University of California
Santa Barbara

Generalization and Optimization in the Interpolation Regime: From Linear Models to Neural Networks

A dissertation submitted in partial satisfaction
of the requirements for the degree

Doctor of Philosophy
in
Electrical and Computer Engineering

by

Hossein Taheri

Committee in charge:

Professor Christos Thrampoulidis, Co-Chair
Professor Haewon Jeong, Co-Chair
Professor Kenneth Rose
Professor Yao Qin

September 2024

The Dissertation of Hossein Taheri is approved.

Professor Kenneth Rose

Professor Yao Qin

Professor Haewon Jeong, Committee Co-Chair

Professor Christos Thrampoulidis, Committee Co-Chair

July 2024

Generalization and Optimization in the Interpolation Regime: From Linear Models to Neural Networks

by

Hossein Taheri

Acknowledgements

Many amazing individuals have helped, inspired, and guided me during the last few years. First, I thank my advisor Prof. Christos Thrampoulidis for his exceptional mentorship and for providing an unstressful and fair work environment during the last five years. His great taste in research, boundless curiosity and passion for understanding have been specially inspirational to me. We explored fascinating concepts and topics across diverse fields, and I was consistently amazed by his deep and extensive knowledge in each area. Beyond research, he was like an older brother to me, offering support in numerous challenging situations. I am grateful for his kindness and generosity.

I would also like to thank other committee members Prof. Haewon Jeong, Prof. Kenneth Rose and Prof. Yao Qin for their insightful questions, suggestions and feedback on this thesis. Their constant support has been essential in the formation of this thesis. I am especially grateful to Prof. Rose for the courses on information theory and pattern recognition; his intuitive and insightful teaching style has greatly inspired me. Likewise I thank Prof. Jeong for the Generative ML course and the thought-provoking questions and discussions we had in class.

I extend my gratitude to my collaborators over the years, whose help has been crucial in shaping my academic experience. I would also like to thank my friends in Santa Barbara who made my Ph.D. journey enjoyable. I am grateful for my friendship with Nate, Rahul, Nima, and many others and for the great memories we shared.

Finally, special thanks go to my parents, my three sisters, and my brother for always being by my side. Even though I could not visit you and there were ten thousands kilometers between us, I felt your presence and positive energy throughout these years. This dissertation is dedicated to all of you.

Curriculum Vitæ

Hossein Taheri

Education

2024	Ph.D. in Electrical and Computer Engineerin (Expected), University of California, Santa Barbara.
2022	M.Sc. in Electrical and Computer Engineering, University of California, Santa Barbara.
2018	B.Sc. in Electrical Engineering and Mathematics (double major), Sharif University of Technology.

Publications

[1] H. Taheri, and C. Thrampoulidis. Generalization and Stability of Interpolating Neural Networks with Minimal Width. In *Journal of Machine Learning Research, 2024.*

[2] P. Deora*, R. Ghaderi*, H. Taheri*, C. Thrampoulidis. On the Optimization and Generalization of Multi-head Attention. (* denotes equal contribution), In *Transactions on Machine learning Research, 2024.*

[3] H. Taheri, and C. Thrampoulidis. On Generalization of Decentralized Learning with Separable Data. In *International Conference on Artificial Intelligence and Statistics, 2023.*

[4] H. Taheri, and C. Thrampoulidis. Fast Convergence in Learning Two-layer Neural Networks with Separable Data. In *AAAI Conference on Artificial Intelligence, 2023.*

[5] H. Taheri, R. Pedarsani, and C. Thrampoulidis. Asymptotic Behavior of Adversarial Training in Binary Linear Classification. In *IEEE transactions on Neural Networks and Learning Systems, 2023.*

[6] 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, 2021.*

[7] H. Taheri, A. Mokhtari, H. Hassani, R. Pedarsani. Quantized Decentralized Stochastic Learning over Directed Graphs. In *International Conference on Machine Learning, 2020.*

[8] 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, 2020.*

[9] H. Taheri, R. Pedarsani, and C. Thrampoulidis. Optimality of Least-squares for classification in Gaussian Mixture Models. In *IEEE International Symposium on Information Theory, 2020.*

[10] A. Reisizadeh, H. Taheri, A. Mokhtari, H. Hassani, R. Pedarsani. Robust and communication-efficient collaborative learning. In *Advances in Neural Information Processing Systems, 2019.*

Abstract

Generalization and Optimization in the Interpolation Regime: From Linear Models to Neural Networks

by

Hossein Taheri

Learning with large models has driven unprecedented advancements across diverse fields of machine learning. As model's size grows the capacity of the model to memorize or interpolate the dataset also increases. Learning under interpolation presents new challenges and opportunities which are not addressed in classical statistical learning theory. In this thesis, we explore the performance of learning methods in the interpolation regime across various models, including linear models and neural networks. Our primary goal is to understand how data and model characteristics influence the convergence behavior of gradient-based methods such as gradient descent and to quantify how well these models generalize to new data.

In the first section, we explore linear models, which are the simplest examples where learning under interpolation can be studied. In particular, we consider empirical risk minimization methods applied on high-dimensional generalized linear models and Gaussian-mixtures. Our goal is to understand the optimal test error performance for such models in an asymptotic set-up where the data-dimension is comparable to the number of training samples. By deriving a system of equations which precisely characterises the test error performance, we are able to find a tight lower-bound on the test error which holds for any convex loss function and ridge-regularization parameter. We then show the bound is tight by proposing a loss function and regularization parameter which achieves the bound. As a corollary, we are able to approximately quantify the sub-optimality of least-squares

depending on the data-model.

Continuing with linear models, we consider adversarial learning with high-dimensional Gaussian-mixture models. Adversarial training, based on empirical risk minimization, currently represents one of the main approaches for defending against adversarial attacks, which involve small but targeted modifications to test data that result in misclassification. We derive precise asymptotic expressions for both standard and adversarial test errors under ℓ_p bounded perturbations within a Gaussian mixture model framework. Our results yield exact error formulas that demonstrate the relationship between adversarial and standard errors and the influence of factors such as the over-parameterization ratio, the data model, and the attack budget.

In the next part of the thesis, we aim to extend our theoretical findings to neural networks. Neural nets are known for their ability to memorize even complex datasets, often achieving near-zero training loss via gradient descent optimization. Despite this capability, they also demonstrate remarkable generalization to new data. We investigate the generalization error (i.e., the gap between training and test errors) of neural networks trained with logistic loss. Our main finding reveals that under a specific data-separability condition, optimal test loss bounds are achievable if the network width is only poly-logarithmically large with respect to the number of training samples. Moreover, our analysis framework which is based on algorithmic stability presents improved generalization bounds and width lower bounds compared to prior works employing alternative methods such as uniform convergence via Rademacher complexity.

Next in chapter five, we again consider the problem of learning two-layer neural networks in the interpolating regime, discussing the role of large-step sizes in speeding up the training. Particularly, we consider the Normalized Gradient Descent (NGD) algorithm where the step-size is chosen inversely proportional to the loss. NGD has proven effective in accelerating the convergence of exponentially-tailed loss functions, such

as exponential and logistic losses, particularly for linear classifiers handling separable data. We demonstrate that for exponentially-tailed losses and two-layer neural nets, NGD achieves a linear convergence rate of the training loss towards the global optimum, provided the iterates identify an interpolating model. This is facilitated by our proof of gradient self-boundedness conditions and the establishment of a log-Lipschitz property. Additionally, we address the generalization capabilities of normalized GD for convex objectives through an algorithmic-stability analysis, showing that it avoids overfitting during training by providing finite-time generalization bounds.

In the final section, we consider the decentralized learning scenario where the data is kept locally among several computing agents which are communicating their parameters over a graph. Our study focuses on decentralized learning in overparameterized settings, where models achieve zero training loss, specifically examining the properties of decentralized gradient descent (DGD) on separable data. Our research provides new finite-time generalization bounds for DGD, extending existing knowledge predominantly focused on centralized learning scenarios. Additionally, we develop enhanced gradient-based methods for decentralized learning with separable data, demonstrating significant orders of magnitude of speed-up compared to previous methods.

These results offer new insights and tools for understanding and improving learning in the interpolation regime across various model architectures and learning paradigms.

Contents

Chapter 1

Introduction

1.1 Motivation

Large-scale learning with big data and large models has shown to be an inseparable part of machine learning in the past few years, achieving breakthrough success in almost all applications from computer vision to language modeling. Using gradient-based methods on empirical risk minimization (ERM) techniques (e.g., logistic regression) is still the the most popular approach for optimization and learning in these scenarios. While implementing an optimization method for a given learning task, the statistician or machine learning engineer is primarily interested in two outcomes: the convergence behavior of the empirical loss and the generalization performance of the learned model to new data. Naturally, obtaining models of small generalization error, that has sufficiently good performance on data beyond the training set, is the ultimate goal of a machine learning task. Nevertheless, the current understanding of the properties of large models and gradient-based optimizers is mostly through heuristics and a theory that explains their properties is only recently emerging.

In this thesis, we present several theoretical and empirical results on the optimization

performance and generalization power of large models, ranging from high-dimensional linear models to neural networks for both centralized and decentralized settings. Our results provide insights on the role of sample-size, model size, loss function and regularization on the performance of learning models in the modern interpolation regime and they complement the findings from classical statistical learning theory.

In the remaining of this chapter, we provide an overview of the content of this thesis.

1.2 Organization of chapters

Chapter 2: In the first chapter of this thesis, we start with high-dimensional linear models which is perhaps the simplest scenario where large-models can be rigorously studied. The study of linear models, nonetheless, provides valuable insight into the general behavior of more complex models. We consider two linear models, namely high-dimensional generalized linear models and Gaussian-mixtures. For the generalized linear models, the goal is recovering the ground signal $w^\star \in \mathbb{R}^d$ from n observations $y_i = \phi(x_i^T w^\star),\ i \in [n]$, where $\phi : \mathbb{R} \to \{\pm 1\}$ is a (possibly random) binary function and $x_i \in \mathbb{R}^d$ denote the data points which are $i.i.d$ sampled from a centered Gaussian distribution. Some examples for ϕ can be the signed model where $\phi(t) = \text{sign}(t)$ and the logistic model where $\phi(t) = 1$ with probability $1/(1+\exp(-t))$ and $\phi(t) = -1$ otherwise. For the Gaussian-mixture model the data points are generated according to $x_i = y_i w^\star + z_i$ where z_i denotes the independent noise. We study the performance of *empirical-risk minimization (ERM)* estimators $\widehat{w}_{f,\lambda}$ that solve the following optimization problem for some *convex* loss function $f : \mathbb{R} \to \mathbb{R}$,

$$\widehat{w}_{f,\lambda} := \arg\min_{w \in \mathbb{R}^d} \frac{1}{n} \sum_{i=1}^{n} f(y_i x_i^\top w) + \lambda \|w\|^2. \tag{1.1}$$

When $\lambda > 0$, choosing convex loss functions such as logistic loss $f(t) = \log(1+\exp(-t))$ and quadratic loss $f(t) = (1-t)^2$ leads to a strongly-convex program. On the other hand, when $\lambda = 0$ it is known (e.g. [1]) that the solution to (1.1) is unique if and only if the over-parameterization is sufficiently small such that $d/n < \delta_f^\star$ where $\delta_f^* = 1$ for the quadratic loss and $\delta_f^\star > 2$ when f is a decreasing loss such as the logistic loss. Our results for the case of zero regularization are valid only given the solution is unique.

Once $\widehat{w}_{f,\lambda}$ is obtained, we can measure its performance with its correlation to $w^\star$ or alternatively by its test accuracy given by,

$$\mathcal{A}(\widehat{w}_{f,\lambda}) := \Pr_{x,y}\left(\text{sign}(x^\top \widehat{w}_{f,\lambda}) = y\right).$$

Certainly, choosing loss functions and regularization parameters that result in estimators with higher values of correlation and test accuracy are more desirable. Our goal in this part is to understand the optimal performance in an asymptotic set-up where the data-dimension (d) is comparable to the number of training samples (n). In particular we are interested in answering the following questions:

How do we quantify the performance of estimators derived by ERM in Eq.(1.1)? What is the optimal performance and how to choose the optimal loss function and regularization parameter in the ERM problem?

In other words, we aim to precisely characterize the optimal loss $f^\star$ and optimal regularization parameter $\lambda^\star$ defined as:

$$(f^\star, \lambda^\star) := \arg \max_{f\in\mathcal{C},\lambda\in\mathbb{R}^+} \mathcal{A}(\widehat{w}_{f,\lambda}), \tag{1.2}$$

where $\mathcal{C}$ denotes the set of convex and real-valued functions. To tackle the problem, first we show the performance of ERM for any convex f and non-negative λ is derived

through a system of three equations. As a demonstration, we note the following which characterizes the error for generalized linear models with label function ϕ:

$$\mathbb{E}\Big[S\,\phi(S)\,\mathcal{M}'_{f,1}\left(\alpha G+\mu S\phi(S);\tau\right)\Big]=-\lambda\mu, \tag{1.3a}$$

$$\tau^2\,\delta\,\mathbb{E}\Big[\left(\mathcal{M}'_{f,1}\left(\alpha G+\mu S\phi(S);\tau\right)\right)^2\Big]=\alpha^2, \tag{1.3b}$$

$$\tau\,\delta\,\mathbb{E}\Big[G\,\mathcal{M}'_{f,1}\left(\alpha G+\mu S\phi(S);\tau\right)\Big]=\alpha(1-\lambda\tau\delta). \tag{1.3c}$$

In the above, $S,G\sim N(0,1)$, $\mathcal{M}'_{f,1}\left(x;\tau\right):=\frac{d\mathcal{M}_f(x;\tau)}{dx}$ and $\mathcal{M}_f(x;\tau):=\min_v \frac{1}{2\tau}(x-v)^2+f(v)$ is known as the Moreau-Envelope of the loss function f (e.g., see [2]). The above equations are derived according to an application of Gaussian comparison inequalities known as the Convex Gaussian Min-max Theorem (CGMT). For more information and background about CGMT we refer the reader to [3, 4, 5]. Denoting $\alpha_{f,\lambda}$ and $\mu_{f,\lambda}$ as the solution obtained from solving the equations in (1.3), we can show that the high-dimensional limit of the test accuracy is then obtained as

$$\lim_{d,m\to\infty}\mathcal{A}(\widehat{w}_{f,\lambda})=\mathbb{P}_{G,S}\left(\frac{\alpha_{f,\lambda}}{\mu_{f,\lambda}}\,G+S\phi(S)>0\right),\quad G,S\sim\mathcal{N}(0,1).$$

This essentially provides an asymptotically precise formula for the performance of ERM solutions based on the chosen loss function and regularization. In general, it can be shown that in order to find the best possible ERM performance, we should find $f^\star\in\mathcal{C}$ and $\lambda^\star>0$ that minimizes $\alpha_{f^\star,\lambda^\star}/\mu_{f^\star,\lambda^\star}$. By exploiting an algebraic structure in the equations, we are able to find a tight lower-bound on the desired quantity which holds for any convex loss function and ridge-regularization parameter. We then show the bound is tight by proposing a loss function and regularization parameter which precisely achieves the bound. This essentially shows that the derived loss and regularization achieve the desired maximality condition of Eq. (1.2). The proposed optimal loss and regularization parameter are

derived based on different problem parameters such as over-parameterization ratio d/n, data model ϕ and the strength of the ground signal (i.e. SNR). As a corollary, we are able to approximately quantify the sub-optimality of quadratic loss. For instance, we show that surprisingly quadratic loss (defined as $f(t) = (1-t)^2$) with optimally-tuned regularization is approximately optimal for logistic data model and "small" signal strength (i.e., small $\|w_0\|$), but the sub-optimality gap grows drastically as signal strength increases.

Chapter 3: Continuing with linear models, we consider the more challenging case of adversarial learning with high-dimensional Gaussian-mixture models. Many modern learning algorithms are known to be susceptible to small carefully crafted perturbations which can fool the classifier to label test points incorrectly. As an example, a traffic sign can be slightly altered such that it is misclassified by a convolutional neural network [6, 7].

One prominent method to robustify the model against such attacks is *adversarial training* [6] which is based on ERM training of adversarial inputs:

$$\widehat{w} = \arg\min_{w\in\mathbb{R}^d} \max_{\|\delta_i\|_q\le\varepsilon_{tr}} \frac{1}{n}\sum_{i=1}^{n} f\left(y_i(x_i+\delta_i)^\top w\right) + \lambda\|w\|^2. \tag{1.4}$$

Once $\widehat{w}$ is obtained, the adversarial test accuracy is defined as:

$$\mathcal{A}(\widehat{w}) := \mathbb{E}_{x,y}\left[\max_{\|\delta\|_q\le\varepsilon_{\rm ts}} \mathbf{1}_{\{y=\mathrm{sign}(\langle x+\delta,\widehat{w}\rangle)\}}\right]. \tag{1.5}$$

In the equations above, $\varepsilon_{tr}, \varepsilon_{ts}$ represent the attack budget during training and test time, respectively. As a side note, we remark that ε_{tr} need not be chosen equal to ε_{ts} and in fact our results show that the optimal value of ε_{tr} is generally larger than ε_{ts} in the *high-dimensional regime*; that is for defending against attacks of strength ε_{ts} during test time, the attack budget during training should be chosen larger than ε_{ts} with the gap increasing as d/n increases.

Adversarial training in Eq. (1.4) currently represents the one of most well-known approaches for defending against adversarial attacks, which involves small but targeted modifications (denoted by δ_i) to training data as a proxy for minimizing the success rate of attacks during test time. While this method has shown practical success, its generalization properties in classification settings remain poorly understood. In particular, we are interested in answering the following questions:

what is the performance of adversarial training in high-dimensional linear models? Does adversarial training lead to degradation in test accuracy on clean data? Is there always a trade-off between standard and adversarial accuracies?

We addresses this questions by offering a detailed analysis of the robustness of adversarial training within the context of binary linear classification. In a similar style as the equations presented for standard training in Eq. (1.3), here we derive precise asymptotic behavior for both standard and adversarial test errors under ℓ_q bounded perturbations where $q \geq 1$. Our approach here allows the use of general Gaussian design beyond isotropic features where $x_i \sim \mathcal{N}(0, \Sigma)$. Our results yield exact error formulas that elucidate the relationship between adversarial and standard errors and the influence of factors such as the over-parameterization ratio, the data model, and the attack budget. Notably, the error curves show that the optimal value of ε_{tr} decreases as the over-parameterization ratio d/n decreases. Perhaps surprisingly our results also illustrate that adversarial training can improve standard accuracy across a range of values for over-parameterization ratio.

Chapter 4: In the next part of the thesis, we consider the more challenging case of binary classification with neural networks. We focus on a one-hidden layer network with

m hidden neurons where

$$\Phi(w, x) := \frac{1}{\sqrt{m}} \sum_{i=1}^{m} a_i \sigma(x^\top w_i). \tag{1.6}$$

In the above, $\sigma : \mathbb{R} \to \mathbb{R}$ is the activation non-linearity, the weight vector $w \in \mathbb{R}^{md}$ is formed by stacking all $w_i \in \mathbb{R}^d$ that is the weight vector entering the i'th hidden neuron. Moreover, $a_i \in \mathbb{R}$ is the second layer weight connecting the neuron i to the output, which is assumed to be fixed during training. The ERM in this case is defined as

$$\min_w F(w) := \frac{1}{n} \sum_{i=1}^{n} f(y_i \Phi(w, x_i)), \tag{1.7}$$

where f is any convex and monotonically decreasing loss such as the logistic loss.

It is a well-known fact that one hidden-layer neural networks are "universal approximators", that is given large enough m, they can approximate any real continuous function arbitrarily well [8]. In practice, large neural networks are renowned for their ability to memorize datasets, often achieving near-zero training loss via gradient descent optimization [9]. In particular, it is often empirically observed that despite the non-convex optimization landscape and no explicit regularization, gradient descent for neural networks of even small width leads the objective to the global optimum. Interestingly, despite the large overparameterization and underlying model's complexity, these solutions also demonstrate sufficiently good generalization capabilities to new data [10, 11]. This motivates the following questions which are the focus of our study:

How wide should a neural network be in the interpolating regime such that convergence to the global optimum is guaranteed? What is the convergence behavior of gradient descent under this over-parameterization condition? How do the solutions found by gradient descent generalize to unseen data?

We address these questions by providing a refined analysis of the generalization behavior of neural networks trained with logistic loss through the lens of algorithmic stability. Algorithmic stability is perhaps the oldest method for studying the generalization abilities of learning algorithms [12, 13]. In particular, we use the *on-average* notion of stability [14] which bounds the expected generalization gap of gradient descent at iteration T as follows:

$$\mathbb{E}\left[F_{\mathcal{D}}\left(w_T\right) - F\left(w_T\right)\right] \lesssim \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n}\left\|w_T - w_T^{\neg i}\right\|\right],$$

where $F_{\mathcal{D}}$ is the test error and both expectations are over data sampling. Additionally, $w_T^{\neg i}$ denotes the output of T iterations of gradient descent when the ith training sample is removed from the training set. Therefore, it suffices to bound the right hand side in the equation above. While algorithmic stability has been proven useful for convex objectives [15, 14], the extensions to non-convex objectives lead to sub-optimal results [16, 17, 18]. Our findings in this chapter, extend the algorithmic stability framework to the non-convex case of neural networks and obtains generalization rates akin to the convex case which hold true even for very small width networks.

We note that the objective in Eq. (1.7) can be highly non-convex depending on the dataset. Therefore unfavorable scenarios where gradient descent may converge to a local optimum or saddle point are plausible. To tackle this issue, in the literature it is often the case that an NTK-separabilty condition [19] for the dataset is assumed which guarantees that the linearization of the network around initialization can separate the dataset with some margin. Formally speaking, the NTK separability condition assumes that for almost surely all n training samples from the data distribution there exists $w^\star \in \mathbb{R}^{md}$ and $\gamma > 0$

such that $\|w^\star\| = 1$ and for all $i \in [n]$,

$$y_i \left\langle \nabla_w \Phi\left(w_0, x_i\right), w^\star \right\rangle \geq \gamma.$$

In other words, the above condition guarantees that at a constant distance from initialization (i.e. $w_0 \in \mathbb{R}^{md}$), there exists a weight vector that can classify the training set with some margin γ.

Our main finding in this section shows that under the NTK separability condition and with standard Gaussian initialization, gradient descent at iteration T decays the training loss at the rate of $\widetilde{O}(\frac{1}{\gamma^2 T})$ and achieves rate-optimal test loss bounds of order $\widetilde{O}(\frac{1}{\gamma^2 n})$, if the network width is only poly-logarithmically large with respect to the number of training samples where,

$$m = \Omega\left(\frac{\log^4(n)}{\gamma^4}\right).$$

This departure from existing generalization outcomes using algorithmic stability, which typically require polynomial width of order $m = \Omega(n^2)$, underscores the significance of our approach. Moreover, our analysis presents improved generalization bounds and width lower bounds compared to prior works employing alternative methods such as uniform convergence via Rademacher complexity. The key to this improvement lies in leveraging the Hessian information of the objective function during gradient descent iterates. We demonstrate that neural networks of poly-logarithmically large width trained by the logistic loss satisfy an *approximate quasi-convexity* property along the gradient descent path. To demonstrate the practical implications of our findings, we specialize our analysis to an XOR-distributed dataset for which we present refined width conditions.

Chapter 5: Continuing with neural networks in the interpolating regime, in chapter 5, we investigate the Normalized Gradient Descent (NGD) algorithm for accelerating the training process. Unlike ordinary GD where the step-size is typically fixed, In NGD the step-size at iteration t is chosen as,

$$\eta_t = \Theta\left(\frac{1}{F(w_t)}\right).$$

Since in the interpolating regime $F(w_t)$ is decaying to zero, the step-size in NGD is growing unboundedly as training progresses. Previous studies on NGD focused on the case of linear classifiers in the interpolating regime [20, 21, 22]. We essentially extend the analysis of NGD to neural networks. While in the previous chapter, we derived the rate of $O(1/T)$ for ordinary gradient descent with neural nets, here we show that under specific separability conditions, the training rate is *exponentially* decaying. Moreover, we study for the first time, the finite-time test error performance of normalized GD for convex objectives. In particular, we provide sufficient conditions for the generalization of NGD and derive bounds of order $O(1/n)$ on the expected generalization error, where n is the training-set size.

Chapter 6: In the final section, we transition from centralized to the decentralized learning scenarios. In decentralized learning, data is kept locally among several computing agents (also called nodes) which are communicating only their parameters over a graph. Formally, the decentralized gradient descent method (DGD) has the following update rule for any node $\ell \in [N]$ at iteration t:

$$w_\ell^{(t+1)} = \sum_{k \in \text{Neigh}(\ell)} A_{\ell k} w_k^{(t)} - \eta_t \nabla F_\ell\Big(w_\ell^{(t)}\Big).$$

In the above, node ℓ computes a weighted summation of its neighbors parameters and then implements one step of gradient descent on its local loss $F_\ell(w_\ell^{(t)})$. Each node has access to n_ℓ training points and overall DGD updates aim at minimizing the global objective $F(\cdot)$ with n training points where $F(w) := \frac{1}{N}\sum_{\ell=1}^{N} F_\ell(w)$.

Our study focuses on decentralized learning for classification tasks in over-parameterized settings, where the model has enough capacity to perfectly classify the whole dataset correctly and thus can achieve zero training loss. Our research provides new finite-time generalization bounds for DGD, extending existing knowledge predominantly focused on centralized learning scenarios. Remarkably, these bounds for DGD are nearly equivalent in order to those of centralized learning. In particular, for logistic regression we find that with $\eta = \frac{1}{T^{1/3}}$ and after T iterations, DGD with n training points reaches the test loss rate of order,

$$\widetilde{O}\left(\frac{1}{T^{2/3}} + \frac{1}{n}\right),$$

which is comparable to the rate of $\widetilde{O}(\frac{1}{T} + \frac{1}{n})$ for centralized settings. Additionally, we develop enhanced gradient methods based on normalized gradient descent for decentralized learning with separable data. Our experiments demonstrate significant improvements in training speed and finite-time generalization performance for our proposed algorithm in a logistic regression task with linearly separable data.

Chapter 2

High-dimensional Linear Models: Sharp Asymptotics and Fundamental Limits

2.1 Introduction

Empirical Risk Minimization (ERM) includes a wide family of statistical inference algorithms that are popular in estimation and learning tasks encountered in a range of applications in signal processing, communications and machine learning. ERM methods are often efficient in implementation, but first one needs to make certain choices: such as, choose an appropriate loss function and regularization function, and tune the regularization parameter. Classical statistics have complemented the practice of ERM with an elegant theory regarding optimal such choices, as well as, fundamental limits, i.e., tight bounds on their performance, e.g., [23]. These classical theories typically assume that the size m of the set of observations is much larger than the dimension n of the parameter to be estimated, i.e., $m \gg n$. In contrast, modern inference problems are typically high-dimensional, i.e. m and n are of the same order and often $n > m$ [24, 25, 26]. This chapter studies the fundamental limits of convex ERM in high-dimensions for generalized

linear models.

Generalized linear models (GLM) relate the response variable y_i to a linear model $\mathbf{a}_i^T\mathbf{x}_0$ via a link function: $y_i = \varphi(\mathbf{a}_i^T\,\mathbf{x}_0)$. Here, $\mathbf{x}_0 \in \mathbb{R}^n$ is a vector of true parameters and $\mathbf{a}_i \in \mathbb{R}^n,\ i \in [m]$ are the feature (or, measurement) vectors. Following the ERM principle, $\mathbf{x}_0$ can be estimated by the minimizer of the empirical risk $\frac{1}{m}\sum_{i=1}^m \mathcal{L}\left(y_i, \mathbf{a}_i^T\mathbf{x}\right)$ for a chosen loss function $\mathcal{L}$. Typically, ERM is combined with a regularization term and among all possible choices arguably the most popular one is ridge regularization, which gives rise to ridge-regularized ERM (RERM, in short):

$$\widehat{\mathbf{x}}_{\mathcal{L},\lambda} = \arg\min_{\mathbf{x}\in\mathbb{R}^n} \frac{1}{m}\sum_{i=1}^m \mathcal{L}\left(y_i, \mathbf{a}_i^T\mathbf{x}\right) + \frac{\lambda}{2}\|\mathbf{x}\|_2^2. \tag{2.1}$$

This chapter aims to provide answers to the following questions on fundamental limits of (2.1): *What is the minimum achievable (estimation/prediction) error of $\widehat{\mathbf{x}}_{\mathcal{L},\lambda}$? How does this depend on the link function φ and how to choose $\mathcal{L}$ and λ to achieve it? What is the sub-optimality gap of popular choices such as ridge-regularized least-squares (RLS)? How do the answers to these questions depend on the over-parameterization ratio n/m?* We provide answers to the questions above for the following two popular instances of GLMs. *Linear models:* $y_i = \mathbf{a}_i^T\mathbf{x}_0 + z_i$, where $z_i \overset{\text{iid}}{\sim} P_Z,\ i \in [m]$. As is typical, for linear models, we measure performance of $\widehat{\mathbf{x}}_{\mathcal{L},\lambda}$ with the squared error: $\|\widehat{\mathbf{x}}_{\mathcal{L},\lambda} - \mathbf{x}_0\|_2^2$.

Binary models: $y_i = f(\mathbf{a}_i^T\mathbf{x}_0),\ i \in [m]$ for a (possibly random) link function outputing values $\{\pm 1\}$, e.g., logistic, probit and signed models. We measure estimation performance in terms of (normalized) correlation $(\widehat{\mathbf{x}}_{\mathcal{L},\lambda}^T\,\mathbf{x}_0)\Big/\|\widehat{\mathbf{x}}_{\mathcal{L},\lambda}\|_2\|\mathbf{x}_0\|_2$ and prediction performance in terms of classification error $\mathbb{P}(y \neq \text{sign}(\widehat{\mathbf{x}}_{\mathcal{L},\lambda}^T\,\mathbf{a}))$ where the probability is over a fresh data point $(\mathbf{a}, y)$.

All our results are valid under the following two assumptions.

Assumption 2.1.1 (High-dimensional asymptotics)**.** *Throughout the chapter, we assume the high-dimensional limit where $m, n \to \infty$ at a fixed ratio $\delta = m/n > 0$.*

Assumption 2.1.2 (Gaussian features)**.** *The feature vectors $\mathbf{a}_i \in \mathbb{R}^n, i \in [m]$ are iid $\mathcal{N}(\mathbf{0}, \mathbf{I}_n)$.*

Overview of Contributions. We are now ready to summarize this chapter's main contributions.

• For linear models, we prove a lower bound on the squared-estimation error of RERM; see Theorem 2.2.1. We start with a system of two nonlinear equations that is parametrized by the loss $\mathcal{L}$ and the regularizer λ, and determines the high-dimensional limit of the error for the corresponding $\mathcal{L}$ and λ [26, **?**]. By identifying an algebraic structure in these equations, we establish a lower bound on their solution that holds for all choices of $\mathcal{L}$ and λ.

• For binary models, we first derive a system a of three nonlinear equations whose unique solution characterizes the statistical performance (correlation or classification error) of RERM under mild assumptions on the loss and link functions $\mathcal{L}$ and f; see Theorem 2.3.1. Previous works have only considered specific loss and link functions or *no* regularization. Second, we use this system of equations to upper bound the accuracy over this class of $(\mathcal{L}, f)$-pairs; see Theorem 2.3.2.

• Importantly, we present a recipe for optimally tuning $\mathcal{L}$ and λ in both linear and binary models; see Lemmas 2.2.1 and 2.3.1. For specific models, such as linear model with additive exponential noise, binary logistic and signed data, we numerically show that the optimal loss function is convex and we use gradient-descent to optimize it. The numerical simulations perfectly match with the theoretical predictions suggesting that our bounds are tight.

- We derive simple closed-form approximations to the aforementioned bounds; see Corollaries 2.2.1 (linear) and 2.3.1 (binary). These simple (yet tight) expressions allow us to precisely quantify the sub-optimality of ridge-regularized least-squares (RLS). For instance, we show that *optimally-tuned RLS is (perhaps surprisingly) approximately optimal for logistic data and small signal strength, but the sub-optimality gap grows drastically as signal strength increases.* In the Appendix, we also include comparisons to ERM without regularization and to a simple averaging method.

2.1.1 Prior Work

Our results fit in the rapidly growing recent literature on *sharp* asymptotics of (possibly non-smooth) convex optimization-based estimators, e.g., [27, 28, 29, 30, 31, 32, 33, 34, 26, 35, 36, 37, 38, 39, 40, 41, 42, 43, 42, 43, 44, 45]. Most of these works study linear models. Extensions to generalized linear models for the special case of regularized LS were studied in [46], while more recently there has been a surge of interest in RERM methods tailored to binary models (such as logistic regression or SVM) [47, 1, 48, 49, 50, 51, 52, 53, 54, 55, 56, 57].

Out of these works relatively few have focused on fundamental limits among families of ERM (rather than specific instances). The papers [58, 36, 59] derive lower bounds and optimal loss functions for the squared error of (unregularized) ERM for linear models. In a related work, [60] studies robustness of these methods to the noise distribution. More recently, [41] performed an in-depth analysis of fundamental limits of convex-regularized LS for linear models of structured signals. For binary models, upper bounds on the correlation of un-regularized ERM were only recently derived in [53]. This chapter contributes to this line work. For linear models, we build on corresponding sharp error characterizations in [35, 37] to extend the results of [58, 36, 59] to ridge-regularized ERM. Specifically, our

results hold for all values of $\delta > 0$ including the, so called, overparameterized regime $\delta < 1$. For binary models, our contribution is twofold: (i) we present sharp asymptotic characterizations for RERM for a wide class of loss and link functions; (ii) we use these to extend the correlation bounds of [53] to the regularized case.

On a technical level, the sharp asymptotics are derived using the convex Gaussian min-max Theorem (CGMT) [33, **?**]. In particular, we follow the machinery introduced in [46, 51, 52, 53, 54] that applies the CGMT to binary models and predicts the performance in terms of a system of few nonlinear equations. Our main technical contribution here is proving existence and uniqueness of the solutions to these equations, which is critical as it guarantees that our performance bounds hold for a wide class of loss and link functions.

Notation. We use boldface notation for vectors. We write $i \in [m]$ for $i = 1, 2, \ldots, m$. For a random variable H with density $p_{_H}(h)$ that has a derivative $p'_{_H}(h), \forall h \in \mathbb{R}$, we define its *Fisher information* $\mathcal{I}(H) := \mathbb{E}[\big(p'_{_H}(h)/p_{_H}(h)\big)^2]$. We write $\mathcal{M}_{\mathcal{L}}\left(x;\tau\right) := \min_v \frac{1}{2\tau}(x-v)^2 + \mathcal{L}(v)$, for the *Moreau envelope function* and $\text{prox}_{\mathcal{L}}\left(x;\tau\right) := \arg\min_v \frac{1}{2\tau}(x-v)^2 + \mathcal{L}(v)$ for the *proximal operator* of the loss $\mathcal{L} : \mathbb{R} \to \mathbb{R}$ at x with parameter $\tau > 0$. We denote the first order derivative of the Moreau-envelope function w.r.t x as: $\mathcal{M}'_{\mathcal{L},1}\left(x;\tau\right) := \frac{\partial \mathcal{M}_{\mathcal{L}}(x;\tau)}{\partial x}$. Finally, for a sequence of random variables $\mathcal{X}_{m,n}$ that converges in probability to some constant c in the high-dimensional asymptotic limit of Assumption 2.1.1, we write $\mathcal{X}_{m,n} \xrightarrow{P} c$.

2.2 Linear Models

Consider data $(y_i, \mathbf{a}_i)$ from an additive noisy linear model: $y_i = \mathbf{a}_i^T \mathbf{x}_0 + z_i,\ z_i \overset{\text{iid}}{\sim} P_Z,\ i \in [m]$.

Assumption 2.2.1 (Noise distribution)**.** *The noise variables z_i are iid distributed as $Z \sim P_Z$, $i \in [m]$, for a distribution P_Z with zero mean and finite nonzero second moment.*

For loss functions that are lower semicontinuous (lsc), proper, and convex we focus on the following version of (2.1) that is tailored to linear models:

$$\widehat{\mathbf{x}}_{\mathcal{L},\lambda} := \arg\min_{\mathbf{x}\in\mathbb{R}^n} \ \frac{1}{m}\sum_{i=1}^{m}\mathcal{L}\left(y_i - \mathbf{a}_i^T\mathbf{x}\right) + \frac{\lambda}{2}\|\mathbf{x}\|^2. \tag{2.2}$$

We assume without loss of generality that $\|\mathbf{x}_0\|_2 = 1$ [1].

2.2.1 Background on Asymptotic Performance

Prior works have investigated the limit of the squared error $\|\widehat{\mathbf{x}}_{\mathcal{L},\lambda} - \mathbf{x}_0\|^2$ [26, ?]. Specifically, consider the following system of two equations in two unknowns α and τ:

$$\mathbb{E}\Big[\Big(\mathcal{M}'_{\mathcal{L},1}\left(\alpha\, G + Z;\tau\right)\Big)^2\Big] = \frac{\alpha^2 - \lambda^2\delta^2\tau^2}{\tau^2\,\delta}, \tag{2.3a}$$

$$\mathbb{E}\Big[G\cdot\mathcal{M}'_{\mathcal{L},1}\left(\alpha\, G + Z;\tau\right)\Big] = \frac{\alpha(1-\lambda\delta\tau)}{\tau\,\delta}, \tag{2.3b}$$

where $G\sim\mathcal{N}(0,1)$ and $Z\sim P_Z$ is the noise variable. It has been shown in [26, ?] that under appropriate regularity conditions on $\mathcal{L}$ and the noise distribution P_Z, the system of equations above has a unique solution $(\alpha_{\mathcal{L},\lambda}>0, \tau_{\mathcal{L},\lambda}>0)$ and $\alpha^2_{\mathcal{L},\lambda}$ is the HD limit of the squared-error, i.e.,

$$\|\widehat{\mathbf{x}}_{\mathcal{L},\lambda} - \mathbf{x}_0\|_2^2 \xrightarrow{P} \alpha^2_{\mathcal{L},\lambda}. \tag{2.4}$$

Here, we derive tight lower bounds on $\alpha^2_{\mathcal{L},\lambda}$ over both the choice of $\mathcal{L}$ and λ. Our starting point is the asymptotic characterization in (2.4), i.e., our results hold for all loss functions and regularizer parameters for which (2.3) has a unique solution that characterizes the

[1] Suppose that $\|\mathbf{x}_0\|_2 = r > 0$. Then, the optimization problem in (2.2) can be transformed to the case $\widetilde{\mathbf{x}}_0 := \mathbf{x}_0/r$ (hence $\|\widetilde{\mathbf{x}}_0\| = 1$) by setting $\widetilde{\mathcal{L}}(t) := \mathcal{L}(rt)$, $\widetilde{\lambda} := r^2\lambda$ and $\widetilde{Z} = Z/r$. This implies that the results of Section 2.2.2 can be reformulated by replacing Z with $\widetilde{Z}$.

HD limit of the square-error. To formalize this, we define the following collection of loss functions $\mathcal{L}$ and noise distributions P_Z:

$$\mathcal{C}_{\text{lin}} := \Big\{ (\mathcal{L}, P_Z) \,\Big|\, \forall \lambda > 0\text{: (2.3) has a unique bounded solution } (\alpha_{\mathcal{L},\lambda} > 0, \tau_{\mathcal{L},\lambda} > 0) \text{ and (2.4) holds} \Big\}. \tag{2.5}$$

We refer the reader to [26, Thm. 1.1] and [?, Thm. 2] for explicit characterizations of $(\mathcal{L}, P_Z)$ that belong to $\mathcal{C}_{\text{lin}}$. We conjecture that some of these regularity conditions (e.g., the differentiability requirement) can in fact be relaxed. While this is beyond the scope of this chapter, if this is shown then automatically the results of this chapter formally hold for a richer class of loss functions.

2.2.2 Fundamental Limits and Optimal Tuning

Our first main result, stated as Theorem 2.2.1 below, establishes a tight bound on the achievable values of $\alpha^2_{\mathcal{L},\lambda}$ for all regularization parameters $\lambda > 0$ and all choices of $\mathcal{L}$ such that $(\mathcal{L}, P_Z) \in \mathcal{C}_{\text{lin}}$.

Theorem 2.2.1 (Lower bound on $\alpha_{\mathcal{L},\lambda}$)**.** *Let Assumptions 2.1.1, 2.1.2 and 2.2.1 hold. For $G \sim \mathcal{N}(0,1)$ and noise random variable $Z \sim P_Z$, consider a new random variable $V_a := a\,G + Z$, parameterized by $a \in \mathbb{R}$. Fix any $\delta > 0$ and define $\alpha_\star = \alpha_\star(\delta, P_Z)$ as follows:*

$$\alpha_\star := \min_{0 \le x < 1/\delta} \left[a > 0 : \; \frac{\delta(a^2 - x^2\,\delta^2)\,\mathcal{I}(V_a)}{(1 - x\,\delta)^2} = 1 \right]. \tag{2.6}$$

For any $\mathcal{L}$ such that $(\mathcal{L}, P_Z) \in \mathcal{C}_{\text{lin}}$, $\lambda > 0$ and $\alpha^2_{\mathcal{L},\lambda}$ denoting the respective high-dimensional limit of the squared-error as in (2.4), it holds that $\alpha_{\mathcal{L},\lambda} \ge \alpha_\star$.

The proof of the theorem is presented in Section 2.5.3.2. This includes showing that

the minimization in (2.6) is feasible for any $\delta > 0$. In general, the lower bound $\alpha_\star$ can be computed by numerically solving (2.6). For special cases of noise distributions (such as Gaussian), it is possible to analytically solve (2.6) and obtain a closed-form formula for $\alpha_\star$, which is easier to interpret. While this is only possible for few special cases, our next result establishes a simple closed-form lower bound on $\alpha_\star$ that is valid under only mild assumptions on P_Z. For convenience, let us define $h_\delta : \mathbb{R}_{>0} \to \mathbb{R}_{>0}$,

$$h_\delta(x) := \frac{1}{2}\Big(1 - x - \delta + \sqrt{(1+\delta+x)^2 - 4\delta}\Big)\,. \tag{2.7}$$

The subscript δ emphasizes the dependence of the function on the oversampling ratio δ. We also note for future reference that h_δ is strictly increasing for all fixed $\delta > 0$.

Corollary 2.2.1 (Closed-form lower bound on $\alpha_\star^2$)**.** *Let $\alpha_\star$ be as in (2.6) under the assumptions of Theorem 2.2.1. Assume that p_Z is differentiable and takes strictly positive values on the real line. Then, it holds that*

$$\alpha_\star^2 \geq h_\delta\left(1/\mathcal{I}(Z)\right).$$

Moreover, the equality holds if and only if $Z \sim \mathcal{N}(0, \zeta^2)$ for $\zeta > 0$.

The proof of Corollary presented in Section 2.5.3.5 shows that the gap between the actual value of $\alpha_\star$ and $h_\delta(1/\mathcal{I}(Z))$ depends solely on the distribution of Z. Informally: the more Z resembles a Gaussian, the smaller the gap. The simple approximation of Corollary 2.2.1 is key for comparing the performance of optimally tuned RERM to optimally-tuned RLS in Section 2.2.3.

A natural question regarding the lower bound of Theorem 2.2.1 is whether it is tight. Indeed, the lower bound cannot be improved in general. This can be argues as follows. Consider the case of additive Gaussian noise $Z \sim \mathcal{N}(0, \zeta^2)$ for which $\mathcal{I}(Z) = 1/\mathbb{E}[Z^2] =$

$1/\zeta^2$. On the one hand, Corollary 2.2.1 shows that $\alpha_\star^2 \geq h_\delta(\zeta^2)$ and on the other hand, we show in Lemma 2.2.2 that optimally-tuned RLS achieves this bound, i.e., $\alpha^2_{\ell_2,\lambda_{\text{opt}}} = h_\delta(\zeta^2)$. Thus, the case of Gaussian noise shows that the bound of Theorem 2.2.1 cannot be improved in general.

Our next result reinforces the claim that the bound is actually tight for a larger class of noise distributions.

Lemma 2.2.1 (Optimal tuning for linear RERM). *For given $\delta > 0$ and P_Z, let $(\alpha_\star > 0, x_\star \in [0, 1/\delta))$ be the optimal solution in the minimization in (2.6). Denote $\lambda_\star = x_\star$ and define $V_\star := \alpha_\star G + Z$. Consider the loss function $\mathcal{L}_\star : \mathbb{R} \to \mathbb{R}$ defined as $\mathcal{L}_\star(v) := -\mathcal{M}_{\frac{\alpha_\star^2 - \lambda_\star^2\delta^2}{1-\lambda_\star\delta}\cdot\log\left(p_{V_\star}\right)}(v;1)$. Then for $\mathcal{L}_\star$ and $\lambda_\star$, the equations (2.3) satisfy $(\alpha, \tau) = (\alpha_\star, 1)$.*

We leave for future work coming up with sufficient conditions on P_Z under which $(\mathcal{L}_\star, P_Z) \in \mathcal{C}_{\text{lin}}$, which would imply that the bound of Theorem 2.2.1 is achieved by choosing $\mathcal{L} = \mathcal{L}_\star$ and $\lambda = \lambda_\star$ in (2.2). In Figures 2.1(Left) and 2.2(Top Left), we numerically (by using gradient descent) evaluate the performance of the proposed loss function $\mathcal{L}_\star$, in the case of Laplacian noise, suggesting that it achieves the lower bound $\alpha_\star$ in Theorem 2.2.1. See also Figure 2.3(Left) for an illustration of $\mathcal{L}_\star$.

2.2.3 The Sub-optimality Gap of RLS in Linear Models

We rely on Theorem 2.2.1 to investigate the statistical gap between least-squares (i.e. $\mathcal{L}(t) = t^2$ in (2.2)) and the optimal choice of $\mathcal{L}$. As a first step, the lemma below computes the high-dimensional limit of optimally regularized RLS.

Lemma 2.2.2 (Asymptotic error of optimally regularized RLS). *Fix $\delta > 0$ and noise distribution P_Z. Let $\widehat{\mathbf{x}}_{\ell_2,\lambda}$ be the solution to λ-regularized least-squares. Further let $\alpha_{\ell_2,\lambda}$ denote the high-dimensional limit of $\|\widehat{\mathbf{x}}_{\ell_2,\lambda} - \mathbf{x}_0\|_2^2$. Then, $\lambda \mapsto \alpha_{\ell_2,\lambda}$ is minimized at*

$\lambda_{\text{opt}} = 2\,\mathbb{E}[Z^2]$ *and*

$$\alpha^2_{\ell_2,\lambda_{\text{opt}}} := h_\delta\left(\mathbb{E}\left[Z^2\right]\right).$$

We combine this result with the closed-form lower bound of Corollary 2.2.1 to find that $\alpha_\star^2/\alpha^2_{\ell_2,\lambda_{\text{opt}}} \in [\omega_\delta, 1]$ for

$$\omega_\delta := \frac{h_\delta\left(1/\mathcal{I}(Z)\right)}{h_\delta\left(\mathbb{E}\left[Z^2\right]\right)}.$$

The fact that $\omega_\delta \leq 1$ follows directly by the increasing nature of the function h_δ and the Cramer-Rao bound $\mathbb{E}[Z^2] \geq 1/\mathcal{I}(Z)$ (see Proposition 2.5.3(c)). Moreover, using analytic properties of the function h_δ it is shown in Section 2.5.3.6 that

$$\alpha_\star^2/\alpha^2_{\ell_2,\lambda_{\text{opt}}} \geq \omega_\delta \geq \max\left\{1-\delta\,,\,\left(\mathcal{I}(Z)\,\mathbb{E}[Z^2]\right)^{-1}\right\}. \tag{2.8}$$

The first term in the lower bound in (2.8) reveals that in the highly over-parameterized regime ($\delta \ll 1$), it holds $\omega_\delta \approx 1$. Thus, optimally-regularized LS becomes optimal. More generally, in the overparameterized regime $0 < \delta < 1$, the squared-error of optimally-tuned LS is no worse than $(1-\delta)^{-1}$ times the optimal performance among all convex ERM.

The second term in (2.8) is more useful in the underparameterized regime $\delta \geq 1$ and captures the effect of the noise distribution via the ratio $\left(\mathcal{I}(Z)\,\mathbb{E}[Z^2]\right)^{-1} \leq 1$ (which is closely related to the classical Fisher information distance studied e.g. in [61]). From this and the fact that $\mathcal{I}(Z) = 1/\mathbb{E}[Z^2]$ iff $Z \sim \mathcal{N}(0,\zeta^2)$ we conclude that ω_δ attains its maximum value 1 (thus, optimally-tuned LS is optimal) when Z is Gaussian. For completeness, we remark that [62] has shown that when $Z \sim \mathcal{N}(0,\zeta^2)$, then the minimum mean square error (MMSE) is also given by $h_\delta(\zeta^2)$. To further illustrate that our results are informative for general noise distributions, consider the case of Laplacian noise, i.e., $Z \sim \texttt{Laplace}(0,b^2)$. Using $\mathbb{E}[Z^2] = 2b^2$ and $\mathcal{I}(Z) = b^{-2}$ in (2.8) we obtain $\omega_\delta \geq 1/2$, for all $b > 0$ and $\delta > 0$. Therefore we find that optimally-tuned RLS achieves squared-error

that is at most twice as large as the optimal error, i.e. if $Z \sim \texttt{Laplace}(0, b^2)$, $b > 0$ then for all $\delta > 0$ it holds that $\alpha^2_{\ell_2, \lambda_{\text{opt}}} \leq 2\,\alpha^2_\star$. See also Figures 2.1 and 2.2 for a numerical comparison.

2.3 Binary Models

Consider data $(y_i, \mathbf{a}_i)$, $i \in [m]$ from a binary model: $y_i = f(\mathbf{a}_i^T \mathbf{x}_0)$ where f is a (possibly random) link function outputting $\{\pm 1\}$.

Assumption 2.3.1 (Link function)**.** *The link function f satisfies $\nu_f := \mathbb{E}\,[S\, f(S)] \neq 0$, for $S \sim \mathcal{N}(0, 1)$.*[2]

Under Assumptions 2.1.1, 2.1.2 and 2.3.1 we study the ridge-regularized ERM for binary measurements:

$$\widehat{\mathbf{w}}_{\mathcal{L},\lambda} := \arg\min_{\mathbf{w} \in \mathbb{R}^n} \; \frac{1}{m} \sum_{i=1}^{m} \mathcal{L}\left(y_i \mathbf{a}_i^T \mathbf{w}\right) + \frac{\lambda}{2} \|\mathbf{w}\|^2. \tag{2.9}$$

We also assume that $\|\mathbf{x}_0\|_2 = 1$ since the signal strength can always be absorbed in the link function, i.e., if $\|\mathbf{x}_0\|_2 = r > 0$ then the results continue to hold for a new link function $\widetilde{f}(t) := f(rt)$.

2.3.1 Asymptotic Performance

In contrast to linear models where we focused on squared error, for binary models, a more relevant performance measure is normalized correlation $\text{corr}\,(\,\widehat{\mathbf{w}}_{\mathcal{L},\lambda}\,,\,\mathbf{x}_0\,)$. Our first result determines the limit of $\text{corr}\,(\,\widehat{\mathbf{w}}_{\mathcal{L},\lambda}\,,\,\mathbf{x}_0\,)$. Specifically, we show that for a wide class

[2] See Section 2.5.4.1 for further discussion.

of loss functions it holds that

$$\rho_{\mathcal{L},\lambda} := \text{corr}\left(\widehat{\mathbf{w}}_{\mathcal{L},\lambda}, \mathbf{x}_0\right) := \frac{|\widehat{\mathbf{w}}_{\mathcal{L},\lambda}^T \mathbf{x}_0|}{\|\widehat{\mathbf{w}}_{\mathcal{L},\lambda}\|_2 \|\mathbf{x}_0\|_2} \xrightarrow{P} \sqrt{\frac{1}{1+\sigma_{\mathcal{L},\lambda}^2}}, \tag{2.10}$$

where $\sigma_{\mathcal{L},\lambda}^2 := \alpha_{\mathcal{L},\lambda}^2/\mu_{\mathcal{L},\lambda}^2$ and $(\alpha_{\mathcal{L},\lambda}, \mu_{\mathcal{L},\lambda})$ are found by solving the following system of three nonlinear equations in three unknowns (α, μ, τ), for $G, S \overset{\text{iid}}{\sim} \mathcal{N}(0,1)$:

$$\mathbb{E}\Big[S\, f(S)\, \mathcal{M}'_{\mathcal{L},1}\left(\alpha G + \mu S f(S); \tau\right)\Big] = -\lambda\mu, \tag{2.11a}$$

$$\tau^2\, \delta\, \mathbb{E}\Big[\left(\mathcal{M}'_{\mathcal{L},1}\left(\alpha G + \mu S f(S); \tau\right)\right)^2\Big] = \alpha^2, \tag{2.11b}$$

$$\tau\, \delta\, \mathbb{E}\Big[G\, \mathcal{M}'_{\mathcal{L},1}\left(\alpha G + \mu S f(S); \tau\right)\Big] = \alpha(1 - \lambda\tau\delta). \tag{2.11c}$$

To formalize this, we define the following collection of loss and link functions:

$$\mathcal{C}_{\text{bin}} := \Big\{(\mathcal{L},\, f)\Big|\, \forall\lambda > 0\text{: (2.11) has a unique bounded solution } (\alpha_{\mathcal{L},\lambda} > 0, \mu_{\mathcal{L},\lambda}, \tau_{\mathcal{L},\lambda} > 0) \text{ and (2.10) holds}\Big\}. \tag{2.12}$$

Theorem 2.3.1 (Asymptotics for binary RERM)**.** *Let Assumptions 2.1.1 and 2.1.2 hold and* $\|\mathbf{x}_0\|_2 = 1$. *Let* $f : \mathbb{R} \to \{-1, +1\}$ *be a link function satisfying Assumption 2.3.1. Further assume a loss function* $\mathcal{L}$ *with the following properties:* $\mathcal{L}$ *is convex, twice differentiable and bounded from below such that* $\mathcal{L}'(0) \neq 0$ *and for* $G \sim \mathcal{N}(0,1)$*, we have* $\mathbb{E}[\mathcal{L}(G)] < \infty$. *Then, it holds that* $(\mathcal{L}, f) \in \mathcal{C}_{\text{bin}}$.

We prove Theorem 2.3.1 in Section 2.5.2. Previous works have considered special instances of this: [48, 52] study unregularized and regularized logistic-loss for the logistic binary model, while [53] studies strictly-convex ERM without regularization. Here, we follow the same approach as in [52, 53], who apply the convex Gaussian min-max theorem

(CGMT) to relate the performance of RERM to an auxiliary optimization (AO) problem whose first-order optimality conditions lead to the system of equations in (2.11). Our technical contribution in proving Theorem 2.3.1 is proving existence and uniqueness of solutions to (2.11) for a broad class of convex losses. As a final remark, the solution to (2.11) (specifically, the parameter $\sigma_{\mathcal{L},\lambda}^2$) further determines the high-dimensional limit of the classification error for a fresh feature vector $\mathbf{a} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_n)$ (see Section 2.5.4.2) :

$$\mathcal{E}_{\mathcal{L},\lambda} := \mathbb{P}\left(f\left(\mathbf{a}^T\mathbf{x}_0\right)\left(\mathbf{a}^T\widehat{\mathbf{w}}_{\mathcal{L},\lambda}\right) < 0\right) \xrightarrow{P} \mathbb{P}\left(\sigma_{\mathcal{L},\lambda}\, G + S f(S) < 0\right), \quad G, S \overset{\text{iid}}{\sim} \mathcal{N}(0,1). \tag{2.13}$$

2.3.2 Fundamental Limits and Optimal Tuning

Thus far, we have shown in (2.10) and (2.13) that $\sigma_{\mathcal{L},\lambda}$ predicts the high-dimensional limit of the correlation and classification-error of the RERM solution $\widehat{\mathbf{w}}_{\mathcal{L},\lambda}$. In fact, smaller values for $\sigma_{\mathcal{L},\lambda}$ result in better performance, i.e. higher correlation and classification accuracy (see Section 2.5.4.2). In this section we derive a lower bound on $\sigma_{\mathcal{L},\lambda}$ characterizing the statistical limits of RERM for binary models.

Theorem 2.3.2 (Lower Bound on $\sigma_{\mathcal{L},\lambda}$)**.** *Let Assumptions 2.1.1, 2.1.2 and 2.3.1 hold. For $G, S \overset{iid}{\sim} \mathcal{N}(0,1)$ define the random variable $W_s := s\,G + S\,f(S)$ parameterized by $s \in \mathbb{R}$. Fix any $\delta > 0$ and define*

$$\sigma_\star = \sigma_\star(\delta, f) := \min_{0 \le x < 1/\delta} \left[s > 0 : \frac{1 - s^2(1 - s^2\mathcal{I}(W_s))}{\delta s^2(s^2\mathcal{I}(W_s) + \mathcal{I}(W_s) - 1)} - 2x + x^2\delta(1 + \frac{1}{s^2}) = 1 \right]. \tag{2.14}$$

For any $(\mathcal{L}, f) \in \mathcal{C}_{\text{bin}}$, $\lambda > 0$ and $\sigma_{\mathcal{L},\lambda}^2$ the respective high-dimensional limit of the error as in (2.10), it holds that $\sigma_{\mathcal{L},\lambda} \ge \sigma_\star$.

We prove Theorem 2.3.2 in Section 2.5.4.3, where we also show that the minimization in (2.14) is always feasible. In view of (2.10) and (2.13) the theorem's lower bound translates to an upper bound on correlation and test accuracy. Note that $\sigma_\star$ depends on the link function only through the Fisher information of the random variable $s\,G + S\,f(S)$. This parallels the lower bound of Theorem 2.2.1 on linear models with the random variable $S\,f(S)$ effectively playing the role of the noise variable Z.

Next we present a useful closed-form lower bound for $\sigma_\star$. For convenience define the function $H_\delta : \mathbb{R}_{>1} \to \mathbb{R}_{>0}$ parameterized by $\delta > 0$ as following,

$$H_\delta(x) := 2\left(-\delta - x + \delta\, x + \sqrt{(-\delta - x + \delta\, x)^2 + 4\delta(x-1)}\right)^{-1}. \tag{2.15}$$

Corollary 2.3.1 (Lower bound on $\sigma_\star$)**.** *Let $\sigma_\star$ be as in (2.14). Fix any $\delta > 0$ and assume that $f(\cdot)$ is such that the random variable $Sf(S)$ has a differentiable and strictly positive probability density on the real line. Then,*

$$\sigma_\star^2 \geq H_\delta\left(\, \mathcal{I}(Sf(S))\, \right).$$

Corollary 2.3.1 can be viewed as an extension of Corollary 2.2.1 to binary models. The proof of the corollary presented in Section 2.5.4.6 further reveals that the more the distribution of $Sf(S)$ resembles a Gaussian distribution, the tighter the gap is, with equality being achieved if and only if $Sf(S)$ is Gaussian.

Our next result strengthens the lower bound of Theorem 2.3.2 by showing existence of a loss function and regularizer parameter for which the system of equations (2.11) has a solution leading to $\sigma_\star$.

Lemma 2.3.1 (Optimal tuning for binary RERM)**.** *For given $\delta > 0$ and binary link function f, let $(\sigma_\star > 0, x_\star \in [0, 1/\delta))$ be the optimal solution in the minimization in (2.14).*

Denote $\lambda_\star = x_\star$ *and define* $W_\star := \sigma_\star G + Sf(S)$. *Consider the loss function* $\mathcal{L}_\star : \mathbb{R} \to \mathbb{R}$

$$\mathcal{L}_\star(x) := -\mathcal{M}_{\frac{\eta(\lambda_\star\delta-1)}{\delta(\eta-\mathcal{I}(W_\star))}Q+\frac{\lambda_\star\delta-1}{\delta(\eta-\mathcal{I}(W_\star))}\log\left(p_{W_\star}\right)}(x;1)\,, \tag{2.16}$$

where $\eta := 1-\mathcal{I}(W_\star)\cdot(\sigma_\star^2-\sigma_\star^2\lambda_\star\delta-\lambda_\star\delta)-\lambda_\star\delta$ *and* $Q(w) := w^2/2$. *Then for* $\mathcal{L}_\star$ *and* $\lambda_\star$, *the equations* (2.11) *satisfy* $(\alpha,\mu,\tau) = (\sigma_\star, 1, 1)$.

Lemma 2.3.1 suggests that if $\mathcal{L}_\star$ satisfies the assumptions of Theorem 2.3.1, then $\sigma_{\mathcal{L}_\star,\lambda_\star} = \sigma_\star$. In Figures 2.1 and 2.2 and for the special cases of Signed and Logistic models, we verify numerically that performance of candidates $\mathcal{L}_\star$ and $\lambda_\star$ reaches the optimal errors . This suggests that for these models, Lemma 2.3.1 yields the optimal choices for $\mathcal{L}$ and λ. See also Figure 2.3(Right) for an illustration of $\mathcal{L}_\star$.

2.3.3 The Sub-optimality Gap of RLS in Binary Models

We use the optimality results of the previous section to precisely quantify the sub-optimality gap of RLS. First, the following lemma characterizes the performance of RLS.

Lemma 2.3.2 (Asymptotic error of RLS)**.** *Let Assumptions 2.1.1, 2.1.2 and 2.3.1 hold. Recall that* $\nu_f = \mathbb{E}[Sf(S)] \neq 0$. *Fix any* $\delta > 0$ *and consider solving* (2.9) *with the square-loss* $\mathcal{L}(t) = (t-1)^2$ *and* $\lambda \geq 0$. *Then, the system of equations in* (2.11) *has a unique solution* $(\alpha_{\ell_2,\lambda}, \mu_{\ell_2,\lambda}, \tau_{\ell_2,\lambda})$ *and*

$$\sigma^2_{\ell_2,\lambda} = \frac{\alpha^2_{\ell_2,\lambda}}{\mu^2_{\ell_2,\lambda}} = \frac{1}{2\delta\nu_f^2}\Big(1-\delta\nu_f^2+\frac{2+2\delta+\lambda\delta+\delta\nu_f^2\left((2+\lambda)\delta-6\right)}{\sqrt{4+4\delta(\lambda-2)+\delta^2(\lambda+2)^2}}\Big). \tag{2.17}$$

Moreover, it holds that $\sigma^2_{\ell_2,\lambda} \geq \sigma^2_{\ell_2,\lambda_{\rm opt}} := H_\delta((1-\nu_f^2)^{-1})$ *with equality attained for the optimal tuning* $\lambda_{\rm opt} = 2(1-\nu_f^2)/(\delta\,\nu_f^2)$.

In resemblance to Lemma 2.2.2 in which RLS performance for linear measurements only depends on the second moment $\mathbb{E}[Z^2]$ of the additive noise distribution, Lemma 2.3.2 reveals that the corresponding key parameter for binary models is $1-\nu_f^2$. Interestingly, the expression for $\sigma^2_{\ell_2,\lambda_{\text{opt}}}$ conveniently matches with the simple bound on $\sigma_\star^2$ in Corollary 2.3.1. Specifically, it holds for any $\delta > 0$ that

$$1 \;\geq\; \frac{\sigma_\star^2}{\sigma^2_{\ell_2,\lambda_{\text{opt}}}} \geq \Omega_\delta := \frac{H_\delta\left(\mathcal{I}(S\,f(S))\right)}{H_\delta\left((1-\nu_f^2)^{-1}\right)}. \tag{2.18}$$

We note that $H_\delta(\cdot)$ is strictly-decreasing in its domain for a fixed $\delta > 0$. Furthermore, the Cramer-Rao bound (see Prop. 2.5.3 (d)) requires that $\mathcal{I}(Sf(S)) \geq (\text{Var}[Sf(S)])^{-1} = (1-\nu_f^2)^{-1}$. Combining these, confirms that $\Omega_\delta \leq 1$. Furthermore $\Omega_\delta = 1$ and thus $\sigma_\star^2 = \sigma^2_{\ell_2,\lambda_{\text{opt}}}$ iff the random variable $Sf(S)$ is Gaussian. However, for any binary link function satisfying Assumption 2.3.1, $Sf(S)$ does not take a Gaussian distribution (see Section 2.5.4.1), thus (2.18) suggests that square-loss cannot be optimal. Nevertheless, one can use (2.18) to argue that square-loss is (perhaps surprisingly) approximately optimal for certain popular models.

For instance consider logistic link function $\widetilde{f}_r$ defined as $\mathbb{P}(\widetilde{f}_r(x) = 1) = (1+\exp(-rx))^{-1}$, where $r := \|\mathbf{x}_0\|_2$. Using (2.18) and maximizing the sub-optimality gap $1/\Omega_\delta$ over $\delta > 0$, we find that if $f = \widetilde{f}_{r=1}$ then for all $\delta > 0$ it holds that

$$\sigma^2_{\ell_2,\lambda_{\text{opt}}} \leq 1.003\;\sigma_\star^2.$$

Thus, for a logistic link function and $\|\mathbf{x}_0\|_2 = 1$ *optimally-tuned RLS is approximately optimal!* This is in agreement with the key message of Corollary 2.3.1 on the critical role played by $Sf(S)$, since for the logistic model and small values of r, its density is "close" to a Gaussian. However, as signal strength increases and $\widetilde{f}_r$ converges to the sign

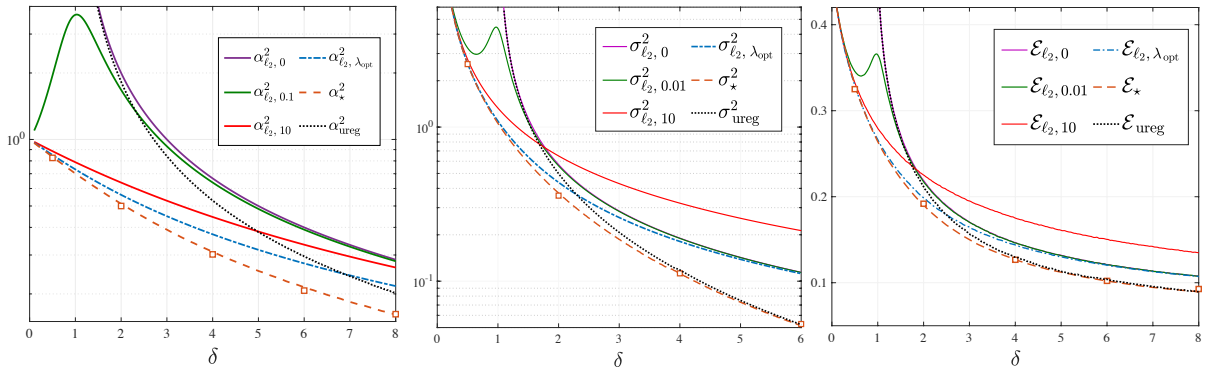


Figure 2.1: The lower bounds on error derived in this paper, compared to RLS for the linear model with $Z \sim \texttt{Laplace}(0,1)$ (Left), and for the binary Signed model (Middle) and binary Logistic model with $\|\mathbf{x}_0\| = 10$ (Right). The red squares denote the performance of the optimally tuned RERM as derived in Lemmas 2.2.1 and 2.3.1. See Section 5.3 for details and additional numerical results.

function ($\widetilde{f}_r(\cdot) \to \text{sign}(\cdot)$), there appears to be room for improvement between RLS and what Theorem 2.3.2 suggests to be possible. This can be precisely quantified using (2.18). For example, for $r = 10$ it can be shown that $\sigma^2_{\ell_2,\lambda_{\rm opt}} \leq 2.442\ \sigma^2_\star,\ \forall \delta > 0$. Lemma 2.3.1 provides the recipe to bridge the gap in this case. Indeed, Figures 2.1 and 2.2 show that the optimal loss function $\mathcal{L}$ predicted by the lemma outperforms RLS for all values δ and its performance matches the best possible one specified by Theorem 2.3.2.

2.4 Conclusion and Future work

This paper derives fundamental lower bounds on the statistical accuracy of ridge-regularized ERM (RERM) for linear and binary models in high-dimensions. It then derives simple closed-form approximations that allow precisely quantifying the sub-optimality gap of RLS. In Section 2.5.6, these bounds are further used to study the benefits of regularization by comparing (RERM) to un-regularized ERM.

Among several interesting directions of future work, we highlight the following. First, our lower bounds make it possible to compare RERM to the optimal Bayes risk [63, 64].

Second, it is interesting to extend the analysis to GLMs for arbitrary link functions beyond linear and binary studied here. A third exciting direction is investigating the fundamental limits of RERM in the presence of correlated (Gaussian) features.

2.5 Proofs and additional results

2.5.1 Useful facts

2.5.1.1 On Moreau Envelopes

In Proposition 2.5.1, some of the differential properties of Moreau-envelope functions, used throughout the paper are summarized (cf. [2]):

Proposition 2.5.1 (Properties of Moreau-envelopes)**.** *Let $\mathcal{L}$ be a lower semi-continuous and proper function. Then*

(a) The value $\mathcal{M}_{\mathcal{L}}(x;\tau)$ is finite and depends continuously on (x,τ), with $\mathcal{M}_{\mathcal{L}}(x;\tau) \to \mathcal{L}(x)$ as $\tau \to 0_+$ and $\mathcal{M}_{\mathcal{L}}(x;\tau) \to \min_{t\in\mathbb{R}} \mathcal{L}(t)$ as $\tau \to +\infty$, for all $x \in \mathbb{R}$.

(b) The first order derivatives of the Moreau-envelope of a function $\mathcal{L}$ are derived as follows:

$$\mathcal{M}'_{\mathcal{L},1}(x;\tau) := \frac{\partial \mathcal{M}_{\mathcal{L}}(x;\tau)}{\partial x} = \frac{1}{\tau}(x - \mathrm{prox}_{\mathcal{L}}(x;\tau)), \tag{2.19}$$

$$\mathcal{M}'_{\mathcal{L},2}(x;\tau) := \frac{\partial \mathcal{M}_{\mathcal{L}}(x;\tau)}{\partial \tau} = -\frac{1}{2\tau^2}(x - \mathrm{prox}_{\mathcal{L}}(x;\tau))^2. \tag{2.20}$$

Also if $\mathcal{L}$ is differentiable then

$$\mathcal{M}'_{\mathcal{L},1}(x;\tau) = \mathcal{L}'(\mathrm{prox}_{\mathcal{L}}(x;\tau)), \tag{2.21}$$

$$\mathcal{M}'_{\mathcal{L},2}(x;\tau) = -\frac{1}{2}(\mathcal{L}'(\mathrm{prox}_{\mathcal{L}}(x;\tau))^2. \tag{2.22}$$

(c) Additionally, based on the relations above, if $\mathcal{L}$ is twice differentiable then the following

is derived for its second order derivatives :

$$\mathcal{M}''_{\mathcal{L},1}(x;\tau) = \frac{\mathcal{L}''(\text{prox}_{\mathcal{L}}(x;\tau))}{1+\tau\mathcal{L}''(\text{prox}_{\mathcal{L}}(x;\tau))}, \tag{2.23}$$

$$\mathcal{M}''_{\mathcal{L},2}(x;\tau) = \frac{\left(\mathcal{L}'(\text{prox}_{\mathcal{L}}(x;\tau))\right)^2 \mathcal{L}''(\text{prox}_{\mathcal{L}}(x;\tau))}{1+\tau\,\mathcal{L}''(\text{prox}_{\mathcal{L}}(x;\tau))}. \tag{2.24}$$

The following proposition gives the recipe for inverting Moreau-envelpe of a convex function:

Proposition 2.5.2 (Inverse of the Moreau envelope)**.** *[59, Result. 23] For $\tau > 0$ and f a convex, lower semi-continuous function such that $g(\cdot) = \mathcal{M}_f(\cdot;\tau)$, the Moreau envelope can be inverted so that $f(\cdot) = -\mathcal{M}_{-g}(\cdot;\tau)$.*

Lemma 2.5.1 (e.g., [53], Lemma A.1.)**.** *The function $H : \mathbb{R}^3 \to \mathbb{R}$ defined as follows*

$$H(x,p,\tau) = \frac{1}{2\tau}(x-p)^2, \tag{2.25}$$

is jointly convex in its arguments.

2.5.1.2 On Fisher Information

In Proposition 2.5.3 we collect some useful properties of the Fisher Information for location. For the proofs and more details, we refer the interested reader to [65].

Proposition 2.5.3 (Properties of Fisher Information, [65])**.** *Let X be a zero-men random variable with probability density p_X satisfying the following conditions: (i) $p_X(x) > 0, -\infty < x < \infty$; (ii) $p'_X(x)$ exists; and (iii) The following integral exists:*

$$\mathcal{I}(X) = \int_{-\infty}^{\infty} \frac{(p'_X(x))^2}{p_X(x)}\,\mathrm{d}x.$$

The Fisher information for location $\mathcal{I}(X)$ defined above satisfies the following properties.

(a) $\mathcal{I}(X) := \mathbb{E}\left[(\xi_X(X))^2\right] = \mathbb{E}\Big[\left(\frac{p'_X(X)}{p_X(X)}\right)^2\Big]$.

(b) For any $c \in \mathbb{R}$, $\mathcal{I}(X+c) = \mathcal{I}(X)$.

(c) For any $c \in \mathbb{R}$, $\mathcal{I}(c\,X) = \mathcal{I}(X)/c^2$.

(d) (Cramer-Rao bound) $\mathcal{I}(X) \geq \frac{1}{\mathbb{E}[X^2]}$, *with equality if and only if* X *is Gaussian.*

(e) For two independent random variables X_1, X_2 *satisfying the three conditions above and any* $\theta \in [0,1]$, *it holds that* $\mathcal{I}(X_1+X_2) \leq \theta^2\mathcal{I}(X_1) + (1-\theta)^2\mathcal{I}(X_2)$.

(f) (Stam's inequality) For two independent random variables X_1, X_2 *satisfying the three conditions above, it holds that*

$$\mathcal{I}(X_1+X_2) \leq \frac{\mathcal{I}(X_1)\cdot\mathcal{I}(X_2)}{\mathcal{I}(X_1)+\mathcal{I}(X_2)}. \tag{2.26}$$

Moreover equality holds if and only if X_1 *and* X_2 *are independent Gaussian random variables.*

Lemma 2.5.2. *Let* $G \sim \mathcal{N}(0,1)$ *and* Z *be a random variable satisfying the assumptions of Proposition 2.5.3. For any* $a \in \mathbb{R}$, *use the shorhand* $V_a := a\,G + Z$. *The following are true:*

(a) $\lim_{a\to 0} a^2\mathcal{I}(V_a) = 0$.

(b) $\lim_{a\to+\infty} a^2\mathcal{I}(V_a) = 1$.

Proof: To show part (a), we use Proposition 2.5.3(e) with $\theta = 0$ to derive that

$$\lim_{a\to 0} a^2\,\mathcal{I}(V_a) \leq \lim_{a\to 0} a^2\,\mathcal{I}(Z) = 0, \tag{2.27}$$

where the second step follows by the fact that $\mathcal{I}(Z)$ is finite for any Z satisfying the assumption of the lemma. In order to prove part (b), we apply Proposition 2.5.3(c) to deduce that :

$$\lim_{a\to+\infty} a^2\,\mathcal{I}(V_a) = \lim_{a\to+\infty} a^2\,\mathcal{I}(a\,G+Z) = \lim_{a\to+\infty} \mathcal{I}(G+\frac{1}{a}Z) = 1, \tag{2.28}$$

2.5.1.3 On Min-max Duality

Theorem 2.5.1 (Sion's min-max theorem [66])**.** *Let X be a compact convex subset of a linear topological space and Y a convex subset of a linear topological space. If f is a real-valued function on $X \times Y$ with $f(x,\cdot)$ upper semicontinuous and quasi-concave on $Y, \forall x \in X$, and $f(\cdot,y)$ lower semicontinuous and quasi-convex on $X, \forall y \in Y$ then,*

$$\min_{x\in X}\sup_{y\in Y} f(x,y) = \sup_{y\in Y}\min_{x\in X} f(x,y).$$

2.5.2 Asymptotics for Binary RERM: Proof of Theorem 2.3.1

In this section, we prove that under the assumptions of Theorem 2.3.1, the system of equations in (2.11) has a unique and bounded solution.

2.5.2.1 Asymptotic Error of RERM via an Auxiliary Min-Max Optimization

As mentioned in Section 2.3, the proof of Theorem 2.3.1 has essentially two parts. The first part of the proof uses the CGMT [**?**] and the machinery developed in [**?**, 46, 52, 67] to relate the properties of the RERM solution to an Auxiliary Optimization (AO). The detailed steps follow mutatis-mutandis analogous derivations in recent works [**?**, 46, 52, 67, 51] and are omitted here for brevity. Instead, we summarize the finding of this analysis in the following proposition.

Proposition 2.5.4. *Consider the optimization problem in (2.9). If the min-max optimization in (2.29) has a unique and bounded solution $(\alpha^\star > 0, \mu^\star, \upsilon^\star > 0, \gamma^\star > 0)$, then the values of $\alpha_{\mathcal{L},\lambda}$ and $\mu_{\mathcal{L},\lambda}$ corresponding to $\mathcal{L}$ and λ defined in (2.66)-(2.67) are derived by setting $\alpha_{\mathcal{L},\lambda} = \alpha_\star$ and $\mu_{\mathcal{L},\lambda} = \mu_\star$, where*

$$(\alpha^\star, \mu^\star, \upsilon^\star, \gamma^\star) = \arg \min_{\substack{(\alpha,\mu,\upsilon)\in \\ \mathbb{R}_{\geq 0}\times\mathbb{R}\times\mathbb{R}_{>0}}} \max_{\gamma\in\mathbb{R}_{>0}} \Bigg[\Theta(\alpha,\mu,\upsilon,\gamma) := \frac{\gamma\upsilon}{2} - \frac{\alpha\gamma}{\sqrt{\delta}} + \frac{\lambda\mu^2}{2} + \frac{\lambda\alpha^2}{2} + \mathbb{E}\left[\mathcal{M}_{\mathcal{L}}\left(\alpha G + \mu S f(S); \frac{\upsilon}{\gamma}\right)\right]\Bigg], \tag{2.29}$$

and $G, S \overset{iid}{\sim} \mathcal{N}(0,1)$.

The system of equations in (2.11) is derived by the first-order optimality conditions of the function Θ based on its arguments $(\alpha, \mu, \upsilon, \gamma)$, i.e., by imposing $\nabla\Theta = \mathbf{0}$. In fact, similar to [53], it only takes a few algebraic steps to simplify the four equations in $\nabla\Theta = \mathbf{0}$ to the three equations in (2.11).

For the rest of this section, we focus on the second part of the proof of Theorem 2.3.1 regarding existence/uniqueness of solutions to (2.11), which has not been previously studied in our setting.

2.5.2.2 Properties of Θ : Strict Convexity-Strict Concavity and Boundedness of Saddle Points

We will show in Lemma 2.5.4 that for proving uniqueness and boundedness of the solutions to (2.11), it suffices to prove uniqueness and boundedness of the saddle point $(\alpha^\star, \mu^\star, \upsilon^\star, \gamma^\star)$ of Θ. In fact, a sufficient condition for uniqueness of solutions in (2.29) is that Θ is (jointly) strictly convex in (α, μ, υ) and strictly-concave in γ (e.g., see [53, Lemma B.2.]). Lemma 2.5.3, which is key to the proof of Theorem 2.3.1, derives sufficient conditions on $\mathcal{L}$ guaranteeing strict convexity-strict concavity of Θ as well as conditions

on $\mathcal{L}$ ensuring boundedness of $(\alpha^\star, \mu^\star, v^\star, \gamma^\star)$.

Lemma 2.5.3 (Properties of Θ). *Let $\mathcal{L}(\cdot)$ be a lower semi-continuous (lsc), proper and convex function and $\lambda > 0$. Then the following statements hold for the function $\Theta : \mathbb{R}_{\geq 0} \times \mathbb{R} \times \mathbb{R}_{>0} \times \mathbb{R}_{>0} \to \mathbb{R}$ in (2.29),*

(a) If $\mathcal{L}$ is bounded from below, then for all solutions $(\alpha^\star, \mu^\star, v^\star, \gamma^\star)$ there exists a constant $C > 0$ such that $\alpha^\star \in [0, C], \mu^\star \in [-C, C]$ and $v^\star \in [0, C]$.

(b) If $\mathcal{L}$ is bounded from below and $\mathbb{E}[\mathcal{L}(G)] < \infty$ for $G \sim \mathcal{N}(0, 1)$, then there exists a constant $C > 0$ such that $\gamma^\star \in [0, C]$.

(c) In addition to the assumptions of parts (a) and (b) assume that $\mathcal{L}'(0) \neq 0$, then $\gamma^\star > 0, \alpha^\star > 0$ and $v^\star > 0$.

(d) If $\mathcal{L}$ is twice differentiable and non-linear, then Θ is jointly strictly-convex in (α, μ, v).

(e) If $\mathcal{L}$ satisfies the assumptions of part (c) then Θ is strictly-concave in γ.

2.5.2.3 Proof of Lemma 2.5.3

Statement (a). Let $\widetilde{\Theta}(\alpha, \mu, v) := \sup_{\gamma \in \mathbb{R}_{>0}} \Theta(\alpha, \mu, v, \gamma)$. For all feasible (α, μ, v) it holds

$$\begin{aligned} \widetilde{\Theta}(\alpha, \mu, v) &\geq \Theta(\alpha, \mu, v, 1) \\ &= \frac{v}{2} - \frac{\alpha}{\sqrt{\delta}} + \frac{\lambda(\alpha^2 + \mu^2)}{2} + \mathbb{E}\Big[\mathcal{M}_{\mathcal{L}}(\alpha G + \mu S f(S); v)\Big]. \end{aligned} \tag{2.30}$$

Recall that $\mathcal{L}$ is bounded from below, i.e., for all $\mathcal{L}(x) \geq B, \forall x \in \mathbb{R}$ for some real B. By definition of Moreau-envelope function the same bound holds for $\mathcal{M}_{\mathcal{L}}$, i.e. for all $x \in \mathbb{R}$

and $y \in \mathbb{R}_{>0}$, we have that $\mathcal{M}_{\mathcal{L}}(x;y) \geq B$. Using this, we proceed from (2.30) to derive that:

$$\widetilde{\Theta}(\alpha,\mu,\upsilon) \;\geq\; B + \frac{\upsilon}{2} - \frac{\alpha}{\sqrt{\delta}} + \frac{\lambda(\alpha^2+\mu^2)}{2}. \tag{2.31}$$

Based on (2.31) that holds for all feasible (α,μ,υ) and using the fact that $\lambda > 0$ it can be readily shown that

$$\lim_{\alpha\to+\infty} \min_{(\mu,\upsilon)\in\mathbb{R}\times\mathbb{R}_{>0}} \widetilde{\Theta}(\alpha,\mu,\upsilon) = +\infty, \qquad \lim_{\upsilon\to+\infty} \min_{(\alpha,\mu)\in\mathbb{R}_{\geq 0}\times\mathbb{R}} \widetilde{\Theta}(\alpha,\mu,\upsilon) = +\infty,$$
$$\lim_{\mu\to\pm\infty} \min_{(\alpha,\upsilon)\in\mathbb{R}_{\geq 0}\times\mathbb{R}_{>0}} \widetilde{\Theta}(\alpha,\mu,\upsilon) = +\infty.$$

Thus, the function $\widetilde{\Theta}(\alpha,\mu,\upsilon)$ is level-bounded in $\mathbb{R}_{\geq 0}\times\mathbb{R}\times\mathbb{R}_{>0}$. This implies the boundedness of solutions $(\alpha^\star,\mu^\star,\upsilon^\star)$ to (2.29) [2, Thm. 1.9], as desired.

Statement (b). Under the assumptions of the lemma, we know from part (a) that the set of solutions to $(\alpha^\star,\mu^\star,\upsilon^\star)$ in (2.29) is bounded. Thus we can apply the Min-Max Theorem 2.5.1 and flip the order of minimum and maximum to write:

$$\min_{\substack{(\alpha,\mu,\upsilon)\\ \in[0,C]\times[-C,C]\times(0,C]}} \max_{\gamma\in\mathbb{R}_{\geq 0}} \Theta(\alpha,\mu,\upsilon,\gamma) = \max_{\gamma\in\mathbb{R}_{\geq 0}} \Big[\widehat{\Theta}(\gamma) := \min_{\substack{(\alpha,\mu,\upsilon)\\ \in[0,C]\times[-C,C]\times(0,C]}} \Theta(\alpha,\mu,\upsilon,\gamma)\Big]. \tag{2.32}$$

Without loss of generality, we assume C large enough such that $C > \max\{1, 1/\sqrt{\delta}\}$. Then, by choosing $\alpha = 1, \mu = 0$ and $\upsilon = 1/\sqrt{\delta}$, we find that for all $\gamma > 0$:

$$\widehat{\Theta}(\gamma) \;\leq\; \Theta\left(1,0,1/\sqrt{\delta},\gamma\right) = -\frac{\gamma}{2\sqrt{\delta}} + \frac{\lambda}{2} + \mathbb{E}\left[\mathcal{M}_{\mathcal{L}}\left(G;\frac{1}{\gamma\sqrt{\delta}}\right)\right]. \tag{2.33}$$

Note that for any $y\in\mathbb{R}$: $\mathcal{M}_{\mathcal{L}}\left(y;\frac{1}{\gamma\sqrt{\delta}}\right) = \min_{x\in\mathbb{R}} \frac{\gamma\sqrt{\delta}}{2}(x-y)^2 + \mathcal{L}(x) \leq \mathcal{L}(y)$. Thus we

derive from (2.33):

$$\widehat{\Theta}(\gamma) \leq -\frac{\gamma}{2\sqrt{\delta}} + \frac{\lambda}{2} + \mathbb{E}\left[\mathcal{L}(G)\right]. \tag{2.34}$$

But $\mathbb{E}\left[\mathcal{L}(G)\right]$ is assumed to be bounded, thus it can be concluded from (2.34) that the function $\widehat{\Theta}(\gamma)$ is level-bounded, i.e.,

$$\lim_{\gamma \to +\infty} \widehat{\Theta}(\gamma) = -\infty. \tag{2.35}$$

This implies boundedness of the set of maximizers $\gamma^\star$, which completes the proof.

Statement (c). First, we show that $\gamma^\star > 0$. On the contrary, assume that $\gamma^\star = 0$. Then based on (2.29) and Proposition 2.5.1(a),

$$(\alpha^\star, \mu^\star, \upsilon^\star) = \arg \min_{\substack{(\alpha,\mu,\upsilon) \\ \in [0,C]\times[-C,C]\times(0,C]}} \left[\frac{\lambda\alpha^2}{2} + \frac{\lambda\mu^2}{2} + \min_{t\in\mathbb{R}} \mathcal{L}(t)\right],$$

implying that $\alpha^\star = \mu^\star = 0$ and $\Theta(\alpha^\star, \mu^\star, \upsilon^\star, \gamma^\star) = \min_{t\in\mathbb{R}} \mathcal{L}(t)$. On the other hand, in this case we find that for any $\widetilde{\gamma} \in (0, C]$,

$$\Theta(\alpha^\star, \mu^\star, \upsilon^\star, \widetilde{\gamma}) = \widetilde{\gamma}\upsilon^\star + \mathcal{M}_{\mathcal{L}}\left(0; \frac{\upsilon^\star}{\widetilde{\gamma}}\right) > \min_{t\in\mathbb{R}} \mathcal{L}(t).$$

To deduce the inequality, we used the fact that $\mathcal{M}_{\mathcal{L}}(0;\tau) = \min_{t\in\mathbb{R}} t^2/(2\tau) + \mathcal{L}(t) > \min_{t\in\mathbb{R}} \mathcal{L}(t)$ for all $\tau \geq 0$, provided that $\mathcal{L}(t)$ does not attain its minimum at $t = 0$. Thus, since by assumption $\mathcal{L}'(0) \neq 0$, we deduce that $\Theta(\alpha^\star, \mu^\star, \upsilon^\star, \widetilde{\gamma}) > \Theta(\alpha^\star, \mu^\star, \upsilon^\star, \gamma^\star)$, which is in contradiction to the optimality of $\gamma^\star$. This shows that $\gamma^\star > 0$ for any loss function satisfying the assumptions of the lemma. Next, we prove that $\alpha^\star > 0$. if $\alpha^\star = 0$, then

based on the optimality of $\alpha^\star$ it holds that

$$\frac{\partial \Theta}{\partial \alpha}\Big|_{(\alpha^\star,\mu^\star,\upsilon^\star,\gamma^\star)} \geq 0,$$

thus based on (2.29),

$$\mathbb{E}\left[G \cdot \mathcal{M}'_{\ell,1}\left(\mu^\star Sf(S); \frac{\upsilon^\star}{\gamma^\star}\right)\right] - \frac{\gamma^\star}{\sqrt{\delta}} \geq 0. \tag{2.36}$$

Since by assumption G and $S\,f(S)$ are independent and $\mathbb{E}[G] = 0$, we deduce from (2.36) that $\gamma^\star = 0$, which is in contradiction to the previously proved fact that $\gamma^\star > 0$. This shows that $\alpha^\star > 0$, as desired. Finally, we note that if $\upsilon^\star = 0$, then based on (2.29) and in light of Proposition 2.5.1(a), we find that,

$$(\alpha^\star, \mu^\star, \gamma^\star) = \arg \min_{\substack{(\alpha,\mu) \\ \in [0,C]\times[-C,C]}} \max_{\gamma \in (0,C]} \left[-\frac{\alpha\gamma}{\sqrt{\delta}} + \frac{\lambda\alpha^2}{2} + \frac{\lambda\mu^2}{2} + \mathbb{E}\Big[\mathcal{L}\left(\alpha G + \mu Sf(S)\right)\Big]\right],$$

which based on the decreasing nature of RHS in terms of γ, implies that either $\gamma^\star = 0$ or $\alpha^\star = 0$. However, we proved that both $\gamma^\star$ and $\alpha^\star$ are positive. This proves the desired result $\upsilon^\star \neq 0$ and completes the proof of this part.

Statement (d). Let $\mathbf{w}_1 := (\alpha_1, \mu_1, \tau_1)$ and $\mathbf{w}_2 := (\alpha_2, \mu_2, \tau_2)$ be two distinct points in the space $\mathbb{R}_{\geq 0} \times \mathbb{R} \times \mathbb{R}_{>0}$. We consider two cases :

Case I : $(\alpha_1, \mu_1) = (\alpha_2, \mu_2)$

In this case, it suffices to show that for fixed $\alpha > 0$ and μ and under the assumptions of the lemma, the function $\mathbb{E}\left[\mathcal{M}_{\mathcal{L}}\left(\alpha G + \mu Sf(S); \tau\right)\right]$ is strictly-convex in τ. Denote by $p(\alpha, \mu, \tau) := \text{prox}_{\mathcal{L}}\left(\alpha G + \mu Sf(S); \tau\right)$. First, we derive second derivate of the Moreau-envelope function with respect to τ by applying (2.24), and further use convexity of $\mathcal{L}$ to

derive that :

$$\frac{\partial^2}{\partial \tau^2}\mathbb{E}\Big[\mathcal{M}_{\mathcal{L}}\left(\alpha G+\mu Sf(S);\tau\right)\Big] = \mathbb{E}\left[\frac{\Big(\mathcal{L}'\left(p\left(\alpha,\mu,\tau\right)\right)\Big)^2\mathcal{L}''\left(p\left(\alpha,\mu,\tau\right)\right)}{1+\tau\,\mathcal{L}''\left(p\left(\alpha,\mu,\tau\right)\right)}\right]\geq 0. \tag{2.37}$$

Next we show that the inequality above is strict if $\mathcal{L}(\cdot)$ is a non-linear function. First we note that combining (2.19) and (2.21) yields that for all $x\in\mathbb{R}$:

$$\mathcal{L}'(\mathrm{prox}_{\mathcal{L}}\left(x;\tau\right)) = \frac{1}{\tau}(x-\mathrm{prox}_{\mathcal{L}}\left(x;\tau\right)),$$
$$\mathcal{L}''(\mathrm{prox}_{\mathcal{L}}\left(x;\tau\right)) = \frac{1-\mathrm{prox}'_{\mathcal{L},1}\left(x;\tau\right)}{\tau\cdot\mathrm{prox}'_{\mathcal{L},1}\left(x;\tau\right)}.$$

Using these relations and denoting by $p'(\alpha,\mu,\tau) := \mathrm{prox}'_{\mathcal{L},1}\left(\alpha G+\mu Sf(S);\tau\right)$, we can rewrite (2.37) as following :

$$\frac{\partial^2}{\partial \tau^2}\mathbb{E}\Big[\mathcal{M}_{\mathcal{L}}\left(\alpha G+\mu Sf(S);\tau\right)\Big] = \frac{1}{\tau^3}\mathbb{E}\left[\frac{\Big(\alpha G+\mu Sf(S)-p(\alpha,\mu,\tau)\Big)^2\Big(1-p'(\alpha,\mu,\tau)\Big)}{p'(\alpha,\mu,\tau)\Big(1+\tau\,\mathcal{L}''(p(\alpha,\mu,\tau))\Big)}\right]. \tag{2.38}$$

It is straightforward to see that if $\alpha>0$, then $\alpha G+\mu Sf(S)$ has positive density in the real line. Thus from (2.38) we find that :

$$\frac{\partial^2}{\partial \tau^2}\mathbb{E}\Big[\mathcal{M}_{\mathcal{L}}\left(\alpha G+\mu Sf(S);\tau\right)\Big]=0 \iff \exists c\in\mathbb{R}\ \text{ s.t. } \forall x\in\mathbb{R}: \quad \mathrm{prox}_{\mathcal{L}}\left(x;\tau\right)=x+c. \tag{2.39}$$

Recalling (2.19), we see that the condition in (2.39) is satisfied if and only if :

$$\exists c_1, c_2 \in \mathbb{R} : \text{s.t. } \forall x \in \mathbb{R} : \mathcal{M}_{\mathcal{L}}(x;\tau) = c_1 x + c_2. \tag{2.40}$$

Using inverse properties of Moreau-envelope in Proposition 2.5.2, we derive that the loss function $\mathcal{L}(\cdot)$ satisfying (2.40) takes the following shape,

$$\forall x \in \mathbb{R} : \quad \mathcal{L}(x) = -\mathcal{M}_{-c_1 I - c_2}(x;\tau) = c_1 x + \frac{\tau c_1^2}{2} + c_2.$$

where $I(\cdot)$ is the identity function i.e. $I(t) = t$, $\forall t \in \mathbb{R}$. Therefore if $\mathcal{L}$ is non-linear function as required by the assumption of the lemma, $\mathbb{E}\left[\mathcal{M}_{\mathcal{L}}(\alpha G + \mu S f(S);\tau)\right]$ has a positive second derivative with respect to τ and consequently Θ is strictly-convex in υ.

Case II : $(\alpha_1, \mu_1) \neq (\alpha_2, \mu_2)$

In this case we use definition of strict-convexity to prove the claim. First, for compactness we define :

$$p_i := \text{prox}_{\mathcal{L}}(\alpha_i G + \mu_i S f(S);\tau_i) = \arg\min_w \frac{1}{2\tau_i}(\alpha_i G + \mu_i S f(S) - w)^2 + \mathcal{L}(w),$$
$$\Omega(\mathbf{w}_i) = \Omega(\alpha_i, \mu_i, \tau_i) := \frac{\lambda \mu_i^2}{2} + \frac{\lambda \alpha_i^2}{2} + \mathbb{E}\Big[\mathcal{M}_{\mathcal{L}}(\alpha_i G + \mu_i S f(S);\tau_i)\Big]$$

for $i = 1, 2$. Based on the way we defined the functions Θ and Ω, one can see that in order to show strict-convexity of Θ in (α, μ, υ) it suffices to prove strict-convexity of Ω in (α, μ, τ). Let $\theta \in (0,1)$, and denote $\tau_\theta := \theta\tau_1 + \overline{\theta}\tau_2, \alpha_\theta := \theta\alpha_1 + \overline{\theta}\alpha_2$ and $\mu_\theta := \theta\mu_1 + \overline{\theta}\mu_2$.

With this notation,

$$\Omega(\theta \mathbf{w}_1 + \overline{\theta}\mathbf{w}_2) \leq \tag{2.41}$$
$$\frac{\lambda\mu_\theta^2}{2} + \frac{\lambda\alpha_\theta^2}{2} + \mathbb{E}\left[\frac{1}{2\tau_\theta}\Big(\alpha_\theta G + \mu_\theta S f(S) - (\theta p_1 + \overline{\theta} p_2)\Big)^2 + \mathcal{L}\Big(\theta p_1 + \overline{\theta} p_2\Big)\right]$$
$$= \frac{\lambda\mu_\theta^2}{2} + \frac{\lambda\alpha_\theta^2}{2} + \mathbb{E}\Big[H\Big(\alpha_\theta G + \mu_\theta S f(S), \theta p_1 + \overline{\theta} p_2, \tau_\theta\Big) + \mathcal{L}\Big(\theta p_1 + \overline{\theta} p_2\Big)\Big]$$
$$\leq \frac{\lambda\mu_\theta^2}{2} + \frac{\lambda\alpha_\theta^2}{2} +$$
$$\mathbb{E}\left[\theta H\Big(\alpha_1 G + \mu_1 S f(S), p_1, \tau_1\Big) + \overline{\theta} H\Big(\alpha_2 G + \mu_2 S f(S), p_2, \tau_2\Big) + \mathcal{L}\Big(\theta p_1 + \overline{\theta} p_2\Big)\right]. \tag{2.42}$$

The first inequality above follows by the definition of the Moreau envelope. The equality in the second line uses the definition of the function $H : \mathbb{R}^3 \to \mathbb{R}$ in (2.25). Finally, the last inequality follows from convexity of H as proved in Lemma 2.5.1.

