Thesis for The Degree of Licentiate of Engineering

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, 2023

Asymptotic Analysis of Machine Learning Models
Comparison Theorems and Universality

David Bosch

ISSN 1652-876X

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 2023.

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

The study of Machine Learning models in asymptotic regimes, has provided insight into many of the properties of ML models, but seemingly contradicts classical statistical wisdom. To solve this mystery, this thesis focuses on the analysis of models such as the LASSO and Random features regression, when the data points and model parameters grow infinite at constant ratios. It provides analysis for the asymptotic behavior of these problems, including characterization of the learning curves; the predicted training and generalization error as a function of the degree of overparameterization.

The papers in this thesis particularly focus on the usage of Gaussian comparison theorems as a methodological tool for the analysis of these problems. In particular, the convex Gaussian min max theorem allows us to study more complex ML optimization problems, by considering alternative models that are simpler to analyze, but asymptotically hold similar properties.

Secondarily, this thesis considers universality, which within the asymptotic context demonstrates that many statistics of ML models are fully determined by lower order statistical moments. This allows us to study surrogate Gaussian models, matching these moments. These surrogate Gaussian models can subsequently be analyzed by means of the Gaussian comparison theorems.

Keywords

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

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.

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.*

[**b**] **David Bosch**, Ashkan Panahi, Babak Hassibi, *Precise Asymptotic Analysis of Deep Random Feature Models*
Submitted to COLT2023.

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. Among them my fellow PhD students; Adam, Alexander, Anton, Arman, Christopher, Daniel, Emil, Emilio, Fazeleh, Filip, Firooz, Hanna, Hannes, Juan, Linus, Lena, Lovisa, Mehrdad, Niklas, Newton, and Tobias. A special thanks also to my office roommates Markus and Hampus. I would also like to express my thanks to my parents, who have been very supportive during my studies. Thank you, Nathan and Adam, for being great brothers who I can always rely on.

Contents

Part I

Summary

Chapter 1

Introductory Chapters

1 Introduction

In contrast to classical statistical theory, the modern day machine learning models that are used in practice generalize well to unseen data, despite being massively overparameterized. Statistical wisdom assumes that an overparameterized statistical model overfits to the training data set, and generalize poorly. Instead, in practice, increasing the model size often improves the generalization further. It is therefore of theoretical interest to analyze machine learning models in the context of statistics to determine the origin of this behavior, and to determine how this knowledge may be leveraged to build more useful models.

The double decent curve, as described by [1] is the most representative of this rising field. Since its proposal, many machine learning algorithms have demonstrated double descent and some of them have been analyzed. This includes linear regression [2]–[4], ridge regression [5], [6], LASSO, Random Features [7], [8], and others [9]–[18]. These models have been analyzed by a number of methods, including the replica technique [19], Gaussian widths [20], as well as the Gaussian comparison theorem [7], [21], [22], which is the focus of this thesis.

As the name suggests, comparison theorems allow us to analyze models, or more specifically optimization problems over models, 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 some limit, similar conclusions maybe drawn for the original problem. The particular theorem, central to this thesis, is the Convex Gaussian Min Max Theorem (CMGT) [23]–[26], which allows for comparisons of optimizations that contain bilinear Gaussian forms.

The CGMT, as well as many other theoretical approaches require Gaussianity of the data or features, and often both to be applicable. This is not representative of real data. However in high dimensional space, we frequently observe the concentration of statistics of random objects, such as vectors,

matrices and tensors. Similarly to the central limit theorem, many statistics of non-Gaussian random variables, will in large limits exhibit a similar behaviour to Gaussian models that match their lower order moments, namely mean and variance. Proving this fact for the interesting statistics of particular ML models, such as training and testing loss, is called universality [27]–[31]. This allow us to study Gaussian surrogate models that have similar asymptotic properties to the models of interest. Universality has been demonstrated for the random features case [32], and under certain assumptions Empirical Risk Minimization [33]. 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 theorem (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 2.4). Here we extend the existing universality results of [32] 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 [34] and ℓ_1 regularization, which could not be previously analyzed, in the random feature context.

2 Background

In this section, we give the relevant background for the works in this thesis. Firstly, we lay out the motivations and the preliminary concepts of what we want to analyze. Next, we discuss comparison theorems, most prominently the Convex Gaussian Min Max Theorem. Then, we discuss universality and the arguments that may be used to obtain it. Finally, we discuss the random features model and how it maybe explored in this regime.

2.1 Problem Setup

In supervised learning problems, we are concerned with datasets, $(\boldsymbol{x}, y) \in \mathbb{R}^d \times \mathbb{R}$ drawn from some joint probability distribution $p_{\boldsymbol{x},y}$, where the $d-$dimensional $\boldsymbol{x}$ are called the data, and y are the labels. Our goal is to find some function $f : \mathbb{R}^d \to \mathbb{R}$ that maps samples of the data to potential labels $\hat{y}_i = f(\boldsymbol{x}_i)$, such that they minimize the mean value of some metric $\ell(y, \hat{y}) : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$, that we call a loss function. Frequently, as in the case of this thesis, we consider an empirical version of this problem where we have a dataset $\mathcal{D} = \left\{(\boldsymbol{x}_i, y_i) \in \mathbb{R}^d \times \mathbb{R}\right\}_{i=1}^n$ of $n-$datapoints sampled from this distribution, and we attempt to minimize the empirical estimator

$$\min_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^{n} \ell(f(x_i), y_i), \tag{1}$$

where $\mathcal{F}$ is some class of functions. We note that we must restrict the class of functions $\mathcal{F}$, otherwise we are able to exactly map each data point to the corresponding label in many different ways. While this minimizes the loss, there is no guarantee that the result effectively generalizes to unseen data. The generalization error is measured by:

$$\mathcal{E}_{gen} = \mathbb{E}_{(\boldsymbol{x}_{new}, y_{new})} \hat{\ell}(y_{new}, h\left(f(\boldsymbol{x}_{new})\right)) \tag{2}$$

where $(\boldsymbol{x}_{new}, y_{new}) \sim p_{\boldsymbol{x},y}$ is a new sample drawn from the same distribution, $\hat{\ell}$ is a (potentially the same) loss function, and h is some potential post-processing function (such as the sign function).

