

Appended publications

This thesis is based on the following publications:

[**Paper I**] **David Bosch**, Ashkan Panahi, Ayca Özcelikkale, *Double Descent in Feature Selection: Revisiting LASSO and Basis Pursuit*
ICML 2021 Workshop on Overparameterization: Pitfalls & Opportunities.

[**Paper II**] **David Bosch**, Ashkan Panahi, Ayca Özcelikkale, Devdatt Dubhashi, *Random Features Model with General Convex Regularization: A Fine Grained Analysis with Precise Asymptotic Learning Curves*
AISTATS 2023.

[**Paper III**] **David Bosch**, Ashkan Panahi, Babak Hassibi, *Precise Asymptotic Analysis of Deep Random Feature Models*
COLT 2023.

[**Paper IV**] **David Bosch**, Danil Akhtiamov, Reza Ghane, Nithin K Varma, Babak Hassibi, *A Novel Gaussian Min-Max Theorem and its Applications*
Submitted to IEEE Transactions On Information Theory.

[**Paper V**] **David Bosch**, Ashkan Panahi, *A Novel Convex Gaussian Min Max Theorem for Repeated Features*
AISTATS 2025.

Other publications

The following publications were published during my PhD studies, or are currently in submission/under revision. However, they are not appended to this thesis, due to contents overlapping that of appended publications or contents not related to the thesis.

[**a**] Firooz Shahriari-Mehr, **David Bosch**, Ashkan Panahi, *Decentralized Constrained Optimization: Double Averaging and Gradient Projection 2021 60th IEEE Conference on Decision and Control.*

Acknowledgment

I would like to express my gratitude to my PhD supervisor, Ashkan Panahi, for his continued advice and support with my research. Without your guidance the work within would not have been possible. I would also like to thank my co-supervisor Devdatt Dubhashi and my examiner Dag Wedelin, for their support, feedback, and insight.

I would also like to express thanks to the people of the DSAI division. Especially my fellow PhD Students. I would also like to express my thanks to my parents, who have been very supportive during my studies, as well as to my brothers, Nathan and Adam.

Contents

Part I

Introductory Chapters

Chapter 1

Introduction

In the last few years, the usage of machine learning (ML) and artificial intelligence (AI) has exploded to unprecedented levels. Modern machine learning models, such as large language models [1], [2] and modern image diffusion models [3], [4] are stunning works of technical innovation; the theory of ML has, however, struggled to keep up with this rapid growth. Many decisions made by practitioners are guided by experiment and empirical observations, but lack rigorous theoretical underpinnings. This has left us with many questions: why do some models generalize better than others? How can initial conditions and algorithms be tuned for optimal performance? Can we predict expected model behavior without having to go through the expensive process of training? This thesis is an attempt to shed new light on these questions and considers the topic of what happens when models grow large, both in terms of their size as well as the amount of data that is used to train them. We examine this "large" regime through statistics.

Statistical models attempt to model real-world objects and their relationships through a limited number of samples from a population (data). For example, we may with to measure the length of an object and perform several measurements, each slightly different due to methodological error, our statistical model would then attempt to model the true length by means of the observations. A broad class of statistical models are parameterized, where we assume that there exists a set of parameters that explains the randomness of our observations. These parameters must be fit to match our collected data. In the context of neural networks (NN), these parameters are the model weights, and we fit them by minimizing some loss function; the minimization is completed using an algorithm like gradient descent.

This thesis is concerned with ML models in the *asymptotic regime*, which is the regime where both the number of data points (observations) and the number of model parameters grow large. This regime is often seen in practice, for example, modern LLMs are both trained on billions of data points and have billions of weights. Until recently, there was little theoretical analysis in this regime. Classical statistics concerns itself primarily with the underparameterized regime, where the number of data points is much smaller than the number

of model parameters. It was common wisdom that increasing the number of model parameters would result in overfitting; a scenario where the model becomes too specialized to training data and fails to generalize well to unseen data. Interest in this field has increased in large part due to the observation that in practice, many ML models can generalize well even when massively overparameterized [5], sometimes even better than in the underparameterized regime.

There exist a number of approaches to analyze models in this regime, including the replica technique [6]–[9], Approximate Message Passing (AMP) [10], [11], Gaussian widths [12], and well as the focus of this thesis Gaussian comparison theorems [13]. As the name suggests, comparison theorems allow us to analyze models, or more specifically optimization problems over model parameters, by comparing them to alternative optimization problems. These alternative optimization problems should be simpler, or more amenable to analysis, than the original problem. Assuming that certain statistics of the alternative problem converge to definite values, in the asymptotic limit, similar conclusions may be drawn for the original problem. The particular theorem, central to this thesis, is the Convex Gaussian Min Max Theorem (CMGT) [14]–[16], which allows for comparisons of optimizations that contain bilinear Gaussian forms. Two papers in this thesis also extend the CGMT to more general setups, such that greater classes of models can be analyzed.

The CGMT, as well as many other theoretical approaches, assume Gaussianity of the data or features to be applicable. This is, however, not representative of real data. Despite this fact, in high-dimensional space (such as the asymptotic regime), we frequently observe that many statistics of the model begin to concentrate. An example of this phenomenon is the central limit theorem [17], where the sample mean of a set of observations from a wide class of probability distributions converges to a normal distribution. Analogously, in many of the interesting statistics of ML models, such as training and testing loss, will under certain conditions also be asymptotically unchanged if the data or features are replaced by Gaussians which share the same first and second moments. This is called *universality* [18]–[24]. As such, for non-Gaussian random data or features, proving universality and applying the Gaussian surrogate model allows for the analysis by means of comparison theorems (or other techniques).

In paper I of this thesis, we extend the existing analysis of the least absolute shrinkage and selection operator (LASSO) and the closely related basis pursuit (BP) problem, which attempt to minimize the ℓ_1 norm of a solution vector of a square-loss optimization. We derive expressions for the asymptotic generalization error for both problems. Furthermore, we consider weak and strong features and demonstrate their impact on generalization. In paper II, we consider the setup of random features regression (see section 3.2). Here we extend the existing universality results of [23] to additional cases, including ℓ_1 regularization, and then make use of a novel nested application of the CGMT to obtain asymptotic expressions for the training, generalization error, as well as the sparsity of the solution vector. We particularly focus on the case of elastic net regularization [25] and ℓ_1 regularization, which could not be previously analyzed, in the random feature context. In Paper III, we prove a

universality result for deep random features models and then obtain asymptotic expressions for the training and generalization error. In Paper IV, we prove an extension of the CGMT to sums of Gaussian bilinear forms that share one optimization variable. We use this extension to examine models such as multi-source regression and binary classification of Gaussian mixture models, and obtain asymptotic expressions. In Paper V, we prove a further generalization of the CGMT to setups in which features are shared or repeated. This allows us to obtain asymptotic results for greater classes of models, including vector-valued regression and regression with convolution.

The rest of the thesis is structured as follows. In Section 2 we discuss comparison theorems in detail. We describe Slepian's and Gordon's lemma and how these comparisons over Gaussian processes can be extended to comparison theorems between optimization problems. We further discuss the Convex Gaussian Min-Max theorem, which is centrally used in papers I, II, and III. Finally, we also outline a statistical physics framework for proving comparison theorems of this form, a framework we make use of in paper V. In section 3, we discuss how comparison theorems are specifically used to analyze machine learning problems. We also discuss random feature models and universality, which are studied in papers II and III. In section 4 we give a summary of the papers included in this thesis, and in section 5 we give our conclusions and future directions. Part II of this thesis includes appended copies of the discussed papers.

Chapter 2

Comparison Theorems

2.1 What are Comparison Theorems

Comparison theorems, within the context of this thesis, refer to a set of probabilistic tools that allow for the comparison between the moments of functions of random variables. The theorems in this work will always consider a pair of processes, the first being called the *primary*, which is the object of interest, whose properties we wish to analyze. The second process will be called the *alternative*, and is the process that will be compared to the primary. In general, the alternative process will be easier to analyze than the primary, and the expected values of many statistics of the alternative process will bound the same statistics of the primary. This ensures that the alternative is a useful proxy for the analysis of properties of the primary process that we wish to study.

We discuss three related comparison theorems in this section, discussed in the order of historical development. The first theorem is Slepian's lemma, which allows for the comparison between two Gaussian processes whose covariance structure satisfies a set of inequalities. There exists a well-known pair of processes that satisfy these inequalities, which gives rise to the Gaussian Max-Max theorem, as it allows for the comparison between two random Max-Max optimization problems.

The second theorem that we consider was given by Gordon. Similarly to Slepian's result, it allows for the comparison between two Gaussian processes whose covariance structures satisfy some set of inequalities. The same well-known pair of processes that satisfy Slepian's lemma can also be shown to satisfy Gordon's lemma, resulting in the Gaussian Min-Max theorem (GMT), which allows for the comparison between two Min-Max problems.

The third theorem, developed most recently, extends the Gaussian Min-Max theorem. While the GMT only provides a one-sided bound, the Convex Gaussian Min-Max Theorem (CGMT) ensures both an upper and lower bound on the value of the primary process, bounds that in most considered cases become asymptotically tight. To establish this bound, there is an additional cost of requiring convexity/concavity assumptions on the considered pair of

processes.
Finally, we discuss how both the Gaussian Max-Max Theorem and the Gaussian Min-Max Theorem can be derived through a single framework based on results from statistical physics involving Gaussian interpolation.

In addition to the three comparison theorems discussed in this section, two of the papers in this thesis prove further generalizations of the CGMT, which may be used to analyze a broader class of models. Further discussion on the generalizations can be found in sections 4.4 and 4.5.