Continuing from (2.42), we use convexity of $\mathcal{L}$ to find that

$$\Omega(\theta \mathbf{w}_1 + \overline{\theta}\mathbf{w}_2) \leq \frac{\lambda\mu_\theta^2}{2} + \frac{\lambda\alpha_\theta^2}{2} +$$
$$\mathbb{E}\Big[\theta H(\alpha_1 G + \mu_1 S f(S), p_1, \tau_1) + \overline{\theta} H(\alpha_2 G + \mu_2 S f(S), p_2, \tau_2) + \theta\, \mathcal{L}(p_1) + \overline{\theta}\mathcal{L}(p_2)\Big] \tag{2.43}$$

Additionally since $\lambda > 0$ and $(\alpha_1, \mu_1) \neq (\alpha_2, \mu_2)$, we find that :

$$\frac{\lambda\mu_\theta^2}{2} + \frac{\lambda\alpha_\theta^2}{2} < \frac{\lambda(\theta\mu_1^2 + \overline{\theta}\mu_2^2)}{2} + \frac{\lambda(\theta\alpha_1^2 + \overline{\theta}\alpha_2^2)}{2}.$$

Thus proceeding from (2.43) we conclude strict-convexity of the function Ω :

$$\Omega(\theta\mathbf{w}_1+\overline{\theta}\mathbf{w}_2) < \frac{\lambda(\theta\mu_1^2+\overline{\theta}\mu_2^2)}{2}+\frac{\lambda(\theta\alpha_1^2+\overline{\theta}\alpha_2^2)}{2}+$$
$$\mathbb{E}\Big[\,\theta H(\alpha_1 G+\mu_1 Sf(S),p_1,\tau_1)+\overline{\theta}H(\alpha_2 G+\mu_2 Sf(S),p_2,\tau_2)+\theta\,\mathcal{L}(p_1)+\overline{\theta}\mathcal{L}(p_2)\Big]$$
$$=\theta\Omega(\mathbf{w}_1)+\overline{\theta}\Omega(\mathbf{w}_2).$$

This completes the proof of part (d).

Statement (e). Based on the proof of part (c) and under the assumptions of the lemma we have $\alpha^\star \neq 0$. Thus we see that the random variable $\alpha G+\mu Sf(S)$ has a positive probability density everywhere in the desired domain of the optimization problem in (2.29). Next, we use the result in [53, Proposition A.6], which states that if the random variable X has a positive density everywhere and $\mathcal{L}$ is continuously differentiable with $\mathcal{L}'(0)\neq 0$ then

$$\mathbb{E}\Big[\mathcal{M}_{\mathcal{L}}\left(X;1/\gamma\right)\Big]$$

is strictly concave in γ. Based on this, Θ is strictly-concave in γ. This completes the proof of the lemma.

2.5.2.4 From (2.29) to (2.11)

The following lemma connects the min-max optimization (2.29) to the system of equations in (2.11)

Lemma 2.5.4 (Uniqueness of solutions to (2.11))**.** *Assume that the optimization problem in* (2.29) *yields a unique and bounded solution* $(\alpha>0,\mu,\upsilon>0,\gamma>0)$. *Then the equations* (2.11) *have a unique and bounded solution* $(\alpha>0,\mu,\tau>0)$ *where* $\tau=\upsilon/\gamma$.

Proof: By direct differentiation with respect to the variables $(\mu,\alpha,\upsilon,\gamma)$, the first

order optimality conditions of the min-max optimization in (2.29) are as follows:

$$\mathbb{E}\left[Sf(S)\mathcal{M}'_{\ell,1}\left(\alpha G+\mu Sf(S);\frac{\upsilon}{\gamma}\right)\right]=-\lambda\mu,\ \lambda\alpha+\mathbb{E}\left[G\mathcal{M}'_{\ell,1}\left(\alpha G+\mu Sf(S);\frac{\upsilon}{\gamma}\right)\right]=\frac{\gamma}{\sqrt{\delta}},$$
$$\frac{1}{\gamma}\mathbb{E}\left[\mathcal{M}'_{\ell,2}\left(\alpha G+\mu Sf(S);\frac{\upsilon}{\gamma}\right)\right]=-\frac{\gamma}{2},\ -\frac{\upsilon}{\gamma^2}\mathbb{E}\left[\mathcal{M}'_{\ell,2}\left(\alpha G+\mu Sf(S);\frac{\upsilon}{\gamma}\right)\right]+\frac{\upsilon}{2}=\frac{\alpha}{\sqrt{\delta}}. \tag{2.44}$$

Assumptions of the lemma imply that the saddle point of the optimization problem in (2.29) is unique and bounded, therefore (2.44) yields a unique bounded solution $(\alpha>0,\mu,\upsilon>0,\gamma>0)$. By denoting $\tau=\upsilon/\gamma$ and using the fact that $\mathcal{M}'_{\mathcal{L},2}(x;\tau)=-\frac{1}{2}(\mathcal{M}'_{\mathcal{L},1}(x;\tau))^2$ (as implied by (2.19)-(2.20)) we reach the Equations (2.11) i.e.,

$$\mathbb{E}\Big[S\,f(S)\cdot\mathcal{M}'_{\mathcal{L},1}\left(\alpha G+\mu Sf(S);\tau\right)\Big]=-\lambda\mu, \tag{2.45a}$$
$$\tau^2\,\delta\cdot\mathbb{E}\Big[\ \left(\mathcal{M}'_{\mathcal{L},1}\left(\alpha G+\mu Sf(S);\tau\right)\right)^2\ \Big]=\alpha^2, \tag{2.45b}$$
$$\tau\,\delta\cdot\mathbb{E}\Big[G\cdot\mathcal{M}'_{\mathcal{L},1}\left(\alpha G+\mu Sf(S);\tau\right)\Big]=\alpha(1-\lambda\tau\delta). \tag{2.45c}$$

The uniqueness of $(\alpha>0,\mu,\tau>0)$ as the solution to (2.45) follows from the uniqueness of the solution $(\alpha>0,\mu,\upsilon>0,\gamma>0)$ to (2.44). In particular if there are two distinct solutions (α_1,μ_1,τ_1) and (α_2,μ_2,τ_2) to the Equations (2.45), then we reach contradiction by noting that $(\alpha_1,\mu_1,\upsilon_1:=\alpha_1/\sqrt{\delta},\gamma_1:=\alpha_1/(\tau_1\sqrt{\delta}))$ and $(\alpha_2,\mu_2,\upsilon_2:=\alpha_2/\sqrt{\delta},\gamma_2:=\alpha_2/(\tau_2\sqrt{\delta}))$ are two distinct points satisfying the Equations (2.44). This completes the proof of the lemma.

2.5.2.5 Completing the proof of Theorem 2.3.1

We are now ready to complete the proof of Theorem 2.3.1. Based on Lemma 2.5.4, for the system of equations in (2.11) to have a unique and bounded solution, it suffices that $(\alpha^\star>0,\mu^\star,\upsilon^\star>0,\gamma^\star>0)$ as the solution of (2.29) is unique and bounded. Since

Θ is convex-concave and the optimality sets are bounded from Lemma 2.5.3(a)-(e), a saddle point of Θ exists [68, Cor. 37.3.2]. Additionally, based on the assumptions of the theorem and in view of Lemma 2.5.3(d),(e), Θ is jointly strictly-convex in (α, μ, v) and strictly-concave in γ which implies the uniqueness of $(\alpha^\star > 0, \mu^\star, v^\star > 0, \gamma^\star > 0)$ as a solution to (2.29). This completes the proof of the theorem.

As mentioned in the main body of the chapter, we conjecture that some of the technical conditions of Theorem 2.3.1, albeit mild in their current form, can be relaxed even further. Refining these conditions can be an interesting topic of future work, but is out of the scope of this chapter. We mention in passing that the conclusions of Theorem 2.3.1 also hold true if we replace the two-times differentiability condition by an assumption that the loss is one-time differentiable and strictly convex.

2.5.3 Fundamental Limits for Linear Models: Proofs for Section 2.2

2.5.3.1 Auxiliary Results

Lemma 2.5.5 (Boundedness of τ in (2.49))**.** *Let $\mathcal{L}(\cdot)$ be a non-linear, convex and twice differentiable function, $\lambda > 0$ and $\delta > 0$ and the pair (α, τ) be a solution to (2.3) where $\alpha > 0$. Then, $0 < \tau < \frac{1}{\lambda\delta}$.*

Proof: Using Stein's lemma (aka Gaussian integration by parts) we find that

$$\mathbb{E}\Big[G \cdot \mathcal{M}'_{\mathcal{L},1}\left(\alpha\, G + Z; \tau\right)\Big] = \alpha\, \mathbb{E}\Big[\mathcal{M}''_{\mathcal{L},1}\left(\alpha\, G + Z; \tau\right)\Big]$$

Therefore the equation in the LHS in (2.3) is equivalent to

$$\tau\delta\, \mathbb{E}\Big[\mathcal{M}''_{\mathcal{L},1}\left(\alpha\, G + Z; \tau\right)\Big] = 1 - \lambda\tau\delta. \tag{2.46}$$

Next we prove that under the assumptions of the lemma, $\mathbb{E}\Big[\mathcal{M}''_{\mathcal{L},1}\left(\alpha G+\mu S f(S);\tau\right)\Big]$ is positive. First using properties of Morea-envelopes in (2.23), we have

$$\mathbb{E}\Big[\mathcal{M}''_{\mathcal{L},1}\left(\alpha\, G+Z;\tau\right)\Big]=\mathbb{E}\left[\frac{\mathcal{L}''(\text{prox}_{\mathcal{L}}\left(\alpha\, G+Z;\tau\right))}{1+\tau\mathcal{L}''(\text{prox}_{\mathcal{L}}\left(\alpha G+Z;\tau\right))}\right]\geq 0. \tag{2.47}$$

In particular, we see that equality is achieved in (2.47) is achieved if and only if

$$\forall x\in\mathbb{R}:\quad \mathcal{M}''_{\mathcal{L},1}\left(x;\tau\right)=0.$$

Or equivalently,

$$\exists\, c_{_1},c_{_2}\in\mathbb{R}:\text{s.t. }\ \forall x\in\mathbb{R}:\mathcal{M}_{\mathcal{L}}\left(x;\tau\right)=c_{_1}x+c_{_2}. \tag{2.48}$$

Finally, using Proposition 2.5.2 to "invert" the Moreau envelope function, we find that the loss function $\mathcal{L}(\cdot)$ satisfying (2.48) is such that

$$\forall x\in\mathbb{R}:\quad \mathcal{L}(x)=-\mathcal{M}_{-c_{_1}I-c_{_2}}\left(x;\tau\right)=c_{_1}x+\frac{\tau c_{_1}^2}{2}+c_{_2},$$

where $I(\cdot)$ is the identity function i.e. $I(t)=t$, $\forall t\in\mathbb{R}$. But according to the assumptions of the lemma, $\mathcal{L}$ is a *non-linear* convex function. Thus, it must hold that $\mathbb{E}\left[\mathcal{M}''_{\mathcal{L},1}\left(\alpha\, G+Z;\tau\right)\right]>0$. Using this and the assumptions on λ and δ, the advertised claim follows directly from (2.46).

2.5.3.2 Proof of Theorem 2.2.1

Fix a convex loss function $\mathcal{L}$ and regularization parameter $\lambda \geq 0$. Let $(\alpha > 0, \tau > 0)$ be the unique solution to

$$\delta\tau^2 \cdot \mathbb{E}\Big[\Big(\mathcal{M}'_{\mathcal{L},1}\left(\alpha\, G + Z; \tau\right)\Big)^2\Big] = \alpha^2 - \lambda^2\delta^2\tau^2, \tag{2.49a}$$

$$\delta\tau \cdot \mathbb{E}\Big[G \cdot \mathcal{M}'_{\mathcal{L},1}\left(\alpha\, G + Z; \tau\right)\Big] = \alpha\,(1 - \lambda\delta\tau). \tag{2.49b}$$

For convenience, let us define the function $\Psi : \mathbb{R}_{\geq 0} \times [0, 1) \to \mathbb{R}$:

$$\Psi(a, x) := \frac{(a^2 - x^2\,\delta^2)\,\mathcal{I}(V_a)}{(1 - x\,\delta)^2}. \tag{2.50}$$

Then, $\alpha_\star > 0$ as in (2.6) is equivalently expressed as

$$\alpha_\star := \min_{0 \leq x < 1/\delta} \left\{ a \geq 0 :\ \Psi(a, x) = \frac{1}{\delta} \right\}. \tag{2.51}$$

Before everything, let us show that $\alpha_\star$ is well-defined, i.e., that the feasible set of the minimization in (2.51) is non-empty for all $\delta > 0$ and random variables Z satisfying Assumption 2.2.1. Specifically, we will show that there exists $a \geq 0$ such that $\Psi\left(a, \frac{a}{(1+a)\delta}\right) = 1/\delta$. It suffices to prove that the range of the function $\widetilde{\Psi}(a) := \Psi\left(a, \frac{a}{(1+a)\delta}\right)$ is $(0, \infty)$. Clearly, the function $\widetilde{\Psi}$ is continuous in $\mathbb{R}_{\geq 0}$. Moreover, it can be checked that $\widetilde{\Psi}(a) = (a^2 + 2a)\Psi_0(a)$ where $\Psi_0(a) := a^2\mathcal{I}(V_a)$. By Lemma 2.5.2, $\lim_{a\to 0} \Psi_0(a) = 0$ and $\lim_{a\to+\infty} \Psi_0(a) = 1$. Hence, we find that $\lim_{a\to 0} \widetilde{\Psi}(a) = 0$ and $\lim_{a\to+\infty} \widetilde{\Psi}(a) = +\infty$, as desired.

We are now ready to prove the main claim of the theorem, i.e.,

$$\alpha \geq \alpha_\star. \tag{2.52}$$

Denote by ϕ_α the density of the Gaussian random variable αG. We start with the following calculation:

$$\begin{aligned}\mathbb{E}\Big[G \cdot \mathcal{M}'_{\mathcal{L},1}\left(V_\alpha;\tau\right)\Big] &= -\alpha \iint \mathcal{M}'_{\mathcal{L},1}\left(u+z;\tau\right)\phi'_\alpha(u)p_Z(z)\mathrm{d}u\mathrm{d}z \\ &= -\alpha \iint \mathcal{M}'_{\mathcal{L},1}\left(v;\tau\right)\phi'_\alpha(u)p_Z(v-u)\mathrm{d}u\mathrm{d}v \\ &= -\alpha \int \mathcal{M}'_{\mathcal{L},1}\left(v;\tau\right)p'_V(v)\mathrm{d}v = -\alpha\,\mathbb{E}\Big[\mathcal{M}'_{\mathcal{L},1}\left(V_\alpha;\tau\right)\cdot\xi_{V_\alpha}(V_\alpha)\Big], \end{aligned} \tag{2.53}$$

where for a random variable V, we denote its score function with $\xi_V(v) := p'_V(v)/p_V(v)$ for $v \in \mathbb{R}$. Using (2.53) and $\alpha > 0$, (2.49b) can be equivalently written as following,

$$1 - \lambda\,\delta\,\tau = -\delta\tau \cdot \mathbb{E}\Big[\mathcal{M}'_{\mathcal{L},1}\left(V_\alpha;\tau\right)\cdot\xi_{V_\alpha}(V_\alpha)\Big]. \tag{2.54}$$

Next, by applying Cauchy-Shwarz inequality, recalling $\mathbb{E}[(\xi_{V_\alpha}(V_\alpha))^2] = \mathcal{I}\left(V_\alpha\right)$ and using (2.49a), we have that

$$\Big(\mathbb{E}\Big[\mathcal{M}'_{\mathcal{L},1}\left(V_\alpha;\tau\right)\cdot\xi_{V_\alpha}(V_\alpha)\Big]\Big)^2 \le \mathbb{E}\Big[\Big(\mathcal{M}'_{\mathcal{L},1}\left(V_\alpha;\tau\right)\Big)^2\Big]\cdot\mathcal{I}\left(V_\alpha\right) = \frac{(\alpha^2-\lambda^2\,\delta^2\,\tau^2)\,\mathcal{I}(V_\alpha)}{\delta\tau^2},$$

where we have also used the fact that $\tau > 0$. To continue, we use (2.54) to rewrite the LHS above and deduce that:

$$\left(\frac{1-\lambda\,\delta\,\tau}{\delta\tau}\right)^2 \le \frac{(\alpha^2-\lambda^2\,\delta^2\,\tau^2)\,\mathcal{I}(V_\alpha)}{\delta\tau^2}. \tag{2.55}$$

By simplifying the resulting expressions we have proved that (α,τ) satisfy the following inequality:

$$\frac{(\alpha^2 - \lambda^2\,\delta^2\,\tau^2)\,\mathcal{I}(V_\alpha)}{(1-\lambda\,\delta\,\tau)^2} \geq \frac{1}{\delta}. \tag{2.56}$$

In the remaining, we use (2.56) to prove (2.52). For the sake of contradiction to (2.52), assume that there exists a valid triplet (α, λ, τ) such that $\alpha < \alpha_\star$. Recall by inequality (2.56) that α satisfies:

$$\Psi\Big(\alpha\,,\,\lambda\,\tau\Big) \geq \frac{1}{\delta}. \tag{2.57}$$

We show first that (2.57) holds with strict inequality. To see this, suppose that $\Psi(\alpha, \lambda\,\tau) = 1/\delta$. From Lemma 2.5.5, it also holds that $\lambda\,\tau \in (0, 1/\delta)$. Hence, the pair $(\alpha, \lambda\,\tau)$ is a feasible point in the minimization in (2.51). Combining this with optimality of $\alpha_\star$ lead to the conclusion that $\alpha_\star \geq \alpha$, which contradicts our assumption $\alpha < \alpha_\star$. Therefore we consider only the case where (2.57) holds with strict inequality i.e., $\Psi(\alpha,\, \lambda\tau) > 1/\delta$.

To proceed, note that $\Psi(0, x) \leq 0$ for all $x \in [0, 1)$. Thus, by continuity of the function $a \mapsto \Psi(a, x)$ for fixed $x \in [0, 1/\delta)$:

$$\exists\,\widetilde{\alpha}:\ \text{s.t.}\ \ 0 \leq \widetilde{\alpha} < \alpha,\ \ \text{and}\ \ \Psi\left(\widetilde{\alpha}\,,\,\lambda\,\tau\right) = \frac{1}{\delta}. \tag{2.58}$$

By recalling our assumption that $\alpha < \alpha_\star$, we can deduce that (2.58) in fact holds for $\widetilde{\alpha} < \alpha_\star$. However, this is in contradiction with the optimality of $\alpha_\star$ defined in (2.51). This shows that for all achievable α it must hold that $\alpha \geq \alpha_\star$. This proves the claim in (2.52) and completes the proof of the theorem.

2.5.3.3 Proof of Lemma 2.2.1

To prove the claim of the lemma, it suffices to show that the proposed loss function and regularization parameter, satisfy the system of equations in (2.49) with $\alpha = \alpha_\star$. For this purpose we show that $(\mathcal{L}, \lambda, \alpha, \tau) = (\mathcal{L}_\star, \lambda_\star, \alpha_\star, 1)$ satisfy (2.49).

First, we recognize that for the candidate optimal loss function in Lemma 2.3.1 we have $\forall v \in \mathbb{R}$ that

$$\mathcal{M}'_{\mathcal{L}_\star,1}(v;1) = -\frac{\alpha_\star^2 - \lambda_\star^2\,\delta^2}{1-\lambda_\star\delta} \cdot \xi_{V_\star}(v). \tag{2.59}$$

Thus by replacing the proposed parameters in (2.49a) we have :

$$\delta\,\mathbb{E}\left[\left(\mathcal{M}'_{\mathcal{L}_\star,1}(V_\star;1)\right)^2\right] = \delta\left(\frac{\alpha_\star^2 - \lambda_\star^2\,\delta^2}{1-\lambda_\star\,\delta}\right)^2 \mathcal{I}(V_\star) = \alpha_\star^2 - \lambda_\star^2\delta^2,$$

where for the last line we used the definitions of $\alpha_\star$ and $\lambda_\star$ in the statement of the lemma. This proves the claim for (2.49a). To show that Equation (2.49b) is satisfied we use its equivalent expression in (2.54) and also replace (2.59) in (2.54). Specifically, this shows that

$$\begin{aligned}\delta\,\mathbb{E}\Big[G\cdot\mathcal{M}'_{\mathcal{L}_\star,1}(V_\star;1)\Big] &= -\delta\,\alpha_\star\,\mathbb{E}\Big[\mathcal{M}'_{\mathcal{L}_\star,1}(V_\star;1)\cdot\xi_{V_\star}(V_\star)\Big]\\ &= \frac{\delta\,\alpha_\star\,(\alpha_\star^2-\lambda_\star^2\,\delta^2)\cdot\mathcal{I}(V_\star)}{1-\lambda_\star\,\delta} = \alpha_\star(1-\lambda_\star\,\delta),\end{aligned}$$

from which we conclude that Equation (2.49b) is satisfied. This completes the proof of the lemma.

2.5.3.4 Proof of Lemma 2.2.2

By letting $\mathcal{L}(t) = t^2$ we find that $\mathcal{M}_{\mathcal{L}}(x;\tau) = \frac{x^2}{2\tau+1}$ for all $x \in \mathbb{R}$ and $\tau \in \mathbb{R}_{>0}$. Using this in Equations (2.49) and a after a few algebraic simplifications we arrive at the following closed-form expression for $\alpha^2_{\ell_2,\lambda}$ for all $\lambda \geq 0$ and random variables Z with finite second moment,

$$\alpha^2_{\ell_2,\lambda} = \frac{1}{2}\left(1 - \mathbb{E}[Z^2] - \delta\right) + \frac{\mathbb{E}[Z^2](\lambda + 2\delta + 2) + 2(\delta-1)^2 + \lambda(\delta+1)}{2\sqrt{(\lambda+2\delta-2)^2 + 8\lambda}}. \tag{2.60}$$

Next, by using direct differentiation to optimize this over $\lambda \geq 0$, we derive $\lambda_{\text{opt}} = 2\mathbb{E}[Z^2]$ and the resulting expression for $\alpha^2_{\ell_2,\lambda_{\text{opt}}}$ in the statement of the lemma.

2.5.3.5 Proof of Corollary 2.2.1

As mentioned in the main body of the chapter, the difficulty in deriving a closed-form expression for $\alpha_\star$ in (2.6) is due to the fact that in general $\mathcal{I}(V_a) = \mathcal{I}(aG + Z)$ may not be expressible in closed-form with respect to a. The core idea behind this corollary is using Stam's inequality (see Proposition 2.5.3) to bound $\mathcal{I}(V_a)$ in terms of $\mathcal{I}(aG) = a^{-2}$ and $\mathcal{I}(Z)$. Specifically, applying (2.26) to the random variables $a\,G$ and Z we find that:

$$\mathcal{I}(V_a) = \mathcal{I}(a\,G + Z) \leq \frac{\mathcal{I}(Z)}{1 + a^2\mathcal{I}(Z)}. \tag{2.61}$$

Substituting the RHS above in place of $\mathcal{I}(V_a)$ in the definition of $\alpha_\star$ in (2.6), let us define $\widehat{\alpha}$ as follows:

$$\widehat{\alpha} := \min_{0 \leq x < 1/\delta} \left\{ a \geq 0 : \frac{(a^2 - x^2\,\delta^2)\,\mathcal{I}(Z)}{(1 - x\,\delta)^2(1 + a^2\mathcal{I}(Z))} \geq \frac{1}{\delta} \right\}. \tag{2.62}$$

The remaining of proof has two main steps. First, we show that

$$\alpha_\star^2 \geq \widehat{\alpha}^2. \tag{2.63}$$

Second, we solve the minimization in (2.62) to yield a closed-form expression for $\widehat{\alpha}$.

Towards proving (2.63), note from the definition of $\alpha_\star$ and inequality (2.61) that there exists $x_\star \in [0, 1/\delta)$ such that

$$\frac{1}{\delta} = \frac{(\alpha_\star^2 - x_\star^2\,\delta^2)\,\mathcal{I}(V_\star)}{(1 - x_\star\,\delta)^2} \leq \frac{(\alpha_\star^2 - x_\star^2\,\delta^2)\,\mathcal{I}(Z)}{(1 - x_\star\,\delta)^2(1 + \alpha_\star^2\,\mathcal{I}(Z))}.$$

Thus, the pair $(\alpha_\star, x_\star)$ is feasible in (2.62). This and optimality of $\widehat{\alpha}$ in (2.62) lead to (2.63), as desired.

The next step is finding a closed-form expression for $\widehat{\alpha}$. Based on (2.62) and few algebraic simplifications we have :

$$\begin{aligned}
\widehat{\alpha}^2 &= \min_{0 \leq x < 1/\delta} \left\{ a^2 : a^2\,\mathcal{I}(Z) \cdot (\delta - (1 - x\,\delta)^2) \geq (1 - x\,\delta)^2 + \delta^3 x^2 \mathcal{I}(Z) \right\} \\
&= \min_{\max\{0, \frac{1-\sqrt{\delta}}{\delta}\} \leq x < 1/\delta} \left\{ a^2 : a^2 \geq \frac{(1 - x\delta)^2 + \delta^3\,x^2\,\mathcal{I}(Z)}{\mathcal{I}(Z) \cdot (\delta - (1 - x\delta)^2)} \right\} \\
&= \min_{\max\{0, \frac{1-\sqrt{\delta}}{\delta}\} \leq x < 1/\delta} \left\{ \frac{(1 - x\delta)^2 + \delta^3\,x^2\,\mathcal{I}(Z)}{\mathcal{I}(Z) \cdot (\delta - (1 - x\delta)^2)} \right\}.
\end{aligned} \tag{2.64}$$

The last equality above is true because the fraction in the constraint in the second line is independent of a. Next, by minimizing with respect to the variable x in (2.64), we reach $\widehat{\alpha}^2 = h_\delta(1/\mathcal{I}(Z))$.

Finally, we know from Proposition 2.5.3(f) that equality in (2.61) is achieved if and only if the noise is Gaussian i.e. $Z \sim \mathcal{N}(0, \zeta^2)$ for some $\zeta > 0$. Thus, if this is indeed the case, then $\alpha_\star = \widehat{\alpha}$ and the lower bound is achieved with replacing the Fisher information

of Z i.e, $\mathcal{I}(Z) = \zeta^{-2}$. This completes the proof of the corollary.

2.5.3.6 Proof of Equation (2.8)

First, we prove the bound $\omega_{_\delta} \geq (\mathcal{I}(Z)\,\mathbb{E}[Z^2])^{-1}$. Fix $\delta > 0$ and consider the function $\widetilde{h}_\delta(x) := h_\delta(x)/x$ for $x \geq 0$. Direct differentiation and some algebra steps suffice to show that $\widetilde{h}_\delta(x)$ is decreasing. Using this and the fact that $1/\mathcal{I}(Z) \leq \mathbb{E}[Z^2]$ (cf. Proposition 2.5.3 (c)), we conclude with the desired.

Next, we prove the lower bound $\omega_\delta \geq 1-\delta$. Fix any $\delta > 0$. First, it is straightforward to compute that $h_\delta(0) = \max\{1-\delta, 0\} \geq 1-\delta$. Also, simple algebra shows that $h_\delta(x) \leq 1, x \geq 0$. From these two and the increasing nature of $h_\delta(x)$ we conclude that $1-\delta \leq h_\delta(x) \leq 1$, for all $x \geq 0$. The desired lower bound follows immediately by applying these bounds to the definition of ω_δ.

2.5.4 Fundametal Limits for Binary Models: Proofs for Section 2.3

2.5.4.1 Discussion on Assumption 2.3.1

As per Assumption 2.3.1, the link function must satisfy $\mathbb{E}[Sf(S)] \neq 0$. This is a rather mild assumption in our setting. For example, it is straightforward to show that it is satisfied for the Signed, Logistic and Probit models. More generally, for a link function $f : \mathbb{R} \to \{\pm 1\}$ and $S \sim \mathcal{N}(0,1)$, the probability density of $Sf(S)$ can be computed as follows for any $x \in \mathbb{R}$:

$$p_{_{Sf(S)}}(x) = \Big(1 + \widehat{f}(x) - \widehat{f}(-x)\Big)\frac{\exp(-x^2/2)}{\sqrt{2\pi}}, \qquad \widehat{f}(x) := \mathbb{P}\left(f(x) = 1\right). \tag{2.65}$$

From this and the fact that $\exp(-x^2/2)$ is an even function of x, we can conclude that Assumption 2.3.1 is valid if $\widehat{f}(x)$ is monotonic and non-constant based on x (e.g., as in the Signed, Logistic and Probit models) . In contrast, Assumption 2.3.1 fails if the function $\widehat{f}$ is even. Finally, we remark that using (2.65), it can be checked that $S\,f(S) \sim \mathcal{N}(\mu, \zeta^2)$ if and only if $(\mu, \zeta) = (0, 1)$, and consequently only if $\widehat{f}$ is an even function. Based on these, we conclude that for all link functions f satisfying Assumption 2.3.1, the resulting distribution of $Sf(S)$ is non-Gaussian. Finally, we remark that $\nu_f = \mathbb{E}[Sf(S)]$ is the first Hermite coefficient of the function f and the requirement $\nu_f \neq 0$ arises in a series of recent works on high-dimensional single-index models, e.g., [69, 70]; see also [71, 72] for algorithms specializing to scenarios in which $\nu_f = 0$.

2.5.4.2 Discussion on the Classification Error (2.13)

First, we prove that for an estimator $\widehat{\mathbf{w}}_{\mathcal{L},\lambda}$, the relation $\mathbb{P}(\sigma_{\mathcal{L},\lambda} G + Sf(S) < 0)$ determines the high-dimensional limit of classification error. Then we show that the classification error is indeed an increasing function of $\sigma_{\mathcal{L},\lambda}$ for most well-known binary models.

For the estimator $\widehat{\mathbf{w}}_{\mathcal{L},\lambda}$ obtained from (2.9), and $\mathbf{x}_0$ denoting the true vector with unit norm, the parameters $\mu_{\mathcal{L},\lambda}$ and $\alpha_{\mathcal{L},\lambda}$ denote the high-dimensional terms of bias and variance,

$$\mathbf{x}_0^T \widehat{\mathbf{w}}_{\mathcal{L},\lambda} \xrightarrow{P} \mu_{\mathcal{L},\lambda}, \tag{2.66}$$

$$\|\widehat{\mathbf{w}}_{\mathcal{L},\lambda} - \mu_{\mathcal{L},\lambda}\, \mathbf{x}_0\|_2^2 \xrightarrow{P} \alpha_{\mathcal{L},\lambda}^2. \tag{2.67}$$

We note that by rotational invariance of Gaussian distribution we may assume without loss of generality that $\mathbf{x}_0 = [1,\, 0,\, 0,\, \cdots,\, 0]^T \in \mathbb{R}^n$. Therefore we deduce from (2.66) and

(2.67) that

$$\widehat{\mathbf{w}}_{\mathcal{L},\lambda}(1) \xrightarrow{P} \mu_{\mathcal{L},\lambda}, \qquad \sum_{i=2}^{n} (\widehat{\mathbf{w}}_{\mathcal{L},\lambda}(i))^2 \xrightarrow{P} \alpha^2_{\mathcal{L},\lambda}.$$

Using these, we derive the following for the classification error :

$$\begin{aligned}\mathcal{E}_{\mathcal{L},\lambda} &= \mathbb{P}\left(f\left(\mathbf{a}^T\mathbf{x}_0\right)\ \mathbf{a}^T\widehat{\mathbf{w}}_{\mathcal{L},\lambda} < 0 \right) \\ &= \mathbb{P}\left(f\left(\mathbf{a}(1)\right) \cdot \left(\widehat{\mathbf{w}}_{\mathcal{L},\lambda}(1)\mathbf{a}(1) + \widehat{\mathbf{w}}_{\mathcal{L},\lambda}(2)\mathbf{a}(2) + \cdots + \widehat{\mathbf{w}}_{\mathcal{L},\lambda}(n)\mathbf{a}(n)\right) < 0 \right).\end{aligned}$$

Recalling Assumption 2.1.2 we have $\mathbf{a} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$. Thus by denoting $S, G \overset{\text{iid}}{\sim} \mathcal{N}(0,1)$ and assuming without loss of generality that $\mu_{\mathcal{L},\lambda} > 0$, we derive (2.13).

Next, we show that for the studied binary models in this chapter, the high-dimensional limit for the classification error is increasing based on effective error term $\sigma > 0$. In particular, we find that if $p_{Sf(S)}(x) > p_{Sf(S)}(-x)$ for $x \in \mathbb{R}_{>0}$ then it is guaranteed that $a \mapsto \mathbb{P}(aG + Sf(S) < 0)$ is an increasing function for $a > 0$. To show this, we denote by ϕ the density of standard normal distribution and let $a_1 > a_2$ to be two positive constants, then under the given condition on $p_{Sf(S)}$, we deduce that,

$$\begin{aligned}&\mathbb{P}\left(Sf(S) < a_1 G\right) - \mathbb{P}\left(Sf(S) < a_2 G\right) = \\ &\qquad \int_0^{+\infty} \int_{a_2 g}^{a_1 g} p_{Sf(S)}(x)\, \phi(g)\, \mathrm{d}x\, \mathrm{d}g - \int_{-\infty}^{0} \int_{a_1 g}^{a_2 g} p_{Sf(S)}(x)\, \phi(g)\, \mathrm{d}x\, \mathrm{d}g > 0.\end{aligned}$$

This shows the desired. Importantly, we remark that in view of (2.65), this condition on the density of $Sf(S)$ is satisfied for many well-known binary models including Logistic, Probit and Signed.

2.5.4.3 Proof of Theorem 2.3.2

We need the following auxiliary result, which we prove first.

Lemma 2.5.6 (Boundedness of τ in (2.45))**.** *Fix $\delta > 0$ and $\lambda > 0$ and let $\mathcal{L}$ be a convex, twice differentiable and non-linear function. Then all solutions τ of the system of equations in (2.45) satisfy $0 < \tau < \frac{1}{\lambda\delta}$.*

Proof: The proof follows directly from the proof of Lemma 2.5.5 by replacing Z with $\mu\, Sf(S)$. Note that the Equation (2.45c) can be obtained by replacing Z with $\mu Sf(S)$ in Equation (2.49b).

Next, we proceed to the proof main of Theorem 2.3.2. For convenience, let us define the function $\Phi : \mathbb{R}_{\geq 0} \times [0, 1/\delta) \to \mathbb{R}$ as following :

$$\Phi(s, x) := \frac{1 - s^2(1 - s^2\mathcal{I}(W_s))}{\delta s^2(s^2\mathcal{I}(W_s) + \mathcal{I}(W_s) - 1)} - 2x + x^2\delta(1 + s^{-2}). \tag{2.68}$$

Then, $\sigma_\star$ as in (2.14) is equivalently expressed as:

$$\sigma_\star := \min_{0 \leq x < 1/\delta} \{s \geq 0 : \ \Phi(s, x) = 1\}, \tag{2.69}$$

Before everything, we show that $\sigma_\star$ is well defined, i.e., the feasible set of the minimization in (2.69) is non-empty for all $\delta > 0$ and link functions $f(\cdot)$ satisfying Assumption 2.3.1. Specifically, we will show that for any $\delta > 0$ there exists $s \geq 0$ such that $\widetilde{\Phi} := \Phi(s, \frac{s}{\delta(1+s)}) = 1$. It suffices to prove that the range of the function $\widetilde{\Phi}$ is $(0, \infty)$. Clearly, the function is continuous in $\mathbb{R}_{\geq 0}$. Moreover, it can be checked that

$$\widetilde{\Phi}(s) = \Phi_0(s) + \frac{2}{\delta(1 + s^2)}, \quad \text{where} \quad \Phi_0(s) := \frac{1 - s^2\mathcal{I}(W_s)}{\delta\, s^2(s^2\mathcal{I}(W_s) + \mathcal{I}(W_s) - 1)}. \tag{2.70}$$

But, by Lemma 2.5.2, $\lim_{s\to 0} \ s^2\,\mathcal{I}(W_s) = 0$ and $\lim_{s\to+\infty} s^2\,\mathcal{I}(W_s) = 1$. Using these, we

can show that $\lim_{s\to 0} \ \Phi_0(s) = +\infty$ and $\lim_{s\to+\infty} \ \Phi_0(s) = 0$. Combined with (2.70), we find that $\lim_{s\to 0} \ \Phi_0(s) = +\infty$ and $\lim_{s\to+\infty} \ s^2\,\mathcal{I}(W_s) = 0$. This concludes the proof of feasibility of the minimization in (2.68).

We are now ready to prove the main claim of the theorem. Fix convex loss function $\mathcal{L}$ and regularization parameter $\lambda \geq$. Let $(\alpha > 0, \mu, \tau > 0)$ be the unique solution to (2.45) and denote $\sigma = \alpha/\mu$. We will prove that

$$\sigma \geq \sigma_\star. \tag{2.71}$$

The first step in the proof will be to transform the equations (2.45) in a more appropriate form. In order to motivate the transformation, note that the performance of the optimization problem in (2.9) is unique up to rescaling. In particular consider the following variant of the optimization problem in (2.9) :

$$\widehat{\mathbf{v}}_{\mathcal{L},\lambda} := \arg\min_{\mathbf{w}} \left[\frac{c_1}{m} \sum_{i=1}^{m} \mathcal{L}\left(c_2\, y_i \mathbf{a}_i^T \mathbf{w}\right) + c_1 \lambda \|c_2 \mathbf{w}\|^2 \right], \quad c_1 > 0,\ c_2 \neq 0.$$

It is straightforward to see that, regardless of the values of c_1 and c_2, $\mathrm{corr}\left(\widehat{\mathbf{w}}_{\mathcal{L},\lambda}, \mathbf{x}_0\right) = \mathrm{corr}\left(\widehat{\mathbf{v}}_{\mathcal{L},\lambda}, \mathbf{x}_0\right)$, where recall that $\widehat{\mathbf{w}}_{\mathcal{L},\lambda}$ solves (2.9). Thus in view of (2.10), we see that the error σ resulting from $\widehat{\mathbf{w}}_{\mathcal{L},\lambda}$ and $\widehat{\mathbf{v}}_{\mathcal{L},\lambda}$ are the same. Motivated by this observation, we consider the following rescaling for the loss function and regularization parameter:

$$\widetilde{\mathcal{L}}(\cdot) := \frac{\tau}{\mu^2}\, \mathcal{L}(\mu\, \cdot), \qquad \widetilde{\lambda} := \tau\lambda, \tag{2.72}$$

From standard properties of Moreau-envelope functions it can be shown that

$$\mathcal{M}'_{\widetilde{\mathcal{L}},1}\left(\cdot/\mu\,; 1\right) = \frac{\tau}{\mu}\, \mathcal{M}'_{\mathcal{L},1}\left(\cdot\,; \tau\right).$$

Using these transformations, we can rewrite the system of equations (2.45) in terms of σ, $\widetilde{\mathcal{L}}$ and $\widetilde{\lambda}$ as follows:

$$\mathbb{E}\Big[Sf(S)\cdot\mathcal{M}'_{\widetilde{\mathcal{L}},1}\left(W_\sigma;1\right)\Big]=-\widetilde{\lambda}, \tag{2.73a}$$

$$\mathbb{E}\Big[\Big(\mathcal{M}'_{\widetilde{\mathcal{L}},1}\left(W_\sigma;1\right)\Big)^2\Big]=\sigma^2/\delta, \tag{2.73b}$$

$$\mathbb{E}\Big[G\cdot\mathcal{M}'_{\widetilde{\mathcal{L}},1}\left(W_\sigma;1\right)\Big]=\sigma(1-\widetilde{\lambda}\delta)/\delta. \tag{2.73c}$$

where we denote $W_\sigma := \sigma G + Sf(S)$.

Next, we further simplify (2.73) as follows. Similar to the procedure leading to (2.53), here also we may deduce that,

$$\mathbb{E}\Big[G\cdot\mathcal{M}'_{\widetilde{\mathcal{L}},1}\left(W_\sigma;1\right)\Big]=-\sigma\,\mathbb{E}\Big[\xi_{W_\sigma}(W_\sigma)\cdot\mathcal{M}'_{\widetilde{\mathcal{L}},1}\left(W_\sigma;1\right)\Big].$$

Thus (2.73c) can be rewritten as

$$\mathbb{E}\Big[\xi_{W_\sigma}(W_\sigma)\cdot\mathcal{M}'_{\widetilde{\mathcal{L}},1}\left(W_\sigma;1\right)\Big]=(\widetilde{\lambda}\delta-1)/\delta. \tag{2.74}$$

Additionally, we linearly combine (2.73a) and (2.73c) (with coefficient σ) to yield :

$$\mathbb{E}\Big[W_\sigma\cdot\mathcal{M}'_{\widetilde{\mathcal{L}},1}\left(W_\sigma;1\right)\Big]=\sigma^2/\delta-\sigma^2\widetilde{\lambda}-\widetilde{\lambda}, \tag{2.75}$$

Putting together (2.73b), (2.74) and (2.75), we have shown that σ satisfies the following

system of equations:

$$\mathbb{E}\Big[W_\sigma \cdot \mathcal{M}'_{\widetilde{\mathcal{L}},1}\left(W_\sigma;1\right)\Big] = \frac{\sigma^2}{\delta} - \sigma^2\widetilde{\lambda} - \widetilde{\lambda}, \tag{2.76a}$$

$$\mathbb{E}\Big[\left(\mathcal{M}'_{\widetilde{\mathcal{L}},1}\left(W_\sigma;1\right)\right)^2\Big] = \frac{\sigma^2}{\delta}, \tag{2.76b}$$

$$\mathbb{E}\Big[\xi_{W_\sigma}(W_\sigma)\cdot \mathcal{M}'_{\widetilde{\mathcal{L}},1}\left(W_\sigma;1\right)\Big] = \widetilde{\lambda} - \frac{1}{\delta}. \tag{2.76c}$$

Next, we will use this fact to derive a lower bound on σ. To this end, let $\beta_1, \beta_2 \in \mathbb{R}$ be two real constants. By combining (2.76a) and (2.76c) we find that

$$\mathbb{E}\Big[(\beta_1 W_\sigma + \beta_2 \xi_{W_\sigma}(W_\sigma))\cdot \mathcal{M}'_{\widetilde{\mathcal{L}},1}\left(W_\sigma;1\right)\Big] = \beta_1(\frac{\sigma^2}{\delta} - \sigma^2\widetilde{\lambda} - \widetilde{\lambda}) + \beta_2(\widetilde{\lambda} - \frac{1}{\delta}). \tag{2.77}$$

Applying Cauchy-Schwarz inequality to the LHS of (2.77) gives :

$$\begin{aligned}\left(\beta_1\left(\frac{\sigma^2}{\delta} - \sigma^2\widetilde{\lambda} - \widetilde{\lambda}\right) + \beta_2(\widetilde{\lambda} - \frac{1}{\delta})\right)^2 &\le \mathbb{E}\Big[\left(\beta_1 W_\sigma + \beta_2\xi_{W_\sigma}(W_\sigma)\right)\big)^2\Big]\cdot \mathbb{E}\Big[\left(\mathcal{M}'_{\widetilde{\mathcal{L}},1}\left(W_\sigma;1\right)\right)^2\Big] \\ &= \mathbb{E}\Big[\left(\beta_1 W_\sigma + \beta_2\xi_{W_\sigma}(W_\sigma)\right)\big)^2\Big]\frac{\sigma^2}{\delta},\end{aligned} \tag{2.78}$$

where we used (2.76b) in the last line. To simplify the expectation in the RHS of (2.78), we use the facts that $\mathbb{E}[W_\sigma^2] = \sigma^2 + 1$ and $\mathbb{E}[(\xi_{W_\sigma}(W_\sigma))^2] = \mathcal{I}(W_\sigma)$. Also by integration by parts one can derive that $\mathbb{E}[W_\sigma \cdot \xi_{W_\sigma}(W_\sigma)] = -1$. Thus we arrive at the following inequality from (2.78):

$$\left(\beta_1\left(\sigma^2/\delta - \sigma^2\widetilde{\lambda} - \widetilde{\lambda}\right) + \beta_2(\widetilde{\lambda} - 1/\delta)\right)^2 \le \beta_1^2\left(\sigma^2+1\right) + \beta_2^2\,\mathcal{I}(W_\sigma) - 2\beta_1\beta_2. \tag{2.79}$$

Now, we choose the coefficients β_1 and β_2 as follows: $\beta_1 = 1-\widetilde{\lambda}\delta-(\sigma^2-\sigma^2\widetilde{\lambda}\delta-\widetilde{\lambda}\delta)\,\mathcal{I}(W_\sigma)$ and $\beta_2 = 1$. (We show later in Theorem 2.3.1, that this choice lead to an achievable lower bound). Substituting these values in (2.79) and simplifying the resulting expressions yield

the following inequality for σ:

$$\frac{1-\sigma^2(1-\sigma^2\mathcal{I}(W_\sigma))}{\delta\sigma^2(\sigma^2\mathcal{I}(W_\sigma)+\mathcal{I}(W_\sigma)-1)}-2\widetilde{\lambda}+\widetilde{\lambda}^2\delta(1+\sigma^{-2})\leq 1. \tag{2.80}$$

We will now finish the proof of the theorem by using (2.80) to prove (2.71). For the sake of contradiction to (2.71), assume that $\sigma<\sigma_\star$. From (2.80) and the notation introduced in (2.68), we have shown that $\Phi\left(\sigma,\widetilde{\lambda}\right)\leq 1$. Recall from (2.72) that $\widetilde{\lambda}=\lambda\tau$. But, from Lemma 2.5.6 it holds that $\widetilde{\lambda}=\lambda\tau<\frac{1}{\delta}$. Therefore, the pair $(\sigma,\widetilde{\lambda})$ is feasible in the minimization problem in (2.69). By this, optimality of $\sigma_\star$ and our assumption that $\sigma<\sigma_\star$ in (2.69) it must hold that $\Phi\left(\sigma,\widetilde{\lambda}\right)<1$. But then, since $\lim_{s\to 0}\Phi\left(s,\widetilde{\lambda}\right)=+\infty$ and by continuity of the function $\Phi(\cdot,x)$ for all fixed $x\in[0,1/\delta)$, we have:

$$\exists\,\sigma_1:\ \text{s.t. } 0<\sigma_1<\sigma,\ \ \text{and}\ \ \Phi(\sigma_1,\widetilde{\lambda})=1. \tag{2.81}$$

Therefore $\Phi(\sigma_1,\widetilde{\lambda})=1$ for $\sigma_1<\sigma_\star$, which contradicts the optimality of $\sigma_\star$ in (2.69) and completes the proof.

2.5.4.4 Proof of Lemma 2.3.1

To prove the claim of the lemma we show that the proposed candidate-optimal loss and regularization parameter pair $(\mathcal{L}_\star,\lambda_\star)$ satisfies the system of equations in (2.45) with $(\alpha,\mu,\tau)=(\sigma_\star,1,1)$. In line with the proof of Theorem 2.3.2 and the equivalent representation of (2.76) for the equations in (2.45), we show that $(\mathcal{L}_\star,\lambda_\star)$ satisfy all three equations in (2.76) with $(\sigma,\mu,\tau)=(\sigma_\star,1,1)$. We emphasize that since $\mu=\tau=1$, based on (2.72) the $\mathcal{L}_\star$ and $\lambda_\star$ remain the same under these changes of parameters thus $(\widetilde{\mathcal{L}}_\star(\cdot),\widetilde{\lambda}_\star)=(\mathcal{L}_\star,\lambda_\star)$.

Note that we need $\mathcal{M}_{\mathcal{L}}(\cdot)$ to be able to assess the equations in (2.76). For this purpose

we use inverse properties of Moreau-envelope functions in Proposition 2.5.2 to derive the following from the definition of $\mathcal{L}_\star$ in (2.16) :

$$\mathcal{M}_{\mathcal{L}_\star}(w;1) = -\frac{\eta(\lambda_\star\delta-1)}{\delta(\eta-\mathcal{I}(W_\star))}Q(w) - \frac{\lambda_\star\delta-1}{\delta(\eta-\mathcal{I}(W_\star))}\log\left(p_{W_\star}(w)\right).$$

Thus,

$$\mathcal{M}'_{\mathcal{L}_\star,1}(w;1) = -\frac{\eta(\lambda_\star\delta-1)}{\delta(\eta-\mathcal{I}(W_\star))}w - \frac{\lambda_\star\delta-1}{\delta(\eta-\mathcal{I}(W_\star))}\xi_{W_\star}(w).$$

Using this and the fact that $\mathbb{E}[W_\star\cdot\xi_{W_\star}(W_\star)] = -1$ (derived by integration by parts), the LHS of the equation (2.76a) changes to

$$\begin{aligned}\mathbb{E}\Big[W_\star\cdot\mathcal{M}'_{\mathcal{L}_\star,1}(W_\star;1)\Big] &= -\frac{\eta(\lambda_\star\delta-1)}{\delta(\eta-\mathcal{I}(W_\star))}\mathbb{E}\left[W_\star^2\right] - \frac{\lambda_\star\delta-1}{\delta(\eta-\mathcal{I}(W_\star))}\mathbb{E}\left[W_\star\cdot\xi_{W_\star}(W_\star)\right]\\ &= -\frac{\eta(\lambda_\star\delta-1)}{\delta(\eta-\mathcal{I}(W_\star))}(\sigma_\star^2+1) + \frac{\lambda_\star\delta-1}{\delta(\eta-\mathcal{I}(W_\star))} = \frac{\sigma_\star^2}{\delta} - \sigma_\star^2\lambda_\star - \lambda_\star,\end{aligned}$$

where for the last step, we replaced η according to the statement of the lemma.

Similarly, for the second equation (2.76b), we begin with replacing the expression for $\mathcal{M}'_{\mathcal{L}_\star,1}(W_\star;1)$ to see that

$$\begin{aligned}\mathbb{E}\left[\left(\mathcal{M}'_{\mathcal{L}_\star,1}(W_\star;1)\right)^2\right] &= \frac{(\lambda_\star\delta-1)^2}{\delta^2(\eta-\mathcal{I}(W_\star))^2}\left(\eta^2\,\mathbb{E}\left[W_\star^2\right] + \mathcal{I}(W_\star) + 2\eta\,\mathbb{E}\left[W_\star\cdot\xi_{W_\star}(W_\star)\right]\right)\\ &= \frac{(\lambda_\star\delta-1)^2}{\delta^2(\eta-\mathcal{I}(W_\star))^2}\left(\eta^2(\sigma_\star^2+1) + \mathcal{I}(W_\star) - 2\eta\right). \qquad (2.82)\end{aligned}$$

After replacing η, we can simplify (2.82) to reach the following

$$\begin{aligned}\mathbb{E}\left[\left(\mathcal{M}'_{\mathcal{L}_\star,1}\left(W_\star;1\right)\right)^2\right] &= \frac{1-\sigma_\star^2(1-\sigma_\star^2\,\mathcal{I}(W_\star))}{\delta^2(\sigma_\star^2\,\mathcal{I}(W_\star)+\mathcal{I}(W_\star)-1)} - \frac{2\lambda_\star\,\sigma_\star^2}{\delta} + \lambda_\star^2(1+\sigma_\star^2)\\ &= \frac{\Phi(\sigma_\star,\lambda_\star)\cdot\sigma_\star^2}{\delta} = \frac{\sigma_\star^2}{\delta},\end{aligned}$$

where the last two steps follow from the definition of $\sigma_\star$ in (2.14) and $\Phi(\cdot,\cdot)$ in (2.68).

For the third Equation (2.76c) we deduce in a similar way that

$$\begin{aligned}\mathbb{E}\Big[\xi_{W_\star}(W_\star)\cdot\mathcal{M}'_{\mathcal{L}_\star,1}\left(W_\star;1\right)\Big] &= -\frac{\eta(\lambda_\star\delta-1)}{\delta(\eta-\mathcal{I}(W_\star))}\mathbb{E}\left[W_\star\cdot\xi_{W_\star}(W_\star)\right] - \frac{\lambda_\star\delta-1}{\delta(\eta-\mathcal{I}(W_\star))}\mathcal{I}(W_\star)\\ &= \lambda_\star - \frac{1}{\delta},\end{aligned}$$

confirming the RHS of Equation (2.76c). This completes the proof.

2.5.4.5 Proof of Lemma 2.3.2

Let $\ell_2(t) = (1-t)^2$ for $t\in\mathbb{R}$. Using the Equations in (2.45) and replacing $\mathcal{M}_{\ell_2}(x;\tau) = \frac{(x-1)^2}{2\tau+1}$ we can solve the equations to find the closed-form formulas for (μ,α,τ) for a fixed $\lambda\geq 0$. For compactness, define $F(\cdot,\cdot):\mathbb{R}_{>0}\times\mathbb{R}_{>0}\to\mathbb{R}_{>0}$ where $F(\delta,\lambda) := \lambda\delta+\sqrt{8\lambda\delta+(\delta(\lambda+2)-2)^2}$. We derive the following for $\mu_{\ell_2,\lambda}$ and $\alpha_{\ell_2,\lambda}$ and for all $\delta>0$,

$$\begin{aligned}\mu_{\ell_2,\lambda} &= \frac{4\delta\,\mathbb{E}[Z^2]}{2+2\delta+F(\delta,\lambda)},\\ \alpha^2_{\ell_2,\lambda} &= \frac{\delta\left(2-2\delta-2\lambda\delta+F(\delta,\lambda)\right)^2\left(2+2\delta+F(\delta,\lambda)\right)\left(1-\frac{8\delta\left(\mathbb{E}[Z^2]\right)^2(2+F(\delta,\lambda))}{(2+2\delta+F(\delta,\lambda))^2}\right)}{2\left(2-2\delta+F(\delta,\lambda)\right)^2\left(F(\delta,\lambda)-\lambda\delta\right)}.\end{aligned}$$

Using these, we reach $\sigma^2_{\ell_2,\lambda} = \alpha^2_{\ell_2,\lambda}/\mu^2_{\ell_2,\lambda}$ as stated in (2.17). By minimizing $\sigma^2_{\ell_2,\lambda}$ with respect to $\lambda\geq 0$ we derive λ_{opt} and the resulting $\sigma^2_{\ell_2,\lambda_{\text{opt}}}$ in the statement of the lemma.

2.5.4.6 Proof of Corollary 2.3.1

The proof is analogous to the proof of Corollary 2.2.1. Here again we use Stam's inequality in Proposition 2.5.3 to provide a bound for $\mathcal{I}(W_\sigma) = \mathcal{I}(\sigma G + Sf(S))$ based on $\mathcal{I}(\sigma G) = \sigma^{-2}$ and $\mathcal{I}(Sf(S))$. First we define

$$\widehat{\sigma} := \min_{x \geq 0} \left\{ s \geq 0 : \frac{1}{\delta} + \frac{1}{\delta s^2 (\mathcal{I}(Sf(S)) - 1)} - 2x + \delta x^2 (1 + s^{-2}) \leq 1 \right\}. \tag{2.83}$$

Next we use Stam's inequality to deduce that :

$$\mathcal{I}(W_\sigma) := \mathcal{I}(\sigma G + Sf(S)) \leq \frac{\mathcal{I}(Sf(S))}{1 + \sigma^2 \mathcal{I}(Sf(S))}.$$

We can use this inequality in the constraint condition of $\sigma_\star$ in (2.14) to deduce that:

$$\frac{1}{\delta} + \frac{1}{\delta \sigma_\star^2 (\mathcal{I}(Sf(S)) - 1)} - 2\lambda_\star + \delta \lambda_\star^2 (1 + \sigma_\star^{-2}) \leq 1, \tag{2.84}$$

Thus we find that $(\sigma, x) = (\sigma_\star, \lambda_\star)$ is a feasible solution of the constraint in (2.83), resulting in :

$$\sigma_\star \geq \widehat{\sigma}. \tag{2.85}$$

To complete the proof of the theorem, we need to find the closed-form $\widehat{\sigma}$. Proceeding from (2.83) we derive the following

$$\begin{aligned}\widehat{\sigma}^2 &= \min_{x\geq 0}\left\{s^2 : \frac{1}{s^2}\left(\frac{1}{\delta(\mathcal{I}(Sf(S))-1)}+x^2\delta\right)\leq 1+2x-\frac{1}{\delta}-x^2\delta\right\}\\ &= \min_{x\geq 0}\left\{s^2 : \frac{1}{s^2}\leq \frac{1+2x-1/\delta-x^2\delta}{\frac{1}{\delta(\mathcal{I}(Sf(S))-1)}+x^2\delta}\right\}\\ &= \left(\max_{x\geq 0}\left\{\frac{1+2x-1/\delta-x^2\delta}{\frac{1}{\delta(\mathcal{I}(Sf(S))-1)}+x^2\delta}\right\}\right)^{-1}.\end{aligned}$$

The first line follows by algebraic simplifications in (2.83). The second line is true since by Cramer-Rao bound (see Proposition 2.5.3 (d)) $\mathcal{I}(Sf(S)) \geq (Var[Sf(S)])^{-1}$; thus $\mathcal{I}(Sf(S)) \geq 1$. Noting that the right hand-side of the inequality is independent of σ and can take positive values for some $x \geq 0$ we conclude the last line. Optimizing with respect to the non-negative variable x in the last line completes the proof and yields the desired result in the statement of the corollary.

2.5.5 Comparison to a Simple Averaging Estimator

In this section, we compare the performance of optimally ridge-regularized ERM to the following simple averaging estimator

$$\widehat{\mathbf{w}}_{\text{ave}} = \frac{1}{m}\sum_{i=1}^{m} y_i \mathbf{a}_i. \tag{2.86}$$

This estimator is closely related to the family of RERM estimators studied in this chapter. To see this, note that $\widehat{\mathbf{w}}_{\text{ave}}$ can be expressed as the solution to ridge-regularized

ERM with $\lambda = 1$ and linear loss function $\mathcal{L}(x) = -x$ for all $x \in \mathbb{R}$:

$$\widehat{\mathbf{w}}_{\text{ave}} = \arg\min_{\mathbf{w}\in\mathbb{R}^n} \frac{1}{2m}\sum_{i=1}^{m} \|y_i\mathbf{a}_i - \mathbf{w}\|_2^2 = \arg\min_{\mathbf{w}\in\mathbb{R}^n} \frac{1}{m}\sum_{i=1}^{m} -y_i\mathbf{a}_i^T\mathbf{w} + \frac{1}{2}\|\mathbf{w}\|_2^2.$$

Moreover, it is not hard to check that the correlation performance of $\widehat{\mathbf{w}}_{\text{ave}}$ is the same as that of the solution of RLS with regularization λ approaching infinity.

It is in fact possible to exploit these relations of the estimator to the RERM family in order to evaluate its asymptotic performance using the machinery of this chapter (i.e., by using the Equations (2.11)). However, a more direct evaluation that uses the closed form expression in (2.86) is preferable here. In fact, it can be easily checked that the following limit is true in the high-dimensional asymptotic regime:

$$\forall \delta > 0 \ : \qquad \text{corr}\left(\widehat{\mathbf{w}}_{\text{ave}}, \mathbf{x}_0\right) \xrightarrow{P} \frac{1}{1+\frac{1}{\delta}\nu_f^{-2}}, \tag{2.87}$$

where recall our notation $\nu_f = \mathbb{E}[Sf(S)],\ S \sim \mathcal{N}(0,1)$. The use of the simple averaging estimator for signal recovery in generalized linear models (also, in single-index models) has been previously investigated for example in [72].

A favorable feature of $\widehat{\mathbf{w}}_{\text{ave}}$ is its computational efficiency. In what follows, we use our lower bounds on the performance of general RERM estimators, to evaluate its suboptimality gap compared to more complicated alternatives. To begin, in view of (2.87) and (2.10) let us define the corresponding "effective error parameter"

$$\sigma_{\text{ave}}^2 = \frac{1}{\delta}\nu_f^{-2}. \tag{2.88}$$

First, we compare this value with the error of regularized LS. Let $\widehat{\mathbf{w}}_{\text{LS}}$ be the solution to

unregularized LS for $n > m$. It can be checked (e.g., [46]) that

$$\operatorname{corr}\left(\widehat{\mathbf{w}}_{\mathrm{LS}}, \mathbf{x}_0\right) \xrightarrow{P} \frac{1}{1+\sigma_{\mathrm{LS}}^2}, \quad \text{where} \quad \sigma_{\mathrm{LS}}^2 := \frac{1}{\delta-1}(\nu_f^{-2}-1). \tag{2.89}$$

Directly comparing this to (2.88), we find that $\frac{\sigma_{\mathrm{LS}}^2}{\sigma_{\mathrm{ave}}^2} = \left(\frac{1}{1-1/\delta}\right)(1-\nu_f^2)$, for all $\delta > 1$. In other words,

$$\sigma_{\mathrm{ave}}^2 \gtrless \sigma_{\mathrm{LS}}^2 \quad \Longleftrightarrow \quad \delta \gtrless \nu_f^{-2}. \tag{2.90}$$

Next, we study the performance gap of the averaging estimator from the optimal RERM. For this, we use Corollary 2.3.1 to compare σ_{ave}^2 to the lower bound $\sigma_\star$. We find that for any $\delta > 0$ and any link function f satisfying the assumptions of Corollary 2.3.1:

$$1 \;\geq\; \frac{\sigma_\star^2}{\sigma_{\mathrm{ave}}^2} \;\geq\; \delta\,\nu_f^2 \cdot H_\delta\Big(\mathcal{I}(Sf(S))\Big). \tag{2.91}$$

We complement these bounds with numerical simulations in Section 5.3.

2.5.6 Gains of Regularization

2.5.6.1 Linear models

In this section, we study the impact of the regularization parameter on the best achievable performance. For this purpose, we compare $\alpha_\star$, the best achievable performance of ridge-regularized case, to the best achievable performance among non-regularized empirical risk minimization with convex losses denoted by α_{ureg}. By definition of α_{ureg}, for all convex losses $\mathcal{L}$, in the regime of $\delta > 1$ it holds that, $\alpha_{\mathrm{ureg}} \;\leq\; \alpha_{\mathcal{L},0}$. In [58], the authors compute a tight lower bound on α_{ureg} and show that it is attained provided that p_Z is log-concave. Our next result bounds the ratio $\alpha_\star^2 \,/\, \alpha_{\mathrm{ureg}}^2$, illustrating the impact of

regularization for a wide range of choices of $Z \sim \mathcal{D}$ and any $\delta > 1$.

Corollary 2.5.1. *Let the assumptions of Corollary 2.2.1 hold and $\delta > 1$. Then it holds that:*

$$\frac{(\delta-1)}{\mathbb{E}[Z^2]}\, h_\delta\left(\frac{1}{\mathcal{I}(Z)}\right) \;\le\; \frac{\alpha_\star^2}{\alpha_{\rm ureg}^2} \;\le\; \min\Big\{(\delta-1)\,\mathcal{I}(Z),\, 1\Big\}. \tag{2.92}$$

Proof: In order to obtain an upper bound for $\alpha_\star^2/\alpha_{\rm ureg}^2$ first we find a lower bound for $\alpha_{\rm ureg}^2$. We have

$$\alpha_{\rm ureg}^2\, \mathcal{I}(V_{\alpha_{\rm ureg}}) = \frac{1}{\delta},$$

thus we may apply the Stam's inequality (as stated in Proposition 2.5.3(f)) for $\mathcal{I}(V_{\alpha_{\rm ureg}})$ to derive the following lower bound :

$$\alpha_{\rm ureg}^2 \ge \frac{1}{(\delta-1)\mathcal{I}(Z)}. \tag{2.93}$$

Also note that it holds that $\alpha_\star^2 \le \alpha_{\ell_2,\lambda_{\rm opt}}^2$. Thus by recalling Lemma 2.2.2 and the fact that the function $h_\delta(\cdot) \le 1$ for all $\delta \ge 0$ we deduce that $\alpha_\star^2 \le 1$. Additionally since $\alpha_\star^2 \le \alpha_{\rm ureg}^2$, we conclude the upper bound in the statement of the Corollary. To proceed, we use the Cramer-Rao bound (see Proposition 2.5.3(d)) for $\mathcal{I}(V_{\alpha_{\rm ureg}})$ to derive the following upper bound for $\alpha_{\rm ureg}^2$ which holds for all $\delta > 1$:

$$\alpha_{\rm ureg}^2 \le \frac{\mathbb{E}[Z^2]}{\delta-1}.$$

This combined with the result of Corollary 2.2.1 derives the lower bound in the statement of the corollary and completes the proof.

Importantly, based on (2.92) we find that as $\delta \to 1$ the ratio $\alpha_\star^2\,/\,\alpha_{\rm ureg}^2$ reaches zero, implying the large gap between $\alpha^\star$ and $\alpha_{\rm ureg}$ in this regime. In the highly under-

parameterized regime where $\delta \to \infty$, by computing the limit in the lower bound our bound gives

$$\frac{1}{\mathbb{E}[Z^2]\,\mathcal{I}(Z)} \;\leq\; \lim_{\delta\to\infty} \frac{\alpha_\star^2}{\alpha_{\text{ureg}}^2} \;\leq\; 1\,. \tag{2.94}$$

For example, we see that in this regime when Z is close to a Gaussian distribution such that $\mathcal{I}(Z) \approx 1/\mathbb{E}[Z^2]$, then provably $\alpha_\star \approx \alpha_{\text{ureg}}$, implying that impact of regularization is infinitesimal in the resulting error. We remark that for other distributions that are far from Gaussian in the sense $\mathcal{I}(Z) \gg 1/\mathbb{E}[Z^2]$ the simple lower bound in (2.94) is not tight; this is because the bound of Corollary 2.2.1 is not tight in this case.

2.5.6.2 Binary models

In order to demonstrate the impact of regularization on the performance of ERM based inference, we compare $\sigma_\star$ with the optimal error of the non-regularized ERM for $\delta > 1$ which we denote by σ_{ureg}. Thus σ_{ureg} satisfies for all convex losses that $\sigma_{\text{ureg}} \leq \sigma_{\mathcal{L},0}$. The general approach for determining σ_{ureg} is discussed in [53] in which the authors also show the achievability of σ_{ureg} for well-known models such as the Signed and Logistic models.

Our next result quantifies the gap between σ_{ureg} and $\sigma_\star$ in terms of the label functions f and $\delta > 1$.

Corollary 2.5.2. *Let the assumptions of Theorem 2.3.2 hold and $\delta > 1$. Further assume the label function f is such that $p_{S\cdot f(S)}(x)$ is differentiable and positive for all $x \in \mathbb{R}$. Then it holds that:*

$$\frac{(\delta-1)\nu_f^2}{1-\nu_f^2}\, H_\delta\Big(\mathcal{I}(\,Sf(S)\,)\Big) \;\leq\; \frac{\sigma_\star^2}{\sigma_{\text{ureg}}^2} \;\leq\; \min\left\{\frac{\delta-1}{\delta}\cdot\frac{\mathcal{I}(Sf(S))-1}{\nu_f^2}\,,\,1\right\}. \tag{2.95}$$

Proof: To provide the bounds of the ratio $\sigma_\star^2/\sigma_{\text{ureg}}^2$, we follow a similar argument stated in the proof of Corollary 2.5.1. First, we use the result in [53] which states that for σ_{ureg}^2 and all $\delta > 1$ it holds that

$$\sigma_{\text{ureg}}^2 \geq \frac{1}{(\delta-1)(\mathcal{I}(Sf(S))-1)}. \tag{2.96}$$

Since it trivially holds that $\sigma_\star^2 \leq \sigma_{\ell_2,\lambda_{\text{opt}}}^2$ and also by noting that $\sigma_{\ell_2,\lambda_{\text{opt}}}^2$ as derived by Lemma 2.3.2 satisfies $\sigma_{\ell_2,\lambda_{\text{opt}}}^2 \leq \frac{1}{\delta\nu_f^2}$ for all $\delta > 0$ (which is followed by the fact that $H_\delta(x) \leq \frac{x}{(x-1)\delta}$), we conclude that

$$\sigma_\star^2 \leq \frac{1}{\delta\nu_f^2}. \tag{2.97}$$

Additionally since it trivially holds that $\sigma_\star^2 \leq \sigma_{\text{ureg}}^2$ we conclude the upper bound in the statement of the corollary. We proceed with proving the lower bound in the statement of the corollary. For this purpose, first we derive an upper bound for σ_{ureg}^2. Using the fact that σ_{ureg}^2 satisfies :

$$\frac{1-\sigma_{\text{ureg}}^2(1-\sigma_{\text{ureg}}^2\mathcal{I}(W_{\text{ureg}}))}{\delta\sigma_{\text{ureg}}^2(\sigma_{\text{ureg}}^2\mathcal{I}(W_{\text{ureg}})+\mathcal{I}(W_{\text{ureg}})-1)} = 1 \tag{2.98}$$

as well as the Cramer-Rao lower bound (Proposition 2.5.3(d)) for $\mathcal{I}(W_{\text{ureg}})$ we may deduce that :

$$\sigma_{\text{ureg}}^2 \leq \frac{(\delta-1)\nu_f^2}{1-\nu_f^2}. \tag{2.99}$$

This combined with the lower bound on $\sigma_\star^2$ as stated in Corollary 2.3.1 proves the lower bound in the statement of the corollary and completes the proof.

Importantly, as shown by (2.95), in the case of δ being close to 1, one can see that

both of the bounds in (2.95) vanish. This shows the large gap between σ_{ureg} and $\sigma_\star$ and further implies the benefit of regularization in this regime. When $\delta \to \infty$ i.e. in the highly under-parameterized regime, by deriving the limits as well as using Proposition 2.5.3 (d), we see that (2.95) yields:

$$\frac{\nu_f^2}{1-\nu_f^2} \cdot \frac{1}{\mathcal{I}(Sf(S))-1} \leq \lim_{\delta\to\infty} \frac{\sigma_\star^2}{\sigma_{\text{ureg}}^2} \leq 1. \tag{2.100}$$

Thus in this case both the values of $\sigma_\star$ and σ_{ureg} are approaching zero with the ratio depending on the properties of $Sf(S)$. For models such as Logistic with small signal strength (i.e. small $\|\mathbf{x}_0\|$) where $\mathcal{I}(Sf(S)) \approx 1/(1-\nu_f^2)$, one can derive that based on (2.100) the ratio reaches 1, which confirms the intuition that for large values of δ the impact of regularization is almost negligible.

2.5.7 Numerical Experiments

2.5.7.1 Details on Figure 2.1

In Fig. 2.1(Left), we compare the lower bound of Theorem 2.2.1 with the error of RLS (see Lemma 2.2.2) for $Z \sim \texttt{Laplace}(0,1)$ and $\|\mathbf{x}_0\|_2 = 1$. To numerically validate that $\alpha_\star$ is achievable by the proposed choices of loss function and regularization parameter in Lemma 2.2.1, we proceed as follows. We generate noisy linear measurements with iid Gaussian feature vectors $\mathbf{a}_i \in \mathbb{R}^{100}$. The estimator $\widehat{\mathbf{x}}_{\mathcal{L}_\star,\lambda_\star}$ is computed by running gradient descent (GD) on the corresponding optimization in (2.2) when the proposed optimal loss and regularizer of Lemma 2.2.1 are used. See Figure 2.3(Left) for an illustration of the optimal loss for this model. The resulting vector $\widehat{\mathbf{x}}_{\mathcal{L}_\star,\lambda_\star}$ is used to compute $\|\widehat{\mathbf{x}}_{\mathcal{L}_\star,\lambda_\star} - \mathbf{x}_0\|^2$. The average of these values over 50 independent Monte-carlo trials is shown in red squares. The close match between the theoretical and empirical values suggest that the fundamental

limits presented in this chapter are accurate even in small dimensions (also see the first and second rows of Table 2.1).

In the next two figures, we present results for binary models. Figure 2.1(Middle) plots the effective error parameter σ for the Signed model and Figure 2.1(Right) plots the classification error '$\mathcal{E}$' for the Logistic model with $\|\mathbf{x}_0\|_2 = 10$. The red squares correspond to the numerical evaluations of ERM with $\mathcal{L} = \mathcal{L}_\star$ and $\lambda = \lambda_\star$ (as in Lemma 2.3.1) derived by running GD on the proposed optimal loss and regularization parameter. See Figure 2.3(Right) for an illustration of the optimal loss in this case. The solution $\widehat{\mathbf{w}}_{\mathcal{L}_\star,\lambda_\star}$ of GD is used to calculate $\sigma_{\mathcal{L}_\star,\lambda_\star}$ and $\mathcal{E}_{\mathcal{L}_\star,\lambda_\star}$ in accordance with (2.10) and (2.13), respectively. Again, note the close match between theoretical and numerical evaluations (also see the third and fourth rows of Table 2.1).

Finally, for all three models studied in Figure 2.1, we also include the theoretical predictions for the error of the following: (i) RLS with small and large regularization (as derived in Equations (2.60) and (2.17)); (ii) optimally tuned RLS (as predicted by Lemmas 2.2.2 and 2.3.2); (iii) optimally-tuned unregularized ERM (marked as $\alpha_{\text{ureg}}, \sigma_{\text{ureg}}, \mathcal{E}_{\text{ureg}}$). The curves for the latter are obtained from [58] and [53] for linear and binary models, respectively. We refer the reader to Sections 2.5.6.1 and 2.5.6.2 for a precise study of the benefits of regularization in view of Theorems 2.2.1 and 2.3.2, for both linear and binary models.

2.5.7.2 Additional Experiments

In this section, we present additional numerical results comparing the bounds of Theorems 2.2.1 and 2.3.2 to the performance of the following: (i) Ridge-regularized Least-Squares (RLS); (ii) optimal unregularized ERM (Section 2.5.6); (iii) a simple averaging estimator (see Section 2.5.5). Figure 2.2(Top Left) plots the asymptotic squared error α^2 of these estimators for linear measurements with $Z \sim \texttt{Laplace}(0, 2)$. Similarly, Figure

Table 2.1: Theoretical and numerical values of $\alpha_\star^2/\alpha_{\mathcal{L},\lambda_{\text{opt}}}^2$ (for linear models) and $\sigma_\star^2/\sigma_{\mathcal{L},\lambda_{\text{opt}}}^2$(for binary models) for different values of δ and for some special cases studied in this chapter. The theoretical results for $\alpha_\star$ and $\sigma_\star$ correspond to Theorems 2.2.1 and 2.3.2. The empirical values of $\alpha_\star$ and $\sigma_\star$ are derived by numerically solving the optimally-tuned RERM (as derived in Lemmas 2.2.1 and 2.3.1) by GD with $n = 100$. Results shown are averages over 50 independent experiments.

	δ	0.5	2	4	6	8
$Z \sim \text{LAPLACE}(0,1)$	THEORY	0.9798	0.9103	0.8332	0.7690	0.7447
	EXPERIMENT	0.9700	0.8902	0.8109	0.7530	0.7438
$Z \sim \text{LAPLACE}(0,2)$	THEORY	0.9832	0.9329	0.8796	0.8371	0.8043
	EXPERIMENT	0.9785	0.9103	0.8550	0.8316	0.7864
$f = \text{SIGN}$	THEORY	0.9934	0.8531	0.6199	0.4602	0.3618
	EXPERIMENT	0.9918	0.8204	0.6210	0.4710	0.3829
$f = \text{LOGISTIC}, \|\mathbf{x}_0\| = 10$	THEORY	0.9826	0.8721	0.7116	0.6211	0.5712
	EXPERIMENT	0.9477	0.8987	0.7112	0.6211	0.6389

2.2(Top Right) and Figure 2.2(Bottom) plot the effective error term σ for Logistic data with $\|\mathbf{x}_0\|_2 = 1$, and the limiting value ρ of the correlation measure for Logistic data with $\|\mathbf{x}_0\|_2 = 10$, respectively. The red squares represent the performance of optimally tuned ERM (as per Lemmas 2.2.1 and 2.3.1) derived numerically by running GD, as previously described in the context of Figure 2.1.

The numerical findings in Figures 2.1 and 2.2 validate the theoretical findings of Sections 2.2.3 and 2.3.3, regarding sub-optimality of RLS for Laplace noise and Logistic binary model (with large $\|\mathbf{x}_0\|$) and optimality of λ-tuned RLS for Logistic model with small $\|\mathbf{x}_0\|$. Furthermore, by comparing the optimal performance of unregularized ERM to the optimal errors of RERM in both Figures 2.1 and 2.2, we confirm the the theoretical guarantees of Section 2.5.6 regarding the impact of regularization in the regime of small δ for both linear and binary models.

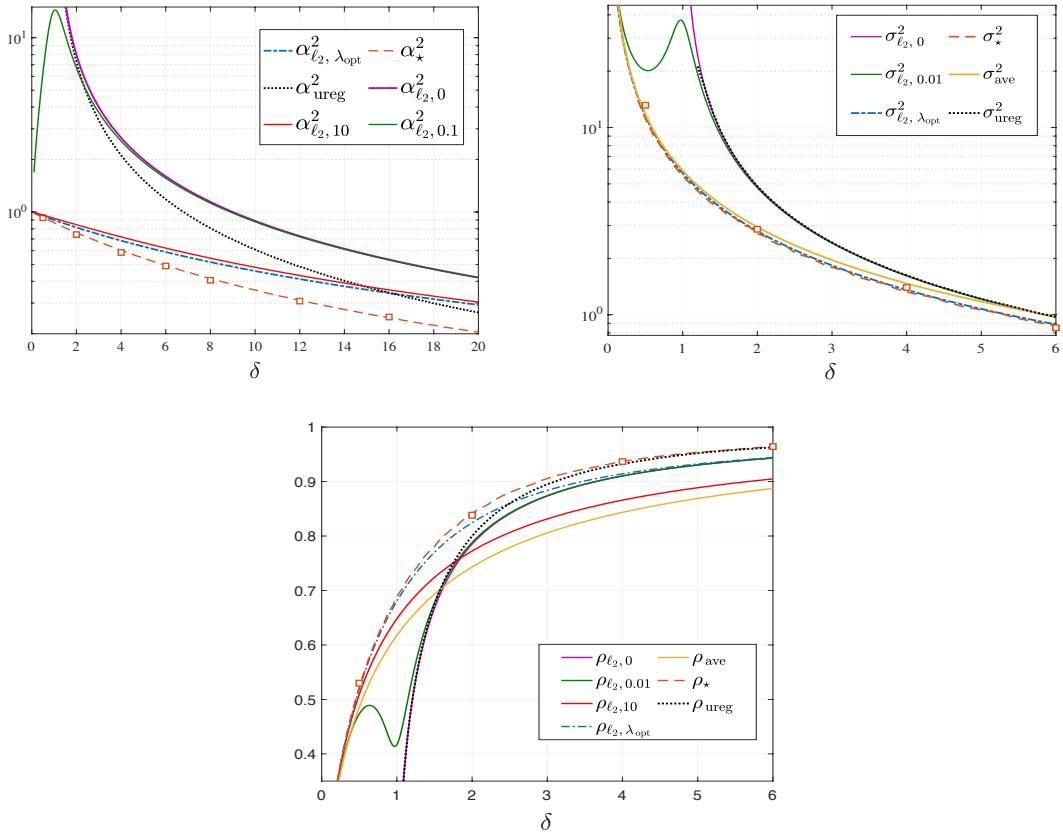


Figure 2.2: Fundamental error bounds derived in this chapter compared to RLS, averaging estimator and optimal unregularized ERM for: (Top Left) a linear model with $Z \sim \texttt{Laplace}(0,2)$, (Top Right) a binary Logistic model with $\|\mathbf{x}_0\|_2 = 1$, (Bottom) a binary Logistic model with $\|\mathbf{x}_0\|_2 = 10$ (here shown is correlation measure (2.10)). The red squares correspond to numerical evaluation of the performance of the optimally tuned RERM as derived in Lemmas 2.2.1 and 2.3.1; see text for details.

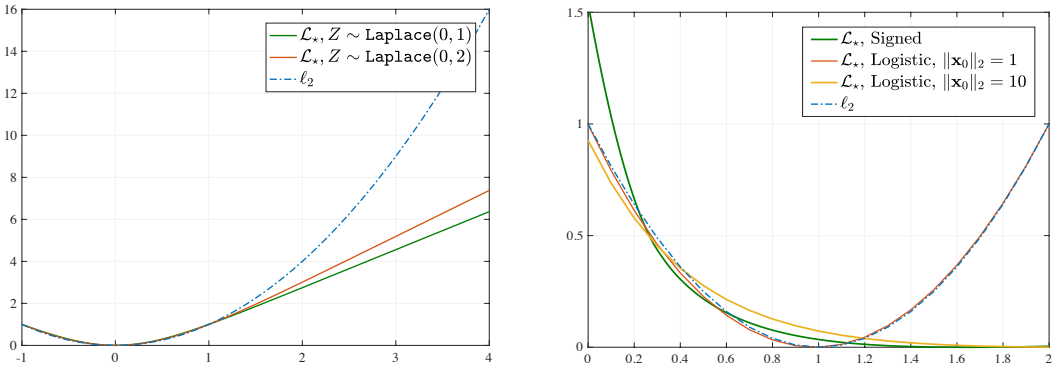


Figure 2.3: Illustrations of the proposed loss functions achieving optimal performance (as in Lemmas 2.2.1 and 2.3.1), for three special cases: a linear model with additive Laplace noise, the binary logistic model and the binary signed model. Here, in both plots, we fix $\delta = 2$. The curves are appropriately shifted and rescaled to allow direct comparison to the least-squares loss function; see text for details.

2.5.7.3 Optimal Tuning in Special Cases

Figure 2.3 depicts the candidate for optimal loss function derived in Lemmas 2.2.1 and 2.3.1, for specific linear and binary models discussed in this chapter. To allow for a direct comparison with the least-squares loss function, the optimal losses for the linear models are shifted such that $\mathcal{L}_\star \geq 0$ and rescaled such that $\mathcal{L}_\star(1) = 1$. Similarly, for the Logistic model with $\|\mathbf{x}_0\| = 1$, the optimal loss is rescaled such that $\mathcal{L}_\star(1) = 0$ and $\mathcal{L}_\star(2) = 1$. Interestingly, for this model, $\mathcal{L}_\star$, when rescaled (which results in no change in performance by appropriately rescaling $\lambda_\star$) is similar to the least-squares loss. This confirms the (approximate) optimality of optimally-tuned RLS for this model and further verifies the numerical observations in Figure 2.2 (Top Right) and the theoretical guarantees of Section 2.3.3 for this model.

Chapter 3

Adversarial Training with High-dimensional Linear Models

3.1 Introduction

Several machine learning models ranging from simple linear classifiers to complex deep neural networks have been shown to be prone to adversarial attacks, i.e., small additive perturbations to the data that cause the model to predict a wrong label [73, 74]. The requirement for robustness against adversaries is crucial for the safety of systems that rely on decisions made by these algorithms (e.g., in self-driving cars). With this motivation, over the past few years, there have been remarkable efforts by the research community to construct defense mechanisms, e.g., see [75, 76] for a survey. Among many proposals in the already rich literature, perhaps the most popular approach is that of adversarial training [6]. Among many favorable properties, adversarial training is flexible and easy-to-adjust to different types of data perturbations and has also been shown to achieve state-of-the-art performance in several tasks [7]. However, despite major recent progress in the study and implementation of adversarial training, its efficacy has been

mainly shown empirically without providing much theoretical understanding. Indeed, many questions regarding its theoretical properties remain open even for simple models. For instance, how does the adversarial/standard error depend on the adversary's budget during training time and test time? How do they depend on the over-parameterization ratio? What is the role of the chosen loss function?

In this chapter, we consider the adversarial training problem for ℓ_q-norm bounded perturbations in classification tasks, which solves the following robust empirical risk minimization (ERM) problem:

$$\min_{\boldsymbol{\theta}\in\mathbb{R}^l} \sum_{i=1}^{m} \max_{\|\boldsymbol{\delta}_i\|_q \le \varepsilon_{\rm tr}} \widetilde{\mathcal{L}}\left(y_i, f_{\boldsymbol{\theta}}(\mathbf{x}_i+\boldsymbol{\delta}_i)\right) + \lambda\|\boldsymbol{\theta}\|_2^2. \tag{3.1}$$

Here, $\{(\mathbf{x}_i, y_i)\}_{i\in[m]} \in \mathbb{R}^n \times \{\pm 1\}$ is the training set, $\boldsymbol{\delta}_i \in \mathbb{R}^n$ are the perturbations with l the dimension of the feature space, $f_{\boldsymbol{\theta}} : \mathbb{R}^n \to \mathbb{R}$ is a model parameterized by a vector $\boldsymbol{\theta} \in \mathbb{R}^l$, $\varepsilon_{\rm tr}$ is a user-specified tunable parameter that can be interpreted as the adversary's budget during training, and λ is the ridge-regularization parameter. Once the robust classifier $\widehat{\boldsymbol{\theta}}$ is obtained by (3.1), the *adversarial error / robust classification error* is given by

$$\mathbb{E}_{\mathbf{x},y}\left[\max_{\|\boldsymbol{\delta}\|_q\le\varepsilon_{\rm ts}} \mathbf{1}_{\left\{y\neq \mathrm{sign}(f_{\widehat{\boldsymbol{\theta}}}(\mathbf{x}+\boldsymbol{\delta}))\right\}}\right],$$

where $\mathbf{1}_{\{.\}}$ is the 0/1-indicator function, $(\mathbf{x}, y) \in \mathbb{R}^n \times \{\pm 1\}$ is a test sample drawn from the same distribution as that of the training dataset, $\varepsilon_{\rm ts}$ is the budget of the adversary, and $f_{\widehat{\boldsymbol{\theta}}}$ uses the trained parameters $\widehat{\boldsymbol{\theta}}$ and the fresh sample $\mathbf{x}$ to output a label guess. The standard classification error is given by the same formula by simply setting $\varepsilon_{\rm ts} = 0$.

The goal of this chapter is to precisely analyze the performance of adversarial training in (3.1) for binary classification with linear models i.e., $f_{\boldsymbol{\theta}}(\mathbf{x}) = \langle\boldsymbol{\theta}, \mathbf{x}\rangle$. In our proof we use the Convex-Gaussian-Min-max-Theorem (CGMT) [**?**, **?**, **?**] and in particular its

applications to the convex ERM that enables its precise analysis, e.g., [5, 55, 52, 53, 77]. However, compared to previous works, we develop a new analysis for robust optimization with correlated data.

Our main contributions are summarized as follows:

• We precisely analyze, for the first time, the performance of adversarial training with ℓ_2 and ℓ_∞ attacks in binary classification for two important data models of Gaussian Mixtures and Generalized Linear Models. See Sections 3.3 and 3.4.

• Our approach is general, allowing us to characterize the role of feature correlation, regularization and general ℓ_q attacks with $q \geq 1$. In particular, our proof technique allows for non-isotropic features, yielding novel theoretical results even for non-adversarial convex regularized ERM settings (i.e., when $\varepsilon_{\rm tr} = \varepsilon_{\rm ts} = 0$). We elaborate on our technical approach in Section 3.3.3.

• Numerical illustrations in Section 3.3.2 show tight agreements between our theoretical and empirical results and also allow us to draw intriguing conclusions regarding the behavior of adversarial and standard errors as functions of key problem parameters such as the sampling ratio $\delta := m/n$, the budget of the adversary $\varepsilon_{\rm ts}$, and the robust-optimization hyper-parameter $\varepsilon_{\rm tr}$ in our studied settings. Moreover, we observe interesting phonemena by comparing our results with the Bayes optimal robust errors.

3.1.1 Prior Works

Relevant to the flavour of our results, the recent work [78] studies precise tradeoffs and performance analysis in adversarial training with linear regression with ℓ_2 perturbations and isotropic Gaussian data. Compared to [78], our results hold for binary models, general ℓ_q perturbations with $q \geq 1$, non-isotropic features with mild assumptions on the covariance matrix. Moreover, we consider regularized ERM allowing us to study the

behavior of adversarial training in the over-parameterized regime in the limit of $\lambda \to 0$. Similar results on the behavior of adversarial training in classification are only derived in a concurrent work by [79]. On the one hand, compared to [79] our analysis applies to both discriminative and generative data models, and also to the *regularized* ERM. Our analysis also allows generic covariance matrices while the analysis of [79] only applies to very specfic structures for the covariance matrix. Additionally, we examine how our formulae on adversarial training compare with those of the Bayes robust estimator. On the other hand, [79] extend their analysis to robust support vector machines (SVM). Note however that we can retrieve the same results regarding the performance of adversarially-robust SVM by evaluating our formulae on regularized ERM with logistic loss and vanishing regularization parameter.

Our analysis of correlated features was motivated by [55], which derives sharp generalization gaurantees for SVM models with correlated data. Very recently, correlated features have been considered in various settings, e.g., [80, 81, 82]. However, none of these works studies the more challenging problem of adversarially-robust ERM as we do here. To see, at a high-level, why this differs from standard ERM or standard SVM analysis note the following complications in the analysis. First, because adversarial training is formulated as a min-max optimization, it is not at all apparent that the machinery of Gaussian comparison theorems applies. Second, the performance metric here is robust error (rather than standard error), and we show that this changes the statistics that needs to be tracked by the CGMT analysis. Third, the primary optimization to which we eventually apply the CGMT involves an "effective" ℓ_p-regularizer (where ℓ_p is the dual norm of the adversary's ℓ_q-norm), which unlike previous works appears inside the argument of the loss function, requiring new techniques to scalarize the auxiliary optimization. Specially, we do this in the presence of non-isotropic features, which yields new results even for standard ERM methods.

The Adversarial Bayes risk for Gaussian-mixtures has been recently characterized in [83, 84, 85]. Here, we combine their results with our precise asymptotics on the practically relevant adversarial training method, allowing us to investigate fundamental limits of the latter. The references [86, 87] discuss optimization landscape of adversarial training, however these works do not address generalization properties of adversarial training, as done in this chapter. The prior work [88] considers adversarial training with linear loss in order to analyze the sample complexity of robust estimators. Instead, here we investigate the more challenging, but practically more relevant, 0/1-loss and its tractable approximations (e.g. hinge, logistic). Another related line of work studies trade-offs between the standard and adversarial errors e.g., see [89, 90, 91, 85], but for simpler algorithms and data models, rather than adversarial training and correlated GLM/GMM, which we focus on here. The benefits of unlabeled data in robustness have been investigated in several works, e.g. [92, 93]. An exciting direction opening up with our analysis is investigating adversarial training performance for random features and neural tangent models. To date, precise asymptotics for such models have been obtained only very recently and for the simpler problem of standard ERM [94, 95, 96, 81, 97]. A preliminary version of this work appeared in [98]. The results presented in [98] only apply to data that follow the isotropic Gaussian mixture model and only to ℓ_∞ attacks. The current manuscript significantly extends the scope of these results: First, we extend the results for GMM to general covariance matrices (not necessarily isotropic). This is important because it better captures data distributions in practice. We also note that the extension is technically nontrivial, requiring several modifications in the proof compared to the isotropic case. Second, in the journal version we describe unifying analysis and results that applies both to discriminative and generative models. Specifically for discriminative models, we present new results for GLM data. Third, we provide a general analysis of ℓ_p-norm attacks. This extends the results of the conference version that only applied to

ℓ_∞-norm. For demonstration, we present results for ℓ_2-attacks in Section 3.4. Finally, we have extended our numerical study by introducing additional experiments in Appendix 3.7.

Notation

Letting $\delta(x)$ denote a Dirac delta mass at x, the empirical distribution of a vector $\mathbf{x} \in \mathbb{R}^n$ is given by $\frac{1}{n}\sum_{i=1}^n \delta(\mathbf{x}_i)$. The empirical joint distribution of $\mathbf{v}, \mathbf{u} \in \mathbb{R}^n$ is given by $\frac{1}{n}\sum_{i=1}^n \delta(\mathbf{v}_i, \mathbf{u}_i)$. The Wasserstein-$k$ distance between two measures ρ_1, ρ_2 is defined as $W_k(\rho_1, \rho_2) \triangleq \left(\inf_{\rho \in \mathrm{P}} \mathbb{E}_{(X,Y)\sim\rho}|X-Y|^k\right)^{1/k}$, where P denotes all couplings of ρ_1 and ρ_2. We say that a sequence of probability distributions μ_n converges in Wasserstein-k distance to a probability distribution μ, if $W_k(\mu_n, \mu) \to 0$ as $n \to \infty$. The Gaussian Q-function is denoted by $Q(\cdot)$. $\odot$ denotes the element-wise multiplication. The function $\|\cdot\|_q^p$ is denoted by ℓ_q^p. For a positive semi-definite matrix S we define $\|\mathbf{v}\|_S \triangleq \sqrt{\mathbf{v}^T S \mathbf{v}}$. Finally, for a sequence of random variables $X_{m,n}$ that converges in probability to some constant c in the proportional asymptotic limit $m, n \to \infty$, $m/n \to \delta$, we write $X_{m,n} \xrightarrow{P} c$.

3.2 Problem Formulation

In this section, we describe the data model, the specific form of (3.1), and the asymptotic regime for which our results hold. After this section, it is understood that all our results hold in the setting described here without any further explicit reference.

3.2.1 Data Model

We study two stylized models for binary classification.

Gaussian Mixture Models. The first model is a Gaussian Mixture model (GMM) where the conditional distribution of the feature vectors is a Gaussian with mean $\pm\boldsymbol{\theta}_n^\star$ (depending on the label $y_i \in \{\pm 1\}$) and with covariance $\boldsymbol{\Sigma}_n$. The subscript n emphasizes the dependence on dimension. Formally, the GMM assumes

$$\mathbb{P}(y_i = 1) = \pi \in [0,1], \quad \mathbf{x}_i | y_i \sim \mathcal{N}(y_i \boldsymbol{\theta}_n^\star, \boldsymbol{\Sigma}_n). \tag{3.2}$$

Generalized Linear Models. The second model is a generalized linear model (GLM) with binary link function. Specifically, assume that the label $y_i \in \{\pm 1\}$ associated with the feature vector $\mathbf{x}_i$ is generated as

$$y_i = \psi\left(\langle \boldsymbol{\theta}_n^\star, \mathbf{x}_i \rangle\right), \quad \mathbf{x}_i \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}_n), \tag{3.3}$$

for a possibly random link function $\psi : \mathbb{R} \to \{\pm 1\}$. This includes the well-known Logistic and Signed models, by letting $\mathbb{P}(\psi(x) = 1) = 1/(1+\exp(-x))$ and $\psi(x) = \text{sign}(x)$, respectively.

We assume that the underlying (unknown) vector of regressors $\boldsymbol{\theta}_n^\star \in \mathbb{R}^n$, and the covariance matrix $\boldsymbol{\Sigma}_n \in \mathbb{R}^{n \times n}$, satisfy the following technical (and mild) assumptions.

Assumption 3.2.1. *The minimum and maximum eigenvalues of the covariance matrices $\boldsymbol{\Sigma}_n$ satisfy $0 < c < \lambda_{\min}(\boldsymbol{\Sigma}_n)$ and $\lambda_{\max}(\boldsymbol{\Sigma}_n) < C < \infty$.*

Assumption 3.2.2. *Denoting $\zeta_n \triangleq (\boldsymbol{\theta}_n^{\star\top} \boldsymbol{\Sigma}_n \boldsymbol{\theta}_n^\star)^{1/2}$ for GLM and $\widetilde{\zeta_n} \triangleq (\boldsymbol{\theta}_n^{\star\top} \boldsymbol{\Sigma}_n^{-1} \boldsymbol{\theta}_n^\star)^{1/2}$ for GMM, we define their high-dimensional limits as ζ and $\widetilde{\zeta}$, i.e., $\zeta_n \xrightarrow{P} \zeta$ and $\widetilde{\zeta_n} \xrightarrow{P} \widetilde{\zeta}$. Moreover, for both models we assume without loss of generality that $\|\boldsymbol{\theta}_n^\star\|_2 \xrightarrow{P} 1$.*

Assumption 3.2.3. *Let $\boldsymbol{\Sigma}_n = \mathbf{U}_n \boldsymbol{\Lambda}_n \mathbf{U}_n^\top$ be the eigen-decomposition of $\boldsymbol{\Sigma}_n$ and let $\lambda_{n,i}$ denote the i'th entry on the diagonal of $\boldsymbol{\Lambda}_\mathbf{n}$. Denote $\mathbf{v}_n \triangleq \mathbf{U}_n^\top \boldsymbol{\theta}_n^\star$. Then the joint*

distribution of $(\sqrt{n}\boldsymbol{\theta}^\star_{n,i}, \lambda_{n,i}, \sqrt{n}\mathbf{v}_{n,i})$, $i \in [n]$, *converges in Wasserstein-2 distance to a probability distribution* Π *in* $\mathbb{R} \times \mathbb{R}_+ \times \mathbb{R}$, *i.e.,*

$$\frac{1}{n}\sum_{i=1}^{n} \delta(\sqrt{n}\boldsymbol{\theta}^\star_{n,i}, \lambda_{n,i}, \sqrt{n}\mathbf{v}_{n,i}) \xrightarrow{W_2} \Pi.$$

The assumption on $\|\boldsymbol{\theta}^\star_n\|_2$ is without loss of generality for GLM since $\|\boldsymbol{\theta}^\star_n\|_2$ can be absorbed in the link function ψ. Similarly for GMM, if $\|\boldsymbol{\theta}^\star_n\|_2 \neq 1$, we can always assume normalized features $\mathbf{x}$, by appropriately scaling the covariance matrix $\boldsymbol{\Sigma}_n$. We remark that while the Gaussian distribution assumption on feature vectors is crucial for our theoretical analysis, our empirical results suggest that this assumption can be relaxed to include at least the family of sub-Gaussian data distributions. We discuss this universality property in Appendix 3.7.