In the scope of this thesis, we will consider a class of functions parameterized by a set of parameters $\boldsymbol{\theta} \in \mathbb{R}^m$, given by

$$\mathcal{F} = \{f_{\boldsymbol{\theta}}(\boldsymbol{x}) = \boldsymbol{\varphi}(\boldsymbol{x})^T \boldsymbol{\theta}\}. \tag{3}$$

Here $\boldsymbol{\varphi}(\cdot) : \mathbb{R}^d \to \mathbb{R}^m$ is a fixed function that maps the input data into a space of some other dimension m. We will generally restrict $\boldsymbol{\varphi}$ to a particular form, such as a random feature mapping, as discussed in section 2.4. To choose the value of the parameters $\boldsymbol{\theta}$, we employ the empirical risk minimization framework, and examine the following minimization problem:

$$\mathcal{E}_{train} = \min_{\boldsymbol{\theta} \in \mathbb{R}^m} \frac{1}{n} \sum_{i=1}^{n} \ell(f_{\boldsymbol{\theta}}(\boldsymbol{x}_i), y_i) + R(\boldsymbol{\theta}), \tag{4}$$

where $R : \mathbb{R}^m \to \mathbb{R}$ is a regularization function. We will generally consider cases where the regularization function is separable:

$$R(\boldsymbol{\theta}) = \sum_{i=1}^{m} r(\theta_i), \tag{5}$$

where $r : \mathbb{R} \to \mathbb{R}$ is applied element-wise to the elements of $\boldsymbol{\theta}_i = (\theta_i)_i$. We furthermore define $\hat{\boldsymbol{\theta}}$ as one of the optimal solutions of problem (4). In general, we will assume that the problem (4) is strongly convex, such that the solution $\hat{\boldsymbol{\theta}}$ is unique.

Similarly in the context of parameterized functions, the generalization error will be a function of the choice of parameter $\boldsymbol{\theta}$. We will consider generalizations of the form:

$$\mathcal{E}_{gen}(\boldsymbol{\theta}) = \mathbb{E}_{(\boldsymbol{x}_{new}, y_{new})} \left[\ell(y_{new}, f_{\boldsymbol{\theta}}(\boldsymbol{x}_{new})) \right]. \tag{6}$$

In other words, we will consider the same loss function, without post processing.

Furthermore, in the papers of this thesis other statistics of the set of parameters $\boldsymbol{\theta}$ will be studied. In general these are functions $T(\boldsymbol{\theta}) : \mathbb{R}^m \to \mathbb{R}$ of the parameters $\boldsymbol{\theta}$. The generalization error is one of these statistics, but we will also study $T(\boldsymbol{\theta}) = \|\boldsymbol{\theta}\|_1$, the ℓ_1 norm of the parameters, which gives a proxy for the sparsity of the solution vector $\boldsymbol{\theta}$.

The papers of this thesis provide theoretical analysis of certain problems of this type, in paper 1 the LASSO problem, and in paper 2 random feature mappings. We aim to give expressions for the training error, generalization error, and certain other statistics, that are accurate in the *asymptotic regime*. By asymptotic regime we mean the regime, where the number of data points n, the number of model parameters m, and the dimension of the input data d all grow infinite at constant ratios:

$$\frac{m}{n} \xrightarrow{n,m\to\infty} \gamma \qquad \frac{m}{d} \xrightarrow{m,d\to\infty} \eta. \tag{7}$$

To compute these expressions, we will first assume an existing true model, or teacher, that relates the data points $\boldsymbol{x}$ and the labels y as follows:

$$y_i = \boldsymbol{\varphi}(\boldsymbol{x}_i)^T \boldsymbol{\theta}^* + \nu_i, \tag{8}$$

where $\boldsymbol{\theta}^* \in \mathbb{R}^m$ is a "true" parameter vector, and ν_i is noise, that is i.i.d drawn from some distribution p_ν. We generally assume noise to be Gaussian. Based on this model, we define the error vector $\boldsymbol{e} = \boldsymbol{\theta} - \boldsymbol{\theta}^*$ which measures the degree of miss-match between the chosen parameters and the true parameters. Our asymptotic expression for various statistics of interest will be frequently expressed as functions of the optimal error vector $\hat{\boldsymbol{e}} = \hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*$.

Our theoretical approach to analyzing problems of this type will be by means of comparison theorem, discussed in section 2.2 below. These theorems, and in particular the Convex Gaussian Min max Theorem (CGMT) allows us to consider another optimization problem, alternative to problem (4), which in the asymptotic limit matches a wide range of the statistics of the original problem;

such as the training error (4) and generalization error (6). This alternative optimization is in general simpler to analyze, and occasionally permits closed form solutions.

The comparison theorems that we consider require that the data embeddings $\boldsymbol{\varphi}_i = \boldsymbol{\varphi}(\boldsymbol{x}_i)$ are Gaussian. In the majority of potential models, this is not the case. As such, in cases like the random feature model, we must first take a preliminary step before the CGMT analysis can be undertaken. We must demonstrate that we can replace $\boldsymbol{\varphi}_i$ with $\tilde{\boldsymbol{\varphi}}_i$ which are Gaussian distributed according to $\mathcal{N}(\boldsymbol{\mu}', \boldsymbol{\Sigma}')$ and asymptotically match the first two moments of $\boldsymbol{\varphi}_i$:

$$\left\|\mathbb{E}_{\boldsymbol{x}}\boldsymbol{\varphi}(\boldsymbol{x}) - \boldsymbol{\mu}'\right\|_2 \xrightarrow{n\to\infty} 0 \tag{9}$$

$$\left\|\mathbb{E}_{\boldsymbol{x}}\boldsymbol{\varphi}(\boldsymbol{x})\boldsymbol{\varphi}(\boldsymbol{x})^T - \boldsymbol{\Sigma}'\right\|_2 \xrightarrow{n\to\infty} 0. \tag{10}$$

We must then show that the statistics of interest, ie the training error, generalization error, etc. are asymptotically the same under this replacement of Gaussians. If this can be demonstrated, it is called *universality*. Universality is discussed in section 2.3 below.

2.2 Comparison Theorems

Recalling the ERM problem as described in equation (4), we note that in many cases this optimization is analytically intractable, and/or computationally expensive to perform. For the sake of analysis it may be prudent to instead consider an alternative optimization, which is hopefully simpler to analyze. If this alternative optimization problem has desirable qualities, such as placing bounds on the the statistics of the original problem, it may serve as a good candidate for such an analysis.

Here, we discuss the Convex Gaussian Min Max Theorem (CGMT), which is a particular form of a Gaussian comparison theorem. It demonstrates that certain optimizations over Gaussian bilinear forms can be analyzed by an alternative optimization that provides an upper and lower bound in probability of the training error of the original problem. If the alternative problem concentrates on some value, we can guarantee that the original problem similarly concentrates.