2.2 Slepian's Lemma and Gaussian Max-Max Theorem

In 1962, Slepian [26] proved the following theorem about Gaussian centered Gaussian processes:

Lemma 1 (Slepian's Lemma [26]) *Let X_i, Y_i for $i = 1, \dots, n$ be two sequences of real-valued centered Gaussian random variables, which satisfy the following inequalities:*

- $\mathbb{E}[X_i^2] = \mathbb{E}[Y_i^2]$ *for all* $i = 1, \dots, n$
- $\mathbb{E}[X_i X_j] \leq E[Y_i Y_j]$ *for all* $i, j \neq i = 1, \dots, n$.

Then for $c_1, c_2, \dots, c_n \in \mathbb{R}$ we have that:

$$\mathbb{P}\left[\bigcup_{i=1}^{n} X_i \geq c_i\right] \geq \mathbb{P}\left[\bigcup_{i=1}^{n} Y_i \geq c_i\right].$$

In the case that $c_1 = c_2 = \cdots c_n = c$ are all equal to some shared c, the union of events $\bigcup_{i=1}^{n} X_i \leq c$ becomes equivalent to the maximization $\max_i X_i \leq c$. As such, Slepian's lemma demonstrates that if there exist two Gaussian processes that have the same variance, but one has greater pairwise covariance, we can find probabilistic bounds on the maximum over the set of all random variables.

One pair of such processes that satisfy these inequalities is the following:

$$p(\boldsymbol{x}, \boldsymbol{y}) = \boldsymbol{x}^T \boldsymbol{G} \boldsymbol{y} + \|\boldsymbol{x}\| \, \|\boldsymbol{y}\| \, \gamma, \tag{1}$$

$$a(\boldsymbol{x}, \boldsymbol{y}) = \|\boldsymbol{x}\| \, \boldsymbol{g}^T \boldsymbol{y} + \|\boldsymbol{y}\| \, \boldsymbol{h}^T \boldsymbol{x}. \tag{2}$$

Here $\boldsymbol{x} \in \mathbb{R}^n, \boldsymbol{y} \in \mathbb{R}^m$ are n and m dimensional vectors respectively and $\|\cdot\|$ denotes the 2-norm, $\boldsymbol{G} \in \mathbb{R}^{n \times m}, \gamma \in \mathbb{R}, \boldsymbol{g} \in \mathbb{R}^m, \boldsymbol{h} \in \mathbb{R}^n$ all have i.i.d standard Gaussian entries and are independent of each other. We can see that for any value of $\boldsymbol{x}, \boldsymbol{y}$, these processes are real valued and centered. It can also readily be shown that $\mathbb{E}_{\boldsymbol{G},\gamma}[p^2(\boldsymbol{x}, \boldsymbol{y})] = \mathbb{E}_{\boldsymbol{g},\boldsymbol{h}}[a^2(\boldsymbol{x}, \boldsymbol{y})]$ and $\mathbb{E}_{\boldsymbol{G},\gamma}[p(\boldsymbol{x}, \boldsymbol{y}) p(\boldsymbol{x}', \boldsymbol{y}')] \leq \mathbb{E}_{\boldsymbol{g},\boldsymbol{h}}[a(\boldsymbol{x}, \boldsymbol{y}) a(\boldsymbol{x}', \boldsymbol{y}')]$ where $\boldsymbol{x} \neq \boldsymbol{x}', \boldsymbol{y} \neq \boldsymbol{y}'$. Using this pair of processes, the following theorem can be proven about comparing these two processes

Theorem 1 (Gaussian Max-Max Theorem) *Let $p(\boldsymbol{x}, \boldsymbol{y})$ and $a(\boldsymbol{x}, \boldsymbol{y})$ be defined in* (1) *and* (2) *respectively. Let $\mathcal{X} \subset \mathbb{R}^n$ and $\mathcal{Y} \subset \mathbb{R}^m$ be two compact sets and let $\psi : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$ be a continuous function. Then, for any $c \in \mathbb{R}$:*

$$\mathbb{P}\left[\max_{\boldsymbol{x}\in\mathcal{X}} \max_{\boldsymbol{y}\in\mathcal{Y}} p(\boldsymbol{x}, \boldsymbol{y}) + \psi(\boldsymbol{x}, \boldsymbol{y}) > c\right] \leq \mathbb{P}\left[\max_{\boldsymbol{x}\in\mathcal{X}} \max_{\boldsymbol{y}\in\mathcal{Y}} a(\boldsymbol{x}, \boldsymbol{y}) + \psi(\boldsymbol{x}, \boldsymbol{y}) > c\right].$$

The Gaussian max-max theorem allows for a probabilistic comparison between two maximization problems. We note that both problems share the same continuous ψ function; this allows us to bound problems of the form $p + \psi$ by instead considering $a + \psi$, which is often easier to analyze.

Slepian developed his lemma to study the maximum singular value of a Gaussian matrix $\boldsymbol{G} \in \mathbb{R}^{m\times n}$. If we denote by $\mathcal{S}_n \subset \mathbb{R}^n$ the unit sphere in n dimensions, we observer that $\sigma_{max}(\boldsymbol{G}) = \max_{\boldsymbol{x}\in\mathcal{S}_n} \|\boldsymbol{G}\boldsymbol{x}\| = \max_{\boldsymbol{x}\in\mathcal{S}_n} \max_{\boldsymbol{y}\in\mathcal{S}_m} \boldsymbol{y}^T\boldsymbol{G}\boldsymbol{x}$, where σ_{max} denotes the maximum singular value. This problem can be expressed as a particular case of the Gaussian Max-Max Theorem where $\psi = 0, \mathcal{X} = \mathcal{S}_n$, and $\mathcal{Y} = \mathcal{S}_m$, the alternative can be solved in this case to find that the expected value of the maximum singular value of $\boldsymbol{G}$ is bounded asymptotically by $\sqrt{n} + \sqrt{m}$, exactly as is predicted by standard results from random matrix theory [27].

2.3 Gordon's Theorem and the Gaussian Min-Max Theorem

In 1985, Gordon [28] developed an extension to Slepian's lemma which may be used to analyze $\min - \max$ optimization problems in contrast to simply maximization problems. Gordon's Comparison Lemma is given as follows:

Lemma 2 (Gordon's Lemma [28]) *Let $X_{i,j}, Y_{i,j}$ for $i = 1, \ldots, n, j = 1, \ldots, m$ be two sequences of real valued centered Gaussian random variables, which satisfy the following inequalities:*

- $\mathbb{E}[X_{i,j}^2] = \mathbb{E}[Y_{i,j}^2]$ *for all* $i = 1, \ldots, n, j = 1, \ldots, m$
- $\mathbb{E}[X_{i,j}X_{i,k}] \leq E[Y_iY_j]$ *for all* $i = 1, \ldots, n,\ j, k = 1, \ldots m.$
- $\mathbb{E}[X_{i,j}X_{l,k}] \geq E[Y_{i,j}Y_{l,k}]$ *for all* $i \neq l = 1, \ldots, n,\ j, k = 1, \ldots m.$

Then for $c_{i,j} \in \mathbb{R}$ for $i = 1, \ldots, n, j = 1, \ldots m$ we have that:

$$\mathbb{P}\left[\bigcap_{i=1}^{n}\bigcup_{j=1}^{m} X_{i,j} \geq c_{i,j}\right] \geq \mathbb{P}\left[\bigcap_{i=1}^{n}\bigcup_{j=1}^{m} Y_{i,j} \geq c_{i,j}\right]. \tag{3}$$

Similarly to the case of Slepian's lemma discussed above when $c_{i,j} = c$ for all i, j then the intersections and unions over the events that $X_{i,j} \geq c$ becomes equivalent to the probability that the event that $\min_i \max_j X_{i,j} \geq c$.

Gordon's Lemma is proven by means of an interpolation between the two Gaussian processes. We can define two covariance matrices $\mathbf{\Gamma}^X \in \mathbb{R}^{nm \times nm}$ and $\mathbf{\Gamma}^Y \in \mathbb{R}^{nm \times nm}$ defined elementwise by:

$$\Gamma^X_{im+j,i'm+j'} = \mathbb{E}[X_{i,j}X_{i',j'}], \qquad \Gamma^Y_{im+j,i'm+j'} = \mathbb{E}[Y_{i,j}Y_{i',j'}],$$

where $i, i' = 1, \ldots, n$ and $j, j' = 1, \ldots, m$. We can then consider a Gaussian Process Z^t_{ij} which has covariance matrix $\mathbf{\Gamma}^t = \sqrt{t}\mathbf{\Gamma}^X + \sqrt{1-t}\mathbf{\Gamma}^Y$ which interpolates between the processes X and Y. We define the function,

$$Q(Z; \mathbf{\Gamma}) = \mathbb{P}\left(\bigcap_{i=1}^{n}\bigcup_{j=1}^{m} Z_{i,j} \geq c_{i,j}\right).$$

It can be shown that

$$\frac{dQ}{dt}(Z; \mathbf{\Gamma}^t) = \sum_{\alpha<\beta}^{nm} \frac{\partial Q}{\partial \Gamma_{\alpha,\beta}}(Z; \mathbf{\Gamma})\bigg|_{\mathbf{\Gamma}=\mathbf{\Gamma}^t} (\Gamma^X_{\alpha,\beta} - \Gamma^Y_{\alpha,\beta}).$$

Gordon then proves that the derivative with respect to $\Gamma_{\alpha,\beta}$ is positive in case 2 of the theorem and negative in case 3 of the theorem. This, combined with the assumption of equality in case 1 of the theorem, shows that $\frac{dQ}{dt} \geq 0$, from which the statement in equation (3) follows trivially.