3.2.2 Asymptotic Regime

We consider the high-dimensional asymptotic regime in which the size m of the training set and the dimension n of the feature space grow large at a proportional rate. Formally, $m, n \to \infty$ at a fixed ratio $\delta = m/n$.

3.2.3 Robust Learning

Let $\widehat{\boldsymbol{\theta}}_n$ be a linear classifier trained on data generated according to either models (3.2) or (3.3). As is typical, given $\widehat{\boldsymbol{\theta}}_n$, a decision is made about the label of $\mathbf{x}$ based on $\text{sign}(\langle \mathbf{x}, \widehat{\boldsymbol{\theta}}_n \rangle)$. Thus, letting y be the label of a fresh sample $\mathbf{x}$, the *standard error* is given by

$$\mathcal{E}(\widehat{\boldsymbol{\theta}}_n) \triangleq \mathbb{E}_{\mathbf{x},y}\left[\mathbf{1}_{\left\{y \neq \text{sign}\left(\langle \mathbf{x}, \widehat{\boldsymbol{\theta}}_n \rangle\right)\right\}}\right]. \tag{3.4}$$

Here, the expectation is over a fresh pair $(\mathbf{x}, y)$ also generated according to either the GLM or the GMM model. Next, we define the adversarial error with respect to a worst-case ℓ_q-norm bounded additive perturbation. Let $\varepsilon_{\text{ts}} \geq 0$ be the budget of the adversary. Then, the *adversarial error* is defined as follows:

$$\mathcal{E}_{\ell_q,\varepsilon_{\text{ts}}}(\widehat{\boldsymbol{\theta}}_n) \triangleq \mathbb{E}_{\mathbf{x},y}\left[\max_{\|\boldsymbol{\delta}\|_q \leq \varepsilon_{\text{ts}}} \mathbf{1}_{\left\{y \neq \text{sign}\left(\langle \mathbf{x}+\boldsymbol{\delta}, \widehat{\boldsymbol{\theta}}_n\rangle\right)\right\}}\right]. \tag{3.5}$$

Adversarial training leads to a classifier $\widehat{\boldsymbol{\theta}}_n$ that solves the robust optimization problem (3.1) with $\widetilde{\mathcal{L}}(y, f_{\boldsymbol{\theta}}(\mathbf{x}+\boldsymbol{\delta}))$ replaced by $\mathcal{L}(y\langle \boldsymbol{\theta}, \mathbf{x}+\boldsymbol{\delta}\rangle)$. The loss function $\mathcal{L} : \mathbb{R} \to [0, \infty)$ is chosen as a convex approximation to the $0/1$ loss. Specifically, throughout the chapter, we assume that $\mathcal{L}$ is convex and decreasing. This includes popular choices such as the logistic, hinge and exponential losses.

3.3 Main Results for ℓ_∞ Perturbations

3.3.1 Asymptotic Behavior

In this section, we focus on the case of bounded ℓ_∞-perturbations, i.e. the adversarial error in (3.5) is considered for $q = \infty$. Specifically, let $\widehat{\boldsymbol{\theta}}_n$ be a solution to the following robust minimization:

$$\min_{\boldsymbol{\theta}_n} \sum_{i=1}^{m} \max_{\|\boldsymbol{\delta}_i\|_\infty \leq \frac{\varepsilon_{\text{tr}}}{\sqrt{n}}} \mathcal{L}\left(y_i \left\langle \mathbf{x}_i + \boldsymbol{\delta}_i, \boldsymbol{\theta}_n \right\rangle\right) + \lambda \|\boldsymbol{\theta}_n\|_2^2. \tag{3.6}$$

In our asymptotic setting, ε_{tr} is of constant order and the factor $1/\sqrt{n}$ in front of it is the proper normalization needed to ensure that the perturbations norm $\|\boldsymbol{\delta}_i\|_2$, is comparable to the norm of the true vector $\|\boldsymbol{\theta}_n^\star\|_2$, i.e., both are constant in the high-dimensional limit $\to n$. We explain this normalization further in Section 3.3.3. Here, we

consider the case of diagonal covariance matrix (i.e., $\mathbf{\Sigma}_n = \mathbf{\Lambda}_n$). Note that this assumption can be made without loss of generality for GMM data. Indeed, instead of features $\mathbf{x}_i$ as in (3.2) and mean vector $\boldsymbol{\theta}_n^\star$, we can equivalently analyze features $\widetilde{\mathbf{x}}_i = \mathbf{U}_n^\top \mathbf{x}_i$ and mean vector $\widetilde{\boldsymbol{\theta}_n^\star} := \mathbf{U}_n^\top \boldsymbol{\theta}_n^\star$. For the GLM data in (3.3), we defer the general case of possibly nondiagonal $\mathbf{\Sigma}_n$ to Appendix 3.8.1 where we also discuss how final expressions simplify in the case of isotropic features.

Before presenting our main result, we need to introduce some necessary definitions. We write

$$\mathcal{M}_f(x;\kappa) \triangleq \min_v \frac{1}{2\kappa}(x-v)^2 + f(v), \tag{3.7}$$

for the Moreau envelope of a function $f : \mathbb{R} \to \mathbb{R}$ at $x \in \mathbb{R}$ with parameter $\kappa > 0$ [2]. We also define the following min-max optimization over eight scalar variables. Denote $\bar{\mathbf{v}} \triangleq (\alpha, \tau_1, w, \mu, \tau_2, \beta, \gamma, \eta)$ and define $f : \mathbb{R}^8 \to \mathbb{R}$ as follows:

$$f(\bar{\mathbf{v}}) \triangleq -\gamma w - \frac{\mu^2 \tau_2}{2\alpha} C^2 - \frac{\alpha\beta^2}{2\delta\tau_2} - \frac{\alpha\tau_2}{2} + \frac{\beta\tau_1}{2} + \eta\mu - \frac{\eta^2\alpha}{2\tau_2 C^2},$$

where $C = \widetilde{\zeta}$ and ζ (defined in Assumption 3.2.2) for GMM and GLM, respectively. We introduce the following min-max objective based on the eight scalars,

$$\begin{aligned} &\min_{\substack{\alpha,\tau_1,w\in\mathbb{R}_+,\\ \mu\in\mathbb{R}}} \max_{\substack{\tau_2,\beta,\gamma\in\mathbb{R}_+,\\ \eta\in\mathbb{R}}} f(\bar{\mathbf{v}}) + \mathbb{E}\left[\mathcal{M}_{\mathcal{L}}\left(Z_{\alpha,\mu} - w; \frac{\tau_1}{\beta}\right)\right] \\ &+ \varepsilon_{\mathrm{tr}}\gamma\, \mathbb{E}\left[\mathcal{M}_{\ell_1 + \frac{r}{\varepsilon_{\mathrm{tr}}\gamma}\ell_2^2}\left(\frac{\alpha\beta}{\tau_2\sqrt{\delta L}} H + \frac{\alpha\eta}{\tau_2 D}\, T; \frac{\alpha\varepsilon_{\mathrm{tr}}\gamma}{\tau_2 L}\right)\right], \end{aligned} \tag{3.8}$$

where $D \triangleq \widetilde{\zeta}^2 L$ and ζ^2 for models (3.2) and (3.3), respectively, $H \sim \mathcal{N}(0,1)$ and

$(T, L, V) \sim \Pi$ where Π was defined in Assumption 3.2.3. We also let for convenience

$$Z_{\alpha,\mu} \triangleq \begin{cases} \sqrt{\alpha^2 + \mu^2 \widetilde{\zeta^2}}\, G + \mu \widetilde{\zeta^2} & \text{for GMM,} \\ \alpha G + \mu \zeta S \cdot \psi(\zeta S) & \text{for GLM,} \end{cases} . \tag{3.9}$$

where $G, S \overset{\text{iid}}{\sim} \mathcal{N}(0,1)$. Notice that the objective function of (3.8) depends explicitly on the sampling ratio δ and on the training parameter ε_{tr}. Moreover, it depends implicitly on $\boldsymbol{\theta}_n^\star$ and $\boldsymbol{\Lambda}_n$ via T and L, respectively, and on the specific loss $\mathcal{L}$ via its Moreau envelope. The nature of allowed perturbations (the ℓ_∞-type) is reflected in (3.8), via the Moreau-envelope of the dual-norm (the ℓ_1 norm).

We are now ready to state our main result in Theorem 3.3.1, which establishes a relation between the solutions of (3.8) and the adversarial risk of the robust classifier $\widehat{\boldsymbol{\theta}}_n$. The proof is deferred to Appendices 3.8.1 and 3.8.1.

Theorem 3.3.1. *Assume that the training dataset $\{(\mathbf{x}_i, y_i)\}_{i=1}^m$, is generated according to either (3.2) or (3.3) with diagonal covariance matrices satisfying Assumptions 3.2.1-3.2.3. Consider the robust classifiers $\{\widehat{\boldsymbol{\theta}}_n\}$, obtained by adversarial training in (3.6). Then, the high-dimensional limit for the adversarial error $\mathcal{E}_{\ell_\infty, \frac{\varepsilon_{\text{ts}}}{\sqrt{n}}}(\widehat{\boldsymbol{\theta}}_n)$, converges to,*

$$\begin{cases} Q\left(\frac{\mu^\star \widetilde{\zeta^2} - w^\star\, \varepsilon_{\text{ts}}/\varepsilon_{\text{tr}}}{\sqrt{\mu^{\star 2}\widetilde{\zeta^2} + \alpha^{\star 2}}}\right) & \textit{for GMM,} \\ \mathbb{P}\left(\mu^\star \zeta\, S\, \psi(\zeta S) + \alpha^\star G < \frac{w^\star \varepsilon_{\text{ts}}}{\varepsilon_{\text{tr}}}\right) & \textit{for GLM,} \end{cases} \tag{3.10}$$

where $Q(\cdot)$ denotes the Gaussian Q-function and $(\alpha^\star, \mu^\star, w^\star)$ is the unique solution to the scalar minimax problem (3.8).

The asymptotics for adversarial error in Theorem 3.3.1 are precise in the sense that they hold with probability 1, as $m, n \to \infty$. In the following section, we demonstrate the

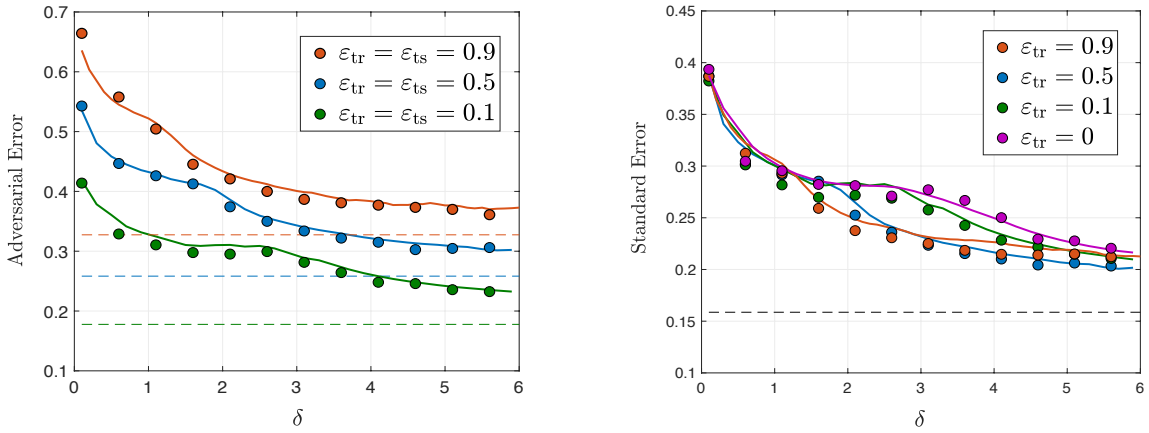
precise theoretical values and the corresponding numerical values.

3.3.2 Numerical Illustrations

In this section, we illustrate the theoretical predictions for various values of the different problem parameters, including $\delta = m/n$ and the attack budgets $\varepsilon_{\rm tr}$ and $\varepsilon_{\rm ts}$. For numerical results here, we focus on the hinge-loss i.e., $\mathcal{L}(t) = \max(1-t, 0)$ and on the GMM with isotropic features; thus L has a unit mass at 1. Additional experiments on GLM are given in Appendix 3.7. We further assume that T is standard normal and fix regularization parameter $\lambda = 10^{-4}$. To solve (3.8), we derive the solution of the corresponding saddle-point equations (derived in Eq. (3.60) in Appendix 3.8.1) by iterating over the equations and finding the fixed-point solution after 100 iterations. For the numerical results, we set $n = 200$ and solve the ERM problem (3.6) by gradient descent. The resulting estimator is used to compute the adversarial test error by evaluating (3.4) on a test set of 3×10^3 samples. We then average the results over 20 independent experiments. The results for both numerical and theoretical values are depicted in Figures 3.1-3.2. Next, we discuss some of the insights obtained from these figures.

Impact of δ on standard/adversarial errors. Figure 3.1 depicts the adversarial and standard errors as a function of $\delta = m/n$. We compare the results of adversarial training with the Bayes optimal error. Formally, the Bayes Adversarial Error is defined as

$$\mathcal{E}_{\ell_q,\varepsilon_{\rm ts}}(\mathrm{OPT}) \triangleq \min_{f_{\boldsymbol{\theta}}} \ \mathbb{E}_{\mathbf{x},y}\left[\max_{\|\boldsymbol{\delta}\|_q \le \varepsilon_{\rm ts}} \mathbf{1}_{\left\{y \neq f_{\boldsymbol{\theta}}(\mathbf{x}+\boldsymbol{\delta})\right\}}\right]. \tag{3.11}$$



(a) (b)

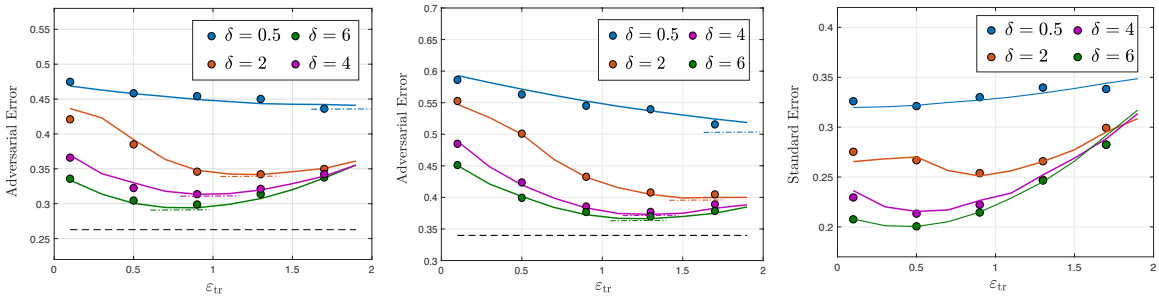
Figure 3.1: Adversarial/Standard test error based on $\delta := m/n$. Solid lines correspond to theoretical predictions while markers denote the empirical results derived by solving ERM with vanishing regularization($r = 10^{-4}$) using gradient descent. The dashed lines denote the Bayes adversarial error (left) and the Bayes standard error (right). Note that the adversarial error of estimators obtained from adversarial training, approaches the Bayes adversarial error as δ grows.

For the Gaussian-mixture model (3.2) under an ℓ_q attack with budget ε, the Bayes adversarial error is derived as follows[83]:

$$\mathcal{E}_{\ell_q,\varepsilon}(\mathrm{OPT}) = Q(\|\boldsymbol{\theta}^\star - \boldsymbol{\mu}^\star\|_{\boldsymbol{\Sigma}_n^{-1}}), \tag{3.12}$$
$$\text{where } \boldsymbol{\mu}^\star \triangleq \arg\min_{\|\boldsymbol{\mu}\|_q \le \varepsilon} \|\boldsymbol{\theta}^\star - \boldsymbol{\mu}\|^2_{\boldsymbol{\Sigma}_n^{-1}}.$$

The dashed lines in Figure 3.1 show the *Bayes Adversarial Error*, derived according to (3.12) for $\varepsilon = \varepsilon_{\rm ts}/\sqrt{n}$.

Note that both errors decrease as the sampling ratio δ grows, with the adversarial error approaching the Bayes adversarial error of the corresponding value of $\varepsilon_{\rm ts}$. In Appendix 3.8.3, we formally prove that for ℓ_2 attacks bounded by $\varepsilon_{\rm ts} \in [0, 1]$, the robust error achieved by adversarial training with any $\varepsilon_{\rm tr} \in [0, 1]$ converges to the Bayes adversarial error in the infinite sample-size limit i.e., when $\delta \to \infty$. More generally, in light of



(a) (b) (c)

Figure 3.2: Theoretical (solid lines) and Empirical (markers) results for the impact of adversarial training on the adversarial test error for $\varepsilon_{\rm ts} = 0.5$ (Left) and $\varepsilon_{\rm ts} = 0.9$ (Middle). The blacked dashed lines denote the Bayes adversarial error for the corresponding values of $\varepsilon_{\rm ts}$. The colored dashed lines depict the optimal value of each curve. Note that the optimal value of $\varepsilon_{\rm tr}$ decreases as δ grows. Right: Impact of adversarial training on the standard test error, illustrating that adversarial training can improve standard accuracy.

comparison between the error formulae of Theorem 3.3.1 and the Bayes adversarial error, Figure 3.1 provides a means to quantify the sub-optimality gap of adversarial training for all values of the oversampling ratio $\delta > 0$ and for different values of the adversary's budget. A related study was performed in [99], but therein the authors derive error bounds for a simple averaging estimator. Instead, our analysis is precise and holds for the broader case of convex decreasing losses. Next, we comment on the shape of the error curves as a function of the sampling ratio. Note that a second sharp decrease in standard and adversarial errors appears right after an separability threshold $\delta_{\frac{\varepsilon_{\rm tr}}{\sqrt{n}},\Pi}$, which we define as the maximum value of δ for which the data-points are $(\ell_\infty, \frac{\varepsilon_{\rm tr}}{\sqrt{n}})$-separable (for definition, see the discussion on Robust Separability in Section 3.5). This constantly decreasing behavior of the error is in contrast to the corresponding behavior in linear regression with ℓ_2 perturbations and ℓ_2 loss analyzed in [78], where error based on δ starts rising after the first decrease and then again decreases as δ grows. This double-descent behavior can be considered as extension of the more familiar double-descent behavior in standard ERM (first observed in numerous high-dimensional machine learning models [100, 11, 101]),

to the adversarial training case. Finally, we highlight the following observation from Figure 3.1a: For highly over-parametrized models (very small δ), standard accuracy remains the same for different choices of ε_{tr}. As δ grows, adversarial training (perhaps surprisingly) seems to (also) improve the standard accuracy. However, for very large δ, increasing ε_{tr} hurts standard accuracy. These observations are consistent and theoretically validate corresponding findings on the role of data-set size on standard accuracy that were empirically observed in [89] for neural network training with non-synthetic datasets such as MNIST.

Impact of ε_{tr} on standard/adversarial errors. Adversarial and Standard error curves based on the hyper-parameter ε_{tr} are illustrated in Figure 3.2. Note that the adversarial error behavior based on ε_{tr} is informative about the role of the data-set size on the optimal value of ε_{tr}. Figures 3.2a-3.2b show that the optimal value of ε_{tr} is typically larger than ε_{ts}. Also note that as δ gets smaller, larger values of ε_{tr} are preferable for robustness. As detailed in Appendix 3.7, this behavior is also observed in real-world experiments with the MNIST dataset. Figure 3.2c illustrates the impact of ε_{tr} on the standard error, where similar to Figure 3.1b, we observe that adversarial training can help standard accuracy. In particular, we observe that in the under-parameterized regime where $\delta > \delta_{\frac{\varepsilon_{\text{tr}}}{\sqrt{n}},\Pi}$ (as we will define in Section 3.5), adversarial training with small values of ε_{tr} is beneficial for accuracy. As δ increases, such gains diminish and indeed adversarial training starts hurting standard accuracy.

3.3.3 Proof Sketch

The complete proof of Theorem 3.3.1 is deferred to the appendix. Here, we provide an outline of the key steps in deriving (3.8) and (3.10).

Reducing (3.6) to a minimization problem. For a decreasing loss function, picking $\boldsymbol{\delta}_i^\star \triangleq -y_i \operatorname{sign}(\boldsymbol{\theta}_n)\, \varepsilon_{\text{tr}}/\sqrt{n}$, optimizes the inner maximization in (3.6). Therefore, (3.6) is equivalent to,

$$\min_{\boldsymbol{\theta}_n} \sum_{i=1}^{m} \mathcal{L}\Big(y_i \left\langle \mathbf{x}_i, \boldsymbol{\theta}_n \right\rangle - \frac{\varepsilon_{\text{tr}}}{\sqrt{n}} \|\boldsymbol{\theta}_n\|_1\Big) + \lambda \|\boldsymbol{\theta}_n\|_2^2. \tag{3.13}$$

From (3.13), we can see now why the specific normalization of ε_{tr} is needed in (3.6). Recall that (for model (3.3), for instance), $\mathbf{x}_i \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}_n)$ and $\|\boldsymbol{\theta}_n^\star\|_2 \xrightarrow{P} 1$. For simplicity assume here that $\boldsymbol{\Sigma}_n = \mathbb{I}_n$. For fixed $\boldsymbol{\theta}$, the argument $y_i \langle \mathbf{x}_i, \boldsymbol{\theta} \rangle$ behaves as $\|\boldsymbol{\theta}\|_2 S\psi(S)$, where $S \sim \mathcal{N}(0,1)$. Thus, for $\boldsymbol{\theta}$s that are such that $\|\boldsymbol{\theta}\|_2 = \Theta(1)$ (which ought to be the case for "good" classifiers in view of $\|\boldsymbol{\theta}_n^\star\|_2 = 1$), the term $y_i \langle \mathbf{x}_i, \boldsymbol{\theta} \rangle$ is an $\Theta(1)$-term. Now, thanks to the normalization $1/\sqrt{n}$ in (3.6), the second term $\frac{\varepsilon_{\text{tr}}}{\sqrt{n}} \|\boldsymbol{\theta}\|_1$ in (3.13) is also of the same order. Here, we used again the intuition that $\|\boldsymbol{\theta}\|_1 = \Theta(\sqrt{n})$, as is the case for the true $\boldsymbol{\theta}^\star$. Our analysis formalizes these heuristic explanations.

The key statistics for the adversarial error. Our key observation is that the asymptotics of the adversarial error of a sequence of arbitrary classifiers $\{\boldsymbol{\theta}_n\}$, depend on the asymptotics of only a few key statistics of $\{\boldsymbol{\theta}_n\}$. This is formalized in the following lemma. The proof is deferred to Appendix 3.8.1. Similar to before, there is nothing special here to $q = \infty$, so we state this result for general q.

Lemma 3.3.1. *Fix $q \geq 1$ and let ℓ_p denote the dual norm of ℓ_q. Let $\widetilde{\boldsymbol{\theta}_n^\star} \triangleq \boldsymbol{\Sigma}_n^{1/2} \boldsymbol{\theta}_n^\star$ for data model (3.3) and $\widetilde{\boldsymbol{\theta}_n^\star} \triangleq \boldsymbol{\Sigma}_n^{-1/2} \boldsymbol{\theta}_n^\star$ for data model (3.2). Further, for both models, define projection matrices Θ_n and $\Theta_n^\perp$ as follows, $\Theta_n \triangleq \widetilde{\boldsymbol{\theta}_n^\star} \widetilde{\boldsymbol{\theta}_n^\star}^\top / \|\widetilde{\boldsymbol{\theta}_n^\star}\|_2^2$, $\Theta_n^\perp \triangleq \mathbb{I}_n - \Theta_n$. Further, let ε and ε' (possibly scaling with the problem dimensions) be the upper-bounds on norm of the adversarial perturbation during training and test time, respectively. With this notation, assume that the sequence of $\{\boldsymbol{\theta}_n\}$ is such that the following limits are true for the statistics*

$\|\boldsymbol{\theta}_n\|_p$, $\|\Theta_n \boldsymbol{\Sigma}_n^{1/2} \boldsymbol{\theta}_n\|_2$ *and* $\|\Theta_n^{\perp} \boldsymbol{\Sigma}_n^{1/2} \boldsymbol{\theta}_n\|_2$,

$$\{\varepsilon \|\boldsymbol{\theta}_n\|_p\} \xrightarrow{P} w, \qquad \frac{1}{C}\{\|\Theta_n \boldsymbol{\Sigma}_n^{1/2} \boldsymbol{\theta}_n\|_2\} \xrightarrow{P} \mu,$$
$$\{\|\Theta_n^{\perp} \boldsymbol{\Sigma}_n^{1/2} \boldsymbol{\theta}_n\|_2\} \xrightarrow{P} \alpha,$$

where $C = \tilde{\zeta}, \zeta$ *for GMM and GLM, respectively. Then, in the high-dimensional limit, the adversarial error converges to,*

$$\begin{cases} Q\left(\frac{\mu\widetilde{\zeta}^2 - w\,\varepsilon'/\varepsilon}{\sqrt{\mu^2\widetilde{\zeta}^2 + \alpha^2}}\right) & \textit{for GMM,} \\ \mathbb{P}\Big(\mu\zeta S\,\psi(\zeta S) + \alpha G - w\varepsilon'/\varepsilon < 0\Big) & \textit{for GLM.} \end{cases} \tag{3.14}$$

The detailed proof of the lemma is deferred to the appendix. There are essentially two steps in establishing the result. The first is to exploit the decreasing nature of the $0/1$-loss to explicitly optimize over $\boldsymbol{\delta}_i$. This optimization gives rise to the dual norm $\|\boldsymbol{\theta}_n\|_p$. The second step is to consider the change of variables $\boldsymbol{\theta}_n \Rightarrow \widetilde{\boldsymbol{\theta}}_n \triangleq \boldsymbol{\Sigma}_n^{1/2} \boldsymbol{\theta}_n$ and decompose $\widetilde{\boldsymbol{\theta}}_n$ on its projection on $\boldsymbol{\Sigma}_n^{1/2} \boldsymbol{\theta}_n^{\star}$ and its complement. In the notation of the lemma, $\widetilde{\boldsymbol{\theta}}_n = \Theta_n \widetilde{\boldsymbol{\theta}}_n + \Theta_n^{\perp} \widetilde{\boldsymbol{\theta}}_n$. The Gaussianity of the feature vectors together with orthogonality of the two components in the decomposition of $\boldsymbol{\theta}_n$ explain the appearance of the Gaussian variables S and G in (3.14). When applied to ℓ_∞-perturbations, Lemma 3.3.1 reduces the goal of computing asymptotics of the adversarial risk of $\widehat{\boldsymbol{\theta}}_n$ to computing asymptotics of the corresponding statistics $\|\boldsymbol{\Sigma}_n^{-1/2} \widetilde{\boldsymbol{\theta}}_n\|_1$, $\|\Theta_n \widetilde{\boldsymbol{\theta}}_n\|_2$, and $\|\Theta_n^{\perp} \widetilde{\boldsymbol{\theta}}_n\|_2$.

Scalarizing the objective function. The previous two steps set the stage for the core of the analysis, which we outline next. Thanks to step 1, we are now asked to analyze the statistical properties of a convex optimization problem. On top of that, due to step 2, the outcomes of the analysis ought to be asymptotic predictions for the

quantities $\|\boldsymbol{\Sigma}_n^{-1/2}\widetilde{\boldsymbol{\theta}}_n\|_1$, $\|\Theta_n\widetilde{\boldsymbol{\theta}}_n\|_2$ and $\|\Theta_n^{\perp}\widetilde{\boldsymbol{\theta}}_n\|_2$. However, note that the term $\|\boldsymbol{\Sigma}_n^{-1/2}\widetilde{\boldsymbol{\theta}}_n\|_1$ appears inside the loss function. In particular, this is a new challenge, specific to robust optimization compared to previous analysis of standard regularized ERM. Moreover, both of the terms $\|\boldsymbol{\Sigma}_n^{-1/2}\widetilde{\boldsymbol{\theta}}_n\|_1$ and $\|\boldsymbol{\Sigma}_n^{-1/2}\widetilde{\boldsymbol{\theta}}_n\|_2^2$ are not decomposable based on $\|\Theta_n\widetilde{\boldsymbol{\theta}}_n\|_2$ and $\|\Theta_n^{\perp}\widetilde{\boldsymbol{\theta}}_n\|_2$, due to the presence of the term $\boldsymbol{\Sigma}_n^{-1/2}$. The first step to overcome these challenges is to identify an appropriate minimax Auxiliary Optimization (AO) problem that is probabilistically equivalent to (3.13). The second crucial step is to scalarize the AO based on an appropriate Lagrangian formulation. Finally, we perform a probabilistic analysis of the scalar AO. This results in the deterministic minimax problem in (3.8). See the appendix for details.

3.4 Main Results for ℓ_2 Perturbations

When $q = 2$, the min-max problem is equivalent to the following, by choosing the optimal choice $\boldsymbol{\delta}_i = -y_i\varepsilon_{\text{tr}}\boldsymbol{\theta}/\|\boldsymbol{\theta}\|_2$,

$$\min_{\boldsymbol{\theta}_n} \frac{1}{m}\sum_{i=1}^{m} \mathcal{L}\left(y_i \left\langle \mathbf{x}_i, \boldsymbol{\theta}_n\right\rangle - \varepsilon_{\text{tr}}\|\boldsymbol{\theta}_n\|_2\right) + \lambda \|\boldsymbol{\theta}_n\|_2^2. \tag{3.15}$$

Here, we assume $\{\boldsymbol{\Sigma}_n\}$ to be a sequence of positive definite matrices. Denote $\widetilde{\mathbf{v}} \triangleq (\alpha, \tau_1, \tau_3, w, \mu, \tau_2, \beta, \gamma, \eta)$ and define $g : \mathbb{R}^9 \to \mathbb{R}$ as follows,

$$\begin{aligned} g(\widetilde{\mathbf{v}}) \triangleq & -\gamma w - \frac{\mu^2\tau_2}{2\alpha}C^2 - \frac{\alpha\beta^2}{2\delta\tau_2} - \frac{\alpha\tau_2}{2} + \frac{\beta\tau_1}{2} \\ & + \eta\mu - \frac{\eta^2\alpha}{2\tau_2 C^2} + \frac{\varepsilon_{\text{tr}}\gamma\tau_3}{2}, \end{aligned}$$

where recall that $C \triangleq \widetilde{\zeta}$ and ζ for GMM and GLM, respectively. With this notation, we introduce the following min-max problem,

$$\min_{\substack{\alpha,\tau_1,\tau_3,w\in\mathbb{R}_+\\ \mu\in\mathbb{R}}} \max_{\substack{\tau_2,\beta,\gamma\in\mathbb{R}_+,\\ \eta\in\mathbb{R}}} g(\widetilde{\mathbf{v}}) + \mathbb{E}\left[\mathcal{M}_{\mathcal{L}}\left(Z_{\alpha,\mu} - w; \frac{\tau_1}{\beta}\right)\right] + \frac{\eta^2\alpha^2}{\tau_2^2 C^4}\left(\frac{\varepsilon_{\rm tr}\gamma}{2\tau_3} + r\right)\mathbb{E}_L\left[\frac{\frac{C^4\beta^2}{\eta^2\delta} + \widetilde{L}}{\frac{\varepsilon_{\rm tr}\gamma\alpha + 2\tau_3 r\alpha}{\tau_2\tau_3} + L}\right], \tag{3.16}$$

where we define $\widetilde{L} \triangleq 1/L$ and L for GMM and GLM, respectively and the random variables L and $Z_{\alpha,\mu}$ are defined same as in (3.8).

Theorem 3.4.1. *Consider the same setting as in Theorem 3.3.1, only here assume that $q = 2$ and $\{\mathbf{\Sigma}_n\}$ are positive definite matrices(not necessarily diagonal) satisfying Assumptions 3.2.1-3.2.3. Let $(\alpha^\star, \mu^\star, w^\star)$ be the unique solution to the minimax problem (3.16). Then, the high-dimensional limit for the adversarial error $(\mathcal{E}_{\ell_2,\varepsilon_{\rm ts}}(\widehat{\boldsymbol{\theta}}_n))$ converges to*

$$\begin{cases} Q\left(\frac{\mu^\star\widetilde{\zeta}^2 - w^\star\,\varepsilon_{\rm ts}/\varepsilon_{\rm tr}}{\sqrt{\mu^{\star 2}\widetilde{\zeta}^2 + \alpha^{\star 2}}}\right) & \text{for GMM,} \\ \mathbb{P}\Big(\mu^\star\zeta S\psi(\zeta S) + \alpha^\star G < w^\star\frac{\varepsilon_{\rm ts}}{\varepsilon_{\rm tr}}\Big) & \text{for GLM.} \end{cases} \tag{3.17}$$

Proof of Theorem 3.4.1 is deferred to Appendix 3.8.2. Compared to Theorem 3.3.1, note here that the asymptotic prediction only depends on the total energy of $\boldsymbol{\theta}_n^\star$(which was assumed to be 1 in Assumption 3.2.2) and not on its empirical distribution T. We present numerical illustrations on ℓ_2-attacks in Appendix 3.7, where we also discuss how the data-set size and attack budgets, affect the adversarial and standard test errors based on Theorem 3.4.1.

3.5 Further Discussions

Remark 3.5.1 (Training with no Regularization and *Robust Separability*). An instance of special interest in practice is solving the *unregularized* version of the min-max problem:

$$\min_{\boldsymbol{\theta}_n} \frac{1}{m} \sum_{i=1}^{m} \max_{\|\boldsymbol{\delta}_i\|_q \leq \varepsilon} \mathcal{L}\left(y_i \left\langle \mathbf{x}_i + \boldsymbol{\delta}_i, \boldsymbol{\theta}_n \right\rangle\right). \tag{3.18}$$

Following the same proof techniques as above, we can show that the formulas predicting the statistical behavior of this unconstrained version are given by the same formulas as in Theorem 3.3.1 with $r = 0$ and also provided that the sampling ration δ is large enough so that a certain robust separability condition holds. In what follows, we describe this condition. We start with some background on (standard) data separability. Recall, that training data $\{(\mathbf{x}_i, y_i)\}$ are linearly separable if and only if $\exists \boldsymbol{\theta} \in \mathbb{R}^n$ such that for all training samples $y_i \langle \mathbf{x}_i, \boldsymbol{\theta} \rangle \geq 1$. Now, we say that data are (ℓ_q, ε)-separable if and only if

$$\exists \boldsymbol{\theta} \in \mathbb{R}^n \quad \text{s.t.} \quad y_i \langle \mathbf{x}_i, \boldsymbol{\theta} \rangle - \varepsilon \|\boldsymbol{\theta}\|_p \geq 1, \ \forall i \in [m].$$

Note that (standard) linear separability is equivalent to $(\ell_q, 0)$-separability as defined above. Moreover, it is clear that (ℓ_q, ε)-separability implies $(\ell_q, 0)$-separability for any $\varepsilon \geq 0$. Recent works have shown that in the proportional limit data from the GLM are $(\ell_q, 0)$-separable if and only if the sampling ratio satisfies $\delta < \delta_\psi$ [1, 48, 55, 102] for some $\delta_\psi > 2$. Here, the subscript ψ denotes dependence of the phase-transition threshold δ_ψ on the link function ψ of the GLM. We conjecture that there is a threshold $\delta_{\psi,\varepsilon,\Pi}$, depending on ε, the link function ψ and the probability distribution Π such that data are (ℓ_q, ε)-separable if and only if $\delta < \delta_{\psi,\varepsilon,\Pi}$. We believe that our techniques can be used to prove this conjecture and determine $\delta_{\psi,\varepsilon,\Pi}$, but we leave this interesting question to future work. Instead here, we simply note that based on the above discussion, if such a

threshold exists, then it must satisfy $\delta_{\psi,\varepsilon,\Pi} \leq \delta_{\psi,0,\Pi}$, for all values of ε, and in fact it is a decreasing function of ε. Now let us see how this notion relates to solving (3.6) and to our asymptotic characterization of its performance. Recall from (3.13) that the robust ERM for decreasing losses reduces to the minimization $\min_{\boldsymbol{\theta}} \sum_{i=1}^{m} \mathcal{L}(y_i\langle \mathbf{x}_i, \boldsymbol{\theta}\rangle - \varepsilon\|\boldsymbol{\theta}\|_p)$. Thus, using again the decreasing nature of the loss, it can be checked that the solution to the objective function above becomes unbounded for $\boldsymbol{\theta}$ such that the argument of the loss is positive for any $i \in [m]$. This is equivalent to the condition of (ℓ_q, ε)-separability. In other words, when data are (ℓ_q, ε)-separable, the robust estimator is unbounded. Recall from Section 3.3.3 that the minimax optimization variables w, μ, α represent the limits of $\|\widehat{\boldsymbol{\theta}}_n\|_p$, $\|\Theta_n \boldsymbol{\Sigma}_n^{1/2} \widehat{\boldsymbol{\theta}}_n\|_2$, and $\|\Theta_n^{\perp} \boldsymbol{\Sigma}_n^{1/2} \widehat{\boldsymbol{\theta}}_n\|_2$. Thus, if $\widehat{\boldsymbol{\theta}}_n$ is unbounded, then $w^\star, \mu^\star, \alpha^\star$ are not well defined. In accordance with this, we conjecture that the minimax problem (3.8) for $r = 0$ (corresponding to (3.18)) has a solution if and only if the data are *not* (ℓ_q, ε)-separable, equivalently, iff $\delta > \delta_{\psi,\varepsilon,\Pi}$. Equivalent results are applicable to the Gaussian-Mixture models.

Remark 3.5.2 (On Statistical Limits in Adversarial Training). The asymptotics in (3.17) imply that for ℓ_2 perturbations and isotropic features, since $w^\star = \sqrt{\alpha^{\star 2} + \mu^{\star 2}}$, the errors depend on the ratio $\alpha^\star/\mu^\star$. In fact, it can be seen that smaller values of the ratio lead to smaller adversarial error. This leads to an interesting conclusion: *In order to find the hyper-parameter ε_{tr} that minimizes the adversarial error, it suffices to tune ε_{tr} to minimize the ratio $\alpha^\star/\mu^\star$.* A similar conclusion can be made for the case of ℓ_∞ perturbations, by noting from (3.10) that the adversarial error is characterized in a closed form in terms of $(\alpha^\star, \mu^\star, w^\star)$. In view of these observations, our sharp guarantees for the performance of adversarial training open the way to answering questions on the statistical limits and optimality of adversarial training, e.g. *how to optimally tune ε_{tr}? How to optimally choose the loss function and what is the best minimum values of adversarial error achieved by the family of robust estimators? How do these answers depend on the adversary budget*

and/or the sampling ratio δ? Fundamental questions of this nature have been recently addressed in the non-adversarial case based on the corresponding saddle-point equations for standard ERM, e.g., [58, 41, 49, 103, 53, 77]. Theorems 3.3.1 and 3.4.1 are the first steps towards such extensions to the adversarial settings.

3.6 Conclusions and Future Directions

We studied the generalization behavior of adversarial training in a binary classification setting. In particular, we derived precise theoretical predictions for the performance of adversarial training for the GLM and GMM. Numerical simulations validate theoretical predictions even for relatively small problem dimensions and demonstrate the role of all problem paramters on adversarial robustness. Finally, we remark that the current analysis can be extended to general convex regularization functions building on our ideas. An interesting future direction is analyzing adversarial training for Random Features [104] and Neural Tangent Kernel [105] models. One other natural question is considering attacks other than ℓ_q-norm attacks considered in this chapter.

3.7 Additional Numerical Experiments

Experiments on ℓ_2 attacks and GLM

In this section, we complement the numerical illustrations of Section 3.3.2, by considering the case of Signed measurements as well as extending to the ℓ_2-perturbations case. We focus on the Hinge-loss and for simulation results we set $n = 200, \lambda = 0, \mathbf{\Sigma}_n = \mathbb{I}_n$ and average the results over 20 experiments. Figures 3.3a-3.3b depict the adversarial/standard errors for the signed measurements. Notably, based on Figure 3.3a, one can observe that adversarial attacks are successful in GLM, as for a fixed δ, adversarial training does not seem to improve noticeably the adversarial error (the error bars are obtained by 10 experiments). However, note the critical role of data-set size on both standard and adversarial errors as depicted in Figure 3.3b. Similar to the GMM, here we also observe that both adversarial and standard errors are decreasing based on δ in both cases of $q = 2, \infty$.

Figures 3.3c-3.3d depict the error curves for the GMM and $q = 2$. Perhaps surprisingly, here we see that more aggressive adversarial training improves the standard error as the error curve is strictly decreasing with respect to ε_{tr}. We also highlight that unlike the $q = \infty$ case where there was a finite optimal choice of ε_{tr}, here increasing ε_{tr}, always helps the robust accuracy. Note also the role of δ on error curves, especially by increasing δ, both errors decrease and notably the adversarial error approaches the Bayes optimal error. For a formal proof of this phenomenon, see also the discussion in Appendix 3.8.3.

Universality in Adversarial Robustness

Thus far, we focused on Gaussian data. One may wonder whether our theoretical results extend to other data distributions. We conjecture that our results enjoy the

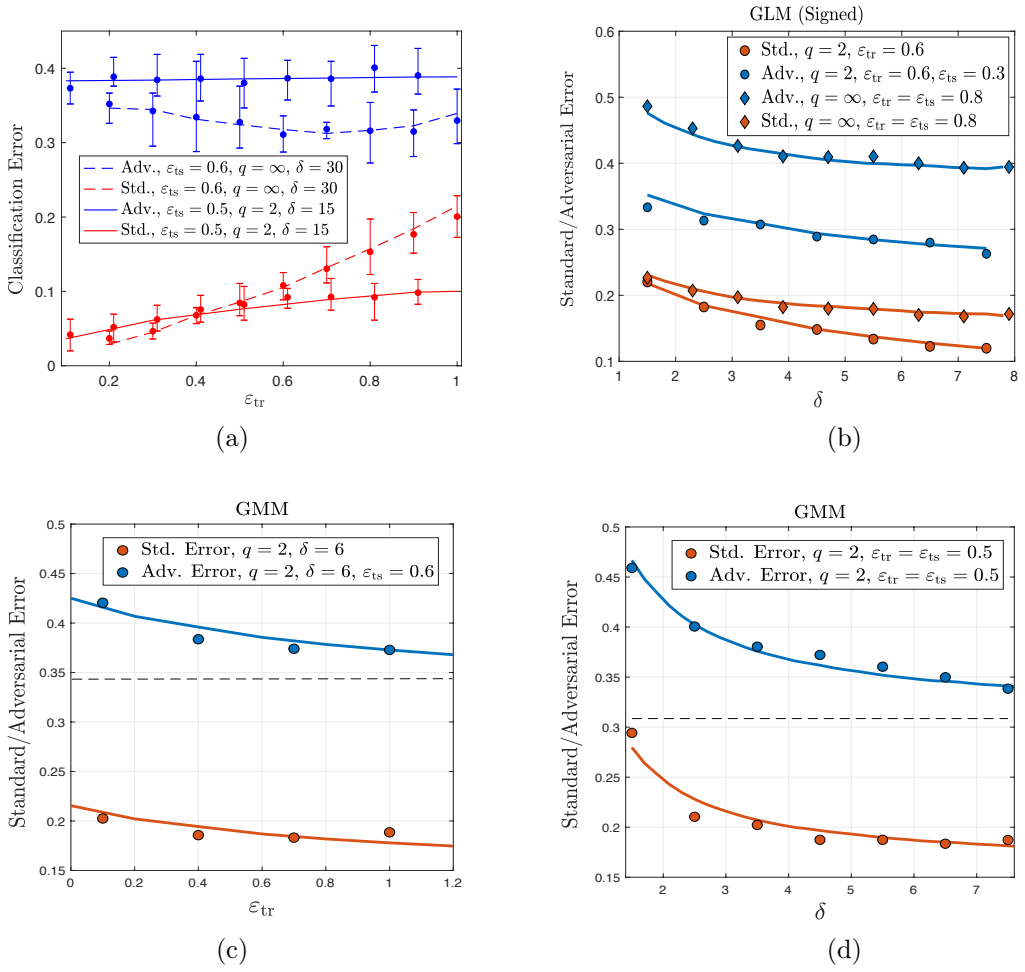


Figure 3.3: Adversarial and Standard Errors for the Signed model (Top) and the Gaussian-mixture model (Bottom). The dashed lines denote the Bayes adversarial error for the corresponding values of ε_{ts}.

universality property, i.e., the same asymptotic formulas in Theorem 3.3.1 and Theorem 3.4.1, hold when data is sampled from a *sub-Gaussian* distribution. Figure 3.4 illustrates the empirical results for the adversarial and standard error of Gaussian-mixture model as well as a model obtained by the mixture of Rademacher distributions, i.e.,

$$\mathbf{x}_i | y_i \;\sim\; y_i \boldsymbol{\theta}_n^\star + \boldsymbol{\rho}_i, \quad \mathbb{P}(y_i = 1) = \pi \in [0, 1],$$

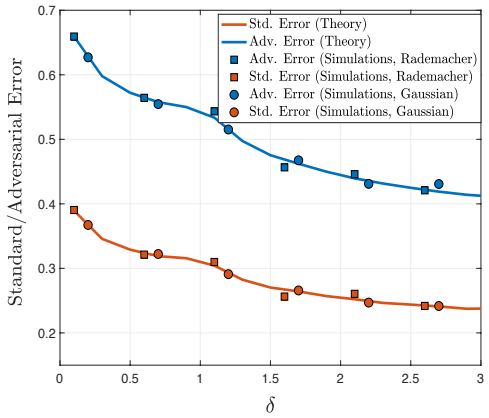
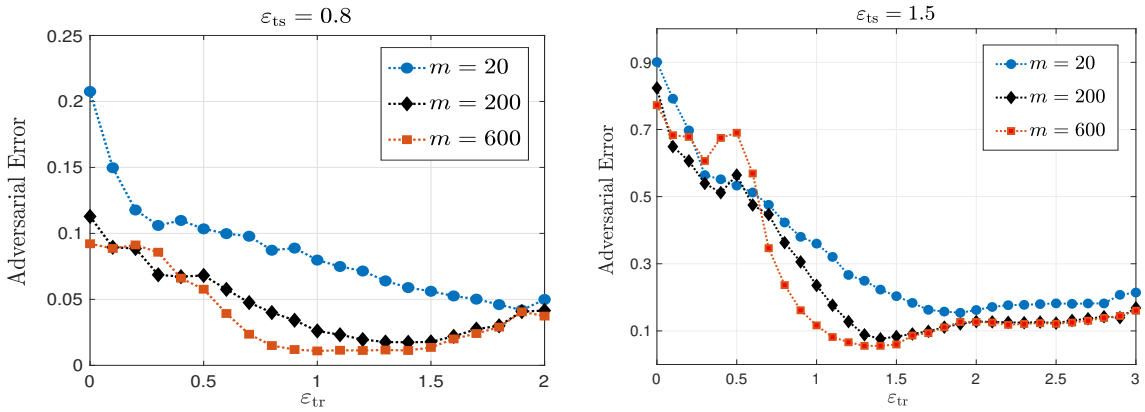


Figure 3.4: Empirical results for Rademacher(squares) and Gaussian(circles) data distributions in a generative data model alongside the theoretical curves. Here $q = \infty$ and $\varepsilon_{\text{tr}} = \varepsilon_{\text{ts}} = 1$. The perfect match between theory and experiments supports the conjectured universality property in adversarial training.

where each entry of $\boldsymbol{\rho}_i \in \mathbb{R}^n$ is distributed iid from Rademacher distribution. Note the perfect agreement between theory and simulation for both standard and adversarial errors, which supports the universality conjecture. For standard ERM, the universality property has been studied in numerous recent works e.g., see [106, 107, 108]. Extending such results to the adversarial training case is left for future work.

Experiments on MNIST

We present experimental results to demonstrate the adversarial error of linear classification beyond synthetic data. Specifically, we consider the ℓ_∞ robust classification with hinge-loss for the '0' and '1' digits of the MNIST dataset, which has a dimension of $n = 784$. Both classes' data points are shifted and normalized, so that $\overline{\mathbf{x}}_{(0)} = -\overline{\mathbf{x}}_{(1)}$ and $\|\overline{\mathbf{x}}_{(0)}\| = \|\overline{\mathbf{x}}_{(1)}\| = 1$, where $\overline{\mathbf{x}}_{(0)}, \overline{\mathbf{x}}_{(1)}$ are the empirical mean vectors of each class. Figure 3.5 displays the results for the adversarial error for two attack budgets at test time, denoted by ε_{ts}. Similar to the findings in Figure 3.2a-3.2b, we observe that increasing the sample size m leads to a decrease in the optimal ε_{tr} (which optimizes the adversarial error).



(a) (b)

Figure 3.5: Adversarial error in binary classification of digits '0' and '1' from the MNIST dataset, while using Hinge-loss and setting $q = \infty$. The results shown represent the averages from 10 experiments. Note that the optimal ε_{tr} decreases as the sample size increases.

Furthermore, we observe that for sufficiently large m, the optimal ε_{tr} is approximately equal to ε_{ts}.

3.8 Proofs

3.8.1 Proofs for Section 3.3

In this section, we provide an asymptotic analysis for adversarial training with ℓ_∞ perturbations. First, we consider the case of general $\mathbf{\Sigma}_n$ and then show how our theoretical results simplify when $\mathbf{\Sigma}_n$ is diagonal and when $\mathbf{\Sigma}_n = \mathbb{I}_n$. We focus on Generalized linear models (3.3). The corresponding analysis for Gaussian-Mixture models (3.2) is deferred to Section 3.8.3.

We begin with proving the key statistics required for the high-dimensional asymptotics.

Adversarial Error of an Arbitrary Estimator

In the following lemma, we characterize the asymptotic adversarial error under $\ell_q, q \geq 1$ perturbations of an arbitrary sequence of estimators $\{\boldsymbol{\theta}_n\}_{n=1}^{\infty}$ (where $\boldsymbol{\theta}_n \in \mathbb{R}^n$), in terms of the high-dimensional limits for the key statistics $\|\boldsymbol{\theta}_n\|_p$, $\|\Theta_n \boldsymbol{\Sigma}_n^{1/2} \boldsymbol{\theta}_n\|_2$ and $\|\Theta_n^{\perp} \boldsymbol{\Sigma}_n^{1/2} \boldsymbol{\theta}_n\|_2$, where p is such that $1/p + 1/q = 1$. We assume that the adversary has budget ε.

First, we formalize the adversarial test error in the next lemma, which is a restatement of Lemma 3.3.1 specialized to GLM.

Lemma 3.8.1. *The high-dimensional limit of the adversarial test error for the Generalized Linear models with a given sequence of classifiers $\{\boldsymbol{\theta}_n\}$ is given as follows,*

$$\{\mathcal{E}_{\ell_q,\varepsilon}^{GLM}(\boldsymbol{\theta}_n)\} \xrightarrow{P} \mathbb{P}\Big(\mu\zeta\, S \cdot \psi(\zeta S) + \alpha G - u\varepsilon < 0\Big) \tag{3.19}$$

where $G, S \overset{iid}{\sim} \mathcal{N}(0,1)$ and provided that

$$\Big\|\boldsymbol{\Sigma}_n^{-1/2}\widetilde{\boldsymbol{\theta}_n}\Big\|_p \xrightarrow{P} u, \quad \langle \widetilde{\boldsymbol{\theta}_n^{\star}}, \widetilde{\boldsymbol{\theta}_n}\rangle / \|\widetilde{\boldsymbol{\theta}_n^{\star}}\|_2^2 \xrightarrow{P} \mu, \quad \Big\|\Theta_n^{\perp}\widetilde{\boldsymbol{\theta}_n}\Big\|_2 \xrightarrow{P} \alpha,$$

for ℓ_p-norm denoting the dual of the ℓ_q-norm, $\widetilde{\boldsymbol{\theta}_n} \triangleq \boldsymbol{\Sigma}_n^{1/2}\boldsymbol{\theta}_n, \widetilde{\boldsymbol{\theta}_n^{\star}} \triangleq \boldsymbol{\Sigma}_n^{1/2}\boldsymbol{\theta}_n^{\star}$ and $\Theta_n^{\perp} \in \mathbb{R}^{n\times n}$ defined as follows:

$$\Theta_n^{\perp} \triangleq \mathbb{I}_n - \Theta_n, \quad \Theta_n \triangleq \frac{\widetilde{\boldsymbol{\theta}_n^{\star}}\widetilde{\boldsymbol{\theta}_n^{\star}}^{\top}}{\|\widetilde{\boldsymbol{\theta}_n^{\star}}\|_2^2}.$$

Moreover, in the special case of $q = 2$ and $\boldsymbol{\Sigma}_n = \mathbb{I}_n$, by denoting $\sigma \triangleq \alpha/\mu$, (3.19) simplifies to,

$$\{\mathcal{E}_{\ell_2,\varepsilon}^{GLM}(\boldsymbol{\theta}_n)\} \xrightarrow{P} \mathbb{P}\left(\frac{S\psi(S) + \sigma G}{\sqrt{\sigma^2 + 1}} < \varepsilon\right). \tag{3.20}$$

Proof: First, for $\boldsymbol{\theta}_n \neq 0$ note the following chain of equalities:

$$\begin{aligned}\max_{\|\boldsymbol{\delta}\|_q \leq \varepsilon} \mathbf{1}_{\{y \neq \operatorname{sign}\langle \mathbf{x}+\boldsymbol{\delta}, \boldsymbol{\theta}_n\rangle\}} &= \max_{\|\boldsymbol{\delta}\|_q \leq \varepsilon} \mathbf{1}_{\{y\langle \mathbf{x}, \boldsymbol{\theta}_n\rangle + y\langle \boldsymbol{\delta}, \boldsymbol{\theta}_n\rangle < 0\}} \\ &= \mathbf{1}_{\{y\langle \mathbf{x}, \boldsymbol{\theta}_n\rangle + \min_{\|\boldsymbol{\delta}\|_q \leq \varepsilon} y\langle \boldsymbol{\delta}, \boldsymbol{\theta}_n\rangle < 0\}} \\ &= \mathbf{1}_{\{y\langle \mathbf{x}, \boldsymbol{\theta}_n\rangle - \varepsilon \|\boldsymbol{\theta}_n\|_p < 0\}},\end{aligned}$$

where in the last line we used the fact that ℓ_p is the dual norm of ℓ_q norm. Thus, we can write

$$\begin{aligned}\mathbb{E}_{\mathbf{x},y}\left[\max_{\|\boldsymbol{\delta}\|_q < \varepsilon} \mathbf{1}_{\{y \neq \operatorname{sign}\langle \mathbf{x}+\delta, \boldsymbol{\theta}_n\rangle\}}\right] &= \mathbb{P}\Big(y \neq \operatorname{sign}\left(\langle \mathbf{x}, \boldsymbol{\theta}_n\rangle - y\varepsilon \|\boldsymbol{\theta}_n\|_p\right)\Big) \\ &= \mathbb{P}\Big(y\langle \mathbf{x}, \boldsymbol{\theta}_n\rangle - \varepsilon \|\boldsymbol{\theta}_n\|_p < 0\Big) \\ &= \mathbb{P}\Big(y\left\langle \bar{\mathbf{x}}, \boldsymbol{\Sigma}_n^{1/2}\boldsymbol{\theta}_n\right\rangle - \varepsilon \|\boldsymbol{\theta}_n\|_p < 0\Big) \\ &= \mathbb{P}\Big(y\left\langle \bar{\mathbf{x}}, \Theta_n\boldsymbol{\Sigma}_n^{1/2}\boldsymbol{\theta}_n\right\rangle + y\left\langle \bar{\mathbf{x}}, \Theta_n^{\perp}\boldsymbol{\Sigma}_n^{1/2}\boldsymbol{\theta}_n\right\rangle - \varepsilon \|\boldsymbol{\theta}_n\|_p < 0\Big) \\ &= \mathbb{P}\Big(y\left\langle \bar{\mathbf{x}}, \Theta_n\widetilde{\boldsymbol{\theta}_n}\right\rangle + y\left\langle \bar{\mathbf{x}}, \Theta_n^{\perp}\widetilde{\boldsymbol{\theta}_n}\right\rangle - \varepsilon \|\boldsymbol{\Sigma}_n^{-1/2}\widetilde{\boldsymbol{\theta}_n}\|_p < 0\Big),\end{aligned} \tag{3.21}$$

Where $\bar{\mathbf{x}}$ is a standard Gaussian vector. Also, for the labels y we have,

$$y = \psi\left(\langle \mathbf{x}, \boldsymbol{\theta}_n^{\star}\rangle\right) = \psi(\langle \bar{\mathbf{x}}, \widetilde{\boldsymbol{\theta}_n^{\star}}\rangle) = \psi(\langle \bar{\mathbf{x}}, \Theta_n\widetilde{\boldsymbol{\theta}_n^{\star}}\rangle).$$

Now, by Gaussianity of $\bar{\mathbf{x}}$ and since $\Theta_n\Theta_n^{\perp} = \mathbf{0}_n$, $\Theta_n + \Theta_n^{\perp} = \mathbb{I}_n$, we find that $\langle \bar{\mathbf{x}}, \Theta_n\widetilde{\boldsymbol{\theta}_n}\rangle$ and y are both independent of $\langle \bar{\mathbf{x}}, \Theta_n^{\perp}\widetilde{\boldsymbol{\theta}_n}\rangle$. Therefore, we can replace $\langle \bar{\mathbf{x}}, \Theta_n^{\perp}\widetilde{\boldsymbol{\theta}_n}\rangle$ by $\langle \bar{\bar{\mathbf{x}}}, \Theta_n^{\perp}\widetilde{\boldsymbol{\theta}_n}\rangle$ for some standard Gaussian vector $\bar{\bar{\mathbf{x}}}$ *independent* of $\bar{\mathbf{x}}$. Then, by rotational invariance

of $\bar{\tilde{\mathbf{x}}}$ and since y is independent of it and takes values ± 1, $y\bar{\tilde{\mathbf{x}}}^\top \Theta_n^\perp \widetilde{\boldsymbol{\theta}_n}$ is distributed as $\bar{\tilde{\mathbf{x}}}^\top \Theta_n^\perp \widetilde{\boldsymbol{\theta}_n}$. But, again by rotational invariance of the gaussian distribution, we have that for $G, S \overset{\text{iid}}{\sim} \mathcal{N}(0,1)$,

$$\begin{aligned}
\left\langle \bar{\mathbf{x}}, \Theta \widetilde{\boldsymbol{\theta}_n^\star} \right\rangle &\sim \|\widetilde{\boldsymbol{\theta}_n^\star}\|_2 \, S, \\
\left\langle \bar{\tilde{\mathbf{x}}}, \Theta_n^\perp \widetilde{\boldsymbol{\theta}_n} \right\rangle &\sim \|\Theta_n^\perp \widetilde{\boldsymbol{\theta}_n}\|_2 G, \\
\left\langle \bar{\mathbf{x}}, \Theta \widetilde{\boldsymbol{\theta}_n} \right\rangle = \frac{\left\langle \widetilde{\boldsymbol{\theta}_n}, \widetilde{\boldsymbol{\theta}_n^\star} \right\rangle}{\|\widetilde{\boldsymbol{\theta}_n^\star}\|_2^2} \left\langle \bar{\mathbf{x}}, \widetilde{\boldsymbol{\theta}_n^\star} \right\rangle &\sim \frac{\left\langle \widetilde{\boldsymbol{\theta}_n}, \widetilde{\boldsymbol{\theta}_n^\star} \right\rangle}{\|\widetilde{\boldsymbol{\theta}_n^\star}\|_2^2} \|\widetilde{\boldsymbol{\theta}_n^\star}\|_2 \, S.
\end{aligned}$$

Next, recall that $\left\|\widetilde{\boldsymbol{\theta}_n^\star}\right\|_2 = \boldsymbol{\theta}_n^{\star\top} \boldsymbol{\Sigma}_n \boldsymbol{\theta}_n^\star \to \zeta$ based on Assumption 3.2.2 and note the lemma's assumptions on convergence of $\|\boldsymbol{\Sigma}_n^{-1/2} \widetilde{\boldsymbol{\theta}_n}\|_p$, $\langle \widetilde{\boldsymbol{\theta}_n}, \widetilde{\boldsymbol{\theta}_n^\star} \rangle / \|\widetilde{\boldsymbol{\theta}_n^\star}\|_2^2$ and $\|\Theta^\perp \widetilde{\boldsymbol{\theta}_n}\|_2$. Combining with the above, we deduce that,

$$y \xrightarrow{P} \psi(\zeta \, S), \quad \left\langle \bar{\mathbf{x}}, \Theta_n \widetilde{\boldsymbol{\theta}_n} \right\rangle \xrightarrow{P} \mu\zeta \, S, \quad \left\langle \bar{\mathbf{x}}, \Theta_n^\perp \widetilde{\boldsymbol{\theta}_n} \right\rangle \xrightarrow{P} \alpha G. \tag{3.22}$$

Putting this together with (3.21) gives the limit in (3.19) for GLM. To derive (3.20), note that when $q = 2$ and $\boldsymbol{\Sigma}_n = \mathbb{I}_n$, it holds that $\zeta = 1$ and $u = \sqrt{\alpha^2 + \mu^2}$ due to

$$\|\widetilde{\boldsymbol{\theta}_n}\|_2 = \|\Theta_n \widetilde{\boldsymbol{\theta}_n} + \Theta_n^\perp \widetilde{\boldsymbol{\theta}_n}\|_2 = \sqrt{\|\Theta_n \widetilde{\boldsymbol{\theta}_n}\|_2^2 + \|\Theta_n^\perp \widetilde{\boldsymbol{\theta}_n}\|_2^2} \xrightarrow{P} \sqrt{\alpha^2 + \mu^2}.$$

This concludes the proof.

Case I: Correlated Features with General Covariance Matrix

For all $\mathbf{x} \in \mathbb{R}^n, \tau, C \in \mathbb{R}_+$ and a PD matrix $\mathbf{S} \in \mathbb{R}^{n \times n}$, we define,

$$\mathcal{M}_{\left(\boldsymbol{\ell}_1 + C\boldsymbol{\ell}_2^2, \mathbf{S}\right)}(\mathbf{x}; \tau) \triangleq \min_{\mathbf{y} \in \mathbb{R}^n} \; \frac{1}{2\tau} \left\|\mathbf{S}^{1/2}(\mathbf{x} - \mathbf{y})\right\|_2^2 + \|\mathbf{y}\|_1 + C\|\mathbf{y}\|_2^2.. \tag{3.23}$$

Assume that the PD covariance matrix $\mathbf{\Sigma}_n$ and the true vector $\boldsymbol{\theta}_n^\star$ satisfy the following limit for all constants $c_1, c_2, c_3, c_4 \in \mathbb{R}^+ \times \mathbb{R}^+ \times \mathbb{R} \times \mathbb{R}^+$, and the standard Gaussian vector $\mathbf{h} \in \mathbb{R}^n$,

$$\frac{1}{n}\mathcal{M}_{\left(\boldsymbol{\ell}_1+c_1\boldsymbol{\ell}_2^2,\,\mathbf{\Sigma}_n\right)}\left(c_2\mathbf{\Sigma}_n^{-1/2}\mathbf{h}+c_3\sqrt{n}\boldsymbol{\theta}_n^\star;c_4\right)\xrightarrow{P}\bar{M}_{c_1}\left(c_2,c_3,c_4\right), \tag{3.24}$$

for a function $\bar{M}:\mathbb{R}^3\to\mathbb{R}$.

Following the same notation as in (3.8), we introduce the following min-max objective based on eight scalars,

$$\begin{aligned}\min_{\substack{\alpha,\tau_1,w\in\mathbb{R}_+,\\ \mu\in\mathbb{R}}}\ \max_{\substack{\tau_2,\beta,\gamma\in\mathbb{R}_+,\\ \eta\in\mathbb{R}}}\ f_{\delta,\mathcal{C}}(\bar{\mathbf{v}})\ +\ \mathbb{E}_{G,S}\left[\mathcal{M}_{\mathcal{L}}\left(\alpha G+\mu\zeta\,S\,\psi(\zeta S)-w;\frac{\tau_1}{\beta}\right)\right]\\ +\varepsilon_{\rm tr}\gamma\,\bar{M}_{\frac{r}{\gamma\varepsilon_{\rm tr}}}\left(\frac{\alpha\beta}{\tau_2\sqrt{\delta}},\frac{\alpha\eta}{\tau_2\zeta^2},\frac{\alpha\gamma\varepsilon_{\rm tr}}{\tau_2}\right).\end{aligned} \tag{3.25}$$

Theorem 3.8.1. *Assume that the training dataset $\{(\mathbf{x}_i,y_i)\}_{i=1}^m$, is generated according to Generalized Linear models (3.3) with PD covariance matrices satisfying Assumptions 3.2.1-3.2.3. Consider the sequence of robust classifiers $\{\widehat{\boldsymbol{\theta}}_n\}$, obtained by adversarial training in (3.6) with a convex decreasing loss function $\mathcal{L}:\mathbb{R}\to\mathbb{R}$. Then, the high-dimensional limit for the adversarial test error $(\mathcal{E}_{\ell_\infty,\frac{\varepsilon_{\rm ts}}{\sqrt{n}}})$ is derived as follows,*

$$\left\{\mathcal{E}^{GLM}_{\ell_\infty,\frac{\varepsilon_{\rm ts}}{\sqrt{n}}}\left(\widehat{\boldsymbol{\theta}}_n\right)\right\}\xrightarrow{P}\mathbb{P}\Big(\mu^\star\zeta\,S\,\psi(\zeta S)+\alpha^\star G<w^\star\varepsilon_{\rm ts}/\varepsilon_{\rm tr}\Big), \tag{3.26}$$

where $(\alpha^\star,\mu^\star,w^\star)$ is the unique solution to the scalar minimax problem (3.25).

Proof: Recall that for the GLM we have $\mathbf{x}_i\overset{\rm iid}{\sim}\mathcal{N}(\mathbf{0},\mathbf{\Sigma}_n)$. Therefore, the decreasing

nature of the loss leads to the following simplification in (3.13):

$$\widehat{\boldsymbol{\theta}}_n := \min_{\boldsymbol{\theta}_n \in \mathbb{R}^n} \max_{\substack{\|\boldsymbol{\delta}_i\|_\infty \le \varepsilon \\ i \in [m]}} \frac{1}{m} \sum_{i=1}^m \mathcal{L}\left(y_i \left\langle \mathbf{x}_i + \boldsymbol{\delta}_i, \boldsymbol{\theta}_n \right\rangle\right) + \lambda \|\boldsymbol{\theta}_n\|_2^2 \tag{3.27}$$

$$= \min_{\boldsymbol{\theta}_n \in \mathbb{R}^n} \frac{1}{m} \sum_{i=1}^m \mathcal{L}(y_i \left\langle \mathbf{x}_i, \boldsymbol{\theta}_n \right\rangle - \varepsilon \|\boldsymbol{\theta}_n\|_1) + \lambda \|\boldsymbol{\theta}_n\|_2^2$$

$$= \min_{\substack{\boldsymbol{\theta}_n \in \mathbb{R}^n, \mathbf{v} \in \mathbb{R}^m \\ v_i = y_i \mathbf{x}_i^\top \boldsymbol{\theta}_n}} \frac{1}{m} \sum_{i=1}^m \mathcal{L}(v_i - \varepsilon \|\boldsymbol{\theta}_n\|_1) + \lambda \|\boldsymbol{\theta}_n\|_2^2. \tag{3.28}$$

In the last expression above, we have introduced additional variables v_i. This redundancy will allow us to write again the optimization as a minimax problem, but this time in a different —more convenient in terms of analysis— form compared to (3.27). Specifically, the minimization in (3.28) is equivalent to the following:

$$\min_{\boldsymbol{\theta}_n \in \mathbb{R}^n, \mathbf{v} \in \mathbb{R}^m} \max_{\mathbf{u} \in \mathbb{R}^m} \quad \frac{1}{m} \sum_{i=1}^m \mathcal{L}(v_i - \varepsilon \|\boldsymbol{\theta}_n\|_1) + \frac{1}{m} \sum_{i=1}^m u_i \left(y_i \left\langle \mathbf{x}_i, \boldsymbol{\theta}_n \right\rangle - v_i\right) + \lambda \|\boldsymbol{\theta}_n\|_2^2. \tag{3.29}$$

We introduce the variable $\widetilde{\boldsymbol{\theta}_n} \triangleq \boldsymbol{\Sigma}_n^{1/2} \boldsymbol{\theta}_n$ and $\bar{\mathbf{x}}_i \triangleq \boldsymbol{\Sigma}_n^{-1/2} \mathbf{x}_i$ thus $\bar{\mathbf{x}}_i \overset{\text{iid}}{\sim} \mathcal{N}(\mathbf{0}, \mathbb{I}_n)$. Based on the new notation, (3.29) can be rewritten as:

$$\min_{\widetilde{\boldsymbol{\theta}_n} \in \mathbb{R}^n, \mathbf{v} \in \mathbb{R}^m} \max_{\mathbf{u} \in \mathbb{R}^m} \quad \frac{1}{m} \sum_{i=1}^m \mathcal{L}\left(v_i - \varepsilon \left\| \boldsymbol{\Sigma}_n^{-1/2} \widetilde{\boldsymbol{\theta}_n} \right\|_1\right) + \frac{1}{m} \sum_{i=1}^m u_i \left(y_i \left\langle \bar{\mathbf{x}}_i, \widetilde{\boldsymbol{\theta}_n} \right\rangle - v_i\right)$$
$$+ \lambda \left\| \boldsymbol{\Sigma}_n^{-1/2} \widetilde{\boldsymbol{\theta}_n} \right\|_2^2. \tag{3.30}$$

Next, we define the projection matrices $\Theta_n, \Theta_n^\perp \in \mathbb{R}^{n \times n}$ based on $\widetilde{\boldsymbol{\theta}_n^\star} \triangleq \boldsymbol{\Sigma}_n^{1/2} \boldsymbol{\theta}_n^\star$ as follows,

$$\Theta_n \triangleq \frac{\widetilde{\boldsymbol{\theta}_n^\star} \widetilde{\boldsymbol{\theta}_n^\star}^\top}{\left\| \widetilde{\boldsymbol{\theta}_n^\star} \right\|_2^2}, \quad \Theta_n^\perp \triangleq \mathbb{I}_n - \Theta_n.$$

Since $\Theta_n + \Theta_n^{\perp} = \mathbb{I}_n$, we deduce that (3.30) is equivalent to,

$$\min_{\widetilde{\boldsymbol{\theta}_n}\in\mathbb{R}^n, \mathbf{v}\in\mathbb{R}^m} \max_{\mathbf{u}\in\mathbb{R}^m} \frac{1}{m}\sum_{i=1}^{m} \mathcal{L}\left(v_i - \varepsilon \left\|\boldsymbol{\Sigma}_n^{-1/2}\widetilde{\boldsymbol{\theta}_n}\right\|_1\right) - \frac{1}{m}\sum_{i=1}^{m} u_i v_i + \frac{1}{m}\sum_{i=1}^{m} u_i y_i \left\langle \bar{\mathbf{x}}_i, \Theta_n \widetilde{\boldsymbol{\theta}_n}\right\rangle \tag{3.31}$$

$$+ \frac{1}{m}\sum_{i=1}^{m} u_i y_i \left\langle \bar{\mathbf{x}}_i, \Theta_n^{\perp}\widetilde{\boldsymbol{\theta}_n}\right\rangle + \lambda \left\|\boldsymbol{\Sigma}_n^{-1/2}\widetilde{\boldsymbol{\theta}_n}\right\|_2^2.$$

Splitting $\widetilde{\boldsymbol{\theta}_n}$ based on $\Theta_n, \Theta_n^{\perp}$ has two purposes. First it immediately reveals the two terms $\|\Theta_n\widetilde{\boldsymbol{\theta}_n}\|_2$ and $\|\Theta_n^{\perp}\widetilde{\boldsymbol{\theta}_n}\|_2$ of interest to us in view of Lemma 3.8.1 . Second, as we will see, it allows the use of the CGMT.

For compactness we write (3.31) in vector notation,

$$\min_{\widetilde{\boldsymbol{\theta}_n}\in\mathbb{R}^n, \mathbf{v}\in\mathbb{R}^m} \max_{\mathbf{u}\in\mathbb{R}^m} \frac{\mathbf{1}_m^{\top}}{m}\left(\mathbf{v} - \varepsilon \left\|\boldsymbol{\Sigma}_n^{-1/2}\widetilde{\boldsymbol{\theta}_n}\right\|_1 \mathbf{1}_m\right) - \frac{\langle \mathbf{u}, \mathbf{v}\rangle}{m} + \frac{\left\langle \mathbf{u}, Y\bar{X}\Theta_n\widetilde{\boldsymbol{\theta}_n}\right\rangle}{m} + \frac{\left\langle \mathbf{u}, Y\bar{X}\Theta_n^{\perp}\widetilde{\boldsymbol{\theta}_n}\right\rangle}{m}$$

$$+ \lambda \left\|\boldsymbol{\Sigma}_n^{-1/2}\widetilde{\boldsymbol{\theta}_n}\right\|_2^2, \tag{3.32}$$

where

$$\begin{aligned}
(\mathbf{v}) &\triangleq [\mathcal{L}(v_1);\ \mathcal{L}(v_2);\ \cdots;\ \mathcal{L}(v_m)] \in \mathbb{R}^{m\times 1},\\
Y &\triangleq \operatorname{diag}(y_1,\ y_2, \cdots,\ y_m) \in \mathbb{R}^{m\times m},\\
\bar{X} &\triangleq [\bar{\mathbf{x}}_1^{\top};\ \bar{\mathbf{x}}_2^{\top}; \cdots;\ \bar{\mathbf{x}}_m^{\top}] \in \mathbb{R}^{m\times n}.
\end{aligned} \tag{3.33}$$

Before proceeding, we recall our main tool the Convex Gaussian Min-max Theorem [?, ?, ?] which relies on Gordon's Gaussian Min-max theorem. The Gordon's Gaussian comparison inequality [3] compares the min-max value of two doubly indexed Gaussian

processes $\mathcal{X}_{\mathbf{w},\mathbf{u}}, \mathcal{Y}_{\mathbf{w},\mathbf{u}}$ based on how their autocorrelation functions compare,

$$\mathcal{X}_{\mathbf{w},\mathbf{u}} \triangleq \mathbf{u}^\top \mathbf{G}\mathbf{w} + \Gamma(\mathbf{w},\mathbf{u}), \tag{3.34a}$$

$$\mathcal{Y}_{\mathbf{w},\mathbf{u}} \triangleq \|\mathbf{w}\|_2 \mathbf{g}^\top \mathbf{u} + \|\mathbf{u}\|_2 \mathbf{h}^\top \mathbf{w} + \Gamma(\mathbf{w},\mathbf{u}), \tag{3.34b}$$

where: $\mathbf{G} \in \mathbb{R}^{m\times n}$, $\mathbf{g} \in \mathbb{R}^m$, $\mathbf{h} \in \mathbb{R}^n$, they all have entries iid Gaussian; the sets $\mathcal{S}_{\mathbf{w}} \subset \mathbb{R}^n$ and $\mathcal{S}_{\mathbf{u}} \subset \mathbb{R}^m$ are compact; and, $\Gamma : \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}$. For these two processes, define the following (random) min-max optimization programs, which we refer to as the *primary optimization* (PO) problem and the *auxiliary optimization* (AO).

$$\Phi(\mathbf{G}) = \min_{\mathbf{w}\in\mathcal{S}_{\mathbf{w}}} \max_{\mathbf{u}\in\mathcal{S}_{\mathbf{u}}} \mathcal{X}_{\mathbf{w},\mathbf{u}}, \tag{3.35a}$$

$$\phi(\mathbf{g},\mathbf{h}) = \min_{\mathbf{w}\in\mathcal{S}_{\mathbf{w}}} \max_{\mathbf{u}\in\mathcal{S}_{\mathbf{u}}} \mathcal{Y}_{\mathbf{w},\mathbf{u}}. \tag{3.35b}$$

According to the version of the CGMT in Theorem 6.1 in [5], if the sets $\mathcal{S}_{\mathbf{w}}$ and $\mathcal{S}_{\mathbf{u}}$ are convex and ψ is continuous *convex-concave* on $\mathcal{S}_{\mathbf{w}} \times \mathcal{S}_{\mathbf{u}}$, then, for any $\nu \in \mathbb{R}$ and $t > 0$, it holds

$$\mathbb{P}\left(|\Phi(\mathbf{G}) - \nu| > t\right) \leq 2\mathbb{P}\left(|\phi(\mathbf{g},\mathbf{h}) - \nu| > t\right). \tag{3.36}$$

In words, concentration of the optimal cost of the AO problem around μ implies concentration of the optimal cost of the corresponding PO problem around the same value μ. Moreover, starting from (3.36) and under strict convexity conditions, the CGMT shows that concentration of the optimal solution of the AO problem implies concentration of the optimal solution of the PO to the same value. For example, if minimizers of (3.35b) satisfy $\|\mathbf{w}^*(\mathbf{g},\mathbf{h})\|_2 \to \zeta^*$ for some $\zeta^* > 0$, then, the same holds true for the minimizers of (3.35a): $\|\mathbf{w}^*(\mathbf{G})\|_2 \to \zeta^*$ (Theorem 6.1(iii) in [5]). Thus, one can analyze the AO to infer corresponding properties of the PO, the premise being of course that the former is

simpler to handle than the latter.

Returning to our minimax problem (3.32), we observe that the objective is convex in $(\widetilde{\boldsymbol{\theta}_n}, \mathbf{v})$ and concave in $\mathbf{u}$. Also note that the term $Y\bar{X}\Theta_n\widetilde{\boldsymbol{\theta}_n}$ is independent of $Y\bar{X}\Theta_n^{\perp}\widetilde{\boldsymbol{\theta}_n}$ as the entries of Y depend only on $\bar{X}\Theta_n$ which is orthogonal to $\bar{X}\Theta_n^{\perp}$, i.e., based on the definition of Θ_n we have

$$y_i = \psi(\mathbf{x}_i^\top \boldsymbol{\theta}_n^\star) = \psi(\bar{\mathbf{x}}_i^\top \widetilde{\boldsymbol{\theta}_n^\star}) = \psi(\bar{\mathbf{x}}_i^\top \Theta_n \widetilde{\boldsymbol{\theta}_n^\star}).$$

Therefore, along the same lines as in the proof of Lemma 3.8.1, we can substitute $\bar{X}\Theta_n^{\perp}$ by $\hat{X}\Theta_n^{\perp}$ for a standard Gaussian matrix $\hat{X}$ that is independent of $\bar{X}$ and everything else in the objective of (3.32). Thus, we can use CGMT for PO in (3.32) with the choice

$$\Gamma\left(\{\widetilde{\boldsymbol{\theta}_n}, \mathbf{v}\}, \mathbf{u}\right) \triangleq \frac{\mathbf{1}_m^\top}{m}\left(\mathbf{v} - \varepsilon \left\|\boldsymbol{\Sigma}_n^{-1/2}\widetilde{\boldsymbol{\theta}_n}\right\|_1 \mathbf{1}_m\right) - \frac{\langle \mathbf{u}, \mathbf{v}\rangle}{m} + \frac{\left\langle \mathbf{u}, Y\bar{X}\Theta_n\widetilde{\boldsymbol{\theta}_n}\right\rangle}{m} + \lambda \left\|\boldsymbol{\Sigma}_n^{-1/2}\widetilde{\boldsymbol{\theta}_n}\right\|_2^2.$$

With this, we derive the following AO for (3.32),

$$\begin{aligned}\min_{\widetilde{\boldsymbol{\theta}_n}\in\mathbb{R}^n, \mathbf{v}\in\mathbb{R}^m} \max_{\mathbf{u}\in\mathbb{R}^m} \; & \frac{\mathbf{1}_m^\top}{m}\left(\mathbf{v} - \varepsilon \left\|\boldsymbol{\Sigma}_n^{-1/2}\widetilde{\boldsymbol{\theta}_n}\right\|_1 \mathbf{1}_m\right) - \frac{\langle \mathbf{u}, \mathbf{v}\rangle}{m} + \frac{\left\langle \mathbf{u}, Y\bar{X}\Theta_n\widetilde{\boldsymbol{\theta}_n}\right\rangle}{m} \\ & + \frac{\mathbf{u}^\top Y\mathbf{g}\left\|\Theta_n^{\perp}\widetilde{\boldsymbol{\theta}_n}\right\|_2}{m} + \frac{\|\mathbf{u}^\top Y\|_2\left\langle \mathbf{h}, \Theta_n^{\perp}\widetilde{\boldsymbol{\theta}_n}\right\rangle}{m} + \lambda\left\|\boldsymbol{\Sigma}_n^{-1/2}\widetilde{\boldsymbol{\theta}_n}\right\|_2^2, \end{aligned} \tag{3.37}$$

where $\mathbf{g} \in \mathbb{R}^m$, $\mathbf{h} \in \mathbb{R}^n$ have entries i.i.d. standard normal. Note that similar to [5] and despite the fact that we are working with finite dimensional matrices now, we will consider the asymptotic limit at the end of the approach. Thus as the final optimization has a bounded solution in the high-dimensional limit, we can relax the assumption of compactness of the domain of optimization which is needed for CGMT.

To proceed, we observe that $Y\mathbf{g} \sim \mathcal{N}(0,1)$ and $\|\mathbf{u}^\top Y\|_2 = \|\mathbf{u}\|_2$. So, next we can

optimize w.r.t $\mathbf{u}$ to find that:

$$\begin{aligned}
&\max_{\mathbf{u}\in\mathbb{R}^m} -\frac{\langle \mathbf{u}, \mathbf{v}\rangle}{m} + \frac{\left\langle \mathbf{u}, Y\bar{X}\Theta_n\widetilde{\boldsymbol{\theta}_n}\right\rangle}{m} + \frac{\langle \mathbf{u}, \mathbf{g}\rangle \left\|\Theta_n^{\perp}\widetilde{\boldsymbol{\theta}_n}\right\|_2}{m} + \frac{\|\mathbf{u}\|_2\left\langle \mathbf{h}, \Theta_n^{\perp}\widetilde{\boldsymbol{\theta}_n}\right\rangle}{m} \\
&= \max_{\mathbf{u}\in\mathbb{R}^m, \frac{\|\mathbf{u}\|_2}{\sqrt{m}}=\beta} \frac{1}{m}\left\langle \mathbf{u}, -\mathbf{v} + Y\bar{X}\Theta_n\widetilde{\boldsymbol{\theta}_n} + \mathbf{g}\left\|\Theta_n^{\perp}\widetilde{\boldsymbol{\theta}_n}\right\|_2\right\rangle + \frac{\beta\left\langle \mathbf{h}, \Theta_n^{\perp}\widetilde{\boldsymbol{\theta}_n}\right\rangle}{\sqrt{m}} \\
&= \max_{\beta\in\mathbb{R}_+} \frac{\beta}{\sqrt{m}}\left\|-\mathbf{v} + Y\bar{X}\Theta_n\widetilde{\boldsymbol{\theta}_n} + \mathbf{g}\left\|\Theta_n^{\perp}\widetilde{\boldsymbol{\theta}_n}\right\|_2\right\|_2 + \frac{\beta\left\langle \mathbf{h}, \Theta_n^{\perp}\widetilde{\boldsymbol{\theta}_n}\right\rangle}{\sqrt{m}}.
\end{aligned}$$

Hence we replace this in (3.37) to simplify the objective as follows,

$$\begin{aligned}
\min_{\widetilde{\boldsymbol{\theta}_n}\in\mathbb{R}^n, \mathbf{v}\in\mathbb{R}^m} \max_{\beta\in\mathbb{R}_+} \; &\frac{\mathbf{1}_m^\top}{m}\left(\mathbf{v} - \varepsilon\left\|\boldsymbol{\Sigma}_n^{-1/2}\widetilde{\boldsymbol{\theta}_n}\right\|_1 \mathbf{1}_m\right) + \frac{\beta}{\sqrt{m}}\left\|-\mathbf{v} + Y\bar{X}\Theta_n\widetilde{\boldsymbol{\theta}_n} + \mathbf{g}\left\|\Theta_n^{\perp}\widetilde{\boldsymbol{\theta}_n}\right\|_2\right\|_2 \\
&+ \frac{\beta\left\langle \mathbf{h}, \Theta_n^{\perp}\widetilde{\boldsymbol{\theta}_n}\right\rangle}{\sqrt{m}} + \lambda\left\|\boldsymbol{\Sigma}_n^{-1/2}\widetilde{\boldsymbol{\theta}_n}\right\|_2^2. \qquad (3.38)
\end{aligned}$$

Next, our trick is to dualize the the term $\varepsilon\left\|\boldsymbol{\Sigma}_n^{-1/2}\widetilde{\boldsymbol{\theta}_n}\right\|_1$ inside the loss function. For this, we first introduce an extra optimization variable $w > 0$ along with the constraint $w = \varepsilon\left\|\boldsymbol{\Sigma}_n^{-1/2}\widetilde{\boldsymbol{\theta}_n}\right\|_1$ and then turn this into an unconstrained min-max problem. This yields the following equivalent formulation of (3.38),

$$\begin{aligned}
\min_{\substack{\widetilde{\boldsymbol{\theta}_n}\in\mathbb{R}^n, \mathbf{v}\in\mathbb{R}^m \\ w=\varepsilon\left\|\boldsymbol{\Sigma}_n^{-1/2}\widetilde{\boldsymbol{\theta}_n}\right\|_1}} \max_{\beta\in\mathbb{R}_+} \; &\frac{\mathbf{1}_m^\top}{m}(\mathbf{v} - w\mathbf{1}_m) + \frac{\beta}{\sqrt{m}}\left\|-\mathbf{v} + Y\bar{X}\Theta_n\widetilde{\boldsymbol{\theta}_n} + \mathbf{g}\left\|\Theta_n^{\perp}\widetilde{\boldsymbol{\theta}_n}\right\|_2\right\|_2 \\
&+ \frac{\beta\left\langle \mathbf{h}, \Theta_n^{\perp}\widetilde{\boldsymbol{\theta}_n}\right\rangle}{\sqrt{m}} + \lambda\left\|\boldsymbol{\Sigma}_n^{-1/2}\widetilde{\boldsymbol{\theta}_n}\right\|_2^2 \qquad (3.39) \\
= \min_{\widetilde{\boldsymbol{\theta}_n}\in\mathbb{R}^n, \mathbf{v}\in\mathbb{R}^m, w\in\mathbb{R}_+} \max_{\beta,\gamma\in\mathbb{R}_+} \; &\frac{\mathbf{1}_m^\top}{m}(\mathbf{v} - w\mathbf{1}_m) + \gamma\left(\varepsilon\left\|\boldsymbol{\Sigma}_n^{-1/2}\widetilde{\boldsymbol{\theta}_n}\right\|_1 - w\right) + \lambda\left\|\boldsymbol{\Sigma}_n^{-1/2}\widetilde{\boldsymbol{\theta}_n}\right\|_2^2 \\
&+ \frac{\beta}{\sqrt{m}}\left\|-\mathbf{v} + Y\bar{X}\Theta_n\widetilde{\boldsymbol{\theta}_n} + \mathbf{g}\left\|\Theta_n^{\perp}\widetilde{\boldsymbol{\theta}_n}\right\|_2\right\|_2 + \frac{\beta\left\langle \mathbf{h}, \Theta_n^{\perp}\widetilde{\boldsymbol{\theta}_n}\right\rangle}{\sqrt{m}}. \qquad (3.40)
\end{aligned}$$

The key reason behind this reformulation is to allow optimization with respect to $\widetilde{\boldsymbol{\theta}_n}$ which is the primary variable of interest in the objective function. As we will see, our goal is optimizing with respect to the direction of $\Theta_n^{\perp}\widetilde{\boldsymbol{\theta}_n}$ and $\Theta_n\widetilde{\boldsymbol{\theta}_n}$, which according to Lemma 3.8.1 comprise the terms parametrizing the adversarial error of the estimator $\widetilde{\boldsymbol{\theta}_n}$. To do this, we introduce the slack variable $\widetilde{\boldsymbol{\rho}_n}$ for $\widetilde{\boldsymbol{\theta}_n}$ (equivalently $\boldsymbol{\rho}_n$ for $\boldsymbol{\theta}_n$ where $\boldsymbol{\rho}_n \triangleq \boldsymbol{\Sigma}_n^{-1/2}\widetilde{\boldsymbol{\rho}_n}$) and rewrite the optimization problem (3.40),

$$\begin{aligned}
&\min_{\substack{\widetilde{\boldsymbol{\theta}_n}\in\mathbb{R}^n,\mathbf{v}\in\mathbb{R}^m,w\in\mathbb{R}_+ \\ \text{s.t. } \boldsymbol{\Sigma}_n^{-1/2}\widetilde{\boldsymbol{\theta}_n}=\boldsymbol{\rho}_n}} \max_{\beta,\gamma\in\mathbb{R}_+} \frac{\mathbf{1}_m^\top}{m}(\mathbf{v}-w\mathbf{1}_m) + \gamma\left(\varepsilon\left\|\boldsymbol{\Sigma}_n^{-1/2}\widetilde{\boldsymbol{\theta}_n}\right\|_1 - w\right) + \lambda\left\|\boldsymbol{\Sigma}_n^{-1/2}\widetilde{\boldsymbol{\theta}_n}\right\|_2^2 \\
&\qquad + \frac{\beta}{\sqrt{m}}\left\|-\mathbf{v} + Y\bar{X}\Theta_n\widetilde{\boldsymbol{\theta}_n} + \mathbf{g}\left\|\Theta_n^{\perp}\widetilde{\boldsymbol{\theta}_n}\right\|_2\right\|_2 + \frac{\beta\left\langle \mathbf{h}, \Theta_n^{\perp}\widetilde{\boldsymbol{\theta}_n}\right\rangle}{\sqrt{m}} \\
&= \min_{\widetilde{\boldsymbol{\rho}_n},\widetilde{\boldsymbol{\theta}_n}\in\mathbb{R}^n,\mathbf{v}\in\mathbb{R}^m,w\in\mathbb{R}} \max_{\beta,\gamma\in\mathbb{R}_+,\boldsymbol{\lambda}\in\mathbb{R}^n} \frac{\mathbf{1}_m^\top}{m}(\mathbf{v}-w\mathbf{1}_m) - \gamma w + \varepsilon\gamma\left\|\boldsymbol{\Sigma}_n^{-1/2}\widetilde{\boldsymbol{\rho}_n}\right\|_1 + r\left\|\boldsymbol{\Sigma}_n^{-1/2}\widetilde{\boldsymbol{\rho}_n}\right\|_2^2 \\
&\qquad + \left\langle\frac{\boldsymbol{\lambda}}{\sqrt{n}}, \widetilde{\boldsymbol{\rho}_n} - \widetilde{\boldsymbol{\theta}_n}\right\rangle + \frac{\beta}{\sqrt{m}}\left\|-\mathbf{v} + Y\bar{X}\Theta_n\widetilde{\boldsymbol{\theta}_n} + \mathbf{g}\left\|\Theta_n^{\perp}\widetilde{\boldsymbol{\theta}_n}\right\|_2\right\|_2 + \frac{\beta\left\langle \mathbf{h}, \Theta_n^{\perp}\widetilde{\boldsymbol{\theta}_n}\right\rangle}{\sqrt{m}}.
\end{aligned} \tag{3.41}$$

In (3.41), we applied the Lagrangian method to both of terms $\left\|\boldsymbol{\Sigma}_n^{-1/2}\widetilde{\boldsymbol{\theta}_n}\right\|_1$ and $\left\|\boldsymbol{\Sigma}_n^{-1/2}\widetilde{\boldsymbol{\theta}_n}\right\|_2^2$. This is essential to scalarizing the objective function based on $\Theta_n\widetilde{\boldsymbol{\theta}_n}$ and $\Theta_n^{\perp}\widetilde{\boldsymbol{\theta}_n}$, which is our next step. As a remark and as we will see in Section 3.8.1, only in the special case of $\boldsymbol{\Sigma}_n = \mathbb{I}_n$, it is possible to apply the Lagrangian to the ℓ_1 norm and simply decompose $\|\widetilde{\boldsymbol{\theta}_n}\|_2^2$ as

$$\|\widetilde{\boldsymbol{\theta}_n}\|_2^2 = \|\Theta_n\widetilde{\boldsymbol{\theta}_n}\|_2^2 + \|\Theta_n^{\perp}\widetilde{\boldsymbol{\theta}_n}\|_2^2.$$

Now, we can finally optimize w.r.t the direction of $\Theta_n^{\perp}\widetilde{\boldsymbol{\theta}_n}$. First, note that

$$\left\langle\boldsymbol{\lambda}, \widetilde{\boldsymbol{\rho}_n} - \widetilde{\boldsymbol{\theta}_n}\right\rangle = \left\langle\boldsymbol{\lambda}, \Theta_n\left(\widetilde{\boldsymbol{\rho}_n} - \widetilde{\boldsymbol{\theta}_n}\right)\right\rangle + \left\langle\boldsymbol{\lambda}, \Theta_n^{\perp}\widetilde{\boldsymbol{\rho}_n}\right\rangle - \left\langle\boldsymbol{\lambda}, \Theta_n^{\perp}\widetilde{\boldsymbol{\theta}_n}\right\rangle.$$

With this decomposition, we can optimize w.r.t. $\Theta_n^{\perp}\widetilde{\boldsymbol{\theta}_n}$ as follows,

$$\begin{aligned}
&\min_{\Theta_n^{\perp}\widetilde{\boldsymbol{\theta}_n}\in\mathbb{R}^n} -\left\langle \frac{\boldsymbol{\lambda}}{\sqrt{n}}, \Theta_n^{\perp}\widetilde{\boldsymbol{\theta}_n}\right\rangle + \frac{\beta}{\sqrt{m}}\left\|-\mathbf{v}+Y\bar{X}\Theta_n\widetilde{\boldsymbol{\theta}_n}+\mathbf{g}\left\|\Theta_n^{\perp}\widetilde{\boldsymbol{\theta}_n}\right\|_2\right\|_2 + \frac{\beta\left\langle \mathbf{h}, \Theta_n^{\perp}\widetilde{\boldsymbol{\theta}_n}\right\rangle}{\sqrt{m}} \\
&= \min_{\Theta_n^{\perp}\widetilde{\boldsymbol{\theta}_n}\in\mathbb{R}^n, \|\Theta_n^{\perp}\widetilde{\boldsymbol{\theta}_n}\|_2=\alpha} \left\langle -\frac{\boldsymbol{\lambda}}{\sqrt{n}} + \frac{\beta\mathbf{h}}{\sqrt{m}}, \Theta_n^{\perp}\widetilde{\boldsymbol{\theta}_n}\right\rangle + \frac{\beta}{\sqrt{m}}\left\|-\mathbf{v}+Y\bar{X}\Theta_n\widetilde{\boldsymbol{\theta}_n}+\alpha\mathbf{g}\right\|_2 \\
&= \min_{\alpha\in\mathbb{R}_+} -\alpha\left\|-\frac{\Theta_n^{\perp}\boldsymbol{\lambda}}{\sqrt{n}} + \frac{\beta}{\sqrt{m}}\Theta_n^{\perp}\mathbf{h}\right\|_2 + \frac{\beta}{\sqrt{m}}\left\|-\mathbf{v}+Y\bar{X}\Theta_n\widetilde{\boldsymbol{\theta}_n}+\alpha\mathbf{g}\right\|_2. \qquad (3.42)
\end{aligned}$$

By replacing (3.42) in (3.41) we have,

$$\begin{aligned}
\min_{\widetilde{\boldsymbol{\rho}_n},\Theta_n\widetilde{\boldsymbol{\theta}_n}\in\mathbb{R}^n,\mathbf{v}\in\mathbb{R}^m,w,\alpha\in\mathbb{R}_+} \max_{\beta,\gamma\in\mathbb{R}_+,\boldsymbol{\lambda}\in\mathbb{R}^n} \; & \frac{\mathbf{1}_m^{\top}}{m}(\mathbf{v}-w\mathbf{1}_m) - \gamma w + \varepsilon\gamma\left\|\boldsymbol{\Sigma}_n^{-1/2}\widetilde{\boldsymbol{\rho}_n}\right\|_1 \\
& + \left\langle \frac{\boldsymbol{\lambda}}{\sqrt{n}}, \Theta_n\left(\widetilde{\boldsymbol{\rho}_n}-\widetilde{\boldsymbol{\theta}_n}\right)\right\rangle + r\left\|\boldsymbol{\Sigma}_n^{-1/2}\widetilde{\boldsymbol{\rho}_n}\right\|_2^2 + \left\langle \frac{\boldsymbol{\lambda}}{\sqrt{n}}, \Theta_n^{\perp}\widetilde{\boldsymbol{\rho}_n}\right\rangle \\
& - \alpha\left\|-\frac{\Theta_n^{\perp}\boldsymbol{\lambda}}{\sqrt{n}} + \frac{\beta}{\sqrt{m}}\Theta_n^{\perp}\mathbf{h}\right\|_2 + \frac{\beta}{\sqrt{m}}\left\|-\mathbf{v}+Y\bar{X}\Theta_n\widetilde{\boldsymbol{\theta}_n}+\alpha\mathbf{g}\right\|_2. \qquad (3.43)
\end{aligned}$$