2.2.1 Gaussian Min Max Theorem

The CGMT is a particular application to optimization problems of a general theorem concerning Gaussian processes by Gordon [35].

Theorem 1 (Gordon [35]) *Let X_{ij} and Y_{ij} be two centered real valued Gaussian Processes indexed by $1 \le i \le I$ and $1 \le j \le J$, assume that the following holds*

$$\begin{aligned} \mathbb{E}X_{ij}^2 &= \mathbb{E}Y_{ij}^2 && \forall i, j \\ \mathbb{E}X_{ij}X_{ik} &\ge \mathbb{E}Y_{ij}Y_{ik} && \forall i, j, k \\ \mathbb{E}X_{ij}X_{lk} &\le \mathbb{E}Y_{ij}Y_{lk} && \forall j, k, i \ne l. \end{aligned} \tag{11}$$

Then for all $\lambda_{ij} \in \mathbb{R}$ we have that

$$\mathbb{P}\left(\bigcap_{1\leq i\leq I}\bigcup_{1\leq j\leq J}[Y_{ij}\geq\lambda_{ij}]\right)\geq\mathbb{P}\left(\bigcap_{1\leq i\leq I}\bigcup_{1\leq j\leq J}[X_{ij}\geq\lambda_{ij}]\right). \tag{12}$$

Gordon's theorem compares two Gaussian processes indexed by discrete sets I, J and examines the probability of events of the form:

$$\bigcap_{1\leq i\leq I}\bigcup_{1\leq j\leq J}[Y_{ij}\geq\lambda_{ij}]. \tag{13}$$

For a fixed value λ, this event may be translated into the language of optimization theory, and can be equivalently expressed as the optimal solution of a min-max problem:

$$\mathbb{P}\left(\min_{1\leq i\leq I}\max_{1\leq j\leq J}Y_{ij}\geq\lambda\right) \tag{14}$$

As such, assuming the conditions on the covariance structure of the two problems given in (11), the optimal solution of the min max over the discrete set of I, J of Y dominates the optimal solutions of the min max over X in probability. In practice, the second condition of equation (11), is often taken to be an equality relation, which is a special case.

To go from Gordon's theorem to the CGMT one must make a particular choice for the Gaussian processes X and Y. The unions and intersections readily transfer to min and max respective over discrete sets. By means of an ϵ-net argument, it may be extended to min-max problems over continuous sets. We now describe the CGMT.

2.2.2 Convex Gaussian Min Max Theorem

The CGMT is described by the following theorem:

Theorem 2 (CGMT [24]) *Let $\boldsymbol{G}\in\mathbb{R}^{n\times m}, \boldsymbol{g}\in\mathbb{R}^n, \boldsymbol{h}\in\mathbb{R}^m$ have i.i.d standard Normal elements, and are independent of each other. Let $\mathcal{S}_1\subset\mathbb{R}^n$ and $\mathcal{S}_2\subset\mathbb{R}^m$ be compact sets, and let $\psi(\cdot,\cdot):\mathbb{R}^{n\times m}\to\mathbb{R}$ be continuous function defined on $\mathcal{S}_1\times\mathcal{S}_2$. Consider the following two problems:*

$$\mathcal{P}_1(\boldsymbol{G}) = \min_{\boldsymbol{x}\in\mathcal{S}_1}\max_{\boldsymbol{y}\in\mathcal{S}_2}\boldsymbol{x}^T\boldsymbol{G}\boldsymbol{y}+\psi(\boldsymbol{x},\boldsymbol{y}), \tag{15}$$

$$\mathcal{P}_2(\boldsymbol{g},\boldsymbol{h}) = \min_{\boldsymbol{x}\in\mathcal{S}_1}\max_{\boldsymbol{y}\in\mathcal{S}_2}\|\boldsymbol{x}\|_2\,\boldsymbol{y}^T\boldsymbol{g}+\|\boldsymbol{y}\|_2\,\boldsymbol{h}^T\boldsymbol{x}+\psi(\boldsymbol{x},\boldsymbol{y}). \tag{16}$$

Then for any $c\in\mathbb{R}$ we have that

$$\mathbb{P}\left(\mathcal{P}_1(\boldsymbol{G})<c\right)\leq 2\mathbb{P}\left(\mathcal{P}_2(\boldsymbol{g},\boldsymbol{h})\leq c\right). \tag{17}$$

Furthermore, if $\mathcal{S}_1,\mathcal{S}_2$ are convex sets and ψ is convex-concave on $\mathcal{S}_1\times\mathcal{S}_2$. Then for any $C\in\mathbb{R}$, it holds that

$$\mathbb{P}\left(\mathcal{P}_1(\boldsymbol{G})>C\right)\leq 2\mathbb{P}\left(\mathcal{P}_2(\boldsymbol{g},\boldsymbol{h})\geq C\right). \tag{18}$$

This theorem demonstrates that the alternative optimization in (16) both upper bounds and lower bounds the primary optimization (15) in probability. This theorem is most useful if the alternative optimization concentrates. Assume that there is some μ such that

$$\mathbb{P}(|\mathcal{P}_2(\boldsymbol{g}, \boldsymbol{h}) - \mu| \geq \epsilon) \xrightarrow{n,m,\to\infty} 0 \qquad \forall \epsilon > 0. \tag{19}$$

Then by the bounds given by the theorem we similarly see that

$$\mathbb{P}(|\mathcal{P}_1(\boldsymbol{G}) - \mu| \geq \epsilon) \xrightarrow{n,m,\to\infty} 0 \qquad \forall \epsilon > 0. \tag{20}$$

As such, we can analyze or solve problem $\mathcal{P}_2$, and assuming concentration, we can determine the optimal value of $\mathcal{P}_1$ by proxy.

We make note of an important pitfall concerning this method of analysis. Primarily, the CGMT only demonstrates convergence of the optimal values of the two optimization problems. The optimal solutions $\hat{\boldsymbol{x}}_1, \hat{\boldsymbol{y}}_1$ of $\mathcal{P}_1$ and $\hat{\boldsymbol{x}}_2, \hat{\boldsymbol{y}}_2$ of $\mathcal{P}_2$ will in general be distinct and independent of each other. This difficulty must be addressed by additional arguments concerning the statistics of the optimal solutions.