The same set of processes as described in equations (1), (2) satisfy this set of relations as well. From this, we can obtain the Gaussian Min-Max Theorem (GMT):

Theorem 2 (Gaussian Min-Max Theorem) *Let $p(\boldsymbol{x}, \boldsymbol{y})$ and $a(\boldsymbol{x}, \boldsymbol{y})$ be defined in* (1) *and* (2) *respectively. Let $\mathcal{X} \subset \mathbb{R}^n$ and $\mathcal{Y} \subset \mathbb{R}^m$ be two compact sets and let $\psi : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$ be a continuous function. Then, for any $c \in \mathbb{R}$:*

$$\mathbb{P}\left[\min_{\boldsymbol{x}\in\mathcal{X}}\max_{\boldsymbol{y}\in\mathcal{Y}} p(\boldsymbol{x}, \boldsymbol{y}) + \psi(\boldsymbol{x}, \boldsymbol{y}) \leq c\right] \leq \mathbb{P}\left[\min_{\boldsymbol{x}\in\mathcal{X}}\max_{\boldsymbol{y}\in\mathcal{Y}} a(\boldsymbol{x}, \boldsymbol{y}) + \psi(\boldsymbol{x}, \boldsymbol{y}) \leq c\right].$$

Once again, this comparison theorem allows us to obtain bounds on the object of interest $p + \psi$ by means of analysis of $a + \psi$ in the cases that $a + \psi$ is easier to analyze.

Both Theorem 1 and Theorem 2 operate on continuous sets, while both Slepian's and Gordon's theorems consider finitely indexed Gaussian Processes. This issue is resolved by means of an $\epsilon-$net argument. The compact sets $\mathcal{X}$ and $\mathcal{Y}$ both admit nets of finitely many elements. On these nets, the theorem holds by Gordon's lemma. An additional proof is necessary to show that probabilistic bounds still hold when not on the net; the continuity of ψ (which is equivalent to uniform continuity on compact sets $\mathcal{X} \times \mathcal{Y}$) is necessary to ensure that these deviations cannot be too large.

2.4 The Convex Gaussian Min-Max Theorem

In 2014, [13] proved that under mild additional conditions, the GMT provides not only an upper bound on the values of the primary but also a corresponding

lower bound. Both bounds are based on the same alternative problem. This Convex Gaussian Min Max Theorem (CGMT) also resolved a secondary issue of the GMT, which limited its application to ML-specific problems, that being the presence of the γ term in (1). This term is problematic because it does not up frequently in the problems of interest. The bilinear term naturally appears when analyzing many machine learning problems. We discuss this in more detail below in section 3, the γ term, however, does not. Requiring the presence of the γ term for the theorem to hold, therefore limits the theorem's applicability. The CGMT, therefore, instead considers the following primary process:

$$r(\boldsymbol{x}, \boldsymbol{y}) = \boldsymbol{x}^T \boldsymbol{G} \boldsymbol{y}, \tag{4}$$

which is equivalent to p as given in (1) but without the γ term. The theorem is given as follows:

Theorem 3 (Convex Gaussian Min-Max Theorem) *Let $r(\boldsymbol{x}, \boldsymbol{y})$ and $a(\boldsymbol{x}, \boldsymbol{y})$ be defined in (4) and (2) respectively. Let $\mathcal{X} \subset \mathbb{R}^n$ and $\mathcal{Y} \subset \mathbb{R}^m$ be two compact sets and let $\psi : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$ be a continuous function. Then, for any $c \in \mathbb{R}$:*

$$\mathbb{P}\left[\min_{\boldsymbol{x}\in\mathcal{X}} \max_{\boldsymbol{y}\in\mathcal{Y}} r(\boldsymbol{x}, \boldsymbol{y}) + \psi(\boldsymbol{x}, \boldsymbol{y}) \leq c\right] \leq 2\mathbb{P}\left[\min_{\boldsymbol{x}\in\mathcal{X}} \max_{\boldsymbol{y}\in\mathcal{Y}} a(\boldsymbol{x}, \boldsymbol{y}) + \psi(\boldsymbol{x}, \boldsymbol{y}) \leq c\right].$$

Furthermore, assume that $\mathcal{X}, \mathcal{Y}$ are both convex sets, and that ψ is convex-concave on $\mathcal{X} \times \mathcal{Y}$. Then for any $c_2 \in \mathbb{R}$:

$$\mathbb{P}\left[\min_{\boldsymbol{x}\in\mathcal{X}} \max_{\boldsymbol{y}\in\mathcal{Y}} r(\boldsymbol{x}, \boldsymbol{y}) + \psi(\boldsymbol{x}, \boldsymbol{y}) \geq c_2\right] \leq 2\mathbb{P}\left[\min_{\boldsymbol{x}\in\mathcal{X}} \max_{\boldsymbol{y}\in\mathcal{Y}} a(\boldsymbol{x}, \boldsymbol{y}) + \psi(\boldsymbol{x}, \boldsymbol{y}) \geq c_2\right].$$

One of the most powerful features of the CGMT is that, if the alternative optimization problem concentrates asymptotically, or in other words, if there is some value $A \in \mathbb{R}$ such that for any $\epsilon > 0$,

$$\lim_{n,m\to\infty} \mathbb{P}\left[\left|\min_{\boldsymbol{x}\in\mathcal{X}} \max_{\boldsymbol{y}\in\mathcal{Y}} a(\boldsymbol{x}, \boldsymbol{y}) + \psi(\boldsymbol{x}, \boldsymbol{y}) - A\right| > \epsilon\right] = 0,$$

then the optimal value of $r + \psi$ must concentrate to the same value of A as well. Furthermore, this result can be strengthened. We denote by $(\hat{\boldsymbol{x}}_r, \hat{\boldsymbol{y}}_r)$ the optimal point of the $\min - \max$ problem over $r + \psi$ and similarly denote by $(\hat{\boldsymbol{x}}_a, \hat{\boldsymbol{y}}_a)$ the optimal point of the $\min - \max$ problem over $a + \psi$. Then, for many functions $f(\boldsymbol{x}) : \mathcal{X} \to \mathbb{R}$ or $g(\boldsymbol{y}) : \mathcal{Y} \to \mathbb{R}$, if the value of $f(\hat{\boldsymbol{x}}_a)$ concentrates asymptotically then $f(\hat{\boldsymbol{x}}_r)$ will concentrate on the same value. This observation allows for the analysis of the statistics of the primary optimization problem by proxy through the study of the statistics of the alternative optimization. For example, in the context of machine learning, this machinery allows us to study the generalization error of the primary objective through the generalization error of the alternative. In section 3 below, we discuss how the CGMT can be used as a tool in the asymptotic analysis of machine learning models.

2.5 Statistical Physics Approach to Comparison Theorems

More recently, Stojnic [29] showed that the Gaussian Max-Max Theorem and the Gaussian Min-Max Theorem can be proven together in a single master theorem based on a different proof framework from statistical physics. This is possible due to the relationship between the solutions of optimization problems and their associated Gibbs (or Boltzmann) distribution. We consider a parameter $\beta > 0$, which is traditionally called the "inverse temperature" (as this is the unit it takes in the context of statistical physics). Next, we consider a function $f(\boldsymbol{x})$ which we will assume to be convex, defined on some compact set $\mathcal{X}$, then it can be proven [30], [31] that

$$\min_{\boldsymbol{x}\in\mathcal{X}} f(\boldsymbol{x}) = \lim_{\beta\to\infty} \frac{1}{\beta} \log \left(\underbrace{\int_{\mathcal{X}} e^{-\beta f(x)} d\boldsymbol{x}}_{\equiv Z} \right).$$

This quantity on the right at a finite value of β is called the free energy function, and the quantity Z is called the partition function. A similar expression for the maximum of a concave function can be obtained by flipping the sign of the exponent. The partition function also allows us to construct a probability distribution called the Gibbs distribution, whose probability density function is given by:

$$\frac{1}{Z} e^{-\beta f(\boldsymbol{x})}.$$

The value of Z in this context acts as a normalization constant to ensure that the probabilities sum to 1. If we assume that $f(\boldsymbol{x})$ has a unique optimal value $\hat{\boldsymbol{x}}$ then the Gibbs distribution will concentrate more of the total probability density around this optimal point, such that in the limit of large β it will converge weakly to a delta function around $\hat{\boldsymbol{x}}$. In other words for another well behaved function $g(\boldsymbol{x})$, we have that:

$$\lim_{\beta\to\infty} \frac{1}{Z} \int_{\mathcal{X}} g(\boldsymbol{x}) e^{-\beta f(\boldsymbol{x})} d\boldsymbol{x} = g(\hat{\boldsymbol{x}}).$$

While it may appear initially that we have increased the complexity of the problem, the benefit of this approach is that for finite values of β, our smoothed version of minimization is continuous and differentiable. Intuitively, this allows us to determine properties of our objective at "finite temperature" (i.e. $\beta < \infty$), and then finally take the β limit to obtain our objects of study.

In the case of the Gaussian comparison theorems, we consider a function

$$H_t(\boldsymbol{x}, \boldsymbol{y}) = \sqrt{t} p(\boldsymbol{x}, \boldsymbol{y}) + \sqrt{1-t} a(\boldsymbol{x}, \boldsymbol{y}) + \psi(\boldsymbol{x}, \boldsymbol{y}),$$

where $t \in [0, 1]$ p and a are defined in (1) and (2) respectively and ψ is a continuous function of interest. The function H_t expresses an interpolation

from our primary to the alternative objective. We then consider the following function

$$\xi(\mathcal{X}, \mathcal{Y}, \beta, s, t) = \mathbb{E}_{\boldsymbol{G},\gamma,\boldsymbol{g},\boldsymbol{h}} \frac{1}{\beta} \log \left(\int_{\mathcal{X}} \left(\int_{\mathcal{Y}} e^{\beta H_t(\boldsymbol{x},\boldsymbol{y})} d\boldsymbol{y} \right)^s d\boldsymbol{x} \right).$$

Here $s \in \{-1, 1\}$ is an additional parameter that defines if the problem is $\max - \max$ or $\min - \max$. Similarly to above, it can be shown under mild conditions that:

$$\lim_{\beta \to \infty} \xi(\mathcal{X}, \mathcal{Y}, \beta, s, t) = \mathbb{E}_{\boldsymbol{G},\gamma,\boldsymbol{g},\boldsymbol{h}} \max_{\boldsymbol{x} \in \mathcal{X}} s \max_{\boldsymbol{y} \in \mathcal{Y}} H_t(\boldsymbol{x}, \boldsymbol{y}).$$

We can note that by setting s to the value of -1 we can obtain a $\min - \max$ optimization, as required for the GMT. The bounds between the expected values of the primary and alternative optimizations can then be obtained by studying the value of $\frac{d\xi}{dt}$, in other words, how the function changes as we interpolate from the primary to the alternative. We make use of a similar statistical physics framework in paper V for a CGMT extension.