We replace ε with $\varepsilon_{\mathrm{tr}}/\sqrt{n}$ specialized to the case of $q=\infty$. Such normalization is necessary to guarantee the boundedness of the solutions to (3.43) when $\varepsilon_{\mathrm{tr}}=\mathcal{O}(1)$. To continue, we will use the same trick as in [?] that $x=\min_{\tau\in\mathbb{R}_+}\frac{x^2}{2\tau}+\frac{\tau}{2}$ for every $x\in\mathbb{R}_+$. Thus we may rewrite the last two terms based on the squared ℓ_2 norm by introducing

two new variables $\tau_1, \tau_2 \in \mathbb{R}_+$ to obtain the following new objective,

$$\min_{\widetilde{\boldsymbol{\rho}_n}, \Theta_n \widetilde{\boldsymbol{\theta}_n} \in \mathbb{R}^n, \mathbf{v} \in \mathbb{R}^m, w, \alpha, \tau_1 \in \mathbb{R}_+} \max_{\tau_2, \beta, \gamma \in \mathbb{R}_+, \boldsymbol{\lambda} \in \mathbb{R}^n} \frac{\mathbf{1}_m^\top}{m}(\mathbf{v} - w\mathbf{1}_m) - \gamma w + \frac{\varepsilon_{\mathrm{tr}}\gamma}{\sqrt{n}} \left\| \boldsymbol{\Sigma}_n^{-1/2} \widetilde{\boldsymbol{\rho}_n} \right\|_1$$
$$+ \left\langle \frac{\boldsymbol{\lambda}}{\sqrt{n}}, \Theta_n \left(\widetilde{\boldsymbol{\rho}_n} - \widetilde{\boldsymbol{\theta}_n} \right) \right\rangle + r \left\| \boldsymbol{\Sigma}_n^{-1/2} \widetilde{\boldsymbol{\rho}_n} \right\|_2^2 + \left\langle \frac{\boldsymbol{\lambda}}{\sqrt{n}}, \Theta_n^\perp \widetilde{\boldsymbol{\rho}_n} \right\rangle$$
$$- \frac{\alpha}{2\tau_2 n} \left\| -\Theta_n^\perp \boldsymbol{\lambda} + \frac{\beta}{\sqrt{\delta}} \Theta_n^\perp \mathbf{h} \right\|_2^2 - \frac{\alpha\tau_2}{2} + \frac{\beta}{2\tau_1 m} \Big\| - \mathbf{v} + Y\bar{X}\Theta_n \widetilde{\boldsymbol{\theta}_n} + \alpha \mathbf{g} \Big\|_2^2 + \frac{\beta\tau_1}{2}, \tag{3.44}$$

where we also used the fact that $m/n = \delta$. By the following chain of equations, we simplify the maximization with respect to $\boldsymbol{\lambda}$,

$$\max_{\boldsymbol{\lambda} \in \mathbb{R}^n} \left\langle \frac{\boldsymbol{\lambda}}{\sqrt{n}}, \Theta_n^\perp \widetilde{\boldsymbol{\rho}_n} \right\rangle - \frac{\alpha}{2\tau_2 n} \left\| -\Theta_n^\perp \boldsymbol{\lambda} + \frac{\beta}{\sqrt{\delta}} \Theta_n^\perp \mathbf{h} \right\|_2^2 + \left\langle \frac{\boldsymbol{\lambda}}{\sqrt{n}}, \Theta_n \left(\widetilde{\boldsymbol{\rho}_n} - \widetilde{\boldsymbol{\theta}_n} \right) \right\rangle$$
$$= \max_{\boldsymbol{\lambda} \in \mathbb{R}^n} -\frac{\alpha}{2n\tau_2} \left\| \Theta_n^\perp \left(\frac{\beta}{\sqrt{\delta}} \mathbf{h} - \boldsymbol{\lambda} + \frac{\tau_2 \widetilde{\boldsymbol{\rho}_n} \sqrt{n}}{\alpha} \right) \right\|_2^2 + \frac{\tau_2}{2n\alpha} \left\| \Theta_n^\perp \left(\widetilde{\boldsymbol{\rho}_n} \sqrt{n} + \frac{\alpha\beta}{\tau_2 \sqrt{\delta}} \mathbf{h} \right) \right\|_2^2$$
$$- \frac{\alpha\beta^2}{2m\tau_2} \left\| \Theta_n^\perp \mathbf{h} \right\|_2^2 + \left\langle \frac{\boldsymbol{\lambda}}{\sqrt{n}}, \Theta_n \left(\widetilde{\boldsymbol{\rho}_n} - \widetilde{\boldsymbol{\theta}_n} \right) \right\rangle \tag{3.45}$$
$$= \max_{\Theta_n \boldsymbol{\lambda} \in \mathbb{R}^n} \max_{\Theta_n^\perp \boldsymbol{\lambda} \in \mathbb{R}^n} -\frac{\alpha}{2n\tau_2} \left\| \Theta_n^\perp \left(\frac{\beta}{\sqrt{\delta}} \mathbf{h} - \boldsymbol{\lambda} + \frac{\tau_2 \widetilde{\boldsymbol{\rho}_n} \sqrt{n}}{\alpha} \right) \right\|_2^2 + \frac{\tau_2}{2n\alpha} \left\| \Theta_n^\perp \left(\widetilde{\boldsymbol{\rho}_n} \sqrt{n} + \frac{\alpha\beta}{\tau_2 \sqrt{\delta}} \mathbf{h} \right) \right\|_2^2$$
$$- \frac{\alpha\beta^2}{2m\tau_2} \left\| \Theta_n^\perp \mathbf{h} \right\|_2^2 + \left\langle \frac{\Theta_n \boldsymbol{\lambda}}{\sqrt{n}}, \Theta_n \left(\widetilde{\boldsymbol{\rho}_n} - \widetilde{\boldsymbol{\theta}_n} \right) \right\rangle \tag{3.46}$$
$$= \max_{\Theta_n \boldsymbol{\lambda} \in \mathbb{R}^n} \frac{\tau_2}{2n\alpha} \left\| \Theta_n^\perp \left(\widetilde{\boldsymbol{\rho}_n} \sqrt{n} + \frac{\alpha\beta}{\tau_2 \sqrt{\delta}} \mathbf{h} \right) \right\|_2^2 - \frac{\alpha\beta^2}{2m\tau_2} \left\| \Theta_n^\perp \mathbf{h} \right\|_2^2 + \left\langle \frac{\Theta_n \boldsymbol{\lambda}}{\sqrt{n}}, \Theta_n \left(\widetilde{\boldsymbol{\rho}_n} - \widetilde{\boldsymbol{\theta}_n} \right) \right\rangle$$
$$= \frac{\tau_2}{2n\alpha} \left\| \Theta_n^\perp \left(\widetilde{\boldsymbol{\rho}_n} \sqrt{n} + \frac{\alpha\beta}{\tau_2 \sqrt{\delta}} \mathbf{h} \right) \right\|_2^2 - \frac{\alpha\beta^2}{2\delta\tau_2}. \tag{3.47}$$

In deriving (3.45) we used completion of squares. In (3.46), we decompose maximization of $\boldsymbol{\lambda}$ into $\Theta_n \boldsymbol{\lambda}$ and $\Theta_n^\perp \boldsymbol{\lambda}$ and used the fact that $\Theta_n^\perp + \Theta_n = \mathbb{I}_n$ and $\Theta_n^\perp \Theta_n = \mathbf{0}_n$. In the

last line we used the fact that $\|\Theta_n^{\perp}\mathbf{h}\|_2^2 \to n$. We note that the last line is true subject to the constraint $\Theta_n\widetilde{\boldsymbol{\rho}_n} = \Theta_n\widetilde{\boldsymbol{\theta}_n}$, which ensures boundedness of the min-max objective. We include this constraint in the next step of the proof. Therefore, inserting (3.47) back in (3.44) we derive,

$$\begin{aligned}\min_{\substack{\widetilde{\boldsymbol{\rho}_n},\Theta_n\widetilde{\boldsymbol{\theta}_n}\in\mathbb{R}^n,\mathbf{v}\in\mathbb{R}^m,w,\alpha,\tau_1\in\mathbb{R}_+\\ \text{s.t. } \Theta_n\widetilde{\boldsymbol{\rho}_n}=\Theta_n\widetilde{\boldsymbol{\theta}_n}}} \max_{\gamma,\tau_2,\beta\in\mathbb{R}_+} \; & \frac{\mathbf{1}_m^\top}{m}(\mathbf{v}-w\mathbf{1}_m) - \gamma w + \frac{\varepsilon_{\text{tr}}\gamma}{\sqrt{n}}\left\|\boldsymbol{\Sigma}_n^{-1/2}\widetilde{\boldsymbol{\rho}_n}\right\|_1 \\ & + \; r\left\|\boldsymbol{\Sigma}_n^{-1/2}\widetilde{\boldsymbol{\rho}_n}\right\|_2^2 + \frac{\tau_2}{2n\alpha}\left\|\Theta_n^{\perp}\left(\widetilde{\boldsymbol{\rho}_n}\sqrt{n} + \frac{\alpha\beta}{\tau_2\sqrt{\delta}}\mathbf{h}\right)\right\|_2^2 - \frac{\alpha\beta^2}{2\delta\tau_2} - \frac{\alpha\tau_2}{2} \\ & + \frac{\beta}{2\tau_1 m}\left\| - \mathbf{v} + Y\bar{X}\Theta_n\widetilde{\boldsymbol{\theta}_n} + \alpha\mathbf{g}\right\|_2^2 + \frac{\beta\tau_1}{2}. \end{aligned} \tag{3.48}$$

Recalling $\Theta_n^{\perp} \triangleq \mathbb{I} - \Theta_n$, we can deduce

$$\begin{aligned}\frac{1}{n}\left\|\Theta_n^{\perp}\left(\widetilde{\boldsymbol{\rho}_n}\sqrt{n} + \frac{\alpha\beta}{\tau_2\sqrt{\delta}}\mathbf{h}\right)\right\|_2^2 = \frac{1}{n}\left\|\widetilde{\boldsymbol{\rho}_n}\sqrt{n} + \frac{\alpha\beta}{\tau_2\sqrt{\delta}}\mathbf{h}\right\|_2^2 - \|\Theta_n\widetilde{\boldsymbol{\rho}_n}\|_2^2 - \frac{\alpha^2\beta^2}{n\tau_2^2\delta}\|\Theta_n\mathbf{h}\|_2^2 \\ - 4\frac{\alpha\beta}{\tau_2\sqrt{m}}\mathbf{h}^\top\Theta_n\widetilde{\boldsymbol{\rho}_n}. \end{aligned} \tag{3.49}$$

Since $\|\widetilde{\boldsymbol{\theta}_n^\star}\|_2 \xrightarrow{P} \zeta$ where $\zeta = \mathcal{O}(1)$ by Assumption 3.2.2, we can see that

$$\|\Theta_n\mathbf{h}\|_2^2 = \mathcal{O}(1), \quad \mathbf{h}^\top\Theta_n\widetilde{\boldsymbol{\rho}_n} = \mathbf{h}^\top\Theta_n\widetilde{\boldsymbol{\theta}_n} = \mu\mathbf{h}^\top\widetilde{\boldsymbol{\theta}_n^\star} = \mathcal{O}(1),$$

which implies that the last two terms in (3.49) vanish asymptotically and we have that,

$$\begin{aligned}\frac{1}{n}\left\|\Theta_n^{\perp}\left(\widetilde{\boldsymbol{\rho}_n}\sqrt{n} + \frac{\alpha\beta}{\tau_2\sqrt{\delta}}\mathbf{h}\right)\right\|_2^2 &= \frac{1}{n}\left\|\widetilde{\boldsymbol{\rho}_n}\sqrt{n} + \frac{\alpha\beta}{\tau_2\sqrt{\delta}}\mathbf{h}\right\|_2^2 - \|\Theta_n\widetilde{\boldsymbol{\rho}_n}\|_2^2 \\ &= \frac{1}{n}\left\|\widetilde{\boldsymbol{\rho}_n}\sqrt{n} + \frac{\alpha\beta}{\tau_2\sqrt{\delta}}\mathbf{h}\right\|_2^2 - \mu^2\left\|\widetilde{\boldsymbol{\theta}_n^\star}\right\|_2^2. \end{aligned}$$

The last line is due to the constraint in (3.48) i.e., $\Theta_n\widetilde{\boldsymbol{\rho}_n} = \Theta_n\widetilde{\boldsymbol{\theta}_n}$ (or equivalently

$\frac{\langle\widetilde{\boldsymbol{\theta}_n^\star},\widetilde{\boldsymbol{\rho}_n}\rangle}{\|\widetilde{\boldsymbol{\theta}_n^\star}\|_2^2}=\mu$, based on the definition of Θ_n and μ). Therefore by plugging this in (3.48) and introducing the Lagrangian multiplier $\eta\in\mathbb{R}$, (3.48) can be equivalently rewritten as follows,

$$\begin{aligned}
&\min_{\substack{\widetilde{\boldsymbol{\rho}_n},\Theta_n\widetilde{\boldsymbol{\theta}_n}\in\mathbb{R}^n,\mathbf{v}\in\mathbb{R}^m,w,\alpha,\tau_1\in\mathbb{R}_+\\ \text{s.t. } \langle\widetilde{\boldsymbol{\theta}_n^\star},\widetilde{\boldsymbol{\rho}_n}\rangle\Big/\|\widetilde{\boldsymbol{\theta}_n^\star}\|_2^2=\mu}} \max_{\gamma,\tau_2,\beta\in\mathbb{R}_+} \frac{\mathbf{1}_m^\top}{m}(\mathbf{v}-w\mathbf{1}_m)-\gamma w+\frac{\varepsilon_{\mathrm{tr}}\gamma}{\sqrt{n}}\left\|\boldsymbol{\Sigma}_n^{-1/2}\widetilde{\boldsymbol{\rho}_n}\right\|_1\\
&\qquad +r\left\|\boldsymbol{\Sigma}_n^{-1/2}\widetilde{\boldsymbol{\rho}_n}\right\|_2^2+\frac{\tau_2}{2n\alpha}\left\|\widetilde{\boldsymbol{\rho}_n}\sqrt{n}+\frac{\alpha\beta}{\tau_2\sqrt{\delta}}\mathbf{h}\right\|_2^2-\frac{\mu^2\tau_2\left\|\widetilde{\boldsymbol{\theta}_n^\star}\right\|_2^2}{2\alpha}-\frac{\alpha\beta^2}{2\delta\tau_2}-\frac{\alpha\tau_2}{2}\\
&\qquad +\frac{\beta}{2\tau_1 m}\left\|-\mathbf{v}+Y\bar{X}\Theta_n\widetilde{\boldsymbol{\theta}_n}+\alpha\mathbf{g}\right\|_2^2+\frac{\beta\tau_1}{2}\\
=&\min_{\substack{\widetilde{\boldsymbol{\rho}_n}\in\mathbb{R}^n,\mathbf{v}\in\mathbb{R}^m,\\ w,\alpha,\tau_1\in\mathbb{R}_+,\mu\in\mathbb{R}}} \max_{\gamma,\tau_2,\beta\in\mathbb{R}_+,\eta\in\mathbb{R}} \frac{\mathbf{1}_m^\top}{m}(\mathbf{v}-w\mathbf{1}_m)-\gamma w+\frac{\varepsilon_{\mathrm{tr}}\gamma}{\sqrt{n}}\left\|\boldsymbol{\Sigma}_n^{-1/2}\widetilde{\boldsymbol{\rho}_n}\right\|_1+r\left\|\boldsymbol{\Sigma}_n^{-1/2}\widetilde{\boldsymbol{\rho}_n}\right\|_2^2\\
&\quad +\frac{\tau_2}{2n\alpha}\left\|\widetilde{\boldsymbol{\rho}_n}\sqrt{n}+\frac{\alpha\beta}{\tau_2\sqrt{\delta}}\mathbf{h}\right\|_2^2-\frac{\mu^2\tau_2\left\|\widetilde{\boldsymbol{\theta}_n^\star}\right\|_2^2}{2\alpha}-\frac{\alpha\beta^2}{2\delta\tau_2}-\frac{\alpha\tau_2}{2}+\frac{\beta}{2\tau_1 m}\left\|-\mathbf{v}+\mu Y\bar{X}\widetilde{\boldsymbol{\theta}_n^\star}+\alpha\mathbf{g}\right\|_2^2\\
&\quad +\frac{\beta\tau_1}{2}+\eta\left(\mu-\frac{\langle\widetilde{\boldsymbol{\theta}_n^\star},\widetilde{\boldsymbol{\rho}_n}\rangle}{\|\widetilde{\boldsymbol{\theta}_n^\star}\|_2^2}\right). \qquad (3.50)
\end{aligned}$$

Minimization w.r.t $\mathbf{v}$ can be written based on the moreau-envelope of :

$$\begin{aligned}
&\min_{\mathbf{v}\in\mathbb{R}^m}\frac{\mathbf{1}_m^\top}{m}(\mathbf{v}-w\mathbf{1}_m)+\frac{\beta}{2\tau_1 m}\left\|-\mathbf{v}+\mu Y\bar{X}\widetilde{\boldsymbol{\theta}_n^\star}+\alpha\mathbf{g}\right\|_2^2\\
&\qquad =\frac{1}{m}\mathcal{M}\left(\mu Y\bar{X}\widetilde{\boldsymbol{\theta}_n^\star}+\alpha\mathbf{g}-w\mathbf{1}_m;\frac{\tau_1}{\beta}\right). \qquad (3.51)
\end{aligned}$$

Our final key step is to write the minimization with respect to $\widetilde{\boldsymbol{\rho}_n}\in\mathbb{R}^n$ based on the Moreau-envelope of the $\boldsymbol{\ell}_1+\boldsymbol{\ell}_2^2$ norms. To this end, we rewrite the terms in (3.50)

consisting of $\widetilde{\boldsymbol{\rho}_n}$ as following,

$$\min_{\widetilde{\boldsymbol{\rho}_n}\in\mathbb{R}^n} \frac{\varepsilon_{\mathrm{tr}}\gamma}{\sqrt{n}} \left\|\boldsymbol{\Sigma}_n^{-1/2}\widetilde{\boldsymbol{\rho}_n}\right\|_1 + r\left\|\boldsymbol{\Sigma}_n^{-1/2}\widetilde{\boldsymbol{\rho}_n}\right\|_2^2 + \frac{\tau_2}{2n\alpha}\left\|\widetilde{\boldsymbol{\rho}_n}\sqrt{n} + \frac{\alpha\beta}{\tau_2\sqrt{\delta}}\mathbf{h}\right\|_2^2 - \eta\frac{\left\langle\widetilde{\boldsymbol{\theta}_n^\star}, \widetilde{\boldsymbol{\rho}_n}\right\rangle}{\left\|\widetilde{\boldsymbol{\theta}_n^\star}\right\|_2^2}$$

$$= \min_{\widetilde{\boldsymbol{\rho}_n}\in\mathbb{R}^n} \frac{\varepsilon_{\mathrm{tr}}\gamma}{\sqrt{n}} \left\|\boldsymbol{\Sigma}_n^{-1/2}\widetilde{\boldsymbol{\rho}_n}\right\|_1 + r\left\|\boldsymbol{\Sigma}_n^{-1/2}\widetilde{\boldsymbol{\rho}_n}\right\|_2^2 + \frac{\tau_2}{2\alpha n}\left\|\widetilde{\boldsymbol{\rho}_n}\sqrt{n} + \frac{\alpha\beta}{\tau_2\sqrt{\delta}}\mathbf{h} - \frac{\eta\alpha\sqrt{n}}{\tau_2\left\|\widetilde{\boldsymbol{\theta}_n^\star}\right\|_2^2}\widetilde{\boldsymbol{\theta}_n^\star}\right\|_2^2$$

$$- \frac{\eta^2\alpha}{2\tau_2\left\|\widetilde{\boldsymbol{\theta}_n^\star}\right\|_2^2} - \frac{\alpha\beta\eta}{\sqrt{m}\tau_2\left\|\widetilde{\boldsymbol{\theta}_n^\star}\right\|_2^2}\left\langle\widetilde{\boldsymbol{\theta}_n^\star}, \mathbf{h}\right\rangle$$

$$= \min_{\widetilde{\boldsymbol{\rho}_n}\in\mathbb{R}^n} \frac{\varepsilon_{\mathrm{tr}}\gamma}{\sqrt{n}} \left\|\boldsymbol{\Sigma}_n^{-1/2}\widetilde{\boldsymbol{\rho}_n}\right\|_1 + r\left\|\boldsymbol{\Sigma}_n^{-1/2}\widetilde{\boldsymbol{\rho}_n}\right\|_2^2 + \frac{\tau_2}{2\alpha n}\left\|\widetilde{\boldsymbol{\rho}_n}\sqrt{n} + \frac{\alpha\beta}{\tau_2\sqrt{\delta}}\mathbf{h} - \frac{\eta\alpha\sqrt{n}}{\tau_2\left\|\widetilde{\boldsymbol{\theta}_n^\star}\right\|_2^2}\widetilde{\boldsymbol{\theta}_n^\star}\right\|_2^2$$

$$- \frac{\eta^2\alpha}{2\tau_2\left\|\widetilde{\boldsymbol{\theta}_n^\star}\right\|_2^2} \tag{3.52}$$

$$= \min_{\widetilde{\boldsymbol{\rho}_n}\in\mathbb{R}^n} \frac{\varepsilon_{\mathrm{tr}}\gamma}{n} \left\|\widetilde{\boldsymbol{\rho}_n}\sqrt{n}\right\|_1 + \frac{r}{n}\left\|\widetilde{\boldsymbol{\rho}_n}\sqrt{n}\right\|_2^2$$
$$+ \frac{\tau_2}{2\alpha n}\left\|\boldsymbol{\Sigma}_n^{1/2}\left(\widetilde{\boldsymbol{\rho}_n}\sqrt{n} + \frac{\alpha\beta}{\tau_2\sqrt{\delta}}\boldsymbol{\Sigma}_n^{-1/2}\mathbf{h} - \frac{\eta\alpha\sqrt{n}}{\tau_2\left\|\widetilde{\boldsymbol{\theta}_n^\star}\right\|_2^2}\boldsymbol{\Sigma}_n^{-1/2}\widetilde{\boldsymbol{\theta}_n^\star}\right)\right\|_2^2 - \frac{\eta^2\alpha}{2\tau_2\left\|\widetilde{\boldsymbol{\theta}_n^\star}\right\|_2^2}. \tag{3.53}$$

Here, the first step follows from the completion of squares while the second step follows from the fact that $\widetilde{\boldsymbol{\theta}_n^\star}^\top\mathbf{h} = \mathcal{O}(\|\widetilde{\boldsymbol{\theta}_n^\star}\|) = \mathcal{O}(1)$ and thus the last term asymptotically vanishes. Now, note that the minimization w.r.t. $\widetilde{\boldsymbol{\rho}_n}$ in (3.53) is equivalent to the following Moreau-Envelope function:

$$\frac{\varepsilon_{\mathrm{tr}}\gamma}{n}\mathcal{M}_{\left(\boldsymbol{\ell}_1+\frac{r}{\varepsilon_{\mathrm{tr}}\gamma}\boldsymbol{\ell}_2^2,\,\boldsymbol{\Sigma}_n\right)}\left(\frac{\alpha\beta}{\tau_2\sqrt{\delta}}\boldsymbol{\Sigma}_n^{-1/2}\mathbf{h} + \frac{\alpha\eta\sqrt{n}}{\tau_2\left\|\widetilde{\boldsymbol{\theta}_n^\star}\right\|_2^2}\boldsymbol{\theta}_n^\star; \frac{\alpha\varepsilon_{\mathrm{tr}}\gamma}{\tau_2}\right),$$

where recall the definition of $\mathcal{M}_{\left(\boldsymbol{\ell}_1+C\,\boldsymbol{\ell}_2^2,\mathbf{s}\right)}(\mathbf{x};\tau)$ in (3.23). Thus, the following objective function is derived by replacing the Moreau-envelopes in (3.50),

$$\min_{\substack{\alpha,\tau_1,w\in\mathbb{R}_+,\\ \mu\in\mathbb{R}}}\ \max_{\substack{\tau_2,\beta,\gamma\in\mathbb{R}_+,\\ \eta\in\mathbb{R}}}\ -\gamma w-\frac{\mu^2\tau_2}{2\alpha}\left\|\widetilde{\boldsymbol{\theta}_n^\star}\right\|_2^2-\frac{\alpha\beta^2}{2\delta\tau_2}-\frac{\alpha\tau_2}{2}+\frac{\beta\tau_1}{2}+\eta\mu-\frac{\eta^2\alpha}{2\tau_2\left\|\widetilde{\boldsymbol{\theta}_n^\star}\right\|_2^2}$$
$$+\frac{1}{m}\mathcal{M}\left(\mu Y\bar{X}\widetilde{\boldsymbol{\theta}_n^\star}+\alpha\mathbf{g}-w\mathbf{1}_m;\frac{\tau_1}{\beta}\right)$$
$$+\frac{\varepsilon_{\mathrm{tr}}\gamma}{n}\mathcal{M}_{\left(\boldsymbol{\ell}_1+\frac{r}{\varepsilon_{\mathrm{tr}}\gamma}\boldsymbol{\ell}_2^2,\,\boldsymbol{\Sigma}_n\right)}\left(\frac{\alpha\beta}{\tau_2\sqrt{\delta}}\boldsymbol{\Sigma}_n^{-1/2}\mathbf{h}+\frac{\alpha\eta\sqrt{n}}{\tau_2\left\|\widetilde{\boldsymbol{\theta}_n^\star}\right\|_2^2}\boldsymbol{\theta}_n^\star;\frac{\alpha\varepsilon_{\mathrm{tr}}\gamma}{\tau_2}\right). \quad (3.54)$$

We note that based on the definition of Θ_n the entry i on the diagonal of Y, denoted by y_i is derived as $y_i=\psi\left(\langle\mathbf{x}_i,\boldsymbol{\theta}_n^\star\rangle\right)=\psi(\langle\bar{\mathbf{x}}_i,\widetilde{\boldsymbol{\theta}_n^\star}\rangle)$, where

$$\left\langle\bar{\mathbf{x}}_i,\widetilde{\boldsymbol{\theta}_n^\star}\right\rangle\sim\mathcal{N}\left(0,\boldsymbol{\theta}_n^{\star\top}\boldsymbol{\Sigma}_n\boldsymbol{\theta}_n^\star\right).$$

Therefore it yields that

$$\mu Y\bar{X}\widetilde{\boldsymbol{\theta}_n^\star}\xrightarrow{P}\mu\zeta\,\left(\mathbf{s}\odot\Psi(\zeta\mathbf{s})\right),$$

where $\Psi(\zeta\mathbf{s})\triangleq[\psi(\zeta s_1);\cdots;\psi(\zeta s_m)]$ for the vector $\mathbf{s}\in\mathbb{R}^m$ with i.i.d standard normal entries s_i and by Assumption 3.2.2, ζ denotes the high-dimensional limit of $\boldsymbol{\theta}_n^{\star\top}\boldsymbol{\Sigma}_n\boldsymbol{\theta}_n^\star$. Therefore based on the separability of the Moreau-envelope $\mathcal{M}$ we have,

$$\frac{1}{m}\mathcal{M}\left(\mu Y\bar{X}\widetilde{\boldsymbol{\theta}_n^\star}+\alpha\mathbf{g}-w\mathbf{1}_m;\frac{\tau_1}{\beta}\right)\xrightarrow{P}\mathbb{E}_{S,G}\left[\mathcal{M}_{\mathcal{L}}\left(\alpha G+\mu\zeta S\cdot\psi(\zeta S)-w;\frac{\tau_1}{\beta}\right)\right],$$

for $S, G \overset{\text{iid}}{\sim} \mathcal{N}(0,1)$. Also, it holds that,

$$\frac{1}{n}\mathcal{M}_{\left(\boldsymbol{\ell}_1+\frac{r}{\varepsilon_{\text{tr}}\gamma}\boldsymbol{\ell}_2^2,\,\boldsymbol{\Sigma}_n\right)}\left(\frac{\alpha\beta}{\tau_2\sqrt{\delta}}\boldsymbol{\Sigma}_n^{-1/2}\mathbf{h}+\frac{\alpha\eta\sqrt{n}}{\tau_2\left\|\widetilde{\boldsymbol{\theta}_n^\star}\right\|_2^2}\boldsymbol{\theta}_n^\star;\frac{\alpha\varepsilon_{\text{tr}}\gamma}{\tau_2}\right)\xrightarrow{P}\bar{M}\left(\frac{\alpha\beta}{\tau_2\sqrt{\delta}},\frac{\alpha\eta}{\tau_2\zeta^2};\frac{\alpha\gamma\varepsilon_{\text{tr}}}{\tau_2}\right). \tag{3.55}$$

Putting these back in (3.54), we conclude with the objective in (3.25). This completes the proof.

Case II: Correlated Features with Diagonal Covariance Matrix (Proof of Theorem 3.3.1)

Note that the Moreau-envelope in (3.55) is not separable in general and thus the computation of $\mathcal{M}_{\left(\boldsymbol{\ell}_1+\frac{r}{\varepsilon_{\text{tr}}\gamma}\boldsymbol{\ell}_2^2,\boldsymbol{\Sigma}_n\right)}$ may not be simplified further. By assuming $\boldsymbol{\Sigma}_n$ to be diagonal i.e., $\boldsymbol{\Sigma}_n=\boldsymbol{\Lambda}_n$ with diagonal entries $\lambda_{n,i}$, $i\in[n]$, it is concluded from (3.53) that the minimization becomes separable over the entires of $\widetilde{\boldsymbol{\rho}_n}$. In fact, it is inferred that in this case:

$$\begin{aligned}&\frac{1}{n}\mathcal{M}_{\left(\boldsymbol{\ell}_1+\frac{r}{\varepsilon_{\text{tr}}\gamma}\boldsymbol{\ell}_2^2,\,\boldsymbol{\Sigma}_n\right)}\left(\frac{\alpha\beta}{\tau_2\sqrt{\delta}}\boldsymbol{\Sigma}_n^{-1/2}\mathbf{h}+\frac{\alpha\eta\sqrt{n}}{\tau_2\left\|\widetilde{\boldsymbol{\theta}_n^\star}\right\|_2^2}\boldsymbol{\theta}_n^\star;\frac{\alpha\varepsilon_{\text{tr}}\gamma}{\tau_2}\right)\\&\qquad=\frac{1}{n}\sum_{i=1}^{n}\mathcal{M}_{\ell_1+\frac{r}{\varepsilon_{\text{tr}}\gamma}\ell_2^2}\left(\frac{\alpha\beta}{\tau_2\sqrt{\delta\lambda_{n,i}}}\mathbf{h}_i+\frac{\alpha\eta\sqrt{n}}{\tau_2\left\|\widetilde{\boldsymbol{\theta}_n^\star}\right\|_2^2}\boldsymbol{\theta}_{n,i}^\star;\frac{\alpha\varepsilon_{\text{tr}}\gamma}{\tau_2\lambda_{n,i}}\right).\end{aligned} \tag{3.56}$$

By Assumption 3.2.1, we know that $\lambda_{n,i}\in(c,C)$, for all $i\in[n]$ and all $n\in\mathbb{N}$, where $c>0$. This results in $\mathcal{M}_{\ell_1+\ell_2^2}(\cdot;\cdot)$ being Pseudo-Lipschitz of order 2. Thus by Assumption 3.2.3,

the expression in (3.56) converges in probability to

$$\mathbb{E}_{L,H,T}\left[\mathcal{M}_{\ell_1+\frac{r}{\varepsilon_{\rm tr}\gamma}\ell_2^2}\left(\frac{\alpha\beta}{\tau_2\sqrt{\delta L}}H+\frac{\alpha\eta}{\tau_2\zeta^2}T;\frac{\alpha\varepsilon_{\rm tr}\gamma}{\tau_2 L}\right)\right], \tag{3.57}$$

for the standard Gaussian random variable H and (L,T) drawn according to distribution Π. Thus in this case, (3.54) converges to the min-max problem in (3.8). This completes the proof of Theorem 3.3.1 for GLM.

Case III: Isotropic Features

When $\mathbf{\Sigma}_n=\mathbb{I}_n$, the final expressions can be further simplified, as the term $\|\mathbf{\Sigma}_n\widetilde{\boldsymbol{\theta}_n}\|_2^2$ becomes decomposable into $\|\Theta_n\widetilde{\boldsymbol{\theta}_n}\|_2^2$ and $\|\Theta_n^{\perp}\widetilde{\boldsymbol{\theta}_n}\|_2^2$. Here we focus on the case of $q=\infty$ for GLM and defer the analysis of $q=2$ to Section 3.8.2. Proceeding with the same notation as in (3.8), consider the following min-max objective,

$$\begin{aligned}\min_{\substack{\alpha,\tau_1,w\in\mathbb{R}_+,\\ \mu\in\mathbb{R}}}\ \max_{\substack{\tau_2,\beta,\gamma\in\mathbb{R}_+,\\ \eta\in\mathbb{R}}}\ & f_{\delta,1}(\bar{\mathbf{v}})+r\alpha^2+r\mu^2+\mathbb{E}_{G,S}\left[\mathcal{M}_{\mathcal{L}}\left(\alpha G+\mu S\,\psi(S)-w;\frac{\tau_1}{\beta}\right)\right]\\ &+\varepsilon_{\rm tr}\gamma\,\mathbb{E}_{H,T}\left[\mathcal{M}_{\ell_1}\left(\frac{\alpha\beta}{\tau_2\sqrt{\delta}}H+\frac{\alpha\eta}{\tau_2}T;\frac{\alpha\varepsilon_{\rm tr}\gamma}{\tau_2}\right)\right].\end{aligned} \tag{3.58}$$

Corollary 3.8.1. *Consider the Generalized Linear models (3.3). Assume the same settings and assumptions as in Theorem 3.8.1, only here assume that $\mathbf{\Sigma}_n=\mathbb{I}_n$. Then, the high-dimensional limit for the adversarial test error $(\mathcal{E}_{\ell_\infty,\frac{\varepsilon_{\rm ts}}{\sqrt{n}}})$ is derived as follows,*

$$\left\{\mathcal{E}^{GLM}_{\ell_\infty,\frac{\varepsilon_{\rm ts}}{\sqrt{n}}}\left(\widehat{\boldsymbol{\theta}}_n\right)\right\}\xrightarrow{P}\mathbb{P}\Big(\mu^\star\,S\,\psi(S)+\alpha^\star G<w^\star\varepsilon_{\rm ts}/\varepsilon_{\rm tr}\Big), \tag{3.59}$$

where $(\alpha^\star,\mu^\star,w^\star)$ is the unique solution to the scalar minimax problem (3.58).

Proof: The proof follows the same steps as Theorem 3.8.1. Note here that $\zeta=1$ and the random variable $L=1$. Also, in deriving (3.41), it suffices to write the Lagrangian

equivalent formulation only for the ℓ_1 loss and write $\|\widetilde{\boldsymbol{\theta}_n}\|_2^2 = \|\Theta_n\widetilde{\boldsymbol{\theta}_n}\|_2^2 + \|\Theta_n^{\perp}\widetilde{\boldsymbol{\theta}_n}\|_2^2 \xrightarrow{P} r\alpha^2 + r\mu^2$, which results in (3.58).

A System of Equations.

We find solutions to the min-max problem in (3.58)(the objective of which we denote by $\bar{L} : \mathbb{R}^8 \to \mathbb{R}$) by forming and solving $\nabla_{\bar{\mathbf{v}}}\bar{L} = \mathbf{0}$. To compute $\nabla_{\bar{\mathbf{v}}}\bar{L}$ we leverage properties of Moreau-envelopes and appropriately combine different equations in the system $\nabla_{\bar{\mathbf{v}}}\bar{L} = \mathbf{0}$, so as to simplify the resulting expressions (details are provided below). This leads to the system of eight equations (3.60). Our experiments suggest that the simplifications that lead to these, are important for a simple iterative fixed-point scheme to obtain the

theoretical values of $(\alpha^\star, \mu^\star, w^\star)$.

$$
\begin{cases}
\eta = \frac{\beta}{\tau_1}\mathbb{E}[Z\left(w + \mathrm{prox}_{\mathcal{L}}\left(\mu Z + \alpha G - w; \tau_1/\beta\right)\right)] - 2\lambda\mu + \frac{\mu\tau_2}{\alpha} - \kappa\frac{\beta\mu}{\tau_1}, \\
\mu = \mathbb{E}\left[T \cdot \mathrm{prox}_{\ell_1}\left(\alpha\beta H/(\tau_2\sqrt{\delta}) + \alpha\eta T/\tau_2; \alpha\varepsilon_{\mathrm{tr}}\gamma/\tau_2\right)\right], \\
\gamma = -\mathbb{E}\left[\mathcal{M}'_{\mathcal{L},1}\left(\mu Z + \alpha G - w; \tau_1/\beta\right)\right], \\
\beta^2 = \mathbb{E}\Big[\Big(\mathcal{M}'_{\mathcal{L},1}\left(\mu Z + \alpha G - w; \tau_1/\beta\right)\Big)^2\Big], \\
\tau_1 = \frac{1}{\sqrt{\delta}}\mathbb{E}\left[H \cdot \mathrm{prox}_{\ell_1}\left(\frac{\alpha\beta H}{\tau_2\sqrt{\delta}} + \alpha\eta T/\tau_2; \alpha\varepsilon_{\mathrm{tr}}\gamma/\tau_2\right)\right], \\
\alpha = \frac{1}{\beta + 2\lambda\tau_1}(\tau_1\tau_2 + \beta\mathbb{E}\left[G\mathrm{prox}_{\mathcal{L}}\left(\alpha G + \mu Z - w; \tau_1/\beta\right)\right]), \\
w = \varepsilon_{\mathrm{tr}}\mathbb{E}\left[\mathcal{M}_{\ell_1}\left(\frac{\alpha\beta H}{\tau_2\sqrt{\delta}} + \alpha\eta T/\tau_2; \alpha\varepsilon_{\mathrm{tr}}\gamma/\tau_2\right)\right] \\
\qquad\qquad -\frac{\varepsilon_{\mathrm{tr}}^2\gamma\alpha}{2\tau_2}\mathbb{E}\Big[\Big(\mathcal{M}'_{\ell_1,1}\left(\frac{\alpha\beta H}{\tau_2\sqrt{\delta}} + \alpha\eta T/\tau_2; \alpha\varepsilon_{\mathrm{tr}}\gamma/\tau_2\right)\Big)^2\Big], \\
\tau_2^2 = \frac{\alpha^2}{\alpha^2+\mu^2}\Big(\beta^2/\delta + \eta^2 + \varepsilon_{\mathrm{tr}}^2\gamma^2\mathbb{E}\Big[\Big(\mathcal{M}'_{\ell_1,1}\left(\frac{\alpha\beta H}{\tau_2\sqrt{\delta}} + \alpha\eta T/\tau_2; \alpha\varepsilon_{\mathrm{tr}}\gamma/\tau_2\right)\Big)^2\Big] \\
\qquad\qquad -\frac{2\beta\varepsilon_{\mathrm{tr}}\gamma}{\sqrt{\delta}}\mathbb{E}\left[H\mathcal{M}'_{\ell_1,1}\left(\frac{\alpha\beta H}{\tau_2\sqrt{\delta}} + \alpha\eta T/\tau_2; \alpha\varepsilon_{\mathrm{tr}}\gamma/\tau_2\right)\right] \\
\qquad\qquad -2\eta\varepsilon_{\mathrm{tr}}\gamma\mathbb{E}\left[T\mathcal{M}'_{\ell_1,1}\left(\frac{\alpha\beta H}{\tau_2\sqrt{\delta}} + \alpha\eta T\tau_2; \alpha\varepsilon_{\mathrm{tr}}\gamma/\tau_2\right)\right]\Big),
\end{cases}
\tag{3.60}
$$

where the random variable $Z = S\,\psi(S)$ for GLM and $Z = S + 1$ for GMM, and the constant $\kappa = 1$ and 2 for GLM and GMM, respectively. Here, the Proximal operator of a function $f : \mathbb{R} \to \mathbb{R}$, at x with parameter $\kappa > 0$, is defined as follows,

$$
\mathrm{prox}_f\left(x; \kappa\right) \triangleq \arg\min_v \frac{1}{2\kappa}(x - v)^2 + f(v). \tag{3.61}
$$

Next, we explain how to derive the Equations (3.60) for GLM. The approach for GMM is similar. Before starting, we recall useful properties of Moreau-envelops which we will leverage in deriving the equations.

Proposition 3.8.1 ([2]). *Let $\mathcal{L}$ be a lower semi-continuous and proper function. Denote $\mathcal{M}'_{\mathcal{L},1}(x;\kappa) \triangleq \frac{\partial \mathcal{M}_{\mathcal{L}}(x;\kappa)}{\partial x}$. and $\mathcal{M}'_{\mathcal{L},2}(x;\kappa) \triangleq \frac{\partial \mathcal{M}_{\mathcal{L}}(x;\kappa)}{\partial \kappa}$. Then the following relations hold between first-order derivatives of Moreau-envelopes and the corresponding proximal operator,*

$$\mathcal{M}'_{\mathcal{L},1}(x;\tau) = \frac{1}{\tau}(x - \mathrm{prox}_{\mathcal{L}}(x;\tau)), \tag{3.62}$$

$$\mathcal{M}'_{\mathcal{L},2}(x;\tau) = -\frac{1}{2\tau^2}(x - \mathrm{prox}_{\mathcal{L}}(x;\tau))^2. \tag{3.63}$$

We proceed with the derivation of the Equations (3.60). First, we start with $\nabla_\mu L$ to find that,

$$\begin{aligned}\nabla_\mu \bar{L} &= -\frac{\mu\tau_2}{\alpha} + \eta + \mathbb{E}\left[S\psi(S)\cdot \mathcal{M}'_{\mathcal{L},1}\left(\mu S\psi(S) + \alpha G - w; \frac{\tau_1}{\beta}\right)\right] + 2\lambda\mu \\ &= -\frac{\mu\tau_2}{\alpha} + \eta + \frac{\beta}{\tau_1}\left(\mu - w\mathbb{E}\left[S\psi(S)\right] - \mathbb{E}\left[S\psi(S)\cdot \mathrm{prox}_{\mathcal{L}}\left(\mu S\psi(S) + \alpha G - w; \frac{\tau_1}{\beta}\right)\right]\right) \\ &\quad + 2\lambda\mu,\end{aligned}$$

which gives rise to the equation below for finding $\eta^\star$:

$$\eta = \frac{\mu\tau_2}{\alpha} - \frac{\beta\mu}{\tau_1} + \frac{\beta w}{\tau_1}\mathbb{E}[S\psi(S)] + \frac{\beta}{\tau_1}\mathbb{E}\left[S\psi(S)\cdot \mathrm{prox}_{\mathcal{L}}\left(\mu S\psi(S) + \alpha G - w; \frac{\tau_1}{\beta}\right)\right] - 2\lambda\mu. \tag{3.64}$$

By taking derivative w.r.t η and rewriting the derivatives based on proximal operators, we derive the equation for $\mu^\star$:

$$\begin{aligned}\nabla_\eta \bar{L} &= \mu - \frac{\eta\alpha}{\tau_2} + \frac{\varepsilon_{\mathrm{tr}}\gamma\alpha}{\tau_2}\mathbb{E}\left[T\mathcal{M}'_{\ell_1,1}\left(\frac{\alpha\beta}{\tau_2\sqrt{\delta}}H + \frac{\alpha\eta}{\tau_2}T; \frac{\alpha\varepsilon_{\mathrm{tr}}\gamma}{\tau_2}\right)\right] \\ &= \mu - \frac{\eta\alpha}{\tau_2} + \frac{\alpha\eta}{\tau_2}\mathbb{E}\left[T^2\right] - \mathbb{E}\left[T\cdot \mathrm{prox}_{\ell_1}\left(\frac{\alpha\beta}{\tau_2\sqrt{\delta}}H + \frac{\alpha\eta}{\tau_2}T; \frac{\alpha\varepsilon_{\mathrm{tr}}\gamma}{\tau_2}\right)\right],\end{aligned}$$

which after noting that $\mathbb{E}[Z^2] = 1$, yields the following equation:

$$\mu = \mathbb{E}\left[T \cdot \mathrm{prox}_{\ell_1}\left(\frac{\alpha\beta}{\tau_2\sqrt{\delta}}H + \frac{\alpha\eta}{\tau_2}T; \frac{\alpha\varepsilon_{\mathrm{tr}}\gamma}{\tau_2}\right)\right]. \tag{3.65}$$

In order to find $\gamma^\star$, we consider $\nabla_w \bar{L}$ to derive that:

$$\gamma = -\mathbb{E}\left[\mathcal{M}'_{\mathcal{L},1}\left(\mu S\psi(S) + \alpha G - w; \frac{\tau_1}{\beta}\right)\right] \tag{3.66}$$

To proceed, we derive $\nabla_{\tau_1}\bar{L}$ and $\nabla_\beta \bar{L}$:

$$\begin{aligned}\nabla_{\tau_1}\bar{L} &= \frac{\beta}{2} + \frac{1}{\beta}\mathbb{E}\left[\mathcal{M}'_{\mathcal{L},2}\left(\mu S\psi(S) + \alpha G - w; \frac{\tau_1}{\beta}\right)\right] \\ &= \frac{\beta}{2} - \frac{1}{2\beta}\mathbb{E}\left[\left(\mathcal{M}'_{\mathcal{L},1}\left(\mu S\psi(S) + \alpha G - w; \frac{\tau_1}{\beta}\right)\right)^2\right]\end{aligned} \tag{3.67}$$

$$\begin{aligned}\nabla_\beta \bar{L} &= -\frac{\alpha\beta}{\delta\tau_2} + \frac{\tau_1}{2} - \frac{\tau_1}{\beta^2}\mathbb{E}\left[\mathcal{M}'_{\mathcal{L},2}\left(\mu S\psi(S) + \alpha G - w; \frac{\tau_1}{\beta}\right)\right] \\ &\quad + \frac{\alpha\varepsilon_{\mathrm{tr}}\gamma}{\tau_2\sqrt{\delta}}\mathbb{E}\left[H \cdot \mathcal{M}'_{\ell_1,1}\left(\frac{\alpha\beta}{\tau_2\sqrt{\delta}}H + \frac{\alpha\eta}{\tau_2}T; \frac{\alpha\varepsilon_{\mathrm{tr}}\gamma}{\tau_2}\right)\right] \\ &= -\frac{\alpha\beta}{\delta\tau_2} - \frac{\tau_1}{2} + \frac{\tau_1}{2\beta^2}\mathbb{E}\left[\left(\mathcal{M}'_{\mathcal{L},1}\left(\mu S\psi(S) + \alpha G - w; \frac{\tau_1}{\beta}\right)\right)^2\right] \\ &\quad + \frac{\alpha\varepsilon_{\mathrm{tr}}\gamma}{\tau_2\sqrt{\delta}}\mathbb{E}\left[H \cdot \mathcal{M}'_{\ell_1,1}\left(\frac{\alpha\beta}{\tau_2\sqrt{\delta}}H + \frac{\alpha\eta}{\tau_2}T; \frac{\alpha\varepsilon_{\mathrm{tr}}\gamma}{\tau_2}\right)\right].\end{aligned} \tag{3.68}$$

(3.67) yields the equation for deriving β i.e.,

$$\beta = \left(\mathbb{E}\left[\left(\mathcal{M}'_{\mathcal{L},1}\left(\mu S\psi(S) + \alpha G - w; \frac{\tau_1}{\beta}\right)\right)^2\right]\right)^{1/2}. \tag{3.69}$$

Next, we combine (3.67) with (3.68) with proper coefficients to simplify the equations

yielding,

$$\begin{aligned}\frac{\nabla_{\tau_1}\bar{L}}{\beta}+\frac{\nabla_{\beta}\bar{L}}{\tau_1}&=1-\frac{\alpha\beta}{\delta\tau_1\tau_2}+\frac{\alpha\varepsilon_{\mathrm{tr}}\gamma}{\sqrt{\delta}\tau_1\tau_2}\mathbb{E}\left[H\cdot\mathcal{M}'_{\ell_1,1}\left(\frac{\alpha\beta}{\tau_2\sqrt{\delta}}H+\frac{\alpha\eta}{\tau_2}T;\frac{\alpha\varepsilon_{\mathrm{tr}}\gamma}{\tau_2}\right)\right]\\&=1-\frac{\alpha\beta}{\delta\tau_1\tau_2}+\frac{1}{\sqrt{\delta}\tau_1}\left(\frac{\alpha\beta}{\tau_2\sqrt{\delta}}-\mathbb{E}\left[H\cdot\mathrm{prox}_{\ell_1}\left(\frac{\alpha\beta}{\tau_2\sqrt{\delta}}H+\frac{\alpha\eta}{\tau_2}T;\frac{\alpha\varepsilon_{\mathrm{tr}}\gamma}{\tau_2}\right)\right]\right),\end{aligned} \tag{3.70}$$

which yields the following equation:

$$\tau_1=\frac{1}{\sqrt{\delta}}\,\mathbb{E}\left[H\cdot\mathrm{prox}_{\ell_p}\left(\frac{\alpha\beta}{\tau_2\sqrt{\delta}}H+\frac{\alpha\eta}{\tau_2}T;\frac{\alpha\varepsilon_{\mathrm{tr}}\gamma}{\tau_2}\right)\right]. \tag{3.71}$$

In a similar way, we derive $\nabla_{\tau_2}\bar{L}$ and $\nabla_{\alpha}\bar{L}$:

$$\begin{aligned}\nabla_{\tau_2}\bar{L}=-\frac{r^2\mu^2}{2\alpha}+\frac{\alpha\beta^2}{2\delta\tau_2^2}-\frac{\alpha}{2}+\frac{\eta^2\alpha r^2}{2\tau_2^2}-\frac{\varepsilon_{\mathrm{tr}}^2\gamma^2\alpha}{\tau_2^2}\mathbb{E}\left[\mathcal{M}'_{\ell_1,2}\left(\frac{\alpha\beta}{\tau_2\sqrt{\delta}}H+\frac{\alpha\eta}{\tau_2}T;\frac{\alpha\varepsilon_{\mathrm{tr}}\gamma}{\tau_2}\right)\right]\\-\frac{\alpha\beta\varepsilon_{\mathrm{tr}}\gamma}{\tau_2^2\sqrt{\delta}}\mathbb{E}\left[H\cdot\mathcal{M}'_{\ell_1,1}\left(\frac{\alpha\beta}{\tau_2\sqrt{\delta}}H+\frac{\alpha\eta}{\tau_2}T;\frac{\alpha\varepsilon_{\mathrm{tr}}\gamma}{\tau_2}\right)\right]\\-\frac{\alpha\eta\varepsilon_{\mathrm{tr}}\gamma}{\tau_2^2}\mathbb{E}\left[T\cdot\mathcal{M}'_{\ell_1,1}\left(\frac{\alpha\beta}{\tau_2\sqrt{\delta}}H+\frac{\alpha\eta}{\tau_2}T;\frac{\alpha\varepsilon_{\mathrm{tr}}\gamma}{\tau_2}\right)\right],\end{aligned} \tag{3.72}$$

$$\begin{aligned}\nabla_{\alpha}\bar{L}=\frac{\mu^2\tau_2}{2\alpha^2}-\frac{\beta^2}{2\delta\tau_2}-\frac{\tau_2}{2}-\frac{\eta^2}{2\tau_2}+\mathbb{E}\left[G\cdot\mathcal{M}'_{\mathcal{L},1}\left(\alpha G+\mu S\psi(S)-w;\frac{\tau_1}{\beta}\right)\right]\\+\frac{\varepsilon_{\mathrm{tr}}^2\gamma^2}{\tau_2}\mathbb{E}\left[\mathcal{M}'_{\ell_1,2}\left(\frac{\alpha\beta}{\tau_2\sqrt{\delta}}H+\frac{\alpha\eta}{\tau_2}T;\frac{\alpha\varepsilon_{\mathrm{tr}}\gamma}{\tau_2}\right)\right]+2\lambda\alpha\\+\frac{\beta\varepsilon_{\mathrm{tr}}\gamma}{\tau_2\sqrt{\delta}}\mathbb{E}\left[H\cdot\mathcal{M}'_{\ell_1,1}\left(\frac{\alpha\beta}{\tau_2\sqrt{\delta}}H+\frac{\alpha\eta}{\tau_2}T;\frac{\alpha\varepsilon_{\mathrm{tr}}\gamma}{\tau_2}\right)\right]\\+\frac{\eta\varepsilon_{\mathrm{tr}}\gamma}{\tau_2}\mathbb{E}\left[T\cdot\mathcal{M}'_{\ell_1,1}\left(\frac{\alpha\beta}{\tau_2\sqrt{\delta}}H+\frac{\alpha\eta}{\tau_2}T;\frac{\alpha\varepsilon_{\mathrm{tr}}\gamma}{\tau_2}\right)\right].\end{aligned} \tag{3.73}$$

First, the following equation is directly followed based on (3.72):

$$\begin{aligned}\tau_2^2 =& \frac{2\alpha}{\alpha^2+\mu^2}\Bigg(\frac{\alpha\beta^2}{2\delta}+\frac{\eta^2\alpha}{2}+\frac{\varepsilon_{\rm tr}^2\gamma^2\alpha}{2}\mathbb{E}\left[\left(\mathcal{M}'_{\ell_1,1}\left(\frac{\alpha\beta}{\tau_2\sqrt{\delta}}H+\frac{\alpha\eta}{\tau_2}T;\frac{\alpha\varepsilon_{\rm tr}\gamma}{\tau_2}\right)\right)^2\right]\\ &-\frac{\alpha\beta\varepsilon_{\rm tr}\gamma}{\sqrt{\delta}}\mathbb{E}\left[H\cdot\mathcal{M}'_{\ell_1,1}\left(\frac{\alpha\beta}{\tau_2\sqrt{\delta}}H+\frac{\alpha\eta}{\tau_2}T;\frac{\alpha\varepsilon_{\rm tr}\gamma}{\tau_2}\right)\right]\\ &-\alpha\eta\varepsilon_{\rm tr}\gamma\mathbb{E}\left[T\cdot\mathcal{M}'_{\ell_1,1}\left(\frac{\alpha\beta}{\tau_2\sqrt{\delta}}H+\frac{\alpha\eta}{\tau_2}T;\frac{\alpha\varepsilon_{\rm tr}\gamma}{\tau_2}\right)\right]\Bigg).\end{aligned} \tag{3.74}$$

In the next step, we combine (3.72) and (3.73) to derive that,

$$\begin{aligned}\frac{\nabla_{\tau_2}\bar{L}}{\alpha}+\frac{\nabla_{\alpha}\bar{L}}{\tau_2} &= \frac{1}{\tau_2}\mathbb{E}\left[G\cdot\mathcal{M}'_{\mathcal{L},1}\left(\alpha G+\mu S\psi(S)-w;\tau_1/\beta\right)\right]-1+2\lambda\alpha/\tau_2\\ &= \frac{\beta}{\tau_1\tau_2}\Big(\alpha-\mathbb{E}\left[G\cdot\mathrm{prox}_{\mathcal{L}}\left(\alpha G+\mu S\psi(S)-w;\tau_1/\beta\right)\right]\Big)-1+2\lambda\alpha/\tau_2.\end{aligned}$$

This gives the following equation, based on the stationary point condition:

$$\alpha=\Big(\tau_1\tau_2+\beta\mathbb{E}\Big[G\cdot\mathrm{prox}_{\mathcal{L}}\left(\alpha G+\mu S\psi(S)-w;\tau_1/\beta\right)\Big]\Big)\Big/(\beta+2\lambda\tau_1). \tag{3.75}$$

Finally, the following equation is derived directly based on $\nabla_\gamma\bar{L}$,

$$\begin{aligned}w=&\varepsilon_{\rm tr}\mathbb{E}\left[\mathcal{M}_{\ell_1}\left(\frac{\alpha\beta}{\tau_2\sqrt{\delta}}H+\frac{\alpha\eta}{\tau_2}T;\frac{\alpha\varepsilon_{\rm tr}\gamma}{\tau_2}\right)\right]\\ &-\frac{\varepsilon_{\rm tr}^2\gamma\alpha}{2\tau_2}\mathbb{E}\left[\left(\mathcal{M}'_{\ell_1,1}\left(\frac{\alpha\beta}{\tau_2\sqrt{\delta}}H+\frac{\alpha\eta}{\tau_2}T;\frac{\alpha\varepsilon_{\rm tr}\gamma}{\tau_2}\right)\right)^2\right].\end{aligned} \tag{3.76}$$

By putting together the equations (3.64), (3.65), (3.66), (3.69), (3.71), (3.74), (3.75) and (3.76), we end up with the system of eight equations in (3.60) for GLM. The steps required for deriving (3.60) for GMM are in a similar fashion.

3.8.2 Proofs for Section 3.4

Case I: Correlated Features with General Covariance Matrix (Proof of Theorem 3.4.1)

First, we note that when $q = p = 2$, the term $\|\boldsymbol{\Sigma}_n^{-1/2}\widetilde{\boldsymbol{\rho}_n}\|_p$ (in (3.50)) can be rewritten as follows,

$$\left\|\boldsymbol{\Sigma}_n^{-1/2}\widetilde{\boldsymbol{\rho}_n}\right\|_2 = \min_{\tau_3\in\mathbb{R}_+} \frac{1}{2\tau_3}\left\|\boldsymbol{\Sigma}_n^{-1/2}\widetilde{\boldsymbol{\rho}_n}\right\|_2^2 + \frac{\tau_3}{2}. \tag{3.77}$$

The reason behind this reformulation is to permit the analysis of the final Moreau-envelope expression. With this, we rewrite the steps previously required to derive (3.53), as follows

$$\begin{aligned}
&\min_{\widetilde{\boldsymbol{\rho}_n}\in\mathbb{R}^n}\left(\frac{\varepsilon_{\mathrm{tr}}\gamma}{2\tau_3}+r\right)\left\|\boldsymbol{\Sigma}_n^{-1/2}\widetilde{\boldsymbol{\rho}_n}\right\|_2^2 + \frac{\tau_2}{2n\alpha}\left\|\widetilde{\boldsymbol{\rho}_n}\sqrt{n}+\frac{\alpha\beta}{\tau_2\sqrt{\delta}}\mathbf{h}\right\|_2^2 - \eta\frac{\left\langle\widetilde{\boldsymbol{\theta}_n^\star},\widetilde{\boldsymbol{\rho}_n}\right\rangle}{\left\|\widetilde{\boldsymbol{\theta}_n^\star}\right\|_2^2}\\
&=\min_{\widetilde{\boldsymbol{\rho}_n}\in\mathbb{R}^n}\left(\frac{\varepsilon_{\mathrm{tr}}\gamma}{2\tau_3}+r\right)\left\|\boldsymbol{\Sigma}_n^{-1/2}\widetilde{\boldsymbol{\rho}_n}\right\|_2^2 + \frac{\tau_2}{2\alpha n}\left\|\widetilde{\boldsymbol{\rho}_n}\sqrt{n}+\frac{\alpha\beta}{\tau_2\sqrt{\delta}}\mathbf{h}-\frac{\eta\alpha\sqrt{n}}{\tau_2\left\|\widetilde{\boldsymbol{\theta}_n^\star}\right\|_2^2}\widetilde{\boldsymbol{\theta}_n^\star}\right\|_2^2\\
&\qquad -\frac{\eta^2\alpha}{2\tau_2\left\|\widetilde{\boldsymbol{\theta}_n^\star}\right\|_2^2} - \frac{\alpha\beta\eta}{\sqrt{m}\tau_2\left\|\widetilde{\boldsymbol{\theta}_n^\star}\right\|_2^2}\left\langle\widetilde{\boldsymbol{\theta}_n^\star},\mathbf{h}\right\rangle\\
&=\min_{\widetilde{\boldsymbol{\rho}_n}\in\mathbb{R}^n}\left(\frac{\varepsilon_{\mathrm{tr}}\gamma}{2\tau_3}+r\right)\left\|\boldsymbol{\Sigma}_n^{-1/2}\widetilde{\boldsymbol{\rho}_n}\right\|_2^2 + \frac{\tau_2}{2\alpha n}\left\|\widetilde{\boldsymbol{\rho}_n}\sqrt{n}+\frac{\alpha\beta}{\tau_2\sqrt{\delta}}\mathbf{h}-\frac{\eta\alpha\sqrt{n}}{\tau_2\left\|\widetilde{\boldsymbol{\theta}_n^\star}\right\|_2^2}\widetilde{\boldsymbol{\theta}_n^\star}\right\|_2^2 - \frac{\eta^2\alpha}{2\tau_2\left\|\widetilde{\boldsymbol{\theta}_n^\star}\right\|_2^2}\\
&=\min_{\widetilde{\boldsymbol{\rho}_n}\in\mathbb{R}^n}\frac{1}{n}\left(\frac{\varepsilon_{\mathrm{tr}}\gamma}{2\tau_3}+r\right)\left\|\widetilde{\boldsymbol{\rho}_n}\sqrt{n}\right\|_2^2\\
&\qquad+\frac{\tau_2}{2\alpha n}\left\|\boldsymbol{\Sigma}_n^{1/2}\left(\widetilde{\boldsymbol{\rho}}_n\sqrt{n}+\frac{\alpha\beta}{\tau_2\sqrt{\delta}}\boldsymbol{\Sigma}_n^{-1/2}\mathbf{h}-\frac{\eta\alpha\sqrt{n}}{\tau_2\left\|\widetilde{\boldsymbol{\theta}_n^\star}\right\|_2^2}\boldsymbol{\Sigma}_n^{-1/2}\widetilde{\boldsymbol{\theta}_n^\star}\right)\right\|_2^2 - \frac{\eta^2\alpha}{2\tau_2\left\|\widetilde{\boldsymbol{\theta}_n^\star}\right\|_2^2}.
\end{aligned}\tag{3.78}$$

To proceed, note that $\mathbf{\Sigma}_n^{1/2} = \mathbf{U}_n \mathbf{\Lambda}_n^{1/2} \mathbf{U}_n^\top$ where $\mathbf{U}_n$ is an orthogonal matrix, therefore with a change of variable $\mathbf{U}_n^\top \widetilde{\boldsymbol{\rho}_n} \Rightarrow \widetilde{\boldsymbol{\rho}_n}$, the minimization based on $\widetilde{\boldsymbol{\rho}_n}$ in (3.78) is equivalent to:

$$\min_{\widetilde{\boldsymbol{\rho}_n} \in \mathbb{R}^n} \frac{1}{n} \left(\frac{\varepsilon_{\rm tr} \gamma}{2\tau_3} + r \right) \left\| \widetilde{\boldsymbol{\rho}_n} \sqrt{n} \right\|_2^2 + \frac{\tau_2}{2\alpha n} \left\| \mathbf{\Lambda}_n^{1/2} \widetilde{\boldsymbol{\rho}_n} \sqrt{n} + \frac{\alpha\beta}{\tau_2 \sqrt{\delta}} \mathbf{U}_n^\top \mathbf{h} - \frac{\eta \alpha \sqrt{n}}{\tau_2 \left\| \widetilde{\boldsymbol{\theta}_n^\star} \right\|_2^2} \mathbf{U}_n^\top \widetilde{\boldsymbol{\theta}_n^\star} \right\|_2^2 . \tag{3.79}$$

It holds that $\mathbf{U}_n^\top \mathbf{h} \sim \mathbf{h}$ and following Assumption 3.2.3, we have

$$\mathbf{U}_n^\top \widetilde{\boldsymbol{\theta}_n^\star} = \mathbf{U}_n^\top \mathbf{\Sigma}_n^{1/2} \boldsymbol{\theta}_n^\star = \mathbf{\Lambda}_n^{1/2} \mathbf{U}_n^\top \boldsymbol{\theta}_n^\star = \mathbf{\Lambda}_n^{1/2} \mathbf{v}_n. \tag{3.80}$$

Therefore optimization over $\widetilde{\boldsymbol{\rho}_n}$ becomes separable over its entries and (3.79) is equivalent to

$$\begin{aligned}
&\frac{1}{n} \sum_{i=1}^n \min_{\widetilde{\boldsymbol{\rho}_{n,i}} \in \mathbb{R}} \left(\frac{\varepsilon_{\rm tr} \gamma}{2\tau_3} + r \right) \widetilde{\boldsymbol{\rho}_{n,i}}^2 + \frac{\tau_2 \lambda_{n,i}}{2\alpha n} \left(\widetilde{\boldsymbol{\rho}_{n,i}} + \frac{\alpha\beta}{\tau_2 \sqrt{\delta \lambda_{n,i}}} \mathbf{h}_i - \frac{\eta \alpha \sqrt{n}}{\tau_2 \left\| \widetilde{\boldsymbol{\theta}_n^\star} \right\|_2^2} \mathbf{v}_{n,i} \right)^2 \\
&= \frac{1}{n} \left(\frac{\varepsilon_{\rm tr} \gamma}{2\tau_3} + r \right) \sum_{i=1}^n \mathcal{M}_{\ell_2^2} \left(\frac{\alpha\beta}{\tau_2 \sqrt{\delta \lambda_{n,i}}} \mathbf{h}_i + \frac{\eta \alpha \sqrt{n}}{\tau_2 \left\| \widetilde{\boldsymbol{\theta}_n^\star} \right\|_2^2} \mathbf{v}_{n,i}; \frac{\varepsilon_{\rm tr} \gamma \alpha + 2\tau_3 r \alpha}{2\tau_2 \tau_3 \lambda_{n,i}} \right) \\
&\xrightarrow{P} \left(\frac{\varepsilon_{\rm tr} \gamma}{2\tau_3} + r \right) \mathbb{E}_{H,V,L} \left[\mathcal{M}_{\ell_2^2} \left(\frac{\alpha\beta}{\tau_2 \sqrt{\delta L}} H + \frac{\eta \alpha}{\tau_2 \zeta^2} V; \frac{\varepsilon_{\rm tr} \gamma \alpha + 2\tau_3 r \alpha}{2\tau_2 \tau_3 L} \right) \right] \qquad (3.81) \\
&= \left(\frac{\varepsilon_{\rm tr} \gamma}{2\tau_3} + r \right) \left(\frac{\eta^2 \alpha^2}{\tau_2^2 \zeta^4} \right) \mathbb{E}_L \left[\frac{\frac{\zeta^4 \beta^2}{\eta^2 \delta} + L}{\frac{\varepsilon_{\rm tr} \gamma \alpha + 2\tau_3 r \alpha}{2\tau_2 \tau_3} + L} \right], \qquad (3.82)
\end{aligned}$$

where H is standard normal and in (3.81), we used Assumption 3.2.3, together with the fact that ℓ_2^2 is pseudo-Lipschitz of order 2. In deriving (3.82), we used the fact that $\mathcal{M}_{\ell_2^2}(x;\tau) = \frac{x^2}{\tau+1}$ and $\mathbb{E}[H^2] = \mathbb{E}[V^2] = 1$. Inserting this back in (3.50) leads to the

following objective,

$$\min_{\substack{\alpha,\tau_1,\tau_3,w\in\mathbb{R}_+,\\ \mu\in\mathbb{R}}} \max_{\substack{\tau_2,\beta,\gamma\in\mathbb{R}_+,\\ \eta\in\mathbb{R}}} -\gamma w - \frac{\mu^2\tau_2}{2\alpha}\zeta^2 - \frac{\alpha\beta^2}{2\delta\tau_2} - \frac{\alpha\tau_2}{2} + \frac{\beta\tau_1}{2} + \eta\mu - \frac{\eta^2\alpha}{2\tau_2\zeta^2} + \frac{\varepsilon_{\rm tr}\gamma\tau_3}{2} + \mathbb{E}_{G,S}\left[\mathcal{M}_{\mathcal{L}}\left(\alpha G + \mu\zeta S\cdot\psi(\zeta S) - w; \frac{\tau_1}{\beta}\right)\right] + \left(\frac{\varepsilon_{\rm tr}\gamma}{2\tau_3} + r\right)\left(\frac{\eta^2\alpha^2}{\tau_2^2\zeta^4}\right)\mathbb{E}_L\left[\frac{\frac{\zeta^4\beta^2}{\eta^2\delta} + L}{\frac{\varepsilon_{\rm tr}\gamma\alpha + 2\tau_3 r\alpha}{\tau_2\tau_3} + L}\right]. \tag{3.83}$$

This completes the proof of Theorem 3.4.1.

Case II: Isotropic Features

Here we derive the minimax objective for $\mathbf{\Sigma}_n = \mathbb{I}_n$. We focus here on GLM, the extensions to GMM are achievable in light of the analysis in Section 3.8.3.

Corollary 3.8.2. *Consider the Generalized Linear model (3.3). Let $\mathbf{\Sigma}_n = \mathbb{I}_n$ and $\|\boldsymbol{\theta}_n^\star\|_2 \xrightarrow{P} 1$. The high-dimensional limit for the adversarial test error takes the following form,*

$$\{\mathcal{E}^{GLM}_{\ell_2,\varepsilon}(\boldsymbol{\theta}_n)\} \xrightarrow{P} \mathbb{P}\left(\frac{\mu^\star S\psi(S) + \alpha^\star G}{\sqrt{\alpha^{\star 2} + \mu^{\star 2}}} < \varepsilon\right), \tag{3.84}$$

where $(\alpha^\star, \mu^\star)$ is the unique solution to the following min-max objective,

$$\min_{\mu\in\mathbb{R},\alpha,\tau\in\mathbb{R}_+} \max_{\beta\in\mathbb{R}_+} \widetilde{L} \triangleq \frac{\beta\tau}{2} - \frac{\alpha\beta}{\sqrt{\delta}} + \lambda\alpha^2 + \lambda\mu^2 + \mathbb{E}_{G,S}\Big[\mathcal{M}_{\mathcal{L}}\left(\mu S\psi(S) + \alpha G - \varepsilon_{\rm tr}\sqrt{\alpha^2+\mu^2}; \tau/\beta\right)\Big]. \tag{3.85}$$

Proof: We know that,

$$\widehat{\boldsymbol{\theta}}_n = \min_{\boldsymbol{\theta}_n\in\mathbb{R}^n} \frac{1}{m}\sum_{i=1}^m \mathcal{L}(y_i\mathbf{x}_i^\top\boldsymbol{\theta}_n - \varepsilon_{\rm tr}\|\boldsymbol{\theta}_n\|_2) + \lambda\|\boldsymbol{\theta}_n\|_2^2.$$

To proceed, we use our approach that derived (3.38), to end at a similar expression, here for $p = 2$. We omit the steps as they are akin to the steps that led to (3.38). We end up with the following objective which is the counterpart of (3.38) for $q = p = 2$.

$$\min_{\boldsymbol{\theta}_n \in \mathbb{R}^n, \mathbf{v} \in \mathbb{R}^m} \max_{\beta \in \mathbb{R}_+} \frac{\mathbf{1}_m^\top}{m} (\mathbf{v} - \varepsilon_{\text{tr}} \|\boldsymbol{\theta}_n\|_2 \mathbf{1}_m) + \frac{\beta}{\sqrt{m}} \left\| -\mathbf{v} + YX\Theta_n \boldsymbol{\theta}_n + \mathbf{g} \|\Theta_n^\perp \boldsymbol{\theta}_n\|_2 \right\|_2 \tag{3.86}$$
$$+ \frac{\beta \mathbf{h}^\top \Theta_n^\perp \boldsymbol{\theta}_n}{\sqrt{m}} + \lambda \|\boldsymbol{\theta}_n\|_2^2 =$$

$$\min_{\boldsymbol{\theta}_n \in \mathbb{R}^n, \mathbf{v} \in \mathbb{R}^m} \max_{\beta, \tau \in \mathbb{R}_+} \frac{\mathbf{1}_m^\top}{m} (\mathbf{v} - \varepsilon_{\text{tr}} \|\boldsymbol{\theta}_n\|_2 \mathbf{1}_m) + \frac{\beta}{m\tau} \left\| -\mathbf{v} + YX\Theta_n \boldsymbol{\theta}_n + \mathbf{g} \|\Theta_n^\perp \boldsymbol{\theta}_n\|_2 \right\|_2^2 + \frac{\beta\tau}{2}$$
$$+ \frac{\beta \mathbf{h}^\top \Theta_n^\perp \boldsymbol{\theta}_n}{\sqrt{m}} + \lambda \|\boldsymbol{\theta}_n\|_2^2 ,$$

where similar to (3.44), here also (3.86) is due to $x = \min_{\tau \in \mathbb{R}_+} \frac{x^2}{2\tau} + \frac{\tau}{2}$. By minimizing w.r.t. $\boldsymbol{\theta}_n$ and denoting $\alpha \triangleq \|\Theta_n^\perp \boldsymbol{\theta}_n\|_2$, $\mu \triangleq \|\Theta_n \boldsymbol{\theta}_n\|_2$ we have,

$$\min_{\mathbf{v} \in \mathbb{R}^m, \mu \in \mathbb{R}, \alpha \in \mathbb{R}_+} \max_{\beta, \tau \in \mathbb{R}_+} \frac{\mathbf{1}_m^\top}{m} \left(\mathbf{v} - \varepsilon_{\text{tr}} \sqrt{\alpha^2 + \mu^2} \mathbf{1}_m \right) + \frac{\beta}{m\tau} \| -\mathbf{v} + \mu YX \boldsymbol{\theta}_n^\star + \alpha \mathbf{g} \|_2^2$$
$$+ \frac{\beta\tau}{2} - \frac{\alpha\beta \mathbf{h}}{\sqrt{m}} + \lambda \|\boldsymbol{\theta}_n\|_2^2.$$

After $m, n \to \infty$, one can easily see that the objective simplifies to (3.85). Additionally, by replacing $(\alpha^\star, \mu^\star)$ derived as the solution of (3.85), in (3.20), we derive the asymptotic error of adversary. This completes the proof

A System of Equations

Now, we present the corresponding fixed-point equations for the ℓ_2 case in (3.87). The equations are obtained by forming $\nabla \widetilde{L} = \mathbf{0}$ based on three variables (α, μ, κ), where

$\kappa := \tau/\beta$.

$$\begin{cases} \mathbb{E}_{G,S}\Big[\Big(\mathcal{M}'_{\mathcal{L},1}\left(\mu S\psi(S) + \alpha G - \varepsilon_{\rm tr}\sqrt{\alpha^2+\mu^2};\kappa\right)\Big)^2\Big] = \frac{\alpha^2}{\kappa^2\delta}, \\ \mathbb{E}_{G,S}\left[S\psi(S)\cdot\mathcal{M}'_{\mathcal{L},1}\left(\mu S\psi(S) + \alpha G - \varepsilon_{\rm tr}\sqrt{\alpha^2+\mu^2};\kappa\right)\right] = -2\lambda\mu \\ \qquad\qquad + \frac{\varepsilon_{\rm tr}\mu}{\sqrt{\alpha^2+\mu^2}}\mathbb{E}_{G,S}\left[\mathcal{M}'_{\mathcal{L},1}\left(\mu S\psi(S) + \alpha G - \varepsilon_{\rm tr}\sqrt{\alpha^2+\mu^2};\kappa\right)\right], \\ \mathbb{E}_{G,S}\left[G\cdot\mathcal{M}'_{\mathcal{L},1}\left(\mu S\psi(S) + \alpha G - \varepsilon_{\rm tr}\sqrt{\alpha^2+\mu^2};\kappa\right)\right] = -2\alpha\lambda \\ \qquad\qquad + \frac{\varepsilon_{\rm tr}\alpha}{\sqrt{\alpha^2+\mu^2}}\mathbb{E}_{G,S}\left[\mathcal{M}'_{\mathcal{L},1}\left(\mu S\psi(S) + \alpha G - \varepsilon_{\rm tr}\sqrt{\alpha^2+\mu^2};\kappa\right)\right] + \frac{\alpha}{\delta\kappa}. \end{cases} \tag{3.87}$$

Next, we show how to derive the saddle-point equations (3.87) from $\nabla\widetilde{L} = \mathbf{0}$. To derive the first equation in (3.87), we can see that based on Proposition 3.8.1,

$$\begin{aligned} \nabla_\tau\widetilde{L} &= \frac{\beta}{2} - \frac{1}{2\beta}\mathbb{E}\Big[\Big(\mathcal{M}'_{\mathcal{L},1}\left(\mu S\psi(S) + \alpha G - \varepsilon_{\rm tr}\sqrt{\alpha^2+\mu^2};\tau/\beta\right)\Big)^2\Big], \\ \nabla_\beta\widetilde{L} &= \frac{\tau}{2} - \frac{\alpha}{\sqrt{\delta}} + \frac{\tau}{2\beta^2}\mathbb{E}\Big[\Big(\mathcal{M}'_{\mathcal{L},1}\left(\mu S\psi(S) + \alpha G - \varepsilon_{\rm tr}\sqrt{\alpha^2+\mu^2};\tau/\beta\right)\Big)^2\Big]. \end{aligned} \tag{3.88}$$

After forming $\frac{\nabla_\tau\widetilde{L}}{\beta} + \frac{\nabla_\beta\widetilde{L}}{\tau} = 0$, we can deduce that $\alpha = \tau\sqrt{\delta}$. Since we defined $\kappa \triangleq \tau/\beta$, it follows that $\beta = \alpha/(\kappa\sqrt{\delta})$. Replacing this in (3.88), yields the first equation in (3.87). The last two equations in (3.87), are obtained directly from $\nabla_\mu\widetilde{L} = 0$ and $\nabla_\alpha\widetilde{L} = 0$.

For GMM (3.2), the min-max objective and the system of equations are obtained by replacing $S\psi(S)$ with $S+1$, in (3.85) and (3.87).

3.8.3 The Gaussian-Mixture Model Analysis

Adversarial Error of an Arbitrary Estimator

Next lemma (restatement of Lemma 3.3.1 for GMM) derives the asymptotic error of a given sequence of estimators for the Gaussian-Mixture model.

Lemma 3.8.2. *The high-dimensional limit of the Adversarial Error for the Gaussian-Mixture model with a sequence of classifiers* $\{\boldsymbol{\theta}_n\}$ *is given as follows,*

$$\left\{\mathcal{E}^{GMM}_{\ell_q,\varepsilon}(\boldsymbol{\theta}_n)\right\} \xrightarrow{P} Q\left(\frac{\mu\widetilde{\zeta}^2 - \varepsilon u}{\sqrt{\alpha^2+\mu^2\widetilde{\zeta}^2}}\right), \tag{3.89}$$

where $Q(\cdot)$ *denotes the Gaussian Q-function and* u, μ *and* α *are derived as follows,*

$$\left\|\boldsymbol{\Sigma}_n^{-1/2}\widetilde{\boldsymbol{\theta}_n}\right\|_p \xrightarrow{P} u, \quad \langle\widetilde{\boldsymbol{\theta}^\star_n}, \widetilde{\boldsymbol{\theta}_n}\rangle/\|\widetilde{\boldsymbol{\theta}^\star_n}\|_2^2 \xrightarrow{P} \mu, \quad \left\|\Theta_n^\perp\widetilde{\boldsymbol{\theta}_n}\right\|_2 \xrightarrow{P} \alpha,$$

for ℓ_p*-norm denoting the dual of the* ℓ_q*-norm,* $\widetilde{\boldsymbol{\theta}_n} \triangleq \boldsymbol{\Sigma}_n^{1/2}\boldsymbol{\theta}_n, \widetilde{\boldsymbol{\theta}^\star_n} \triangleq \boldsymbol{\Sigma}_n^{-1/2}\boldsymbol{\theta}^\star_n$ *and* $\Theta_n^\perp \in \mathbb{R}^{n\times n}$ *defined as follows:*

$$\Theta_n^\perp \triangleq \mathbb{I}_n - \Theta_n, \quad \Theta_n \triangleq \frac{\widetilde{\boldsymbol{\theta}^\star_n}\widetilde{\boldsymbol{\theta}^\star_n}^\top}{\left\|\widetilde{\boldsymbol{\theta}^\star_n}\right\|_2^2}.$$

Moreover, in the special case of $q=2$ *and* $\boldsymbol{\Sigma}_n = \mathbb{I}_n$*, by denoting* $\sigma \triangleq \alpha/\mu$*, (3.89) simplifies to,*

$$\left\{\mathcal{E}^{GMM}_{\ell_2,\varepsilon}(\boldsymbol{\theta}_n)\right\} \xrightarrow{P} Q\left(\frac{\mu}{\sqrt{\alpha^2+\mu^2}} - \varepsilon\right), \tag{3.90}$$

Proof: Note that here $\widetilde{\boldsymbol{\theta}^\star_n}$ is defined rather differently in GLM. Based on the definition of GMM, we have $\mathbf{x} = y\boldsymbol{\theta}^\star_n + \mathbf{z}$ for $\mathbf{z} \sim \mathcal{N}(\mathbf{0}_n, \boldsymbol{\Sigma}_n)$ and $\mathbf{z} = \boldsymbol{\Sigma}_n^{1/2}\bar{\mathbf{z}}$ for standard Gaussian

vector $\bar{\mathbf{z}}$. We can write

$$\mathbb{E}_{\mathbf{x},y}\left[\max_{\|\boldsymbol{\delta}\|_q<\varepsilon} \mathbf{1}_{\{y\neq \operatorname{sign}\langle \mathbf{x}+\boldsymbol{\delta},\boldsymbol{\theta}_n\rangle\}}\right] = \mathbb{P}\Big(y\langle \mathbf{x},\boldsymbol{\theta}_n\rangle - \varepsilon\|\boldsymbol{\theta}_n\|_p< 0\Big)$$

$$= \mathbb{P}\Big(y\langle \mathbf{z},\boldsymbol{\theta}_n\rangle + \langle \boldsymbol{\theta}_n,\boldsymbol{\theta}_n^\star\rangle - \varepsilon\|\boldsymbol{\theta}_n\|_p< 0\Big) \tag{3.91}$$

$$= \mathbb{P}\left(y\langle \bar{\mathbf{z}},\widetilde{\boldsymbol{\theta}_n}\rangle + \langle \widetilde{\boldsymbol{\theta}_n},\widetilde{\boldsymbol{\theta}_n^\star}\rangle - \varepsilon\left\|\boldsymbol{\Sigma}_n^{-1/2}\widetilde{\boldsymbol{\theta}_n}\right\|_p\right)$$

$$= \mathbb{P}\left(\langle \bar{\mathbf{z}},\Theta_n\widetilde{\boldsymbol{\theta}_n}\rangle + \langle \bar{\mathbf{z}},\Theta_n^\perp\widetilde{\boldsymbol{\theta}_n}\rangle + \langle \widetilde{\boldsymbol{\theta}_n},\widetilde{\boldsymbol{\theta}_n^\star}\rangle - \varepsilon\left\|\boldsymbol{\Sigma}_n^{-1/2}\widetilde{\boldsymbol{\theta}_n}\right\|_p < 0\right), \tag{3.92}$$

where (3.91) and (3.92) follow from the definition of the Gaussian-Mixture model and noting that $\mathbf{z}$ is independent of y. Since $\langle \bar{\mathbf{z}},\Theta_n^\perp\widetilde{\boldsymbol{\theta}_n}\rangle$ and $\langle \bar{\mathbf{z}},\Theta_n\widetilde{\boldsymbol{\theta}_n}\rangle$ are independent, we can deduce that for $G,S \overset{\text{iid}}{\sim} \mathcal{N}(0,1)$, it holds that

$$\left\langle \bar{\mathbf{z}},\Theta_n\widetilde{\boldsymbol{\theta}_n}\right\rangle = \frac{\left\langle \widetilde{\boldsymbol{\theta}_n^\star},\widetilde{\boldsymbol{\theta}_n}\right\rangle}{\left\|\widetilde{\boldsymbol{\theta}_n^\star}\right\|_2^2}\left\langle \bar{\mathbf{z}},\widetilde{\boldsymbol{\theta}_n^\star}\right\rangle \xrightarrow{P} \mu\widetilde{\zeta}S,$$

$$\left\langle \bar{\mathbf{z}},\Theta_n^\perp\widetilde{\boldsymbol{\theta}_n}\right\rangle \sim \left\|\Theta_n^\perp\widetilde{\boldsymbol{\theta}_n}\right\|_2\bar{\mathbf{z}} \xrightarrow{P} \alpha G,$$

$$\left\langle \widetilde{\boldsymbol{\theta}_n},\widetilde{\boldsymbol{\theta}_n^\star}\right\rangle \xrightarrow{P} \mu\widetilde{\zeta}^2.$$

where recall that $\left\|\widetilde{\boldsymbol{\theta}_n^\star}\right\|_2 = \boldsymbol{\theta}_n^{\star\top}\boldsymbol{\Sigma}_n^{-1}\boldsymbol{\theta}_n^\star \to \widetilde{\zeta}$ by Assumption 3.2.2. Therefore, from (3.92), we infer that,

$$\{\mathcal{E}_{\ell_q,\varepsilon}^{\text{GMM}}(\boldsymbol{\theta}_n)\} \xrightarrow{P} \mathbb{P}\Big(\mu\widetilde{\zeta}\left(S+\widetilde{\zeta}\right) + \alpha G - u\varepsilon < 0\Big).$$

This leads to (3.89). When $q=2$ and $\boldsymbol{\Sigma}_n = \mathbb{I}_n$, we have that $\widetilde{\zeta}=1$ and noting that

$u = \sqrt{\alpha^2 + \mu^2}$, leads to (3.90). This completes the proof.

Proofs for the Gaussian-Mixture Model

In this section, we outline the approach to the proof of Theorem 3.3.1 for GMM. In light of the previously described steps for GLM, here we only need to derive the corresponding min-max scalar problem for GMM. For the Gaussian-Mixture model we have by definition that $\mathbf{x}_i \sim \mathcal{N}(y_i \boldsymbol{\theta}_n^\star, \boldsymbol{\Sigma}_n)$. Thus, the min-max ERM can be equivalently written as follows,

$$
\begin{aligned}
&\min_{\boldsymbol{\theta}_n \in \mathbb{R}^n} \max_{\substack{\|\boldsymbol{\delta}_i\|_\infty \le \varepsilon \\ i \in [m]}} \frac{1}{m} \sum_{i=1}^m \mathcal{L}\Big(y_i \Big\langle \mathbf{x}_i + \boldsymbol{\delta}_i, \boldsymbol{\theta}_n \Big\rangle\Big) + \lambda \|\boldsymbol{\theta}_n\|_2^2 \\
&= \min_{\boldsymbol{\theta}_n \in \mathbb{R}^n} \frac{1}{m} \sum_{i=1}^m \mathcal{L}\Big(y_i \Big\langle \mathbf{x}_i, \boldsymbol{\theta}_n \Big\rangle - \varepsilon \|\boldsymbol{\theta}_n\|_1\Big) + \lambda \|\boldsymbol{\theta}_n\|_2^2 \\
&= \min_{\widetilde{\boldsymbol{\theta}_n} \in \mathbb{R}^n} \frac{1}{m} \sum_{i=1}^m \mathcal{L}\left(\Big\langle \bar{\mathbf{z}}_i, \widetilde{\boldsymbol{\theta}_n} \Big\rangle + \Big\langle \widetilde{\boldsymbol{\theta}_n}, \widetilde{\boldsymbol{\theta}_n^\star} \Big\rangle - \varepsilon \left\| \boldsymbol{\Sigma}_n^{-1/2} \widetilde{\boldsymbol{\theta}_n} \right\|_1\right) + \lambda \left\| \boldsymbol{\Sigma}_n^{-1/2} \widetilde{\boldsymbol{\theta}_n} \right\|_2^2. \\
&= \min_{\widetilde{\boldsymbol{\theta}_n} \in \mathbb{R}^n} \frac{1}{m} \sum_{i=1}^m \mathcal{L}\left(\Big\langle \bar{\mathbf{z}}_i, \Theta_n \widetilde{\boldsymbol{\theta}_n} \Big\rangle + \Big\langle \bar{\mathbf{z}}_i, \Theta_n^\perp \widetilde{\boldsymbol{\theta}_n} \Big\rangle + \Big\langle \widetilde{\boldsymbol{\theta}_n}, \widetilde{\boldsymbol{\theta}_n^\star} \Big\rangle - \varepsilon \left\| \boldsymbol{\Sigma}_n^{-1/2} \widetilde{\boldsymbol{\theta}_n} \right\|_1\right) + \lambda \left\| \boldsymbol{\Sigma}_n^{-1/2} \widetilde{\boldsymbol{\theta}_n} \right\|_2^2
\end{aligned}
\tag{3.93}
$$

The second step is due to the fact that $\widetilde{\boldsymbol{\theta}_n^\star} \triangleq \boldsymbol{\Sigma}_n^{-1/2} \boldsymbol{\theta}_n^\star$, $\widetilde{\boldsymbol{\theta}_n} \triangleq \boldsymbol{\Sigma}_n^{1/2} \boldsymbol{\theta}_n$ and that y_i and $\bar{\mathbf{z}}_i \overset{\text{iid}}{\sim} \mathcal{N}(\mathbf{0}, \mathbb{I}_n)$ are independent for all i. In the last step we used the matrices $\Theta_n \triangleq \widetilde{\boldsymbol{\theta}_n^\star} \widetilde{\boldsymbol{\theta}_n^\star}^\top / \|\widetilde{\boldsymbol{\theta}_n^\star}\|_2^2$ and $\Theta_n^\perp \triangleq \mathbb{I}_n - \Theta_n$, to allow scalarization w.r.t. desired quantities α, μ and also to allow using CGMT as the random variables $\langle \bar{\mathbf{z}}_i, \Theta_n \widetilde{\boldsymbol{\theta}_n} \rangle$ are $\langle \bar{\mathbf{z}}_i, \Theta_n^\perp \widetilde{\boldsymbol{\theta}_n} \rangle$ are independent. Next, similar to (3.32), we can use the Lagrangian multiplier method to

obtain that (3.93) is equivalent to

$$\min_{\widetilde{\boldsymbol{\theta}_n}\in\mathbb{R}^n,\mathbf{v}\in\mathbb{R}^m}\ \max_{\mathbf{u}\in\mathbb{R}^m}\ \frac{\mathbf{1}_m^\top}{m}\left(\mathbf{v}-\varepsilon\left\|\boldsymbol{\Sigma}_n^{-1/2}\widetilde{\boldsymbol{\theta}_n}\right\|_1\mathbf{1}_m\right)-\frac{\langle\mathbf{u},\mathbf{v}\rangle}{m}+\frac{\left\langle\mathbf{u},\bar{Z}\Theta_n\widetilde{\boldsymbol{\theta}_n}\right\rangle}{m}+\frac{\left\langle\mathbf{u},\bar{Z}\Theta_n^{\perp}\widetilde{\boldsymbol{\theta}_n}\right\rangle}{m}$$
$$+\frac{\langle\mathbf{u},\mathbf{1}_m\rangle}{m}\left\langle\widetilde{\boldsymbol{\theta}_n},\widetilde{\boldsymbol{\theta}_n^\star}\right\rangle+\lambda\left\|\boldsymbol{\Sigma}_n^{-1/2}\widetilde{\boldsymbol{\theta}_n}\right\|_2^2, \tag{3.94}$$

The objective in (3.94) bears close similarity to its GLM counterpart in (3.32). Note that here $\bar{Z}\Theta_n\widetilde{\boldsymbol{\theta}_n}$ and $\bar{Z}\Theta_n^{\perp}\widetilde{\boldsymbol{\theta}_n}$ have the same role as $Y\bar{X}\Theta_n\widetilde{\boldsymbol{\theta}_n}$ and $Y\bar{X}\Theta_n^{\perp}\widetilde{\boldsymbol{\theta}_n}$ in (3.32), respectively. Here we also have an additional term $\frac{\langle\mathbf{u},\mathbf{1}_m\rangle}{m}\langle\widetilde{\boldsymbol{\theta}_n},\widetilde{\boldsymbol{\theta}_n^\star}\rangle$ compared to (3.32). We recall that based on the definition, it holds that $\langle\widetilde{\boldsymbol{\theta}_n},\widetilde{\boldsymbol{\theta}_n^\star}\rangle\xrightarrow{P}\mu\widetilde{\zeta}^2$. Continuing with the same technique described in Section 3.8.1 that led to the objective (3.54), we find that for GMM, (3.94) is equivalent to the following min-max problem (details are omitted for brevity):

$$\min_{\substack{\alpha,\tau_1,w\in\mathbb{R}_+,\\ \mu\in\mathbb{R}}}\ \max_{\substack{\tau_2,\beta,\gamma\in\mathbb{R}_+,\\ \eta\in\mathbb{R}}}\ -\gamma w-\frac{\mu^2\tau_2}{2\alpha}\left\|\widetilde{\boldsymbol{\theta}_n^\star}\right\|_2^2-\frac{\alpha\beta^2}{2\delta\tau_2}-\frac{\alpha\tau_2}{2}+\frac{\beta\tau_1}{2}+\eta\mu-\frac{\eta^2\alpha}{2\tau_2\left\|\widetilde{\boldsymbol{\theta}_n^\star}\right\|_2^2}$$
$$+\frac{1}{m}\mathcal{M}\left(\mu\bar{Z}\widetilde{\boldsymbol{\theta}_n^\star}+\alpha\mathbf{g}+\mu\widetilde{\zeta}^2\mathbf{1}_m-w\mathbf{1}_m;\frac{\tau_1}{\beta}\right)$$
$$+\frac{\varepsilon_{\mathrm{tr}}\gamma}{n}\mathcal{M}_{\left(\boldsymbol{\ell}_1+\frac{r}{\varepsilon_{\mathrm{tr}}\gamma}\boldsymbol{\ell}_2^2,\boldsymbol{\Sigma}_n\right)}\left(\frac{\alpha\beta}{\tau_2\sqrt{\delta}}\boldsymbol{\Sigma}_n^{-1/2}\mathbf{h}+\frac{\alpha\eta\sqrt{n}}{\tau_2\left\|\widetilde{\boldsymbol{\theta}_n^\star}\right\|_2^2}\boldsymbol{\Sigma}_n^{-1}\boldsymbol{\theta}_n^\star;\frac{\alpha\varepsilon_{\mathrm{tr}}\gamma}{\tau_2}\right). \tag{3.95}$$

We have $\bar{Z}\widetilde{\boldsymbol{\theta}_n^\star}\sim\widetilde{\zeta}\mathbf{s}$ for a standard Gaussian vector $\mathbf{s}$ independent of $\mathbf{g}$. This leads to

$$\frac{1}{m}\mathcal{M}\left(\mu\bar{Z}\widetilde{\boldsymbol{\theta}_n^\star}+\alpha\mathbf{g}+\mu\widetilde{\zeta}^2\mathbf{1}_m-w\mathbf{1}_m;\frac{\tau_1}{\beta}\right)\xrightarrow{P}\mathbb{E}_G\left[\mathcal{M}_{\mathcal{L}}\left(\sqrt{\alpha^2+\mu^2\widetilde{\zeta}^2}G+\mu\widetilde{\zeta}^2-w;\frac{\tau_1}{\beta}\right)\right],$$

for standard Gaussian random variable G. In particular, when $\boldsymbol{\Sigma}_n$ is a diagonal matrix, we end up with the following min-max problem based on eight scalars:

$$\min_{\substack{\alpha,\tau_1,w\in\mathbb{R}_+,\\ \mu\in\mathbb{R}}} \max_{\substack{\tau_2,\beta,\gamma\in\mathbb{R}_+,\\ \eta\in\mathbb{R}}} -\gamma w - \frac{\mu^2\tau_2}{2\alpha}\widetilde{\zeta}^2 - \frac{\alpha\beta^2}{2\delta\tau_2} - \frac{\alpha\tau_2}{2} + \frac{\beta\tau_1}{2} + \eta\mu - \frac{\eta^2\alpha}{2\tau_2\widetilde{\zeta}^2}$$
$$+ \mathbb{E}_G\left[\mathcal{M}_{\mathcal{L}}\left(\sqrt{\alpha^2+\mu^2\widetilde{\zeta}^2}\,G + \mu\widetilde{\zeta}^2 - w; \frac{\tau_1}{\beta}\right)\right]$$
$$+ \varepsilon_{\rm tr}\gamma\,\mathbb{E}_{L,H,T}\left[\mathcal{M}_{\ell_1+\frac{r}{\varepsilon_{\rm tr}\gamma}\ell_2^2}\left(\frac{\alpha\beta}{\tau_2\sqrt{\delta L}}H + \frac{\alpha\eta}{\tau_2\widetilde{\zeta}^2 L}T; \frac{\alpha\varepsilon_{\rm tr}\gamma}{\tau_2 L}\right)\right], \tag{3.96}$$

as desired by Theorem 3.3.1.

Proof of Theorem 3.4.1 for GMM, follows the same steps as GLM, however note that here, due to the definition of $\widetilde{\boldsymbol{\theta}_n^\star}$, (3.80) changes into

$$\mathbf{U}_n^\top\widetilde{\boldsymbol{\theta}_n^\star} = \mathbf{U}_n^\top\boldsymbol{\Sigma}_n^{-1/2}\boldsymbol{\theta}_n^\star = \boldsymbol{\Lambda}_n^{-1/2}\mathbf{U}_n^\top\boldsymbol{\theta}_n^\star = \boldsymbol{\Lambda}_n^{-1/2}\mathbf{v}_n. \tag{3.97}$$

Thus, the resulting min-max objective has the following form,

$$\min_{\substack{\alpha,\tau_1,\tau_3,w\in\mathbb{R}_+,\\ \mu\in\mathbb{R}}} \max_{\substack{\tau_2,\beta,\gamma\in\mathbb{R}_+,\\ \eta\in\mathbb{R}}} -\gamma w - \frac{\mu^2\tau_2}{2\alpha}\widetilde{\zeta}^2 - \frac{\alpha\beta^2}{2\delta\tau_2} - \frac{\alpha\tau_2}{2} + \frac{\beta\tau_1}{2} + \eta\mu - \frac{\eta^2\alpha}{2\tau_2\widetilde{\zeta}^2} + \frac{\varepsilon_{\rm tr}\gamma\tau_3}{2}$$
$$+ \mathbb{E}_G\left[\mathcal{M}_{\mathcal{L}}\left(\sqrt{\alpha^2+\mu^2\widetilde{\zeta}^2}\,G + \mu\widetilde{\zeta}^2 - w; \frac{\tau_1}{\beta}\right)\right]$$
$$+ \left(\frac{\varepsilon_{\rm tr}\gamma}{2\tau_3} + r\right)\left(\frac{\eta^2\alpha^2}{\tau_2^2\widetilde{\zeta}^4}\right)\mathbb{E}_L\left[\frac{\frac{\widetilde{\zeta}^4\beta^2}{\eta^2\delta} + L^{-1}}{\frac{\varepsilon_{\rm tr}\gamma\alpha+2\tau_3 r\alpha}{2\tau_2\tau_3} + L}\right].$$

This together with Lemma 3.8.2, yields the proof of Theorem 3.4.1 for GMM.