For example, under mild conditions, it can be shown that $\|\hat{\boldsymbol{x}}_1\|$ will converge to the value of $\|\hat{\boldsymbol{x}}_2\|$, for a number of different norms. In general, we can frequently obtain convergence information about the statistics of the solutions, despite the solutions themselves being different.

As an example of the application of this theorem, we may note that for any loss function ℓ and Gaussian $\boldsymbol{\varphi}_i$ for $i = 1, \ldots, n$ we can express the ERM problem (4) as:

$$\min_{\boldsymbol{\theta} \in \mathbb{R}^m} \sum_i \ell(\boldsymbol{\varphi}_i^T \boldsymbol{\theta}, y_i) + R(\boldsymbol{\theta}) = \min_{\boldsymbol{\theta} \in \mathbb{R}^m} \max_{\boldsymbol{z} \in \mathbb{R}^n} \sum_i z_i \boldsymbol{\varphi}_i^T \boldsymbol{\theta} - \ell^*(z_i, y_i) + R(\boldsymbol{\theta}), \tag{21}$$

where ℓ^* is the Legendre transform of ℓ with respect to the first element, and $\boldsymbol{z} = (z_i)_i$. If it can be shown that $\hat{\boldsymbol{\theta}}, \hat{\boldsymbol{z}}$ can both be restricted to compact sets, then the CGMT can be applied. To see this, we may recall that by the definition of the Legendre transform ℓ^* is convex in the first element, and R is generally convex. Furthermore, we note that the conditions are trivially satisfied if the problem is strongly convex-concave, as this allows for an implicit ℓ_2^2 ball around the unique solutions of the optimization problems, therefore allowing the CGMT to be applied. This occurs for example if $R(\boldsymbol{\theta})$ is strongly convex and ℓ is $L-$smooth, which implies that ℓ^* is strongly convex.

2.3 Universality

2.3.1 Motivation

We may note that the CGMT only works when the matrix $\boldsymbol{G}$ in (15) is *i.i.d* Gaussian, which is not realistic in the majority of practically considered ML use cases.

Letting $\boldsymbol{y} = (y_i)_i$ and $\boldsymbol{X}_{ij} = \boldsymbol{\varphi}_j(\boldsymbol{x}_i)$, where the elements of $\boldsymbol{X}$ are i.i.d but no longer Gaussian, we consider the following problem:

$$\begin{aligned} \mathcal{P}(\boldsymbol{X}) = & \min_{\boldsymbol{\theta}} \|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\theta}\|_2^2 + R(\boldsymbol{\theta}) \\ = & \min_{\boldsymbol{e}} \|\boldsymbol{\nu} - \boldsymbol{X}\boldsymbol{e}\|_2^2 + R(\boldsymbol{e} + \boldsymbol{\theta}^*), \end{aligned} \tag{22}$$

where we introduce the error vector $\boldsymbol{e} = \boldsymbol{\theta} - \boldsymbol{\theta}^*$, recalling the label definition in (8). This optimization problem may be expressed in terms of a min max problem by taking the Legendre transform of the 2−norm, as in (21). Then, we obtain:

$$\mathcal{P}(\boldsymbol{X}) = \min_{\boldsymbol{e}} \max_{\boldsymbol{z}} \boldsymbol{z}^T\boldsymbol{\nu} - \boldsymbol{z}^T\boldsymbol{X}\boldsymbol{e} - \frac{1}{2}\|\boldsymbol{z}\|_2^2 + R(\boldsymbol{e} + \boldsymbol{\theta}^*). \tag{23}$$

The result is in the form assumed by CGMT. However, the matrix $\boldsymbol{X}$ is no longer i.i.d Gaussian.

The CGMT might appear to be inapplicable, at this point, however we are dealing with the asymptotic limit in which $n, m, d \to \infty$. As such, we may be able to exploit the "blessing of dimensionality" by recalling that the properties of many random matrices and vectors concentrate in this large limit.

To gain a deeper insight, we may recall the classical example of the central limit theorem for an independent sequence of random variables X_i with mean μ and variance σ^2. We recall that as $n \to \infty$,

$$\frac{1}{n}\sum_i^n X_i \to_{\mathrm{d}} \mathcal{N}(\mu, \sigma^2). \tag{24}$$

As such, regardless of many of the complexities and higher order moments of a probability distribution X_i, the empirical mean is only dependent on the first two moments, μ, σ^2. We may hope that in (22) a similar phenomenon occurs and that the statistics of interest, such as the training and generalization error, may similarly depend only on the first few moments of the distribution $\boldsymbol{X}$, in the large limit. Then, we may instead consider another random vector $\tilde{\boldsymbol{x}}$ which has the same first and second moments as a row $\boldsymbol{x}$ of $\boldsymbol{X}$. In other words,

$$\mathbb{E}\boldsymbol{x} = \mathbb{E}\tilde{\boldsymbol{x}} = \boldsymbol{\mu} \qquad \mathbb{E}(\boldsymbol{x} - \boldsymbol{\mu})(\boldsymbol{x} - \boldsymbol{\mu})^T = \mathbb{E}(\tilde{\boldsymbol{x}} - \boldsymbol{\mu})(\tilde{\boldsymbol{x}} - \boldsymbol{\mu})^T \tag{25}$$

Accordingly, we ask if there is a class of test functions ϕ such that

$$\left|\phi(\mathcal{P}(\boldsymbol{X})) - \phi(\mathcal{P}(\tilde{\boldsymbol{X}}))\right| \xrightarrow{n\to\infty} 0. \tag{26}$$

In other words, under the set of test functions, the values of the original problem and the Gaussian surrogate are asymptotically indistinguishable. More generally, let $\boldsymbol{\theta}_{\boldsymbol{X}}$ and $\boldsymbol{\theta}_{\tilde{\boldsymbol{X}}}$ be the optimal solutions of the two problems. We hope to find a set of test functions such that for relevant statistics T we have:

$$|\phi(T(\boldsymbol{\theta}_{\boldsymbol{X}})) - \phi(T(\boldsymbol{\theta}_{\tilde{\boldsymbol{X}}}))| \xrightarrow{n\to\infty} 0. \tag{27}$$

Establishing such relations, we can show a chain of relationships for analyzing the ML problem. Firstly, by means of universality we can prove that

the training and generalization error of the original problem, and the Gaussian surrogate problem, will be asymptotically equivalent.