Chapter 3

Comparison Theorems in Machine learning

Within the context of machine learning, we assume that a dataset is a set of random samples drawn from some distribution $\mathcal{D}$, which describes the likelihood of observing a particular sample. Our goal is then to train a model that performs a certain task based on the data in the dataset. For example, in supervised learning, we attempt to predict a label corresponding to each observation. As the dataset is a random variable, with samples drawn from the underlying data distribution $\mathcal{D}$, our model is also random and dependent upon which particular set of data points is drawn. In machine learning, we are generally interested in particular statistics of this model, most saliently the expected training error and generalization error with respect to the data distribution $\mathcal{D}$. However, we are also interested in additional statistics such as the expected model sparsity or quantization.

Throughout this thesis, we will focus on the empirical risk minimization framework in the supervised learning case. In this framework, we consider a dataset $D = \{ (\boldsymbol{x}_i, y_i) \}_{i=1}^n$ of n data points and labels. The model of interest $f_{\boldsymbol{\theta}}(\boldsymbol{x})$ will be parameterized by a set of parameters $\boldsymbol{\theta}$, and the goal will be to choose a particular choice of model parameters which minimizes a given risk function. In general, we will consider a loss function $\ell(\cdot, \cdot)$ which will measure how well the model can predict the label y_i from the data point $\boldsymbol{x}_i$, as well as a regularization function $R(\boldsymbol{\theta})$ which will impose some additional penalty on the undesired model parameters. In general, our optimization problem will take the form:

$$\min_{\boldsymbol{\theta}} \frac{1}{n} \sum_{i=1}^{n} \ell(f_{\boldsymbol{\theta}}(\boldsymbol{x}_i), y_i) + R(\boldsymbol{\theta}). \tag{1}$$

We note that in theory we would like to minimize the parameters with respect to the dataset distribution $\mathcal{D}$, and consider $\mathbb{E}_{(\boldsymbol{x},y)\sim\mathcal{D}}[\ell(f_{\boldsymbol{\theta}}(\boldsymbol{x}), y)]$ instead of the sum in (1). However, in practice, the distribution $\mathcal{D}$ is often either unknown or computationally intractable. As such, we limit ourselves to the empirical distribution over the n collected samples and use this empirical distribution to

compute the expected value of the loss. 1 To apply the comparison theorem discussed in section 2 to problems in the form of (1), in general, two conditions need to hold. Firstly, we must be in the student-teacher framework. In this framework, the labels y_i are generated by means of a known function of $\boldsymbol{x}_i$. It is generally assumed that there exists some set of teacher parameters $\boldsymbol{\theta}^*$ that characterize how the labels are generated. A common case studied in the literature is $y_i = g(\boldsymbol{\theta}^{*T}\boldsymbol{x}_i)$, where $g : \mathbb{R} \to \mathbb{R}$ is a suitable function, such as the sign function (for binary classification) or label corruption in the form of additive noise.

The second condition is Gaussianity. Our comparison theorems of interest rely on the objective function being Gaussian. In simple cases, such as linear models, this requires assuming that $\boldsymbol{x}_i$ are drawn from a Gaussian distribution. For more complex models, we need to prove that our objective can be approximated asymptotically by a Gaussian model. This type of analysis is referred to as universality and is discussed in more detail below. The chain of analysis for a given ML problem then becomes first proving universality and finding an asymptotically equivalent Gaussian model, and then subsequently analyzing that Gaussian model, such as by means of a comparison theorem, studies such as [32]–[34] follow this approach.

This chapter presents few well-known studies carried out by universality and Gaussian comparison results. In section 3.1 we will discuss linear models and how these may be broached by the CGMT. In section 3.2 we will consider random feature models, which can be considered as proxies for 2-layer fully connected neural networks, as well as deep random features. Both shallow and deep random feature models are non-Gaussian, but can be shown to be asymptotically similar to Gaussian models by means of universality arguments.

3.1 Linear Models

In the case of linear models, our ERM equation takes the form:

$$\min_{\boldsymbol{\theta}} \frac{1}{n}\sum_{i=1}^{n} \ell(\boldsymbol{x}_i^T\boldsymbol{\theta}, y_i) + R(\boldsymbol{\theta}).$$

By means of the convex conjugate of the loss function ℓ with respect to the first argument, this model can be expressed in a form amenable to the CGMT:

$$\min_{\boldsymbol{\theta}} \max_{\boldsymbol{z}} \frac{1}{n}\boldsymbol{z}^T\boldsymbol{X}\boldsymbol{\theta} - \frac{1}{n}\sum_{i=1}^{n} \ell^*(z_i, y_i) + R(\boldsymbol{\theta}).$$

where $\boldsymbol{z} = [z_1, z_2, \ldots, z_n]^T$ is the dual variable introduced in the convex conjugate ℓ^* of ℓ and $\boldsymbol{X}$ is the matrix of data points with rows $\boldsymbol{x}_i$. If $\boldsymbol{X}$ is a Gaussian matrix (i.e., the data is sampled from a Gaussian distribution), problems of this form can be analyzed by the CGMT. There exist results in the literature that analyze different choices for the loss and regularization function [19], [35], [36] and how these choices impact the performance of the model.

3.2 Random Features

Random feature models [37], which are the subject of analysis of a number of works in this thesis, are two-layer neural networks in which the weights from the input layer to the hidden layer are not trained. The randomly initialized weights, in essence, map the input features into a different space using a random embedding. For input features $\boldsymbol{x}_i \in \mathbb{R}^d$, the model is given by $f_{\boldsymbol{\theta}}(\boldsymbol{x}_i) = \boldsymbol{\theta}^T \sigma(\boldsymbol{W}\boldsymbol{x}_i)$, where $\boldsymbol{W} \in \mathbb{R}^{m\times d}$ are a set of i.i.d Gaussian random weights, and σ is a non-linear activation function. The random features model has been studied extensively [36]–[39] as a proxy for neural networks.

The random feature model is, however, not Gaussian and therefore is not directly amenable to applications of the CGMT. Despite this fact, when the activation function σ is odd, it has been proven [23] that the empirical risk minimization optimization problem with respect to the RF model is asymptotically equivalent to a Gaussian model which shares the same first and second moments. In practice this result can be strengthened further by approximating the odd activation $\sigma(\boldsymbol{W}\boldsymbol{x}_i) \approx \rho_1 \boldsymbol{W}\boldsymbol{x}_i + \rho_2 \boldsymbol{z}$, where $\rho_1 = \mathbb{E}_{x\sim\mathcal{N}(0,1)}[\sigma'(x)]$, $\rho_* = \sqrt{\mathbb{E}_{x\sim\mathcal{N}(0,1)}[\sigma^2(x)] - \rho_1^2}$ and $\boldsymbol{z}$ is a standard normal vector. This approximation can be obtained by means of a truncated Hermite polynomial expansion of the activation function. We note that this approximation results in a linear model, which may be analyzed by means of the CGMT. Demonstrating that a model is asymptotically equivalent to a Gaussian model with respect to a set of test functions is called *universality*. Papers II and III in this thesis prove universality results.

Universality in this work, as well as in many others [19], [23], [24], [40], is generally proven by means of Lindeberg's method. The principle of Lindeberg's method is to step by step replace parts of the feature matrix by a Gaussian surrogate, and then to bound the difference in the value of the test functions under this change. For example, let $\boldsymbol{X} \in \mathbb{R}^{n\times m}$ be a data matrix of n data points of dimension m, where each data point $\boldsymbol{x}_i \sim P$ for some probability distribution P with mean $\boldsymbol{\mu}$ and covariance $\boldsymbol{\Sigma}$. Furthermore, let $T(\boldsymbol{X})$ be some function of this data, for example, the training loss of a model trained on this data.

To apply Lindeberg's argument, we consider another set of data points $\tilde{\boldsymbol{x}}_i \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, and consider a set of intermediate matrices

$$\boldsymbol{X}_r = \left[\boldsymbol{x}_1\ \boldsymbol{x}_2\ \cdots \boldsymbol{x}_{r-1}\ \tilde{\boldsymbol{x}}_r\ \cdots \tilde{\boldsymbol{x}}_n\right]^T, \qquad r = 0, \ldots, n.$$

We observe that $\boldsymbol{X}_0 = \boldsymbol{X}$ and $\boldsymbol{X}_n = \tilde{\boldsymbol{X}}$. Now, we note that

$$\left|T(\boldsymbol{X}) - T(\tilde{\boldsymbol{X}})\right| = \left|\sum_{r=0}^{n-1} T(\boldsymbol{X}_r) - T(\boldsymbol{X}_{r+1})\right| \leq \sum_{r=0}^{n-1} \left|T(\boldsymbol{X}_r) - T(\boldsymbol{X}_{r+1})\right|,$$

where the first equality is obtained by a telescoping sum, and the second by the triangle inequality. If we demonstrate that $|T(\boldsymbol{X}_r) - T(\boldsymbol{X}_{r+1})| \leq \frac{C}{n^{3/2}}$, for

some constant $C > 0$, we will be able to show that:

$$\left|T(\boldsymbol{X}) - T(\tilde{\boldsymbol{X}})\right| \leq \sum_{r=0}^{n-1} \frac{C}{n^{3/2}} \leq \frac{C}{\sqrt{n}} \xrightarrow{n\to\infty} 0.$$

As such, by bounding the difference between two successive terms of the replacement, we can prove that the statistics T of $\boldsymbol{X}$ and a Gaussian surrogate that matches the first and second moments are asymptotically equivalent.