The Large Sample-size Limit

In this section, we focus on the $\delta = m/n \to \infty$ limit. In particular, we consider the Exponential loss $\mathcal{L}(t) = \exp(-t)$ and the isotropic Gaussian-mixture model and set $r = 0$. We prove that for $q = 2$, when $\delta \to \infty$, the adversarial test error, exactly achieves the

Bayes adversarial error derived by [83]. The results are summarized in the following corollary.

Corollary 3.8.3. *Consider the Gaussian-mixture model under the same settings as Corollary 3.8.2. Let the loss function $\mathcal{L}$, be the Exponenttal loss and let $\delta \to \infty$. Fix $\varepsilon_{\rm ts} < 1$. Then if $\varepsilon_{\rm tr} < 1$, the adversarial test error of estimators derived by adversarial training and the Bayes adversarial test error are equal.*

Proof: To see this, note that under these conditions, (3.85) takes the following form

$$\min_{\mu\in\mathbb{R},\alpha,\tau\in\mathbb{R}_+} \max_{\beta\in\mathbb{R}_+} \widetilde{L}_{\delta\to\infty} = \frac{\beta\tau}{2} + \mathbb{E}_G\Big[\mathcal{M}_{\mathcal{L}}\left(\sqrt{\alpha^2+\mu^2}\,G + \mu - \varepsilon_{\rm tr}\sqrt{\alpha^2+\mu^2}; \tau/\beta\right)\Big]. \tag{3.98}$$

In light of Proposition 3.8.1, $\mathcal{M}_{\mathcal{L}}(x;\cdot)$ is a decreasing function for all x. This gives,

$$\lim_{\delta\to\infty} \beta^\star(\delta) = \infty, \qquad \lim_{\delta\to\infty} \tau^\star(\delta) = 0. \tag{3.99}$$

Since $\lim_{\kappa\to\infty} \mathcal{M}_{\mathcal{L}}(x;\kappa) = \mathcal{L}(x)$ for all x, we deduce that,

$$\begin{aligned}(\alpha^\star,\mu^\star) &= \arg\min_{\alpha\in\mathbb{R}_+,\mu\in\mathbb{R}} \ \mathbb{E}_G\left[\exp\left(\varepsilon_{\rm tr}\sqrt{\alpha^2+\mu^2} - \mu + \sqrt{\alpha^2+\mu^2}G\right)\right]\\ &= \arg\min_{\alpha\in\mathbb{R}_+,\mu\in\mathbb{R}} \ \exp\left(\varepsilon_{\rm tr}\sqrt{\alpha^2+\mu^2} - \mu + (\alpha^2+\mu^2)/2\right)\\ &= \arg\min_{\alpha\in\mathbb{R}_+,\mu\in\mathbb{R}} \ \varepsilon_{\rm tr}\sqrt{\alpha^2+\mu^2} - \mu + (\alpha^2+\mu^2)/2,\end{aligned}$$

which results in $(\alpha^\star,\mu^\star) = (0, 1-\varepsilon_{\rm tr})$. Plugging these in (3.90), we derive the following for the large sample-size limit of the generalization error of adversarial training, conditioned on $\varepsilon_{\rm tr} < 1$,

$$\lim_{\delta\to\infty} \mathcal{E}^{\rm GMM}_{\ell_2,\varepsilon_{\rm ts}}(\widehat{\boldsymbol{\theta}}_n) = Q\Big(1-\varepsilon_{\rm ts}\Big). \tag{3.100}$$

On the other hand, based on [83], the Bayes adversarial error for isotropic GMM is derived as follows,

$$\mathcal{E}^{\mathrm{GMM}}_{\ell_2,\varepsilon_{\mathrm{ts}}}(\mathrm{OPT}) = Q\left(\min_{\|\mathbf{z}\|_q \le \varepsilon_{\mathrm{ts}}} \|\boldsymbol{\theta}^\star - \mathbf{z}\|_2\right).$$

In particular, noting that $\|\boldsymbol{\theta}^\star\|_2 \xrightarrow{P} 1$ (by Assumption 3.2.2) and $q = 2$, we find that the Bayes adversarial error in this case is

$$\mathcal{E}^{\mathrm{GMM}}_{\ell_2,\varepsilon_{\mathrm{ts}}}(\mathrm{OPT}) \xrightarrow{P} Q\Big(\max\{1-\varepsilon_{\mathrm{ts}}, 0\}\Big). \tag{3.101}$$

Comparing (3.101) with (3.100) reveals that in the infinite sample size limit, if $\varepsilon_{\mathrm{ts}} < 1$, by choosing any $\varepsilon_{\mathrm{tr}} < 1$, the test error of adversarial training reaches the Bayes adversarial error. As a remark, it can be readily shown that for the general case of $\|\boldsymbol{\theta}^\star\|_2 \xrightarrow{P} c$, the same results hold for $\varepsilon_{\mathrm{tr}} < c$ and $\varepsilon_{\mathrm{ts}} < c$.

Chapter 4

Generalization and Optimization in Interpolating Neural Networks

4.1 Introduction

Neural networks have remarkable expressive capabilities and can memorize a complete dataset even with mild overparameterization. In practice, using gradient descent (GD) on neural networks with logistic or cross-entropy loss can result in the objective reaching zero training error and close to zero training loss. Zero training error, often referred to as "interpolating" the data, indicates perfect classification of the dataset. Despite their strong memorization ability, these networks also exhibit remarkable generalization capabilities to new data. This has motivated a surge of studies in recent years exploring the optimization and generalization properties of first-order gradient methods in overparameterized neural networks, with a specific focus in the so-called Neural Tangent Kernel (NTK) regime. In the NTK regime, the model operates as the first-order approximation of the network at a sufficiently large initialization or at the large-width limit [105, 109]. Prior works on this topic mostly focused on quadratic-loss minimization and their optimization/generalization

guarantees required network widths that increased polynomially with the sample size n. This, however, is not in line with practical experience. Improved results were obtained more recently by [110] who have investigated the optimization and generalization of ReLU neural networks with logistic loss, which is more suitable for classification tasks. Assuming that the NTK with respect to the model can interpolate the data (i.e. separate them with positive margin γ), they showed through a Rademacher complexity analysis that GD on neural networks with polylogarithmic width can achieve generalization guarantees that decrease with the sample size n at a rate of $\tilde{O}(\frac{1}{\sqrt{n}})$.

In this chapter, we provide rate-optimal optimization and generalization analyses of GD for shallow neural networks of minimal width assuming that the model itself can interpolate the data. We focus on two-layer networks with smooth activations that can almost surely separate n training samples from the data distribution. Concretely, we consider a realizability condition where data and initialization are such that model weights can achieve arbitrarily small training error ε while their distance from initialization is $g(\varepsilon)$ for some function $g : \mathbb{R}_+ \to \mathbb{R}_+$. Under this condition, we demonstrate generalization guarantees of order $O(\frac{g(\frac{1}{T})^2}{n})$. More generally, for any iteration T of GD and assuming network width $m = \Omega(g(\frac{1}{T})^4)$, we obtain an expected test-loss rate $O(\frac{g(\frac{1}{T})^2}{T} + \frac{g(\frac{1}{T})^2}{n})$. Additional to the generalization bounds, we provide optimization guarantees under the same setting by showing that the training loss approaches zero at rate $O(\frac{g(\frac{1}{T})^2}{T})$. We note that these results are derived without NTK-type analyses. For demonstration and also for connection to prior works on neural-tangent data models, we specialize our generalization and optimization results to the class of NTK-separable data. We show this is possible because the NTK-data separability assumption implies our realizability condition holds. Thus, for logistic-loss minimization on NTK-separable data, we show that the expected test loss of GD is $\tilde{O}(\frac{1}{T} + \frac{1}{n})$ provided polylogarithmic number of neurons $m = \Omega(\log^4(T))$. This further suggests that a network of width $m = \Omega(\log^4(n))$, attains expected test loss

$\tilde{O}(\frac{1}{n})$ after $T \approx n$ iterations.

In contrast to prior optimization and generalization analyses that often depend on the NTK framework, which requires the first-order approximation of the model, we build on the algorithmic stability approach [13] for shallow neural-network models of finite width. Although the stability analysis has been utilized in previous studies to derive generalization bounds for (stochastic) gradient descent in various models, most results that are rate-optimal heavily rely on the convexity assumption. Specifically, the stability-analysis framework has been successful in achieving optimal generalization bounds for convex objectives in [14, 111, 112]. On the other hand, previous studies on non-convex objectives either resulted in suboptimal bounds or relied on assumptions that are not in line with the actual practices of neural network training. For instance, [17] derived a generalization bound of $O(\frac{T^{\beta c/(\beta c+1)}}{n})$ for general β-smooth and non-convex objectives, but this required a time-decaying step-size $\eta_t \leq c/t$, which can degrade the training performance. More recently, [16] explored the use of the stability approach specifically for logistic-loss minimization of a two-layer network. By refining the model-stability analysis framework introduced by [14], they derived generalization-error bounds provided the hidden width increases polynomially with the sample size. In comparison, our analysis leads to improved generalization and optimization rates and under standard separability conditions such as NTK-separability, only requires a polylogarithmic width for both global convergence and generalization.

Notation

We denote $[n] := \{1, 2, \cdots, n\}$. We use the standard notation $O(\cdot), \Omega(\cdot)$ and use $\tilde{O}(\cdot), \tilde{\Omega}(\cdot)$ to hide polylogarithmic factors. Occasionally we use $\lesssim$ to hide numerical constants. The Gradient and Hessian of a function $\Phi : \mathbb{R}^{d_1 \times d_2} \to \mathbb{R}$ with respect to the

ith input ($i = 1, 2$) are denoted by $\nabla_i \Phi$ and $\nabla_i^2 \Phi$, respectively. All logarithms are in base e. We use $\|\cdot\|$ for the ℓ_2 norm of vectors and the operator norm of matrices. We denote $[w_1, w_2] := \{w \,:\, w = \alpha w_1 + (1-\alpha) w_2, \alpha \in [0,1]\}$ the line segment between $w_1, w_2 \in \mathbb{R}^{d'}$.

4.2 Problem Setup

Given n i.i.d. samples $(x_i, y_i) \sim \mathcal{D}, i \in [n]$ from data distribution $\mathcal{D}$, we study unconstrained empirical risk minimization with objective $\widehat{F} : \mathbb{R}^{d'} \to \mathbb{R}$:

$$\min_{w \in \mathbb{R}^{d'}} \left\{ \widehat{F}(w) := \frac{1}{n} \sum_{i=1}^{n} \widehat{F}_i(w) = \frac{1}{n} \sum_{i=1}^{n} f\left(y_i \Phi\left(w, x_i\right)\right) \right\}. \tag{4.1}$$

This serves as a proxy for minimizing the *test loss* $F : \mathbb{R}^{d'} \to \mathbb{R}$:

$$F(w) := \mathbb{E}_{(x,y)\sim\mathcal{D}} \left[f\left(y \Phi(w, x)\right) \right]. \tag{4.2}$$

We introduce our assumptions on the data (x, y), the model $\Phi(\cdot, x)$, and the loss function $f(\cdot)$, below. We start by imposing the following mild assumption on the data distribution.

Assumption 4.2.1 (Bounded features)**.** *Assume any $(x, y) \sim \mathcal{D}$ has almost surely bounded features, i.e. $\|x\| \leq R$, and binary label $y \in \{\pm 1\}$.*

The model $\Phi : \mathbb{R}^{d'} \times \mathbb{R}^d \to \mathbb{R}$ is parameterized by trainable weights $w \in \mathbb{R}^{d'}$ and takes input $x \in \mathbb{R}^d$. For our main results, we assume Φ is a one-hidden layer neural-net of m neurons, i.e.

$$\Phi(w, x) := \frac{1}{\sqrt{m}} \sum_{j=1}^{m} a_j \, \sigma(\langle w_j, x \rangle), \tag{4.3}$$

where $\sigma : \mathbb{R} \to \mathbb{R}$ is the activation function, $w_j \in \mathbb{R}^d$ denotes the weight vector of the jth

hidden neuron and $\frac{a_j}{\sqrt{m}}, j \in [m]$ are the second-layer weights. For the second layer weights, we assume that they are fixed during training taking values $a_j \in \{\pm 1\}$. We assume that for half of second layer weights we have $a_j = 1$ and for the other half $a_j = -1$. On the other hand, all the first-layer weights are updated during training. Thus, the total number of trainable parameters is $d' = md$ and we denote $w = [w_1; w_2; \ldots; w_m] \in \mathbb{R}^{d'}$ the vector of trainable weights. Throughout, we make the following assumptions on the activation function.

Assumption 4.2.2 (Lipschitz and smooth activation)**.** *The activation function $\sigma : \mathbb{R} \to \mathbb{R}$ satisfies the following for non-negative constants ℓ, L:*

$$|\sigma'(u)| \leq \ell, \quad |\sigma''(u)| \leq L, \qquad \forall u \in \mathbb{R}.$$

We note that the smoothness assumption which is required by our framework excludes the use of ReLU. Examples of activation functions that satisfy the smoothness condition include Softplus $\sigma(u) = \log(1 + e^u)$, Gaussian error linear unit (GELU) $\sigma(u) = \frac{1}{2}u(1 + \text{erf}(\frac{u}{\sqrt{2}}))$, and Hyperbolic-Tangent where $\sigma(u) = \frac{e^u - e^{-u}}{e^u + e^{-u}}$. On the other hand, Lipschitz assumption is rather mild, since it is possible to restrict the parameter space to a bounded domain.

Next, we discuss conditions on the loss function. Of primal interest is the commonly used logistic loss function $f(u) = \log(1 + e^{-u})$. However, our results hold for a broader class of convex, non-negative and monotonically decreasing functions ($\lim_{u\to\infty} f(u) = 0$) that satisfy the following:

Assumption 4.2.3 (Lipschitz and smooth loss)**.** *The convex loss function $f : \mathbb{R} \to \mathbb{R}_+$ satisfies for all $u \in \mathbb{R}$*

4.2.3.A: *Lipschitzness: $|f'(u)| \leq G_f$.*

4.2.3.B: *Smoothness:* $f''(u) \leq L_f$.

Assumption 4.2.4 (Self-bounded loss)**.** *The convex loss function* $f : \mathbb{R} \to \mathbb{R}_+$ *is self-bounded with some constant* $\beta_f > 0$*, i.e.,* $|f'(u)| \leq \beta_f f(u), \forall u \in \mathbb{R}$.

The self-boundedness Assumption 4.2.4 is the key property of the loss that drives our analysis and justifies the polylogarithmic width requirement, as will become evident. Note that the logistic loss naturally satisfies Assumptions 4.2.3.A and 4.2.3.B (with $G_f = 1, L_f = 1/4$), as well as, Assumption 4.2.4 with $\beta_f = 1$. Other interesting examples of loss functions satisfying those assumptions include polynomial losses, with the tail behavior $f(u) = 1/u^\beta$ for $\beta > 0$, which we discuss in Remark 4.4.1. To lighten the notation and without loss of generality, we set $G_f = L_f = \beta_f = 1$ for the rest of the chapter. We remark that our training-loss results also hold for the exponential loss e^{-u}. The exponential loss is self-bounded and while it is not Lipschitz or smooth it satisfies a second-order self-bounded property $f''(u) \leq f(u)$, which we can leverage instead; see Appendix 4.8.1 for details.

4.3 Main Results

We present bounds on the train loss and generalization gap of gradient-descent (GD) under the setting of Section 4.2. Formally, GD with step-size $\eta > 0$ optimizes (6.1) by performing the following updates starting from an initialization w_0:

$$\forall t \geq 0 \ : \ w_{t+1} = w_t - \eta \nabla \widehat{F}(w_t).$$

4.3.1 Key properties

The key challenge in both the optimization and generalization analysis is the non-convexity of $f(y\Phi(\cdot, x))$, and consequently of the train loss $\widehat{F}(\cdot)$. Despite non-convexity, we derive bounds analogous to the convex setting, e.g. corresponding bounds on linear logistic regression in [113, 114, 112]. We show this is possible provided the loss satisfies the following key property, which we call *self-bounded weak convexity*.

Definition 4.3.1 (Self-bounded weak convexity)**.** We say a function $\widehat{F} : \mathbb{R}^{d'} \to \mathbb{R}$ is self-bounded weakly convex if there exists constant $\kappa > 0$ such that for all w,

$$\lambda_{\min}\left(\nabla^2 \widehat{F}(w)\right) \geq -\kappa\, \widehat{F}(w)\,. \tag{4.4}$$

Recall a function $G : \mathbb{R}^{d'} \to \mathbb{R}$ is weakly convex if $\exists \kappa \geq 0$ such that uniformly over all $w \in \mathbb{R}^{d'}$, $\lambda_{\min}\left(\nabla^2 G(w)\right) \geq -\kappa$. If $\kappa = 0$, the function is convex. Instead, property (4.4) lower bounds the curvature by $-\kappa\, G(w)$ that changes proportionally with the function value $G(w)$. We explain below how this is exploited in our setting.

To begin with, the following lemma shows that property (4.4) holds for the train loss under the setting of Section 4.2: training of a two-layer net with smooth activation and self-bounded loss. The lemma also shows that the gradient of the train loss is self bounded. Those two properties together summarize the key ingredients for which our analysis applies.

Lemma 4.3.1 (Key self-boundedness properties)**.** *Consider the setup of Section 4.2 and let Assumptions 4.2.1-4.2.2 hold. Further assume the loss is self-bounded as per Assumption 4.2.4. Then, the objective satisfies the following self-boundedness properties for its Gradient and Hessian:*

1. *Self-bounded gradient:* $\left\|\nabla \widehat{F}_i(w)\right\| \leq \ell R\, \widehat{F}_i(w), \quad \forall i \in [n].$

2. *Self-bounded weak convexity:* $\lambda_{\min}\left(\nabla^2 \widehat{F}(w)\right) \geq -\frac{LR^2}{\sqrt{m}}\widehat{F}(w)$.

Both of these properties follow from the self-boundedness of the convex loss f combined with Lipshitz and smoothness of σ. The self-boundedness of the gradient is used for generalization analysis and in particular in obtaining the model stability bound. The self-bounded weak convexity plays an even more critical role for our optimization and generalization results. In particular, the wider the network the closer the loss to having convex-like properties. Moreover, the "self-bounded" feature of this property provides another mechanism that favors convex-like optimization properties of the loss. To see this, consider the minimum Hessian eigenvalue $\lambda_{\min}(\nabla^2 \widehat{F}(w_t))$ at gradient descent iterates $\{w_t\}_{t\geq 1}$: As training progresses, the train loss $\widehat{F}(w_t)$ decreases, and thanks to the self-bounded weak convexity property, the gap to convexity also decreases. We elaborate on the role of self-bounded weak convexity in our proofs in Section 4.5.

4.3.2 Training loss

We begin with a general bound on the training loss and the parameter's norm, which is also required for our generalization analysis.

Theorem 4.3.1 (Training loss – General bound)**.** *Suppose Assumptions 4.2.1-4.2.4 hold. Fix any training horizon $T \geq 0$ and any step-size $\eta \leq 1/L_{\widehat{F}}$ where $L_{\widehat{F}}$ is the objective's smoothness parameter. Assume any $w \in \mathbb{R}^{d'}$ and hidden-layer width m such that $\|w - w_0\|^2 \geq \max\{\eta T \widehat{F}(w), \eta \widehat{F}(w_0)\}$ and $m \geq 18^2 L^2 R^4 \|w - w_0\|^4$. Then, the training loss and the parameters' norm satisfy*

$$\widehat{F}(w_T) \;\leq\; \frac{1}{T}\sum_{t=1}^{T} \widehat{F}(w_t) \;\leq\; 2\widehat{F}(w) + \frac{5\|w - w_0\|^2}{2\eta T}, \tag{4.5}$$

$$\forall t \in [T] \quad : \quad \|w_t - w_0\| \;\leq\; 4\|w - w_0\|.$$

A few remarks are in place regarding the theorem. First, Eq. (4.5) upper bounds the running average (also known as regret) of train loss for iterations $1, \dots, T$ by the value, at an arbitrarily chosen point w, of a ridge-regularized objective with regularization parameter inversely proportional to ηT. Because of smoothness and Lipschitz Assumption 4.2.3 of f, it turns out that the training objective is $L_{\widehat{F}}$-smooth. Hence, by the descent lemma of GD for smooth functions, the same upper bound holds in Eq. (4.5) for the value of the loss at time T, as well. Moreover, the theorem provides a uniform upper bound of the norm of all GD iterates in terms of $\|w - w_0\|$. Notably, and despite the non-convexity in our setting, our bounds are same up to constants to analogous bounds for logistic linear regression in [114, 112]. As discussed in Sec. 4.3.1 this is possible thanks to the self-bounded weak convexity property.

The condition $m \gtrsim \|w - w_0\|^4$ on the norm of the weights controls the maximum deviations of weights w from initialization (with respect to network width) required for our results to guarantee arbitrarily small train loss. Specifically, to get the most out of Theorem 4.3.1 we need to choose appropriate w that satisfies both the condition $m \gtrsim \|w - w_0\|^4$ and keeps the associated ridge-regularized loss $\widehat{F}(w) + \|w - w_0\|^2/(\eta T)$ small. This combined requirement is formalized in the neural-net realizability Assumption 4.3.1 below. As we will discuss later in Section 4.4, this assumption translates into an assumption on the underlying data distribution that ultimately enables the application of Theorem 4.3.1 to achieve vanishing training error.

Assumption 4.3.1 (NN–Realizability)**.** *There exists a decreasing function $g : \mathbb{R}_+ \to \mathbb{R}_+$ which measures the norm of deviations from initialization of models that achieve arbitrarily small training error.*

Formally, for almost surely all n training samples and for any sufficiently small $\varepsilon > 0$

there exists $w^{(\varepsilon)} \in \mathbb{R}^{d'}$ *such that*

$$\widehat{F}(w^{(\varepsilon)}) \leq \varepsilon, \quad \textit{and} \quad g(\varepsilon) = \left\| w^{(\varepsilon)} - w_0 \right\|.$$

Since Assumption 4.3.1 holds for arbitrarily small ε, it guarantees that the model has enough capacity to interpolate the data, i.e., attain train error that is arbitrarily small (ε). Additionally, this is accomplished for model weights whose distance from initialization is managed by the function $g(\varepsilon)$. By using these model weights to select w in Theorem 4.3.1 we obtain train loss bounds for interpolating models.

Theorem 4.3.2 (Training loss under interpolation)**.** *Let Assumptions 4.2.1-4.3.1 hold. Let* $\eta \leq \min\{\frac{1}{L_{\widehat{F}}}, g(1)^2, \frac{g(1)^2}{\widehat{F}(w_0)}\}$ *and assume the width satisfies* $m \geq 18^2 L^2 R^4 \, g(\frac{1}{T})^4$ *for a fixed training horizon* T*. Then,*

$$\begin{aligned} \widehat{F}(w_T) \; &\leq \frac{2}{T} + \frac{5\, g(\frac{1}{T})^2}{2\eta T}, \\ \forall t \in [T] \; : \quad \|w_t - w_0\| \; &\leq \; 4\, g(\frac{1}{T}). \end{aligned} \tag{4.6}$$

To interpret the theorem's conclusions suppose that the function $g(\cdot)$ of Assumption 4.3.1 is at most logarithmic; i.e., $g(\frac{1}{T}) = O(\log(T))$. Then, Theorem 4.3.2 implies that $m = \Omega(\log^4(T))$ neurons suffice to achieve train loss $\tilde{O}(\frac{1}{T})$ while GD iterates at all iterations satisfy $\|w_t - w_0\| = O(\log(T))$. In Section 4.4 (see also Remark 4.3.1), we will give examples of data separability conditions that guarantee the desired logarithmic growth of $g(\cdot)$ for logistic loss minimization, which in turn imply the favorable convergence guarantees described above. Under the same conditions we will show that the step-size requirement simplifies to $\eta \leq \min\{3, 1/L_{\widehat{F}}\}$ (see Corollary 4.4.1). Finally, we remark that Theorem 4.3.2 provides sufficient parameterization conditions under which GD with $T = \tilde{\Omega}(n)$ iterations finds weights w_T that yield an interpolating classifier and thus,

achieve zero training error. To see this, assume logistic loss and observe setting $T \gtrsim n$ in Eq. (4.6) gives $\widehat{F}(w_T) \leq \log(2)/n$. This in turn implies that every sample loss satisfies $\widehat{F}_i(w_T) \leq \log(2)$, equivalently $y_i = \text{sign}\left(\Phi(w_T, x_i)\right)$.

4.3.3 Generalization

Our main result below bounds the generalization gap of GD for training two-layer nets with self-bounded loss functions. We remark that all expectations that appear below are over the training set.

Theorem 4.3.3 (Generalization gap – General bound)**.** *Suppose Assumptions 4.2.1-4.2.4 hold. Fix any time horizon $T \geq 1$ and any step size $\eta \leq 1/L_{\widehat{F}}$ where $L_{\widehat{F}}$ is the objective's smoothness parameter. Let any $w \in \mathbb{R}^{d'}$ such that $\|w - w_0\|^2 \geq \max\{\eta T\, \widehat{F}(w), \eta \widehat{F}(w_0)\}$. Suppose hidden-layer width m satisfies $m \geq 64^2 L^2 R^4 \|w - w_0\|^4$. Then, the generalization gap of GD at iteration T is bounded as*

$$\mathbb{E}\Big[F(w_T) - \widehat{F}(w_T)\Big] \leq \frac{8\ell^2 R^2}{n}\, \mathbb{E}\left[\eta T\, \widehat{F}(w) + 2\|w - w_0\|^2\right].$$

A few remarks regarding the theorem are in place. The theorem's assumptions are similar to those in Theorem 4.3.1, which bounds the training loss. The condition $\|w - w_0\|^2 \geq \max\{\eta T \widehat{F}(w), \eta \widehat{F}(w_0)\}$ needs to hold almost surely over the training data, which is non-restrictive, as in later applications of the theorem, the choice of w arises from Assumption 4.3.1. The condition $m \geq 64^2 L^2 R^4 \|w - w_0\|^4$ on the width of the network, is also the same as that of Theorem 4.3.1 but with a larger constant. This means that the last-iterate train loss bound from Theorem 4.3.1 (Eq. (4.5)) holds under the setting of Theorem 4.3.3. Hence, it applies to the expected train loss $\mathbb{E}[\widehat{F}(w_T)]$ and, combined with the generalization-gap bound, yields a bound on the expected test loss $\mathbb{E}[F(w_T)]$.

To optimize the bound, a proper w must be selected by minimizing the population

version of a ridge-regularized training objective. In interpolation settings, the procedure for selecting w follows the same guidelines as in Assumption 4.3.1 and in a similar style as obtaining Theorem 4.3.2.

Theorem 4.3.4 (Generalization gap under interpolation)**.** *Let Assumptions 4.2.1-4.3.1 hold. Fix $T \geq 1$ and let $m \geq 64^2 L^2 R^4 \, g(\frac{1}{T})^4$. Then, for any $\eta \leq \min\{\frac{1}{L_{\widehat{F}}}, g(1)^2, \frac{g(1)^2}{\widehat{F}(w_0)}\}$ the expected generalization gap at iteration T satisfies*

$$\mathbb{E}\Big[F(w_T) - \widehat{F}(w_T)\Big] \leq \frac{24\ell^2 R^2 \, g(\frac{1}{T})^2}{n}\,. \tag{4.7}$$

Note the width condition is similar in order to that of Theorem 4.3.2. Thus, provided $g(\frac{1}{T}) \lesssim \log(T)$ (see Remark 4.3.1 and Section 4.4 for examples), we have generalization gap of order $\tilde{O}(\frac{1}{n})$ with $m = \Omega(\log^4(T))$ neurons. Combined with the training loss guarantees from Theorem 4.3.2, we have test loss rate $\tilde{O}(\frac{1}{T} + \frac{1}{n})$. This further implies that with $m \approx \log^4(n)$ neurons and $T = n$ iterations, the test loss reaches the optimal rate of $\tilde{O}(\frac{1}{n})$. On the other hand, previous stability-based generalization bounds (e.g., [16]) required polynomial width $m \gtrsim T^2$ and eventually obtained sub-optimal generalization rates of order $O(\frac{T}{n})$. We further discuss the technical novelties resulting in these improvements in Section 4.5.

Remark 4.3.1 (Example: Linearly-separable data). Consider logistic-loss minimization, tanh activation $\sigma(u) = \frac{e^u - e^{-u}}{e^u + e^{-u}}$ and data distribution that is linearly separable with margin γ, i.e., for almost surely all n samples there exists unit-norm vector $v^\star \in \mathbb{R}^d$ such that $\forall i \in [n] : y_i \langle v^\star, x_i \rangle \geq \gamma$. We initialize the weights to zero, i.e. $w_0 = 0$ and show that the realizability Assumption 4.3.1 naturally holds in this setting. To see this, for any fixed $\varepsilon > 0$, set $\alpha = \frac{2(\log(1/\varepsilon))}{\gamma\sqrt{m}}$ and assume $m \geq 4\log^2(1/\varepsilon)$. With this choice, select weights $w_j^{(\varepsilon)} := \alpha v^\star, a_j = \frac{1}{\sqrt{m}}$ for $j \in [1, \cdots, \frac{m}{2}]$ and $w_j^{(\varepsilon)} := -\alpha v^\star, a_j = \frac{-1}{\sqrt{m}}$ for $j \in \{\frac{m}{2}+1, \cdots, m\}$.

Then, the model output for any sample (x_i, y_i) satisfies

$$y_i\Phi(w^{(\varepsilon)}, x_i) = \frac{y_i\sqrt{m}}{2}\left(\sigma(\alpha\langle v^\star, x_i\rangle) - \sigma(-\alpha\langle v^\star, x_i\rangle)\right) = y_i\sqrt{m}\sigma(\alpha\langle v^\star, x_i\rangle)$$
$$\geq \sqrt{m}\sigma(\alpha\gamma) \geq \frac{\sqrt{m}}{2}\alpha\gamma = \log(\frac{1}{\varepsilon})$$

where the first equality uses the fact that *tanh* is odd, the first inequality follows by the increasing nature of *tanh* and data separability, and the last inequality follows since $\alpha\gamma \leq 1$ and $\sigma(u) \geq u/2$ for all $u \in [0, 1]$. Thus, the loss satisfies $\widehat{F}(w^{(\varepsilon)}) \leq \varepsilon$ since for the logistic function $\log(1 + e^u) \leq e^u$. Moreover, our choice of α implies $g(\varepsilon) = \|w^{(\varepsilon)} - w_0\| = \|w^{(\varepsilon)}\| = \alpha\sqrt{m} = 2\log(1/\varepsilon)/\gamma$. To conclude, the NN-Realizability Assumption 4.3.1 holds with $g(\varepsilon) = 2\log(1/\varepsilon)/\gamma$ and thus applying Theorems 4.3.2, 4.3.4 shows that with $m = \Omega(\log^4(T))$ neurons, the training loss and generalization gap are bounded by $\tilde{O}(\frac{1}{\gamma^2 T})$ and $\tilde{O}(\frac{1}{\gamma^2 n})$, respectively. We note that the same conclusion as above holds for other smooth activations such as Softmax or GELU.

4.4 On Realizability of NTK-Separable Data

In this section, we interpret our results for NTK-separable data by showing that our realizability condition holds for this class. We recall the definition of NTK-separability below [19, 115].

Assumption 4.4.1 (Separability by NTK)**.** *For almost surely all n training samples from the data distribution there exists $w^\star \in \mathbb{R}^{d'}$ and $\gamma > 0$ such that $\|w^\star\| = 1$ and for all $i \in [n]$,*

$$y_i\left\langle \nabla_1\Phi(w_0, x_i), w^\star \right\rangle \geq \gamma. \tag{4.8}$$

We also assume a bound on the model's output at initialization. Similar assumptions, but for the value of the loss, also appear in prior works that study generalization using the algorithmic stability framework [18, 116].

Assumption 4.4.2 (Initialization bound)**.** *There exists parameter C such that $\forall i \in [n]$: $|\Phi(w_0, x_i)| \leq C$, for almost surely all n training samples from the data distribution*

The next proposition relates the NTK-separability assumption to our realizability assumption. The proofs for this section are given in Appendix 4.8.3.

Proposition 4.4.1 (Realizability of NTK-separable data)**.** *Let Assumptions 4.2.1-4.2.2,4.4.1-4.4.2 hold. Assume $f(\cdot)$ to be the logistic loss. Fix $\varepsilon > 0$ and let $m \geq \frac{L^2 R^4}{4\gamma^4 C^2}(2C + \log(1/\varepsilon))^4$. Then the realizability Assumption 4.3.1 holds with $g(\varepsilon) = \frac{1}{\gamma}(2C + \log(1/\varepsilon))$. In other words, there exists $w^{(\varepsilon)}$ such that*

$$\widehat{F}(w^{(\varepsilon)}) \leq \varepsilon, \quad \textit{and} \quad \left\|w^{(\varepsilon)} - w_0\right\| = \frac{1}{\gamma}\left(2C + \log(1/\varepsilon)\right). \tag{4.9}$$

Having established realizability, the following is an immediate corollary of the general results presented in the last section.

Corollary 4.4.1 (Results under NTK-separability)**.** *Let Assumptions 4.2.1-4.2.2,4.4.1-4.4.2 hold and assume logistic loss. Suppose $m \geq \frac{64^2 L^2 R^4}{\gamma^4}(2C + \log(T))^4$ for a fixed training horizon T. Then for any $\eta \leq \min\{3, \frac{1}{L_{\widehat{F}}}\}$, the training loss and generalization gap are bounded as follows:*

$$\widehat{F}(w_T) \leq \frac{5(2C + \log(T))^2}{\gamma^2 \eta T},$$
$$\mathbb{E}\left[F(w_T) - \widehat{F}(w_T)\right] \leq \frac{24\ell^2 R^2}{\gamma^2 n}(2C + \log(T))^2.$$

A few remarks are in place regarding the corollary. By Corollary 4.4.1, we can conclude

that the expected generalization rate of GD on logistic loss and NTK-separable data as per Assumption 4.4.1 is $\tilde{O}(\frac{1}{n})$ provided width $m = \Omega(\log^4(T))$. Moreover, the expected training loss is $\mathbb{E}[\widehat{F}(w_T)] = (\frac{1}{T})$. Thus, the expected test loss after T steps is $(\frac{1}{T} + \frac{1}{n})$. In particular for $T = \Omega(n)$, the expected test loss becomes $\tilde{O}(\frac{1}{n})$. This rate is optimal with respect to sample size and only requires polylogarithmic hidden width with respect to n, specifically, $m = \Omega(\log^4(n))$. Notably, it represents an improvement over prior stability results, e.g., [16] which required polynomial width and yielded suboptimal generalization rates of order $O(T/n)$. It is worth noting that the test loss bound's dependence on the margin, particularly the $\frac{1}{\gamma^2 n}$-rate obtained in our analysis, bears similarity to the corresponding results in the convex setting of linearly separable data recently established in [114, 112]. Additionally, our results improve upon corresponding bounds for neural networks obtained via Rademacher complexity analysis [110] which yield generalization rates $\tilde{O}(\frac{1}{\sqrt{n}})$. Moreover, these works have a γ^{-8} dependence on margin for the minimum network width, whereas in Corollary 4.4.1 this is reduced to γ^{-4}. We also note that in general, both γ and C may depend on the data distribution, the data dimension, or the nature of initialization. This is demonstrated in the next section where we apply the corollary above to the noisy XOR data distribution and Gaussian initialization.

Remark 4.4.1 (Benefits of exponential tail). We have stated Corollary 4.4.1 for the logistic loss, which has an exponential tail behavior. For general self-bounded loss functions and by following the same steps, we can show a bound on generalization gap of order $O(\frac{1}{n}(f^{-1}(\frac{1}{T}))^2)$ provided $m = \Omega((f^{-1}(\frac{1}{T}))^4)$. Hence, the tail behavior of f controls both the generalization gap and minimum width requirement. In particular, under Assumption 4.4.1, polynomial losses with tail behavior $f(u) \sim 1/u^\beta$ result in generalization gap $O(T^{2/\beta}/n)$ for $m = \Omega(T^{4/\beta})$. Thus, increasing the rate of decay β for the loss, improves both bounds on generalization and width. This suggests the benefits of self-bounded fast-decaying losses such as exponentially-tailed loss functions for which the dependence

on T is indeed only logarithmic.

Example: Noisy XOR data

Next, we specialize the results of the last section to the noisy XOR data distribution [117] and derive the corresponding margin and test-loss bounds. Consider the following 2^d points,

$$x_i = (x_i^1, x_i^2, \cdots, x_i^d) \in \{(1,0),(0,1),(-1,0),(0,-1)\} \times \{-1,1\}^{d-2},$$

where $\times$ denotes the Cartesian product and the labels are determined as $y_i = -1$ if $x_i^1 = 0$ and $y_i = 1$ if $x_i^1 = \pm 1$. Moreover, consider normalization $\overline{x}_i = \frac{1}{\sqrt{d-1}} x_i$ so that $R = 1$. The noisy XOR data distribution is the uniform distribution over the set with elements $(\overline{x}_i, y_i)$. For this dataset and Gaussian initialization, [110] have shown for ReLU activation that the NTK-separability assumption holds with margin $\gamma = \Omega(1/d)$. In the next result, we compute the margin for activation functions that are convex, Lipshitz and locally strongly convex.

Proposition 4.4.2 (Margin)**.** *Consider the noisy XOR data $(\overline{x}_i, y_i) \in \mathbb{R}^d \times \{\pm 1\}$. Assume the activation function is convex, ℓ-Lipschitz and μ-strongly convex in the interval $[-2,2]$ for some $\mu > 0$, i.e., $\min_{t\in[-2,2]} \sigma''(t) \geq \mu$. Moreover, assume Gaussian initialization $w_0 \in \mathbb{R}^{d'}$ with entries iid $N(0,1)$. If $m \geq \frac{80^2 d^3 \ell^2}{2\mu^2} \log(2/\delta)$, then with probability at least $1-\delta$ over the initialization, the NTK-separability Assumption 4.4.1 is satisfied with margin $\gamma = \frac{\mu}{80d}$.*

An interesting example of an activation function that satisfies the mentioned assumptions is the Softplus activation where $\sigma(u) = \log(1+e^u)$. This activation function has $\mu = 0.1$ and $\ell = 1$, and it is also smooth with $L = 1/4$. Therefore, the results on

generalization and training loss presented in Corollary 4.4.1 hold for it. For noisy XOR data, Proposition 4.4.2 shows the margin in Assumption 4.4.1 is $\gamma \gtrsim 1/d$. Additionally, for standard Gaussian initialization we have by Lemma 4.8.9 that with high-probability the initialization bound in Assumption 4.4.2 satisfies $C \lesssim \sqrt{d}$. Putting these together, and applying Corollary 4.4.1 shows that GD with n training samples reaches test loss rate $\tilde{O}(\frac{d^3}{n})$ after $T \approx n$ iterations and given $m = \tilde{\Omega}(d^6)$ neurons. It is worth noting that the number of training samples can be exponentially large with respect to d. In this case the minimum width requirement is only polylogarithmic in n.

4.5 Proof Sketches

We discuss here high-level proof ideas for both optimization and generalization bounds of Theorems 4.3.1 and 4.3.3. Formal proofs are deferred to Appendices 4.8.1 and 4.8.2.

4.5.1 Training loss

As already discussed in Section 4.3.1, the key insight we use to obtain bounds that are analogous to results for optimizing convex objectives, is to exploit the self-bounded weak convexity property of the objective in Eq. (4.4). Thanks to this property, the Hessian minimum eigenvalue $\lambda_{\min}(\nabla^2 \widehat{F}(w_t))$ becomes less negative at the same rate at which the train loss $\widehat{F}(w_t)$ decreases.

The technical challenge at formalizing this intuition arises as follows. Controlling the rate at which $\widehat{F}(w_t)$ converges to $\widehat{F}(w)$ for the theorem's w requires controlling the Hessian at *all* intermediate points $w_{\alpha t} := \alpha w_t + (1-\alpha)w, \alpha \in [0,1]$ between w and GD iterates w_t. This is due to Taylor's theorem used to relate $\widehat{F}(w_t)$ to the target value $\widehat{F}(w)$

as follows:

$$\widehat{F}(w) \geq \widehat{F}(w_t) + \left\langle \nabla \widehat{F}(w_t), w - w_t \right\rangle + \frac{1}{2} \lambda_{\min} \left(\nabla^2 \widehat{F}(w_{\alpha t}) \right) \left\| w - w_t \right\|^2.$$

Thus from self-bounded weak convexity, to control the last term above we need to control $\widehat{F}(w_{\alpha t})$ for any intermediate point $w_{\alpha t}$ along the GD trajectory. This is made possible by establishing the following generalized local quasi-convexity property.

Proposition 4.5.1 (Generalized Local Quasi-Convexity)**.** *Suppose $\widehat{F} : \mathbb{R}^{d'} \to \mathbb{R}$ satisfies the self-bounded weak convexity property in Eq. (4.4) with parameter κ. Let $w_1, w_2 \in \mathbb{R}^{d'}$ be two arbitrary points with distance $\|w_1 - w_2\| \leq D < \sqrt{2/\kappa}$. Set $\tau := (1 - \kappa D^2/2)^{-1}$. Then,*

$$\max_{v \in [w_1, w_2]} \widehat{F}(v) \leq \tau \cdot \max\{\widehat{F}(w_1), \widehat{F}(w_2)\}. \tag{4.10}$$

Recall that quasi-convex functions satisfy Eq. (4.10) with $\tau = 1$ and D can be unboundedly large. The Proposition 4.5.1 indicates that our neural-net objective function is approximately quasi-convex (since $\tau > 1$) and this property holds locally, i.e. provided that w_1, w_2 are sufficiently close.

Applying (4.10) for $w_1 = w_t, w_2 = w$ allows controlling $\widehat{F}(w_{\alpha t})$ in terms of the train loss $\widehat{F}(w_t)$ and the target loss $\widehat{F}(w)$. The only additional requirement in Proposition 4.5.1 for this to hold is that

$$1/\kappa \propto \sqrt{m} \gtrsim \|w_t - w\|^2. \tag{4.11}$$

This condition exactly determines the required neural-net width. Formally, we have the following.

Corollary 4.5.1 (GLQC of sufficiently wide neural nets)**.** *Let Assumptions 4.2.1,4.2.2, 4.2.4 hold. Fix arbitrary $w_1, w_2 \in \mathbb{R}^{d'}$, any constant $\lambda > 1$, and m large enough such that $\sqrt{m} \geq \lambda \frac{LR^2}{2} \|w_1 - w_2\|^2$. Then,*

$$\max_{v \in [w_1, w_2]} \widehat{F}(v) \leq (1 - 1/\lambda)^{-1} \cdot \max\{\widehat{F}(w_1), \widehat{F}(w_2)\}. \tag{4.12}$$

To conclude, using Corollary 4.5.1, we can show the regret bound in Eq. (4.5) provided (by (4.11)) that $\sqrt{m} \gtrsim \|w_t - w\|^2$ is true for all $t \in [T]$. To make the width requirement independent of w_t, we then use a recursive argument to prove that $\|w_t - w\| \leq 3\|w - w_0\|$. These things put together, lead to the parameter bound $\|w_t - w_0\| \leq 4\|w - w_0\|$ and the width requirement $\sqrt{m} \gtrsim \|w - w_0\|^2$ in the theorem's statement. We note that the GLQC property is also crucially required for the generalization analysis which we discuss next.

4.5.2 Generalization gap

We bound the generalization gap using stability analysis [13, 17]. In particular, we use [14, Thm. 2] that relates the generalization gap to the "on average model stability". Formally, let $w_t^{\neg i}$ denote the t-th iteration of GD on the leave-one-out loss $\widehat{F}^{\neg i}(w) := \frac{1}{n} \sum_{j \neq i} \widehat{F}_j(w)$. As before, w_t denotes the GD output on full-batch loss $\widehat{F}$. We will use the fact (see Corollary 4.8.3) that $f(y\Phi(\cdot, x))$ is $G_{\widehat{F}}$-Lipschitz with $G_{\widehat{F}} = \ell R$ under Assumptions 4.2.2 and 4.2.3.A. Then, using [14, Thm. 2(a)] (cf. Lemma 4.8.6) it holds that

$$\mathbb{E}\Big[F(w_T) - \widehat{F}(w_T)\Big] \leq 2G_{\widehat{F}}\, \mathbb{E}\Big[\frac{1}{n} \sum_{i=1}^{n} \|w_T - w_T^{\neg i}\|\Big]. \tag{4.13}$$

In order to bound the on-average model-stability term on the right-hand side above we need to control the degree of expansiveness of GD. Recall that for convex objectives GD

is non-expansive (e.g. [17]), that is $\|(w - \eta\nabla\widehat{F}(w)) - (w' - \eta\nabla\widehat{F}(w'))\| \leq \|w - w'\|$ for any w, w'. For the non-convex objective in our setting, the lemma below establishes a generalized non-expansiveness property via leveraging the structure of the objective's Hessian for the two-layer net.

Lemma 4.5.1 (GD-Expansiveness)**.** *Let Assumptions 4.2.1 and 4.2.2 hold. For any $w, w' \in \mathbb{R}^{d'}$, any step-size $\eta > 0$, and $w_\alpha := \alpha w + (1-\alpha)w'$ it holds for $H(w) := \eta\frac{LR^2}{\sqrt{m}}\widehat{F}'(w) + \max\left\{1, \eta\ell^2R^2\widehat{F}''(w)\right\}$ that*

$$\left\|\left(w - \eta\nabla\widehat{F}(w)\right) - \left(w' - \eta\nabla\widehat{F}(w')\right)\right\| \leq \max_{\alpha\in[0,1]} H(w_\alpha)\ \|w - w'\|\,,$$

where we define $\widehat{F}'(w) := \frac{1}{n}\sum_{i=1}^n |f'(y_i\Phi(w, x_i))|$ and $\widehat{F}''(w) := \frac{1}{n}\sum_{i=1}^n f''(y_i\Phi(w, x_i))$.

This lemma can be further simplified for the class of self-bounded loss functions. Specifically, using $|f'(u)| \leq f(u)$ and $f''(u) \leq 1$ from Assumptions 4.2.4 and 4.2.3.B, we immediately deduce the following.

Corollary 4.5.2 (Expansiveness for self-bounded losses)**.** *In the setting of Lemma 4.5.1, further assume the loss satisfies Assumptions 4.2.3.B and 4.2.4. Provided $\eta \leq 1/(\ell^2R^2)$, it holds for all $w, w' \in \mathbb{R}^{d'}$ that*

$$\left\|\left(w - \eta\nabla\widehat{F}(w)\right) - \left(w' - \eta\nabla\widehat{F}(w')\right)\right\| \leq \left(1 + \eta\frac{LR^2}{\sqrt{m}}\max_{\alpha\in[0,1]}\widehat{F}(w_\alpha)\right)\left\|w - w'\right\|. \quad (4.14)$$

In Eq. (4.14) the expansiveness is weaker than in a convex scenario, where the coefficient would be 1 instead of $1 + \frac{\eta LR^2}{\sqrt{m}}\max_{\alpha\in[0,1]}\widehat{F}(w_\alpha)$. However, for self-bounded losses (i.e. $|f'(u)| \leq f(u)$) the "gap to convexity" $\frac{\eta LR^2}{\sqrt{m}}\max_{\alpha\in[0,1]}\widehat{F}(w_\alpha)$ in Corollary 4.5.2 is better than the gap from Lemma 4.5.1 for 1-Lipschitz losses (i.e. $|f'(u)| \leq 1$), which would be $\frac{\eta LR^2}{\sqrt{m}}$. Indeed, after unrolling the GD iterates, the latter eventually leads to polynomial width requirements [16].

Instead, to obtain a polylogarithmic width, we use the expansiveness bound in Eq. (4.14) for self-bounded losses together with the generalized-local quasi-convexity property in Corollary 4.5.1 as follows. From Corollary 4.5.1, if m is large enough such that

$$\sqrt{m} \geq LR^2 \|w_t - w_t^{\neg i}\|^2, \qquad \forall t \in [T],\ \forall i \in [n],$$

then Eq. (4.12) holds on the GD path. This further simplifies the result of Corollary 4.5.2 applied for $w = w_t, w' = w_t^{\neg i}$ into

$$\left\| (w_t - \eta \nabla \widehat{F}^{\neg i}(w_t)) - (w_t^{\neg i} - \eta \nabla \widehat{F}^{\neg i}(w_t^{\neg i})) \right\| \leq \widetilde{H}_t^i \left\| w_t - w_t^{\neg i} \right\|,$$

where $\widetilde{H}_t^i := 1 + \frac{2\eta LR^2}{\sqrt{m}} \max\{\widehat{F}^{\neg i}(w_t), \widehat{F}^{\neg i}(w_t^{\neg i})\}$. Now from the optimization analyses in Sec. 4.5.1, we know intuitively that $\widehat{F}^{\neg i}(w_t) \leq \widehat{F}(w_t)$ decays at rate $\tilde{O}(1/t)$; thus, so does $\widehat{F}^{\neg i}(w_t^{\neg i})$. Therefore, for all $i \in [n]$ the expansivity coefficient $\widetilde{H}_t^i$ in the above display is decaying to 1 as GD progresses.

To formalize all these and connect them to the model-stability term in (4.13), note using triangle inequality and the Gradient Self-boundedness property of Lemma 4.3.1 that

$$\left\| w_{t+1} - w_{t+1}^{\neg i} \right\| \leq \left\| (w_t - \eta \nabla \widehat{F}^{\neg i}(w_t)) - (w_t^{\neg i} - \eta \nabla \widehat{F}^{\neg i}(w_t^{\neg i})) \right\| + \frac{\eta \ell R}{n} \widehat{F}_i(w_t)\,.$$

Unrolling this display over $t \in [T]$, averaging over $i \in [n]$, and using our expansiveness bound above we show in Appendix 4.8.2 the following bound for the model stability term

$$\frac{1}{n} \sum_{i=1}^{n} \left\| w_T - w_T^{\neg i} \right\| \leq \frac{\eta \ell R e^{\beta}}{n} \sum_{t=0}^{T-1} \widehat{F}(w_t)\,, \tag{4.15}$$

where $\beta \lesssim \left(\sum_{t=1}^T \widehat{F}(w_t) + \sum_{t=1}^T \widehat{F}^{\neg i}(w_t^{\neg i}) \right) / \sqrt{m}$. But, we know from training-loss bounds in Theorem 4.3.1 that $\sum_{t=1}^T \widehat{F}(w_t) \lesssim \|w - w_0\|^2$ (and similar for $\sum_{t=1}^T \widehat{F}^{\neg i}(w_t^{\neg i})$). Thus,

$\beta \lesssim \|w - w_0\|^2/\sqrt{m}$. At this point, the theorem's conditions guarantees $\sqrt{m} \gtrsim \|w - w_0\|^2$, so that $\beta = O(1)$. Plugging back in (4.15) we conclude with the following stability bound: $\frac{1}{n}\sum_{i=1}^{n}\|w_T - w_T^{\neg i}\| \lesssim \sum_{t=0}^{T}\widehat{F}(w_t)/n$. Applying the train-loss bounds of Theorem 4.3.1 once more completes the proof.

4.6 Prior Works

The theoretical study of generalization properties of neural networks (NN) is more than two decades old [118, 119]. Recently, there has been an increased interest in understanding and improving generalization of SGD/GD on over-parameterized neural networks, e.g. [120, 121, 122, 16]. These results however typically require very large width where $m = \text{poly}(n)$. We discuss most-closely related-works below.

Quadratic loss. For quadratic loss, [123, 124, 125, 121, 126] showed that sufficiently over-parameterized neural networks of polynomial width satisfy a local Polyak-Łojasiewicz (PL) condition $\|\nabla\widehat{F}(w)\|^2 \geq 2\mu(\widehat{F}(w) - \widehat{F}^\star)$, where μ is at least the smallest eigenvalue of the neural tangent kernel matrix. The PL property in this case implies that the training loss converges linearly with the rate $\widehat{F}(w_t) = O((1-\eta\mu)^t)$ if the GD iterates remain in the PL region. Moreover, [127, 128], have used the PL condition to further characterize stability properties of corresponding non-convex models. Notably, [128] derived order-optimal rates $O(\frac{1}{\mu n})$ for the generalization loss. However these rates only apply to quadratic loss. Models trained with logistic or exponential loss on separable data do *not* satisfy the PL condition even for simple interpolating linear models. Aside from the PL condition-related results, but again for quadratic loss, [129] showed under specific assumptions on the data translating to low-rank NTK, that logarithmic width is sufficient to obtain classification error of order $O(n^{-1/4})$. In general, they achieve error rate $O(n^{-1/2})$, but for $m = \tilde{\Omega}(n^2)$.

Logistic-loss minimization with linear models. Logistic-loss minimization is more appropriate for classification and rate-optimal generalization bounds for GD have been obtained recently in the linear setting, where the training objective is convex. In particular, for linear logistic regression on data that are linearly separable with margin $\gamma > 0$, [114] proved a finite-time test-error bound $O(\frac{\log^2 T}{\gamma^2 T} + \frac{\log^2 T}{\gamma^2 n})$. Ignoring log factors, this is order-optimal with the sample size n and training horizon T. Their proof uses exponential-decaying properties of the logistic loss to control the norm of gradient iterates, which it cleverly combines with Markov's inequality to bound the fraction of well-separated datapoints at any iteration. This in turn translates to a test-error bound by standard margin-based generalization bounds. More recently, [112] used algorithmic-stability analysis proving same rates (up to log factors) for the test loss. Their results hold for general convex, smooth, self-bounded and decreasing objectives under a realizability assumption suited for convex objectives (analogous to Assumption 4.3.1). Specifically, this includes linear logistic regression with linearly separable data. Here, we show that analogous rates on the test loss hold true for more complicated nonconvex settings where data are separable by shallow neural networks.

Stability of GD in NN. State-of-the-art generalization bounds on shallow neural networks via the stability-analysis framework have appeared very recently in [16, 18, 116]. For Lipschitz losses, [16] shows that the empirical risk is weakly convex with a weak-convexity parameter that improves as the neural-network width m increases. Leveraging this observation, they establish stability bounds for GD iterates at time T provided sufficient parameterization $m = \tilde{\Omega}(T^2)$. Since the logistic loss is Lipschitz, these bounds also apply to our setting. Nevertheless, our work improves upon [16] in that: (i) we require significantly smaller width, poly-logarithmic rather than polynomial, and (ii) we show $\tilde{O}(1/n)$ test loss bounds in the realizable setting, while their bounds are $O(T/n)$. Central to our improvements is a largely refined analysis of the curvature of the loss via

identifying and proving a generalized quasi-convexity property for neural networks of polylogarithmic width trained with self-bounded losses (see Section 4.5 for details). Our results also improve upon the other two works [18, 116], which both require polynomial widths. However, we note that these results are not directly comparable since [18, 116] focus on quadratic-loss minimization. See also Appendix **??**.

Uniform convergence in NN. Uniform bounds on the generalization loss have been derived in literature via Rademacher complexity analysis [130]; see for example [131, 132, 133, 134, 135] for a few results in this direction. These works typically obtain the bounds of order $O(\frac{\mathcal{R}}{\sqrt{n}})$, where $\mathcal{R}$ depends on the Rademacher complexity of the hypothesis space. Recent works by [110] also utilized Rademacher complexity analysis to obtain test loss rates of $O(1/\sqrt{n})$ under an NTK separability assumption (see also [19]) with polylogarithmic width requirement for shallow and deep networks, respectively. Instead, while maintaining minimal width requirements, we obtain test-loss rates $\tilde{O}(1/n)$, which are order-optimal. Our approach, which is based on algorithmic-stability, is also different and uncovers new properties of the optimization landscape, including a generalized local quasi-convexity property. On the other hand, the analysis of [110] applies to ReLU activation and bounds the test loss with high-probability over the sampling of the training set. Instead, we require smooth activations similar to other studies such as [129, 136, 137, 19, 16, 18, 116] and we bound the test loss in expectation over the training set.

Convergence/implicit bias of GD. Convergence and implicit bias of GD for logistic/exponential loss functions on linear models and neural networks have been investigated in [113, 138, 20, 139, 140, 136]. In particular, [139, 141] have shown for homogeneous neural-networks that GD converges in direction to a max-margin solution. While certainly powerful, this implicit-bias convergence characterization becomes relevant only when the number T of GD iterations is exponentially large. Instead, our convergence bounds apply for finite T (on the order of sample size), thus are more practically relevant. Moreover,

their results assume a GD iterate t_0 such that $\widehat{F}(w_{t_0}) \leq \log(2)/n$. Similar assumption appears in [136], which require initialization $\widehat{F}(w_0) \leq 1/n^{1+C}$ for constant $C > 0$. Our approach is entirely different: we prove that sufficient parameterization benefits the loss curvature and suffices for GD steps to find an interpolating model and attain near-zero training loss, provided data satisfy an appropriate realizability condition.

4.7 Conclusions

In this chapter we study smooth shallow neural networks trained with self-bounded loss functions, such as logistic loss. Under interpolation, we provide minimal sufficient parameterization conditions to achieve rate-optimal generalization and optimization bounds. These bounds improve upon prior results which require substantially large over-parameterization or obtain sub-optimal generalization rates. Specifically, we significantly improve previous stability-based analyses in terms of both relaxing the parameterization requirements and obtaining improved rates. Although our focus was on binary classification with shallow net- works, our approach can potentially be extended to other architectures such as transformers; for preliminary results in this direction see [142]. Extending our results to the stochastic case by analyzing SGD is another important future direction. Moreover, while our current treatment relies on smoothness of the activation function to exploit properties of the curvature of the training objective, we aim to examine the potential of our results to extend to non-smooth activations. Finally, our generalization analysis bounds the expectation of the test loss (over data sampling) and it is an important future direction extending these guarantees to a high-probability setting.

4.8 Proofs

4.8.1 Training Loss Analysis

This section includes the proofs of the results stated in Section 4.3.2.

Proof of Theorem 4.3.1

We begin with proving the general train-loss and parameter-norm bounds of Theorem 4.3.1. In fact, we state and prove a slightly more general statement of the theorem which includes non-smooth and non-Lipschitz losses (such as expoential loss) that satisfy a second order self-bounded property described below.

Assumption 4.8.1 (2nd order self-boundedness)**.** *The convex loss function $f : \mathbb{R} \to \mathbb{R}_+$ satisfies the 2nd order self-boundedness property, i.e.*

$$f''(u) \leq f(u), \forall u \in \mathbb{R}.$$

Theorem 4.8.1 (General statement of Theorem 4.3.1)**.** *Let Assumptions 4.2.1-4.2.2 hold. Assume the loss function satisfies self-bounded Assumption 4.2.4. Moreover, suppose either Assumption 4.2.3 or Assumption 4.8.1 hold. Fix any $T \geq 0$. Let the step-size satisfy the assumptions of the descent lemma (Lemma 4.8.1). Assume any w and m such that $\|w - w_0\|^2 \geq \max\left\{\eta T \widehat{F}(w), \eta \widehat{F}(w_0)\right\}$ and $m \geq 18^2 L^2 R^4 \|w - w_0\|^4$. Then, the training loss and the parameters' norm satisfy*

$$\frac{1}{T}\sum_{t=1}^{T} \widehat{F}(w_t) \leq 2\widehat{F}(w) + \frac{5\|w - w_0\|^2}{2\eta T}, \tag{4.16}$$

$$\forall t \in [T] \quad : \quad \|w_t - w_0\| \leq 4\|w - w_0\|.$$

To prove Theorem 4.8.1, we first state our descent lemma for both self-bounded losses and lipschitz-smooth losses.

Lemma 4.8.1 (Descent lemma)**.** *Let Assumptions 4.2.1-4.2.2 hold. Assume the loss function satisfies self-boundedness Assumptions 4.2.4,4.8.1. Then, for any* $\eta < \frac{1}{R^2\,\widehat{F}(w_t)} \min\{\frac{1}{\ell^2+L}, \frac{1}{\sqrt{L}\ell}\}$ *the descent property holds, i.e.,*

$$\widehat{F}(w_{t+1}) \leq \widehat{F}(w_t) - \frac{\eta}{2}\|\nabla \widehat{F}(w_t)\|^2.$$

Moreover, if f *satisfies Assumption 4.2.3 then the descent property holds for any* $\eta \leq 1/L_{\widehat{F}}$ *where* $L_{\widehat{F}} := \ell^2 R^2 + \frac{LR^2}{\sqrt{m}}$ *is the smoothness parameter of the training objective.*

Proof: Due to self-boundedness Assumption 4.8.1, as well as Assumptions 4.2.1-4.2.2 the objective is also self-bounded according to Corollary 4.8.3, i.e., $\|\nabla^2 \widehat{F}(w)\| \leq \left(\ell^2 R^2 + \frac{LR^2}{\sqrt{m}}\right) \widehat{F}(w), \|\nabla \widehat{F}(w)\| \leq \ell R\, \widehat{F}(w)$.

By Taylor's expansion, there exists a $w' \in [w_t, w_{t+1}]$ such that,

$$\begin{aligned}
\widehat{F}(w_{t+1}) &= \widehat{F}(w_t) + \left\langle \nabla \widehat{F}(w_t), w_{t+1} - w_t \right\rangle + \frac{1}{2}\left\langle w_{t+1} - w_t, \nabla^2 \widehat{F}(w')\,(w_{t+1} - w_t) \right\rangle \\
&\leq \widehat{F}(w_t) + \left\langle \nabla \widehat{F}(w_t), w_{t+1} - w_t \right\rangle + \frac{1}{2} \max_{v \in [w_t, w_{t+1}]} \left\| \nabla^2 \widehat{F}(v) \right\| \cdot \|w_{t+1} - w_t\|^2 \\
&\leq \widehat{F}(w_t) - \eta \|\nabla \widehat{F}(w_t)\|^2 + \frac{\eta^2 \left(\ell^2 R^2 + \frac{LR^2}{\sqrt{m}}\right)}{2} \max_{v \in [w_t, w_{t+1}]} \widehat{F}(v) \cdot \left\| \nabla \widehat{F}(w_t) \right\|^2.
\end{aligned}$$

By Corollary 4.8.1, for $\sqrt{m} \geq \eta^2 L \ell^2 R^4 \widehat{F}^2(w_t) \geq LR^2 \|\eta \nabla \widehat{F}(w_t)\|^2 = LR^2 \|w_{t+1} - w_t\|^2$ it holds that

$$\max_{v \in [w_t, w_{t+1}]} \widehat{F}(v) \leq 2 \max\{\widehat{F}(w_t), \widehat{F}(w_{t+1})\},$$

which yields

$$\widehat{F}(w_{t+1}) \leq \widehat{F}(w_t) - \eta\|\nabla\widehat{F}(w_t)\|^2 + \eta^2\left(\ell^2R^2 + \frac{LR^2}{\sqrt{m}}\right)\max\left\{\widehat{F}(w_t), \widehat{F}(w_{t+1})\right\}\cdot\|\nabla\widehat{F}(w_t)\|^2. \tag{4.17}$$

We note that the condition on m simplifies to $m \geq 1$ if $\eta \leq \frac{1}{\sqrt{L}\ell R^2}\frac{1}{\widehat{F}(w_t)}$.

Back to (4.17), if $\widehat{F}(w_{t+1}) \geq \widehat{F}(w_t)$ by our condition $\eta < \frac{1}{\ell^2R^2+LR^2/\sqrt{m}}\frac{1}{\widehat{F}(w_t)}$ it holds that

$$\begin{aligned}\widehat{F}(w_{t+1}) &\leq \widehat{F}(w_t) + \eta\|\nabla\widehat{F}(w_t)\|^2\left(\frac{\widehat{F}(w_{t+1})}{\widehat{F}(w_t)} - 1\right)\\ &\leq \widehat{F}(w_t) + \eta\ell^2R^2\widehat{F}^2(w_t)\left(\frac{\widehat{F}(w_{t+1})}{\widehat{F}(w_t)} - 1\right).\end{aligned}$$

Since $\eta < \frac{1}{\ell^2R^2}\frac{1}{\widehat{F}(w_t)}$,

$$\begin{aligned}\widehat{F}(w_{t+1}) &< \widehat{F}(w_t) + \widehat{F}(w_t)\left(\frac{\widehat{F}(w_{t+1})}{\widehat{F}(w_t)} - 1\right)\\ &= \widehat{F}(w_{t+1}),\end{aligned}$$

which is a contradiction. Thus it holds that $\widehat{F}(w_{t+1}) < \widehat{F}(w_t)$. Continuing from Eq. (4.17) with the assumption $\eta < \frac{1}{\ell^2R^2+LR^2/\sqrt{m}}\frac{1}{\widehat{F}(w_t)}$, we conclude that

$$\begin{aligned}\widehat{F}(w_{t+1}) &\leq \widehat{F}(w_t) - \eta\|\nabla\widehat{F}(w_t)\|^2 + \frac{1}{2}\eta^2\left(\ell^2R^2 + \frac{LR^2}{\sqrt{m}}\right)\widehat{F}(w_t)\cdot\|\nabla\widehat{F}(w_t)\|^2\\ &\leq \widehat{F}(w_t) - \frac{\eta}{2}\|\nabla\widehat{F}(w_t)\|^2.\end{aligned}$$

This completes the proof for self-bounded losses.

Next, suppose f is 1-smooth and 1-Lipschitz. Then, as per Corollary 4.8.3, $\widehat{F}$ is

smooth with the constant

$$L_{\widehat{F}} := \ell^2 R^2 + \frac{LR^2}{\sqrt{m}}.$$

Following similar steps as in the beginning of proof and assuming step-size $\eta \leq 1/L_{\widehat{F}}$ we immediately conclude that,

$$\begin{aligned}\widehat{F}(w_{t+1}) &\leq \widehat{F}(w_t) - \eta\|\nabla\widehat{F}(w_t)\|^2 + \frac{\eta^2 L_{\widehat{F}}}{2}\|\nabla\widehat{F}(w_t)\|^2 \\ &\leq \widehat{F}(w_t) - \frac{\eta}{2}\|\nabla\widehat{F}(w_t)\|^2.\end{aligned}$$

This completes the proof.

As a remark, the descent property implies that the loss decreases by each step, i.e., $\widehat{F}(w_t) \leq \widehat{F}(w_0)$. Thus for self-bounded losses the condition $\eta < \frac{1}{R^2\,\widehat{F}(w_0)} \min\{\frac{1}{\ell^2+L}, \frac{1}{\sqrt{L}\ell}\}$ is sufficient. We also note that the Lipschitz-smoothness and 2nd order self-bounded assumptions are only required for the descent lemma above, which results in conditions on the step-size based on the properties of loss. In the rest of the proof we only use the self-bounded Assumption 4.2.4 in order to use the self-bounded weak convexity property of the objective (see Def. 4.3.1).

Next lemma finds a general relation for the training loss in terms of an arbitrary point $w \in \mathbb{R}^{d'}$ and the fluctuations of loss between w and GD iterates w_t.

Lemma 4.8.2. *Let Assumptions 4.2.1-4.2.2 hold. Assume the loss function satisfies the self-bounded Assumption 4.2.4. Moreover, suppose $\widehat{F}$ and step-size η are such that the following descent condition is satisfied for all $t \geq 0$:*

$$\widehat{F}(w_{t+1}) \leq \widehat{F}(w_t) - \frac{\eta}{2}\|\nabla\widehat{F}(w_t)\|^2. \tag{4.18}$$

Then, for any $w \in \mathbb{R}^{d'}$ it holds that

$$\frac{1}{T}\sum_{t=1}^{T}\widehat{F}(w_t) \le \widehat{F}(w) + \frac{\|w-w_0\|^2}{\eta T} + \frac{1}{2}\frac{LR^2}{\sqrt{m}}\frac{1}{T}\sum_{t=0}^{T-1}\max_{\alpha\in[0,1]}\widehat{F}(w_{\alpha t})\,\|w-w_t\|^2,$$

where we set $w_{\alpha t} := \alpha w_t + (1-\alpha)w$.

Proof:

Fix any w. By Taylor, there exists $w_{\alpha t}, \alpha \in [0,1]$ such that

$$\begin{aligned}
\widehat{F}(w) &= \widehat{F}(w_t) + \left\langle \nabla\widehat{F}(w_t), w-w_t\right\rangle + \frac{1}{2}\left\langle w-w_t, \nabla^2\widehat{F}(w_{\alpha t})\,(w-w_t)\right\rangle \\
&\ge \widehat{F}(w_t) + \left\langle \nabla\widehat{F}(w_t), w-w_t\right\rangle + \frac{1}{2}\lambda_{\min}\left(\nabla^2\widehat{F}(w_{\alpha t})\right)\|w-w_t\|^2 \\
&\ge \widehat{F}(w_t) + \left\langle \nabla\widehat{F}(w_t), w-w_t\right\rangle - \frac{1}{2}\frac{LR^2}{\sqrt{m}}\widehat{F}(w_{\alpha t})\,\|w-w_t\|^2.
\end{aligned}$$

The last line is true by Corollary 4.8.3. Thus, for any w,

$$\widehat{F}(w) \ge \widehat{F}(w_t) + \left\langle \nabla\widehat{F}(w_t), w-w_t\right\rangle - \frac{1}{2}\frac{LR^2}{\sqrt{m}}\max_{\alpha\in[0,1]}\widehat{F}(w_{\alpha t})\,\|w-w_t\|^2.$$

Plugging this in (4.18) gives

$$\begin{aligned}
\widehat{F}(w_{t+1}) &\le \widehat{F}(w) - \left\langle \nabla\widehat{F}(w_t), w-w_t\right\rangle - \frac{\eta}{2}\left\|\nabla\widehat{F}(w_t)\right\|^2 + \frac{1}{2}\frac{LR^2}{\sqrt{m}}\max_{\alpha\in[0,1]}\widehat{F}(w_{\alpha t})\,\|w-w_t\|^2 \\
&= \widehat{F}(w) + \frac{1}{\eta}\left(\|w-w_t\|^2 - \|w-w_{t+1}\|^2\right) + \frac{1}{2}\frac{LR^2}{\sqrt{m}}\max_{\alpha\in[0,1]}\widehat{F}(w_{\alpha t})\,\|w-w_t\|^2.
\end{aligned} \tag{4.19}$$

where the second line follows by completion of squares using $w_{t+1} - w_t = -\eta\nabla\widehat{F}(w_t)$.

Telescoping the above display for $t = 0, \dots, T-1$, we arrive at the desired.

Next, when m is large enough so that we can invoke the generalized-local quasi-convexity property, the bound of Lemma 4.8.2 takes the following convenient form

Lemma 4.8.3. *Let the assumptions of Lemma 4.8.2 hold. Assume w and m such that $\sqrt{m} \geq 2LR^2\|w - w_t\|^2$ for all $t \in [T-1]$ then*

$$\frac{1}{T}\sum_{t=1}^{T}\widehat{F}(w_t) \leq 2\widehat{F}(w) + \frac{2\|w - w_0\|^2}{\eta T} + \frac{\widehat{F}(w_0)}{2T}. \tag{4.20}$$

Proof: We invoke Corollary 4.8.1 with $\lambda = 4$ to deduce that for all $t \in [T-1]$

$$\max_{\alpha\in[0,1]}\widehat{F}(w_{\alpha t}) \leq \frac{4}{3}\max\{\widehat{F}(w), \widehat{F}(w_t)\} < \frac{4}{3}\widehat{F}(w_t) + \frac{4}{3}\widehat{F}(w). \tag{4.21}$$

Noting the assumption on m and recalling Lemma 4.8.2,

$$\begin{aligned}
\frac{1}{T}\sum_{t=1}^{T}\widehat{F}(w_t) &\leq \widehat{F}(w) + \frac{\|w - w_0\|^2}{\eta T} + \frac{1}{2}\frac{LR^2}{\sqrt{m}}\frac{1}{T}\sum_{t=0}^{T-1}\max_{\alpha\in[0,1]}\widehat{F}(w_{\alpha t})\,\|w - w_t\|^2 \\
&\leq \frac{4}{3}\widehat{F}(w) + \frac{\|w - w_0\|^2}{\eta T} + \frac{1}{3T}\sum_{t=0}^{T-1}\widehat{F}(w_t) \\
&\leq \frac{4}{3}\widehat{F}(w) + \frac{\|w - w_0\|^2}{\eta T} + \frac{1}{3T}\sum_{t=0}^{T}\widehat{F}(w_t).
\end{aligned}$$

Arranging terms yields the desired result.

Finally, using the about bounds on the training loss, we can bound the parameter-norm using a recursive argument presented in the lemma below.

Lemma 4.8.4 (Iterates-norm bound)**.** *Suppose the assumptions of Lemma 4.8.2 hold. Fix any $T \geq 0$ and assume any w and m such that*

$$\|w - w_0\|^2 \geq \max\{\eta T\widehat{F}(w), \eta\widehat{F}(w_0)\}. \tag{4.22}$$

and

$$\sqrt{m} \geq 18LR^2\|w - w_0\|^2, \tag{4.23}$$

Then, for all $t \in [T]$,

$$\|w_t - w\| \leq 3\|w - w_0\|. \tag{4.24}$$

Proof: Denote $A_t = \|w_t - w\|$. Start by recalling from (4.19) that for all t:

$$A_{t+1}^2 \leq A_t^2 + \eta\widehat{F}(w) - \eta\widehat{F}(w_{t+1}) + \eta\,\frac{LR^2}{2\sqrt{m}} \max_{\alpha\in[0,1]} \widehat{F}(w_{\alpha t})\, A_t^2. \tag{4.25}$$

We will prove the desired statement (4.24) using induction. For $t = 0$, $A_0 = \|w - w_0\|$. Thus, the assumption of induction holds. Now assume (4.24) is correct for $t \in [T-1]$, i.e. $A_t \leq 3\|w - w_0\|, \forall t \in [T-1]$. We will then prove it holds for $t = T$.

The first observation is that by induction hypothesis $\sqrt{m} \geq 18LR^2\|w - w_0\|^2 \geq 2LR^2A_t^2$ for all $t \in [T-1]$. Thus, for all $t \in [T-1]$, the condition of the generalized local quasi-convexity Corollary 4.5.1 holds for $\lambda = 4$ implying (see also (4.21))

$$\forall t \in [T-1] \;:\; \max_{\alpha\in[0,1]} \widehat{F}(w_{\alpha t}) \leq \frac{4}{3}\widehat{F}(w_t) + \frac{4}{3}\widehat{F}(w).$$

Using this in (4.25) we find for all $t \in [T-1]$ that

$$\begin{aligned} A_{t+1}^2 &\leq A_t^2 + \eta\widehat{F}(w) - \eta\widehat{F}(w_{t+1}) + \eta\,\frac{LR^2 \cdot A_t^2}{2\sqrt{m}} \left(\frac{4}{3}\widehat{F}(w_t) + \frac{4}{3}\widehat{F}(w)\right) \\ &\leq A_t^2 + \eta\widehat{F}(w) - \eta\widehat{F}(w_{t+1}) + \eta\left(\frac{1}{3}\widehat{F}(w_t) + \frac{1}{3}\widehat{F}(w)\right) \end{aligned}$$

where in the second inequality we used again that $\sqrt{m} \geq 2LR^2A_t^2$. We proceed by

telescoping the above display over $t = 0, 1, \ldots, T-1$ to get

$$\begin{aligned} A_T^2 &\leq A_0^2 + \frac{4}{3}\eta T \widehat{F}(w) + \frac{1}{3}\eta \widehat{F}(w_0) + \frac{1}{3}\eta \sum_{t=0}^{T-1} \widehat{F}(w_t) - \eta \widehat{F}(w_T) \\ &\leq A_0^2 + \frac{4}{3}\eta T \widehat{F}(w) + \frac{2}{3}\eta \widehat{F}(w_0) + \frac{1}{3}\eta \sum_{t=1}^{T} \widehat{F}(w_t), \end{aligned}$$

where the second line follows by nonegativity of the loss.

Now, to bound the last term above, observe that the condition of Lemma 4.8.3 holds since $\sqrt{m} \geq 2LR^2A_t^2$ for all $t \in [T-1]$ by induction hypothesis. Hence, using (4.20), we conclude that

$$\begin{aligned} A_T^2 &\leq A_0^2 + \frac{4}{3}\eta T \widehat{F}(w) + \frac{2}{3}\eta \widehat{F}(w_0) + \frac{1}{3}\eta T \left(2\widehat{F}(w) + \frac{2A_0^2}{\eta T} + \frac{\widehat{F}(w_0)}{2T} \right) \\ &= \frac{5}{3}A_0^2 + 2\eta T \widehat{F}(w) + \frac{5}{6}\eta \widehat{F}(w_0) \\ &\leq \frac{5}{3}\|w - w_0\|^2 + 2\|w - w_0\|^2 + \frac{5}{6}\|w - w_0\|^2 = \frac{9}{2}\|w - w_0\|^2 \qquad \implies \quad A_T \leq 3\|w - w_0\|. \end{aligned} \tag{4.26}$$

In the last inequality, we used the assumptions of the lemma on $\|w - w_0\|$ and $A_0 = \|w - w_0\|$. This completes the proof.

Completing the proof of Theorem 4.8.1.

The proof follows from combining the bounds on the training loss and parameters' growth from Lemmas 4.8.3-4.8.4 and noting that with condition on $\|w - w_0\|^2$ from Lemma 4.8.4 we have $\widehat{F}(w_0) \leq \|w - w_0\|^2/\eta$ to derive (4.16). Moreover, we have $\|w_t - w_0\| \leq \|w_t - w\| + \|w - w_0\| \leq 4\|w - w_0\|$.

Proof of Theorem 4.3.2

Here we prove training loss bound for interpolating NN as asserted by Theorem 4.3.2. Similar to the previous section, we prove a more general result where the loss is not necessarily Lipschitz or smooth. We are now ready to prove Theorem 4.3.2 for general self-bounded losses. In particular, Theorem 4.3.2 follows directly from the next result by choosing f to be Lipschitz and smooth.

Theorem 4.8.2 (General statement of Theorem 4.3.2)**.** *Suppose Assumptions 4.2.1-4.2.2, 4.2.4 hold. Moreover, assume the objective and data satisfy the Assumption 4.3.1. Let the step-size satisfy the assumptions of Descent Lemma 4.8.1. Moreover, assume* $\eta \leq \min\{g(1)^2, \frac{1}{L_{\widehat{F}}}, \frac{g(1)^2}{\widehat{F}(w_0)}\}$ *and* $m \geq 18^2 L^2 R^4\, g(\frac{1}{T})^4$ *for a fixed training horizon* T*. Then,*

$$\widehat{F}(w_T) \;\leq \frac{2}{T} + \frac{5\, g(\frac{1}{T})^2}{2\eta T},$$
$$\forall t \in [T] \;:\; \|w_t - w_0\| \;\leq\; 4\, g(\frac{1}{T}).$$

Proof: According to Assumption 4.3.1, for any sufficiently small $\varepsilon > 0$, there exists a $w^{(\varepsilon)}$ such that $\widehat{F}(w^{(\varepsilon)}) \leq \varepsilon$ and $\|w^{(\varepsilon)} - w_0\| = g(\varepsilon)$. Pick $\varepsilon = 1/T$. With the condition $\eta \leq \min\{g(1)^2, g(1)^2/\widehat{F}(w_0)\}$ we have

$$\max\left\{\eta T \widehat{F}(w^{(1/T)}), \eta \widehat{F}(w_0)\right\} \leq g(1)^2 \leq g(\frac{1}{T})^2 = \|w^{(1/T)} - w_0\|^2,$$

where in the second inequality we used the fact that g is a decreasing function. The desired result is obtained by Theorem 4.8.1.

Generalized local quasi-convexity property

In the remainder of this section, we prove the generalized local quasi-convexity property.

Proposition 4.8.1 (Restatement of Proposition 4.5.1)**.** *Suppose* $\widehat{F} : \mathbb{R}^{d'} \to \mathbb{R}$ *satisfies the self-bounded weak convexity property in Eq. 4.4 with parameter* κ*. Let* $w_1, w_2 \in \mathbb{R}^{d'}$ *be two arbitrary points with distance* $\|w_1 - w_2\| \le D < \sqrt{2/\kappa}$ *. Set* $\tau := (1 - \kappa D^2/2)^{-1}$*. Then,*

$$\max_{v \in [w_1, w_2]} \widehat{F}(v) \le \tau \cdot \max\{\widehat{F}(w_1), \widehat{F}(w_2)\}. \tag{4.27}$$

Proof: Assume the claim of the proposition is incorrect, then

$$\max_{v \in [w_1, w_2]} \widehat{F}(v) > \tau \cdot \max\{\widehat{F}(w_1), \widehat{F}(w_2)\} > \max\{\widehat{F}(w_1), \widehat{F}(w_2)\}. \tag{4.28}$$

Define $w_\star := \arg\max_{v \in [w_1, w_2]} \widehat{F}(v)$. Note that $w_\star$ is an interior point. Thus by the optimality condition it holds

$$\left\langle \nabla \widehat{F}(w_\star), w_1 - w_2 \right\rangle = 0. \tag{4.29}$$

By Taylor's approximation theorem for two points $w_1, w \in \mathbb{R}^{d'}$, there exists a $w_\beta \in [w, w_1]$, such that

$$\widehat{F}(w_1) = \widehat{F}(w) + \left\langle \nabla \widehat{F}(w), w_1 - w \right\rangle + \frac{1}{2} \left\langle w - w_1, \nabla^2 \widehat{F}(w_\beta)\,(w - w_1) \right\rangle \tag{4.30}$$

Pick $w = w_\star = \alpha_\star w_1 + (1 - \alpha_\star) w_2$ in Eq. (4.30), and note that

$$\left\langle \nabla \widehat{F}(w_\star), w_1 - w_\star \right\rangle = -(1 - \alpha_\star) \left\langle \nabla \widehat{F}(w_\star), w_1 - w_2 \right\rangle = 0.$$

Therefore,

$$\begin{aligned}\widehat{F}(w_1) &= \widehat{F}(w_\star) + \frac{1}{2}\left\langle w_\star - w_1, \nabla^2 \widehat{F}(w_\beta)\,(w_\star - w_1)\right\rangle \\ &\geq \widehat{F}(w_\star) + \frac{1}{2}\lambda_{\min}(\nabla^2 \widehat{F}(w_\beta))\left\|w_\star - w_1\right\|^2 \\ &\geq \widehat{F}(w_\star) - \frac{1}{2}\kappa\,\widehat{F}(w_\beta)\left\|w_\star - w_1\right\|^2.\end{aligned}$$

where in the last line we used the self-bounded weak convexity property i.e., $\lambda_{\min}\left(\nabla^2 \widehat{F}(w_\beta)\right) \geq -\kappa \widehat{F}(w_\beta)$.

This leads to

$$\begin{aligned}\widehat{F}(w_1) &\geq \widehat{F}(w_\star) - \frac{(1-\alpha_\star)^2}{2}\kappa\,\widehat{F}(w_\beta)\left\|w_1 - w_2\right\|^2 \\ &> \widehat{F}(w_\star) - \frac{1}{2}\kappa\,\widehat{F}(w_\beta)\left\|w_1 - w_2\right\|^2.\end{aligned}$$

Note that $w_\beta \in [w_\star, w_1] \subset [w_1, w_2]$, thus $\widehat{F}(w_\beta) \leq \widehat{F}(w_\star)$ by definition of $w_\star$. Therefore,

$$\begin{aligned}\widehat{F}(w_\star) &< \frac{1}{1 - \frac{1}{2}\kappa\,\|w_1 - w_2\|^2}\widehat{F}(w_1) \\ &\leq \frac{1}{1 - \frac{1}{2}\kappa D^2}\widehat{F}(w_1),\end{aligned}$$

which is in contradiction with (4.28). This proves the statement of the proposition.

Specializing this property to two-layer neural networks yields the following.

Corollary 4.8.1 (Restatement of Corollary 4.5.1)**.** *Let Assumptions 4.2.1,4.2.2, 4.2.4 hold. Fix arbitrary $w_1, w_2 \in \mathbb{R}^{d'}$, any constant $\lambda > 1$, and m large enough such that $\sqrt{m} \geq \lambda \frac{LR^2}{2}\|w_1 - w_2\|^2$. Then,*

$$\max_{v \in [w_1, w_2]} \widehat{F}(v) \leq (1 - 1/\lambda)^{-1} \cdot \max\{\widehat{F}(w_1), \widehat{F}(w_2)\}. \tag{4.31}$$

Proof: By our assumptions and Corollary 4.8.3 the objective's Hessian satisfies

$$\lambda_{\min}\left(\nabla^2 \widehat{F}(w)\right) \geq -\frac{LR^2}{\sqrt{m}}\widehat{F}(w).$$

Invoking Proposition 4.8.1 with $\kappa := \frac{LR^2}{\sqrt{m}}$ concludes the claim.

4.8.2 Generalization Error Analysis

This section includes the proofs of the generalization results stated in Section 4.3.3.

Proof of Theorem 4.3.3

We prove the generalization gap of Theorem 4.3.3 for Lipshitz-smooth losses. The proof follows the steps of our proof sketch in Sec. 4.5.2.

First, the proofs of exansiveness of GD in NN (Lemma 4.5.1) and the corresponding model stability bound are given next.

Lemma 4.8.5 (GD-Expansivieness)**.** *Let Assumptions 4.2.1-4.2.2 hold. For any w, w' and $w_\alpha = \alpha w + (1-\alpha)w'$ it holds that*

$$\left\|\left(w - \eta\nabla\widehat{F}(w)\right) - \left(w' - \eta\nabla\widehat{F}(w')\right)\right\| \leq \max_{\alpha\in[0,1]} H(w_\alpha)\,\|w - w'\|\,,$$
$$H(w) := \eta\frac{LR^2}{\sqrt{m}}\widehat{F}'(w) + \max\left\{1, \eta\ell^2R^2\widehat{F}''(w)\right\},$$

where we define $\widehat{F}'(w) := \frac{1}{n}\sum_{i=1}^n |f'(y_i\Phi(w,x_1))|$ and $\widehat{F}''(w) := \frac{1}{n}\sum_{i=1}^n f''(y_i\Phi(w,x_1))$.

Proof: Fix $u : \|u\| = 1$ and define $g_u : \mathbb{R}^{d'} \to \mathbb{R}$:

$$g_u(w) := \langle u, w\rangle - \eta\langle u, \nabla\widehat{F}(w)\rangle.$$

Note

$$\left\|w - \nabla\widehat{F}(w) - (w' - \nabla\widehat{F}(w'))\right\| = \max_{\|u\|=1} |g_u(w) - g_u(w')|\,.$$

For any w, w', we have

$$\begin{aligned} g_u(w) - g_u(w') &= \int_0^1 u^\top \left(I - \eta\nabla^2\widehat{F}(w' + \alpha(w - w'))\right)(w - w')\mathrm{d}\alpha \\ &\leq \max_{\alpha\in[0,1]} \left\|\left(I - \eta\nabla^2\widehat{F}(w' + \alpha(w - w'))\right)\right\| \left\|w - w'\right\|. \end{aligned} \tag{4.32}$$

For convenience denote $w_\alpha := \alpha w + (1-\alpha)w'$ and $A_\alpha := \nabla^2\widehat{F}(w_\alpha)$. Then, for any $\alpha \in [0,1]$ we have that

$$\left\|I - \eta\nabla^2\widehat{F}(w_\alpha)\right\| = \max\left\{|1 - \eta\lambda_{\min}(A_\alpha)|, |1 - \eta\lambda_{\max}(A_\alpha)|\right\}. \tag{4.33}$$

For convenience, let $\beta := \frac{1}{\sqrt{m}}LR^2\widehat{F}'(w_\alpha) \geq 0$ and note from Lemma 4.8.11 that $\lambda_{\min}(A_\alpha) \geq -\beta$. Using this, we will show that

$$|1 - \eta\lambda_{\min}(A_\alpha)| \leq \max\{1 + \eta\beta, \eta\lambda_{\max}(A_\alpha)\}. \tag{4.34}$$

To show this consider two cases. First, if $\eta\lambda_{\min}(A_\alpha) \in [-\eta\beta, 1]$, then

$$|1 - \eta\lambda_{\min}(A_\alpha)| = 1 - \eta\lambda_{\min}(A_\alpha) \leq 1 + \eta\beta.$$

On the other hand, if $\eta\lambda_{\min}(A_\alpha) \geq 1$, then

$$|1 - \eta\lambda_{\min}(A_\alpha)| = \eta\lambda_{\min}(A_\alpha) - 1 \leq \eta\lambda_{\min}(A_\alpha) \leq \eta\lambda_{\max}(A_\alpha),$$

which shows (4.34).

Next, we will show that

$$|1 - \eta\lambda_{\max}(A_\alpha)| \leq \max\{1 + \eta\beta, \eta\lambda_{\max}(A_\alpha)\}. \tag{4.35}$$

We consider again three cases. First, if $\eta\lambda_{\max}(A_\alpha) \in [0, 1]$, then

$$|1 - \eta\lambda_{\max}(A_\alpha)| = 1 - \eta\lambda_{\max}(A_\alpha) \leq 1.$$

Second, if $\eta\lambda_{\max}(A_\alpha) \geq 1$

$$|1 - \eta\lambda_{\max}(A_\alpha)| = \eta\lambda_{\max}(A_\alpha) - 1 \leq \eta\lambda_{\max}(A_\alpha).$$

Otherwise, it must be that $-\beta \leq \lambda_{\min}(A_\alpha) \leq \lambda_{\max}(A_\alpha) \leq 0$. Thus,

$$|1 - \eta\lambda_{\max}(A_\alpha)| = 1 - \eta\lambda_{\max}(A_\alpha) \leq 1 - \eta\lambda_{\min}(A_\alpha) \leq 1 + \eta\beta.$$

To complete the proof of the lemma combine (4.33) with (4.34) and (4.35):

$$\|I - \eta\nabla^2\widehat{F}(w_\alpha)\| \leq \max\{1 + \eta\beta, \eta\lambda_{\max}(A_\alpha)\},$$

and further use from Lemma 4.8.11 that $\eta\lambda_{\max}(A_\alpha) \leq \eta\ell^2R^2\widehat{F}''(w) + \eta\beta$.

For the stability analysis below, recall the definition of the leave-one-out (loo) training loss for $i \in [n]$: $\widehat{F}^{\neg i}(w) := \frac{1}{n}\sum_{j\neq i}\widehat{F}_j(w)$. With these, define the loo model updates of GD on the loo loss:

$$w_{t+1}^{\neg i} := w_t^{\neg i} - \eta\nabla\widehat{F}^{\neg i}(w_t^{\neg i}),\ t \geq 0, \qquad w_0^{\neg i} = w_0.$$

Theorem 4.8.3 (Model stability bound)**.** *Suppose Assumptions 4.2.1, 4.2.2, 4.2.3, 4.2.4*

hold. Fix any time horizon $T \geq 1$ and any step size $\eta > 0$. Set the regret and the leave-one-out regrets of GD updates as follows:

$$\texttt{Reg} := \frac{1}{T}\sum_{t=1}^{T} \widehat{F}(w_t) \qquad \textit{and} \qquad \texttt{Reg}_{loo} := \frac{1}{T}\max_{i\in[n]} \sum_{t=1}^{T} \widehat{F}^{\neg i}(w_t^{\neg i}).$$

Suppose that the width m is large enough so that it satisfies the following two conditions:

$$\sqrt{m} \geq 4LR^2 \max\left\{\|w_t - w_0\|^2, \|w_t^{\neg i} - w_0\|^2\right\}, \quad \forall i \in [n], t \in [T], \tag{4.36}$$

and

$$\sqrt{m} \geq 6LR^2\eta T \max\left\{\texttt{Reg}, \texttt{Reg}_{loo}\right\}. \tag{4.37}$$

Then, the leave-one-out model stability is bounded as follows:

$$\frac{1}{n}\sum_{i=1}^{n} \left\|w_T - w_T^{\neg i}\right\| \leq \frac{2\eta\ell R}{n}\left(\widehat{F}(w_0) + T\cdot\texttt{Reg}\right).$$

Proof: Using self-boundedness Assumption 4.2.4 together with Corollary 4.5.2 it holds for all $i \in [n]$:

$$\begin{aligned}
\left\|w_{t+1} - w_{t+1}^{\neg i}\right\| &\leq \left\| \left(w_t - \eta\nabla\widehat{F}^{\neg i}(w_t)\right) - \left(w_t^{\neg i} - \eta\nabla\widehat{F}^{\neg i}(w_t^{\neg i})\right) \right\| + \frac{\eta}{n}\left\|\nabla\widehat{F}_i(w_t)\right\| \\
&\leq \left\| \left(w_t - \eta\nabla\widehat{F}^{\neg i}(w_t)\right) - \left(w_t^{\neg i} - \eta\nabla\widehat{F}^{\neg i}(w_t^{\neg i})\right) \right\| + \frac{\eta\ell R}{n}\widehat{F}_i(w_t) \\
&\leq \left(1 + \eta\frac{LR^2}{\sqrt{m}}\max_{\alpha\in[0,1]}\widehat{F}^{\neg i}(w_{\alpha t}^{\neg i})\right)\left\|w_t - w_t^{\neg i}\right\| + \frac{\eta\ell R}{n}\widehat{F}_i(w_t),
\end{aligned} \tag{4.38}$$

where we denote for convenience $w_{\alpha t}^{\neg i} = \alpha w_t + (1-\alpha)w_t^{\neg i}$.

Moreover, by the theorem's condition in Eq. (4.36), it holds for all $t \in [T]$ and all

$i \in [n]$ that

$$\sqrt{m} \geq 2LR^2(\|w_t - w_0\|^2 + \|w_t^{\neg i} - w_0\|^2) \geq LR^2 \left\|w_t - w_t^{\neg i}\right\|^2 .$$

Thus, we can apply Corollary 4.5.1 for $\lambda = 2$, which gives the following generalized-local quasi-convexity property for the loo objective:

$$\max_{\alpha \in [0,1]} \widehat{F}^{\neg i}(w_{\alpha t}^{\neg i}) \leq 2 \max \left\{ \widehat{F}^{\neg i}(w_t), \widehat{F}^{\neg i}(w_t^{\neg i}) \right\} .$$

In turn applying this back in (4.38) we have shown that

$$\left\|w_{t+1} - w_{t+1}^{\neg i}\right\| \leq \left(1 + \eta \frac{2LR^2}{\sqrt{m}} \max \left\{ \widehat{F}^{\neg i}(w_t), \widehat{F}^{\neg i}(w_t^{\neg i}) \right\} \right) \left\|w_t - w_t^{\neg i}\right\| + \frac{\eta \ell R}{n} \widehat{F}_i(w_t) \tag{4.39}$$

To continue, denote for convenience

$$\beta_t^i := \eta \frac{2LR^2}{\sqrt{m}} \max \left\{ \widehat{F}^{\neg i}(w_t), \widehat{F}^{\neg i}(w_t^{\neg i}) \right\} \quad \text{and} \quad \rho := \eta \ell R,$$

so that:

$$\left\|w_{t+1} - w_{t+1}^{\neg i}\right\| \leq \left(1 + \beta_t^i\right) \left\|w_t - w_t^{\neg i}\right\| + \frac{\rho}{n} \widehat{F}_i(w_t), \qquad \forall i \in [n], t \in [T] .$$

By unrolling the iterations over $t \in [T]$ and noting $w_0 = w_0^{\neg i}$, we obtain the following for

the leave-one-out parameter distance at iteration T:

$$\begin{aligned}
\left\|w_T - w_T^{\neg i}\right\| &\leq \frac{\rho}{n}\sum_{t=0}^{T-1}\left(\prod_{\tau=t+1}^{T-1}(1+\beta_\tau^i)\right)\widehat{F}_i(w_t) \\
&\leq \frac{\rho}{n}\sum_{t=0}^{T-1}\exp\left(\sum_{\tau=t+1}^{T-1}\beta_\tau^i\right)\widehat{F}_i(w_t) \\
&\leq \frac{\rho}{n}\sum_{t=0}^{T-1}\exp\left(\sum_{\tau=1}^{T-1}\beta_\tau^i\right)\widehat{F}_i(w_t) = \exp\left(\sum_{\tau=1}^{T-1}\beta_\tau^i\right)\frac{\rho}{n}\sum_{t=0}^{T-1}\widehat{F}_i(w_t) \\
&\leq \frac{\rho}{n}\exp\left(\max_{j\in[n]}\sum_{\tau=1}^{T-1}\beta_\tau^j\right)\sum_{t=0}^{T-1}\widehat{F}_i(w_t), \qquad \forall i\in[n]\,. \qquad (4.40)
\end{aligned}$$

It remains to bound $\beta := \max_{i\in[n]}\sum_{\tau=1}^{T-1}\beta_\tau^i$. We do this as follows:

$$\begin{aligned}
\beta &= \frac{2\eta LR^2}{\sqrt{m}}\max_{i\in[n]}\left\{\max\left\{\sum_{t=1}^{T}\widehat{F}^{\neg i}(w_t)\,,\,\sum_{t=1}^{T}\widehat{F}^{\neg i}(w_t^{\neg i})\right\}\right\} \\
&\leq \frac{2\eta LR^2}{\sqrt{m}}\max_{i\in[n]}\left\{\max\left\{\sum_{t=1}^{T}\widehat{F}(w_t)\,,\,\sum_{t=1}^{T}\widehat{F}^{\neg i}(w_t^{\neg i})\right\}\right\} \\
&= \frac{2\eta LR^2}{\sqrt{m}}\max\left\{\sum_{t=1}^{T}\widehat{F}(w_t)\,,\,\max_{i\in[n]}\sum_{t=1}^{T}\widehat{F}^{\neg i}(w_t^{\neg i})\right\} \\
&= \frac{2\eta LR^2}{\sqrt{m}}T\,\max\left\{\mathtt{Reg}\,,\,\mathtt{Reg}_{\mathtt{loo}}\right\} \leq 2/3\,,
\end{aligned}$$

where: (i) in the first inequality we used nonnegativity of $f(\cdot)$ to conclude for any $i\in[n]$ and any w that $\widehat{F}^{\neg i}(w)\leq\widehat{F}(w)$; (ii) in the last line, we recalled the definition of the regret terms and we used the theorem's condition (4.43) on large enough m.