Then by considering the Gaussian surrogate problem, we can make use of the CGMT to analyze it. By means of the CGMT analysis, we find alternative expressions for training and generalization error for the Gaussian surrogate model that hold asymptotically. From this chain, we see that the CGMT values will converge to their corresponding values in the original problem, in the asymptotic limit. This allows us to analyze problems such as random feature embeddings, by means of an alternative CGMT problem.

2.3.2 Proving Universality

While there exist a number of possible approaches to proving universality of models, here we focus on two that are used in the literature; first being Lindeberg's Method and the second being Stein's method. Lindeberg's method is used in the works of [28], [32], [33], while steins method is additionally used in [32], [33].

Lindebergs 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.

For 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

$$\boldsymbol{X}_r = [\boldsymbol{x}_1\ \boldsymbol{x}_2\ \cdots \boldsymbol{x}_{r-1}\ \tilde{\boldsymbol{x}}_r\ \cdots \tilde{\boldsymbol{x}}_n]^T \qquad r = 0, \ldots, n. \tag{28}$$

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

$$\begin{aligned}\left\|T(\boldsymbol{X}) - T(\tilde{\boldsymbol{X}})\right\|_2^2 &= \left\|\sum_{r=0}^{n-1} T(\boldsymbol{X}_r) - T(\boldsymbol{X}_{r+1})\right\|_2^2 \\ &\leq \sum_{r=0}^{n-1} \left\|T(\boldsymbol{X}_r) - T(\boldsymbol{X}_{r+1})\right\|_2^2,\end{aligned} \tag{29}$$

where the first equality is obtained by a telescoping sum, and the second by the triangle inequality. If we demonstrate that $\|T(\boldsymbol{X}_r) - T(\boldsymbol{X}_{r+1})\|_2^2 \leq \frac{C}{n^{3/2}}$, for some constant $C > 0$, we will be able to show that:

$$\left\|T(\boldsymbol{X}) - T(\tilde{\boldsymbol{X}})\right\|_2^2 \leq \sum_{r=0}^{n-1} \frac{C}{n^{3/2}} \leq \frac{C}{\sqrt{n}} \xrightarrow{n\to\infty} 0 \tag{30}$$

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 moment, are asymptotically equivalent.

Stein's method: Stein's method for determining universality relies on a property of Gaussians, known as Stein's lemma. Steins lemma, establishes for a function g of a Gaussian vector $\tilde{x} \sim \mathcal{N}(0, \sigma^2)$ that

$$\mathbb{E}\tilde{x}^T g(\tilde{x}) = \sigma^2 \mathbb{E}\nabla g(\tilde{x}). \tag{31}$$

Assuming that universality holds, we can then assume that in the asymptotic limit a function of interest T of a non Gaussian vector x will begin to act like a function of Gaussian variable. Recalling what we want to bound

$$|\mathbb{E}_x T(x) - \mathbb{E}_{\tilde{x}} T(\tilde{x})| = |\mathbb{E}_x \left[T(x) - \mathbb{E}_{\tilde{x}}[T(\tilde{x})]\right]|, \tag{32}$$

we aim to find a function $f(x)$ such that

$$f'(x) - xf(x) = T(x) - \mathbb{E}_{\tilde{x}}[T(\tilde{x})], \tag{33}$$

in which case we may argue that

$$|\mathbb{E}_x T(x) - \mathbb{E}_{\tilde{x}} T(\tilde{x})| = |\mathbb{E}_x(f'(x)) - \mathbb{E}_x x f(x)|. \tag{34}$$

This translation step can be easier to perform as the right hand side can be simpler to bound than the left hand side. Often, this is performed by exploiting the fact that $f(x)$ can be Taylor expanded to match the $f'(x)$ term. We may also note that $f(x)$ has a definite solution in terms of T, for 1 dimensional data, it is given by

$$f(x) = e^{x^2/2} \int_{-\infty}^{x} e^{-t^2/2}[T(t) - \mathbb{E}_{\tilde{x}}[T(\tilde{x})]]dt. \tag{35}$$

Solutions for higher dimensional cases also exist, see [36].

2.4 Random feature and Hermite Polynomials

Much of the analysis in the papers in the sequel concern a particular model called the random features model. Consider a dataset $\{(\boldsymbol{z}_i, y_i) \in \mathbb{R}^d \times \mathbb{R}\}_{i=1}^n$, where we assume that $\boldsymbol{z}_i \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_d)$. We then let $\boldsymbol{W} \in \mathbb{R}^{m \times d}$ be a weight matrix, with i.i.d standard normal elements, and with rows $\boldsymbol{w}_i \in \mathbb{R}^d$ where $i = 1, \ldots, m$. Finally, let σ be a non-linear activation function applied elementwise. We then define the random feature embedding as

$$X_{ij} = \sigma(\boldsymbol{w}_i^T \boldsymbol{z}_j) \qquad i = 1, \ldots, m \quad j = 1, \ldots, n. \tag{36}$$

Our goal is to find a mapping from the elements of $\boldsymbol{X} = (X_{ij})$ to the labels $\boldsymbol{y} = (y_i)_i$ by considering a class of parametric functions

$$\mathcal{F} = \left\{ f_{\boldsymbol{\theta}}(\boldsymbol{X}) = \frac{1}{\sqrt{m}} \boldsymbol{X}\boldsymbol{\theta} \right\}, \tag{37}$$

where each function $f_{\boldsymbol{\theta}}$ is parameterized by $\boldsymbol{\theta}$. Our goal is to find the optimal value of $\boldsymbol{\theta}$ by minimizing the objective value of the problem, given by:

$$\mathcal{E}_{train} = \min_{\boldsymbol{\theta} \in \mathbb{R}^m} \frac{1}{2n} \left\| \boldsymbol{y} - \frac{1}{\sqrt{m}} \boldsymbol{X}\boldsymbol{\theta} \right\| + R(\boldsymbol{\theta}). \tag{38}$$

Furthermore, let $\hat{\boldsymbol{\theta}}$ be the optimal solution to this problem. For the sake of analysis, we must also assume an existing model on the relationship between the data and the lables, we choose

$$\boldsymbol{y} = \frac{1}{\sqrt{m}}\boldsymbol{X}\boldsymbol{\theta}^* + \boldsymbol{\nu} \tag{39}$$

Where $\boldsymbol{\theta}^* \in \mathbb{R}^m$ is the "true" solution vector, which may be either deterministic or random according to some distribution $p_{\boldsymbol{\theta}^*}$, and $\boldsymbol{\nu}$ is noise. We generally assume noise to be Gaussian, distributed according to $\boldsymbol{\nu} \sim \mathcal{N}(\mathbf{0}, \sigma_{\boldsymbol{\nu}}^2 \boldsymbol{I}_n)$.