The random feature model can be naturally generalized to a deep random feature model. In the deep RF model, we consider an $L-$layer deep neural network where only the final output layer is trained. More formally we consider L matrices $W^{(1)}, \ldots, W^{(L)}$, each of dimensions $\boldsymbol{W}^{(l)} \in \mathbb{R}^{m_l \times m_{l-1}}$. Then, we recursively define a sequence of models, where $\boldsymbol{z}_i^{(0)} = \boldsymbol{x}_i \in \mathbb{R}^{m_0}$ and $\boldsymbol{z}^{(l+1)} = \sigma(\boldsymbol{W}^{(l+1)}\boldsymbol{z}^{(l)})$, with σ being a a non-linear activation function applied elementwise. The final model of interest is then given by $f_{\boldsymbol{\theta}}(\boldsymbol{x}_i) = \boldsymbol{\theta}^T \boldsymbol{z}^{(L)}$. Similarly to the shallow random features case, we can prove universality results for this model. We prove this in paper III.

3.3 Limitations of Existing Comparison Theorems

Both the linear models and random feature models discussed above assume that the labels are 1-dimensional objects. While this allows for the analysis of single regression problems as well as binary classification, single-dimensional output cannot express the behavior of more complex models. The central difficulty is that an application of the CGMT necessarily requires the bilinear form $\boldsymbol{z}^T \boldsymbol{X} \boldsymbol{\theta}$, where $\boldsymbol{X}$ is an i.i.d. Gaussian matrix. For more complex models, attempting to express the model as a bilinear form will result in the repetition of Gaussian elements. For example, the natural extension to the bilinear form is $\mathrm{Tr}[\boldsymbol{Z}^T \boldsymbol{X} \boldsymbol{\Theta}]$ where $\boldsymbol{Z}, \boldsymbol{\Theta}$ are now matrices. This form shows up in the analysis of multiclass models (of which some cases have been analyzed [41]), however linearizing this form results in $(\mathrm{vec}(\boldsymbol{Z}))^T (\boldsymbol{I} \otimes \boldsymbol{X})(\mathrm{vec}(\boldsymbol{\Theta}))$, where $\mathrm{vec}(\cdot)$ is the vectorization operation that stacks all columns of a matrix into a single vector, and $\otimes$ denotes the Kronecker product. The matrix $(\boldsymbol{I} \otimes \boldsymbol{X})$ is clearly not i.i.d. Gaussian, as such the CGMT cannot naively be used to analyze this problem. Furthermore, other models of interest, such as convolutional neural networks and time series, also cannot be linearized into the requisite bilinear form. This limitation of the CGMT is in part addressed by our work in paper V, where we prove a generalization of the CGMT. Our more general form can handle problems such as certain convolutional filters and time series, and we discuss some examples in the paper.

Chapter 4

Summary of the Included Papers

4.1 Paper I

In this paper, we consider the Least Absolute Shrinkage and Selection Operator (LASSO) and the closely related Basis Pursuit optimization problem. For a given data set $\{(\boldsymbol{x}_i, y_i) \in \mathbb{R}^m \times \mathbb{R}\}_{i=1}^n$, the LASSO problem is given by

$$\min_{\boldsymbol{\theta}\in\mathbb{R}^m} \frac{1}{2n}\sum_{i=1}^{n}(y_i - \boldsymbol{\theta}^T\boldsymbol{x}_i)^2 + \frac{\lambda}{\sqrt{m}}\left\|\boldsymbol{\theta}\right\|_1, \tag{1}$$

where $\lambda \geq 0$ is the parameter that controls regularization strength. The basis pursuit problem is defined in the limit of $\lambda \to 0$, when $m > n$, as

$$\begin{aligned} &\min_{\boldsymbol{\theta}\in\mathbb{R}^m} \left\|\boldsymbol{\theta}\right\|_1 \\ &\text{s.t.} \\ y_i = \boldsymbol{\theta}^T\boldsymbol{x}_i \qquad &i = 1, \ldots, n. \end{aligned}$$

We consider the case where $\boldsymbol{x}_i$ are normally distributed with zero mean and covariance matrix $\boldsymbol{R}$, and where the labels are given by

$$y_1 = \frac{1}{\sqrt{m}}\boldsymbol{x}_i^T\boldsymbol{\theta}^* + \nu_i, \qquad i = 1, \ldots, n.$$

Here, ν_i is i.i.d Gaussian noise with variance $\sigma_{\boldsymbol{\nu}}^2$. It is specifically assumed that $\boldsymbol{\theta}^*$ is nearly sparse. By this, we mean that a small subset A of indices of $\boldsymbol{\theta}^*$ exists such that $\boldsymbol{\theta}_A^*$, i.e. $\boldsymbol{\theta}^*$ restricted to the indices in A, has values much larger than $\boldsymbol{\theta}_{A^c}^*$. For this problem, the generalization error, as a function of the regularization parameter, can be expressed as

$$\begin{aligned} \mathcal{E}_{gen}(\lambda) = \mathbb{E}_{\boldsymbol{x},y}(y - \hat{\boldsymbol{\theta}}_\lambda^T\boldsymbol{x})^2 - \mathbb{E}(y - \boldsymbol{x}^T\boldsymbol{\theta}^*)^2 \\ = \boldsymbol{e}_\lambda^T\boldsymbol{R}\boldsymbol{e}_\lambda, \end{aligned}$$

where $\hat{\boldsymbol{\theta}}_\lambda$ is the solution to (1) for a given value of regularization strength $\lambda \geq 0$, and $\boldsymbol{e}_\lambda = \hat{\boldsymbol{\theta}}_\lambda - \boldsymbol{\theta}^*$ is the error vector.

In Theorem 1 of this paper, we demonstrate, by means of the CGMT, that the optimization problem (1) can asymptotically be expressed as

$$\min_{\boldsymbol{e}} \frac{1}{2}\boldsymbol{e}^T\boldsymbol{R}\boldsymbol{e} + \frac{q}{\sqrt{n}}\boldsymbol{e}^T\boldsymbol{h} + \frac{q\lambda}{\beta\sqrt{m}}\left\|\frac{\boldsymbol{\theta}^*}{\sqrt{m}} + \boldsymbol{e}\right\|_1,$$

where $\boldsymbol{h} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_m)$ and β, q are constants satisfying:

$$q^2 = \boldsymbol{e}^T\boldsymbol{R}\boldsymbol{e} + \sigma_{\boldsymbol{\nu}}^2, \qquad \beta = q + \frac{1}{n}\boldsymbol{e}^T\boldsymbol{h}.$$

In this paper, we consider the case that $\boldsymbol{R}$ is diagonal, with entries r_j for $j = 1, \ldots, m$. The values of r_j give the strength of the given features. We consider combinations of strong and weak features, such that for some set $r_1 = r_2 = \cdots r_{m_1} = R$ for some larger value R and for the remainder $r_{m_1+1} = \cdots = r_m = r$ where $R > r$. This gives us m_1 strong features and $m - m_1$ weak features. Theoretically, we determine an expression for the generalization error in terms of these weak features given by:

$$\mathcal{E}_{gen}(\lambda) = \frac{1}{m}\sum_{j=1}^{m} r_j \mathbb{E}_\phi\left[\mathcal{T}_{\frac{\lambda q}{\beta r_j}}\left(\theta_j^* + \frac{q\phi}{\sqrt{r_j\gamma}} - \theta_j^*\right)^2\right],$$

where $\gamma = \frac{n}{m}$, ϕ is a standard Gaussian random variable, and $\mathcal{T}$ is a soft thresholding operator, defined as

$$\mathcal{T}_a(b) = \begin{cases} b - a & b > a \\ b + a & b < -a \\ 0 & |b| \leq a \end{cases}.$$

We also give theoretical expressions for the predicted sparsity of the solution vector. We experimentally verify the claims made and explore the impact of the regularization strength and strength of the features, and the generalization and sparsity of the solution vectors.

4.2 Paper II

In this paper, we consider the case of random features regression as described in section 3.2. We make two contributions to this problem. The first is an extension of the results for universality, and the second is a novel nested application of the CGMT that allows us to express the original optimization as a 4-dimensional scalar optimization. Previous results involved optimizations of m-dimensional proximal operators, which were, in many cases, intractable.

For universality, we extended the existing results in [23]. [23] had given universality results for random feature models, under a number of assumptions. The main assumption we improve upon is the necessity of the regularization

function to be strongly convex, and to have a third derivative that is uniformly bounded over all $\mathbb{R}$.

We extend this result in two ways. Firstly, we deal with regularization functions that are not differentiable at all points. We prove that if we can construct a sequence of functions $R_k(\boldsymbol{\theta})$ converging uniformly to $R(\boldsymbol{\theta})$ as $k \to \infty$, and if all of those functions R_k are thrice differentiable, then universality holds for $R(\boldsymbol{\theta})$ as well. This allows us to prove universality for the elastic net regularization function:

$$R(\boldsymbol{\theta}) = \frac{\alpha}{2} \|\boldsymbol{\theta}\|^2 + \lambda \|\boldsymbol{\theta}\|_1 .$$

Here α, λ are two regularization strength parameters. Secondly, we extend the universality results to ℓ_1 regularization. To prove this, we make use of a similar technique as [19] and consider elastic net regularization at very small values of α. We demonstrate that with high probability the feature matrix $\boldsymbol{X}$ (as described in section 3.2) satisfies the restricted isometry property [42]. We make use of this to show that the difference in solution vector between the cases of α small and $\alpha = 0$ is negligible, and therefore the solution is stable, despite the lack of strong convexity. We make use of this argument to demonstrate the universality of ℓ_1 regularization.