Using this in (4.40) and averaging over $i\in[n]$ yields

$$\begin{aligned}
\frac{1}{n}\sum_{i\in[n]}\left\|w_T - w_T^{\neg i}\right\| &\leq \frac{\rho e^{\beta}}{n}\sum_{t=0}^{T-1}\frac{1}{n}\sum_{i=1}^{n}\widehat{F}_i(w_t) \\
&\leq \frac{\eta\ell Re^{2/3}}{n}\sum_{t=0}^{T-1}\widehat{F}(w_t)\,.
\end{aligned}$$

The advertised bound follows by using $e^{2/3} \leq 2$ and writing

$$\frac{1}{T}\sum_{t=0}^{T-1}\widehat{F}(w_t) \leq \frac{1}{T}\sum_{t=0}^{T}\widehat{F}(w_t) = \frac{\widehat{F}(w_0)}{T} + \texttt{Reg}.$$

To bound the generalization gap in terms of model stability we rely on the following result.

Lemma 4.8.6 ([14]). *Suppose the sample loss $f(\cdot, z)$ is $G_{\widehat{F}}$-Lipschitz for almost surely all data points $z \sim \mathcal{D}$. Then, the following relation holds between expected generalization loss and model stability at any iterate T,*

$$\mathbb{E}\Big[F(w_T)\Big] - \mathbb{E}\Big[\widehat{F}(w_T)\Big] \leq 2G_{\widehat{F}}\,\mathbb{E}\Big[\frac{1}{n}\sum_{i=1}^{n}\|w_T - w_T^{\neg i}\|\Big]. \tag{4.41}$$

With the two results above, we are ready to prove Theorem 4.3.3.

Theorem 4.8.4 (Restatement of Theorem 4.3.3). *Suppose Assumptions 4.2.1- 4.2.4 hold. Fix any time horizon $T \geq 1$ and any step size $\eta \leq 1/L_{\widehat{F}}$ where $L_{\widehat{F}}$ is the objective's smoothness parameter. Let any $w \in \mathbb{R}^{d'}$ such that $\|w - w_0\|^2 \geq \max\{\eta T\,\widehat{F}(w), \eta\widehat{F}(w_0)\}$. Suppose hidden-layer width m satisfies $m \geq 64^2L^2R^4\|w - w_0\|^4$. Then, the generalization gap of GD at iteration T is bounded as*

$$\mathbb{E}\Big[F(w_T) - \widehat{F}(w_T)\Big] \leq \frac{8\ell^2R^2}{n}\,\mathbb{E}\left[\eta T\,\widehat{F}(w) + 2\|w - w_0\|^2\right],$$

where all expectations are over the training set.

Proof: The proof essentially follows by combining Theorem 4.8.3 with Theorem 4.3.1. Note that the assumptions of Theorem 4.3.1 are met. Thus, the regret and

parameter-norm are bounded as follows:

$$\texttt{Reg} \leq 2\widehat{F}(w) + \frac{5\|w - w_0\|^2}{2\eta T} \qquad \text{and} \qquad \max_{t\in[T]} \|w_t - w_0\| \leq 4\|w - w_0\| \,. \tag{4.42}$$

We can also use Theorem 4.3.1 to the leave-one-out objective $\widehat{F}^{\neg i}$ and the corresponding loo GD updates $w_t^{\neg i}$. This bounds the loo regret and the norm of the loo parameter, as follows:

$$\texttt{Reg}_{\texttt{loo}} \leq 2\widehat{F}(w) + \frac{5\|w - w_0\|^2}{2\eta T} \qquad \text{and} \qquad \max_{i\in[n]} \max_{t\in[T]} \|w_t^{\neg i} - w_0\| \leq 4\|w - w_0\| \,.$$

We use these two displays to show that m is by assumption large enough so that Eqs. (4.36) and (4.43) hold. Indeed, we have

$$\sqrt{m} \geq 64LR^2\|w - w_0\|^2 = 4LR^2 \left(4\|w - w_0\|\right)^2 \geq 4LR^2 \max\left\{\|w_t - w_0\|^2, \|w_t^{\neg i} - w_0\|^2\right\}$$

and

$$\begin{aligned}
\sqrt{m} \geq 64LR^2\|w - w_0\|^2 &> 6LR^2 \cdot 5\|w - w_0\|^2 \\
&> 6LR^2 \cdot (2\eta T\widehat{F}(w) + 5\|w - w_0\|^2/2) \\
&\geq 6LR^2\eta T \max\left\{\texttt{Reg}, \texttt{Reg}_{\texttt{loo}}\right\} \,.
\end{aligned}$$

In the second display we also used the theorem's assumption that $\|w - w_0\|^2 \geq \eta T\widehat{F}(w)$.

Thus, we can apply Theorem 4.8.3 to find that

$$\begin{aligned}\frac{1}{n}\sum_{i=1}^{n}\left\|w_T - w_T^{\neg i}\right\| &\leq \frac{2\ell R}{n}\left(\eta\widehat{F}(w_0) + \eta T\cdot \texttt{Reg}\right)\\ &\leq \frac{2\ell R}{n}\left(\eta\widehat{F}(w_0) + 2\eta T\widehat{F}(w) + 5\|w-w_0\|^2/2\right)\\ &\leq \frac{2\ell R}{n}\left(2\eta T\widehat{F}(w) + 7\|w-w_0\|^2/2\right)\end{aligned}$$

where in the penultimate line we used (4.42) and in the last line we used the theorem's assumption that $\|w-w_0\|^2 \geq \eta\widehat{F}(w_0)$.

To conclude the proof, simply take expectations over the train set on the above display and apply Lemma 4.8.6 recalling $G_{\widehat{F}} = \ell R$.

Proof of Theorem 4.3.4

Here we prove the generalization gap for interpolating neural networks as per Theorem 4.3.4.

Theorem 4.8.5 (Restatement of Theorem 4.3.4)**.** *Let Assumptions 4.2.1-4.3.1 hold. Fix $T \geq 1$ and let $m \geq 64^2 L^2 R^4\, g(\frac{1}{T})^4$. Then, for any $\eta \leq \min\{\frac{1}{L_{\widehat{F}}}, g(1)^2, \frac{g(1)^2}{\widehat{F}(w_0)}\}$ the expected generalization gap at iteration T satisfies*

$$\mathbb{E}\Big[F(w_T) - \widehat{F}(w_T)\Big] \leq \frac{24\ell^2 R^2\, g(\frac{1}{T})^2}{n}\,. \tag{4.43}$$

Proof: According to Assumption 4.3.1, for any sufficiently small $\varepsilon > 0$, there exists $w^{(\varepsilon)}$ such that $\widehat{F}(w^{(\varepsilon)}) \leq \varepsilon$ and $\|w^{(\varepsilon)} - w_0\| = g(\varepsilon)$. Recall from Theorem 4.3.3 that,

$$\mathbb{E}\left[F(w_T) - \widehat{F}(w_T)\right] \leq \frac{8\ell^2 R^2}{n}\left(\eta T\widehat{F}(w) + 2\|w-w_0\|^2\right)\,. \tag{4.44}$$

In particular let $\varepsilon = 1/T$ and replace w with $w^{(\varepsilon)}$. This is possible since after $T \geq 1$

steps and with the decreasing nature of g and the condition on step-size it holds that $\|w^{(1/T)} - w_0\|^2 = g(1/T)^2 \geq g(1)^2 \geq \max\{\eta T \widehat{F}(w^{(1/T)}), \eta \widehat{F}(w_0)\}$. Thus continuing from (4.44) we have,

$$\mathbb{E}\left[F(w_T) - \widehat{F}(w_T)\right] \leq \frac{8\ell^2 R^2}{n}\left(\eta + 2g(\frac{1}{T})^2\right).$$

Recalling $\eta \leq g(1)^2 \leq g(\frac{1}{T})^2$ leads to the claim of the theorem.

4.8.3 Proofs for Section 4.4

We first prove proposition 4.4.1, which we repeat here for convenience.

Proposition 4.8.2 (Restatement of Proposition 4.4.1)**.** *Let Assumptions 4.2.1-4.2.2,4.4.1-4.4.2 hold. Assume $f(\cdot)$ to be the logistic loss. Fix $\varepsilon > 0$ and let $m \geq \frac{L^2R^4}{4\gamma^4 C^2}(2C + \log(1/\varepsilon))^4$. Then the realizability Assumption 4.3.1 holds with $g(\varepsilon) = \frac{1}{\gamma}(2C + \log(1/\varepsilon))$. In other words, there exists $w^{(\varepsilon)}$ such that*

$$\widehat{F}(w^{(\varepsilon)}) \leq \varepsilon, \quad \textit{and} \quad \left\|w^{(\varepsilon)} - w_0\right\| = \frac{1}{\gamma}\left(2C + \log(1/\varepsilon)\right). \tag{4.45}$$

Proof: By Taylor there exists $w' \in [w, w_0]$ such that,

$$y_i\Phi(w, x_i) = y_i\Phi(w_0, x_i) + y_i\Big\langle \nabla_1\Phi(w_0, x_i), w - w_0\Big\rangle + \frac{1}{2}y_i\Big\langle w - w_0, \nabla_1^2\Phi(w', x_i)(w - w_0)\Big\rangle \tag{4.46}$$

Pick $w = w^{(\varepsilon)} := w_0 + \frac{w^\star}{\gamma}(2C + \log(1/\varepsilon))$ for $w^\star$ defined in Assumption 4.4.1. Since $\|w^\star\| = 1$, we automatically derive the desired for $\|w^{(\varepsilon)} - w_0\|$. Next, we show that $\widehat{F}_i(w^{(\varepsilon)}) \leq \varepsilon$. Based on Lemma 4.8.10, $\|\nabla_1^2\Phi(w', x_i)\| \leq \frac{LR^2}{\sqrt{m}}$. Continuing from Eq. (4.46),

we deduce the following,

$$\begin{aligned} y_i\Phi(w,x_i) &\geq -\left|y_i\Phi(w_0,x_i)\right| + y_i\left\langle \nabla_1\Phi(w_0,x_i), w^{(\varepsilon)}-w_0\right\rangle - \frac{1}{2}\left\|\nabla_1^2\Phi(w',x_i)\right\|\left\|w^{(\varepsilon)}-w_0\right\|^2 \\ &\geq -C + 2C + \log(1/\varepsilon) - \frac{LR^2}{2\gamma^2\sqrt{m}}(2C+\log(1/\varepsilon))^2 \\ &\geq \log(1/\varepsilon). \end{aligned}$$

The last step is due to the condition on m. The inequality above implies that $\widehat{F}_i(w) := f(y_i\Phi(w,x_i)) \leq \log(1+\varepsilon) \leq \varepsilon$, and thus $\widehat{F}(w) \leq \varepsilon$ as desired. This completes the proof.

With this, we may now prove Corollary 4.4.1.

Corollary 4.8.2 (Restatement of Corollary 4.4.1)**.** *Let Assumptions 4.2.1-4.2.2,4.4.1-4.4.2 hold and assume logistic loss. Suppose* $m \geq \frac{64^2L^2R^4}{\gamma^4}(2C+\log(T))^4$ *for a fixed training horizon* T*. Then, for any* $\eta \leq \min\{3, \frac{1}{L_{\widehat{F}}}\}$ *the training loss and generalization gap are bounded as follows:*

$$\begin{aligned} \widehat{F}(w_T) &\leq \frac{5(2C+\log(T))^2}{\gamma^2\eta T}, \\ \mathbb{E}\Big[F(w_T) - \widehat{F}(w_T)\Big] &\leq \frac{24\ell^2R^2}{\gamma^2 n}(2C+\log(T))^2. \end{aligned}$$

Proof: The given assumption on m satisfies the conditions of Proposition 4.4.1 for $\varepsilon = \frac{1}{T}$, $g(1/T) = \frac{1}{\gamma}(2C+\log(T))$. We can apply the results of our optimization and generalization results from Theorems 4.3.2 and 4.3.4 for a fixed T which satisfies $T \geq 1$. Note that we can assume without loss of generality that $\gamma \leq 1$ which implies that $g(1)^2 = 4C^2/\gamma^2 \geq 4$. Moreover, for logistic loss it holds $g(1)^2/\widehat{F}(w_0) \geq \frac{4C^2}{\gamma^2\log(1+e^C)} \geq 3$ for all $C \geq 1$. Therefore the condition on step-size simplifies to $\eta \leq \min\{3, 1/L_{\widehat{F}}\}$. This completes the proof.

Proof of Proposition 4.4.2

The proof of Proposition 4.4.2 has the following steps: First, we consider an infinite-width NTK separability assumption (Assumption 4.8.2) and show in Lemma 4.8.7 that it is equivalent with high-probability to the NTK-separability in Assumption 4.4.1 given logarithmic number of neurons. We then prove that the noisy-XOR dataset satisfies Assumption 4.8.2 for convex and locally strongly-convex activations. The result of Proposition 4.4.2 then follows by combining the two lemmas.

Assumption 4.8.2 (Infinite-width NTK-separability)**.** *There exists* $\overline{w}(\cdot) : \mathbb{R}^d \to \mathbb{R}^d$ *and* $\gamma > 0$ *such that* $\|\overline{w}(z)\|_2 \leq 1$ *for all* $z \in \mathbb{R}^d$*, and for all* $(x, y) \sim \mathcal{D}$*,*

$$y \int_{\mathbb{R}^d} \sigma'\left(\langle z, x\rangle\right) \cdot \langle \overline{w}(z), x\rangle \, \mathrm{d}\mu_{\mathrm{N}}(z) \geq \gamma,$$

where $\mu_N(\cdot)$ *denotes the standard Gaussian measure.*

Lemma 4.8.7. *Let* $\{(x_i, y_i)\}$ *be any dataset of size* $\tilde{n}$ *under Assumption 4.2.1, satisfying the separability condition of Assumption 4.8.2 with some margin* $\tilde{\gamma} > 0$*. Consider initialization* $w_0 \in \mathbb{R}^{d'}$ *where* $w_0 \sim N(0, I_{d'})$*. Then, with probability at least* $1 - \delta$ *the dataset is separable under Assumption 4.4.1 with margin at least* $\gamma = \tilde{\gamma} - \frac{\ell R}{\sqrt{2m}} \log^{1/2}(\tilde{n}/\delta)$*, i.e., there exists unit norm* $w^\star$ *such that for all* $i \in [\tilde{n}] : y_i \langle \nabla_1 \Phi(w_0, x_i), w^\star \rangle \geq \gamma$*.*

Proof: By the model's gradient we have for any $w^\star \in \mathbb{R}^{d'}$,

$$\phi_i := y_i \left\langle \nabla_1 \Phi(w_0, x_i), w^\star \right\rangle = y_i \sum_{j=1}^{m} \frac{a_j}{\sqrt{m}} \sigma'(\langle w_{0,j}, x_i\rangle)\langle x_i, w_j^\star\rangle. \tag{4.47}$$

Let $w_j^\star = \frac{a_j}{\sqrt{m}} \overline{w}(w_{0,j})$. Then $\|w^\star\| \leq 1$ and by Hoeffding's inequality it holds for all $t \geq 0$,

$$\Pr\left(\phi_i \geq \tilde{\gamma} - t\right) \geq 1 - \exp\left(\frac{-2t^2 m}{\ell^2 R^2}\right). \tag{4.48}$$

This leads to the desired result with an extra union bound over $i \in [\tilde{n}]$.

Lemma 4.8.8. *Consider the noisy XOR data distribution* $\{(\bar{x}_i, y_i)\}$ *and two-layer neural network with a convex activation which is* μ*-strongly convex in* $[-2, 2]$ *i.e.,* $\min_{t\in[-2,2]} \sigma''(t) \geq \mu$ *for some* $\mu > 0$*. Then the separability assumption 4.8.2 is satisfied with margin* $\gamma = \frac{\mu}{40d}$.

Proof: The proof is essentially similar to [110, Prop. 5.3] and thus we follow their notation and omit the details for brevity. While their proof relies rather crucially on the ReLU activation, it can be appropriately modified to obtain a similar margin bound under our different assumptions on the activation function. To see this, note that due to convexity of activation function, the integrand in the line above Eq. (D.4) is non-negative. Therefore, we can lower-bound the integral (which evaluates the margin) by restricting A_1 to $|p_1|< 1$. With this restriction we can use the local strong convexity of activation function to lower-bound the margin, i.e., to uniformly lower-bound $y_i \int_{\mathbb{R}^d} \sigma'\left(\langle z, x_i\rangle\right) \cdot \langle \bar{w}(z), x_i\rangle \, \mathrm{d}\mu_{\mathrm{N}}(z)$ for all $i \in [n]$. Specifically, note that with strong convexity in $[-2, 2]$, Eq. (D.4) in [110] changes to $\geq \frac{2p_1}{d-1} U(p_1) \min_{t\in[-2,2]} \sigma''(t) \geq \frac{2p_1\mu}{d-1} U(p_1)$ where $U(t) := \int_{-t}^{t} \varphi(\tau)\mathrm{d}\tau$ is the probability that a standard Gaussian random variable falls in $[-t, t]$. This leads to the final value for margin being $\frac{2\mu}{d-1} \int p_1 \, U(p_1) \, \mathbf{1}\left[p \in A_1\right] \mathrm{d}\mu_N(p) \geq \frac{8\mu}{(2\pi e)^{3/2}(d-1)} \int_0^1 p_1^3 \mathrm{d}p_1 \geq \frac{\mu}{40d}$, as desired.

Proposition 4.8.3 (Restatement of Proposition 4.4.2)**.** *Consider the noisy XOR data distribution* $\{(\bar{x}_i, y_i)\}$*. Assume the activation function is convex,* ℓ*-Lipschitz and* μ*-strongly convex in the interval* $[-2, 2]$ *for some* $\mu > 0$*, i.e.,* $\min_{t\in[-2,2]} \sigma''(t) \geq \mu$*. Moreover, assume Gaussian initialization* $w_0 \in \mathbb{R}^{d'}$ *with entries iid* $N(0, 1)$*. If* $m \geq \frac{80^2 d^3 \ell^2}{2\mu^2} \log(2/\delta)$*, then with probability at least* $1 - \delta$ *over the initialization, the NTK-separability Assumption 4.4.1 is satisfied with margin* $\gamma = \frac{\mu}{80d}$.

Proof: The claim follows by combining the last two lemmas. In particular, we derive the infinite width NTK-separability for the entire data distribution (of size 2^d) with

margin $\tilde{\gamma} = \frac{\mu}{40d}$ and by the assumption on width and noting $\tilde{n} = 2^d$, we have γ-separability by NTK for the entire distribution with probability $1-\delta$ where $\gamma = \tilde{\gamma} - \frac{\ell R}{\sqrt{2m}} \log^{1/2}(\tilde{n}/\delta) = \frac{\mu}{40d} - \frac{\ell R\sqrt{d}}{\sqrt{2m}} \log^{1/2}(1/\delta) \geq \frac{\mu}{80d}$. This completes the proof.

Finally, we show how to control the parameter C that bounds the model output at Gaussian initialization.

Lemma 4.8.9 (Initialization bound)**.** *Let Assumption 4.2.1 hold and assume the activation function to be ℓ-Lipschitz. Consider initialization $w_0 \in \mathbb{R}^{d'}$ where $w_0 \sim N(0, I_{d'})$. Given any $\delta \in (0,1)$, then with probability at least $1-\delta$, it holds for all $i \in [\tilde{n}]$ that*

$$|\Phi(w_0, x_i)| \leq \ell R\sqrt{2\log(2\tilde{n}/\delta)}. \tag{4.49}$$

Proof: Recall that if a function $\phi : \mathbb{R}^{d'} \to \mathbb{R}$ is G-Lipschitz then for Gaussian vector $Z = (Z_1, Z_2, \cdots, Z_{d'})$ where each component is i.i.d. standard Gaussian $Z_i \sim N(0,1)$, it holds for all $t \geq 0$ that $\Pr[|\phi(Z) - \mathbb{E}[\phi(Z)]| \geq t] \leq 2\exp(-\frac{t^2}{2G^2})$. Note that according to Lemma 4.8.10, $\Phi(\cdot, x_i)$ is (ℓR)-Lipschitz for any data point x_i. Therefore, with the given initialization for w_0, we have

$$\Pr\left[\left|\Phi(w_0, x_i) - \mathbb{E}[\Phi(w_0, x_i)]\right| \geq t\right] \leq 2\exp\left(-\frac{t^2}{2\ell^2 R^2}\right).$$

It also holds that $\mathbb{E}[\Phi(w_0, x_i)] = 0$. This is true since for half of second layer weights $a_j = 1$ and for the rest $a_j = -1$. Thus, we have $\Pr[|\Phi(w_0, x_i)| \geq t] \leq 2\exp(-\frac{t^2}{2\ell^2 R^2})$. A union bound yields that uniformly over $i \in [\tilde{n}]$, we have $\Pr[|\Phi(w_0, x_i)| \geq t] \leq 2\tilde{n} \cdot \exp(-\frac{t^2}{2\ell^2 R^2})$ which concludes the claim of lemma.

4.8.4 Gradients and Hessian calculations

Definitions

Assume IID data $(x, y) \sim \mathcal{D}$, $x \in \mathbb{R}^d, y \in \{\pm 1\}$. Denote for convenience $z := yx$. Suppose two-layer neural network model

$$\Phi(w, x_i) = \frac{1}{\sqrt{m}} \sum_{j \in [m]} a_j \sigma(\langle w_j, x \rangle) \tag{4.50}$$

$a_j \in \{\pm 1\}, j \in [m]$ and first-layer weights trained by GD on

$$\widehat{F}(w) = \frac{1}{n} \sum_{i \in [n]} f(y_i \Phi(w, x_i)) =: \frac{1}{n} \sum_{i \in [n]} f(w, z_i)\,. \tag{4.51}$$

for loss function $f : \mathbb{R} \to \mathbb{R}$.

For convenience define

$$\widehat{F}'(w) = \frac{1}{n} \sum_{i \in [n]} |f'(y_i \Phi(w, x_i))| \tag{4.52a}$$

$$\widehat{F}''(w) = \frac{1}{n} \sum_{i \in [n]} |f''(y_i \Phi(w, x_i))| \tag{4.52b}$$

Model's Gradient/Hessian

Lemma 4.8.10. *The following are true for the model* (4.50) *under Assumption 4.2.2.*

1. $\|\nabla_1 \Phi(w, x)\| \leq \ell R$.
2. $\|\nabla_1^2 \Phi(w, x)\| \leq \frac{LR^2}{\sqrt{m}}$.

Proof: Direct calculation yields that,

$$\nabla_1 \Phi(w,x) = \frac{1}{\sqrt{m}} \begin{bmatrix} a_1 \sigma'(\langle w_1, x\rangle) x \\ \cdot \\ \cdot \\ a_m \sigma'(\langle w_m, x\rangle) x \end{bmatrix}$$

Noting that $\sigma'(\cdot) \leq \ell$,

$$\begin{aligned} \|\nabla_1 \Phi(w,x)\|^2 &= \frac{1}{m} \sum_{j=1}^{m} \sum_{i=1}^{d} (x(i)\sigma'(\langle w_j, x\rangle))^2 \\ &\leq \ell^2 \|x\|^2 \\ &\leq \ell^2 R^2. \end{aligned} \tag{4.53}$$

For the Hessian,

$$\frac{\partial^2 \Phi(w,x)}{\partial w_{ij} \partial w_{k\ell}} = \frac{1}{\sqrt{m}} x(j) x(\ell) a_i \sigma''(\langle w_i, x\rangle) \mathbf{1}_{\{i=k\}}. \tag{4.54}$$

Thus,

$$\nabla_1^2 \Phi(w,x) = \frac{1}{\sqrt{m}} \operatorname{diag}\left(a_1 \sigma''(\langle w_1, x\rangle) x x^T, \ldots, a_m \sigma''(\langle w_m, x\rangle) x x^T\right)$$

for any unit norm vector $u \in \mathbb{R}^{md}$, define $\bar{u}_i := [u_{(i-1)m+1} : u_{im}] \in \mathbb{R}^d$. Moreover,

define the matrix $\nabla^2_{w_i}\Phi(w,x) \in \mathbb{R}^{d\times d}$ such that $[\nabla^2_{w_i}\Phi(w,x)]_{j\ell} = \frac{\partial^2\Phi(w,x)}{\partial w_{ij}\partial w_{i\ell}}$

$$\begin{aligned}\left\|u^\top \nabla_1^2\Phi(w,x)\right\|^2 &= \sum_{i=1}^m \left\|u_i^\top \nabla^2_{w_i}\Phi(w,x)\right\|^2 \\ &\le \sum_{i=1}^m \left\|\nabla^2_{w_i}\Phi(w,x)\right\|^2 \|\bar{u}_i\|^2 \\ &\le \sum_{i=1}^m \frac{L^2}{m}\|x\|^4\|\bar{u}_i\|^2 \\ &\le \frac{L^2R^4}{m}.\end{aligned}$$

This completes the proof.

Objective's Gradient/Hessian

Lemma 4.8.11. *Let Assumption 4.2.2 hold. Then, the following are true for the loss gradient and Hessian:*

1. $\|\nabla\widehat{F}(w)\| \le \ell R\,\widehat{F}'(w)$.
2. $\|\nabla^2\widehat{F}(w)\| \le \ell^2R^2\widehat{F}''(w) + \frac{LR^2}{\sqrt{m}}\widehat{F}'(w)$.
3. $\lambda_{\min}\left(\nabla^2\widehat{F}(w)\right) \ge -\frac{LR^2}{\sqrt{m}}\widehat{F}'(w)$.

Proof: The loss gradient is derived as follows,

$$\nabla\widehat{F}(w) = \frac{1}{n}\sum_{i=1}^n f'(y_i\Phi(w,x_i))y_i\nabla_1\Phi(w,x_i)$$

Recalling that $y_i \in \{\pm 1\}$, we can write

$$\begin{aligned}\left\|\nabla \widehat{F}(w)\right\| &= \frac{1}{n}\left\|\sum_{i=1}^{n} f'(y_i\Phi(w,x_i))y_i\nabla_1\Phi(w,x_i)\right\| \\ &\le \frac{1}{n}\sum_{i=1}^{n}|f'(y_i\Phi(w,x_i))|\left\|\nabla_1\Phi(w,x_i)\right\|. \\ &\le \ell R\, F'(w). \end{aligned} \tag{4.55}$$

For the Hessian of loss, note that

$$\nabla^2 \widehat{F}(w) = \frac{1}{n}\sum_{i=1}^{n} f''(y_i\Phi(w,x_i))\nabla_1\Phi(w,x_i)\nabla_1\Phi(w,x_i)^\top + f'(y_i\Phi(w,x_i))y_i\nabla_1^2\Phi(w,x_i). \tag{4.56}$$

It follows that

$$\begin{aligned}\left\|\nabla^2 \widehat{F}(w)\right\| &= \left\|\frac{1}{n}\sum_{i=1}^{n} f'(y_i\Phi(w,x_i))y_i\nabla_1^2\Phi(w,x_i) + f''(y_i\Phi(w,x_i))\nabla_1\Phi(w,x_i)\nabla_1\Phi(w,x_i)^\top\right\| \\ &\le \frac{1}{n}\sum_{i=1}^{n}|f'(y_i\Phi(w,x_i))|\left\|\nabla_1^2\Phi(w,x_i)\right\| + |f''(y_i\Phi(w,x_i))|\left\|\nabla_1\Phi(w,x_i)\nabla_1\Phi(w,x_i)^\top\right\| \\ &\le \frac{1}{n}\sum_{i=1}^{n}|f'(y_i\Phi(w,x_i))|\left\|\nabla_1^2\Phi(w,x_i)\right\| + |f''(y_i\Phi(w,x_i))|\left\|\nabla_1\Phi(w,x_i)\right\|^2 \\ &\le \frac{LR^2}{\sqrt{m}}F'(w) + \ell^2R^2F''(w). \end{aligned} \tag{4.57}$$

To lower-bound the minimum eigenvalue of Hessian, note that f is convex and thus $f''(\cdot) \ge 0$. Therefore the first term in (4.56) is positive semi-definite and the second term

can be lower-bounded as follows,

$$\begin{aligned}\lambda_{\min}(\nabla^2 \widehat{F}(w)) &\geq -\left\| \frac{1}{n}\sum_{i=1}^{n} y_i f'(y_i \Phi(w,x_i)) \nabla_1^2 \Phi(w,x_i) \right\| \\ &\geq -\frac{1}{n}\sum_{i=1}^{n} |y_i f'(y_i \Phi(w,x_i))| \left\| \nabla_1^2 \Phi(w,x_i) \right\| \\ &\geq -\frac{LR^2}{\sqrt{m}} F'(w).\end{aligned}$$

Corollary 4.8.3 (Self-boundedness of Objective)**.** *Let Assumption 4.2.2 hold. If the loss satisfies Assumptions 4.2.4 (with $\beta_f = 1$) and 4.8.1, then*

1. $\|\nabla \widehat{F}(w)\| \leq \ell R\, \widehat{F}(w)$.
2. $\|\nabla^2 \widehat{F}(w)\| \leq \left(\ell^2 R^2 + \frac{LR^2}{\sqrt{m}}\right) \widehat{F}(w)$.
3. $\lambda_{\min}\left(\nabla^2 \widehat{F}(w)\right) \geq -\frac{LR^2}{\sqrt{m}} \widehat{F}(w)$.

If in addition the loss satisfies Assumptions 4.2.3.A and 4.2.3.B with $L_f = G_f = 1$, then

6. $\|\nabla \widehat{F}(w)\| \leq \ell R$.
7. $\|\nabla^2 \widehat{F}(w)\| \leq \ell^2 R^2 + \frac{LR^2}{\sqrt{m}}$.

Proof: For self-bounded losses we have $\widehat{F}'(w) \leq \widehat{F}(w)$ and $\widehat{F}''(w) \leq \widehat{F}(w)$. If the loss is 1-Lipschitz and 1-smooth we have $\widehat{F}'(w) \leq 1$ and $\widehat{F}''(w) \leq 1$. Thus, the claims immediately follow from Lemma 4.8.11.

Chapter 5

Fast Convergence in Learning Neural Networks with Separable Data

5.1 Introduction

5.1.1 Motivation

A wide variety of machine learning algorithms for classification tasks rely on learning a model using monotonically decreasing loss functions such as logistic loss or exponential loss. In modern practice these tasks are often accomplished using over-parameterized models such as large neural networks where the model can interpolate the training data, i.e., it can achieve perfect classification accuracy on the samples. In particular, it is often the case that the training of the model is continued until achieving approximately zero training loss [143].

Over the last decade there has been remarkable progress in understanding or improving the convergence and generalization properties of over-parameterized models trained by various choices of loss functions including logistic loss and quadratic loss. For the quadratic

loss it has been shown that over-parameterization can result in significant improvements in the training convergence rate of (stochastic)gradient descent on empirical risk minimization algorithms. Notably, quadratic loss on two-layer ReLU neural networks is shown to satisfy the Polyak-Łojasiewicz(PL) condition [127, 144, 126]. In fact, the PL property is a consequence of the observation that the tangent kernel associated with the model is a non-singular matrix. Moreover, in this case the PL parameter, which specifies the rate of convergence, is the smallest eigenvalue of the tangent kernel[126]. The fact that over-parameterized neural networks trained by quadratic loss satisfy the PL condition, guarantees that the loss convergences exponentially fast to a global optimum. The global optimum in this case is a model which "perfectly" interpolates the data, where we recall that perfect interpolation requires that the model output for every training input is precisely equal to the corresponding label.

On the other hand, gradient descent using un-regularized logistic regression with linear models and separable data is biased toward the max-margin solution. In particular, in this case the parameter converges in direction with the rate $O(1/\log(t))$ to the solution of hard margin SVM problem, while the training loss converges to zero at the rate $\tilde{O}(1/t)$ [138, 113]. More recently, normalized gradient descent has been proposed as a promising approach for fast convergence of exponentially tailed losses. In this method, at any iteration the step-size is chosen proportionally to the inverse of value of training loss function [20]. This results in choosing unboundedly increasing step-sizes for the iterates of gradient descent. This choice of step-size leads to significantly faster rates for the parameter's directional convergence. In particular, for linear models with separable data, it is shown that normalized GD with decaying step-size enjoys a rate of $O(1/\sqrt{t})$ in directional parameter convergence to the max-margin separator [20]. This has been improved to $O(1/t)$ with normalized GD using fixed step-size [21].

Despite remarkable progress in understanding the behavior of normalized GD with

separable data, these results are only applicable to the implicit bias behavior of "linear models". In this chapter, we aim to discover for the first time, the dynamics of learning a two-layer neural network with normalized GD trained on separable data. We also wish to realize the iterate-wise test error performance of this procedure. We show that using normalized GD on an exponentially-tailed loss with a two layered neural network leads to exponentially fast convergence of the loss to the global optimum. This is comparable to the convergence rate of $O(1/t)$ for the global convergence of neural networks trained with exponentially-tailed losses. Compared to the convergence analysis of standard GD which is usually carried out using smoothness of the loss function, here for normalized GD we use the Taylor's expansion of the loss and use the fact the operator norm of the Hessian is bounded by the loss. Next, we apply a lemma in our proof which shows that exponentially-tailed losses on a two-layered neural network satisfy a log-Lipschitzness condition throughout the iterates of normalized GD. Moreover, crucial to our analysis is showing that the ℓ_2 norm of the gradient at every point is upper-bounded and lower-bounded by constant factors of the loss under given assumptions on the activation function and the training data. Subsequently, the log-Lipschitzness property together with the bounds on the norm of Gradient and Hessian of the loss function ensures that normalized GD is indeed a descent algorithm. Moreover, it results in the fact that the loss value decreases by a constant factor after each step of normalized GD, resulting in the promised geometric rate of decay for the loss.

5.1.2 Contributions

In Section 5.2.1 we introduce conditions –namely log-Lipschitz and self-boundedness assumptions on the gradient and the Hessian– under which the training loss of the normalized GD algorithm converges exponentially fast to the global optimum. More

importantly, in Section 5.2.2 we prove that the aforementioned conditions are indeed satisfied by two-layer neural networks trained with an exponentially-tailed loss function. This yields the first theoretical guarantee on the convergence of normalized GD for non-linear models. We also study a stochastic variant of normalized GD and investigate its training loss convergence in Section 5.2.4.

In Section 5.2.3 we study, for the first time, the finite-time test loss and test error performance of normalized GD for convex losses. In particular, we derive bounds of order $O(1/n)$ on the expected generalization error of normalized GD, where n is the training-set size.

5.1.3 Prior Works

The theoretical study of the optimization landscape of over-parameterized models trained by GD or SGD has been the subject of several recent works. The majority of these works study over-parameterized models with specific choices of loss functions, mainly quadratic or logistic loss functions. For quadratic loss, the exponential convergence rate of over-parameterized neural networks is proved in several recent works e.g., [127, 144, 145, 125, 132, 146, 121, 147, 126]. These results naturally relate to the Neural Tangent Kernel(NTK) regime of infinitely wide or sufficiently large initialized neural networks [105] in which the iterates of gradient descent stay close to the initialization. The NTK approach can not be applied to our setting as the parameters' norm in our setting is growing as $\Theta(t)$ with the NGD updates.

The majority of the prior results apply to the quadratic loss. However, the state-of-the-art architectures for classification tasks use unregularized ERM with logistic/exponential loss functions. Notably, for these losses over-parameterization leads to infinite norm optimizers. As a result, the objective in this case does not satisfy strong convexity or the

PL condition even for linear models. The analysis of loss and parameter convergence of logistic regression on separable data has attracted significant attention in the last five years. Notably, a line of influential works have shown that gradient descent provably converges in direction to the max-margin solution for linear models and two-layer homogenous neural networks. In particular, the study of training loss and implicit bias behavior of GD on logistic/exponential loss was first initiated in the settings of linear classifiers [148, 149, 138, 113, 20]. The implicit bias behavior of GD with logistic loss in two-layer neural networks was later studied by [150, 140, 141]. The loss landscape of logistic loss for over-parameterized neural networks and structured data is analyzed in [151, 136], where it is proved that GD converges to a global optima at the rate $O(1/t)$. The majority of these results hold for standard GD while we focus on normalized GD.

The generalization properties of GD/SGD with binary and multi-class logistic regression is studied in [114, 152] for linear models and in [123, 153, 115] for neural networks. Recently, [154] studied the generalization error of decentralized logistic regression through a stability analysis. For our generalization analysis we use an algorithmic stability analysis [13, 17, 14]. However, unlike these prior works we consider normalized GD and derive the first generalization analysis for this algorithm.

The benefits of normalized GD for speeding up the directional convergence of GD for linear models was suggested by [20, 21]. This chapter contributes to this line of work. Compared to the prior works which are focused on implicit behavior of linear models, we study non-linear models and derive training loss convergence rates. We also study, the generalization performance of normalized GD for convex objectives.

Notation

We use $\|\cdot\|$ to denote the operator norm of a matrix and also to denote the ℓ_2-norm of a vector. The Frobenius norm of a matrix W is shown by $\|W\|_F$. The Gradient and the Hessian of a function $F : \mathbb{R}^d \to \mathbb{R}$ are denoted by ∇F and $\nabla^2 F$. Similarly, for a function $F : \mathbb{R}^d \times \mathbb{R}^{d'} \to \mathbb{R}$ that takes two input variables, the Gradient and the Hessian with respect to the ith variable (where $i = 1, 2$) are denoted by $\nabla_i F$ and $\nabla_i^2 F$, respectively. For functions $F, G : \mathbb{R} \to \mathbb{R}$, we write $F(t) = O(G(t))$ when $|F(t)| \leq m\, G(t)$ after $t \geq t_0$ for positive constants m, t_0. We write $F(t) = \tilde{O}(G(t))$ when $F(t) = O(G(t)H(t))$ for a polylogarithmic function H. Finally, we denote $F(t) = \Theta(G(t))$ if $|F(t)| \leq m_1 G(t)$ and $|F(t)| \geq m_2 G(t)$ for all $t \geq t_0$ for some positive constants m_1, m_2, t_0.

5.1.4 Problem Setup

We consider unconstrained and unregularized empirical risk minimization (ERM) on n samples,

$$\min_{w \in \mathbb{R}^{\widetilde{d}}} F(w) := \frac{1}{n} \sum_{i=1}^{n} f\left(y_i \Phi(w, x_i)\right). \tag{5.1}$$

The ith sample $z_i := (x_i, y_i)$ consists of a data point $x_i \in \mathbb{R}^d$ and its associated label $y_i \in \{\pm 1\}$. The function $\Phi : \mathbb{R}^{\widetilde{d}} \times \mathbb{R}^d \to \mathbb{R}$ represents the model taking the weights vector w and data point x to approximate the label. In this section, we take Φ as a neural network with one hidden layer and m neurons,

$$\Phi(w, x) := \sum_{j=1}^{m} a_j \sigma(\langle w_j, x \rangle).$$

Here $\sigma : \mathbb{R} \to \mathbb{R}$ is the activation function and $w_j \in \mathbb{R}^d$ denotes the input weight vector of the jth hidden neuron. $w \in \mathbb{R}^{\widetilde{d}}$ represents the concatenation of these weights

i.e., $w = [w_1; w_2; \ldots; w_m]$. In our setting the total number of parameters and hence the dimension of w is $\widetilde{d} = md$. We assume that only the first layer weights w_j are updated during training and the second layer weights $a_j \in \mathbb{R}$ are initialized randomly and are maintained fixed during training. The function $f : \mathbb{R} \to \mathbb{R}$ is non-negative and monotonically decreases such that $\lim_{t\to+\infty} f(t) = 0$. In this section, we focus on the exponential loss $f(t) = \exp(-t)$, but we expect that our results apply to a broader class of loss functions that behave similarly to the exponential loss for large t, such as logistic loss $f(t) = \log(1 + \exp(-t))$.

We consider activation functions with bounded absolute value for the first and second derivatives.

Assumption 5.1.1 (Activation function)**.** *The activation function $\sigma : \mathbb{R} \to \mathbb{R}$ is smooth and for all $t \in \mathbb{R}$*

$$|\sigma''(t)| \leq L.$$

Moreover, there are positive constants α, ℓ such that σ satisfies for all $t \in \mathbb{R}$,

$$\alpha \leq \sigma'(t) \leq \ell.$$

An example satisfying the above condition is the activation function known as smoothed-leaky-ReLU which is a smoothed variant of the leaky-ReLU activation $\sigma(t) = \ell t\, \mathbb{I}(t \geq 0) + \alpha t\, \mathbb{I}(t \leq 0)$, where $\mathbb{I}(\cdot)$ denotes the 0–1 indicator function.

Throughout the chapter we let R and a denote the maximum norm of data points and second layer weights, respectively, i.e.,

$$R := \max_{i\in[n]} \|x_i\|, \qquad a := \max_{j\in[m]} |a_j|.$$

Throughout the chapter we assume $R = \Theta(1)$ w.r.t. problem parameters and $a = \frac{1}{m}$. We also denote the *training loss* of the model by F, defined in (5.1) and define the *train error* as misclassification error over the training data, or formally by $F_{0-1}(w) := \frac{1}{n}\sum_{i=1}^{n} \mathbb{I}(\text{SIGN}(\Phi(w, x_i)) \neq y_i)$.

Normalized GD. We consider the iterates of normalized GD as follows,

$$w_{t+1} = w_t - \eta_t \nabla F(w_t). \tag{5.2}$$

The step size is chosen inversely proportional to the loss value i.e., $\eta_t = \eta / F(w_t)$, implying that the step-size is growing unboundedly as the algorithm approaches the optimum solution. Since the gradient norm decays proportionally to the loss, one can equivalently choose $\eta_t = \eta / \|\nabla F(w_t)\|$.

5.2 Main Results

For convergence analysis in our case study, we introduce a few definitions.

Definition 5.2.1 (log-Lipschitz Objective)**.** The training loss $F : \mathbb{R}^{\tilde{d}} \to \mathbb{R}$ satisfies the log-Lipschitzness property if for all $w, w' \in \mathbb{R}^{\tilde{d}}$,

$$\max_{v \in [w, w']} F(v) \leq F(w) \cdot \tilde{c}_{w,w'},$$

where $[w, w']$ denotes the line between w and w' and we define $\tilde{c}_{w,w'} := \exp\left(c(\|w - w'\| + \|w - w'\|^2)\right)$ where the positive constant c is independent of w, w'.

As we will see in the following sections, log-Lipschitzness is a property of neural

networks trained with exponentially tailed losses with $c = \Theta(\frac{1}{\sqrt{m}})$. We also define the property "log-Lipschitzness in the gradient path" if for all w_t, w_{t-1} in Eq. (5.2) there exists a constant C such that,

$$\max_{v \in [w_t, w_{t+1}]} F(v) \leq C\, F(w_t).$$

Definition 5.2.2 (Self lower-bounded gradient)**.** The loss function $F : \mathbb{R}^{\widetilde{d}} \to \mathbb{R}$ satisfies the self-lower bounded Gradient condition for a function, if these exists a constant μ such that for all w,

$$\|\nabla F(w)\| \geq \mu\, F(w).$$

Definition 5.2.3 (Self-boundedness of the gradient)**.** The loss function $F : \mathbb{R}^{\widetilde{d}} \to \mathbb{R}$ satisfies the self-boundedness of the gradient condition for a constant h, if for all w

$$\|\nabla F(w)\| \leq h\, F(w).$$

The above two conditions on the upper-bound and lower bound of the gradient norm based on loss can be thought as the equivalent properties of smoothness and the PL condition but for our studied case of exponential loss. To see this, note that smoothness and PL condition provide upper and lower bounds for the square norm of gradient. In particular, by L-smoothness one can deduce that $\|\nabla F(w)\|^2 \leq 2L(F(w) - F^\star)$ (e.g., [155]) and by the definition of μ-PL condition $\|\nabla F(w)\|^2 \geq 2\mu(F(w) - F^\star)$ [156, 157].

The next necessary condition is an upper-bound on the operator norm of the Hessian of loss.

Definition 5.2.4 (Self-boundedness of the Hessian)**.** The loss function $F : \mathbb{R}^{\widetilde{d}} \to \mathbb{R}$

satisfies the self-boundedness of the Hessian property for a constant H, if for all w,

$$\|\nabla^2 F(w)\| \leq H\, F(w),$$

where $\|\cdot\|$ denotes the operator norm.

It is worthwhile to mention here that in the next sections of the chapter, we prove all the self lower and upper bound in Definitions 5.2.1-5.2.4 are satisfied for a two-layer neural network under some regularity conditions.

5.2.1 Convergence Analysis of Training Loss

The following theorem states that under the conditions above, the training loss converges to zero at an exponentially fast rate.

Theorem 5.2.1 (Convergence of Training Loss)**.** *Consider normalized gradient descent update rule with loss F and step-size η_t. Assume F and the normalized GD algorithm satisfy log-Lipschitzness in the gradient path with parameter C, as well as self-boundedness of the Gradient and the Hessian and the self-lower bounded Gradient properties with parameters h, H and μ, respectively. Let $\eta_t = \frac{\eta}{F(w_t)}$ for all $t \in [T]$ and for any positive constant η satisfying $\eta \leq \frac{\mu^2}{HCh^2}$. Then for the training loss at iteration T the following bound holds:*

$$F(w_{_T}) \leq (1 - \frac{\eta\mu^2}{2})^T F(w_0). \tag{5.3}$$

Remark 5.2.1. The proof of Theorem 5.2.1 is provided in Appendix 5.5.1, where we use a Taylor expansion of the loss and apply the conditions of the theorem. It is worth noting that the rate obtained for normalized GD in Theorem 5.2.1 is significantly faster than the rate of $\widetilde{O}(\frac{1}{T})$ for standard GD with logistic or exponential loss in neural networks (e.g.,

[151, Thm 4.4], and [158, Thm 2]). Additionally, for a continuous-time perspective on the training convergence of normalized GD, we refer to Proposition 5.5.1 in the appendix, which presents a convergence analysis based on *normalized Gradient Flow*. The advantage of this approach is that it does not require the self-bounded Hessian property and can be used to show exponential convergence of normalized Gradient Flow for leaky-ReLU activation.

5.2.2 Two-Layer Neural Networks

In this section, we prove that the conditions that led to Theorem 5.2.1 are in fact satisfied by a two-layer neural network. Consequently, this implies that the training loss bound in Eq.(5.3) is valid for this class of functions. We choose $f(t) = \exp(-t)$ for simpler proofs, however an akin result holds for the broader class of exponentially tailed loss functions.

First, we start with verifying the log-Lipschitzness condition (Definition 5.2.1). In particular, here we prove a variation of this property for the iterates of normalized GD i.e., where w, w' are chosen as w_t, w_{t+1}. The proof is included in Appendix 5.5.2.1.

Lemma 5.2.1 (log-Lipschitzness in the gradient path)**.** *Let F be as in (5.1) for the exponential loss f and let Φ be a two-layer neural network with the activation function satisfying Assumption 5.1.1. Consider the iterates of normalized GD with the step-size $\eta_t = \frac{\eta}{F(w_t)}$. Then for any $\lambda \in [0, 1]$ the following inequality holds:*

$$F(w_t + \lambda(w_{t+1} - w_t)) \leq \exp(\lambda c)\, F(w_t), \tag{5.4}$$

for a positive constant c independent of λ, w_t and w_{t+1}.

As a direct consequence, it follows that,

$$\max_{v\in[w_t,w_{t+1}]} F(v) \leq C\, F(w_t), \tag{5.5}$$

for a numerical constant C.

The next two lemmas state sufficient conditions for F to satisfy the self-lower boundedness for its gradient (Definition 5.2.2). The proofs are deferred to Appendices 5.5.2.2-5.5.2.3.

Lemma 5.2.2 (Self lower-boundedness of gradient)**.** *Let F be as in (5.1) for the exponential loss f and let Φ be a two-layer neural network with the activation function satisfying Assumption 5.1.1. Assume the training data is linearly separable with margin γ. Then F satisfies the self-lower boundedness of gradient with the constant $\mu = \frac{\alpha\gamma}{\sqrt{m}}$ for all w, i.e.,* $\|\nabla F(w)\| \geq \mu F(w)$.

Next, we aim to show that the condition $\|\nabla F(w)\| \geq \mu F(w)$, holds for training data separable by a two-layer neural network during gradient descent updates. In particular, we assume the Leaky-ReLU activation function taking the following form,

$$\sigma(t) = \begin{cases} \ell\, t & t \geq 0, \\ \alpha\, t & t < 0. \end{cases} \tag{5.6}$$

for arbitrary non-negative constants α, ℓ. This includes the widely-used ReLU activation as a special case. Next lemma shows that when the weights are such that the neural network separates the training data, the self-lower boundedness condition holds.

Lemma 5.2.3. *Let F be in (5.1) for the exponential loss f and let Φ be a two-layer neural network with activation function in Eq.(5.6). Assume the first layer weights $w \in \mathbb{R}^{\widetilde{d}}$ are*

such that the neural network separates the training data with margin γ. Then F satisfies the self- lower boundedness of gradient, i.e, $\|\nabla F(w)\| \geq \mu F(w)$, where $\mu = \gamma$.

A few remarks are in place. The result of Lemma 5.2.3 is relevant for w that can separate the training data. Especially, this implies the self lower-boundedness property after GD iterates succeed in finding an interpolator. However, we should also point out that the non-smoothness of leaky-ReLU activation functions precludes the self-bounded Hessian property and it remains an interesting future direction to prove the self lower-boundedness property with general smooth activations. On the other hand, the convergence of normalized "Gradient-flow" does not require the self-bounded Hessian property, as demonstrated in Proposition 5.5.1. This suggests that Lemma 5.2.3 can be applied to prove the convergence of normalized Gradient-flow with leaky-ReLU activations. It is worth highlighting that we have not imposed any specific initialization conditions in our analysis as the self-lower bounded property is essentially sufficient to ensure global convergence.

Next lemma derives the self-boundedness of the gradient and Hessian (c.f. Definitions 5.2.3-5.2.4) for our studied case. The proof of Lemma 5.2.4 (in Appendix 5.5.2.4) follows rather straight-forwardly from the closed-form expressions of gradient and Hessian and using properties of the activation function.

Lemma 5.2.4 (Self-boundedness of the gradient and Hessian)**.** *Let F be in (5.1) for the exponential loss f and let Φ be a two-layer neural network with the activation function satisfying Assumption 5.1.1. Then F satisfies the self-boundedness of gradient and Hessian with constants $h = \frac{\ell R}{\sqrt{m}}, H := \frac{LR^2}{m^2} + \frac{\ell^2 R^2}{m}$ i.e.,*

$$\|\nabla F(w)\| \leq hF(w), \quad \|\nabla^2 F(w)\| \leq HF(w).$$

We conclude this section by offering a few remarks regarding our training convergence

results. We emphasize that combining Theorem 5.2.1 and Lemmas 5.2.1-5.2.4 achieves the convergence of training loss of normalized Gradient Descent for two-layer networks. Moreover, in Appendix 5.5.4, we refer to Proposition 5.5.1 which presents a continuous time convergence analysis of normalized GD based on Gradient Flow. This result is especially relevant in the context of leaky-ReLU activation, where Proposition 5.5.1 together with Lemma 5.2.3 shows exponential convergence of normalized Gradient-flow. Finally, we remark that experiments of the training performance of normalized GD are deferred to Section 5.3.

5.2.3 Generalization Error

In this section, we study the generalization performance of normalized GD algorithm. Formally, the *test loss* for the data distribution $\mathcal{D}$ is defined as follows,

$$\widetilde{F}(w) := \mathbb{E}_{(x,y)\sim\mathcal{D}}\Big[f(y\Phi(w,x))\Big].$$

Depending on the choice of loss f, the test loss might not always represent correctly the classification performance of a model. For this, a more reliable standard is the *test error* which is based on the $0-1$ loss,

$$\widetilde{F}_{0-1}(w) := \mathbb{E}_{(x,y)\sim\mathcal{D}}\Big[\mathbb{I}(y \neq \textsc{sign}(\Phi(w,x)))\Big].$$

We also define the *generalization loss* as the gap between training loss and test loss. Likewise, we define the *generalization error* based on the train and test errors.

With these definitions in place, we are ready to state our results. In particular, in this section we prove that under the normalized GD update rule, the generalization loss at step T is bounded by $O(\frac{T}{n})$ where recall that n is the training sample size. While, the

dependence of generalization loss on T seems unappealing, we show that this is entirely due to the fact that a convex-relaxation of the $0-1$ loss, i.e. the loss function f, is used for evaluating the generalization loss. In particular, we can deduce that under appropriate conditions on loss function and data (c.f. Corollary 5.2.1), the test error is related to the test loss through,

$$\widetilde{F}_{0-1}(w_{_T}) = O(\frac{\widetilde{F}(w_{_T})}{\|w_{_T}\|}).$$

As we will see in the proof of Corollary 5.2.1, for normalized GD with exponentially tailed losses the weights norm $\|w_{_T}\|$ grows linearly with T. Thus, this relation implies that the test error satisfies $\widetilde{F}_{0-1}(w_{_T}) = O(\frac{1}{n})$. Essentially, this bound on the misclassification error signifies the fast convergence of normalized GD on test error and moreover, it shows that normalized GD never overfits during its iterations.

It is worthwhile to mention that our generalization analysis is valid for any model Φ such that $f(y\Phi(\cdot, x))$ is convex for any $(x, y) \sim \mathcal{D}$. This includes linear models i.e., $\Phi(w, x) = \langle w, x\rangle$ or the Random Features model [159], i.e., $\Phi(w, x) = \langle w, \sigma(Ax)\rangle$ where $\sigma(\cdot)$ is applied element-wise on its entries and the matrix $A \in \mathbb{R}^{m\times d}$ is initialized randomly and kept fixed during train and test time. Our results also apply to neural networks in the NTK regime due to the convex-like behavior of optimization landscape in the infinite-width limit.

We study the generalization performance of normalized GD, through a stability analysis [13]. The existing analyses in the literature for algorithmic stability of $\tilde{L}-$smooth losses, rely on the step-size satisfying $\eta_t = O(1/\tilde{L})$. This implies that such analyses can not be employed for studying increasingly large step-sizes as in our case η_t is unboundedly growing. In particular, the common approach in the stability analysis [17, 14] uses the "non-expansiveness" property of standard GD with smooth and convex losses, by showing that for $\eta \leq 2/\tilde{L}$ and for any two points $w, v \in \mathbb{R}^{\tilde{d}}$, it holds that $\|w-\eta\nabla F(w)-(v-\eta\nabla F(v))\|\leq$

$\|w - v\|$. Central to our stability analysis is showing that under the assumptions of self-boundedness of Gradient and Hessian, the normalized GD update rule satisfies the non-expansiveness condition with any step-size satisfying both $\eta \lesssim \frac{1}{F(w)}$ and $\eta \lesssim \frac{1}{F(v)}$. The proof is included in Appendix 5.5.3.1.

Lemma 5.2.5 (Non-expansiveness of normalized GD)**.** *Assume the loss F to satisfy convexity and self-boundedness for the gradient and the Hessian with parameter $h \leq 1$ (Definitions 5.2.3-5.2.4). Let $v, w \in \mathbb{R}^d$. If $\eta \leq \frac{1}{h \cdot \max(F(v), F(w))}$, then*

$$\|w - \eta \nabla F(w) - (v - \eta \nabla F(v))\| \leq \|w - v\|.$$

The next theorem characterizes the test loss for both Lipschitz and smooth objectives. Before stating the theorem, we need to define δ. For the leave-one-out parameter $w_t^{\neg i}$ and loss $F^{\neg i}(\cdot)$ defined as

$$w_{t+1}^{\neg i} = w_t^{\neg i} - \eta_t \nabla F^{\neg i}(w_t^{\neg i}),$$

and

$$F^{\neg i}(w) := \frac{1}{n} \sum_{\substack{j=1 \\ j \neq i}}^{n} f(w, z_j),$$

we define $\delta \geq 1$ to be any constant which satisfies for all $t \in [T], i \in [n]$, the following

$$F^{\neg i}(w_t^{\neg i}) \leq \delta \, F^{\neg i}(w_t).$$

While this condition seems rather restrictive, we prove in Lemma 5.5.1 in Appendix 5.5.3.3 that the condition on δ is satisfied by two-layer neural networks with sufficient

over-parameterization. With these definitions in place, we are ready to state the main theorem of this section.

Theorem 5.2.2 (Test loss). *Consider normalized GD update rule with* $\eta_t = \frac{\eta}{F(w_t)}$ *where* $\eta \leq \frac{1}{h\delta}$. *Assume the loss* F *to be convex and to satisfy the self-bounded gradient and Hessian property with a parameter* h *(Definitions 5.2.3-5.2.4). Then the following statements hold for the test loss: (i) if the loss* F *is* G*-Lipschitz, then the generalization loss at step* T *satisfies*

$$\mathbb{E}[\widetilde{F}(w_{_T}) - F(w_{_T})] \leq \frac{2GT}{n}.$$

(ii) if the loss F *is* $\tilde{L}$*-smooth, then the test loss at step* T *satisfies,*

$$\mathbb{E}[\widetilde{F}(w_{_T})] \leq 4\mathbb{E}[F(w_{_T})] + \frac{3\tilde{L}^2T}{n},$$

where all expectations are over training sets.

The proof of Theorem 5.2.2 is deferred to the Appendix. As discussed earlier in this section, the test loss dependence on T is due to the rapid growth of the ℓ_2 norm of w_t. As a corollary, we show that the generalization error is bounded by $O(\frac{1}{n})$. For this, we assume the next condition.

Assumption 5.2.1 (Margin). *There exists a constant* $\tilde{\gamma}$ *such that after sufficient iterations the model satisfies* $|\Phi(w_t, x)| \geq \tilde{\gamma}\|w_t\|$ *almost surely over the data distribution* $(x, y) \sim \mathcal{D}$.

Assumption 5.2.1 implies that the absolute value of the margin is $\tilde{\gamma}$ i.e., $\frac{|\Phi(w_t,x)|}{\|w_t\|} \geq \tilde{\gamma}$ for almost every x after sufficient iterations. This assumption is rather mild, as intuitively it requires that data distribution is not concentrating around the decision boundaries.

For the loss function, we consider the special case of logistic loss $f(t) = \log(1+\exp(-t))$

for simplicity of exposition and more importantly due to its Lipschitz property. The use of Lipschitz property is essential in view of Theorem 5.2.2.

Corollary 5.2.1 (Test error)**.** *Suppose the assumptions of Theorem 5.2.2 hold. Consider the neural network setup under Assumptions 5.1.1 and 5.2.1 and let the loss function f be the logistic loss. Then the test error at step T of normalized GD satisfies the following:*

$$\mathbb{E}[\widetilde{F}_{0-1}(w_{_T})] = O(\frac{1}{T}\mathbb{E}[F(w_{_T})] + \frac{1}{n})$$

The proof of Corollary 5.2.1 is provided in Appendix 5.5.3.4. In the proof, we use that $\|w_t\|$ grows linearly with t as well as Assumption 5.2.1 to deduce $\widetilde{F}_{0-1}(w_{_T}) = O(\frac{\widetilde{F}(w_{_T})}{T})$. Hence, the statement of the corollary follows from Theorem 5.2.2 (i). Finally, we remark the expected test error is decreasing with the rate $1/n$, which is optimal in the realizable setting we consider in this chapter.

5.2.4 Stochastic Normalized GD

In this section we consider a stochastic variant of normalized GD algorithm, Assume z_t to be the batch selected randomly from the dataset at iteration t. The stochastic normalized GD takes the form,

$$w_{t+1} = w_t - \eta_t \nabla F_{z_t}(w_t), \tag{5.7}$$

where $\nabla F_{z_t}(w_t)$ is the gradient of loss at w_t by using the batch of training points z_t at iteration t. We assume η_t to be proportional to $1/F(w_t)$. Our result in this section states that under the following strong growth condition [160, 161], the training loss converges at an exponential rate to the global optimum.

Assumption 5.2.2 (Strong Growth Condition)**.** *The training loss $F : \mathbb{R}^{\widetilde{d}} \to \mathbb{R}$ satisfies the strong growth condition with a parameter ρ,*

$$\mathbb{E}_z[\|\nabla F_z(w)\|^2] \leq \rho \|\nabla F(w)\|^2.$$

Notably, we show in Appendix 5.5.5.1 that the strong growth condition holds for our studied case under the self-bounded and self-lower bounded gradient property.

The next theorem characterizes the rate of decay for the training loss. The proof and numerical experiments are deferred to Appendices 5.5.5.2 and 5.6, respectively.

Theorem 5.2.3 (Convergence of Training Loss)**.** *Consider stochastic normalized GD update rule in Eq.(5.7). Assume F satisfies Assumption 5.2.2 as well as the log-Lipschitzness in the GD path, self-boundedness of the Gradient and the Hessian and the self-lower bounded Gradient properties (Definitions 5.2.1-5.2.4). Let $\eta_t = \eta/F(w_t)$ for all $t \in [T]$ and for any positive constant η satisfying $\eta \leq \frac{\mu^2}{HC\rho h^2}$. Then for the training loss at iteration T the following bound holds:*

$$F(w_{_T}) \leq (1 - \frac{\eta\mu^2}{2})^T F(w_0).$$

5.3 Numerical Experiments

In this section, we demonstrate the empirical performance of normalized GD. Figure 5.1 illustrates the training loss (Left), the test error % (middle), and the weight norm (Right) of GD with normalized GD. The experiments are conducted on a two-layer neural network with $m = 50$ hidden neurons with leaky-ReLU activation function in (5.6) where $\alpha = 0.2$ and $\ell = 1$. The second layer weights are chosen randomly from $a_j \in \{\pm\frac{1}{m}\}$ and kept fixed during training and test time. The first layer weights are initialized from

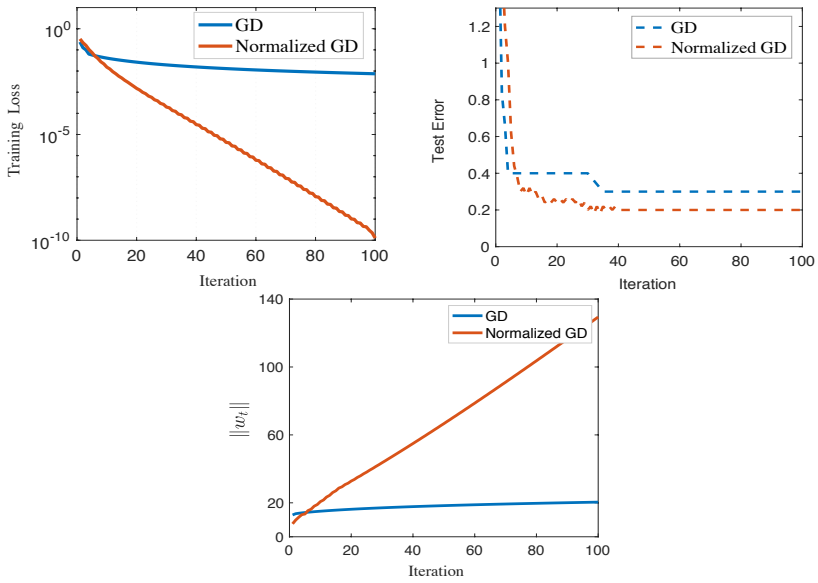


Figure 5.1: Comparison of the training loss, test error (in percentage), and weight norm (i.e., $\|w_t\|$) between gradient descent and normalized gradient descent algorithms. The experiments were conducted on two classes of the MNIST dataset using exponential loss and a two-layer neural network with $m = 50$ hidden neurons. The results demonstrate the performance advantages of normalized gradient descent over traditional gradient descent in terms of both the training loss and test error.

standard Gaussian distribution and then normalized to unit norm. We consider binary classification with exponential loss using digits "0" and "1" from the MNIST dataset ($d = 784$) and we set the sample size to $n = 1000$. The step-size are fine-tuned to $\eta = 30$ and 5 for GD and normalized GD, respectively so that each line represents the best of each algorithm. We highlight the significant speed-up in the convergence of normalized GD compared to standard GD. For the training loss, normalized GD decays exponentially fast to zero while GD converges at a remarkably slower rate. We also highlight that $\|w_t\|$ for normalized GD grows at a rate $\Theta(t)$ while it remains almost constant for GD. In fact this was predicted by Corollary 5.2.1 where in the proof we showed that the weight norm grows linearly with the iteration number. In Figure 5.2, we generate two synthetic dataset according to a realization of a zero-mean Gaussian-mixture model with $n - 40$ and $d = 2$ where the two classes have different covariance matrices (top) and a zero-mean

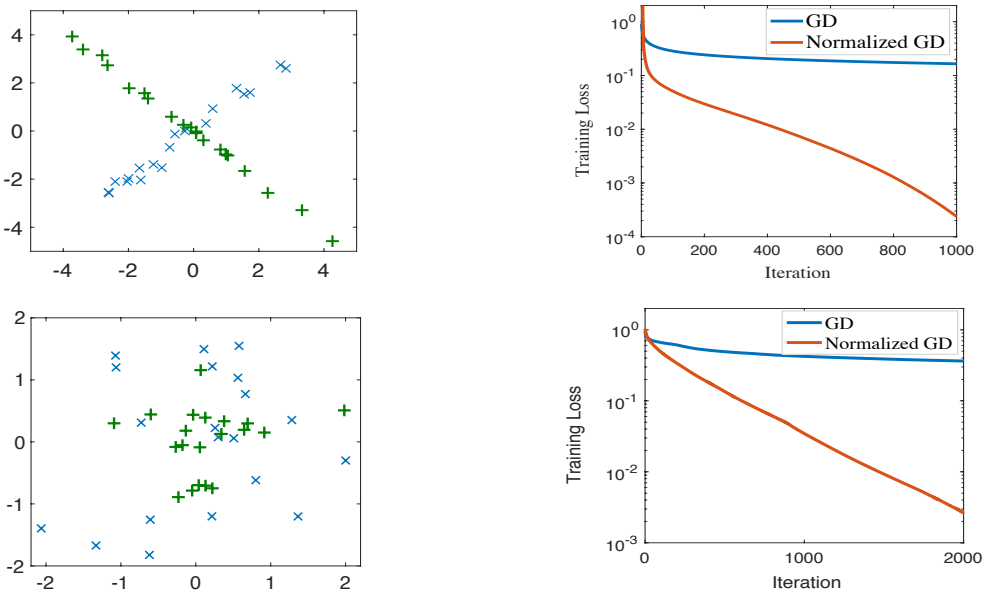


Figure 5.2: The left plot depicts two synthetic datasets, each consisting of $n = 40$ data points. On the right, we present the training loss results of gradient descent and normalized gradient descent algorithms applied to a two-layer neural network with $m = 50$ (top) and 100 (bottom) hidden neurons.

Gaussian-mixture model with $n = 40, d = 5$ (only the first two entires are depicted in the figure) where $\Sigma_1 = \mathbf{I}, \Sigma_2 = \frac{1}{4}\mathbf{I}$ (Bottom). Note that none of the datasets is linearly separable. We consider the same settings as in Figure 5.1 and compared the performance of GD and normalized GD in the right plots. The step-sizes are fine-tuned to $\eta = 80, 350$ and $30, 20$ for GD and normalized GD, respectively. Here again the normalized GD algorithm demonstrates a superior rate in convergence to the final solution.

5.4 Conclusions

We presented the first theoretical evidence for the convergence of normalized gradient methods in non-linear models. While previous results on standard GD for two-layer neural networks trained with logistic/exponential loss proved a rate of $\widetilde{O}(1/t)$ for the training loss, we showed that normalized GD enjoys an exponential rate. We also studied for

the first time, the stability of normalized GD and derived bounds on its generalization performance for convex objectives. We also briefly discussed the stochastic normalized GD algorithm. As future directions, we believe extensions of our results to deep neural networks is interesting. Notably, we expect several of our results to be still true for deep neural networks. Extending the self lower-boundedness property in Lemma 5.2.3 for smooth activation functions is another important direction. Another promising avenue for future research is the derivation of generalization bounds for non-convex objectives by extending the approach used for GD (in [158]) to normalized GD.

5.5 Proofs

5.5.1 Proof of Theorem 5.2.1

Based on the conditions of the theorem we have,

$$\max_{v\in[w_t,w_{t+1}]} F(v) \leq C\, F(w_t),$$
$$\|\nabla^2 F(w)\| \leq H F(w) \quad \text{and} \quad \|\nabla F(w)\| \in [\mu F(w), h F(w)]$$

Then by Taylor's expansion and using the assumptions of the theorem we can deduce,

$$\begin{aligned} F(w_{t+1}) &\leq F(w_t) + \langle \nabla F(w_t), w_{t+1} - w_t \rangle + \frac{1}{2} \max_{v\in[w_t,w_{t+1}]} \|\nabla^2 F(v)\| \cdot \|w_{t+1} - w_t\|^2 \\ &\leq F(w_t) - \eta_t \|\nabla F(w_t)\|^2 + \frac{\eta_t^2}{2} \max_{v\in[w_t,w_{t+1}]} \|\nabla^2 F(v)\| \cdot \|\nabla F(w_t)\|^2 \\ &\leq F(w_t) - \eta_t \|\nabla F(w_t)\|^2 + \frac{\eta_t^2 H}{2} \max_{v\in[w_t,w_{t+1}]} F(v) \cdot \|\nabla F(w_t)\|^2 \\ &\leq F(w_t) - \mu^2 \eta_t (F(w_t))^2 + \frac{\eta_t^2 H C h^2}{2} (F(w_t))^3 \end{aligned}$$

Let $\eta_t = \frac{\eta}{F(w_t)}$,

$$F(w_{t+1}) \leq (1 - \eta\mu^2 + \frac{HCh^2\eta^2}{2}) F(w_t)$$

Then condition on the step-size $\eta \leq \frac{\mu^2}{HCh^2}$, ensures that $1 - \eta\mu^2 + \frac{HCh^2\eta^2}{2} \leq 1 - \frac{\eta\mu^2}{2}$. Thus,

$$F(w_{t+1}) \leq (1 - \frac{\eta\mu^2}{2}) F(w_t).$$

Thus $F(w_{_T}) \leq (1 - \frac{\eta\mu^2}{2})^T F(w_0)$. This completes the proof.

5.5.2 Proofs for Section 5.2.2

5.5.2.1 Proof of Lemma 5.2.1

For a sample point $x \in \mathbb{R}^d$ and two weight vectors $w, w' \in \mathbb{R}^{\widetilde{d}}$, since the activation function satisfies $\sigma' < \ell, \sigma'' < L$, we can deduce that,

$$\begin{aligned}|\Phi(w,x) - \Phi(w',x)| &= |\sum_{j=1}^{m} a_j \sigma(\langle w_j, x\rangle) - a_j \sigma(\langle w'_j, x\rangle)| \\ &\leq \sum_{j=1}^{m} |a_j| \cdot |\sigma(\langle w_j, x\rangle) - \sigma(\langle w'_j, x\rangle)|\end{aligned}$$

By L-smoothness of the activation function and recalling that $\sigma'(\cdot) \leq \ell$ we can write,

$$\begin{aligned}\sigma(\langle w_j, x\rangle) - \sigma(\langle w'_j, x\rangle) &\leq \sigma'(\langle w'_j, x\rangle)\langle w_j - w'_j, x\rangle + \frac{L}{2}|\langle w_j - w'_j, x\rangle|^2 \\ &\leq |\sigma'(\langle w'_j, x\rangle)| \cdot |\langle w_j - w'_j, x\rangle| + \frac{L}{2}|\langle w_j - w'_j, x\rangle|^2 \\ &\leq \ell \|w_j - w'_j\| \|x\| + \frac{L}{2}\|w_j - w'_j\|^2 \|x\|^2 \\ &\leq \ell R \|w_j - w'_j\| + \frac{LR^2}{2}\|w_j - w'_j\|^2.\end{aligned}$$

Since by assumption $|a_j| \leq a$,

$$\begin{aligned}|\Phi(w,x) - \Phi(w',x)| &\leq \sum_{j=1}^{m} |a_j| (\ell R \|w_j - w'_j\| + \frac{LR^2}{2}\|w_j - w'_j\|^2) \\ &\leq aR \sum_{j=1}^{m} (\ell \|w_j - w'_j\| + LR \|w_j - w'_j\|^2).\end{aligned}$$

Hence, for a label $y \in \{\pm 1\}$ we have

$$\begin{aligned} -y\Phi(w,x) + y\Phi(w',x) &\leq |\Phi(w,x) - \Phi(w',x)| \\ &\leq aR \sum_{j=1}^{m} (\ell \|w_j - w'_j\| + LR\|w_j - w'_j\|^2). \end{aligned}$$

Noting the use of exponential loss and by taking $\exp(\cdot)$ of both sides,

$$\begin{aligned} \frac{f(y\Phi(w,x))}{f(y\Phi(w',x))} &= \exp\left(-y\Phi(w,x) + y\Phi(w',x)\right) \\ &\leq \exp\left(aR \sum_{j=1}^{m} (\ell \|w_j - w'_j\| + LR\|w_j - w'_j\|^2)\right) \\ &\leq \exp\left(aR(\sqrt{m}\,\ell \|w - w'\| + LR\|w - w'\|^2)\right) \end{aligned} \tag{5.8}$$

Thus for any two points w, w' it holds,

$$f(y\Phi(w,x)) \leq f(y\Phi(w',x)) \cdot \exp\left(aR(\sqrt{m}\,\ell \|w - w'\| + LR\|w - w'\|^2)\right) \tag{5.9}$$

Therefore, for a sample loss with $(x_i, y_i) \in \mathbb{R}^d \times \{\pm 1\}$ and $v \in [w_t, w_{t+1}]$ i.e, $v =$

$w_t + \lambda(w_{t+1} - w_t)$ for some $\lambda \in [0,1]$, we have,

$$
\begin{aligned}
f(y_i\Phi(v,x_i)) &= f(y_i\Phi(w_t + \lambda(w_{t+1}-w_t),x_i)) \\
&\le f(y_i\Phi(w_t.x_i)) \cdot \exp\left(aR(\sqrt{m}\,\ell\|v-w_t\| + LR\|v-w_t\|^2)\right) \\
&= f(y_i\Phi(w_t.x_i)) \cdot \exp\left(aR(\sqrt{m}\,\ell\lambda\|w_{t+1}-w_t\| + LR\lambda^2\|w_{t+1}-w_t\|^2)\right) \\
&= f(y_i\Phi(w_t.x_i)) \cdot \exp\left(aR(\sqrt{m}\,\ell\lambda\eta_t\|\nabla F(w_t)\| + LR\lambda^2\eta_t^2\|\nabla F(w_t)\|^2)\right) \\
&= f(y_i\Phi(w_t.x_i)) \cdot \exp\left(aR(\sqrt{m}\,\ell\lambda\frac{\eta}{F(w_t)}\|\nabla F(w_t)\| + LR\lambda^2(\frac{\eta}{F(w_t)})^2\|\nabla F(w_t)\|^2)\right) \\
&\le f(y_i\Phi(w_t.x_i)) \cdot \exp\left(\sqrt{m}\,aR\,\ell\lambda h\eta + aLR^2\lambda^2h^2\eta^2\right),
\end{aligned}
$$

where for the last step we used the assumption that $\eta_t = \frac{\eta}{F(w_t)}$ for any constant $\eta \le \frac{\mu^2}{HCh^2}$ and the assumption that $\|\nabla F(w)\| \le hF(w)$. This proves the inequality (5.4) in the statement of the lemma.

To derive (5.5), note that since $\lambda \le 1$,

$$
\begin{aligned}
\max_{v\in[w_t,w_{t+1}]} f(y_i\Phi(v,x_i)) &= \max_{\lambda\in[0,1]} f(y_i\Phi(w_t+\lambda(w_{t+1}-w_t),x_i)) \\
&\le f(y_i\Phi(w_t.x_i)) \cdot \exp\left(\sqrt{m}\,aR\ell\lambda h\eta + aLR^2\lambda^2h^2\eta^2\right)
\end{aligned}
$$

Noting that this holds for all $i \in [n]$, we deduce that the following holds for the training loss:

$$
\begin{aligned}
\max_{v\in[w_t,w_{t+1}]} F(v) &\le \frac{1}{n}\sum_{i=1}^{n} \max_{v\in[w_t,w_{t+1}]} f(y_i\Phi(v,x_i)) \\
&\le F(w_t) \cdot \exp\left(\sqrt{m}\,aR\ell\lambda h\eta + aLR^2\lambda^2h^2\eta^2\right).
\end{aligned}
$$

Recalling that $a \le \frac{1}{m}$ and choosing $C = \exp(\frac{R\ell\lambda h\eta}{\sqrt{m}} + \frac{LR^2\lambda^2h^2\eta^2}{m})$ leads to (5.5) and completes

the proof.

5.5.2.2 Proof of Lemma 5.2.2

For the lower bound on the gradient norm, we can write

$$\|\nabla F(w)\| = \frac{1}{n}\|\sum_{i=1}^{n} f(y_i\Phi(w,x_i))y_i\nabla_1\Phi(w,x_i)\|$$

where $\forall w \in \mathbb{R}^{\widetilde{d}}, x \in \mathbb{R}^d$ the gradient of Φ with respect to the first argument satisfies the following:

$$\nabla_1\Phi(w,x) = [xa_1\sigma'(\langle w_1,x\rangle); xa_2\sigma'(\langle w_2,x\rangle); \cdots; xa_m\sigma'(\langle w_m,x\rangle)] \in \mathbb{R}^{\widetilde{d}}.$$

Equivalently, we can write

$$\|\nabla F(w)\| = \sup_{v\in\mathbb{R}^{\widetilde{d}},\|v\|_2=1} \left\langle \frac{1}{n}\sum_{i=1}^{n} f(y_i\Phi(w,x_i))y_i\nabla_1\Phi(w,x_i), v\right\rangle$$

Choose the candidate vector v as follows

$$\bar{v} = [a_1w^\star; a_2w^\star; \cdots; a_mw^\star] \in \mathbb{R}^{\widetilde{d}} \qquad v = \bar{v}/\|\bar{v}\|,$$

where $w^\star$ is the max-margin separator that satisfies for all $i \in [n]$, $\frac{y_i\langle x_i,w^\star\rangle}{\|w^\star\|} \geq \gamma$, where γ denotes the margin. We have $\|\bar{v}\| = \|\tilde{a}\|\|w^*\|$ where $\tilde{a} \in \mathbb{R}^m$ is the concatenation of second

layer weights a_j. Recalling $\sigma'(\cdot) \geq \alpha$,

$$\begin{aligned}
\|\nabla F(w)\| &\geq \frac{1}{\|\tilde{a}\|\|w^*\|}\frac{1}{n}\sum_{i=1}^{n} f(y_i\Phi(w,x_i)) \cdot y_i\langle x_i, w^\star\rangle\Big(\sum_{j=1}^{m} a_j^2\sigma'(\langle w_j, x_i\rangle)\Big) \\
&\geq \|\tilde{a}\|\frac{\alpha}{n}\sum_{i=1}^{n} f(y_i\Phi(w,x_i)) \cdot \frac{y_i\langle x_i, w^\star\rangle}{\|w^*\|} \\
&\geq \|\tilde{a}\|\alpha \cdot (\min_{j\in[n]} \frac{y_j\langle x_j, w^\star\rangle}{\|w^*\|}) \cdot \frac{1}{n}\sum_{i=1}^{n} f(y_i\Phi(w,x_i)) \\
&\geq \|\tilde{a}\|\alpha\gamma \cdot F(w).
\end{aligned}$$

This completes the proof of the lemma.

5.5.2.3 Proof of Lemma 5.2.3

Recall that,

$$\|\nabla F(w)\|_2 = \sup_{v\in\mathbb{R}^{\tilde{d}}, \|v\|_2=1} \Big\langle \frac{1}{n}\sum_{i=1}^{n} f(y_i\Phi(w,x_i))y_i\nabla_1\Phi(w,x_i), v\Big\rangle$$

where,

$$\nabla_1\Phi(w,x) = [xa_1\sigma'(\langle w_1, x\rangle); xa_2\sigma'(\langle w_2, x\rangle); \cdots; xa_m\sigma'(\langle w_m, x\rangle)] \in \mathbb{R}^{\tilde{d}}$$

Also, assume $w \in \mathbb{R}^{\tilde{d}}$ separates the dataset with margin γ, i.e., for all $i \in [n]$

$$\frac{y_i\Phi(w,x_i)}{\|w\|} \geq \gamma.$$

choose

$$v = \frac{w}{\|w\|}$$

then

$$\begin{aligned}\|\nabla F(w)\| &\geq \Big\langle \frac{1}{n}\sum_{i=1}^{n} f(y_i\Phi(w,x_i))y_i\nabla_1\Phi(w,x_i), v\Big\rangle \\ &= \frac{1}{\|w\|}\frac{1}{n}\sum_{i=1}^{n} f(y_i\Phi(w,x_i))\cdot y_i\sum_{j=1}^{m} a_j\langle w_j,x_i\rangle\sigma'(\langle w_j,x_i\rangle)\end{aligned}$$

Based on the activation function,

$$\langle w_j,x_i\rangle\sigma'(\langle w_j,x\rangle) = \begin{cases}\ell\langle w_j,x_i\rangle & \langle w_j,x_i\rangle \geq 0 \\ \alpha\langle w_j,x_i\rangle & \langle w_j,x_i\rangle < 0.\end{cases}$$

which is equal to $\sigma(\langle w_j,x_i\rangle)$.

Thus,

$$\begin{aligned}\|\nabla F(w)\| &\geq \frac{1}{\|w\|}\frac{1}{n}\sum_{i=1}^{n} f(y_i\Phi(w,x_i))\cdot y_i\sum_{j=1}^{m} a_j\sigma(\langle w_j,x_i\rangle) \\ &= \frac{1}{n}\sum_{i=1}^{n} f(y_i\Phi(w,x_i))\cdot\frac{y_i\Phi(w,x_i)}{\|w\|} \\ &\geq F(w)\cdot\gamma\end{aligned}$$

This completes the proof.

5.5.2.4 Proof of Lemma 5.2.4

Recall that,

$$F(w) := \frac{1}{n}\sum_{i=1}^{n} f(y_i \Phi(w, x_i)),$$
$$\Phi(w, x) := \sum_{j=1}^{m} a_j \sigma(\langle w_j, x\rangle)$$

where $x_i \in \mathbb{R}^d, w_j \in \mathbb{R}^d, a_j \in \mathbb{R}, w = [w_1 w_2 ... w_m] \in \mathbb{R}^{\widetilde{d}}$. Then noting the exponential nature of the loss function we can write,

$$\begin{aligned}\|\nabla F(w)\| &= \frac{1}{n}\Big\| \sum_{i=1}^{n} f'(y_i \Phi(w, x_i)) y_i \nabla_1 \Phi(w, x_i)\Big\| \\ &\leq \frac{1}{n}\sum_{i=1}^{n} f(y_i \Phi(w, x_i)) \|\nabla_1 \Phi(w, x_i)\|.\end{aligned}$$

Noting that $\sigma'(\cdot) \leq \ell$,

$$\|\nabla_1 \Phi(w, x)\|^2 = \sum_{j=1}^{m}\sum_{i=1}^{d} (a_j x(i) \sigma'(\langle w_j, x\rangle))^2 \leq \frac{\ell^2 \|x\|^2}{m}$$

Thus $\forall w \in \mathbb{R}^{\widetilde{d}}$ and $h = \frac{\ell R}{\sqrt{m}}$

$$\|\nabla F(w)\| \leq h F(w).$$

For the Hessian, note that since $|\sigma''(\cdot)| \leq L$ and

$$\nabla_1^2 \Phi(w, x) = \frac{1}{m} \operatorname{diag}\left(a_1 \sigma''(\langle w_1, x\rangle) x x^T, \ldots, a_m \sigma''(\langle w_m, x\rangle) x x^T\right), \tag{5.10}$$

then the operator norm of model's Hessian satisfies,

$$\|\nabla_1^2\Phi(w,x)\|^2 \leq L^2R^4a^2.$$

Thus, for the objective's Hessian $\nabla^2 F(w) \in \mathbb{R}^{\tilde{d}\times\tilde{d}}$, we have

$$\begin{aligned}
\|\nabla^2 F(w)\| &= \|\frac{1}{n}\sum_{i=1}^{n} f(y_i\Phi(w,x_i))y_i\nabla_1^2\Phi(w,x_i) + f(y_i\Phi(w,x_i))\nabla_1\Phi(w,x_i)\nabla_1\Phi(w,x_i)^\top\| \\
&\leq \frac{1}{n}\sum_{i=1}^{n} f(y_i\Phi(w,x_i))(\|\nabla_1^2\Phi(w,x_i)\|+\|\nabla_1\Phi(w,x_i)\nabla_1\Phi(w,x_i)^\top\|) \\
&= \frac{1}{n}\sum_{i=1}^{n} f(y_i\Phi(w,x_i))(\|\nabla_1^2\Phi(w,x_i)\|+\|\nabla_1\Phi(w,x_i)\|_2^2) \\
&\leq (\frac{LR^2}{m^2} + \frac{\ell^2R^2}{m})F(w).
\end{aligned}$$

Denoting $H := \frac{LR^2}{m^2} + \frac{\ell^2R^2}{m}$, we have $\|\nabla^2 F(w)\| \leq HF(w)$. This concludes the proof.

5.5.3 Proofs for Section 5.2.3

5.5.3.1 Proof of Lemma 5.2.5

Define $G(w,v) : \mathbb{R}^{\tilde{d}} \times \mathbb{R}^{\tilde{d}} \to \mathbb{R}$ as follows,

$$G(w,v) := F(w) - \langle\nabla F(v), w\rangle$$

Note that

$$\|\nabla_1^2 G(w,v)\| = \|\nabla^2 F(w)\| \leq hF(w).$$

Thus by Taylor's expansion of G around its first argument and noting the self-boundedness of Hessian and the convexity of F, we have for all $w, \tilde{w} \in \mathbb{R}^d$,

$$\begin{aligned} G(w,v) &\leq G(\tilde{w}) + \langle \nabla_1 G(\tilde{w},v), w-\tilde{w}\rangle + \frac{1}{2}\max_{v\in[w,\tilde{w}]} \|\nabla^2 F(v)\| \|w-\tilde{w}\|^2 \\ &\leq G(\tilde{w}) + \langle \nabla_1 G(\tilde{w},v), w-\tilde{w}\rangle + \frac{h}{2}\max_{v\in[w,\tilde{w}]} F(v) \|w-\tilde{w}\|^2 \\ &\leq G(\tilde{w}) + \langle \nabla_1 G(\tilde{w},v), w-\tilde{w}\rangle + \frac{h}{2}\max(F(w),F(\tilde{w}))\|w-\tilde{w}\|^2. \end{aligned}$$

Taking minimum of both sides

$$\begin{aligned} \min_{w\in\mathbb{R}^d} G(w,v) &\leq \min_{w\in\mathbb{R}^d} G(\tilde{w},v) + \langle \nabla_1 G(\tilde{w},v), w-\tilde{w}\rangle + \max(F(w),F(\tilde{w}))\frac{h\|w-\tilde{w}\|^2}{2} \\ &\leq G(\tilde{w},v) - r\|\nabla_1 G(\tilde{w},v)\|^2 + \max(F(\tilde{w}-r\nabla_1 G(\tilde{w},v)),F(\tilde{w}))\frac{hr^2\|\nabla_1 G(\tilde{w},v)\|^2}{2} \\ &\leq G(\tilde{w},v) - (r-2r^2 hF(\tilde{w}))\|\nabla_1 G(\tilde{w},v)\|^2. \end{aligned} \tag{5.11}$$

In the second step, we chose $w = \tilde{w} - r\nabla_1 G(\tilde{w},v)$ for a positive constant r. Moreover, for the last step we used the following inequality (which we will prove hereafter) that holds under $r \leq \frac{1}{h(\max(F(v),F(\tilde{w})))}$,

$$F(\tilde{w} - r\nabla_1 G(\tilde{w},v)) \leq 4F(\tilde{w}). \tag{5.12}$$

The inequality in (5.12) can be proved according to the following steps. First consider the convexity of F and the self-boundedness of Hessian to derive the Taylor's expansion

of F in the following style:

$$\begin{aligned} F(\tilde{w} - r\nabla_1 G(\tilde{w}, v)) &= F(\tilde{w} - r\nabla F(\tilde{w}) + r\nabla F(v)) \\ &\leq F(\tilde{w} - r\nabla F(\tilde{w})) + r\langle \nabla F(\tilde{w} - r\nabla F(\tilde{w})), \nabla F(v)\rangle \\ &+ \frac{hM(w,v)}{2} r^2 \|\nabla F(v)\|^2, \end{aligned} \tag{5.13}$$

where we define,

$$M(w, v) := \max(F(\tilde{w} - r\nabla F(\tilde{w}) + r\nabla F(v)), F(\tilde{w} - r\nabla F(\tilde{w}))). \tag{5.14}$$

We have that if $r \leq 1/(hF(\tilde{w}))$, then

$$F(\tilde{w} - r\nabla F(\tilde{w})) \leq F(\tilde{w})$$

Now, suppose that the assumption in (5.12) is false and on the contrary $F(\tilde{w} - r\nabla_1 G(\tilde{w}, v)) > 4F(\tilde{w})$, then

$$M(w, v) = F(\tilde{w} - r\nabla_1 G(\tilde{w}, v)).$$

By using Cauchy-Shwarz inequality in (5.13) together with the self-boundedness properties

we deduce that

$$\begin{aligned}
&F(\tilde{w} - r\nabla_1 G(\tilde{w}, v)) \\
&\leq F(\tilde{w}) + r\|\nabla F(\tilde{w} - r\nabla F(\tilde{w}))\|\|\nabla F(v)\| + \frac{hr^2}{2}\|\nabla F(v)\|^2 F(\tilde{w} - r\nabla_1 G(\tilde{w}, v)) \\
&\leq F(\tilde{w}) + rh^2 F(\tilde{w} - r\nabla F(\tilde{w}))F(v) + \frac{r^2h^3}{2}F^2(v)F(\tilde{w} - r\nabla_1 G(\tilde{w}, v)) \\
&\leq F(\tilde{w}) + rh^2 F(\tilde{w})F(v) + \frac{r^2h^3}{2}F^2(v)F(\tilde{w} - r\nabla_1 G(\tilde{w}, v)) \\
&\leq 2F(\tilde{w}) + \frac{1}{2}F(\tilde{w} - r\nabla_1 G(\tilde{w}, v)),
\end{aligned}$$

The last step is derived by the condition on r and the fact that $h \leq 1$. The last inequality leads to contradiction. This proves (5.12). Thus, continuing from (5.11) and assuming $r \leq \frac{1}{2hF(\tilde{w})}$

$$F(v) - \langle \nabla F(v), v\rangle \leq F(\tilde{w}) - \langle \nabla F(v), \tilde{w}\rangle - \frac{r}{2}\|\nabla F(\tilde{w}) - \nabla F(v)\|^2$$

Exchanging v and $\tilde{w}$ in the above and noting that under our assumptions it holds that $r \leq \frac{1}{2hF(v)}$, we can write

$$F(\tilde{w}) - \langle \nabla F(\tilde{w}), \tilde{w}\rangle \leq F(v) - \langle \nabla F(\tilde{w}), v\rangle - \frac{r}{2}\|\nabla F(\tilde{w}) - \nabla F(v)\|^2$$

Combining these two together, we end up with the following inequality:

$$r\|\nabla F(\tilde{w}) - \nabla F(v)\| \leq \langle \nabla F(v) - \nabla F(\tilde{w}), v - \tilde{w}\rangle.$$

Therefore $\forall w, v \in \mathbb{R}^d$ if $\eta \leq 2r$ (which the RHS itself is smaller than $\frac{1}{h \max(F(v),F(w))}$),

$$\begin{aligned}\|w - \eta\nabla F(w) - (v - \eta\nabla F(v))\|^2 &= \|v - w\|^2 - 2\eta\langle\nabla F(v) - \nabla F(w), v - w\rangle \\ &\quad + \eta^2\|\nabla F(v) - \nabla F(w)\|^2 \\ &\leq \|v - w\|^2 - \left(2\eta r - \eta^2\right)\|\nabla F(v) - \nabla F(w)\|^2 \\ &\leq \|v - w\|^2.\end{aligned}$$

This completes the proof.

5.5.3.2 Proof of Theorem 5.2.2

Fix $i \in [n]$ and let $w_t^{\neg i} \in \mathbb{R}^d$ be the vector obtained at the step t of normalized GD with the following iterations,

$$w_{k+1}^{\neg i} = w_k^{\neg i} - \eta_k \nabla F^{\neg i}(w_k^{\neg i}),$$

where η_k denotes the step-size at step k which satisfies $\eta_k \leq \frac{1}{hF^{\neg i}(w_k^{\neg i})}$ for all $k \in [t-1]$. Also, we define the leave-one-out training loss for $i \in [n]$ as follows:

$$F^{\neg i}(w) := \frac{1}{n}\sum_{\substack{j=1 \\ j \neq i}}^{n} f(w, z_j).$$

In words, $w_t^{\neg i}$ is the output of normalized GD at iteration t when the ith sample is left out while the step-size is chosen independent of the i th sample. Thus, we can write

$$\begin{aligned}\mathbb{E}[\widetilde{F}(w_t) - F(w_t)] &= \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}[f(w_t, z) - f(w_t^{\neg i}, z)] + \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}[f(w_t^{\neg i}, z_i) - f(w_t, z_i)] \\ &\leq \frac{2G}{n}\sum_{i=1}^{n}\mathbb{E}[\|w_t - w_t^{\neg i}\|] \end{aligned} \tag{5.15}$$

Since the loss function is non-negative, $F^{\neg i}(w_t) \leq F(w_t)$ for all i. Thus, by assumption of the theorem the step-size satisfies $\eta_t \leq \frac{1}{h\delta F(w_t)} \leq \frac{1}{h\delta F^{\neg i}(w_t)}$, $\forall i \in [n]$. By the definition of δ, this choice of step-size guarantees that $\eta_t \leq \frac{1}{hF^{\neg i}(w_t^{\neg i})}$. Recalling that $\delta \geq 1$, we deduce that $\eta_t \leq \frac{1}{h\max(F^{\neg i}(w_t), F^{\neg i}(w_t^{\neg i}))}$, which allows us to apply Lemma 5.2.5. In particular, by unrolling w_{t+1} and $w_{t+1}^{\neg i}$, and using our result from Lemma 5.2.5 on the non-expansiveness of normalized GD we can write,

$$\begin{aligned}\left\|w_{t+1} - w_{t+1}^{\neg i}\right\| &= \left\|w_t - \frac{1}{n}\eta_t\sum_{j=1}^{n}\nabla f(w_t, z_j) - w_t^{\neg i} + \frac{1}{n}\eta_t\sum_{j\neq i}^{n}\nabla f(w_t^{\neg i}, z_j)\right\| \\ &= \left\|w_t - \eta_t\nabla F^{\neg i}(w_t) - \frac{1}{n}\eta_t\nabla f(w_t, z_i) - w_t^{\neg i} + \eta_t\nabla F^{\neg i}(w_t^{\neg i})\right\| \\ &\leq \left\|w_t - \eta_t\nabla F^{\neg i}(w_t) - w_t^{\neg i} + \eta_t\nabla F^{\neg i}(w_t^{\neg i})\right\| + \frac{1}{n}\eta_t\|\nabla f(w_t, z_i)\| \\ &\leq \left\|w_t - w_t^{\neg i}\right\| + \frac{1}{n}\eta_t\left\|\nabla f\left(w_t, z_i\right)\right\| \\ &\leq \left\|w_t - w_t^{\neg i}\right\| + \frac{1}{n}h\eta_t f(w_t, z_i). \end{aligned} \tag{5.16}$$

This result holds for all $i \in [n]$. By averaging over all training samples,

$$\frac{1}{n}\sum_{i=1}^{n}\|w_{t+1} - w_{t+1}^{\neg i}\| \leq \frac{1}{n}\sum_{i=1}^{n}\|w_t - w_t^i\| + \frac{h}{n}\eta_t F(w_t).$$

Thus, by telescoping sum over t, for the last iteration we have,

$$\frac{1}{n}\sum_{i=1}^{n}\|w_{_T}-w_{_T}^{\neg i}\|\leq \frac{h}{n}\sum_{t=0}^{T-1}\eta_t F(w_t)$$

Next, we recall (5.15) which allows us to bound the generalization gap,

$$\begin{aligned}\mathbb{E}[\widetilde{F}(w_{_T})-F(w_{_T})] &\leq \frac{2Gh}{n}\sum_{t=0}^{T-1}\eta_t F(w_t)\\ &\leq \frac{2GT}{n}.\end{aligned}$$

This completes the poof for L- Lipschitz losses.