According to the steps of the analysis laid out above, we would first like to replace the given problem by a Gaussian equivalent. However, this presents a difficulty due to the choice of non-linearity. We may note that, for a given row $\boldsymbol{x}$ of $\boldsymbol{X}$ that the mean and covariance, with respect to the data, are given by

$$\boldsymbol{\mu} = \mathbb{E}_{\boldsymbol{z}}\sigma\left(\frac{1}{\sqrt{d}}\boldsymbol{W}\boldsymbol{z}\right) \qquad \boldsymbol{\Sigma} = \mathbb{E}_{\boldsymbol{z}}\sigma\left(\frac{1}{\sqrt{d}}\boldsymbol{W}\boldsymbol{z}\right)\sigma\left(\frac{1}{\sqrt{d}}\boldsymbol{W}\boldsymbol{z}\right)^T \tag{40}$$

While the values of $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ can be computed numerically, they are often intractable to compute analytically, and as such difficult to work with. Instead, a clever trick by means of Hermite polynomial expansion can be used.

2.4.1 Hermite Polynomials

A Hermite polynomial expansion for a function f produces an expansion of the function in terms of a function basis. This function basis consists of polynomials that are orthogonal with respect to the Gaussian measure. More specifically for any function $f(x)$, its Hermite polynomial expansion is given by:

$$\begin{aligned} f(x) &= \sum_{n=0}^{\infty} \frac{1}{n!\sqrt{2\pi}} b_n H_n(\boldsymbol{x}) \\ b_n &= \int_{-\infty}^{\infty} f(x) H_n(x) e^{-x^2/2}\, \mathrm{d}x, \end{aligned} \tag{41}$$

where $H_n(x)$ are the Hermite polynomials, explicitly defined by

$$H_n(x) = (-1)^n e^{x^2/2} \frac{d^n}{dx^n} e^{-x^2/2}. \tag{42}$$

These polynomials have interesting properties with respect to the Gaussian measure. Let $\langle \cdot, \cdot \rangle$ be an inner product between functions, defined as

$$\langle a(x), b(x) \rangle = \int_{-\infty}^{\infty} a(x) b(x) e^{-x^2/2} \mathrm{d}x. \tag{43}$$

Then we may note that

$$\langle H_n(x), H_m(x) \rangle = \delta_{nm} \sqrt{2\pi} n!. \tag{44}$$

As such, they form an orthogonal basis, with respect to a Gaussian measure, $e^{-x^2/2}$. Furthermore, for a given function we note that

$$\|f(x)\|^2 = \langle f(x), f(x)\rangle = \langle \sum_n \frac{1}{n!\sqrt{2\pi}} b_n H_n(x), \sum_n \frac{1}{n!\sqrt{2\pi}} b_n H_n(x)\rangle = \sum_n b_n^2. \quad (45)$$

Now, we may recall that $H_0(x) = 1$ and that $H_1(x) = x$, and hence for any function $f(x)$:

$$b_0 = \frac{1}{\sqrt{2\pi}} \langle H_0(x), f(x)\rangle = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} f(x) e^{-x^2/2} \, \mathrm{d}x = \mathbb{E}_x f(x) \quad (46)$$

$$b_1 = \frac{1}{\sqrt{2\pi}} \langle H_1(x), f(x)\rangle = \mathbb{E}_x x f(x) = \mathbb{E}_x f'(x) \quad (47)$$

where the final relation is due to Stein's lemma. Although this must be proven, by the principals of universality discussed above, we might assume that the majority of the information relevant for our statistics, may be contained only within the first two terms of this expansion. Operating under this assumption we could truncate the polynomial expansion to the following form:

$$f(x) \approx b_0 + b_1 x + b_* z, \quad (48)$$

where $z \sim \mathcal{N}(0, 1)$ is a Gaussian that roughly captures all of the higher order information of the expansion. By the conservation of energy of the polynomial we can see that b_* must be given by:

$$b_*^2 = \mathbb{E} f^2(x) - b_0^2 - b_1^2 = \sum_{n=2}^{n} b_n^2. \quad (49)$$

Returning to the Random features model, this Hermite approximation allows us to consider:

$$\sigma\left(\frac{1}{\sqrt{d}} \boldsymbol{w}^T \boldsymbol{z}\right) \approx b_0 + b_1 \frac{1}{\sqrt{d}} \boldsymbol{w}^T \boldsymbol{z} + b_* \bar{z} \quad (50)$$

$$b_0 = \mathbb{E}\sigma(x) \qquad b_1 = \mathbb{E}\sigma'(x) \qquad b_* = \sqrt{\mathbb{E}\sigma^2(x) - b_0^2 - b_1^2} \quad \bar{z} \sim \mathcal{N}(0, 1). \quad (51)$$

Hence, we obtain:

$$\boldsymbol{\mu} = \mathbb{E}\sigma\left(\frac{1}{\sqrt{d}} \boldsymbol{W} \boldsymbol{z}\right) \approx b_0 \mathbf{1} \stackrel{def}{=} \boldsymbol{\mu}', \quad (52)$$

and with covariance

$$\boldsymbol{\Sigma} = \mathbb{E}\sigma\left(\frac{1}{\sqrt{d}} \boldsymbol{w}^T \boldsymbol{z}\right) \sigma\left(\frac{1}{\sqrt{d}} \bar{\boldsymbol{w}}^T \boldsymbol{z}\right)^T \approx b_0 \mathbf{1}\mathbf{1}^T + \frac{b_1}{d} \boldsymbol{W}\boldsymbol{W}^T + b_*^2 \boldsymbol{I} \stackrel{def}{=} \boldsymbol{\Sigma}'. \quad (53)$$

In the papers below, we assume that σ is an odd function, and this results in $b_0 = 0$. Furthermore, if we assume mild conditions on $\boldsymbol{W}$, [32] has demonstrated that in the asymptotic limit,

$$\|\boldsymbol{\Sigma} - \boldsymbol{\Sigma}'\|_2 \leq \frac{C\text{polylog}(m)}{\sqrt{m}} \xrightarrow{m\to\infty} 0, \tag{54}$$

where $\|\cdot\|_2$ is the operator norm (largest singular value), $C > 0$ is a positive constant and polylog m is a polynomial of the logarithm of m. As such the approximation asymptotically matches the first two moments of the true random features. It requires further work, to prove that a Gaussian replacement of the features with this approximation also satisfies universality, but it can be established, as shown in [32].