We then consider the Gaussian equivalent random feature problem for the case of generic strongly convex regularization or ℓ_1 regularization, and find an alternative optimization problem by means of a nested CGMT argument. We note that there are two sources of randomness in the RF problem, the randomness of the Gaussian input data $\boldsymbol{z}$ and secondly that of the Gaussian weight matrix $\boldsymbol{W}$. The two applications of the CGMT are applied to both sources of randomness, successively. The resulting alternative optimization problem is given by:

$$\max_{\beta>0} \min_{q>0} \max_{\xi>0} \min_{t>0} \frac{1}{m} \mathbb{E}\left[\mathcal{M}_{\frac{1}{2c_1} R}\left(\boldsymbol{\theta}^* - \frac{c_2\sqrt{\gamma}}{2c_1}\boldsymbol{\phi}\right)\right] - \frac{c_2^2\gamma}{4c_1} + \frac{\xi t}{2} + \frac{\beta q}{2} + \frac{\beta\sigma_{\boldsymbol{\nu}}^2}{2q} + \frac{\xi\beta^2}{2t\eta} - \frac{\beta\xi^2}{2q} - \frac{q\beta}{2\eta} - \frac{\beta^2}{2},$$

where $\boldsymbol{\phi}$ is a standard Gaussian vector, c_1, c_2 are functions of β, q, ξ, and t, and $\mathcal{M}_{\frac{1}{2c_1} R}$ is the Moreau envelope over the function R. The Moreau envelope with step size τ over a function f is given by:

$$\mathcal{M}_{\tau f}(\boldsymbol{y}) = \min_{\boldsymbol{x}} \frac{1}{2\tau} \|\boldsymbol{x} - \boldsymbol{y}\|^2 + f(\boldsymbol{x}).$$

In the case that the regularization function is separable, in many cases, the Moreau envelope can be solved explicitly, which allows us to obtain a four-dimensional scalar optimization function that converges to the training error of random feature regression. We can similarly obtain an expression for the generalization error that is asymptotically exact. Experimentally, we consider the cases of elastic net and ℓ_1 regularization, and verify our claims. Similarly to paper I above, we also obtain asymptotic expressions for the sparsity of the solution vector.

4.3 Paper III

In this paper, we consider the deep random feature model as described in section 3.2. We prove two results. Firstly, we prove a universality result which states that a deep RF is asymptotically equivalent to a Gaussian model which matches the first and second moments at each layer. Secondly, we analyze this result using the CGMT to obtain asymptotic expressions for this model. For the universality, we recall that the deep random features model is defined recursively. For a dataset $\{ (\boldsymbol{x}_i, y_i) \in \mathbb{R}^{p_0} \times \mathbb{R} \}_{i=1}^n$, we define $\boldsymbol{z}_i^{(0)} = \boldsymbol{x}_i$ and $\boldsymbol{z}_i^{(l+1)} = \sigma(\boldsymbol{W}^{(l+1)} \boldsymbol{z}^{(l)_i})$ where $\boldsymbol{W}^{(l)}$ for l in $1, \dots, L$ are $p_l \times p_{l-1}$ standard Gaussian random matrices. For odd activation functions σ, we find the following recursive Gaussian equivalent feature map:

$$\boldsymbol{\gamma}_i^{(0)} = \boldsymbol{x}_i, \qquad \boldsymbol{\gamma}^{(l+1)} = \rho_{1,l+1} \boldsymbol{W}^{(l+1)} \boldsymbol{\gamma}^{(l)} + \rho_{2,l+1} \boldsymbol{g}^{(l)},$$

where each $\boldsymbol{g}^{(l)}$ is an independent standard Gaussian vector and $\rho_{1,l}, \rho_{2,l}$ are constants also defined recursively as follows:

$$\rho_{1,l} = \mathbb{E}_{z \sim \mathcal{N}(0,1)} \left[\sigma'(\alpha_{l-1} z) \right], \qquad \rho_{2,l} = \sqrt{\mathbb{E}_{z \sim \mathcal{N}(0,1)} \left[\sigma^2(\alpha_{l-1} z) \right] - \alpha_{l-1}^2 \rho_{1,l}^2}.$$

Where σ' is the derivative of the activation function σ and α_l are constants also defined recursively as:

$$\alpha_0 = 1, \qquad \alpha_l = \sqrt{\rho_{1,l}^2 \alpha_{l-1}^2 + \rho_{2,l}^2}.$$

We prove that under a set of assumptions (given in section 3 of the paper) that the training error of the empirical risk minimization (1) for the deep random feature model and the Gaussian equivalent feature map is bounded by order $\mathcal{O}\left(\frac{1}{\sqrt{n}}\right)$. As such, in the asymptotic limit these the training error of the two problems will be identical. In other words, with respect to the training error, the deep random feature model, asymptotically, behaves like a Gaussian process with a complex covariance structure.

We prove our universality result by means of Lindeberg's method, which involves constructing a series of steps that interpolate between the deep RF and the Gaussian equivalent model and then bounding each step. Unlike most approaches of Linderberg's method, we apply this method in the dual space to the original ERM optimization. Mathematically, we can note that for the empirical risk minimization framework, as discussed above in section 3, that:

$$\begin{aligned}
&\min_{\boldsymbol{\theta} \in \mathbb{R}_L^p} \frac{1}{n} \sum_{k=1}^n \ell(\boldsymbol{z}_k^T \boldsymbol{\theta}, y_k) + R(\boldsymbol{\theta}) \\
&= \min_{\boldsymbol{\theta} \in \mathbb{R}_L^p, \boldsymbol{a} \in \mathbb{R}^n} \max_{\boldsymbol{d} \in \mathbb{R}^n} \frac{1}{n} \left(\sum_{k=1}^n \ell(a_k, y_k) + d_k (a_k - \boldsymbol{z}_k^T \boldsymbol{\theta}) \right) + R(\boldsymbol{\theta}) \\
&= - \min_{\boldsymbol{d} \in \mathbb{R}^n} \frac{1}{n} \sum_{k=1}^n \ell^*(-d_k, y_k) + R^* \left(\frac{1}{n} \boldsymbol{Z} \boldsymbol{d} \right)
\end{aligned}$$

here ℓ^* is the convex conjugate of ℓ with respect to the first element, R^* is the convex conjugate of the regularization function R and $\boldsymbol{Z}$ is the matrix with columns $\boldsymbol{z}_k$. In our proof, we create a chain that replaces the elements of $\boldsymbol{Z}$ with their Gaussian equivalents and prove that this difference is bounded.

We further analyze these resulting Gaussian equivalent models and find an alternative optimization by means of the CGMT, which will share the same optimal value. We experimentally verify that these expressions are accurate.

We also note that the covariance matrix of the Gaussian equivalent features $\boldsymbol{\gamma}^{(l)}$ forms a Lyapunov recursion, the recursion is given by:

$$\boldsymbol{R}^{(0)} = \boldsymbol{I}, \qquad \boldsymbol{R}^{(l)} = \rho_{1,l}^2 \boldsymbol{W}^{(l)} \boldsymbol{R}^{(l-1)} \boldsymbol{W}^{(l)T} + \rho_{2,l} \boldsymbol{I}.$$

We examine the eigenvalue distribution of this matrix using standard techniques from free probability theory and find a recursion that describes the Stieltjes transform of the covariance matrix. We note that the recursion suggests that there is a limiting distribution of the eigenvalues of $\boldsymbol{R}$ as we increase the number of layers.

4.4 Paper IV

In this paper, we prove a generalization of the Convex Gaussian Min Max Theorem. In section 2.3 we discuss the pair of processes, (1) and (2), which satisfy Gordon's comparison lemma and are used to prove the comparison lemma. Here we find another pair of primary and alternative processes that fulfill Gordon's lemma's requirements. The processes are:

$$p(\boldsymbol{x}, \boldsymbol{y}_1, \ldots, \boldsymbol{y}_k) = \sum_{l=1}^{k} \boldsymbol{y}_l^T \boldsymbol{G}_l \boldsymbol{\Sigma}_l^{1/2} \boldsymbol{x} + \gamma_l \left\| \boldsymbol{\Sigma}_l^{1/2} \boldsymbol{x} \right\| \|\boldsymbol{y}_l\|,$$

$$a(\boldsymbol{x}, \boldsymbol{y}_1, \ldots, \boldsymbol{y}_k) = \sum_{l=1}^{k} \|\boldsymbol{y}_l\| \, \boldsymbol{g}_l^T \boldsymbol{\Sigma}_l^{1/2} \boldsymbol{x} + \left\| \boldsymbol{\Sigma}_l^{1/2} \boldsymbol{x} \right\| \boldsymbol{h}_l^T \boldsymbol{y}_l.$$

Here $\boldsymbol{G}_l, \boldsymbol{\gamma}_l, \boldsymbol{h}_l, \boldsymbol{g}_l$ for l in $1, \ldots, k$ have i.i.d standard Normal entries and $\boldsymbol{\Sigma}_l$ are all positive semi-definite covariance matrices. As described in sections 2.3 and 2.4, a pair of equations that satisfy Gordon's lemma can be used to prove a Gaussian Min-Max Theorem, and can then be extended to a Convex Gaussian Min-Max theorem. The CGMT theorem for this new pair of processes (Discussed in detail in section III of the paper), proves that for an arbitrary function $\psi(\boldsymbol{x}, \boldsymbol{y}_1, \ldots, \boldsymbol{y}_k)$ which is continuous, convex in $\boldsymbol{x}$ and concave in all $\boldsymbol{y}_l$, that for any $\epsilon, c \in \mathbb{R}$ where $\epsilon > 0$:

$$\mathbb{P}\left[\left| \min_{\boldsymbol{x}} \max_{\boldsymbol{y}_1, \ldots, \boldsymbol{y}_k} r + \psi - c \right| > \epsilon \right] \leq 2^k \mathbb{P}\left[\left| \min_{\boldsymbol{x}} \max_{\boldsymbol{y}_1, \ldots, \boldsymbol{y}_k} a + \psi - c \right| > \epsilon \right],$$

where r is defined, analogously to (4) CGMT discussed in section 2.4, as:

$$r(\boldsymbol{x}, \boldsymbol{y}_1, \ldots, \boldsymbol{y}_k) = \sum_{l=1}^{k} \boldsymbol{y}_l^T \boldsymbol{G}_l \boldsymbol{\Sigma}_l^{1/2} \boldsymbol{x}.$$

In other words, as in the case of the CGMT, if $a + \psi$ concentrates on some definite value, then $p + \psi$ will concentrate on the same value for any finite value of k. We prove additional results such that if the solutions of $a + \psi$ have a high probability of belonging to a ball of a fixed radius, then the solutions of $r + \psi$ will share this property. This allows us to consider the results of the generalization performance of $r + \psi$ by means of the solutions of $a + \psi$.