For $\tilde{L}$-smooth losses, the following relation holds between test and train loss and the leave-one-out distance (e.g., see [152, Lemma 7], [14, Theorem2]):

$$\mathbb{E}[\widetilde{F}(w)]\leq 4\mathbb{E}[F(w)]+\frac{3\tilde{L}^2}{n}\sum_{i=1}^{n}\mathbb{E}[\|w-w^{\neg i}\|^2]. \tag{5.17}$$

Note the dependence on $\|w-w^{\neg i}\|^2$. Recalling (5.16), we had

$$\left\|w_{t+1}-w_{t+1}^{\neg i}\right\|\leq\left\|w_t-w_t^{\neg i}\right\|+\frac{1}{n}\eta_t\, h\, f(w_t,z_i)$$

By telescoping summation,

$$\|w_{_T}-w_{_T}^{\neg i}\|\leq \frac{h}{n}\sum_{t=0}^{T-1}\eta_t f(w_t.z_i)$$

this gives the following upper bound on the averaged squared norm,

$$\begin{aligned}\frac{1}{n}\sum_{i=1}^{n}\|w_{_T}-w_{_T}^{\neg i}\|^2 &\leq \frac{h^2}{n^3}\sum_{i=1}^{n}(\sum_{t=1}^{T-1}\eta_t f(w_t.z_i))^2\\ &\leq \frac{h^2}{n^3}(\sum_{i=1}^{n}\sum_{t=0}^{T-1}\eta_t f(w_t.z_i))^2\\ &= \frac{h^2}{n}(\sum_{t=0}^{T-1}\frac{\eta_t}{n}\sum_{i=1}^{n} f(w_t.z_i))^2\\ &= \frac{h^2}{n}(\sum_{t=0}^{T-1}\eta_t F(w_t))^2.\end{aligned}$$

Hence, replacing these back in (5.17),

$$\begin{aligned}\mathbb{E}[\widetilde{F}(w_{_T})] &\leq 4\mathbb{E}[F(w_{_T})]+\frac{3\tilde{L}^2h^2}{n}(\sum_{t=0}^{T-1}\eta_t F(w_t))^2\\ &\leq 4\mathbb{E}[F(w_{_T})]+\frac{3\tilde{L}^2}{n}T.\end{aligned}$$

This gives the desired result for $\tilde{L}$-smooth losses in part (ii) of the lemma and completes the proof.

5.5.3.3 On δ in Theorem 5.2.2

Lemma 5.5.1. *Assume the iterates of normalized GD with $\eta \leq 1/h$, zero initialization (w.l.o.g) and $m = \beta T^2$ hidden neurons for any constant $\beta > 0$. Then δ in the statement of Theorem 5.2.2 is satisfied with $\delta = \exp(\frac{2R\ell}{\sqrt{\beta}} + \frac{4LR^2}{\beta})$.*

Proof: By the log-Lipschitzness property in (5.9) and recalling $a = 1/m$,

$$F^{\neg i}(w_T^{\neg i}) \le F^{\neg i}(w_T) \cdot \exp\Big(\frac{R\ell}{\sqrt{m}}\|w_T^{\neg i} - w_T\| + \frac{LR^2}{m}\|w_T^{\neg i} - w_T\|^2\Big)$$
$$\le F^{\neg i}(w_T) \cdot \exp\Big(\frac{R\ell}{\sqrt{m}}(\|w_T^{\neg i}\| + \|w_T\|) + \frac{2LR^2}{m}(\|w_T^{\neg i}\|^2 + \|w_T\|^2)\Big). \tag{5.18}$$

Now we note that the weight-norm can be upper bounded as following:

$$\begin{aligned}\|w_T\| &= \Big\|w_{T-1} - \frac{\eta}{F(w_{T-1})}\nabla F(w_{T-1})\Big\| \\ &= \Big\|w_0 - \eta\sum_{t=0}^{T-1}\frac{\nabla F(w_t)}{F(w_t)}\Big\| \\ &\le \eta\sum_{t=0}^{T-1}\Big\|\frac{\nabla F(w_t)}{F(w_t)}\Big\| \\ &\le \eta h T.\end{aligned}$$

Similarly, we can show that $\|w_T^{\neg i}\| \le \eta h T$. Therefore by $m = \beta T^2$ and (5.18),

$$\begin{aligned}F^{\neg i}(w_T^{\neg i}) &\le F^{\neg i}(w_T) \cdot \exp\Big(\frac{R\ell}{\sqrt{m}}(\|w_T^{\neg i}\| + \|w_T\|) + \frac{2LR^2}{m}(\|w_T^{\neg i}\|^2 + \|w_T\|^2)\Big) \\ &\le F^{\neg i}(w_T) \cdot \exp\Big(\frac{2R\ell}{\sqrt{m}}(\eta h T) + \frac{4LR^2}{m}\eta^2 h^2 T^2\Big) \\ &\le F^{\neg i}(w_T) \cdot \exp\Big(\frac{2R\ell}{\sqrt{\beta}} + \frac{4LR^2}{\beta}\Big),\end{aligned}$$

where the last step follows by $\eta h \le 1$ as per assumptions on the step-size. This completes the proof.

5.5.3.4 Proof of Corollary 5.2.1

First, note that if $F(w) < \delta$, then $\|w\| \geq \frac{1}{\ell R}(\log(\frac{1}{\delta}) - \sigma_0)$, where $\sigma_0 = |\sigma(0)|$, since if the lower-bound on $\|w\|$ is incorrect then,

$$\begin{aligned} F(w) &= \frac{1}{n}\sum_{i=1}^{n} \exp(-y_i \Phi(w, x_i)) \\ &\geq \frac{1}{n}\sum_{i=1}^{n} \exp(-\|w\|\|x_i\| - \sigma_0) \\ &\geq \frac{1}{n}\sum_{i=1}^{n} \exp(-\log(1/\delta)) \\ &= \delta. \end{aligned}$$

where we used,

$$\begin{aligned} y\Phi(w, x) &= \sum_{j=1}^{m} y a_j \sigma(\langle w_j, x\rangle) \\ &\leq \sum_{j=1}^{m} |a_j| \cdot |\sigma(\langle w_j, x\rangle)| \\ &\leq \sum_{j=1}^{m} |a_j| (\sigma_0 + \ell |\langle w_j, x\rangle|) \\ &\leq \sigma_0 ||\tilde{a}||_1 + \ell \|x\|_2 \sum_{j=1}^{m} |a_j| \cdot \|w_j\| \\ &\leq \sigma_0 \|\tilde{a}\|_1 + \ell \|x\|_2 \|\tilde{a}\|_2 \|w\|_2 \end{aligned}$$

This is true due to ℓ-Lipschitz activation and our assumption that $\|\tilde{a}\|_1 \leq m\|\tilde{a}\|_\infty = 1$, where $\tilde{a} \in \mathbb{R}^m$ is the concatenation of second layer weights.

Now, note that due to the convergence of training loss there exists a $\tau > 0$ such that

at iteration t the following holds:

$$F(w_t) \leq (1-\tau)^t F(w_0).$$

Hence the weight's norm at iteration t satisfies,

$$\|w_t\| \geq \frac{t}{R}\log(\frac{1}{1-\tau}) - \frac{\sigma_0}{R} = \Theta(t). \tag{5.19}$$

For the test error, by defining $\mathcal{F}$ to be the set of data points labeled incorrectly by $\Phi(w_t, \cdot)$, we can write

$$\begin{aligned}
\mathbb{E}_{(x,y)\sim\mathcal{D}}[f(y\Phi(w_t,x))] &= \lim_{n\to\infty}\frac{1}{n}\sum_{i=1}^{n} f(y_i\Phi(w_t,x_i)) \\
&\geq \lim_{n\to\infty}\frac{1}{n}\sum_{i\in\mathcal{F}} f(y_i\Phi(w_t,x_i)) \\
&= \lim_{n\to\infty}\frac{1}{n}\sum_{i\in\mathcal{F}} f(-|\Phi(w_t,x_i)|) \\
&= \lim_{n\to\infty}\frac{1}{n}\sum_{i\in\mathcal{F}} \log(1+\exp(|\Phi(w_t,x_i)|)) \\
&\geq \frac{1}{3}\|w_t\|\cdot\lim_{n\to\infty}\frac{1}{n}\sum_{i\in\mathcal{F}}\frac{|\Phi(w_t,x_i)|}{\|w_t\|} \\
&\geq \frac{1}{3}\gamma\|w_t\|\mathbb{E}_{(x,y)\sim\mathcal{D}}[\mathbb{I}(\text{SIGN}(\Phi(w_t,x))\neq y)] \\
&= \Theta(t)\mathbb{E}_{(x,y)\sim\mathcal{D}}[\mathbb{I}(\text{SIGN}(\Phi(w_t,x))\neq y)]
\end{aligned}$$

Where we used the fact that $\log(1+\exp(t)) \geq \frac{1}{3}t$ and the one to the last line inequality is due to Assumption 5.2.1 i.e., $\frac{|\Phi(w_t,x_i)|}{\|w_t\|} \geq \gamma$ with high probability over $(x_i, y_i) \overset{\text{iid}}{\sim} \mathcal{D}$. Hence the test error satisfies,

$$\mathbb{E}[\mathbb{I}(y \neq \text{SIGN}(\Phi(w_t,x)))] = O(\frac{F(w_t)}{t}).$$

This together with the test loss bound in Theorem 5.2.2 yields the statement of the corollary and completes the proof.

5.5.4 Normalized Gradient Flow

Proposition 5.5.1 (Normalized GD in continuous time)**.** *Let the loss function F satisfy self-lower boundedness of the gradient with parameter μ (Definition 5.2.2) and the self-bounded gradient property with parameter h (Definition 5.2.3). Consider normalized gradient descent with the Gradient flow differential equation given by $\frac{d}{dt}w_t = -\nabla F(w_t)/F(w_t)$. Then the training loss at time T satisfies*

$$F(w_0) \cdot \exp(-h^2 T) \leq F(w_{_T}) \leq F(w_0) \cdot \exp(-\mu^2 T).$$

Proof: Based on the assumptions, we have

$$\dot{w}_t := \frac{d}{dt} w_t = -\frac{\nabla F(w_t)}{F(w_t)}.$$

Then,

$$\frac{d}{dt} F(w_t) = \nabla F(w_t)^\top \dot{w}_t = -\frac{\|\nabla F(w_t)\|^2}{F(w_t)}$$

By self-lower bounded property we have $\frac{d}{dt}F(w_t) \leq -\mu^2 F(w_t)$. Thus,

$$\frac{d}{dt} \log(F(w_t)) = \frac{\frac{d}{dt}F(w_t)}{F(w_t)} \leq -\mu^2.$$

By integrating from $t = 0$ to $t = T$ one can deduce that,

$$\log(F(w_{_T})) - \log(F(w_0)) \leq -\mu^2 T.$$

This leads to the desired upper-bound for $F(w_{_T})$. A similar approach by using the self-bounded gradient property leads to the lower bound. This concludes the proof.

5.5.5 Proofs for Section 5.2.4

5.5.5.1 On the Strong Growth Condition

Proposition 5.5.2. *Under the self-bounded gradient property (Definitions 5.2.2-5.2.3) there exists a ρ such that the strong growth condition is satisfied i.e.,*

$$\mathbb{E}_z[\|\nabla F_z(w)\|^2] \leq \rho \|\nabla F(w)\|^2.$$

Proof: By the self-bounded gradient property and noting the non-negativity of f we have,

$$\begin{aligned}
\mathbb{E}_z[\|\nabla F_z(w)\|^2] &\leq h^2 \mathbb{E}[(F_z(w))^2] \\
&\leq h^2 n (F(w))^2 \\
&\leq \frac{h^2 n}{\mu^2} \|\nabla F(w)\|^2.
\end{aligned}$$

This completes the proof.

5.5.5.2 Proof of Theorem 5.2.3

Following the proof of Theorem 5.2.1 and noting the log-Lipschitzness and the self-bounded Hessian property we derive that,

$$\begin{aligned} F(w_{t+1}) &\leq F(w_t) + \langle \nabla F(w_t), w_{t+1} - w_t \rangle + \frac{1}{2} HC\, F(w_t)\, \|w_{t+1} - w_t\|^2 \\ &= F(w_t) - \eta_t \langle \nabla F(w_t), \nabla F_{z_t}(w_t) \rangle + \frac{1}{2} HC \eta_t^2 F(w_t) \|\nabla F_{z_t}(w_t)\|^2 \end{aligned} \tag{5.20}$$

Taking expectation with respect to z_t and using self-boundedness property yields,

$$\begin{aligned} \mathbb{E}_{z_t}[F(w_{t+1})] &\leq F(w_t) - \eta_t \|\nabla F(w_t)\|^2 + \frac{1}{2} HC \eta_t^2 F(w_t) \mathbb{E}_{z_t}[\|\nabla F_{z_t}(w_t)\|^2] \\ &\leq F(w_t) - \eta_t \|\nabla F(w_t)\|^2 + \frac{1}{2} \rho HC \eta_t^2 F(w_t) \|\nabla F(w_t)\|^2 \\ &\leq F(w_t) - \mu^2 \eta_t (F(w_t))^2 + \frac{1}{2} \rho H h^2 C \eta_t^2 (F(w_t))^3 \end{aligned}$$

Let $\eta_t = \frac{\eta}{F(w_t)}$, since $\eta \leq \frac{\mu^2}{HC\rho h^2}$

$$\begin{aligned} \mathbb{E}_{z_t}[F(w_{t+1})] &\leq F(w_t)(1 - \eta\mu^2 + \frac{1}{2} \rho H h^2 C \eta^2) \\ &\leq (1 - \frac{\eta\mu^2}{2}) F(w_t). \end{aligned}$$

This completes the proof.

5.6 Experiments on stochastic normalized GD

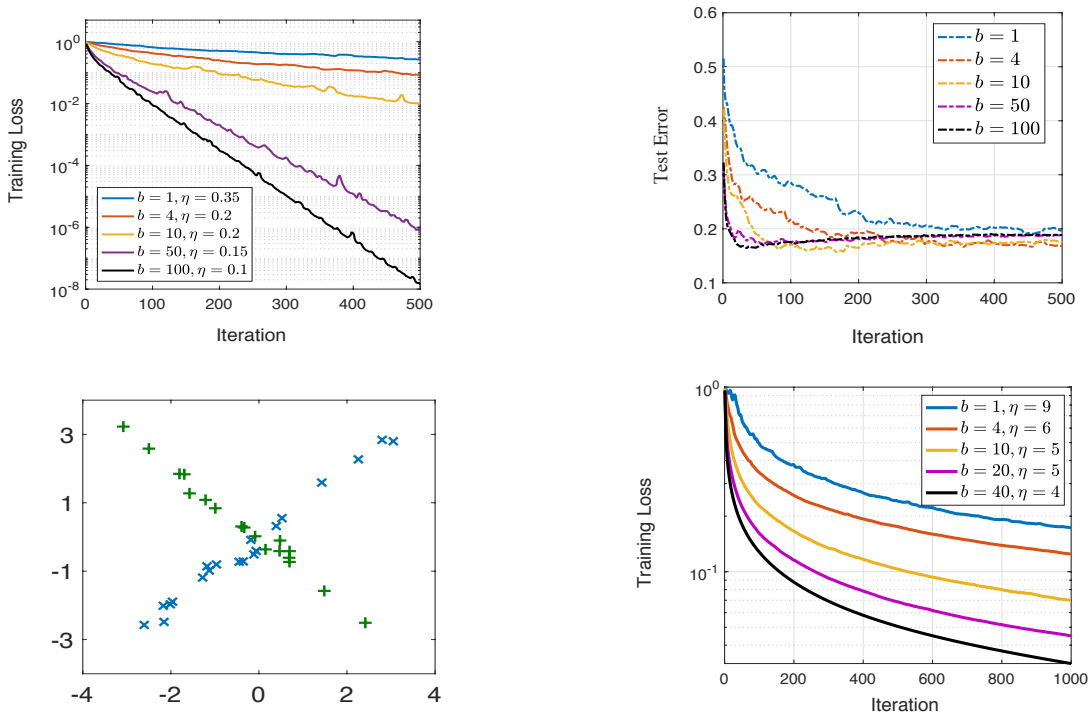


Figure 5.3: (Top) Training loss and Test error of stochastic normalized GD (Eq.(5.7)) on linear classification with signed measurements $y = \text{sign}(x^\top w^\star)$ with $d = 50$, $n = 100$. Here 'b' denotes the batch-size and 'η' is the fine-tuned step-size. (Bottom) Training loss of stochastic normalized GD on the dataset depicted in the left figure ($d = 2, n = 40$) for a two-layer neural network with $m = 50$ hidden neurons.

In this section, we evaluate the performance of stochastic normalized GD in Eq.(5.7) for linear and non-linear models. In Figure 5.3 (Top), we consider binary linear classification on signed data with the exponential loss and plot the training loss and test error performance based on iteration number. b denotes the batch-size from the sample dataset size of $n = 100$. The weight vector is initialized at zero for all curves ($w_0 = 0_d$). The right plot shows the test error for the same setup, where the optimal test error ($\tilde{F}^\star_{0-1} \approx 0.17$) is reached at various iteration numbers for each batch-size. In particular, for $b = 10$(yellow

line) stochastic normalized GD achieves the final test accuracy at almost the same time as the full-batch normalized GD (black line) while using 1/10 th gradient computations.

Figure 5.3 (Bottom) depicts the synthetic dataset of size $n = 40$ in $\mathbb{R}^2$ alongside with the training loss performance for each choice of batch-size b. Here we used a leaky-ReLU activation function as in Eq.(5.6) with $\ell = 1, \alpha = 0.2$.

Chapter 6

Decentralized Learning in the Interpolation Regime

6.1 Introduction

6.1.1 Motivation

Machine learning tasks often revolve around inference from data using empirical risk minimization (ERM):

$$\min_{w\in\mathbb{R}^d} \hat{F}(w) := \frac{1}{n}\sum_{i=1}^{n} f\left(w, x_i\right). \tag{6.1}$$

Here $f : \mathbb{R}^d \times \mathbb{R}^{d'} \to \mathbb{R}$ is a loss function and $x_i := y_i a_i$, where $(a_i, y_i)_{i=1}^n \overset{\text{iid}}{\sim} \mathcal{D}$ represent features and labels, sampled from a distribution $\mathcal{D}$. In large scale machine learning, due to privacy concerns and communication constraints, data points are often distributed on a set of local computing agents. Decentralized learning methods aim at minimizing the global loss function (6.1) while agents communicate their parameters on an underlying connected

graph. The most ubiquitous of these algorithms is Decentralized Gradient Descent (DGD). Here the ℓ th agent runs a step of gradient descent followed by an averaging step in which every agent replaces its parameter with the average of its neighbors [162]:

$$w_\ell^{(t+1)} = \sum_{k \in \mathcal{N}_\ell} A_{\ell k} w_k^{(t)} - \eta_t \nabla \hat{F}_\ell(w_\ell^{(t)}). \tag{6.2}$$

The superscripts signify the iteration number and $A_{\ell k}$ refers to the averaging weights used by agent ℓ for the parameter of agent $k \in \mathcal{N}_\ell$ where $\mathcal{N}_\ell$ is the set of neighbors of agent ℓ. The global loss $\hat{F}$ is the average of local loss functions $\hat{F}_\ell$, $\ell \leq N$, where each $\hat{F}_\ell$ is formed as the average empirical risk evaluated on the local training dataset $\mathcal{S}_\ell$ of the ℓ th agent:

$$\hat{F}(w) = \frac{1}{N} \sum_{\ell=1}^{N} \hat{F}_\ell(w), \quad \hat{F}_\ell(w) = \frac{1}{n_\ell} \sum_{x_j \in \mathcal{S}_\ell} f(w, x_j), \tag{6.3}$$

where n_ℓ denotes the dataset size of agent ℓ. Convergence properties of the train loss $\hat{F}(\cdot)$ in DGD have been studied extensively in literature, e.g., [162, 163, 164, 165, 166]. The bulk of these studies build upon classical optimization theory [155] suited for studying the train loss per iteration. In particular, it is well-stablished in the literature that DGD converges at the rate $\frac{1}{T}\sum_{t=1}^{T} \hat{F}(\bar{w}^{(t)}) - \hat{F}^\star = O(\frac{1}{\sqrt{T}})$ for smooth convex functions [163]. Here $\bar{w}^{(t)}$ is the average of local parameters $w_\ell^{(t)}$. Our results in Sections 6.2.1-6.2.2 show a rate of $\hat{F}(\bar{w}^{(T)}) = O(\frac{(\log T)^2}{T})$ and $\|W^{(T)} - \bar{W}^{(T)}\|_F^2 = O(\frac{(\log T)^4}{T^2})$ for the training loss and consensus error of DGD over separable data with "exponentially tailed" losses.

The study of generalization performance of DGD algorithms in the literature is mostly limited to empirical observations e.g., [167, 168, 169], making the theory behind test error performance largely unexplored. Moreover, the traditional wisdom in convergence analysis of DGD algorithms assumes the existence of a finite norm minimizer, which is often the case for ERM with non-separable training data, e.g. [170]. However, modern

machine learning models operate in over-parameterized settings where the model perfectly interpolates the training data, i.e., it achieves perfect accuracy on the training data [143]. Understanding the challenges imposed by over-parameterization and the behavior of gradient descent on separable data has been the subject of several recent works [138, 113, 132, 20, 140, 114, 22, 21, 152]. Yet, they are all focused on centralized GD, while here we study the impact of the consensus error of DGD on both training and generalization errors.

Our first goal is to complement prior general results on the convergence of training loss in DGD by considering specific, but commonly encountered, settings in ERM over separable data. This includes the analysis of non-smooth objectives such as the exponential loss, analysis of logistic regression in the separable regime where the optimum is achieved at infinity, and analysis of objectives satisfying the PL condition. The second goal is to study, for the first time in these settings, convergence rates of the DGD test loss $F(\bar{w}^{(t)}) := \mathbb{E}_{x\sim\mathcal{D}}[f(\bar{w}^{(t)}, x)]$. Finally, we leverage recent advances in the study of centralized learning with separable data to design fast algorithms for decentralized learning. We discuss our contributions below.

Contributions. In Sections 6.2.1 and 6.2.3, we derive convergence rates for the training and test loss of DGD over separable data. Our results hold for convex losses satisfying realizability and self-boundedness, as well as, convex losses satisfying self-boundedness and the PL condition. In Section 6.2.2, we prove under additional self-boundedness assumptions on the Hessian and gradient, which hold for exponentially tailed losses, that the test loss bound can be improved to approximately match the test loss bounds of centralized GD. When specialized to decentralized logistic regression on separable data, our results provide the first generalization guarantees of DGD. In Section 6.2.4, we propose two algorithms for speeding up the convergence of decentralized learning under separable

data. Numerical experiments demonstrate that our proposed algorithms significantly improve both the train test error of decentralized logistic regression.

6.1.2 Further related works

Decentralized learning. Over the last few years there have been numerous research works which consider the convergence of first order methods for decentralized learning; an incomplete list includes [162, 163, 164, 165, 167, 171, 172, 173, 170, 174, 175, 176]. While DGD is suboptimal for strongly-convex objectives [163, 166], alternative algorithms, namely EXTRA and Grading Tracking, for achieving exponential rate appeared in [177, 178] and were studied further in [179, 174]. More recently, [180] proposes accelerated methods for improving generalization and training accuracy of decentralized algorithms; however, their study of generalization error is empirical. The concurrent works [181, 182] study the generalization bounds of decentralized methods for Lipschitz convex losses (see also [183, 184]). However, we consider exponentially tailed losses under the separable data regime and prove faster convergence and generalization rates under these conditions. Compared to these works, we also propose improved algorithms for learning with separable data. Finally, we highlight that our rates on the train loss are comparable to [170, Theorem 2]. While [170] also derives convergence of DGD train loss on separable data, their analysis is valid only for bounded optimizers. In contrast, we derive training loss bounds which are true for the case of unbounded optimizers as is the case for logistic regression over separable data.

Implicit bias of GD. An early work on the behavior of ERM with vanishing regularization on separable data appeared in [148]. Closely related, a line of recent works [138, 113, 20, 141, 21, 114, 152] studies the parameter convergence, as well as training and test loss convergence, of gradient descent on separable data, showing that for (a class

of) monotonic losses the solution to ERM and the max-margin solution are the same in direction., i.e., $\|\hat{w}^{(t)} - \hat{w}_{\text{MM}}\| \to 0$. Here $\hat{w}^{(t)} := w^{(t)}/\|w^{(t)}\|$ and $\hat{w}_{\text{MM}} := w_{\text{MM}}/\|w_{\text{MM}}\|$, where the vector w_{MM} is the solution to the hard-margin support vector machine problem,

$$w_{\text{MM}} := \arg\min_{w \in \mathbb{R}^d} \|w\| \quad \text{s.t.} \quad y_i w^\top a_i \geq 1, \quad \forall i \in [n].$$

Notably, [113, 138] characterized the rate of directional convergence to be $\|\hat{w}^{(T)} - \hat{w}_{\text{MM}}\| = O(1/\log(T))$ and for the training loss to be $\hat{F}(w^{(T)}) = O(\frac{1}{\eta T})$. Recently, Shamir [114] and Schliserman and Koren [152] showed that the test loss of GD for logistic regression on linearly separable data satisfies $F(w^{(T)}) = \tilde{O}(\frac{1}{\eta T} + \frac{1}{n})$ signifying that overfitting does not happen during the iterates of GD. In Section 6.2.2 (Remark 6.2.5), we show that the test loss of DGD with logistic regression on linearly separable data satisfies $\mathbb{E}[F(\bar{w}^{(T)})] = \tilde{O}(\frac{1}{\eta T} + \frac{1}{n} + \eta^2)$, where the expectation is taken over training samples chosen i.i.d. from the dataset. As we explain, the term η^2 captures the impact of consensus error (i.e., decentralization) on the generalization rate.

While directional convergence is significantly slow for gradient descent, following the update rule $w^{(t+1)} = w^{(t)} - \eta_t \frac{\nabla \hat{F}(w^{(t)})}{\|\nabla \hat{F}(w^{(t)})\|}$, it can be improved to $1/\sqrt{t}$ with decaying η_t at rate $1/\sqrt{t}$ for linear models [20]. Furthermore [185] proved improved training convergence of this algorithm for two-layer neural networks, suggesting the benefits extend to non-linear settings. These results apply to centralized optimization scenarios. However, in decentralized learning settings, the local loss functions are kept private and any information about the global loss, such as its gradient $\|\nabla \hat{F}(\bar{w}^{(t)})\|$ is hidden from the agents. In Section 6.2.4, we propose algorithms which address these challenges and extend the normalized GD update rule to decentralized learning scenarios. Furthermore, we prove the asymptotic convergence of normalized local parameters $w_i^{(t)}/\|w_i^{(t)}\|$ to the solution of centralized GD.

Notation We use $\|\cdot\|$ to denote the ℓ_2-norm of vectors and the operator norm of matrices. The Frobenius norm of a matrix W is shown by $\|W\|_F$. The set $\{i \in \mathbb{N} : i \leq N\}$ is denoted by $[N]$. The gradient and hessian of a function $F : \mathbb{R}^d \to \mathbb{R}$ are denoted by $\nabla F(\cdot)$ and $\nabla^2 F(\cdot)$, respectively. For functions $f, g : \mathbb{R} \to \mathbb{R}$, we write $f(t) = O(g(t))$ when $|f(t)| \leq Mg(t)$ after $t \geq t_0$ for positive constants M, t_0. Finally, we write $f(t) = \tilde{O}(g(t))$ when $f(t) = O(g(t)h(t))$ for a polylogarithmic function h.

6.2 Main Results

Throughout this chapter we make the following standard assumption on the mixing matrix $A = [A_{ij}]_{N\times N}$ corresponding to the underlying connected network.

Assumption 6.2.1 (Mixing matrix)**.** *The mixing matrix $A \in \mathbb{R}^{N\times N}$ is symmetric, doubly stochastic with bounded spectrum i.e., $|\lambda_i(A)| \in (0, 1]$ and $\lambda_2(A) < 1$.*

First, we state a lemma which relates the generalization loss of DGD at iteration t to its train loss and consensus error up to iteration t. The lemma is derived based on a stability analysis [13, 17, 14]. Specifically we use a self-boundedness and a realizability assumption [152] which makes the stability analysis feasible for settings such as logistic regression on separable data. Additionally, we assume convexity and L-smoothness of the loss function. Formally, we assume the following, where for simplicity, we use the shorthand $f_x(w) := f(w, x)$ for the loss incurred at a generic $x \in \mathcal{D}$ in the data distribution $\mathcal{D}$.

Assumption 6.2.2 (Convexity)**.** *The loss functions $f_x : \mathbb{R}^d \to \mathbb{R}$ are convex and differentiable, satisfying, $f_x(w) \leq f_x(v) + \langle \nabla f_x(w), w - v\rangle$.*

Assumption 6.2.3 (Smoothness)**.** *The loss functions $f_x : \mathbb{R}^d \to \mathbb{R}$ are L-smooth and differentiable, i.e. $f_x(w) \leq f_x(v) + \langle \nabla f_x(v), w - v\rangle + \frac{L}{2}\|w - v\|^2$.*

Assumption 6.2.4 (Self-boundedness of the gradient). *The loss functions $f_x : \mathbb{R}^d \to \mathbb{R}$ satisfy the self-boundedness property with the parameters $c > 0$ and $\alpha \in [\frac{1}{2}, 1]$, i.e.,*

$$\|\nabla f_x(w)\| \leq c\,(f_x(w))^{\alpha}\,.$$

Assumption 6.2.4 is weaker than Assumption 6.2.3, since an L-smooth non-negative function f satisfies $\|\nabla f(w)\|^2 \leq 2L(f(w) - f^\star) \leq 2Lf(w)$, where $f^\star := \inf_w f(w) \geq 0$. However, we make use of the smoothness property whenever it suits the analysis, particularly to bound training loss.

Additionally, we make the following assumptions: All local parameters are initiated at zero i.e, $w_\ell^{(1)} = 0$ for all $\ell \leq N$. We assume for simplicity of exposition, that each agent has access to n/N ($n_\ell = n/N$) samples from the dataset. The general case can be treated with minor modifications. We also assume that $f_x(w) \geq 0$ for all w and the minimum of each loss is zero i.e., $f_i^\star = 0$.

Before our key lemma, we introduce a few necessary notations. We define matrix $W^{(t)} \in \mathbb{R}^{N\times d}$ as the concatenation of all agents' parameters at iteration t, i.e., $W = [w_1^{(t)}, \cdots, w_N^{(t)}]^\top$. We also denote by $\bar{w}^{(t)} := \frac{1}{N}\sum_{\ell=1}^N w_\ell^{(t)}$ the average of local parameters, and denote by $\bar{W}^{(t)} = [\bar{w}^{(t)}, \cdots, \bar{w}^{(t)}] \in \mathbb{R}^{N\times d}$ its concatenated matrix.

Lemma 6.2.1 (Key lemma, Informal version). *Let Assumptions 6.2.1-6.2.4 hold. Consider the iterates of decentralized gradient descent in Eq.(6.2) with a fixed positive step-size*

$\eta \leq \frac{2}{L}$. *Then, for the test loss* F *at iteration* $T \geq 1$, *it holds that*

$$\begin{aligned}\mathbb{E}\left[F(\bar{w}^{(T)})\right] \lesssim \; & \mathbb{E}\left[\hat{F}(\bar{w}^{(T)})\right] \\ & + \frac{\eta^2 L^2 c^2 T^2}{n^{3-2\alpha}} \mathbb{E}\Big[(\frac{1}{T}\sum_{t=1}^{T} \hat{F}(\bar{w}^{(t)}))^{2\alpha}\Big] \\ & + \frac{\eta^2 L^4}{N} \mathbb{E}\Big[(\sum_{t=1}^{T} \|W^{(t)} - \bar{W}^{(t)}\|_F)^2\Big], \end{aligned} \tag{6.4}$$

where the expectation is over the training set of n *i.i.d samples.*

The precise statement and the proof of Lemma 6.2.1 are deferred to Appendix 6.5.1. Lemma 6.2.1 bounds the test loss with respect to the train loss and the consensus error. In the following sections, we show how Lemma 6.2.1 yields test loss bounds on DGD by establishing bounds on the train loss and consensus errors under different assumptions on the loss function.

It is worth remarking that Eq. (6.4) is in fact valid not only for DGD, but also for Decentralized Gradient Tracking (DGT). DGT is another popular algorithm for distributed learning that can accelerate train error convergence over DGD by modifying the update in Eq. (6.2) such that each agent keeps a running estimate of the global gradient [178]. The reason why (6.4) continues to hold for DGD is that the proof of Lemma 6.2.1 only relies on the updates of the "averaged" parameter $\bar{w}^{(t)} := \frac{1}{N}\sum_{\ell=1}^{N} w_\ell$ and that the update rule of $\bar{w}^{(t)}$ for both DGD and DGT is derived as $\bar{w}^{(t)} = \bar{w}^{(t-1)} - \frac{\eta}{N}\sum_{\ell=1}^{N} \nabla \hat{F}_\ell(w_\ell^{(t-1)})$. Thus, starting with Eq.(6.4) one can also obtain test loss bounds of DGT after replacing appropriate bounds of DGT for the training loss and consensus error. We leave this to future work.

6.2.1 Convergence with general convex losses

The upper-bound in Eq.(6.4) shows how the consensus error and train loss of DGD affect the test loss.

The next lemma bounds the training loss and consensus error of DGD for general convex losses. The proof is deferred to Appendix 6.5.2.1

Lemma 6.2.2 (Training bounds for convex losses)**.** *Under Assumptions 6.2.1-6.2.3, for any $w \in \mathbb{R}^d$ and for a fixed step-size*

$$\eta < \frac{1}{L} \min \left\{ 1 - \alpha_1, \sqrt{\frac{1-\alpha_1}{2\alpha_2}} \right\},$$

where $\alpha_1 \in (3/4, 1), \alpha_2 > 4$ are parameters that depend only on the mixing matrix, the train loss and consensus error of DGD (6.2) satisfy:

$$\frac{1}{T} \sum_{t=1}^{T} \hat{F}(\bar{w}^{(t)}) \leq \frac{2\|w\|^2}{\eta T} + 4\hat{F}(w), \tag{6.5}$$

$$\frac{1}{NT} \sum_{t=1}^{T} \|W^{(t)} - \bar{W}^{(t)}\|_F^2 \leq \frac{\alpha_2 \eta^2 L^2}{1-\alpha_1} \left(\frac{2\|w\|^2}{\eta T} + 4\hat{F}(w) \right).$$

To bound the training loss for functions $f(\cdot)$ where the optimum is attained at infinity we need a realizability assumption. In particular, we choose $w \in \mathbb{R}^d$ (in Lemma 6.2.2) using the following.

Assumption 6.2.5 (Realizability)**.** *The loss functions $f_x : \mathbb{R}^d \to \mathbb{R}$ satisfy the realizability condition, i.e. $\exists$ decreasing function $\rho : \mathbb{R}_+ \to \mathbb{R}_+$ such that for every $\varepsilon > 0$ there exists $\hat{w} \in \mathbb{R}^d$ with $\|\hat{w}\| \leq \rho(\varepsilon)$ that satisfies $f_x(\hat{w}) \leq \varepsilon$.*

The set of Assumptions 6.2.2-6.2.5 covers classification over linearly separable data with logistic loss, in addition to losses with other exponential-type tails $\exp(-w^r)$ and polynomial tail w^{-r}, for $r > 0$.

Remark 6.2.1 (Training loss of DGD on separable data). The realizability assumption as stated appeared recently in [152] (and was implicitly used in [113, 114]). It can be checked that for linearly separable training data with margin γ, loss functions with an exponential tail such as logistic loss satisfy this assumption with $\rho(\varepsilon) = \frac{1}{\gamma}\log(\frac{1}{\varepsilon})$ (e.g., see Proposition 6.6.3 and [152, Lemma 4]). Based on Lemma 6.2.2, this leads to the following bound for DGD training loss for all $\varepsilon > 0$,

$$\frac{1}{T}\sum_{t=1}^{T}\hat{F}(\bar{w}^{(t)}) \leq \frac{2\log(1/\varepsilon)^2}{\gamma^2\eta T} + 4\varepsilon. \tag{6.6}$$

In particular, choosing $\varepsilon = 1/T$, gives a rate of $O(\frac{(\log T)^2}{\eta T})$, surprisingly matching up to logarithmic factors the corresponding rate for centralized GD in [113, Theorem 1.1].

Remark 6.2.2. The bounds of Lemma 6.2.2 are true for any dataset $\{x_i\}_{i\in[n]}$ provided that Assumptions 6.2.2 and 6.2.3 hold for all $f_x = f_{x_i} = f(w, x_i) := f_i(w), i \in [n]$. Similarly, (6.6) holds provided Assumption 6.2.5 is true over the training set (i.e. provided the training dataset is separable). However, bounding the test loss in Lemma 6.2.1, requires bounding the *expectation over all datasets* of the train/consensus errors. This is guaranteed by Assumptions 6.2.2-6.2.5 as they hold for any point x in the distribution.

Theorem 6.2.1 (Test loss with convex losses). *Under Assumptions 6.2.1-6.2.5, by choosing*

$$\eta < \frac{1}{L\sqrt{T}}\min\left\{1-\alpha_1, \sqrt{\frac{1-\alpha_1}{2\alpha_2}}\right\}$$

where $\alpha_1 \in (3/4, 1), \alpha_2 > 4$ are parameters that depend only on the mixing matrix and

assuming $\varepsilon \leq \frac{\rho(\varepsilon)^2}{\eta T}$*, the test error of DGD for iteration* $T \geq 1$ *satisfies:*

$$\frac{1}{T}\sum_{t=1}^{T} \mathbb{E}\left[F(\bar{w}^{(t)})\right] = O\Big(\frac{\rho(\varepsilon)^2}{\sqrt{T}} + \frac{L^2 c^2 \rho(\varepsilon)^{4\alpha}}{n^{3-2\alpha}} T^{1-\alpha} + \frac{L^4 \rho(\varepsilon)^2}{\sqrt{T}}\Big), \tag{6.7}$$

where the expectation is over the training set of n *i.i.d samples.*

Remark 6.2.3 (DGD with logistic regression never overfits). The proof of Theorem 6.2.1 is delayed to Appendix 6.5.2.2. As in Remark 6.2.1, we take logistic regression on separable data with margin $\gamma > 0$ as our case study. For logistic regression (as well as other loss functions with an exponential tail), it can be verified that the self-boundedness assumption holds with $\alpha = 1$. Similar to Remark 6.2.1 it holds that $\rho(\varepsilon) = \frac{1}{\gamma}\log(\frac{1}{\varepsilon})$, thus choosing $\varepsilon = 1/\sqrt{T}$ results in a test loss rate $\tilde{O}(\frac{1}{\sqrt{T}} + \frac{1}{n})$ by Eq.(6.7). This indicates that the upper-bound decreases at a rate of $\tilde{O}(\frac{1}{\sqrt{T}})$ until after $T = n^2 \cdot (\max(\frac{1}{Lc}, \frac{L}{c}))^4$ iterations where the upper bound essentially reduces to $\tilde{\mathcal{O}}(\frac{L^2 c^2}{n})$. Additionally, the fact that the upper-bound is decreasing proves that with appropriate choice of step-size, overfitting never happens along the path of DGD at any iteration.

Remark 6.2.4 (Log factors). The attentive reader will have recognized in Remarks 6.2.1 and 6.2.3 that due to the "$\rho(\varepsilon) = \mathcal{O}(\log(T))$" factor, the upper bound on the test loss in Eq. (6.7) increases (very) slowly with $\log^4(T)$. Note that this term becomes dominant only when T is exponentially large with respect to the sample size n and the margin γ. Our experiments in Sec. 6.3.2 confirm this slow logarithmic increase late in the training phase. Analogous behavior, but for centralized GD training, are discussed in [138, 152].

6.2.2 On the convergence of DGD with exponentially-tailed losses

In this section, we show that our guarantees can be improved for exponentially tailed losses. First, we note that the bounds in Lemma 6.2.2 and Theorem 6.2.1 hold for the average loss across iterations $t \leq T$. It is straight-forward to see that if DGD is a descent algorithm i.e., $\hat{F}(\bar{w}^{(t+1)}) \leq \hat{F}(\bar{w}^{(t)})$ for all $t \leq T$, then $\hat{F}(\bar{w}^{(T)}) \leq \frac{1}{T}\sum_{t=1}^{T} \hat{F}(\bar{w}^{(t)})$; thus implying that the upper-bounds on training and test loss hold for the last iterate of DGD. We will prove that DGD is indeed a "descent algorithm" for a class of convex losses which include popular choices such as the logistic loss and even non-smooth choices including the exponential loss. Moreover, we show that the consensus error of Lemma 6.2.2 as well as the test loss bounds of Theorem 6.2.1 can be improved compared to the results of the previous section.

In particular, we use the following assumptions together with the self-boundedness gradient assumption (Assumption 6.2.4) with $\alpha = 1$ as well as the convexity assumption.

Assumption 6.2.6 (Self-bounded Hessian)**.** *The local losses $\hat{F}_\ell : \mathbb{R}^d \to \mathbb{R}$ satisfy the following for the Hessian matrices $\nabla^2 \hat{F}_\ell$ and a positive constant h,*

$$\|\nabla^2 \hat{F}_\ell(w)\| \leq h\, \hat{F}_\ell(w).$$

Assumption 6.2.7 (Self-lowerbounded gradient)**.** *The global loss satisfies for a constant τ that*

$$\|\nabla \hat{F}(w)\| \geq \tau \hat{F}(w).$$

Assumptions 6.2.2, 6.2.4, 6.2.6 and 6.2.7 include linear classification with non-smooth losses such as the exponential loss, losses with super-exponential tails ($\exp(-x^r), r > 1$) and the logistic loss; e.g., see Proposition 6.6.1 in the appendix.

Theorem 6.2.2 (Last iterate convergence of DGD)**.** *Consider DGD with the loss functions*

and mixing matrix satisfying Assumptions 6.2.1,6.2.2,6.2.6,6.2.7 and Assumption 6.2.4 with $\alpha = 1$ *and* $c = h$*. Assume that the step-size satisfies* $\eta < \frac{\delta}{\hat{F}(1)}$*, for a constant* δ *depending only on the mixing matrix and on* τ, h*, then DGD is a descent algorithm i.e, for all* $t \geq 1$ *it holds that* $\hat{F}(\bar{w}^{(t+1)}) \leq \hat{F}(\bar{w}^{(t)})$*. Moreover, the train loss and the consensus error of DGD at iteration* T *satisfy the following for all* $w \in \mathbb{R}^d$*,*

$$\begin{aligned} \hat{F}(\bar{w}^{(T)}) &\leq 4\hat{F}(w) + \frac{2\|w\|^2}{\eta T}, \\ \left\|W^{(T)} - \bar{W}^{(T)}\right\|_F^2 &= O\left(h^2\eta^2\hat{F}^2(w) + \frac{h^2\|w\|^4}{T^2}\right). \end{aligned} \tag{6.8}$$

The proof of Theorem 6.2.2 is included in Appendix 6.5.3.1. In the following remark, we discuss the implications of this result.

Remark 6.2.5 (Improved rates). While similar to Lemma 6.2.2, for logistic regression we have $\hat{F}(\bar{w}^{(T)}) = \tilde{O}(\frac{1}{\eta T} + \frac{1}{T})$, for the consensus error rate we have by applying Theorem 6.2.2 and noting that $\rho(\varepsilon) = \log(1/\varepsilon)/\gamma$,

$$\left\|W^{(T)} - \bar{W}^{(T)}\right\|_F^2 = O\left(h^2\eta^2\varepsilon^2 + \frac{h^2(\log(1/\varepsilon))^4}{\gamma^4 T^2}\right).$$

After choosing $\varepsilon = 1/T$, we have the improved rate $\|W^{(T)} - \bar{W}^{(T)}\|_F^2 = \tilde{O}(\frac{1}{T^2})$, which is superior over the rate $\tilde{O}(\frac{1}{T})$ for general convex losses with constant η (Lemma 6.2.2). For the test loss, employing Lemma 6.2.1 with the new rates for the consensus error leads to the following rate for DGD with logistic regression,

$$\mathbb{E}\left[F(\bar{w}^{(T)})\right] = \tilde{O}\Big(\frac{1}{\eta T} + \frac{1}{n} + \eta^2\Big). \tag{6.9}$$

In accordance to Remark 6.2.2, we can conclude the above from Lemma 6.2.1 provided Assumptions 6.2.4 and 6.2.7. Thus, the bounds of Theorem 6.2.2 remain true for all

training sets within the data distribution. We note that the resulting bound in (6.9) is a superior rate for the test loss of logistic regression, compared to the rate of Remark 6.2.3. Concretely, setting $\eta = 1/T^{1/3}$ gives a rate of $\tilde{O}(1/T^{2/3} + 1/n)$, faster than the $\tilde{O}(1/\sqrt{T} + 1/n)$ rate in Remark 6.2.3. On the other hand, it is slightly slower compared to its centralized counterpart $\tilde{O}(1/T + 1/n)$ in [114, 152]. As revealed by Lemma 6.2.1, the additional η^2 factor in (6.9) captures impact of the consensus term, which is unavoidable in decentralized learning.

6.2.3 Convergence under the PL condition

Next, we show how our previous results change when the global loss satisfies the μ-PL condition. Formally, the PL condition [156, 157] is defined as follows.

Assumption 6.2.8 (PL condition)**.** *The loss function $\hat{F} : \mathbb{R}^d \to \mathbb{R}$ satisfies the Polyak-Lojasiewic(PL) condition with parameter $\mu > 0$: $\|\nabla \hat{F}(w)\|^2 \geq 2\mu(\hat{F}(w) - \hat{F}^\star)$.*

The next lemma shows that DGD enjoys an exponential rate under the PL condition and smoothness. and data separability (i.e., $\hat{F}^\star = 0$). See Appendix 6.5.4.1 for a proof.

Lemma 6.2.3 (Train loss under the PL condition)**.** *Let Assumptions 6.2.1,6.2.3 and 6.2.8 hold and let the step-size $\eta \leq \min\{\frac{1-\alpha_1}{\mu}, \frac{1}{2L^2}\sqrt{\frac{(1-\alpha_1)\mu}{\alpha_2}}, \frac{1}{L}\}$, where the constants $\alpha_1 \in (3/4, 1)$ and $\alpha_2 > 4$ depend only on the mixing matrix. Define $\zeta := 1 - \frac{\eta\mu}{2}$, then under the data separability assumption, the iterates of DGD satisfy for all $t \geq 1$,*

$$\hat{F}(\bar{w}^{(t)}) \leq \zeta^{t-1}\hat{F}(\bar{w}^{(1)}),$$
$$\frac{1}{N}\left\|W^{(t)} - \bar{W}^{(t)}\right\|_F^2 \leq \frac{2\alpha_2\eta^2L^2\hat{F}(\bar{w}^{(1)})}{1-\alpha_1}\zeta^{t-1}.$$

We use this lemma combined with our key lemma 6.2.1 to obtain the test loss bound in the next theorem. The proof is provided in Appendix 6.5.4.2.

Theorem 6.2.3 (Test loss under the PL condition)**.** *Let Assumptions 6.2.1-6.2.4 hold. Further assume 6.2.8 holds for all training sets in the distribution. Let η and ζ be as in Lemma 6.2.3. Then the iterates of DGD satisfy for all $T \geq 1$,*

$$\mathbb{E}\left[F(\bar{w}^{(T)})\right] = O\Big(\zeta^T + \frac{L^2c^2}{n^{3-2\alpha}\mu^{2\alpha}}(\eta T)^{2-2\alpha} + \frac{\eta^2 L^4}{\mu^2}\Big).$$

Remark 6.2.6. The bound above involves $(\eta T)^{2-2\alpha}$. When $\alpha < 1$, as in the case of smooth functions such as highly over-parameterized Least-squares $f(w, x) = (1 - w^\top x)^2$ where $d \gg n$, the bound becomes vacuous as it is increasing with T. This suggests the existence of overfitting in DGD under such scenarios; with the optimal value of T achieved at the very early steps of training. See Appendix 6.7.1 for experiments that confirm this behavior.

6.2.4 Improved Algorithms: Fast Distributed Logistic Regression(FDLR)

In this section, we consider decentralized learning with exponentially tail losses on separable data and propose modifications to the DGD algorithm for improving the convergence rates based on the normalized GD mechanism.

Our first proposed algorithm –Fast Distributed Logistic Regression(FDLR)– is summarized in Algorithm 1. Each agent keeps two local variables $w_\ell, v_\ell \in \mathbb{R}^d$ which are also communicated to neighbor agents at each round. In matrix notation, Algorithm 1 has the following updates:

$$W^{(t+1)} = A(W^{(t)} - \eta \widetilde{V}^{(t)}),$$
$$V^{(t+1)} = A\, V^{(t)} + \nabla \hat{F}(W^{(t+1)}) - \nabla \hat{F}(W^{(t)}).$$

As in (6.2), $A \in \mathbb{R}^{N\times N}$ is the mixing matrix of the undirected network of agents, which

Algorithm 1: FDLR

Input: Initial values $w_\ell^{(1)}, v_\ell^{(1)} \in \mathbb{R}^d$ for all agents $\ell \in [N]$, step size η_t and mixing matrix $A = [A_{\ell k}]_{N\times N}$

1 **for** $t = 1, \ldots, T$ *all agents* $\ell \in [N]$ *in parallel* **do**

2 $\quad w_\ell^{(t+\frac{1}{2})} = w_\ell^{(t)} - \eta_t \frac{v_\ell^{(t)}}{\|v_\ell^{(t)}\|}$

3 $\quad$ send and receive local variables $w_\ell^{(t+\frac{1}{2})}$ and $v_\ell^{(t)}$

4 $\quad w_\ell^{(t+1)} = \sum_{k\in\mathcal{N}_\ell} A_{\ell k} w_k^{(t+\frac{1}{2})}$

5 $\quad v_\ell^{(t+1)} = \sum_{k\in\mathcal{N}_\ell} A_{\ell k} v_k^{(t)} + \nabla\hat{F}_\ell(w_\ell^{(t+1)}) - \nabla\hat{F}_\ell(w_\ell^{(t)})$

Algorithm 2: FDLR with Nesterov momentum

Input: Initial values $w_\ell^{(1)}, v_\ell^{(1)}, z^{(1)} \in \mathbb{R}^d$ for all agents $\ell \in [N]$, hyper-parameters η_t, γ_t and mixing matrix $A = [A_{\ell k}]_{N\times N}$

1 **for** $t = 1, \ldots, T$ *all agents* $i \in [N]$ *in parallel* **do**

2 $\quad z_\ell^{(t+1)} = \gamma_t(z_\ell^{(t)} + \frac{v_\ell^{(t)}}{\|v_\ell^{(t)}\|})$

3 $\quad w_\ell^{(t+\frac{1}{2})} = w_\ell^{(t)} - \eta_t(z_\ell^{(t+1)} + \frac{v_\ell^{(t)}}{\|v_\ell^{(t)}\|})$

4 $\quad$ send and receive local variables $w_\ell^{(t+\frac{1}{2})}$ and $v_\ell^{(t)}$

5 $\quad w_\ell^{(t+1)} = \sum_{k\in\mathcal{N}_\ell} A_{\ell k} w_k^{(t+\frac{1}{2})}$

6 $\quad v_\ell^{(t+1)} = \sum_{k\in\mathcal{N}_\ell} A_{\ell k} v_k^{(t)} + \nabla\hat{F}_\ell(w_\ell^{(t+1)}) - \nabla\hat{F}_\ell(w_\ell^{(t)})$

satisfies the regularity conditions in Assumption 6.2.1. Furthermore $W^{(t)}, V^{(t)}, \nabla\hat{F}(W^{(t)}) \in \mathbb{R}^{N\times d}$ are formed by stacking $w_\ell^{(t)}, v_\ell^{(t)}$ and local gradients $\nabla\hat{F}_\ell(w_\ell^{(t)})$ for all $\ell \in [N]$ as their rows. The matrix $\widetilde{V}^{(t)} \in \mathbb{R}^{N\times d}$ is formed by concatenation of the vectors $v_\ell^{(t)}/\|v_\ell^{(t)}\|$ as its rows. In step (2) of Algorithm 1, every agent ℓ runs in parallel an update rule which resembles the distributed gradient descent update rule (aka Eq. (6.2)), with the difference that the local gradient $\nabla\hat{F}_\ell(w_\ell^{(t)})$ is replaced by $v_\ell^{(t)}/\|v_\ell^{(t)}\|$. In the next step, agents send their local parameters $w_\ell^{(t)}, v_\ell^{(t)}$ to their neighbors. Step (4) is the consensus step at which agent ℓ computes a weighted average of $w_k^{(t+1/2)}$ sent from neighbor agents k, in order to update $w_\ell^{(t)}$. Step (5) uses the newly computed local gradient $\nabla\hat{F}_\ell(w_\ell^{(t+1)})$

and the gradient computed in the previous step to updates the local parameter $v_\ell^{(t)}$. The purpose behind introducing the variable $v_\ell^{(t)}$ is to estimate the global gradient. This idea is previously used in the gradient tracking algorithm (e.g. see [178]) and the idea also relates to stochastic variance reduced gradient (SVRG) [186]. The following theorem proves that for exponentially decaying loss functions and separable data, FDRL with time-decaying step-size $\eta_t = 1/\sqrt{t}$ converges successfully in direction to the solution of centralized gradient descent. The proof is provided in Appendix 6.5.5.

Theorem 6.2.4 (Asymptotic convergence of FDLR)**.** *Let the sequence* $\{w_\ell^{(t)}\}$ *be generated by* FDRL *(Algorithm 1) trained with logistic or exponential loss on a separable dataset with* $\eta_t = O(1/\sqrt{t})$*. Then, for all* $\ell \in [N]$*,* $\lim_{t\to\infty} w_\ell^{(t)}/\|w_\ell^{(t)}\| = w_{\text{MM}}/\|w_{\text{MM}}\|$*, where* w_{MM} *is the solution to max-margin problem.*

Based on the above result, we anticipate that FDRL has good test performance. In fact, we will show in Section 6.3 that FDRL achieves good test performance orders of magnitude faster than DGD. To get some insight on this and also on the nature of the FDLR updates consider the infinite time limit. In this limit, when the matrix A satisfies the mixing Assumption 6.2.1, it can be checked that $V^{(\infty)} = \frac{1}{n}\mathbf{1}\mathbf{1}^\top \nabla \hat{F}(W^{(\infty)})$. Hence, as $t \to \infty$ the variables $v_\ell^{(t)}$ for all agents converge to the same global gradient $\sum_{\ell=1}^N \nabla \hat{F}_\ell(w_\ell^{(t)})$. Realizing this, we can see that Step (2) of FDLR is asymptotically approximating a normalized GD update, i.e., for large t, each agent performs an update $w_\ell^{(t+1/2)} \approx w_\ell^{(t)} - \eta_t \frac{\sum_{\ell=1}^N \nabla \hat{F}_\ell(w_\ell^{(t)})}{\|\sum_{\ell=1}^N \nabla \hat{F}_\ell(w_\ell^{(t)})\|_2}$. Previously, normalized gradient descent has been used to speed up convergence in centralized logistic regression over separable data[20]. Here, we essentially extend this idea to a decentralized setting and argue that FDLR is the canonical way to do so. In particular, the idea of introducing additional variables v_ℓ that keep track of the global gradient is critical for the algorithm's success. That is, a naive implementation with updates $w_\ell^{(t+1/2)} = w_\ell^{(t)} - \eta_t \nabla \hat{F}_\ell(w_\ell^{(t)})/\|\nabla \hat{F}_\ell(w_\ell^{(t)})\|_2$ based only on

the local gradients would fail. At the other end, just introducing variables v_ℓ without performing a normalized gradient update (i.e. implementing gradient tracking) also fails to give significant speed ups over DGD. See Section 6.3 for experiments in support of this claim.

We also present a yet improved Algorithm 2, which combines FDLR with Nesterov Momentum. The key innovation of Algorithm 2 compared to FDLR is its step (3), where now the local parameter $w_\ell^{(t)}$ is updated by a weighted average ($z_\ell^{(t+1)}$) of normalized gradients from previous iterations. Similar to our previous remarks regarding FDLR, extending the Nesterov accelarated variant of normalized GD for centralized logistic regression [22] to the distributed setting is more subtle as now each agent has access only to local gradients. Our experiments in Section 6.3 verify the correctness of the proposed implementation of Algorithm 2 as it achieves significant speed ups over both DGD and FDLR.

6.3 Numerical Experiments

In this section, we present numerical experiments to verify our theoretical results and demonstrate the benefits of our proposed algorithms. We begin with a numerical study of the performance of FDLR.

6.3.1 Experiments on FDLR

In Fig. 6.1(Left), we compare the performance of FDLR and its momentum variant to DGD and gradient tracking (GT) for exponential loss with signed measurements (i.e., $y = \text{sign}(a^\top w^\star)$ for samples a, labels y and the true vector of regressors $w^\star$) with $n = 100$, $d = 25$. The underlying graph is selected as an Erdos-Rènyi graph with $N = 50$ agents and connectivity probability $p_c = 0.3$. On the $y-$axis, directional convergece represents

the distance of normalized $w_\ell^{(t)}$ to the normalized final solution for agent $\ell = 1$, i.e., $\| \frac{w_1^{(t)}}{\|w_1^{(t)}\|} - \frac{w_{\text{MM}}}{\|w_{\text{MM}}\|} \|$ (see Theorem 6.2.4). The hyper-parameters η_t, γ_t are fine-tuned for each algorithm to represent the best of each algorithm and the final values are $\eta_t = 0.1, 0.05, 0.5$ and 0.2 for Distributed GD, GT, Alg. 1 and Alg. 2, respectively and $\gamma_t = 0.8$ for Alg. 2. Our algorithms significantly outperform the well-known distributed learning algorithms in directional convergence to the final solution. Regardless, in this case we noticed that the gain obtained by including the momentum is small. In Fig. 6.1(Right), we consider a binary classification task on a real-world dataset (two classes of the UCI WINE dataset [187]) where $d = 13$ and $n = 107$. We compare the performance of FDLR (blue line) and its momentum variant (red line) with DGD and DGT on an Erdos-Rènyi graph with $N = 10$ and $p_c = 0.4$. The hyper-parameters are fine-tuned to $\eta_t = 12, 1, 0.9, 2$ for DGD, DGT, Alg. 1 and Alg. 2 respectively, and $\gamma_t = 0.88$ for Alg. 2. Notably, while Alg. 1 significantly outperforms both DGD and DGT, the benefits of adding the momentum are also significant in this case as Alg. 2 demonstrates a faster rate of convergence than Alg. 1.

The two plots in Fig. 6.2 (Left) illustrate the train and test errors of DGD/DGT and our proposed algorithms for the same setting as Fig. 6.1(Left) with $n = 800$ and $d = 50$. Here the hyper-parameters are fine-tuned to be $\eta_t = 0.01, 0.01, 0.4, 0.5$ for DGD,DGT, Alg. 1 and Alg. 2, respectively and $\gamma_t = 0.5$ for Alg. 2. Fig. 6.2(Right) shows the training and test losses. Here, we use the same dataset with $N = 10, p_c = 0.4$ and an exponential loss. The hyper-parameters are fine-tuned to $\eta_t = 0.01, 0.012, 0.4, 0.6$ and $\gamma_t = 0.9$. Both of our algorithms outperform the commonly used DGD and DT, in both training error and test error performance. Also, the gains of adding the momentum are significant, since FDLR with Nesterov momentum (Algorithm 2) reaches an approximation of its final test accuracy in 50 iterations, while the same happens for FDLR with approximately 300 iterations. An interesting phenomenon in Fig. 6.2 (see also Fig. 6.3(Right)) is

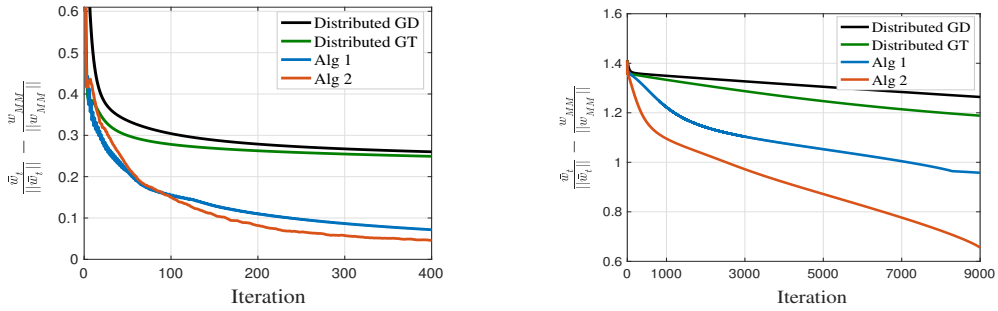


Figure 6.1: Directional parameter convergence of our proposed Algorithms 1-2 compared to the vanilla distributed gradient descent and gradient tracking algorithms on (Left) synthetic data $y = \text{sign}(a^\top w^\star)$ and (Right) on two classes from the UCI WINE dataset.

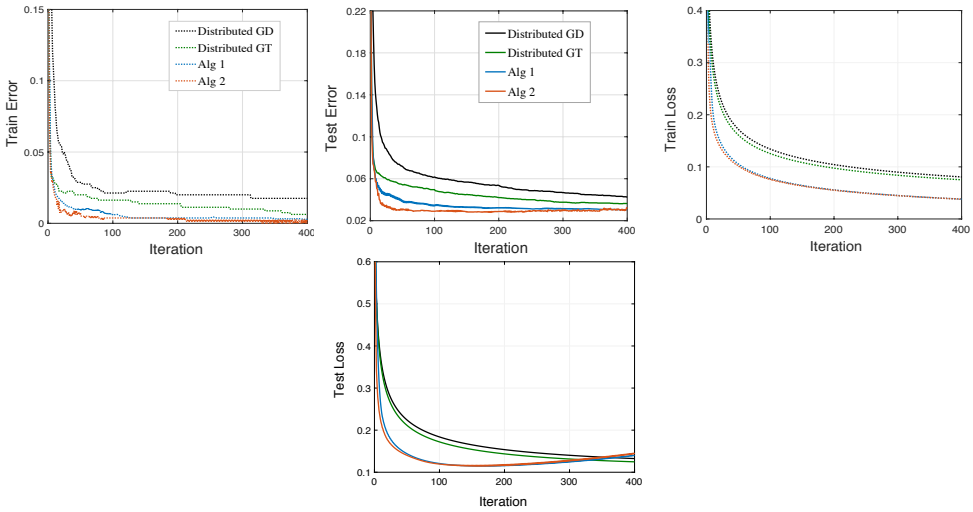


Figure 6.2: Training/test misclassification errors and train/test losses for our proposed algorithms compared to the decentralized gradient descent and gradient tracking algorithms on synthetic data $y = \text{sign}(a^\top w^\star)$ with $n = 800, d = 50$.

the behavior of test loss: while during the starting phase the test loss is monotonically decreasing, after sufficient iterations the test loss starts increasing. This behavior of test loss is indeed captured by the bounds on the test loss of DGD in Theorems 6.2.1-6.2.2 and Remarks 6.2.3-6.2.5. In particular, the increase in test loss is observed in the bound for the test loss $O(\frac{(\log T)^2}{\sqrt{T}} + \frac{(\log T)^2}{n})$ in Remark 6.2.3, where the presence of the term $\frac{(\log T)^2}{n}$ suggests that the bound after sufficient iterations starts to slowly increase. See

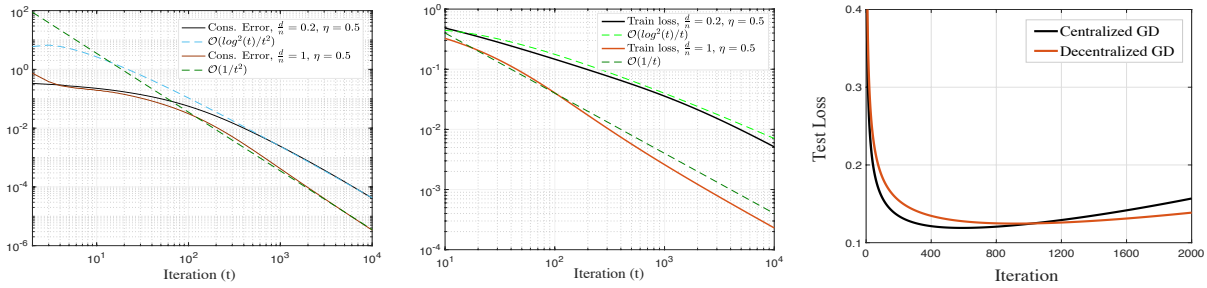


Figure 6.3: Consensus error, train loss and test loss for DGD with exponential loss and linearly separable data. Left and middle plots verify the rates $\tilde{O}(1/t^2)$ and $\tilde{O}(1/t)$ (Theorem 6.2.2 and Remark 6.2.5) for *consensus error* and *training loss* of DGD. Right plot shows test loss for DGD and GD show approximately similar convergence behavior under separable data.

also Remark 6.2.4.

6.3.2 Experiments on convergence of DGD

Next, we investigate the convergence behavior of DGD for the train loss and the consensus error. We consider the same network topology, mixing matrix and data setup as in the last figure. Left and middle plots in Fig. 6.3 show the consensus error $\frac{1}{N}\|W^{(t)} - \bar{W}^{(t)}\|_F^2$ and the train loss $\hat{F}(\bar{w}^{(t)})$ in solid lines, for two over-parameterization ratios d/n. We recall that d and n represent the dimension of the ambient space and the dataset size, respectively. The dashed lines help to show for each solid line, the approximate rate of convergence after sufficient number of iterations. Notably, we observe that the convergence rates on the consensus error ($\tilde{O}(1/t^2)$) and on the train error ($\tilde{O}(1/t)$) stated in Remark 6.2.5 are attained in both cases (recall that $\tilde{O}(\cdot)$ hides logarithmic factors). Fig. 6.3 (Right) depicts the Test loss of DGD for $d/n = 0.05$. For comparison, the corresponding curve for centralized GD is also shown. Here step-sizes are fine-tuned to represent the best of each algorithm. In agreement with our findings in Remark 6.2.5, we observe approximately similar behavior between the convergence behavior of two algorithms. As before, the slight increase in the curves of test loss are due to the

logarithmic factors in test loss upperbouds.

6.4 Conclusions

We studied the behavior of train loss and test loss of decentralized gradient descent (DGD) methods when training dataset is separable. To the best of our knowledge, this yields the first rigorous guarantees for the generalization error of DGD in such a setting. For the same setting, we also proposed fast algorithms and empirically verified that they accelarate both training and test accuracy. We believe our work opens several directions, with perhaps the most exciting one being the analysis of non-convex objectives. We are also interested in extending our results to other distributed settings such as federated learning [188] and Gradient Tracking e.g., [178].

6.5 Proofs

In this section, we present the proofs of all theorems and lemmas stated in the main body. We organize the appendix as follows,

The formal statement and proof of Lemma 6.2.1 are included in Appendix 6.5.1.

The proofs for Section 6.2.1 are included in Appendix 6.5.2.

The proofs for Section 6.2.2 are included in Appendix 6.5.3.

The proofs for Section 6.2.3 are included in Appendix 6.5.4.

The proof of Theorem 6.2.4 is included in Appendix 6.5.5.

Auxiliary results on our assumptions are included in Appendix 6.6.

Finally, we conduct complementary experiments in Appendix 6.7.

Notation

Throughout the appendix we use the following notations:

$$\bar{w} := \frac{1}{N}\sum_{\ell=1}^{N} w_\ell, \quad \bar{W} := [\bar{w}, \bar{w}, \cdots, \bar{w}]^\top \in \mathbb{R}^{N\times d},$$

$$W := [w_1, \cdots, w_N]^\top \in \mathbb{R}^{N\times d},$$

$$\hat{F}(W) := \frac{1}{N}\sum_{\ell=1}^{N} \hat{F}_\ell(w_\ell),$$

$$\nabla \hat{F}(w) := \frac{1}{n}\sum_{i=1}^{n} \nabla f(w, x_i),$$

$$\nabla \hat{F}(W) := [\nabla \hat{F}_1(w_1), \nabla \hat{F}_2(w_2), \cdots, \nabla \hat{F}_N(w_N)]^\top \in \mathbb{R}^{N\times d},$$

$$\bar{\nabla} \hat{F}(W) := \frac{1}{N}\sum_{\ell=1}^{N} \nabla \hat{F}_\ell(w_\ell),$$

$$\nabla \hat{F}_\ell(w_\ell) := \frac{N}{n}\sum_{x_j\in\mathcal{S}_\ell} \nabla f(w_\ell, x_j).$$

where recall that d is the dimension of ambient space, n is the total sample size, N is the number of agents and each agent has access to n/N samples.

6.5.1 Proof of Lemma 6.2.1

Lemma 6.5.1 (Formal statement of Lemma 6.2.1)**.** *Consider the iterates of decentralized gradient descent in Eq.(6.2) with a fixed positive step-size $\eta \leq \frac{2}{L}$. Let Assumptions 6.2.1-6.2.4 hold. Then for the test loss F at iteration $T \geq 1$, it holds that*

$$\mathbb{E}\left[F(\bar{w}^{(T)})\right] \leq 4\mathbb{E}\left[\hat{F}(\bar{w}^{(T)})\right] + \frac{9L^2c^2\eta^2T^2}{n^{3-2\alpha}}\mathbb{E}\left[\left(\frac{1}{T}\sum_{t=1}^{T}\hat{F}(\bar{w}^{(t)})\right)^{2\alpha}\right] \tag{6.10}$$
$$+ \frac{9L^4\eta^2}{N}\mathbb{E}\left[\left(\sum_{t=1}^{T}\|W^{(t)} - \bar{W}^{(t)}\|_F\right)^2\right] + \frac{9L^4\eta^2}{N}\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\left[\left(\sum_{t=1}^{T}\|W_{\neg i}^{(t)} - \bar{W}_{\neg i}^{(t)}\|_F\right)^2\right].$$

where the expectation is over training samples and $W_{\neg i}^{(t)}, \bar{W}_{\neg i}^{(t)}$ denote the parameter matrix and averaged parameter matrix at iteration t for the DGD algorithm when the i-th data sample is left out.

Proof: The proof relies on algorithmic stability[13, 17]. Specifically, we build on the framework introduced by [14] (and also used recently by [152]). Unlike these works, our analysis is for decentralized gradient descent.

We define $w_{\ell,\neg i}^{(t)}$ as the parameter of agent ℓ resulting from decentralized gradient descent at iteration t, when the i^{th} training sample $i \leq n$ is left out during training. We emphasize that the i^{th} sample may or may not belong to the dataset of agent ℓ.

We define $\bar{w}_{\neg i}^{(t)} \in \mathbb{R}^d$

$$\bar{w}_{\neg i}^{(t)} := \frac{1}{N}\sum_{j=1}^{N} w_{j,\neg i}^{(t)},$$

as the average of all agents' parameters at iteration t, when the i^{th} sample is left out of the algorithm. Thus, the parameter matrices $W_{\neg i}^{(t)}, \bar{W}_{\neg i}^{(t)} \in \mathbb{R}^{N\times d}$ are defined as follows,

$$W_{\neg i}^{(t)} := [w_{1,\neg i}^{(t)}, w_{2.\neg i}^{(t)}, \cdots, w_{N,\neg i}^{(t)}],$$
$$\bar{W}_{\neg i}^{(t)} := [\bar{w}_{\neg i}^{(t)}, \bar{w}_{\neg i}^{(t)}, \cdots, \bar{w}_{\neg i}^{(t)}].$$

The first step in the proof is to bound the term $\frac{1}{n}\sum_{i=1}^{n}\|\bar{w}^{(t)} - \bar{w}_{\neg i}^{(t)}\|^2$. By definition of DGD in Eq.(6.2), we have the following update rule for the averaged parameter,

$$\bar{w}^{(t+1)} = \bar{w}^{(t)} - \eta\bar{\nabla}\hat{F}(W^{(t)}).$$

Analogously,

$$\bar{w}_{\neg i}^{(t+1)} = \bar{w}_{\neg i}^{(t)} - \eta\bar{\nabla}\hat{F}(W_{\neg i}^{(t)}) = \bar{w}_{\neg i}^{(t)} - \frac{\eta}{n}\sum_{\ell=1}^{N}\sum_{x_j\in S_\ell, x_j\neq x_i}\nabla f(w_{\ell,\neg i}^{(t)}, x_j).$$

Thus by adding and subtracting $\bar{\nabla}\hat{F}(\bar{W}^{(t)})$ and $\bar{\nabla}\hat{F}(\bar{W}_{\neg i}^{(t)})$, we have

$$\begin{aligned}
&\|\bar{w}^{(t+1)} - \bar{w}_{\neg i}^{(t+1)}\| \\
&\quad= \left\|\bar{w}^{(t)} - \eta\bar{\nabla}\hat{F}(W^{(t)}) - (\bar{w}_{\neg i}^{(t)} - \eta\bar{\nabla}\hat{F}(W_{\neg i}^{(t)}))\right\| \\
&\quad= \Big\|\bar{w}^{(t)} - \eta\bar{\nabla}\hat{F}(\bar{W}^{(t)}) + \eta(\bar{\nabla}\hat{F}(W^{(t)}) - \bar{\nabla}\hat{F}(\bar{W}^{(t)})) \\
&\qquad\qquad - (\bar{w}_{\neg i}^{(t)} - \eta\nabla\hat{F}(\bar{W}_{\neg i}^{(t)})) + \eta\bar{\nabla}\hat{F}(W_{\neg i}^{(t)}) - \eta\nabla\hat{F}(\bar{W}_{\neg i}^{(t)})\Big\| \\
&\quad\le \left\|\bar{w}^{(t)} - \eta\bar{\nabla}\hat{F}(\bar{W}^{(t)}) - (\bar{w}_{\neg i}^{(t)} - \eta\bar{\nabla}\hat{F}(\bar{W}_{\neg i}^{(t)}))\right\| \\
&\qquad\qquad + \eta\left\|\bar{\nabla}\hat{F}(W_{\neg i}^{(t)}) - \nabla\hat{F}(\bar{W}_{\neg i}^{(t)})\right\| + \eta\left\|\bar{\nabla}\hat{F}(W^{(t)}) - \nabla\hat{F}(\bar{W}^{(t)})\right\|.
\end{aligned}$$

For the last term, using smoothness, we can write

$$\begin{aligned}\left\|\bar{\nabla}\hat{F}(W^{(t)}) - \nabla\hat{F}(\bar{W}^{(t)})\right\| &= \frac{1}{N}\left\|\sum_{\ell=1}^{N}\nabla\hat{F}_\ell(w_\ell^{(t)}) - \nabla\hat{F}_\ell(\bar{w}^{(t)})\right\| \\ &\leq \frac{1}{N}\sum_{\ell=1}^{N}\left\|\nabla\hat{F}_\ell(w_\ell^{(t)}) - \nabla\hat{F}_\ell(\bar{w}^{(t)})\right\| \\ &\leq \frac{L}{N}\sum_{\ell=1}^{N}\|w_\ell^{(t)} - \bar{w}^{(t)}\| \\ &\leq \frac{L}{\sqrt{N}}(\sum_{\ell=1}^{N}\|w_\ell^{(t)} - \bar{w}^{(t)}\|^2)^{1/2} = \frac{L}{\sqrt{N}}\|W^{(t)} - \bar{W}^{(t)}\|_F.\end{aligned}$$

The second term is upper-bounded similarly. Using these bounds, splitting the gradient $\nabla\hat{F}(\bar{W}^{(t)}) = \frac{1}{n}\sum_{i'\neq i} f(\bar{w}^{(t)}, x_{i'}) + \frac{1}{n}f(\bar{w}^{(t)}, x_i)$, using smoothness and convexity $\|w + \eta\nabla f(w) - v - \eta\nabla f(v)\| \leq \|w - v\|$ for $\eta \leq 2/L$ [155] and employing Assumption 6.2.4 we can write

$$\begin{aligned}\|\bar{w}^{(t+1)} - \bar{w}_{\neg i}^{(t+1)}\| &\leq \|\bar{w}^{(t)} - \eta\nabla\hat{F}(\bar{W}^{(t)}) - (\bar{w}_{\neg i}^{(t)} - \eta\nabla\hat{F}(\bar{W}_{\neg i}^{(t)}))\| \\ &\quad + \frac{\eta L}{\sqrt{N}}\|W_{\neg i}^{(t)} - \bar{W}_{\neg i}^{(t)}\|_F + \frac{\eta L}{\sqrt{N}}\|W^{(t)} - \bar{W}^{(t)}\|_F \\ &\leq \frac{1}{n}\sum_{i'\neq i}\|\bar{w}^{(t)} - \eta\nabla f(\bar{w}^{(t)}, x_{i'}) - \bar{w}_{\neg i}^{(t)} + \eta\nabla f(\bar{w}_{\neg i}^{(t)}, x_{i'})\| + \frac{1}{n}\|\bar{w}^{(t)} - \eta\nabla f(\bar{w}^{(t)}, x_i) - \bar{w}_{\neg i}^{(t)}\| \\ &\quad + \frac{\eta L}{\sqrt{N}}\|W_{\neg i}^{(t)} - \bar{W}_{\neg i}^{(t)}\|_F + \frac{\eta L}{\sqrt{N}}\|W^{(t)} - \bar{W}^{(t)}\|_F \\ &\leq \|\bar{w}^{(t)} - \bar{w}_{\neg i}^{(t)}\| + \frac{\eta}{n}\|\nabla f(\bar{w}^{(t)}, x_i)\| + \frac{\eta L}{\sqrt{N}}\|W_{\neg i}^{(t)} - \bar{W}_{\neg i}^{(t)}\|_F + \frac{\eta L}{\sqrt{N}}\|W^{(t)} - \bar{W}^{(t)}\|_F \\ &\leq \|\bar{w}^{(t)} - \bar{w}_{\neg i}^{(t)}\| + \frac{c\eta}{n}\,(f(\bar{w}^{(t)}, x_i))^\alpha + \frac{\eta L}{\sqrt{N}}\|W_{\neg i}^{(t)} - \bar{W}_{\neg i}^{(t)}\|_F + \frac{\eta L}{\sqrt{N}}\|W^{(t)} - \bar{W}^{(t)}\|_F.\end{aligned}$$

By summing over $t \in [T]$,

$$\|\bar{w}^{(T+1)} - \bar{w}_{\neg i}^{(T+1)}\| \leq \frac{c\eta}{n} \sum_{t=1}^{T} (f(\bar{w}^{(t)}, x_i))^{\alpha} + \frac{\eta L}{\sqrt{N}} \sum_{t=1}^{T} \|W_{\neg i}^{(t)} - \bar{W}_{\neg i}^{(t)}\|_F + \frac{\eta L}{\sqrt{N}} \sum_{t=1}^{T} \|W^{(t)} - \bar{W}^{(t)}\|_F.$$

We define for the ease of notation the following two consensus terms,

$$e^{(T)} := \Big(\sum_{t=1}^{T} \|W^{(t)} - \bar{W}^{(t)}\|_F \Big)^2 \qquad \text{and} \qquad e_{\neg i}^{(T)} := \Big(\sum_{t=1}^{T} \|W_{\neg i}^{(t)} - \bar{W}_{\neg i}^{(t)}\|_F \Big)^2. \tag{6.11}$$

Thus the bound for the squared term can be written as follows

$$\|\bar{w}^{(T+1)} - \bar{w}_{\neg i}^{(T+1)}\|^2 \leq \frac{3c^2\eta^2}{n^2} \left(\sum_{t=1}^{T} \big(f(\bar{w}^{(t)}, x_i) \big)^{\alpha} \right)^2 + \frac{3\eta^2 L^2}{N} e^{(T)} + \frac{3\eta^2 L^2}{N} e_{\neg i}^{(T)}.$$

By averaging over $i \in [n]$ and noting that $\alpha \in [1/2, 1]$ so that x^{α} is concave, we conclude that

$$\begin{aligned} \frac{1}{n} \sum_{i=1}^{n} \|\bar{w}^{(T+1)} - \bar{w}_{\neg i}^{(T+1)}\|^2 &\leq \frac{3c^2\eta^2}{n^3} \sum_{i=1}^{n} \left(\sum_{t=1}^{T} \big(f(\bar{w}^{(t)}, x_i) \big)^{\alpha} \right)^2 + \frac{3\eta^2 L^2}{Nn} \sum_{i=1}^{n} e_{\neg i}^{(T)} + \frac{3\eta^2 L^2}{N} e^{(T)} \\ &\leq \frac{3c^2\eta^2 T^2}{n^3} \sum_{i=1}^{n} \left(\frac{1}{T} \sum_{t=1}^{T} f(\bar{w}^{(t)}, x_i) \right)^{2\alpha} + \frac{3\eta^2 L^2}{Nn} \sum_{i=1}^{n} e_{\neg i}^{(T)} + \frac{3\eta^2 L^2}{N} e^{(T)} \\ &\leq \frac{3c^2\eta^2 T^{2(1-\alpha)}}{n^{3-2\alpha}} \left(\sum_{t=1}^{T} \hat{F}(\bar{w}^{(t)}) \right)^{2\alpha} + \frac{3\eta^2 L^2}{Nn} \sum_{i=1}^{n} e_{\neg i}^{(T)} + \frac{3\eta^2 L^2}{N} e^{(T)}. \end{aligned}$$

Thus we have for iteration T:

$$\frac{1}{n} \sum_{i=1}^{n} \|\bar{w}^{(T)} - \bar{w}_{\neg i}^{(T)}\|^2 \leq \frac{3c^2\eta^2 T^2}{n^{3-2\alpha}} \left(\frac{1}{T} \sum_{t=1}^{T} \hat{F}(\bar{w}^{(t)}) \right)^{2\alpha} + \frac{3\eta^2 L^2}{Nn} \sum_{i=1}^{n} e_{\neg i}^{(T)} + \frac{3\eta^2 L^2}{N} e^{(T)}. \tag{6.12}$$

Next we use [152, Lemma 7] (see also [14, Theorem 2]), which states that for the L-smooth loss f, the test error of the output w of an algorithm taking as input a dataset $(x_1, \ldots, x_n)$ size n, satisfies the following,

$$\mathbb{E}[F(w)] \leq 4\mathbb{E}[\hat{F}(w)] + \frac{3L^2}{n}\sum_{i=1}^{n}\mathbb{E}[\|w - w_{\neg i}\|^2],$$

where expectations are taken over the training set $(x_1, x_2, \cdots, x_n)$. We replace w with $\bar{w}^{(T)}$ and by using (6.12) (which we can do because it holds true for all datasets since Assumptions 6.2.1-6.2.4 hold for every sample x in the distribution),

$$\begin{aligned}\mathbb{E}[F(\bar{w}^{(T)})] &\leq 4\mathbb{E}[\hat{F}(\bar{w}^{(T)})] + \frac{3L^2}{n}\sum_{i=1}^{n}\mathbb{E}[\|\bar{w}^{(T)} - \bar{w}_{\neg i}^{(T)}\|^2] \\ &\leq 4\mathbb{E}[\hat{F}(\bar{w}^{(T)})] + \frac{9L^2c^2\eta^2T^2}{n^{3-2\alpha}}\mathbb{E}[(\frac{1}{T}\sum_{t=1}^{T}\hat{F}(\bar{w}^{(t)}))^{2\alpha}] \\ &\quad + \frac{9L^4\eta^2}{Nn}\sum_{i=1}^{n}\mathbb{E}[e_{\neg i}^{(T)}] + \frac{9L^4\eta^2}{N}\mathbb{E}[e^{(T)}].\end{aligned}$$

This leads to (6.10) and completes the proof.

Finally, we explain the informal version of the lemma presented in the main body (Lemma 6.2.1). Compared to the bound in Eq. (6.10), the informal Lemma 6.2.1 combines the consensus-error term $e^{(T)}$ with the average leave-one-out consensus-error term $\frac{1}{n}\sum_{i\in[n]} e_{\neg i}^{(T)}$ (recall the definitions in (6.11)). It is convenient doing that for the following reason. To apply Lemma 6.5.1, we need upper bounds on $e^{(T)}$ and $e_{\neg i}^{(T)}$ (for specific assumptions on the function class that is optimized). We do this in the section that follows. It turns out that the bounds we obtain for the consensus-error term $e^{(T)}$ also holds for the leave-one-out consensus error terms $e_{\neg i}^{(T)}, i \leq [n]$. The reason for that is that our bounds are not affected by the sample-size, but rather they depend crucially only on the smoothness parameter of the train loss. It is easy to see that the smoothness parameter

of the leave-one-out train loss $\frac{1}{n}\sum_{i'\neq i} f(w, x_{i'})$ is upper-bounded by the smoothness parameter of $\frac{1}{n}\sum_{i'} f(w, x_{i'})$. See Remark 6.5.1 for more details.

6.5.2 Proofs for Section 6.2.1

Lemma 6.5.2 (Recursions for the consensus error)**.** *Let the step-size $\eta \leq (1-\lambda)/4L$ where $\lambda := \max((|\lambda_2(A)|, |\lambda_N(A)|))^2$. The consensus error of DGD under Assumptions 6.2.1,6.2.3 satisfies the following:*

$$\|W^{(t)} - \bar{W}^{(t)}\|_F^2 < \alpha_1 \|W^{(t-1)} - \bar{W}^{(t-1)}\|_F^2 + \alpha_2 N \eta^2 L^2 \hat{F}(\bar{w}^{(t-1)}), \tag{6.13}$$

where $\alpha_1 := \frac{3+\lambda}{4}, \alpha_2 := 4(\frac{2}{1-\lambda} - 1)$.

Proof: Denoting $A^\infty := \lim_{t\to\infty} A^t = \frac{1}{N}\mathbf{1}\mathbf{1}^T$, it holds by Assumption 6.2.1,

$$\begin{aligned}\|AW - \bar{W}\|_F^2 = \|(A - A^\infty)(W - \bar{W})\|_F^2 &= \sum_{i=1}^N \|(A - A^\infty)(W_i - \bar{W}_i)\|^2 \\ &\leq \sum_{i=1}^N \|A - A^\infty\|^2 \|W_i - \bar{W}_i\|^2 \leq \max(\lambda_2^2(A), \lambda_N^2(A)) \cdot \|W - \bar{W}\|_F^2, \end{aligned} \tag{6.14}$$

where W_i is the i th column of W. By Assumption 6.2.1, $\lambda = \max((|\lambda_2(A)|, |\lambda_N(A)|))^2 < 1$. For the consensus error, we can write,

$$\begin{aligned}\|W^{(t)} - \bar{W}^{(t)}\|_F^2 &= \|W^{(t)} - \bar{W}^{(t-1)} - \bar{W}^{(t)} + \bar{W}^{(t-1)}\|_F^2 \\ &\leq \|W^{(t)} - \bar{W}^{(t-1)}\|_F^2 \\ &= \|AW^{(t-1)} - \eta \nabla \hat{F}(W^{(t-1)}) - \bar{W}^{(t-1)}\|_F^2 \\ &\leq (1+\beta)\|AW^{(t-1)} - \bar{W}^{(t-1)}\|_F^2 + (1+\beta^{-1})\eta^2 \|\nabla \hat{F}(W^{(t-1)})\|_F^2,\end{aligned}$$

where the second step is due to $\|X - \bar{X}\|_F \leq \|X\|_F$ [169, 170]. The last line holds for any

$\beta > 0$, due to $\|a+b\|^2 \leq (1+\beta^{-1})\|a\|^2 + (1+\beta^{-1})\|b\|^2$.

Based on this inequality and by noting (6.14) and using the $L-$smoothness assumption, we can deduce that,

$$\begin{aligned}
\|W^{(t)} - \bar{W}^{(t)}\|_F^2 &\leq (1+\beta)\lambda \|W^{(t-1)} - \bar{W}^{(t-1)}\|_F^2 + (1+\beta^{-1})\eta^2 \|\nabla \hat{F}(W^{(t-1)})\|_F^2 \\
&\leq (1+\beta)\lambda \|W^{(t-1)} - \bar{W}^{(t-1)}\|_F^2 + 2(1+\beta^{-1})\eta^2 \|\nabla \hat{F}(W^{(t-1)}) - \nabla \hat{F}(\bar{W}^{(t-1)})\|_F^2 \\
&\qquad + 2(1+\beta^{-1})\eta^2 \|\nabla \hat{F}(\bar{W}^{(t-1)})\|_F^2 \qquad (6.15) \\
&\leq (1+\beta)\lambda \|W^{(t-1)} - \bar{W}^{(t-1)}\|_F^2 + 2(1+\beta^{-1})\eta^2 L^2 \|W^{(t-1)} - \bar{W}^{(t-1)}\|_F^2 \\
&\qquad + 4(1+\beta^{-1})\eta^2 LN \hat{F}(\bar{w}_{t-1}),
\end{aligned}$$

where the last step is due to $L-$smoothness and the non-negativity of $\hat{F}_\ell$, i.e.

$$\|\nabla \hat{F}(\bar{W}^{(t-1)})\|_F^2 = \sum_{\ell=1}^N \|\nabla \hat{F}_\ell(\bar{w}^{(t-1)})\|^2 \leq 2L \sum_{\ell=1}^N (\hat{F}_\ell(\bar{w}^{(t-1)}) - \hat{F}_\ell^\star) \leq 2LN \hat{F}(\bar{w}^{(t-1)}).$$

Thus,

$$\begin{aligned}
\|W^{(t)} - \bar{W}^{(t)}\|_F^2 &< ((1+\beta)\lambda + 2(1+\beta^{-1})\eta^2 L^2) \|W^{(t-1)} - \bar{W}^{(t-1)}\|_F^2 \\
&\qquad + 4(1+\beta^{-1})\eta^2 LN\, \hat{F}(\bar{w}^{(t-1)}).
\end{aligned}$$

Next, choose $\beta = (1-\lambda)/(2\lambda)$. Then, it follows from the assumption $\eta \leq (1-\lambda)/4L$ that

$$\begin{aligned}
(1+\beta)\lambda + 2(1+\beta^{-1})\eta^2 L^2 &< (3+\lambda)/4 = \alpha_1, \\
4(1+\beta^{-1}) &< 4(2/(1-\lambda) - 1) = \alpha_2.
\end{aligned}$$

This concludes the lemma.

By telescoping summation over the iterates $t = 1, \cdots, T$ of the consensus error in

Eq.(6.13), we end up with the consensus error at iteration T. The final expression is stated in the next lemma.

Lemma 6.5.3. *Under the assumptions of Lemma 6.5.2, it holds for $T > 1$ that,*

$$\|W^{(T)} - \bar{W}^{(T)}\|_F^2 < \alpha_1^{T-1}\|W_1 - \bar{W}_1\|_F^2 + (\alpha_2\eta^2 LN)\sum_{t=1}^{T-1}\alpha_1^{t-1}\hat{F}(\bar{w}^{(T-t)}).$$

Lemma 6.5.4. *Under the assumptions of Lemma 6.5.2 and the zero initialization assumption for all agents, the average consensus error satisfies,*

$$\frac{1}{NT}\sum_{t=1}^{T}\|W^{(t)} - \bar{W}^{(t)}\|_F^2 \leq \frac{\alpha_2\eta^2 L}{(1-\alpha_1)T}\sum_{t=1}^{T-1}\hat{F}(\bar{w}^{(t)}).$$

Proof: By Lemma 6.5.3 and the zero initialization and non-negativity assumptions, we have

$$\begin{aligned}\frac{1}{NT}\sum_{t=1}^{T}\|W^{(t)} - \bar{W}^{(t)}\|_F^2 &\leq \frac{\alpha_2\eta^2 L}{T}\sum_{t=2}^{T}\sum_{\tau=1}^{t-1}\alpha_1^{\tau-1}\hat{F}(\bar{w}^{(t-\tau)}) \leq \frac{\alpha_2\eta^2 L}{T}\sum_{\tau=1}^{T-1}\alpha_1^{\tau-1}\sum_{t=1}^{T-\tau}\hat{F}(\bar{w}^{(t)}) \\ &\leq \frac{\alpha_2\eta^2 L}{T}\sum_{\tau=1}^{T-1}\alpha_1^{\tau-1}\sum_{t=1}^{T-1}\hat{F}(\bar{w}^{(t)}) \leq \frac{\alpha_2\eta^2 L}{(1-\alpha_1)T}\sum_{t=1}^{T-1}\hat{F}(\bar{w}^{(t)}).\end{aligned}$$

Lemma 6.5.5. *Under Assumptions 6.2.2,6.2.3 and for all $w \in \mathbb{R}^d$, the DGD updates satisfy the following recursions:*

$$\frac{2\eta - 4L\eta^2}{T}\sum_{t=1}^{T-1}\hat{F}(\bar{w}^{(t)}) \leq \frac{\|\bar{w}^{(1)} - w\|^2}{T} + 2\eta\hat{F}(w) + \frac{2L^2\eta^2 + \eta L}{NT}\sum_{t=1}^{T-1}\|W^{(t)} - \bar{W}^{(t)}\|_F^2$$

Proof: We start by upper bounding the following quantity:

$$\begin{aligned}\|\bar{w}^{(t+1)} - w\|^2 &= \|\bar{w}^{(t)} - \eta\bar{\nabla}\hat{F}(W^{(t)}) - w\|^2 \\ &= \|\bar{w}^{(t)} - w\|^2 + \eta^2\|\bar{\nabla}\hat{F}(W^{(t)})\|^2 - 2\eta\langle\bar{w}^{(t)} - w, \bar{\nabla}\hat{F}(W^{(t)})\rangle\end{aligned}$$

For the second term above, using $L-$smoothness and non-negativity of the loss, we obtain:

$$\begin{aligned}\|\bar{\nabla}\hat{F}(W^{(t)})\|^2 &= \|\bar{\nabla}\hat{F}(W^{(t)}) - \nabla\hat{F}(\bar{w}^{(t)}) + \nabla\hat{F}(\bar{w}^{(t)})\|^2 \\ &\leq 2\|\bar{\nabla}\hat{F}(W^{(t)}) - \nabla\hat{F}(\bar{w}^{(t)})\|^2 + 2\|\nabla\hat{F}(\bar{w}^{(t)})\|^2 \\ &\leq \frac{2L^2}{N}\sum_{i=1}^{N}\|w_\ell^{(t)} - \bar{w}^{(t)}\|^2 + 4L\hat{F}(\bar{w}^{(t)}).\end{aligned}$$

For the third term, by using $L-$smoothness and convexity properties we can write,

$$\begin{aligned}\langle \bar{w}^{(t)} - w, \bar{\nabla}\hat{F}(W^{(t)})\rangle &= \frac{1}{N}\sum_{\ell=1}^{N}\langle \bar{w}^{(t)} - w, \nabla\hat{F}_\ell(w_\ell^{(t)})\rangle \\ &= \frac{1}{N}\sum_{\ell=1}^{N}\langle \bar{w}^{(t)} - w_\ell^{(t)}, \nabla\hat{F}_\ell(w_\ell^{(t)})\rangle + \frac{1}{n}\sum_{\ell=1}^{N}\langle w_\ell^{(t)} - w, \nabla\hat{F}_\ell(w_\ell^{(t)})\rangle \\ &\geq \frac{1}{N}\sum_{\ell=1}^{N}(\hat{F}_\ell(\bar{w}^{(t)}) - \hat{F}_\ell(w_\ell^{(t)})) - \frac{L}{2}\|w_\ell^{(t)} - \bar{w}^{(t)}\|^2 + \frac{1}{N}\sum_{\ell=1}^{N}(\hat{F}_\ell(w_\ell^{(t)}) - \hat{F}_\ell(w)) \\ &= \hat{F}(\bar{w}^{(t)}) - \hat{F}(w) - \frac{L}{2N}\|W^{(t)} - \bar{W}^{(t)}\|_F^2.\end{aligned}$$

Combining these inequalities we derive the following:

$$\begin{aligned}\|\bar{w}^{(t+1)} - w\|^2 \leq \|\bar{w}^{(t)} - w\|^2 &+ \eta^2\Big(2L^2\|W^{(t)} - \bar{W}^{(t)}\|_F^2/N + 4L\hat{F}(\bar{w}^{(t)})\Big) \\ &- 2\eta\Big(\hat{F}(\bar{w}^{(t)}) - \hat{F}(w) - \frac{L}{2}\|W^{(t)} - \bar{W}^{(t)}\|_F^2/N\Big).\end{aligned}$$

Summing these equations for $t = 1, 2, ..., T-1$,

$$\begin{aligned}\|\bar{w}^{(T)} - w\|^2 \leq \|\bar{w}^{(1)} - w\|^2 &+ \sum_{t=1}^{T-1}\frac{2L^2\eta^2 + \eta L}{N}\|W^{(t)} - \bar{W}^{(t)}\|_F^2 \\ &+ \sum_{t=1}^{T-1}(4L\eta^2 - 2\eta)\hat{F}(\bar{w}^{(t)}) + 2\sum_{t=1}^{T-1}\eta\hat{F}(w).\end{aligned}$$

We conclude that,

$$\frac{2\eta - 4L\eta^2}{T}\sum_{t=1}^{T-1}\hat{F}(\bar{w}^{(t)}) \leq \frac{\|\bar{w}^{(1)} - w\|^2}{T} + \frac{2L^2\eta^2 + \eta L}{NT}\sum_{t=1}^{T-1}\|W^{(t)} - \bar{W}^{(t)}\|_F^2 + 2\eta\hat{F}(w).$$

6.5.2.1 Proof of Lemma 6.2.2

Lemma 6.5.6 (Restatement of Lemma 6.2.2)**.** *Under Assumptions 6.2.1-6.2.4 and zero initialization, for any $w \in \mathbb{R}^d$ and for a fixed step-size $\eta < \min\{\frac{1-\alpha_1}{L}, \frac{1}{L}\sqrt{\frac{1-\alpha_1}{2\alpha_2}}\}$, where $\alpha_1 \in (3/4, 1), \alpha_2 > 4$ are parameters that depend only on the mixing matrix, the following holds for the train loss and consensus error of DGD:*

$$\frac{1}{T}\sum_{t=1}^{T}\hat{F}(\bar{w}^{(t)}) \leq \frac{2\|w\|^2}{\eta T} + 4\hat{F}(w), \tag{6.16}$$

$$\frac{1}{NT}\sum_{t=1}^{T}\|W^{(t)} - \bar{W}^{(t)}\|_F^2 \leq \frac{\alpha_2\eta^2 L}{(1-\alpha_1)}\left(\frac{2\|w\|^2}{\eta T} + 4\hat{F}(w)\right). \tag{6.17}$$

Proof: Recalling the initialization $w_\ell^{(1)} = 0 \Rightarrow \bar{w}^{(1)} = 0$ and using $\eta < 1/(4L)$, we deduce from Lemma 6.5.5 that,

$$\frac{1}{T}\sum_{t=1}^{T-1}\hat{F}(\bar{w}^{(t)}) \leq \frac{\|w\|^2}{\eta T} + 2\hat{F}(w) + \frac{2L^2\eta + L}{NT}\sum_{t=1}^{T-1}\|W^{(t)} - \bar{W}^{(t)}\|_F^2$$

By Lemma 6.5.4,

$$\begin{aligned}\frac{1}{T}\sum_{t=1}^{T-1}\hat{F}(\bar{w}^{(t)}) &\leq \frac{\|w\|^2}{\eta T} + 2\hat{F}(w) + \frac{(2L^2\eta + L)\alpha_2\eta^2 L}{T(1-\alpha_1)}\sum_{t=1}^{T-1}\hat{F}(\bar{w}^{(t)}) \\ &\leq \frac{\|w\|^2}{\eta T} + 2\hat{F}(w) + \frac{1}{2T}\sum_{t=1}^{T-1}\hat{F}(\bar{w}^{(t)}).\end{aligned} \tag{6.18}$$

where the condition on η on the lemma's statement ensures that $(2L^2\eta+L)\alpha_2\eta^2 L/(1-\alpha_1) < 1/2$. This gives the statement of the lemma for the training loss in Eq.(6.16). Appealing

again to Lemma 6.5.4 for the consensus error yields (6.17).