The above discussion suggests that for the sake of analysis, we may examine an alternative optimization problem

$$\mathcal{E}'_{train} = \min_{\boldsymbol{\theta}} \frac{1}{2n} \left\| \tilde{\boldsymbol{y}} - \frac{1}{\sqrt{m}} \tilde{\boldsymbol{X}} \boldsymbol{\theta} \right\| + R(\boldsymbol{\theta}), \tag{55}$$

where the rows $\tilde{\boldsymbol{x}}_i \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{\Sigma}')$ and $\tilde{y}_i = \tilde{\boldsymbol{x}}_i{}^T \boldsymbol{\theta}^* + \nu_i$. With respect to the statistics of interest, namely training and generalization error, this problem will asymptotically have the same values as problem (38). As for additional statistics, such as the sparsity, $\|\hat{\boldsymbol{\theta}}\|_1$. We concretely derive universality for this case in paper 2 of this thesis.

3 Summary of the Included Papers

3.1 Paper 1

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

$$\min_{\boldsymbol{\theta}\in\mathbb{R}^m} \frac{1}{2n}\sum_{i=1}^{n}(y_i - \boldsymbol{\theta}^T\boldsymbol{x}_i)^2 + \frac{\lambda}{\sqrt{m}}\left\|\boldsymbol{\theta}\right\|_1, \tag{56}$$

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

$$\begin{aligned} &\min_{\boldsymbol{\theta}\in\mathbb{R}^m} \left\|\boldsymbol{\theta}\right\|_1 \\ &\qquad\qquad \text{s.t} \\ y_i = \boldsymbol{\theta}^T\boldsymbol{x}_i \qquad &i = 1, \ldots, n. \end{aligned} \tag{57}$$

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

$$y_1 = \frac{1}{\sqrt{m}}\boldsymbol{x}_i^T\boldsymbol{\theta}^* + \nu_i \qquad i = 1, \ldots, n. \tag{58}$$

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 $\boldsymbol{\theta}^*$ exist such that $\boldsymbol{\theta}_A^*$, ie $\boldsymbol{\theta}^*$ restricted to the indices in A, has values much large than $\boldsymbol{\theta}_{A^c}^*$. For this problem, the generalization error, as a function of the regularization parameter can be expressed as

$$\begin{aligned} \mathcal{E}_{gen}(\lambda) = \mathbb{E}_{\boldsymbol{x},y}(y - \hat{\boldsymbol{\theta}}_\lambda^T\boldsymbol{x})^2 - \mathbb{E}(y - \boldsymbol{x}^T\boldsymbol{\theta}^*)^2 \\ = \boldsymbol{e}_\lambda^T\boldsymbol{R}\boldsymbol{e}_\lambda \end{aligned} \tag{59}$$

where $\hat{\boldsymbol{\theta}}_\lambda$ is the solution to (56) 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 (56), can asymptotically be expressed as

$$\min_{\boldsymbol{e}} \frac{1}{2}\boldsymbol{e}^T\boldsymbol{R}\boldsymbol{e} + \frac{q}{\sqrt{n}}\boldsymbol{e}^T\boldsymbol{h} + \frac{q\lambda}{\beta\sqrt{m}}\left\|\frac{\boldsymbol{\theta}^*}{\sqrt{m}} + \boldsymbol{e}\right\|_1, \tag{60}$$

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

$$q^2 = \boldsymbol{e}^T\boldsymbol{R}\boldsymbol{e} + \sigma_{\boldsymbol{\nu}}^2, \qquad \beta = q + \frac{1}{n}\boldsymbol{e}^T\boldsymbol{h}. \tag{61}$$

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 gives 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:

$$\mathcal{E}_{gen}(\lambda) = \frac{1}{m}\sum_{j=1}^{m} r_j \mathbb{E}_\phi \left[\mathcal{T}_{\frac{\lambda q}{\beta r_j}} \left(\theta_j^* + \frac{q\phi}{\sqrt{r_j \gamma}} - \theta_j^* \right)^2 \right], \tag{62}$$

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

$$\mathcal{T}_a(b) = \begin{cases} b - a & b > a \\ b + a & b < -a \\ 0 & |b| \leq a \end{cases} \tag{63}$$

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 generalization and sparsity of the solution vectors.

3.2 Paper 2

In this paper we consider the case of random features regression as described in section 2.4. 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 [32]. [32] had given universality results for random feature models, subject to the hermite trick as explained in 2.4.1, 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 the elastic net regularization function:

$$R(\boldsymbol{\theta}) = \frac{\alpha}{2} \|\boldsymbol{\theta}\|^2 + \lambda \|\boldsymbol{\theta}\|_1 . \tag{64}$$

Here α, λ are two regularization strength parameters. Secondarily, we extend the universality results to ℓ_1 regularization. To prove this, we make use of a similar technique as [28] 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 2.4) satisfies the restricted isometry property [37]. 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 randomess of the Gaussian input data $\boldsymbol{z}$ and secondarily 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:

$$\max_{\beta>0}\min_{q>0}\max_{\xi>0}\min_{t>0}\frac{1}{m}\mathbb{E}\left[\mathcal{M}_{\frac{1}{2c_1}\ R}\left(\boldsymbol{\theta}^* - \frac{c_2\sqrt{\gamma}}{2c_1}\boldsymbol{\phi}\right)\right] \\ -\frac{c_2^2\gamma}{4c_1} + \frac{\xi t}{2} + \frac{\beta q}{2} + \frac{\beta\sigma_{\boldsymbol{\nu}}^2}{2q} + \frac{\xi\beta^2}{2t\eta} - \frac{\beta\xi^2}{2q} - \frac{q\beta}{2\eta} - \frac{\beta^2}{2}, \tag{65}$$

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:

$$\mathcal{M}_{\tau\ f}(\boldsymbol{y}) = \min_{\boldsymbol{x}}\frac{1}{2\tau}\left\|\boldsymbol{x}-\boldsymbol{y}\right\|^2 + f(\boldsymbol{x}). \tag{66}$$

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 4d 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 Discussion and Future Work

In this thesis, we have considered the asymptotic analysis of machine learning models. In paper I, we considered the case of the ℓ_1 regularized least squares, or LASSO problem. We considered the case of Gaussian features, with various feature strengths. We gave expressions for the asymptotic training and generalization loss as a function of feature strength. We also gave asymptotic expressions for the sparsity of the solution vector and considered the effect of feature strength of the sparsity of the LASSO solution. In Paper II, we considered the random features model. We extended the universality results to more general strongly convex regularizers, as well as the specific case of ℓ_1 regularization. Making use of a novel nested application of the CGMT on the covariance of the Gaussian surrogate model, we obtained a 4d scalar optimization problem which is readily computable, and from which the training and generalization error, as well as the sparsity of the solution vector may be obtained.