We make use of this new CGMT to examine the asymptotic behavior of two different problems, the first is multi-source Gaussian regression and binary classification for Gaussian mixture models, and experimentally verify these results.

4.5 Paper V

In this paper, we prove another generalization of the Convex Gaussian Min-Max Theorem. As discussed in section 3.3, a central limitation of the CGMT is that the matrix $\boldsymbol{G}$ in the primary process (see section 2.4 and (4)) must be i.i.d. standard normal. This condition fails to hold in more complex models, such as vector-valued regression. As an example, considering just vector-valued linear regression where the labels are of dimension k, over a dataset $\{ (\boldsymbol{x}_i, \boldsymbol{y}_i) \in \mathbb{R}^m \times \mathbb{R}^k \}_{i=1}^n$:

$$\min_{\boldsymbol{\Theta} \in \mathbb{R}^{m \times k}} \frac{1}{2n} \left\| \boldsymbol{X}\boldsymbol{\Theta} - \boldsymbol{Y} \right\|_F^2,$$

where $\boldsymbol{X} \in \mathbb{R}^{n \times m}, \boldsymbol{Y} \in \mathbb{R}^{n \times k}$ are have columns $\boldsymbol{x}_i$ and $\boldsymbol{y}_i$ respectively and $\|\cdot\|_F$ denotes the Frobenius norm. Expressing this optimization as a min-max problem, we can see that:

$$\begin{aligned} &\min_{\boldsymbol{\Theta} \in \mathbb{R}^{m \times k}} \max_{\boldsymbol{Z} \in \mathbb{R}^{n \times k}} \frac{1}{n} \mathrm{Tr}[\boldsymbol{Z}^T \boldsymbol{X}\boldsymbol{\Theta} - \boldsymbol{Z}^T \boldsymbol{Y}] - \frac{1}{2n} \left\| \boldsymbol{Z} \right\|_F^2 \\ &= \min_{\boldsymbol{\theta} \in \mathbb{R}^{mk}} \max_{\boldsymbol{z} \in \mathbb{R}^{nk}} \frac{1}{n} \boldsymbol{z}^T (\boldsymbol{I}_k \otimes \boldsymbol{X}) \boldsymbol{\theta} - \boldsymbol{z}^T \boldsymbol{y} - \frac{1}{2n} \left\| \boldsymbol{z} \right\|^2. \end{aligned}$$

In the first line, we have introduced $\boldsymbol{Z} \in \mathbb{R}^{n \times k}$ and have taken the convex conjugate of the Frobenius norm, and in the second line we introduce $\boldsymbol{z} = \mathrm{vec}(Z)$, $\boldsymbol{\theta} = \mathrm{vec}(\boldsymbol{\Theta})$, $\boldsymbol{y} = \mathrm{vec}(\boldsymbol{Y})$. We can see that $\boldsymbol{I}_k \otimes \boldsymbol{X}$ is not i.i.d. Gaussian but instead repeats the features of $\boldsymbol{X}$ multiple times over along the diagonal. This means that the CGMT cannot be applied. In this paper, we extend the CGMT to cases like this.

We formalize the idea of sharing weights. We let $\tilde{\boldsymbol{G}}$ be a Gaussian matrix and let $\boldsymbol{A}_k, \boldsymbol{B}_k$ for k in $1, \ldots, K$ be sets of K-deterministic matrices, we then consider *Gaussian Matrix Sum* (GMS), as

$$\boldsymbol{G} = \sum_{k=1}^{K} \boldsymbol{A}_k^T \tilde{\boldsymbol{G}} \boldsymbol{B}_k.$$

In other words, the matrices $\boldsymbol{A}_k, \boldsymbol{B}_k$ encode how the elements of $\tilde{\boldsymbol{G}}$ are repeated in the matrix $\boldsymbol{G}$. As an example for the case of vector-valued regression

as discussed above, we have k matrices $\boldsymbol{A}_a \in \mathbb{R}^{nk\times n}, \boldsymbol{B}_a \in \mathbb{R}^{mk\times m}$ where $a = 1,\dots,k$. These matrices are defined element-wise as $(\boldsymbol{A}_a)_{bn+c,d} = \delta_{a,b}\delta_{c,d}$ and $(\boldsymbol{B}_a)_{bn+e,f} = \delta_{a,b}\delta_{e,f}$ where $a, b = 1,\dots,k,\ c, d = 1\dots n$ and $e, f = 1,\dots m$ and $\delta_{a,b}$ is the Kronecker delta.

For matrices $\boldsymbol{G}$ that can be expressed as a GMS, we find a new pair of primary and alternative equations

$$\begin{aligned} p(\boldsymbol{x},\boldsymbol{y}) = &\quad \boldsymbol{x}^T\boldsymbol{G}\boldsymbol{y} + \mathrm{Tr}[\boldsymbol{P}^{1/2}\boldsymbol{\gamma}\boldsymbol{Q}^{1/2}], \\ a(\boldsymbol{x},\boldsymbol{y}) = &\quad \textstyle\sum_{k=1}^{K} \boldsymbol{f}_k^T\boldsymbol{B}_k\boldsymbol{y} + \boldsymbol{h}_k^T\boldsymbol{A}_k\boldsymbol{x}. \end{aligned}$$

Here $\boldsymbol{G}$ is a GMS, $\boldsymbol{\gamma} \in \mathbb{R}^{K\times K}$ is a standard Gaussian matrix, and $\boldsymbol{F}, \boldsymbol{H}$ are matrices with columns $\boldsymbol{f}_k, \boldsymbol{h}_k$ respectively. $\boldsymbol{F} = \tilde{\boldsymbol{F}}\boldsymbol{P}^{1/2}$ and $\boldsymbol{H} = \tilde{\boldsymbol{H}}\boldsymbol{Q}^{1/2}$ where $\tilde{\boldsymbol{F}}, \tilde{\boldsymbol{H}}$ are i.i.d standard Gaussian matrices. Finally, $\boldsymbol{P}, \boldsymbol{Q}$ are positive semi-definite matrices defined element-wise as:

$$P_{k,k'} = \boldsymbol{x}^T\boldsymbol{A}_k^T\boldsymbol{A}_{k'}\boldsymbol{x}, \qquad Q_{k,k'} = \boldsymbol{y}^T\boldsymbol{B}_k^T\boldsymbol{B}_{k'}\boldsymbol{y}.$$

We can note that this pair of equations is directly equal to the CGMT pair given in (1) and (2) if $K = 1$. We prove the CGMT version of this theorem using a statistical physics proof using Gaussian interpolation as discussed in section 2.5. We further make use of our new theorem to examine the asymptotic behavior of vector-valued regression and regression with matrix convolution and experimentally verify our results.

Chapter 5

Concluding Remarks and Future Directions

In this thesis, we have used comparison theorems and universality results to study machine learning models in the asymptotic regime. That being the regime where both the model parameters and the number of data points grow infinitely, but at a finite ratio. We have shown that comparison theorems are powerful tools for asymptotic analysis as they allow us to consider easier to analyze proxy problems that share important statistics with the problems of interest. We have extended the literature on comparison theorems by extending the convex Gaussian min max theorem to both setups where the primary problem has independent but not identically distributed rows and to setups where features are shared. We have also considered random feature models, both shallow and deep, and have proven that these models can be analyzed through the same CGMT machinery by proving that they are asymptotically equivalent to Gaussian problems, which share moments, i.e., universality.

5.1 Future Directions

A number of future directions exist for the present research. Firstly, we can use the novel CGMT introduced in paper V to analyze a wide class of models, including linear time series models. These models also involve sharing weights between different instances of time, which can be captured using the weight-sharing machinery in paper V.

Another important direction is the analysis of dynamics in the asymptotic regime. While some results exist [43], full gradient descent dynamics, even for simple models, are difficult to analyze. As far as the authors are aware, no theoretical tools exist that can handle multiple discrete time steps of gradient descent without introducing a fresh dataset at each step. Capturing the full dynamics of training would allow for substantially better theoretical understanding of the impact of initial conditions, order of introduced data points (in the case of stochastic methods), and optimal hyperparameter values.

Finally, the replica method, the CGMT, and approximate message passing methods all seem to give identical results when applied to the same problem. There should be a general theorem that proves the relationship between these three methods. There has been some (unpublished) work proving an equivalence between the replica and another CGMT generalization [44]–[46], which claims to make the replica method rigorous. We are not aware of any similar results between the CGMT and message passing methods.

Bibliography

[1] H. Touvron, T. Lavril, G. Izacard *et al.*, 'Llama: Open and efficient foundation language models,' *arXiv preprint arXiv:2302.13971*, 2023 (cit. on p. 3).

[2] J. Achiam, S. Adler, S. Agarwal *et al.*, 'Gpt-4 technical report,' *arXiv preprint arXiv:2303.08774*, 2023 (cit. on p. 3).

[3] J. Sohl-Dickstein, E. Weiss, N. Maheswaranathan and S. Ganguli, 'Deep unsupervised learning using nonequilibrium thermodynamics,' in *International conference on machine learning*, pmlr, 2015, pp. 2256–2265 (cit. on p. 3).

[4] P. Dhariwal and A. Nichol, 'Diffusion models beat gans on image synthesis,' *Advances in neural information processing systems*, vol. 34, pp. 8780–8794, 2021 (cit. on p. 3).

[5] M. Belkin, S. Ma and S. Mandal, 'To understand deep learning we need to understand kernel learning,' in *International Conference on Machine Learning*, PMLR, 2018, pp. 541–549 (cit. on p. 4).

[6] M. Mézard, G. Parisi and M. A. Virasoro, *Spin glass theory and beyond: An Introduction to the Replica Method and Its Applications*. World Scientific Publishing Company, 1987, vol. 9 (cit. on p. 4).

[7] H. S. Seung, H. Sompolinsky and N. Tishby, 'Statistical mechanics of learning from examples,' *Physical review A*, vol. 45, no. 8, p. 6056, 1992 (cit. on p. 4).

[8] T. L. Watkin, A. Rau and M. Biehl, 'The statistical mechanics of learning a rule,' *Reviews of Modern Physics*, vol. 65, no. 2, p. 499, 1993 (cit. on p. 4).

[9] A. Engel, *Statistical mechanics of learning*. Cambridge University Press, 2001 (cit. on p. 4).

[10] D. L. Donoho, A. Maleki and A. Montanari, 'Message-passing algorithms for compressed sensing,' *Proceedings of the National Academy of Sciences*, vol. 106, no. 45, pp. 18 914–18 919, 2009 (cit. on p. 4).

[11] M. Bayati and A. Montanari, 'The dynamics of message passing on dense graphs, with applications to compressed sensing,' *IEEE Transactions on Information Theory*, vol. 57, no. 2, pp. 764–785, 2011 (cit. on p. 4).

[12] V. Chandrasekaran, B. Recht, P. A. Parrilo and A. S. Willsky, 'The convex geometry of linear inverse problems,' *Foundations of Computational mathematics*, vol. 12, pp. 805–849, 2012 (cit. on p. 4).

[13] C. Thrampoulidis, S. Oymak and B. Hassibi, 'The Gaussian min-max theorem in the Presence of Convexity,' *arXiv e-prints*, arXiv:1408.4837, arXiv:1408.4837, Aug. 2014. arXiv: `1408.4837 [cs.IT]` (cit. on pp. 4, 10).

[14] S. Oymak, C. Thrampoulidis and B. Hassibi, 'The squared-error of generalized lasso: A precise analysis,' in *2013 51st Annual Allerton Conference on Communication, Control, and Computing (Allerton)*, IEEE, 2013, pp. 1002–1009 (cit. on p. 4).

[15] C. Thrampoulidis, A. Panahi and B. Hassibi, 'Asymptotically exact error analysis for the generalized equation-lasso,' in *2015 IEEE International Symposium on Information Theory (ISIT)*, IEEE, 2015, pp. 2021–2025 (cit. on p. 4).

[16] C. Thrampoulidis, S. Oymak and B. Hassibi, 'Regularized linear regression: A precise analysis of the estimation error,' in *Conference on Learning Theory*, PMLR, 2015, pp. 1683–1709 (cit. on p. 4).

[17] J. W. Lindeberg, 'Eine neue herleitung des exponentialgesetzes in der wahrscheinlichkeitsrechnung,' *Mathematische Zeitschrift*, vol. 15, no. 1, pp. 211–225, 1922 (cit. on p. 4).

[18] S. B. Korada and A. Montanari, 'Applications of the lindeberg principle in communications and statistical learning,' *IEEE transactions on information theory*, vol. 57, no. 4, pp. 2440–2450, 2011 (cit. on p. 4).

[19] A. Panahi and B. Hassibi, 'A universal analysis of large-scale regularized least squares solutions,' in *NIPS*, 2017, pp. 3384–3393 (cit. on pp. 4, 16, 17, 21).

[20] A. Montanari and P.-M. Nguyen, 'Universality of the elastic net error,' in *2017 IEEE International Symposium on Information Theory (ISIT)*, IEEE, 2017, pp. 2338–2342 (cit. on p. 4).

[21] S. Oymak and J. A. Tropp, 'Universality laws for randomized dimension reduction, with applications,' *Information and Inference: A Journal of the IMA*, vol. 7, no. 3, pp. 337–446, 2018 (cit. on p. 4).

[22] E. Abbasi, F. Salehi and B. Hassibi, 'Universality in learning from linear measurements,' *Advances in Neural Information Processing Systems*, vol. 32, 2019 (cit. on p. 4).

[23] H. Hu and Y. M. Lu, 'Universality laws for high-dimensional learning with random features,' *IEEE Transactions on Information Theory*, 2022 (cit. on pp. 4, 17, 20).

[24] A. Montanari and B. N. Saeed, 'Universality of empirical risk minimization,' in *Conference on Learning Theory*, PMLR, 2022, pp. 4310–4312 (cit. on pp. 4, 17).

[25] H. Zou and T. Hastie, 'Regularization and variable selection via the elastic net,' *Journal of the royal statistical society: series B (statistical methodology)*, vol. 67, no. 2, pp. 301–320, 2005 (cit. on p. 4).

[26] D. Slepian, 'The one-sided barrier problem for gaussian noise,' *Bell System Technical Journal*, vol. 41, no. 2, pp. 463–501, 1962 (cit. on p. 8).

[27] R. Vershynin, *High-dimensional probability: An introduction with applications in data science*. Cambridge university press, 2018, vol. 47 (cit. on p. 9).

[28] Y. Gordon, 'Some inequalities for gaussian processes and applications,' *Israel Journal of Mathematics*, vol. 50, no. 4, pp. 265–289, 1985 (cit. on p. 9).

[29] M. Stojnic, *Fully bilinear generic and lifted random processes comparisons*, 2016. arXiv: 1612.08516 [math.PR]. [Online]. Available: https://arxiv.org/abs/1612.08516 (cit. on p. 12).

[30] C.-R. Hwang, 'Laplace's method revisited: Weak convergence of probability measures,' *The Annals of Probability*, pp. 1177–1182, 1980 (cit. on p. 12).

[31] K. B. Athreya and C.-R. Hwang, 'Gibbs measures asymptotics,' *Sankhya A*, vol. 72, pp. 191–207, 2010 (cit. on p. 12).

[32] O. Dhifallah and Y. M. Lu, 'A precise performance analysis of learning with random features,' *arXiv preprint arXiv:2008.11904*, 2020 (cit. on p. 16).

[33] S. Goldt, B. Loureiro, G. Reeves, F. Krzakala, M. Mézard and L. Zdeborová, 'The gaussian equivalence of generative models for learning with shallow neural networks,' in *Mathematical and Scientific Machine Learning*, PMLR, 2022, pp. 426–471 (cit. on p. 16).

[34] D. Schröder, H. Cui, D. Dmitriev and B. Loureiro, *Deterministic equivalent and error universality of deep random features learning*, 2023. DOI: 10.48550/ARXIV.2302.00401. [Online]. Available: https://arxiv.org/abs/2302.00401 (cit. on p. 16).

[35] C. Thrampoulidis, E. Abbasi and B. Hassibi, 'Precise error analysis of regularized M -estimators in high dimensions,' *IEEE Transactions on Information Theory*, vol. 64, no. 8, pp. 5592–5628, 2018. DOI: 10.1109/TIT.2018.2840720 (cit. on p. 16).

[36] B. Loureiro, C. Gerbelot, H. Cui *et al.*, *Learning curves of generic features maps for realistic datasets with a teacher-student model*, 2021. DOI: 10.48550/ARXIV.2102.08127. [Online]. Available: https://arxiv.org/abs/2102.08127 (cit. on pp. 16, 17).

[37] A. Rahimi and B. Recht, 'Random features for large-scale kernel machines,' *Advances in neural information processing systems*, vol. 20, 2007 (cit. on p. 17).

[38] S. Mei and A. Montanari, 'The generalization error of random features regression: Precise asymptotics and the double descent curve,' *Communications on Pure and Applied Mathematics*, 2019. DOI: https://doi.org/10.1002/cpa.22008. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/cpa.22008. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1002/cpa.22008 (cit. on p. 17).

[39] O. Dhifallah and Y. M. Lu, *A precise performance analysis of learning with random features*, 2020. DOI: 10.48550/ARXIV.2008.11904. [Online]. Available: https://arxiv.org/abs/2008.11904 (cit. on p. 17).

[40] Q. Han and Y. Shen, 'Universality of regularized regression estimators in high dimensions,' *The Annals of Statistics*, vol. 51, no. 4, pp. 1799–1823, 2023 (cit. on p. 17).

[41] C. Thrampoulidis, S. Oymak and M. Soltanolkotabi, *Theoretical insights into multiclass classification: A high-dimensional asymptotic view*, 2020. DOI: 10.48550/ARXIV.2011.07729. [Online]. Available: https://arxiv.org/abs/2011.07729 (cit. on p. 18).

[42] E. J. Candes and T. Tao, 'Decoding by linear programming,' *IEEE transactions on information theory*, vol. 51, no. 12, pp. 4203–4215, 2005 (cit. on p. 21).

[43] H. Cui, L. Pesce, Y. Dandi *et al.*, 'Asymptotics of feature learning in two-layer networks after one gradient-step,' *arXiv preprint arXiv:2402.04980*, 2024 (cit. on p. 27).

[44] M. Stojnic, *Bilinearly indexed random processes* – stationarization *of fully lifted interpolation*, 2023. arXiv: 2311.18097 [math.PR]. [Online]. Available: https://arxiv.org/abs/2311.18097 (cit. on p. 28).

[45] M. Stojnic, *Fully lifted interpolating comparisons of bilinearly indexed random processes*, 2023. arXiv: 2311.18092 [math.PR]. [Online]. Available: https://arxiv.org/abs/2311.18092 (cit. on p. 28).

[46] M. Stojnic, *Fully lifted random duality theory*, 2023. arXiv: 2312.00070 [math.PR]. [Online]. Available: https://arxiv.org/abs/2312.00070 (cit. on p. 28).