There are a number of future directions for the research presented in this thesis. One direction is to consider the case of deep random feature models, where the data is embedded through multiple weight matrices and nonlinearities. For this case, universality must first be established, then our technique of applying the CGMT in a nested manner allows us to examine how depth impacts the generalization of RF models.

A current limitation of the CGMT as a method of analysis is that it only works on bilinear Gaussian forms. This makes it very difficult to analyse ML models that have vector output, such as classification. Some work has been done is this field [38], however most likely a new tool must be developed, potentially extending the CGMT to cover these cases.

Another possible direction is to consider additional types of layer for asymptotic analysis, such as convolutional layers. While convolutional layers can be expressed as binlinear forms, the feature matrix is no longer i.i.d Gaussian as weights are now shared within a filter. Addressing this problems would also require us to develop some different technique to analyse these cases. Finally, it is interesting to consider the case of a 2 layer NN instead of a RF models, in which the weights of both layers are trained.

Bibliography

[1] 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. 3).

[2] A. Tsigler and P. L. Bartlett, "Benign overfitting in ridge regression," *arXiv e-prints*, arXiv:2009.14286, arXiv:2009.14286, Sep. 2020. arXiv: 2009.14286 [math.ST] (cit. on p. 3).

[3] P. L. Bartlett, P. M. Long, G. Lugosi and A. Tsigler, "Benign overfitting in linear regression," *arxiv:1906.11300*, 2020 (cit. on p. 3).

[4] M. Belkin, D. Hsu and J. Xu, "Two models of double descent for weak features," *SIAM Journal on Mathematics of Data Science*, vol. 2, no. 4, pp. 1167–1180, 2020 (cit. on p. 3).

[5] C. Louart, Z. Liao and R. Couillet, "A random matrix approach to neural networks," *The Annals of Applied Probability*, vol. 28, no. 2, pp. 1190–1248, 2018 (cit. on p. 3).

[6] 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. 3).

[7] 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 p. 3).

[8] O. Dhifallah and Y. M. Lu, "A precise performance analysis of learning with random features," *arXiv preprint arXiv:2008.11904*, 2020 (cit. on p. 3).

[9] V. Muthukumar, K. Vodrahalli, V. Subramanian and A. Sahai, *Harmless interpolation of noisy data in regression*, 2019. DOI: 10.48550/ARXIV.1903.09139. [Online]. Available: https://arxiv.org/abs/1903.09139 (cit. on p. 3).

[10] D. Kobak, J. Lomond and B. Sanchez, "The optimal ridge penalty for real-world high-dimensional data can be zero or negative due to the implicit ridge regularization.," *J. Mach. Learn. Res.*, vol. 21, pp. 169–1, 2020 (cit. on p. 3).

[11] Z. Deng, A. Kammoun and C. Thrampoulidis, "A model of double descent for high-dimensional binary linear classification," *arXiv preprint arXiv:1911.05822*, 2019 (cit. on p. 3).

[12] H. Taheri, R. Pedarsani and C. Thrampoulidis, "Fundamental limits of ridge-regularized empirical risk minimization in high dimensions," in *International Conference on Artificial Intelligence and Statistics*, PMLR, 2021, pp. 2773–2781 (cit. on p. 3).

[13] P. Lolas, "Regularization in high-dimensional regression and classification via random matrix theory," *arXiv preprint arXiv:2003.13723*, 2020 (cit. on p. 3).

[14] F. Mignacco, F. Krzakala, Y. Lu, P. Urbani and L. Zdeborova, "The role of regularization in classification of high-dimensional noisy gaussian mixture," in *International Conference on Machine Learning*, PMLR, 2020, pp. 6874–6883 (cit. on p. 3).

[15] G. R. Kini and C. Thrampoulidis, "Analytic study of double descent in binary classification: The impact of loss," in *2020 IEEE International Symposium on Information Theory (ISIT)*, IEEE, 2020, pp. 2527–2532 (cit. on p. 3).

[16] T. Liang and P. Sur, "A precise high-dimensional asymptotic theory for boosting and minimum-ℓ_1-norm interpolated classifiers," *arXiv preprint arXiv:2002.01586*, 2020 (cit. on p. 3).

[17] H. Taheri, R. Pedarsani and C. Thrampoulidis, "Sharp asymptotics and optimal performance for inference in binary models," in *International Conference on Artificial Intelligence and Statistics*, PMLR, 2020, pp. 3739–3749 (cit. on p. 3).

[18] F. Salehi, E. Abbasi and B. Hassibi, "The impact of regularization on high-dimensional logistic regression," *Advances in Neural Information Processing Systems*, vol. 32, 2019 (cit. on p. 3).

[19] 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. 3).

[20] 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. 3).

[21] 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 p. 3).

[22] C. Thrampoulidis, E. Abbasi and B. Hassibi, *Precise error analysis of regularized m-estimators in high-dimensions*, 2016. DOI: 10.48550/ARXIV.1601.06233. [Online]. Available: https://arxiv.org/abs/1601.06233 (cit. on p. 3).

[23] 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. 3).

[24] 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 pp. 3, 8).

[25] C. Thrampoulidis, S. Oymak and M. Soltanolkotabi, "Theoretical insights into multiclass classification: A high-dimensional asymptotic view," *Advances in Neural Information Processing Systems*, vol. 33, pp. 8907–8920, 2020 (cit. on p. 3).

[26] 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. 3).

[27] 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).

[28] 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, 11, 17).

[29] 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).

[30] 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).

[31] 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).

[32] 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, 11, 15, 17).

[33] 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, 11).

[34] 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).

[35] Y. Gordon, "Some inequalities for gaussian processes and applications," *Israel Journal of Mathematics*, vol. 50, no. 4, pp. 265–289, 1985 (cit. on p. 7).

[36] R. E. Gaunt, "Stein's method for functions of multivariate normal random variables," *arXiv preprint arXiv:1507.08688*, 2015 (cit. on p. 12).

[37] 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. 17).

[38] 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. 19).