Remark 6.5.1 (Bounds for leave-one-out consensus error). The bound in Eq. (6.17) also applies to the leave-one-out consensus-error term $\frac{1}{T}\sum_{t=1}^{T}\|W_{\neg i}^{(t)} - \bar{W}_{\neg i}^{(t)}\|_F^2$. To see this starting from Lemma 6.5.4 note that we still have

$$\frac{1}{T}\sum_{t=1}^{T}\|W_{\neg i}^{(t)} - \bar{W}_{\neg i}^{(t)}\|_F^2 \leq \frac{\alpha_2\eta^2 LN}{(1-\alpha_1)T}\sum_{t=1}^{T-1}\hat{F}_{\neg i}(\bar{w}_{\neg i,t}), \tag{6.19}$$

where we denote the leave-one-out train loss $\hat{F}_{\neg i}(w) := \frac{1}{n}\sum_{i'\neq i} f(w, x_{i'})$. This is true because the smoothness parameter of $\hat{F}_{\neg i}(w)$ is $(1-1/n)L \leq L$. Moreover, applying Lemma 6.5.5 to the leave-one-out loss (and using again that it's smoothness parameter is upper bounded by L), we have for all w that

$$\frac{2\eta - 4L\eta^2}{T}\sum_{t=1}^{T-1}\hat{F}_{\neg i}(\bar{w}_{\neg i}^{(t)}) \leq \frac{\|\bar{w}_{\neg i}^{(1)} - w\|^2}{T} + 2\eta\hat{F}_{\neg i}(w) + \frac{2L^2\eta^2 + \eta L}{NT}\sum_{t=1}^{T-1}\|W_{\neg i}^{(t)} - \bar{W}_{\neg i}^{(t)}\|_F^2$$

But, from the initialization assumption $\bar{w}_{\neg i}^{(1)} = 0$ and also $\hat{F}_{\neg i}(w) \leq \hat{F}(w)$ since the functions are assumed non-negative. Hence, and also using (6.19), shows that

$$\frac{2\eta - 4L\eta^2}{T}\sum_{t=1}^{T-1}\hat{F}_{\neg i}(\bar{w}_{\neg i}^{(t)}) \leq \frac{\|w\|^2}{T} + 2\eta\hat{F}(w) + \frac{(2L^2\eta^2 + \eta L)\alpha_2\eta^2 L}{T(1-\alpha_1)}\sum_{t=1}^{T-1}\hat{F}_{\neg i}(\bar{w}_{\neg i,t}).$$

Note that after using $\eta < 1/(4L)$ this is exactly analogous to Eq. (6.18) for the train loss, which leads to the same bound $\hat{F}_{\neg i}(\bar{w}_{\neg i}^{(t)}) \leq \frac{2\|w\|^2}{\eta T} + 4\hat{F}(w)$ for the leave-one-out loss. Plugging this back to Eq. (6.19) shows that the bound in (6.17) also holds for the leave-one-out consensus term.

6.5.2.2 Proof of Theorem 6.2.1

We are ready to prove Theorem 6.2.1 by combining our results from Lemmas 6.2.2 and 6.2.1. We state the proof for general choice of step-size η. In particular, Theorem 6.2.1 follows by the next theorem after choosing $\eta = O(1/\sqrt{T})$.

Theorem 6.5.1 (Theorem 6.2.1 for general η)**.** *Consider DGD under Assumptions 6.2.1-6.2.5, and choose* $\eta < \min\{\frac{1-\alpha_1}{L}, \frac{1}{L}\sqrt{\frac{1-\alpha_1}{2\alpha_2}}\}$. *The following bound holds for the averaged test error of DGD with separable data up to iteration* T, *assuming* $\varepsilon \leq \rho(\varepsilon)^2/\eta T$,

$$\frac{1}{T}\sum_{t=1}^{T} F(\bar{w}^{(t)}) = O\left(\frac{\rho(\varepsilon)^2}{\eta T} + \frac{L^2 c^2 \rho(\varepsilon)^{4\alpha}}{n^{3-2\alpha}}(\eta T)^{2-2\alpha} + L^4 \rho(\varepsilon)^4 \eta^3 T\right).$$

Proof: *By Lemma 6.2.1,*

$$\mathbb{E}\Big[F(\bar{w}^{(t)})\Big] = O\Big(\mathbb{E}\Big[\hat{F}(\bar{w}^{(t)})\Big] + \frac{L^2 c^2 \eta^2 t^2}{n^{3-2\alpha}}\mathbb{E}\Big[(\frac{1}{t}\sum_{\tau=1}^{t}\hat{F}(\bar{w}^{(\tau)}))^{2\alpha}\Big] + \frac{L^4\eta^2}{N}\mathbb{E}\Big[(\sum_{\tau=1}^{t}\|W^{(\tau)} - \bar{W}^{(\tau)}\|)^2\Big]\Big).$$

Thus, by Lemma 6.2.2,

$$\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\Big[F(\bar{w}^{(t)})\Big] = O\left(\frac{\|w\|^2}{\eta T} + \hat{F}(w) + \frac{L^2c^2\eta^2}{n^{3-2\alpha}}\frac{1}{T}\sum_{t=1}^{T} t^2(\frac{\|w\|^2}{\eta t} + \hat{F}(w))^{2\alpha} + \frac{L^4\eta^4}{N}\frac{1}{T}\sum_{t=1}^{T} t^2(\frac{\|w\|^2}{\eta t} + \hat{F}(w))\right).$$

By Assumption 6.2.5 and assuming $\varepsilon \leq \rho(\varepsilon)^2/\eta T$, *the statement of the theorem follows.*

6.5.3 Proofs for Section 6.2.2

Lemma 6.5.7 (Iterates of consensus error)**.** *Consider DGD with the loss functions and mixing matrix satisfying Assumptions 6.2.1,6.2.2, 6.2.6 and Assumption 6.2.4 with* $\alpha = 1$ *and* $c = h$*. By choosing* $\eta_{t-1} \leq \frac{1-\lambda}{4hNM_{(t-1)}}$ *the consensus error at iteration* $t > 1$ *satisfies*

$$\left\|W^{(t)} - \bar{W}^{(t)}\right\|_F^2 < \beta_1 \left\|W^{(t-1)} - \bar{W}^{(t-1)}\right\|_F^2 + \beta_2 \eta_{t-1}^2 h^2 N^2 \hat{F}^2(\bar{w}^{(t-1)}) \tag{6.20}$$

where we define

$$\beta_1 := (3+\lambda)/4, \beta_2 := (4/(1-\lambda) - 2), \lambda := \max\{|\lambda_2(A)|^2, |\lambda_N(A)|^2\}$$

and

$$M_{(t-1)} := \max\{\hat{F}(W^{(t-1)}), \hat{F}(\bar{w}^{(t-1)})\}.$$

Proof: By Lemma 6.5.2 and the inequality (6.15), the consensus error satisfies for any $\beta > 0$,

$$\begin{aligned}\left\|W^{(t)} - \bar{W}^{(t)}\right\|_F^2 &< (1+\beta)\lambda \left\|W^{(t-1)} - \bar{W}^{(t-1)}\right\|_F^2 \\ &+ 2(1+\beta^{-1})\eta_{t-1}^2 \left\|\nabla \hat{F}(W^{(t-1)}) - \nabla \hat{F}(\bar{W}^{(t-1)})\right\|_F^2 + 2(1+\beta^{-1})\eta_{t-1}^2 \left\|\nabla \hat{F}(\bar{W}^{(t-1)})\right\|_F^2.\end{aligned} \tag{6.21}$$

For the second term in (6.21), we have the following chain of inequalities,

$$
\begin{aligned}
&\left\|\nabla \hat{F}(W^{(t-1)}) - \nabla \hat{F}(\bar{W}^{(t-1)})\right\|_F^2 \\
&\qquad = \sum_{\ell=1}^{N} \|\nabla \hat{F}_\ell(w_i^{(t-1)}) - \nabla \hat{F}_\ell(\bar{w}^{(t-1)})\|^2 \\
&\qquad \leq \sum_{\ell=1}^{N} \max_{v_\ell \in [w_\ell^{(t-1)}, \bar{w}^{(t-1)}]} \|\nabla^2 \hat{F}_\ell(v_\ell)\|^2 \|w_\ell^{(t-1)} - \bar{w}^{(t-1)}\|^2 \qquad (6.22) \\
&\qquad \leq h^2 \sum_{\ell=1}^{N} \max_{v_\ell \in [w_\ell^{(t-1)}, \bar{w}^{(t-1)}]} (\hat{F}_\ell(v_\ell))^2 \|w_\ell^{(t-1)} - \bar{w}^{(t-1)}\|^2 \qquad (6.23) \\
&\qquad = h^2 \sum_{\ell=1}^{N} \Big(\max_{v_\ell \in [w_\ell^{(t-1)}, \bar{w}^{(t-1)}]} \hat{F}_\ell(v_\ell) \Big)^2 \|w_\ell^{(t-1)} - \bar{w}^{(t-1)}\|^2 \\
&\qquad \leq h^2 \sum_{\ell=1}^{N} \max\{\hat{F}_\ell^2(w_\ell^{(t-1)}), \hat{F}_\ell^2(\bar{w}^{(t-1)})\} \|w_\ell^{(t-1)} - \bar{w}^{(t-1)}\|^2 \qquad (6.24) \\
&\qquad \leq h^2 \max\{\max_{k \leq N} \hat{F}_k^2(w_k^{(t-1)}), \max_{k \leq N} \hat{F}_k^2(\bar{w}^{(t-1)})\} \sum_{\ell=1}^{N} \|w_\ell^{(t-1)} - \bar{w}^{(t-1)}\|^2 \\
&\qquad \leq h^2 N^2 M_{(t-1)}^2 \left\|W^{(t-1)} - \bar{W}^{(t-1)}\right\|_F^2. \qquad (6.25)
\end{aligned}
$$

The Taylor's remainder theorem gives (6.22) and $v_i \in [w_i^{(t-1)}, \bar{w}^{(t-1)}]$ denotes a point that lies on the line connecting $w_i^{(t-1)}$ and $\bar{w}^{(t-1)}$. Also, (6.23) is valid due to the self-boundedness of the Hessian stated in Assumption 6.2.6. The inequality (6.24) follows by the assumption of convexity of $\hat{F}_i$, due to the fact that for a convex function $f : \mathbb{R}^d \to \mathbb{R}$ and any two points $w_1, w_2 \in \mathbb{R}^d$, it holds that $\max_{v \in [w_1, w_2]} f(v) \leq \max\{f(w_1), f(w_2)\}$. To derive (6.25), we used $\max_{i \leq N} \hat{F}_i(w_i) \leq N \hat{F}(W)$ and $\max_{i \leq N} \hat{F}_i(\bar{w}) \leq N \cdot \hat{F}(\bar{w})$, which hold since the loss functions are non-negative.

In order to derive an upper-bound on the last term in (6.21), we use Assumption 6.2.4

(with $\alpha = 1, c = h$):

$$\left\|\nabla \hat{F}(\bar{W}^{(t-1)})\right\|_F^2 = \sum_{\ell=1}^{N} \left\|\nabla \hat{F}_\ell(\bar{w}^{(t-1)})\right\|^2 \leq h^2 \sum_{\ell=1}^{N} (\hat{F}_\ell(\bar{w}^{(t-1)}))^2 \leq h^2 N^2 \hat{F}^2(\bar{w}^{(t-1)}).$$

Replacing the upper-bounds back in (6.21), we conclude

$$\begin{aligned}\left\|W^{(t)} - \bar{W}^{(t)}\right\|_F^2 < ((1+\beta)\lambda + 2(1+\beta^{-1})\eta_{t-1}^2 h^2 N^2 M_{(t-1)}^2)\left\|W^{(t-1)} - \bar{W}^{(t-1)}\right\|_F^2 \\ + 2(1+\beta^{-1})\eta_{t-1}^2 h^2 N^2 \hat{F}^2(\bar{w}^{(t-1)}).\end{aligned}$$

Choose $\beta = \frac{1-\lambda}{2\lambda}$. Then by lemma's assumption $\eta_{t-1} \leq \frac{1-\lambda}{4hNM_{(t-1)}}$, we can verify the following two inequalities:

$$\begin{aligned}&(1+\beta)\lambda + 2(1+\beta^{-1})\eta_{t-1}^2 h^2 N^2 M_{(t-1)}^2 \leq \frac{3+\lambda}{4}, \\ &2(1+1/\beta) \leq \frac{4}{1-\lambda} - 2.\end{aligned}$$

This concludes the proof.

By recursively evaluating (6.20), we obtain a bound on the consensus error at iteration T, which we present next.

Lemma 6.5.8 (Last iterate consensus error)**.** *Under the assumptions and notations of Lemma 6.5.7, the consensus error at iteration T satisfies*

$$\left\|W^{(T)} - \bar{W}^{(T)}\right\|_F^2 < \beta_1^{T-1}\left\|W^{(1)} - \bar{W}^{(1)}\right\|_F^2 + \beta_2 h^2 N^2 \sum_{t=1}^{T-1} \beta_1^{t-1} \eta_{T-t}^2 \hat{F}^2(\bar{w}_{T-t}).$$

The next lemma obtains a sandwich relation between $F(\bar{w}^{(T)})$ and $\hat{F}(W^{(T)})$. This is convenient as it allows replacing $M_{(t)} := \max(\hat{F}(W^{(t)}), \hat{F}(\bar{w}^{(t)}))$ by either of the two terms with only paying a constant factor of two. See also the remark after the statement

of the theorem.

Lemma 6.5.9. *Under the assumptions and notations of Lemma 6.5.7, with zero initialization* $W^{(1)} = \bar{W}^{(1)} = 0$ *and by choosing* $\eta_t \leq \frac{(1-\lambda)\sqrt{1-\beta_1}}{8h^2NM_{(t)}\sqrt{\beta_2}}$ *for* $t \in [T-1]$*, it holds at iteration* T *that*

$$\frac{1}{2}\hat{F}(\bar{w}^{(T)}) \leq \hat{F}(W^{(T)}) \leq 2\hat{F}(\bar{w}^{(T)}). \tag{6.26}$$

Proof: First, we prove $\hat{F}(W^{(T)}) \leq 2\hat{F}(\bar{w}^{(T)})$*. If* $\hat{F}(W^{(T)}) \leq \hat{F}(\bar{w}^{(T)})$*, there is nothing to prove. Thus, assume* $\hat{F}(W^{(T)}) \geq \hat{F}(\bar{w}^{(T)})$*. Then by applying Taylor's remainder theorem, the self-boundedness Assumption 6.2.4 with* $c = h, \alpha = 1$*, convexity of* $\hat{F}$*, Lemma 6.5.8 and the restriction on the step-size, in respective order, we have the following inequalities,*

$$\begin{aligned}
\hat{F}(W^{(T)}) &\leq |\hat{F}(W^{(T)}) - \hat{F}(\bar{w}^{(T)})| + \hat{F}(\bar{w}^{(T)}) \\
&\leq \max_{v \in [\bar{W}^{(T)}, W^{(T)}]} \|\nabla \hat{F}(v)\| \cdot \|W^{(T)} - \bar{W}^{(T)}\| + \hat{F}(\bar{w}^{(T)}) \\
&\leq h \cdot \max_{v \in [\bar{W}^{(T)}, W^{(T)}]} \hat{F}(v) \cdot \|W^{(T)} - \bar{W}^{(T)}\| + \hat{F}(\bar{w}^{(T)}) \\
&\leq h \cdot \max\{\hat{F}(W^{(T)}), \hat{F}(\bar{w}^{(T)})\} \cdot \|W^{(T)} - \bar{W}^{(T)}\| + \hat{F}(\bar{w}^{(T)}) \\
&\leq \hat{F}(W^{(T)}) \Big(\beta_2 h^4 N^2 \sum_{t=1}^{T-1} \beta_1^{t-1} \eta_{T-t}^2 \hat{F}^2(\bar{w}_{T-t})\Big)^{1/2} + \hat{F}(\bar{w}^{(T)}) \\
&\leq \frac{1}{2}\hat{F}(W^{(T)}) + \hat{F}(\bar{w}^{(T)}).
\end{aligned}$$

Thus $\hat{F}(W^T) \leq 2\hat{F}(\bar{w}^{(T)})$*. By exchanging* $W^{(T)}$ *and* $\bar{w}^{(T)}$ *and in a similar style we derive* $\hat{F}(W^T) \geq \frac{1}{2}\hat{F}(\bar{w}^{(T)})$*. This completes the proof of the lemma.*

Remark 6.5.2. Lemma 6.5.8 above requires tuning $\eta_t \propto 1/M_{(t)} := 1/\max(\hat{F}(W^{(t)}), \hat{F}(\bar{w}^{(t)}))$. Lemma 6.5.9 shows that abiding by this choice for $t = 1, \ldots, T-1$ and any $T > 1$ guar-

antees $\hat{F}(W^{(T)}) \leq 2\hat{F}(\bar{w}^{(T)})$. Hence, $M_{(T)} \geq 2\hat{F}(\bar{w}^{(T)})$. Since this holds for all T and at $t = 1, \hat{F}(W^{(1)}) = \hat{F}(\bar{w}^{(1)})$, it follows by recursion that Lemma 6.5.8 holds provided $\eta_t \propto 1/\hat{F}(\bar{w}^{(t)})$. We use observation in the proofs below.

We are ready to prove Theorem 6.2.2. First, we prove that DGD is a descent algorithm in the next lemma.

Lemma 6.5.10 (Descent lemma)**.** *Consider DGD under the assumptions and notations of Lemma 6.5.7. Moreover, let Assumption 6.2.7 hold, then by choosing* $\eta_t \leq \frac{\delta}{\hat{F}(\bar{w}^{(t)})}$ *for* $t \leq T$*, where*

$$\delta := 1 \Big/ \max\left\{\frac{4h^3N}{\tau^2}, h^2, \frac{6h^2\beta_2}{1-\beta_1}, \frac{4h^2\sqrt{\beta_2}}{\tau(1-\beta_1)}\right\}, \tag{6.27}$$

DGD is a descent algorithm, i.e., for all $T \geq 1$*.*

$$\hat{F}(\bar{w}^{(T+1)}) \leq \hat{F}(\bar{w}^{(T)}).$$

Proof: With the self-boundedness assumption on the Hessian (Assumption 6.2.6) and applying the Taylor's remainder theorem for step $T+1$ of DGD, we obtain the

following,

$$
\begin{aligned}
&\hat{F}(\bar{w}^{(T+1)}) \\
&\leq \hat{F}(\bar{w}^{(T)}) + \langle \nabla \hat{F}(\bar{w}^{(T)}), \bar{w}^{(t+1)} - \bar{w}^{(T)} \rangle + \frac{1}{2} \max_{v \in [\bar{w}^{(T)}, \bar{w}^{(T+1)}]} \|\nabla^2 \hat{F}(v)\| \|\bar{w}^{(T+1)} - \bar{w}^{(T)}\|^2 \\
&\leq \hat{F}(\bar{w}^{(T)}) - \eta_T \langle \nabla \hat{F}(\bar{w}^{(T)}), \bar{\nabla} \hat{F}(W^{(T)}) \rangle + \frac{\eta_T^2}{2} \max_{v \in [\bar{w}^{(T)}, \bar{w}^{(T+1)}]} \|\nabla^2 \hat{F}(v)\| \|\bar{\nabla} \hat{F}(W^{(T)})\|^2 \\
&\leq \hat{F}(\bar{w}^{(T)}) - \eta_T \langle \nabla \hat{F}(\bar{w}^{(T)}), \bar{\nabla} \hat{F}(W^{(T)}) \rangle + \frac{h\eta_T^2}{2} \max_{v \in [\bar{w}^{(T)}, \bar{w}^{(T+1)}]} \hat{F}(v) \|\bar{\nabla} \hat{F}(W^{(T)})\|^2 \\
&\leq \hat{F}(\bar{w}^{(T)}) - \eta_T \langle \nabla \hat{F}(\bar{w}^{(T)}), \bar{\nabla} \hat{F}(W^{(T)}) \rangle + \frac{h\eta_T^2}{2} \max\{\hat{F}(\bar{w}^{(T)}), \hat{F}(\bar{w}^{(T+1)})\} \|\bar{\nabla} \hat{F}(W^{(T)})\|^2,
\end{aligned}
\tag{6.28}
$$

where for the third step we used

$$
\|\nabla^2 \hat{F}(w)\| \leq \frac{1}{N} \sum_{\ell=1}^{N} \|\nabla^2 \hat{F}_\ell(w)\| \leq \frac{h}{N} \sum_{\ell=1}^{N} \hat{F}_\ell(w) = h\, \hat{F}(w).
$$

In the next step of the proof, we upper-bound the second and third terms in (6.28). For the second term, by noting that $2\langle a, b \rangle = \|a\|^2 + \|b\|^2 - \|a - b\|^2$, we can write

$$
\langle \nabla \hat{F}(\bar{w}^{(T)}), \bar{\nabla} \hat{F}(W^{(T)}) \rangle = \frac{1}{2} \|\nabla \hat{F}(\bar{w}^{(T)})\|^2 + \frac{1}{2} \|\bar{\nabla} \hat{F}(W^{(T)})\|^2 - \frac{1}{2} \|\nabla \hat{F}(\bar{w}^{(T)}) - \bar{\nabla} \hat{F}(W^{(T)})\|^2.
\tag{6.29}
$$

By recalling (6.25) and Lemma 6.5.8 (which we can apply because of Remark 6.5.2), we find an upper-bound the last term in (6.29) as follows,

$$
\begin{aligned}
\|\nabla \hat{F}(\bar{w}^{(T)}) - \bar{\nabla} \hat{F}(W^{(T)})\|^2 &= \frac{1}{N^2} \|\nabla \hat{F}(W^{(T)}) - \nabla \hat{F}(\bar{W}^{(T)})\|_F^2 \\
&\leq h^2 M_{(T)}^2 \|W^{(T)} - \bar{W}^{(T)}\|_F^2 \\
&\leq h^4 M_{(T)}^2 N^2 \beta_2 \sum_{t=1}^{T-1} \beta_1^{t-1} \eta_{T-t}^2 \hat{F}^2(\bar{w}_{T-t}).
\end{aligned}
\tag{6.30}
$$

Returning back to (6.28), thus far we have derived the following,

$$\hat{F}(\bar{w}^{(T+1)}) \leq \hat{F}(\bar{w}^{(T)}) - \frac{\eta_T}{2}\|\nabla \hat{F}(\bar{w}^{(T)})\|^2 - \frac{\eta_T}{2}\|\bar{\nabla}\hat{F}(W^{(T)})\|^2 + \frac{1}{2}\eta_T h^4 N^2 M_{(T)}^2 \beta_2 \sum_{t=1}^{T-1} \beta_1^{t-1}\eta_{T-t}^2 \hat{F}^2(\bar{w}_{T-t}) + \frac{h\eta_T^2}{2}\|\bar{\nabla}\hat{F}(W^{(T)})\|^2 \cdot \max\{\hat{F}(\bar{w}^{(T)}), \hat{F}(\bar{w}^{(T+1)})\}. \tag{6.31}$$

We aim to prove that $\hat{F}(\bar{w}^{(T+1)}) \leq \hat{F}(\bar{w}^{(T)})$ for all $T \geq 1$. If $\hat{F}(\bar{w}^{(T+1)}) > \hat{F}(\bar{w}^{(T)})$, applying (6.31) with the assumption $\eta_t < \frac{\delta}{\hat{F}(\bar{w}^{(t)})}$ yields,

$$\hat{F}(\bar{w}^{(T+1)}) \leq \hat{F}(\bar{w}^{(T)}) - \frac{\eta_T}{2}\|\nabla \hat{F}(\bar{w}^{(T)})\|^2 + \frac{1}{2(1-\beta_1)}\eta_T \delta^2 h^4 N^2 M_{(T)}^2 \beta_2 + \frac{h\eta_T^2}{2}\|\bar{\nabla}\hat{F}(W^{(T)})\|^2 \cdot \hat{F}(\bar{w}^{(T+1)}). \tag{6.32}$$

Note that it holds due to (6.30) that,

$$\begin{aligned}\|\bar{\nabla}\hat{F}(W^{(T)})\|^2 &\leq 2\|\nabla \hat{F}(\bar{w}^{(T)})\|^2 + 2\|\bar{\nabla}\hat{F}(W^{(T)}) - \nabla \hat{F}(\bar{w}^{(T)})\|^2 \\ &\leq 2\|\nabla \hat{F}(\bar{w}^{(T)})\|^2 + \frac{2}{1-\beta_1}\delta^2 h^4 N^2 M_{(T)}^2 \beta_2.\end{aligned}$$

Replacing this in (6.32) and noting that $M_{(T)} \leq 2\hat{F}(\bar{w}^{(T)})$ by Lemma 6.5.9, we can simplify the inequality (6.32) as follows,

$$\begin{aligned}&\hat{F}(\bar{w}^{(T+1)}) \\ &\leq \hat{F}(\bar{w}^{(T)}) + \eta_T\|\nabla \hat{F}(\bar{w}^{(T)})\|^2(h\eta_T \hat{F}(\bar{w}^{(T+1)}) - \frac{1}{2}) \\ &\qquad + C' h^2 N^2 \delta^2 \eta_T \hat{F}^2(\bar{w}^{(T)})\,(1 + \eta_T \hat{F}(\bar{w}^{(T+1)})) \\ &\leq \hat{F}(\bar{w}^{(T)}) + \eta_T h^2 \hat{F}^2(\bar{w}^{(T)})((h\eta_T + \delta^2 \eta_T N^2 C')\hat{F}(\bar{w}^{(T+1)}) - \frac{\tau^2}{2h^2} + \delta^2 N^2 C'),\end{aligned}$$

where for the ease of notation we define $C' := 4h^2\beta_2(1-\beta_1)^{-1}$. Recalling $\eta_T < \frac{\delta}{\hat{F}(\bar{w}^{(T)})}$ and noting that by the assumption of the lemma $\delta \leq \frac{\tau^2}{4h^3}$, $\delta < \frac{1}{h^2}$ and $\delta < \frac{\tau}{2hN\sqrt{C'}}$ we conclude that,

$$\hat{F}(\bar{w}^{(T+1)}) \leq \hat{F}(\bar{w}^{(T)}) + \frac{\tau}{2h}\hat{F}(\bar{w}^{(T)})(\frac{\hat{F}(\bar{w}^{(T+1)})}{\hat{F}(\bar{w}^{(T)})} - 1).$$

Dividing both sides by $\hat{F}(\bar{w}^{(T)})$ leads to the contradiction due to the fact that $\tau \leq h$ and thus $\tau/2h < 1$. Thus $\hat{F}(\bar{w}^{(T+1)}) \leq \hat{F}(\bar{w}^{(T)})$. This completes the proof.

6.5.3.1 Proof of Theorem 6.2.2

Theorem 6.5.2 (Restatement of Theorem 6.2.2)**.** *Consider DGD with the loss functions and mixing matrix satisfying Assumptions 6.2.1, 6.2.2, 6.2.6, 6.2.7 and Assumption 6.2.4 with $\alpha = 1$ and $c = h$. Assume that the step-size satisfies $\eta < \frac{\delta}{\hat{F}(1)}$ for δ defined in (6.27). Also, recall positive constants β_1, β_2 depending only on the mixing matrix as defined in Lemma 6.5.7. Then DGD is a descent algorithm i.e, for all $T \geq 1$ it holds that*

$$\hat{F}(\bar{w}^{(T+1)}) \leq \hat{F}(\bar{w}^{(T)}).$$

Moreover, the train loss and the consensus error of DGD at iteration T satisfy the following for all $w \in \mathbb{R}^d$,

$$\hat{F}(\bar{w}^{(T)}) \leq 4\hat{F}(w) + \frac{2\|w\|^2}{\eta T}, \tag{6.33}$$

$$\frac{1}{N^2}\|W^{(T)} - \bar{W}^{(T)}\|_F^2 \leq \frac{8\beta_2 h^2}{1-\beta_1}(4\eta^2\hat{F}^2(w) + \frac{\|w\|^4}{T^2}). \tag{6.34}$$

Proof: First, we note that by Lemma 6.5.10, under the assumption $\eta_t \leq \delta/\hat{F}(\bar{w}^{(t)})$ for $t \leq T$, we have $\hat{F}(\bar{w}^{(T+1)}) \leq \hat{F}(\bar{w}^{(T)})$. Thus fixing $\eta \leq \delta/\hat{F}(\bar{w}^{(1)})$, ensures that

$\hat{F}(\bar{w}^{(T+1)}) \leq \hat{F}(\bar{w}^{(T)})$, for all T.

Next, we derive the train loss and consensus error under the assumptions of the theorem. Start with,

$$\|\bar{w}^{(t+1)} - w\|^2 = \|\bar{w}^{(t)} - w\|^2 + \eta^2 \|\bar{\nabla}\hat{F}(W^{(t)})\|^2 - 2\eta\langle \bar{w}^{(t)} - w, \bar{\nabla}\hat{F}(W^{(t)})\rangle. \tag{6.35}$$

For the second term, by self-boundedness of gradient, we can write,

$$\|\bar{\nabla}\hat{F}(W^{(t)})\| = \frac{1}{n}\|\sum_{\ell=1}^{n} \nabla\hat{F}_\ell(w_\ell^{(t)})\| \leq \frac{h}{n}\sum_{\ell=1}^{n} \hat{F}_\ell(w_\ell^{(t)}) = h\hat{F}(W^{(t)}).$$

For the third term in (6.35), we have,

$$\begin{aligned}
-\langle \bar{w}^{(t)} - w, \bar{\nabla}\hat{F}(W^{(t)})\rangle &= -\frac{1}{N}\sum_{\ell=1}^{N}\langle \bar{w}^{(t)} - w, \nabla\hat{F}_\ell(w_\ell^{(t)})\rangle \\
&= -\frac{1}{N}\sum_{\ell=1}^{N}\langle \bar{w}^{(t)} - w_\ell^{(t)}, \nabla\hat{F}_\ell(w_\ell^{(t)})\rangle - \frac{1}{N}\sum_{\ell=1}^{N}\langle w_\ell^{(t)} - w, \nabla\hat{F}_\ell(w_\ell^{(t)})\rangle \\
&\leq \frac{1}{N}\sum_{\ell=1}^{N}\|w_\ell^{(t)} - \bar{w}^{(t)}\|\|\nabla\hat{F}_i(w_\ell^{(t)})\| - \frac{1}{N}\sum_{\ell=1}^{N}\langle w_\ell^{(t)} - w, \nabla\hat{F}_\ell(w_\ell^{(t)})\rangle \\
&\leq \frac{1}{N}\sum_{\ell=1}^{N}\|w_\ell^{(t)} - \bar{w}^{(t)}\|\|\nabla\hat{F}_\ell(w_\ell^{(t)})\| + \frac{1}{N}\sum_{\ell=1}^{N}(\hat{F}_\ell(w) - \hat{F}_\ell(w_\ell^{(t)})) \qquad (6.36) \\
&\leq \frac{h}{N}\sum_{\ell=1}^{N}\hat{F}_\ell(w_\ell^{(t)})\|w_\ell^{(t)} - \bar{w}^{(t)}\| + \frac{1}{N}\sum_{\ell=1}^{N}(\hat{F}_\ell(w) - \hat{F}_\ell(w_\ell^{(t)})) \qquad (6.37) \\
&\leq h\hat{F}(W^{(t)})\|W^{(t)} - \bar{W}^{(t)}\|_F + \hat{F}(w) - \hat{F}(W^{(t)}).
\end{aligned}$$

Here (6.36) follows by convexity of $\hat{F}_i$, and (6.37) follows by the assumption on self-boundedness of the gradient.

Thus, the inequality (6.35) can be written as follows,

$$\|\bar{w}^{(t+1)} - w\|^2 \leq \|\bar{w}^{(t)} - w\|^2 + \eta^2 h^2 \hat{F}^2(W^{(t)}) + 2\eta h \hat{F}(W^{(t)}) \|W^{(t)} - \bar{W}^{(t)}\|_F + 2\eta \hat{F}(w) - 2\eta \hat{F}(W^{(t)}). \tag{6.38}$$

Moreover, by Lemma 6.5.8 and the assumption on η,

$$\|W^{(t)} - \bar{W}^{(t)}\|_F \leq (\beta_2 h^2 \sum_{t=1}^{T-1} \beta_1^{t-1} \eta_{T-t}^2 \hat{F}^2(\bar{w}_{T-t}))^{1/2} \leq \frac{1}{4h}$$

and

$$\eta \hat{F}(W^{(t)}) \leq \frac{1}{2h^2}.$$

Thus (6.38) changes into,

$$\|\bar{w}^{(t+1)} - w\|^2 \leq \|\bar{w}^{(t)} - w\|^2 - \eta \hat{F}(W^{(t)}) + 2\eta \hat{F}(w).$$

Telescoping sum leads to

$$\frac{1}{T} \sum_{t=1}^{T} \hat{F}(W^{(t)}) \leq 2\hat{F}(w) + \frac{\|\bar{w}^{(1)} - w\|^2}{\eta T}. \tag{6.39}$$

By Lemma 6.5.9, we have $\hat{F}(\bar{w}^{(t)}) \leq 2\hat{F}(W^{(t)})$. Finally, as we proved in the beginning, DGD is a descent algorithm, implying

$$\hat{F}(\bar{w}^{(T)}) \leq \frac{1}{T} \sum_{t=1}^{T} \hat{F}(\bar{w}^{(t)})$$

In view of (6.39), this yields the claim of the theorem for the train loss (6.33). Finally,

appealing to Lemma 6.5.8, gives (6.34). This completes the proof of the theorem.

6.5.4 Proofs for Section 6.2.3

Lemma 6.5.11 (Train loss under PL condition)**.** *Let Assumptions 6.2.1,6.2.3 and 6.2.8 hold, and let* $\eta \leq \min\{\frac{1}{\mu}, \sqrt{\frac{(1-\alpha_1)\mu}{4L^4\alpha_2}}, \frac{1}{L}\}$ *and* $\bar{\zeta} := \max\{\frac{1+\alpha_1}{2}, 1-\frac{\eta\mu}{2}\}$*, where* $\alpha_1 := \frac{3+\lambda}{4}, \alpha_2 := 4(\frac{2}{1-\lambda}-1)$ *same as in Lemma 6.5.2, Then for* $t \geq 1$

$$\hat{F}(\bar{w}^{(t)}) \leq \bar{\zeta}^{t-1}\hat{F}(\bar{w}^{(1)}). \tag{6.40}$$

Proof: By $L-$ smoothness we have

$$\begin{aligned}
\hat{F}(\bar{w}^{(t+1)}) &\leq \hat{F}(\bar{w}^{(t)}) - \eta\langle\nabla\hat{F}(\bar{w}^{(t)}), \bar{\nabla}\hat{F}(W^{(t)})\rangle + \frac{\eta^2 L}{2}\|\bar{\nabla}\hat{F}(W^{(t)})\|^2 \\
&= \hat{F}(\bar{w}^{(t)}) - \frac{\eta-\eta^2 L}{2}\|\bar{\nabla}\hat{F}(W^{(t)})\|^2 - \frac{\eta}{2}\|\nabla\hat{F}(\bar{w}^{(t)})\|^2 + \frac{\eta}{2}\|\bar{\nabla}\hat{F}(W^{(t)}) - \nabla\hat{F}(\bar{w}^{(t)})\|^2 \\
&\leq \hat{F}(\bar{w}^{(t)}) - \frac{\eta}{2}\|\nabla\hat{F}(\bar{w}^{(t)})\|^2 + \frac{\eta}{2}\|\bar{\nabla}\hat{F}(W^{(t)}) - \nabla\hat{F}(\bar{w}^{(t)})\|^2 \\
&\leq \hat{F}(\bar{w}^{(t)}) - \frac{\eta}{2}\|\nabla\hat{F}(\bar{w}^{(t)})\|^2 + \frac{\eta L^2}{N}\|W^{(t)} - \bar{W}^{(t)}\|_F^2.
\end{aligned}$$

By $\mu-$PL condition we have,

$$\hat{F}(\bar{w}^{(t+1)}) \leq (1-\eta\mu)\hat{F}(\bar{w}^{(t)}) + \frac{\eta L^2}{N}\|W^{(t)} - \bar{W}^{(t)}\|_F^2.$$

By Lemma 6.5.2,

$$\frac{1}{N}\|W^{(t)} - \bar{W}^{(t)}\|_F^2 < \alpha_2\eta^2 L\sum_{i=1}^{t-1}\alpha_1^{i-1}\hat{F}(\bar{w}^{(t-i)}). \tag{6.41}$$

which results in,

$$\hat{F}(\bar{w}^{(t+1)}) \leq (1-\eta\mu)\hat{F}(\bar{w}^{(t)}) + \alpha_2 L^3\eta^3 \sum_{i=1}^{t-1} \alpha_1^{i-1}\hat{F}(\bar{w}^{(t-i)}).$$

By induction assume $\hat{F}(\bar{w}^{(t)}) \leq \bar{\zeta}^{t-1}\hat{F}(\bar{w}^{(1)})$ then using the assumptions on $\bar{\zeta}$ and η yield the following inequalities,

$$\begin{aligned}
\hat{F}(\bar{w}^{(t+1)}) &\leq (1-\eta\mu)\bar{\zeta}^{t-1}\hat{F}(\bar{w}^{(1)}) + \alpha_2\eta^3 L^3\hat{F}(\bar{w}^{(1)}) \sum_{i=1}^{t-1} \alpha_1^{i-1}\bar{\zeta}^{t-i-1} \\
&\leq (1-\eta\mu)\bar{\zeta}^{t-1}\hat{F}(\bar{w}^{(1)}) + \alpha_2\eta^3 L^3\hat{F}(\bar{w}^{(1)}) \frac{\bar{\zeta}^{t-2}}{1-\alpha_1/\bar{\zeta}} \\
&= (1-\eta\mu+\alpha_2\eta^3 L^3/(\bar{\zeta}-\alpha_1))\bar{\zeta}^{t-1}\hat{F}(\bar{w}^{(1)}) \\
&\leq (1-\eta\mu+2\alpha_2\eta^3 L^3/(1-\alpha_1))\bar{\zeta}^{t-1}\hat{F}(\bar{w}^{(1)}) \\
&\leq (1-\eta\mu+\eta\mu/2)\bar{\zeta}^{t-1}\hat{F}(\bar{w}^{(1)}) \\
&\leq \bar{\zeta}^{t}\hat{F}(\bar{w}^{(1)}).
\end{aligned}$$

This completes the proof of the lemma.

6.5.4.1 Proof of Lemma 6.2.3

Lemma 6.5.12 (Restatement of Lemma 6.2.3)**.** *Let Assumptions 6.2.1,6.2.3 and 6.2.8 hold and let the step-size* $\eta \leq \min\{\frac{1-\alpha_1}{\mu}, \frac{1}{2L^2}\sqrt{\frac{(1-\alpha_1)\mu}{\alpha_2}}, \frac{1}{L}\}$*, where the constants* $\alpha_1 \in (0,1)$ *and* $\alpha_2 > 0$ *are defined same as in Lemma 6.5.11. Define* $\zeta := 1 - \frac{\eta\mu}{2}$*, then under the data separability assumption, the iterates of DGD satisfy for all* $t \geq 1$*,*

$$\begin{aligned}
\hat{F}(\bar{w}^{(t)}) &\leq \zeta^{t-1}\hat{F}(\bar{w}^{(1)}), \\
\frac{1}{N}\|W^{(t)} - \bar{W}^{(t)}\|_F^2 &\leq \frac{2\alpha_2\eta^2 L^2\hat{F}(\bar{w}^{(1)})}{1-\alpha_1}\zeta^{t-1}.
\end{aligned}$$

Proof: The bound on the train loss follows directly by Lemma 6.5.11, after noting that $\eta \leq \frac{1-\alpha_1}{\mu}$ implies $\frac{1+\alpha_1}{2} \leq 1 - \eta\mu/2$. The consensus error is derived by (6.41) and using the bound on $\hat{F}(\bar{w}^{(t)})$.

6.5.4.2 Proof of Theorem 6.2.3

Theorem 6.5.3 (Restatement of Theorem 6.2.3)**.** *Let Assumptions 6.2.1-6.2.4 and 6.2.8 hold, and let η and ζ be as in Lemma 6.5.11. Then the iterates of DGD under the data separability assumption satisfy for all $T \geq 1$,*

$$\mathbb{E}\Big[F(\bar{w}^{(T)})\Big] = O\Big(\zeta^T + \frac{L^2c^2}{n^{3-2\alpha}\mu^{2\alpha}}(\eta T)^{2-2\alpha} + \frac{\eta^2 L^4}{\mu^2 N}\Big).$$

Proof: By simplifying Lemma 6.2.1 using the convergence bounds in Lemma 6.5.12, we end up with the following,

$$\mathbb{E}\Big[F(\bar{w}^{(T)})\Big] = O\Big(\zeta^T + \frac{L^2c^2\eta^2}{n^{3-2\alpha}(1-\zeta)^{2\alpha}}T^{2-2\alpha} + \frac{\eta^2 L^4}{(1-\sqrt{\zeta})^2 N}\Big).$$

Based on the definition of ζ, we have $(\frac{\eta}{1-\zeta})^{2\alpha} = (\frac{2}{\mu})^{2\alpha}$ and $\frac{\eta^2}{(1-\sqrt{\zeta})^2} \leq \frac{4}{\mu^2}$. This proves the statement of the theorem.

6.5.5 Proof of Theorem 6.2.4

Theorem 6.5.4 (Restatement of Theorem 6.2.4)**.** *Consider* FDRL *(Algorithm 1) on separable dataset, and choose $\eta = O(1/\sqrt{t})$. Then for all $\ell \in [N]$*

$$\lim_{t\to\infty} \frac{w_\ell^{(t)}}{\|w_\ell^{(t)}\|} = \frac{w_{\text{MM}}}{\|w_{\text{MM}}\|},$$

where recall that w_{MM} denotes the solution to hard-margin SVM problem.

Proof: Replace $\frac{v_\ell^{(t)}}{\|v_\ell^{(t)}\|}$ in step 2 of Algorithm 1 by arbitrary perturbations $\varepsilon_\ell^{(t)}$ of unit norm. Then note that the sequence $\{w_\ell^{(t)}\}$ generated by step 2 is identical to decentralized GD with $\eta\|\varepsilon_\ell^{(t)}\| \to 0$. Thus by [163, Lemma 1], consensus is asymptotically achieved for all $\ell \in [N]$, i.e.,

$$\lim_{t\to\infty} \|w_\ell^{(t)} - \bar{w}^{(t)}\| = 0.$$

Thus

$$\lim_{t\to\infty} \|w_\ell^{(t+1)} - w_\ell^{(t)}\| = \lim_{t\to\infty} \|\bar{w}^{(t+1)} - \bar{w}^{(t)}\| = \lim_{t\to\infty} \eta\|\bar{\varepsilon}^{(t)}\| = 0.$$

This implies that for all $i \in [N]$ we have $\lim_{t\to\infty}\|\nabla\hat{F}_\ell(w_\ell^{(t+1)}) - \nabla\hat{F}_\ell(w_\ell^{(t)})\| = 0$, thus by appealing again to [163, Lemma 1] and applying it to step (5) of Algorithm 1, we find that,

$$\lim_{t\to\infty} \|v_\ell^{(t)} - \bar{v}^{(t)}\| = 0.$$

Aggregations of gradients in step (5) imply that $\bar{v}^{(t)} = \frac{\mathbf{1}^\top \nabla\mathcal{L}(W^{(t)})}{N} \to \nabla\hat{F}(\bar{w}^{(t)})$. Thus step (2) of FDLR for every agent i converges to $\bar{w}^{(t)} - \eta\frac{\nabla\hat{F}(\bar{w}^{(t)})}{\|\nabla\hat{F}(\bar{w}^{(t)})\|}$, i.e.,

$$\left\| \left(w_\ell^{(t)} - \eta\frac{v_\ell^{(t)}}{\|v_\ell^{(t)}\|} \right) - \left(\bar{w}^{(t)} - \eta\frac{\nabla\hat{F}(\bar{w}^{(t)})}{\|\nabla\hat{F}(\bar{w}^{(t)})\|} \right) \right\| \overset{t\to\infty}{\Longrightarrow} 0.$$

Thus for all ℓ, the sequence $\{w_\ell^{(t)}\}$ converges to the solution of normalized GD, i.e., the max-margin separator w_{MM}, for linearly separable datasets ([20, Theorem 5]). This leads to the statement of the theorem.

6.6 Auxiliary Results

Proposition 6.6.1 (Bounds on the exponential loss)**.** *Consider linear classification with the exponential loss* $f(w,(a,y)) = \exp(-y \cdot w^\top a)$ *over linearly separable dataset* $(a_i, y_i)_{i=1}^n$ *with binary labels* y_i *and with* $\max_i \|a_i\| \leq r$ *for a constant* r*. The training loss in this case satisfies for all* $w \in \mathbb{R}^d$*,*

$$\|\nabla \hat{F}(w)\| \in [c' F(w), cF(w)], \quad \|\nabla^2 \hat{F}(w)\| \leq hF(w), \tag{6.42}$$

for constants c, c' *and* h *independent of* w*.*

Proof: Using $\hat{F}(w) = \frac{1}{n}\sum_{i=1}^n \exp(-y_i \cdot w^\top a_i)$, one can deduce that,

$$\nabla \hat{F}(w) = -\frac{1}{n}\sum_{i=1}^n y_i a_i \exp(-y_i\, w^\top a_i),$$
$$\nabla^2 \hat{F}(w) = \frac{1}{n}\sum_{i=1}^n a_i a_i^\top \exp(-y_i\, w^\top a_i).$$

Therefore it holds that,

$$\begin{aligned}\|\nabla \hat{F}(w)\| &= \frac{1}{n}\|\sum_{i=1}^n y_i a_i \exp(-y_i w^\top a_i)\| \\ &\leq \frac{1}{n}\sum_{i=1}^n \|y_i a_i \exp(-y_i w^\top a_i)\| \\ &= \frac{1}{n}\sum_{i=1}^n \|y_i a_i\| \exp(-y_i w^\top a_i) \leq r\hat{F}(w).\end{aligned}$$

A similar approach for the Hessian of $\hat{F}$ results in the following inequality,

$$\|\nabla^2 \hat{F}(w)\| \leq r^2 \hat{F}(w).$$

Moreover, due to linear separability there exists a $w^\star \in \mathbb{R}^d$ such that,

$$\frac{y_i {w^\star}^\top a_i}{\|w^\star\|} \geq \gamma, \quad \forall i \in [N],$$

where $\gamma > 0$ denotes the margin. Therefore, using the supremum definition of norm we can write,

$$\begin{aligned}
\|\nabla \hat{F}(w)\| &= \frac{1}{n} \left\| \sum_{i=1}^n y_i a_i \exp(-y_i w^\top a_i) \right\| \\
&= \sup_{\substack{v \in \mathbb{R}^d \\ \text{s,t. } \|v\|=1}} \left\langle \frac{1}{n} \sum_{i=1}^n y_i a_i \exp(-y_i w^\top a_i), v \right\rangle \\
&\geq \left\langle \frac{1}{n} \sum_{i=1}^n y_i a_i \exp(-y_i w^\top a_i), \frac{w^\star}{\|w^\star\|} \right\rangle \\
&\geq \frac{1}{n} \sum_{i=1}^n \gamma \cdot \exp(-y_i w^\top a_i) \\
&= \gamma \hat{F}(w).
\end{aligned}$$

This completes the proof.

Proposition 6.6.2 (Bounds on the logistic loss)**.** *Consider linear classification with the logistic loss $f(w,(a,y)) = \log(1+\exp(-y \cdot w^\top a))$ over linearly separable dataset $(a_i, y_i)_{i=1}^n$ with binary labels y_i and with $\max_i \|a_i\| \leq r$ for a constant r. The training loss in this case satisfies for all $w \in \mathbb{R}^d$,*

$$\|\nabla \hat{F}(w)\| \in [c' \Phi(w), cF(w)], \quad \|\nabla^2 \hat{F}(w)\| \leq hF(w),$$

for $\Phi(w) := \frac{1}{n} \sum_{i=1}^n \frac{\exp(-y_i w^\top a_i)}{1+\exp(-y_i w^\top a_i)}$ and constants c, c' and h independent of w.

Proof: The training loss is now $\hat{F}(w) = \frac{1}{n}\sum_{i=1}^{n} \log(1+\exp(-y_i \cdot w^\top a_i))$. Thus,

$$\nabla \hat{F}(w) = \frac{1}{n}\sum_{i=1}^{n}(-y_i a_i)\frac{\exp(-y_i\, w^\top a_i)}{1+\exp(-y_i \cdot w^\top a_i)},$$
$$\nabla^2 \hat{F}(w) = \frac{1}{n}\sum_{i=1}^{n} a_i a_i^\top \frac{\exp(-y_i\, w^\top a_i)}{(1+\exp(-y_i\, w^\top a_i))^2}.$$

By considering the norm and noting that $\exp(t)/(1+\exp(t)) \leq \log(1+\exp(t))$,

$$\begin{aligned}\|\nabla \hat{F}(w)\| &= \frac{1}{n}\left\|\sum_{i=1}^{n}(-y_i a_i)\frac{\exp(-y_i\, w^\top a_i)}{1+\exp(-y_i \cdot w^\top a_i)}\right\| \\ &\leq \frac{1}{n}\sum_{i=1}^{n}\|y_i a_i\|\,\frac{\exp(-y_i\, w^\top a_i)}{1+\exp(-y_i \cdot w^\top a_i)} \\ &\leq \frac{r}{n}\sum_{i=1}^{n}\log(1+\exp(-y_i w^\top a_i)) = r\hat{F}(w).\end{aligned}$$

Likewise, since $\exp(t)/(1+\exp(t))^2 \leq 2\log(1+\exp(t))$, we can conclude that the operator norm of the Hessian satisfies,

$$\nabla^2 \hat{F}(w) \leq 2r^2 \hat{F}(w).$$

This completes the proof of upper-bounds for the gradient and Hessian. For the lower-bound on gradient note that by using the supremum definition of norm and recalling the

max-margin separator satisfies $\frac{y_i w^{\star\top} a_i}{\|w^\star\|} \geq \gamma$ for the margin $\gamma > 0$ and all $i \in [n]$, we obtain,

$$\begin{aligned}
\|\nabla \hat{F}(w)\| &= \frac{1}{n}\left\|\sum_{i=1}^{n} y_i a_i \frac{\exp(-y_i\, w^\top a_i)}{1+\exp(-y_i \cdot w^\top a_i)}\right\| \\
&= \sup_{\substack{v\in\mathbb{R}^d \\ \text{s,t. } \|v\|=1}} \left\langle \frac{1}{n}\sum_{i=1}^{n} y_i a_i \frac{\exp(-y_i\, w^\top a_i)}{1+\exp(-y_i \cdot w^\top a_i)}, v \right\rangle \\
&\geq \left\langle \frac{1}{n}\sum_{i=1}^{n} y_i a_i \frac{\exp(-y_i\, w^\top a_i)}{1+\exp(-y_i \cdot w^\top a_i)}, \frac{w^\star}{\|w^\star\|} \right\rangle \\
&\geq \frac{1}{n}\sum_{i=1}^{n} \gamma \frac{\exp(-y_i\, w^\top a_i)}{1+\exp(-y_i \cdot w^\top a_i)}.
\end{aligned}$$

This yields the lower bound $\gamma\Phi(w)$ on the norm of gradient and completes the proof.

Proposition 6.6.3 (Realizability of the exponential and logistic loss [152])**.** *On linearly separable data with margin $\gamma > 0$, the exponential loss function satisfies the realizability assumption (Assumption 6.2.5) with $\rho(\varepsilon) = -\frac{1}{\gamma}\log(\varepsilon)$, where γ denotes the margin. Moreover, the logistic loss function satisfies the realizability assumption with $\rho(\varepsilon) = -\frac{1}{\gamma}\log(\exp(\varepsilon) - 1)$.*

6.7 Additional Experiments

6.7.1 Experiments on over-parameterized Least-squares

In Fig. 6.4, we conduct experiments for highly over-parameterized Least-squares ($f(w, x) = (1 - w^\top x)^2$), where d is typically significantly larger than n to ensure perfect interpolation of dataset. Note that, the train loss is not strongly-convex in this case, instead it satisfies the PL condition(Assumption 6.2.8). Notably, as predicted by Lemma 6.2.3, we notice the linear convergence of the train loss and the consensus error in Fig. 6.4 (Left). On the other hand, for the test loss, we observe its remarkably fast convergence

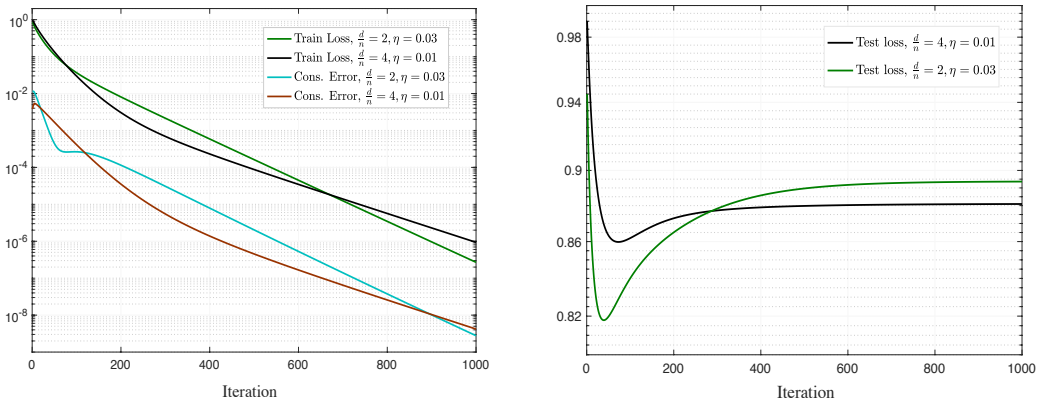


Figure 6.4: Consensus error, train loss and test loss for DGD with over-parameterized least-squares(square loss). The test loss achieves its optimum at the very early stages of DGD.

(after approximately 50 iterations) to the optimal value, which is followed by a sharp increase in the subsequent iterations.

6.7.2 On the update rule of FDLR

In the final section of this chapter, we state a remark regarding the update rule of FDLR. Recall the update rule of DGD,

$$w_\ell^{(t+1)} = \sum_{k \in \mathcal{N}_\ell} A_{\ell k} w_k^{(t)} - \eta_t \nabla \hat{F}_\ell(w_\ell^{(t)}). \tag{6.43}$$

Notably, we expect FDLR to be perhaps the simplest approach for accommodating normalized gradients in decentralized learning setting since in DGD the agents only have access to their local gradients. In particular consider a Normalized DGD algorithm with the same update as in (6.43) but with $\nabla \hat{F}_i(w_i^{(t)})$ replaced by $\nabla \hat{F}_i(w_i^{(t)})/\|\nabla \hat{F}_i(w_i^{(t)})\|$, i.e.,

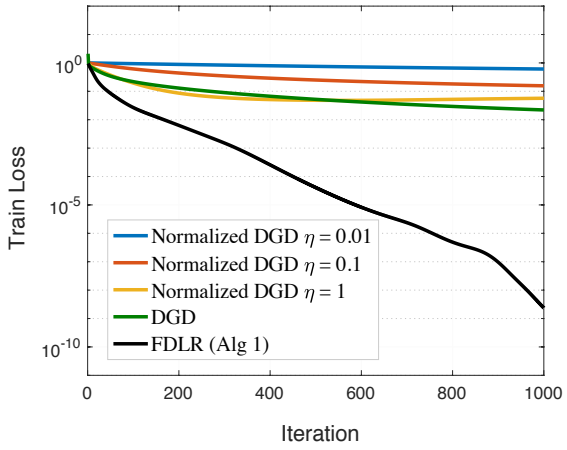


Figure 6.5: Normalized DGD with the update rule in Eq.(6.44) for different step-sizes η compared to DGD (Eq.(6.43)) and to FDLR (Alg 1). The step-sizes for DGD and FDLR are fine-tuned so that best of each algorithm is depicted. Normalized DGD cannot outperform DGD while FDLR is significantly faster than DGD. Here we consider linear classification with the exponential loss function and the dataset is generated according to signed measurements with Gaussian features and $n = 100, d = 50$.

$$w_\ell^{(t+1)} = \sum_{k \in \mathcal{N}_\ell} A_{\ell k} w_k^{(t)} - \eta_t \frac{\nabla \hat{F}_\ell(w_\ell^{(t)})}{\|\nabla \hat{F}_\ell(w_\ell^{(t)})\|} \tag{6.44}$$

The Normalized DGD algorithm above does not lead to faster convergence. This is due to the fact that in DGD the local gradient norm $\|\nabla \hat{F}_i(w_i^{(t)})\|$ can be different than the global gradient norm $\|\nabla \hat{F}(w_i^{(t)})\|$. Thus even if with the update rule (6.44) the local parameters $w_i^{(t)}$ converge to the global optimal solution, still the update rule for the averaged parameter $\bar{w}^{(t)}$ is different than the update rule of centralized normalized GD. Our numerical experiment in Fig. 6.5 demonstrates the incapability of Normalized DGD in speeding up DGD. In particular, we note that for any choice of step-size Normalized DGD does not lead to acceleration compared to DGD whereas FDLR massively outperforms DGD.

A note about convergence rates of DGD

As mentioned in the chapter's introduction, many prior works on investigate convergence of DGD and of its stochastic variant decentralized stochastic gradient descent (DSGD) under various assumptions, e.g. [167, 168, 169, 170] and many references therein. Most recently, [170] has presented a powerful unifying analysis of DSGD under rather weak assumptions. Specialized to convex L-smooth functions for which there exists w^* such that $\|\nabla f_i(w^*)\|= 0$ (i.e. interpolation) [170, Thm. 2] shows a rate of $\mathcal{O}(LR_0/T)$ for average DSGD updates. Here, $R_0 = \|w_1 - w^*\|_2$. Ignoring logarithmic factors, this rate is the same as what we obtained in (6.6) (as a consequence of Lemma 6.2.2) for DGD specifically applied to logistic loss over separable data. However, our result does *not* directly follow from [170, Thm. 2]. The reason is that logistic loss on separable data does *not* attain a bounded estimator. In fact, we believe the $\log^2 T$ dependence of the rate that shows up in our analysis (see Eq. (6.5)), is a consequence of the infinitely normed-optimizers in our setting and we expect the bound to be tight as suggested by our experiments (see Fig 6.3) and in agreement with convergence bounds for logistic regression on separable data in the centralized case derived recently in [113, Theorem. 1.1]. On the other hand, the results of [170] are applicable to finite optimizers, which yields $\mathcal{O}(1/T)$ convergence rates without log factor. Besides the above, in Theorem 6.2.2, we prove novel last-iterate (as opposed to averaged in the literature) convergence bounds for the train loss and faster consensus error rates of $\tilde{\mathcal{O}}(1/T^2)$. This is possible by leveraging additional Hessian self-bounded (Assumption 6.2.6) and self-lower-boudned (Assumption 6.2.7) assumptions, which hold for example for the exponential loss. Finally, we recall that our main focus is on studying finite time *generalization* bounds for DGD (e.g. Theorem 6.2.1), which to the best of our knowledge are new in this setting. Having discussed these, it is worth noting that the analysis of [170] applies under a relaxed assumption

on the mixing matrix (see [170, Assumption 4]) than the corresponding assumptions (e.g. Assumption 6.2.1) in the literature. For example, this relaxed assumption covers decentralized local SGD (with multiple local updates per iteration) as a special case and is interesting to extend our results (on logistic regression over separable data) to such settings.

Chapter 7

Conclusions

In this thesis, we have explored various aspects of learning in the interpolation regime across linear models and neural networks. Our findings provide both theoretical insights and practical implications for improving the performance of these models in over-parameterized settings.

For linear models, we derived sharp guarantees that accurately predict the performance of high-dimensional linear classifiers. These precise results allowed us to design optimal loss functions and regularization parameters, thereby achieving the theoretical lower bound on test error. We extended this framework to the adversarial training scenario, deriving exact asymptotic expressions for both standard and adversarial test errors under ℓ_p-bounded perturbations in Gaussian mixture models.

In the context of neural networks, we established non-asymptotic bounds on the training and test error performance in the interpolating regime. Our analysis revealed an exponential improvement in the lower bound on network width necessary for optimal performance. Additionally, the resulting generalization bounds enhance the results obtained from well-established methods such as uniform convergence, providing a more refined understanding of neural network behavior in over-parameterized settings.

Finally, we studied the behavior of train loss and test loss of decentralized gradient descent (DGD) methods in the interpolation regime and proposed two algorithms for speeding up the training.

As a future direction, we aim to extend our sharp analysis framework to more complex neural network architectures, such as deep neural networks and transformers. Additionally, we hope to derive the fundamental limits for adversarial training, offering precise characterizations for the optimal loss, regularization and attack budget that could guide the development of robust learning models. Another potential direction is to extend our algorithmic-stability analysis from neural networks to more complex architectures and training paradigms, such as next-token prediction in transformers.

Bibliography

[1] E. J. Candès and P. Sur, *The phase transition for the existence of the maximum likelihood estimate in high-dimensional logistic regression*, *arXiv preprint arXiv:1804.09753* (2018).

[2] R. T. Rockafellar and R. J.-B. Wets, *Variational analysis*, vol. 317. Springer Science & Business Media, 2009.

[3] Y. Gordon, *Some inequalities for gaussian processes and applications*, *Israel Journal of Mathematics* **50** (1985), no. 4 265–289.

[4] C. Thrampoulidis, S. Oymak, and B. Hassibi, *Regularized linear regression: A precise analysis of the estimation error*, in *Proceedings of The 28th Conference on Learning Theory*, pp. 1683–1709, 2015.

[5] C. Thrampoulidis, E. Abbasi, and B. Hassibi, *Precise error analysis of regularized m-estimators in high dimensions*, *IEEE Transactions on Information Theory* **64** (2018), no. 8 5592–5628.

[6] I. J. Goodfellow, J. Shlens, and C. Szegedy, *Explaining and harnessing adversarial examples*, *arXiv preprint arXiv:1412.6572* (2014).

[7] A. Madry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu, *Towards deep learning models resistant to adversarial attacks*, *arXiv preprint arXiv:1706.06083* (2017).

[8] G. V. Cybenko, *Approximation by superpositions of a sigmoidal function*, *Mathematics of Control, Signals and Systems* **2** (1989) 303–314.

[9] C. Zhang, S. Bengio, M. Hardt, B. Recht, and O. Vinyals, *Understanding deep learning requires rethinking generalization*, *arXiv preprint arXiv:1611.03530* (2016).

[10] M. Belkin, D. Hsu, S. Ma, and S. Mandal, *Reconciling modern machine-learning practice and the classical bias–variance trade-off*, *Proceedings of the National Academy of Sciences* **116** (2019), no. 32 15849–15854.

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

[12] L. Devroye and T. Wagner, *Distribution-free performance bounds for potential function rules*, *Information Theory, IEEE Transactions on* **IT-25** (10, 1979) 601 – 604.

[13] O. Bousquet and A. Elisseeff, *Stability and generalization*, *The Journal of Machine Learning Research* **2** (2002) 499–526.

[14] Y. Lei and Y. Ying, *Fine-grained analysis of stability and generalization for stochastic gradient descent*, in *International Conference on Machine Learning*, pp. 5809–5819, PMLR, 2020.

[15] S. Shalev-Shwartz and S. Ben-David, *Understanding machine learning: From theory to algorithms*. Cambridge university press, 2014.

[16] D. Richards and M. Rabbat, *Learning with gradient descent and weakly convex losses*, in *International Conference on Artificial Intelligence and Statistics*, pp. 1990–1998, PMLR, 2021.

[17] M. Hardt, B. Recht, and Y. Singer, *Train faster, generalize better: Stability of stochastic gradient descent*, in *International conference on machine learning*, pp. 1225–1234, PMLR, 2016.

[18] D. Richards and I. Kuzborskij, *Stability & generalisation of gradient descent for shallow neural networks without the neural tangent kernel*, *Advances in Neural Information Processing Systems* **34** (2021) 8609–8621.

[19] A. Nitanda, G. Chinot, and T. Suzuki, *Gradient descent can learn less over-parameterized two-layer neural networks on classification problems*, *arXiv preprint arXiv:1905.09870* (2019).

[20] M. S. Nacson, J. Lee, S. Gunasekar, P. H. P. Savarese, N. Srebro, and D. Soudry, *Convergence of gradient descent on separable data*, in *The 22nd International Conference on Artificial Intelligence and Statistics*, pp. 3420–3428, PMLR, 2019.

[21] Z. Ji and M. Telgarsky, *Characterizing the implicit bias via a primal-dual analysis*, in *Algorithmic Learning Theory*, pp. 772–804, PMLR, 2021.

[22] Z. Ji, N. Srebro, and M. Telgarsky, *Fast margin maximization via dual acceleration*, in *International Conference on Machine Learning*, pp. 4860–4869, PMLR, 2021.

[23] P. J. Huber, *Robust statistics*. Springer, 2011.

[24] E. J. Candès, *Mathematics of sparsity (and a few other things)*, in *Proceedings of the International Congress of Mathematicians, Seoul, South Korea*, vol. 123, Citeseer, 2014.

[25] A. Montanari, *Statistical estimation: from denoising to sparse regression and hidden cliques*, *Statistical Physics, Optimization, Inference and Message-passing Algorithms: Lecture Notes of the Les Houches School of Physics-Special Issue, October 2013* (2015) 127.

[26] N. E. Karoui, *Asymptotic behavior of unregularized and ridge-regularized high-dimensional robust regression estimators: rigorous results*, *arXiv preprint arXiv:1311.2445* (2013).

[27] D. L. Donoho, A. Maleki, and A. Montanari, *The noise-sensitivity phase transition in compressed sensing*, *Information Theory, IEEE Transactions on* **57** (2011), no. 10 6920–6941.

[28] M. Stojnic, *Various thresholds for ℓ_1-optimization in compressed sensing*, *arXiv preprint arXiv:0907.3666* (2009).

[29] M. Bayati and A. Montanari, *The lasso risk for gaussian matrices*, *Information Theory, IEEE Transactions on* **58** (2012), no. 4 1997–2017.

[30] V. Chandrasekaran, B. Recht, P. A. Parrilo, and A. S. Willsky, *The convex geometry of linear inverse problems*, *Foundations of Computational Mathematics* **12** (2012), no. 6 805–849.

[31] D. Amelunxen, M. Lotz, M. B. McCoy, and J. A. Tropp, *Living on the edge: A geometric theory of phase transitions in convex optimization*, *arXiv preprint arXiv:1303.6672* (2013).

[32] S. Oymak and B. Hassibi, *Sharp mse bounds for proximal denoising*, *Foundations of Computational Mathematics* **16** (2016), no. 4 965–1029.

[33] M. Stojnic, *A framework to characterize performance of lasso algorithms*, *arXiv preprint arXiv:1303.7291* (2013).

[34] S. Oymak, C. Thrampoulidis, and B. Hassibi, *The squared-error of generalized lasso: A precise analysis*, *arXiv preprint arXiv:1311.0830* (2013).

[35] N. El Karoui, *On the impact of predictor geometry on the performance on high-dimensional ridge-regularized generalized robust regression estimators*, *Probability Theory and Related Fields* **170** (2018), no. 1-2 95–175.

[36] D. Donoho and A. Montanari, *High dimensional robust m-estimation: Asymptotic variance via approximate message passing*, *Probability Theory and Related Fields* **166** (2016), no. 3-4 935–969.

[37] C. Thrampoulidis, S. Oymak, and B. Hassibi, *Regularized linear regression: A precise analysis of the estimation error*, in *Conference on Learning Theory*, pp. 1683–1709, 2015.

[38] S. Oymak and J. A. Tropp, *Universality laws for randomized dimension reduction, with applications*, *Information and Inference: A Journal of the IMA* **7** (2017), no. 3 337–446.

[39] L. Miolane and A. Montanari, *The distribution of the lasso: Uniform control over sparse balls and adaptive parameter tuning*, *arXiv preprint arXiv:1811.01212* (2018).

[40] S. Wang, H. Weng, and A. Maleki, *Does slope outperform bridge regression?*, *arXiv preprint arXiv:1909.09345* (2019).

[41] M. Celentano and A. Montanari, *Fundamental barriers to high-dimensional regression with convex penalties*, *arXiv preprint arXiv:1903.10603* (2019).

[42] H. Hu and Y. M. Lu, *Asymptotics and optimal designs of slope for sparse linear regression*, *arXiv preprint arXiv:1903.11582* (2019).

[43] Z. Bu, J. Klusowski, C. Rush, and W. Su, *Algorithmic analysis and statistical estimation of slope via approximate message passing*, in *Advances in Neural Information Processing Systems*, pp. 9361–9371, 2019.

[44] H. Taheri, R. Pedarsani, and C. Thrampoulidis, *Optimality of least-squares for classification in gaussian-mixture models*, in *2020 IEEE International Symposium on Information Theory (ISIT)*, pp. 2515–2520, IEEE, 2020.

[45] H. Taheri, R. Pedarsani, and C. Thrampoulidis, *Asymptotic behavior of adversarial training in binary linear classification*, *IEEE Transactions on Neural Networks and Learning Systems* (2023).

[46] C. Thrampoulidis, E. Abbasi, and B. Hassibi, *Lasso with non-linear measurements is equivalent to one with linear measurements*, in *Advances in Neural Information Processing Systems*, pp. 3420–3428, 2015.

[47] H. Huang, *Asymptotic behavior of support vector machine for spiked population model*, *The Journal of Machine Learning Research* **18** (2017), no. 1 1472–1492.

[48] P. Sur and E. J. Candès, *A modern maximum-likelihood theory for high-dimensional logistic regression*, *Proceedings of the National Academy of Sciences* (2019) 201810420.

[49] X. Mai, Z. Liao, and R. Couillet, *A large scale analysis of logistic regression: Asymptotic performance and new insights*, in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 3357–3361, IEEE, 2019.

[50] X. Mai, Z. Liao, and R. Couillet, *A large scale analysis of logistic regression: asymptotic performance and new insights*, in *ICASSP*, 2019.

[51] A. Kammoun and M.-S. Alouini, *On the precise error analysis of support vector machines*, *arXiv preprint arXiv:2003.12972* (2020).

[52] F. Salehi, E. Abbasi, and B. Hassibi, *The impact of regularization on high-dimensional logistic regression*, *arXiv preprint arXiv:1906.03761* (2019).

[53] 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*, pp. 3739–3749, PMLR, 2020.

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

[55] A. Montanari, F. Ruan, Y. Sohn, and J. Yan, *The generalization error of max-margin linear classifiers: High-dimensional asymptotics in the overparametrized regime*, *arXiv preprint arXiv:1911.01544* (2019).

[56] T. Liang and P. Sur, *A precise high-dimensional asymptotic theory for boosting and min-l1-norm interpolated classifiers*, *arXiv preprint arXiv:2002.01586* (2020).

[57] F. Mignacco, F. Krzakala, Y. M. Lu, and L. Zdeborová, *The role of regularization in classification of high-dimensional noisy gaussian mixture*, *arXiv preprint arXiv:2002.11544* (2020).

[58] D. Bean, P. J. Bickel, N. El Karoui, and B. Yu, *Optimal m-estimation in high-dimensional regression*, *Proceedings of the National Academy of Sciences* **110** (2013), no. 36 14563–14568.

[59] M. Advani and S. Ganguli, *Statistical mechanics of optimal convex inference in high dimensions*, *Physical Review X* **6** (2016), no. 3 031034.

[60] D. L. Donoho and A. Montanari, *Variance breakdown of huber (m)-estimators: n/p \rightarrow m\in (1,\infty)*, *arXiv preprint arXiv:1503.02106* (2015).

[61] O. Johnson and A. Barron, *Fisher information inequalities and the central limit theorem*, *Probability Theory and Related Fields* **129** (2004), no. 3 391–409.

[62] Y. Wu and S. Verdú, *Optimal phase transitions in compressed sensing*, *Information Theory, IEEE Transactions on* **58** (2012), no. 10 6241–6263.

[63] J. Barbier, F. Krzakala, N. Macris, L. Miolane, and L. Zdeborová, *Optimal errors and phase transitions in high-dimensional generalized linear models*, *Proceedings of the National Academy of Sciences* **116** (2019), no. 12 5451–5460.

[64] G. Reeves and H. D. Pfister, *The replica-symmetric prediction for random linear estimation with gaussian matrices is exact*, *IEEE Transactions on Information Theory* **65** (2019), no. 4 2252–2283.

[65] N. Blachman, *The convolution inequality for entropy powers*, *IEEE Transactions on Information Theory* **11** (1965), no. 2 267–271.

[66] M. Sion, *On general minimax theorems.*, *Pacific J. Math.* **8** (1958), no. 1 171–176.

[67] H. Taheri, R. Pedarsani, and C. Thrampoulidis, *Sharp guarantees for solving random equations with one-bit information*, in *2019 57th Annual Allerton Conference on Communication, Control, and Computing (Allerton)*, pp. 765–772, 2019.

[68] R. T. Rockafellar, *Convex analysis*, vol. 28. Princeton university press, 1997.

[69] Y. Plan and R. Vershynin, *The generalized lasso with non-linear observations*, *arXiv preprint arXiv:1502.04071* (2015).

[70] M. Genzel, *High-dimensional estimation of structured signals from non-linear observations with general convex loss functions*, *IEEE Transactions on Information Theory* **63** (2016), no. 3 1601–1619.

[71] M. Mondelli and A. Montanari, *Fundamental limits of weak recovery with applications to phase retrieval*, *arXiv preprint arXiv:1708.05932* (2017).

[72] Y. M. Lu and G. Li, *Phase transitions of spectral initialization for high-dimensional nonconvex estimation*, *arXiv preprint arXiv:1702.06435* (2017).

[73] C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. Goodfellow, and R. Fergus, *Intriguing properties of neural networks. arxiv 2013*, *arXiv preprint arXiv:1312.6199* (2013).

[74] S.-M. Moosavi-Dezfooli, A. Fawzi, and P. Frossard, *Deepfool: a simple and accurate method to fool deep neural networks*, in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2574–2582, 2016.

[75] S. H. Silva and P. Najafirad, *Opportunities and challenges in deep learning adversarial robustness: A survey*, *arXiv preprint arXiv:2007.00753* (2020).

[76] A. Chakraborty, M. Alam, V. Dey, A. Chattopadhyay, and D. Mukhopadhyay, *Adversarial attacks and defences: A survey*, *arXiv preprint arXiv:1810.00069* (2018).

[77] 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*, pp. 2773–2781, PMLR, 2021.

[78] A. Javanmard, M. Soltanolkotabi, and H. Hassani, *Precise tradeoffs in adversarial training for linear regression*, *arXiv preprint arXiv:2002.10477* (2020).

[79] A. Javanmard and M. Soltanolkotabi, *Precise statistical analysis of classification accuracies for adversarial training*, *arXiv preprint arXiv:2010.11213* (2020).

[80] M. Celentano, A. Montanari, and Y. Wei, *The lasso with general gaussian designs with applications to hypothesis testing*, *arXiv preprint arXiv:2007.13716* (2020).

[81] O. Dhifallah and Y. M. Lu, *On the inherent regularization effects of noise injection during training*, *arXiv preprint arXiv:2102.07379* (2021).

[82] M. Celentano and A. Montanari, *Cad: Debiasing the lasso with inaccurate covariate model*, *arXiv preprint arXiv:2107.14172* (2021).

[83] A. N. Bhagoji, D. Cullina, and P. Mittal, *Lower bounds on adversarial robustness from optimal transport*, in *Advances in Neural Information Processing Systems*, pp. 7498–7510, 2019.

[84] C. Dan, Y. Wei, and P. Ravikumar, *Sharp statistical guarantees for adversarially robust gaussian classification*, *arXiv preprint arXiv:2006.16384* (2020).

[85] E. Dobriban, H. Hassani, D. Hong, and A. Robey, *Provable tradeoffs in adversarially robust classification*, *arXiv preprint arXiv:2006.05161* (2020).

[86] Z. Charles, S. Rajput, S. Wright, and D. Papailiopoulos, *Convergence and margin of adversarial training on separable data*, *arXiv preprint arXiv:1905.09209* (2019).

[87] Z. Allen-Zhu and Y. Li, *Feature purification: How adversarial training performs robust deep learning*, *arXiv preprint arXiv:2005.10190* (2020).

[88] Y. Min, L. Chen, and A. Karbasi, *The curious case of adversarially robust models: More data can help, double descend, or hurt generalization*, *arXiv preprint arXiv:2002.11080* (2020).

[89] D. Tsipras, S. Santurkar, L. Engstrom, A. Turner, and A. Madry, *Robustness may be at odds with accuracy*, *arXiv preprint arXiv:1805.12152* (2018).

[90] A. Raghunathan, S. M. Xie, F. Yang, J. C. Duchi, and P. Liang, *Adversarial training can hurt generalization*, *arXiv preprint arXiv:1906.06032* (2019).

[91] H. Zhang, Y. Yu, J. Jiao, E. P. Xing, L. E. Ghaoui, and M. I. Jordan, *Theoretically principled trade-off between robustness and accuracy*, *arXiv preprint arXiv:1901.08573* (2019).

[92] A. Raghunathan, S. M. Xie, F. Yang, J. Duchi, and P. Liang, *Understanding and mitigating the tradeoff between robustness and accuracy*, *arXiv preprint arXiv:2002.10716* (2020).

[93] Y. Carmon, A. Raghunathan, L. Schmidt, P. Liang, and J. C. Duchi, *Unlabeled data improves adversarial robustness*, *arXiv preprint arXiv:1905.13736* (2019).

[94] S. Mei and A. Montanari, *The generalization error of random features regression: Precise asymptotics and double descent curve*, *arXiv preprint arXiv:1908.05355* (2019).

[95] S. Goldt, M. Mézard, F. Krzakala, and L. Zdeborová, *Modeling the influence of data structure on learning in neural networks: The hidden manifold model*, *Physical Review X* **10** (2020), no. 4 041044.

[96] O. Dhifallah and Y. M. Lu, *A precise performance analysis of learning with random features*, *arXiv preprint arXiv:2008.11904* (2020).

[97] B. Ghorbani, S. Mei, T. Misiakiewicz, and A. Montanari, *Limitations of lazy training of two-layers neural network.*, in *NeurIPS*, 2019.

[98] H. Taheri, R. Pedarsani, and C. Thrampoulidis, *Asymptotic behavior of adversarial training in binary linear classification*, *IEEE International Symposium on Information Theory (ISIT) (to appear)* (2022).

[99] L. Schmidt, S. Santurkar, D. Tsipras, K. Talwar, and A. Madry, *Adversarially robust generalization requires more data*, in *Advances in Neural Information Processing Systems*, pp. 5014–5026, 2018.

[100] M. Belkin, D. Hsu, S. Ma, and S. Mandal, *Reconciling modern machine learning and the bias-variance trade-off*, *arXiv preprint arXiv:1812.11118* (2018).

[101] T. Hastie, A. Montanari, S. Rosset, and R. J. Tibshirani, *Surprises in high-dimensional ridgeless least squares interpolation*, *arXiv preprint arXiv:1903.08560* (2019).

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

[103] X. Mai and Z. Liao, *High dimensional classification via regularized and unregularized empirical risk minimization: Precise error and optimal loss*, *arXiv preprint arXiv:1905.13742* (2019).

[104] A. Rahimi and B. Recht, *Random features for large-scale kernel machines*, *Advances in neural information processing systems* **20** (2007).

[105] A. Jacot, F. Gabriel, and C. Hongler, *Neural tangent kernel: Convergence and generalization in neural networks*, *Advances in neural information processing systems* **31** (2018).

[106] M. Bayati, M. Lelarge, A. Montanari, *et. al.*, *Universality in polytope phase transitions and message passing algorithms*, *Annals of Applied Probability* **25** (2015), no. 2 753–822.

[107] S. Oymak and J. A. Tropp, *Universality laws for randomized dimension reduction, with applications*, *Information and Inference: A Journal of the IMA* **7** (2018), no. 3 337–446.

[108] E. Abbasi, F. Salehi, and B. Hassibi, *Universality in learning from linear measurements*, *arXiv preprint arXiv:1906.08396* (2019).

[109] L. Chizat, E. Oyallon, and F. Bach, *On lazy training in differentiable programming*, *Advances in neural information processing systems* **32** (2019).

[110] Z. Ji and M. Telgarsky, *Polylogarithmic width suffices for gradient descent to achieve arbitrarily small test error with shallow relu networks*, in *International Conference on Learning Representations*, 2020.

[111] R. Bassily, V. Feldman, C. Guzmán, and K. Talwar, *Stability of stochastic gradient descent on nonsmooth convex losses*, *Advances in Neural Information Processing Systems* **33** (2020) 4381–4391.

[112] M. Schliserman and T. Koren, *Stability vs implicit bias of gradient methods on separable data and beyond*, in *Proceedings of Thirty Fifth Conference on Learning Theory* (P.-L. Loh and M. Raginsky, eds.), vol. 178 of *Proceedings of Machine Learning Research*, pp. 3380–3394, PMLR, 02–05 Jul, 2022.

[113] Z. Ji and M. Telgarsky, *Risk and parameter convergence of logistic regression*, *arXiv preprint arXiv:1803.07300* (2018).

[114] O. Shamir, *Gradient methods never overfit on separable data*, *Journal of Machine Learning Research* **22** (2021), no. 85 1–20.

[115] Y. Cao and Q. Gu, *Generalization error bounds of gradient descent for learning over-parameterized deep relu networks*, in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, pp. 3349–3356, 2020.

[116] Y. Lei, R. Jin, and Y. Ying, *Stability and generalization analysis of gradient methods for shallow neural networks*, in *Advances in Neural Information Processing Systems*, 2022.

[117] C. Wei, J. D. Lee, Q. Liu, and T. Ma, *Regularization matters: Generalization and optimization of neural nets vs their induced kernel*, *Advances in Neural Information Processing Systems* **32** (2019).

[118] P. Bartlett, *For valid generalization the size of the weights is more important than the size of the network*, in *Advances in Neural Information Processing Systems*, vol. 9, MIT Press, 1996.

[119] P. L. Bartlett, V. Maiorov, and R. Meir, *Almost linear vc dimension bounds for piecewise polynomial networks*, NIPS'98, p. 190–196, MIT Press, 1998.

[120] Z. Allen-Zhu, Y. Li, and Y. Liang, *Learning and generalization in overparameterized neural networks, going beyond two layers*, *Advances in neural information processing systems* **32** (2019).

[121] S. Oymak and M. Soltanolkotabi, *Toward moderate overparameterization: Global convergence guarantees for training shallow neural networks*, *IEEE Journal on Selected Areas in Information Theory* **1** (2020), no. 1 84–105.

[122] A. Javanmard, M. Mondelli, and A. Montanari, *Analysis of a two-layer neural network via displacement convexity*, *The Annals of Statistics* **48** (2020), no. 6.

[123] Y. Li and Y. Liang, *Learning overparameterized neural networks via stochastic gradient descent on structured data*, *Advances in neural information processing systems* **31** (2018).

[124] M. Soltanolkotabi, A. Javanmard, and J. D. Lee, *Theoretical insights into the optimization landscape of over-parameterized shallow neural networks*, *IEEE Transactions on Information Theory* **65** (2018), no. 2 742–769.

[125] Z. Allen-Zhu, Y. Li, and Z. Song, *A convergence theory for deep learning via over-parameterization*, in *International Conference on Machine Learning*, pp. 242–252, PMLR, 2019.

[126] C. Liu, L. Zhu, and M. Belkin, *Loss landscapes and optimization in over-parameterized non-linear systems and neural networks*, *Applied and Computational Harmonic Analysis* **59** (2022) 85–116.

[127] Z. Charles and D. Papailiopoulos, *Stability and generalization of learning algorithms that converge to global optima*, in *International Conference on Machine Learning*, pp. 745–754, PMLR, 2018.

[128] Y. Lei and Y. Ying, *Sharper generalization bounds for learning with gradient-dominated objective functions*, in *International Conference on Learning Representations*, 2020.

[129] S. Oymak, Z. Fabian, M. Li, and M. Soltanolkotabi, *Generalization guarantees for neural networks via harnessing the low-rank structure of the jacobian*, *arXiv preprint arXiv:1906.05392* (2019).

[130] P. L. Bartlett and S. Mendelson, *Rademacher and gaussian complexities: Risk bounds and structural results*, *Journal of Machine Learning Research* **3** (2002), no. Nov 463–482.

[131] B. Neyshabur, R. Tomioka, and N. Srebro, *Norm-based capacity control in neural networks*, in *Conference on Learning Theory*, pp. 1376–1401, PMLR, 2015.

[132] S. Arora, S. Du, W. Hu, Z. Li, and R. Wang, *Fine-grained analysis of optimization and generalization for overparameterized two-layer neural networks*, in *International Conference on Machine Learning*, pp. 322–332, PMLR, 2019.

[133] N. Golowich, A. Rakhlin, and O. Shamir, *Size-independent sample complexity of neural networks*, *Information and Inference: A Journal of the IMA* **9** (2020), no. 2 473–504.

[134] G. Vardi, O. Shamir, and N. Srebro, *The sample complexity of one-hidden-layer neural networks*, in *Advances in Neural Information Processing Systems* (A. H. Oh, A. Agarwal, D. Belgrave, and K. Cho, eds.), 2022.

[135] S. Frei, N. S. Chatterji, and P. L. Bartlett, *Random feature amplification: Feature learning and generalization in neural networks*, *arXiv preprint arXiv:2202.07626* (2022).

[136] N. S. Chatterji, P. M. Long, and P. Bartlett, *When does gradient descent with logistic loss interpolate using deep networks with smoothed relu activations?*, in *Conference on Learning Theory*, pp. 927–1027, PMLR, 2021.

[137] Y. Bai and J. D. Lee, *Beyond linearization: On quadratic and higher-order approximation of wide neural networks*, in *International Conference on Learning Representations*, 2020.

[138] D. Soudry, E. Hoffer, M. S. Nacson, S. Gunasekar, and N. Srebro, *The implicit bias of gradient descent on separable data*, *The Journal of Machine Learning Research* **19** (2018), no. 1 2822–2878.

[139] K. Lyu and J. Li, *Gradient descent maximizes the margin of homogeneous neural networks*, in *International Conference on Learning Representations*, 2020.

[140] L. Chizat and F. Bach, *Implicit bias of gradient descent for wide two-layer neural networks trained with the logistic loss*, in *Conference on Learning Theory*, pp. 1305–1338, PMLR, 2020.

[141] Z. Ji and M. Telgarsky, *Directional convergence and alignment in deep learning*, *Advances in Neural Information Processing Systems* **33** (2020) 17176–17186.

[142] P. Deora, R. Ghaderi, H. Taheri, and C. Thrampoulidis, *On the optimization and generalization of multi-head attention*, *Transactions on Machine Learning Research*.

[143] C. Zhang, S. Bengio, M. Hardt, B. Recht, and O. Vinyals, *Understanding deep learning (still) requires rethinking generalization*, *Communications of the ACM* **64** (2021), no. 3 107–115.

[144] R. Bassily, M. Belkin, and S. Ma, *On exponential convergence of sgd in non-convex over-parametrized learning*, *arXiv preprint arXiv:1811.02564* (2018).

[145] S. Du, J. Lee, H. Li, L. Wang, and X. Zhai, *Gradient descent finds global minima of deep neural networks*, in *International conference on machine learning*, pp. 1675–1685, PMLR, 2019.

[146] S. Oymak and M. Soltanolkotabi, *Overparameterized nonlinear learning: Gradient descent takes the shortest path?*, in *International Conference on Machine Learning*, pp. 4951–4960, PMLR, 2019.

[147] I. M. Safran, G. Yehudai, and O. Shamir, *The effects of mild over-parameterization on the optimization landscape of shallow relu neural networks*, in *Conference on Learning Theory*, pp. 3889–3934, PMLR, 2021.

[148] S. Rosset, J. Zhu, and T. J. Hastie, *Margin maximizing loss functions*, in *NIPS*, 2003.

[149] M. Telgarsky, *Margins, shrinkage, and boosting*, in *International Conference on Machine Learning*, pp. 307–315, PMLR, 2013.

[150] K. Lyu and J. Li, *Gradient descent maximizes the margin of homogeneous neural networks*, *arXiv preprint arXiv:1906.05890* (2019).

[151] D. Zou, Y. Cao, D. Zhou, and Q. Gu, *Gradient descent optimizes over-parameterized deep relu networks*, *Machine learning* **109** (2020), no. 3 467–492.

[152] M. Schliserman and T. Koren, *Stability vs implicit bias of gradient methods on separable data and beyond*, *arXiv preprint arXiv:2202.13441* (2022).

[153] Y. Cao and Q. Gu, *Generalization bounds of stochastic gradient descent for wide and deep neural networks*, *Advances in neural information processing systems* **32** (2019).

[154] H. Taheri and C. Thrampoulidis, *On generalization of decentralized learning with separable data*, in *International Conference on Artificial Intelligence and Statistics*, pp. 4917–4945, PMLR, 2023.

[155] Y. Nesterov, *Introductory lectures on convex optimization: A basic course*, vol. 87. Springer Science & Business Media, 2003.

[156] B. Polyak, *Gradient methods for the minimisation of functionals*, *Ussr Computational Mathematics and Mathematical Physics* **3** (1963) 864–878.

[157] S. Lojasiewicz, *A topological property of real analytic subsets*, *Coll. du CNRS, Les equations aux derive es partielles* (1963).

[158] H. Taheri and C. Thrampoulidis, *Generalization and stability of interpolating neural networks with minimal width*, *Journal of Machine Learning Research* **25** (2024), no. 156 1–41.

[159] A. Rahimi and B. Recht, *Random features for large-scale kernel machines*, in *Advances in Neural Information Processing Systems* (J. Platt, D. Koller, Y. Singer, and S. Roweis, eds.), vol. 20, Curran Associates, Inc., 2007.

[160] M. Schmidt and N. L. Roux, *Fast convergence of stochastic gradient descent under a strong growth condition*, *arXiv preprint arXiv:1308.6370* (2013).

[161] S. Vaswani, F. Bach, and M. Schmidt, *Fast and faster convergence of sgd for over-parameterized models and an accelerated perceptron*, in *The 22nd international conference on artificial intelligence and statistics*, pp. 1195–1204, PMLR, 2019.

[162] A. Nedic and A. Ozdaglar, *Distributed subgradient methods for multi-agent optimization*, *IEEE Transactions on Automatic Control* **54** (2009), no. 1 48–61.

[163] A. Nedić and A. Olshevsky, *Distributed optimization over time-varying directed graphs*, *IEEE Transactions on Automatic Control* **60** (2014), no. 3 601–615.

[164] K. Yuan, Q. Ling, and W. Yin, *On the convergence of decentralized gradient descent*, *SIAM Journal on Optimization* **26** (2016), no. 3 1835–1854.

[165] X. Lian, C. Zhang, H. Zhang, C.-J. Hsieh, W. Zhang, and J. Liu, *Can decentralized algorithms outperform centralized algorithms? a case study for decentralized parallel stochastic gradient descent*, in *Advances in Neural Information Processing Systems*, pp. 5330–5340, 2017.

[166] A. Nedić and A. Olshevsky, *Stochastic gradient-push for strongly convex functions on time-varying directed graphs*, *IEEE Transactions on Automatic Control* **61** (2016), no. 12 3936–3947.

[167] Z. Jiang, A. Balu, C. Hegde, and S. Sarkar, *Collaborative deep learning in fixed topology networks*, *Advances in Neural Information Processing Systems* **30** (2017).

[168] J. Wang, V. Tantia, N. Ballas, and M. Rabbat, *Slowmo: Improving communication-efficient distributed sgd with slow momentum*, *arXiv preprint arXiv:1910.00643* (2019).

[169] A. Koloskova, T. Lin, S. U. Stich, and M. Jaggi, *Decentralized deep learning with arbitrary communication compression*, *arXiv preprint arXiv:1907.09356* (2019).

[170] A. Koloskova, N. Loizou, S. Boreiri, M. Jaggi, and S. Stich, *A unified theory of decentralized sgd with changing topology and local updates*, in *International Conference on Machine Learning*, pp. 5381–5393, PMLR, 2020.

[171] M. Assran, N. Loizou, N. Ballas, and M. Rabbat, *Stochastic gradient push for distributed deep learning*, *arXiv preprint arXiv:1811.10792* (2018).

[172] S. Pu, W. Shi, J. Xu, and A. Nedic, *Push-pull gradient methods for distributed optimization in networks*, *IEEE Transactions on Automatic Control* (2020).

[173] H. Taheri, A. Mokhtari, H. Hassani, and R. Pedarsani, *Quantized decentralized stochastic learning over directed graphs*, in *International Conference on Machine Learning*, pp. 9324–9333, PMLR, 2020.

[174] R. Xin, U. A. Khan, and S. Kar, *An improved convergence analysis for decentralized online stochastic non-convex optimization*, *IEEE Transactions on Signal Processing* **69** (2021) 1842–1858.

[175] M. T. Toghani and C. A. Uribe, *Communication-efficient distributed cooperative learning with compressed beliefs*, *IEEE Transactions on Control of Network Systems* (2022).

[176] M. T. Toghani and C. A. Uribe, *Scalable average consensus with compressed communications*, in *2022 American Control Conference (ACC)*, pp. 3412–3417, IEEE, 2022.

[177] W. Shi, Q. Ling, G. Wu, and W. Yin, *Extra: An exact first-order algorithm for decentralized consensus optimization*, *SIAM Journal on Optimization* **25** (2015), no. 2 944–966.

[178] A. Nedic, A. Olshevsky, and W. Shi, *Achieving geometric convergence for distributed optimization over time-varying graphs*, *SIAM Journal on Optimization* **27** (2017), no. 4 2597–2633.

[179] A. Koloskova, T. Lin, and S. U. Stich, *An improved analysis of gradient tracking for decentralized machine learning*, *Advances in Neural Information Processing Systems* **34** (2021).

[180] T. Lin, S. P. Karimireddy, S. Stich, and M. Jaggi, *Quasi-global momentum: Accelerating decentralized deep learning on heterogeneous data*, in *International Conference on Machine Learning*, pp. 6654–6665, PMLR, 2021.

[181] T. Sun, D. Li, and B. Wang, *Stability and generalization of decentralized stochastic gradient descent*, in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 35, pp. 9756–9764, 2021.

[182] D. Richards *et. al.*, *Graph-dependent implicit regularisation for distributed stochastic subgradient descent*, *Journal of Machine Learning Research* **21** (2020), no. 2020.

[183] D. Richards, P. Rebeschini, and L. Rosasco, *Decentralised learning with random features and distributed gradient descent*, in *International Conference on Machine Learning*, pp. 8105–8115, PMLR, 2020.

[184] Y. Sun, M. Maros, G. Scutari, and G. Cheng, *High-dimensional inference over networks: Linear convergence and statistical guarantees*, *arXiv preprint arXiv:2201.08507* (2022).

[185] H. Taheri and C. Thrampoulidis, *Fast convergence in learning two-layer neural networks with separable data*, in *AAAI Conference on Artificial Intelligence*, 2023.

[186] R. Johnson and T. Zhang, *Accelerating stochastic gradient descent using predictive variance reduction*, *Advances in neural information processing systems* **26** (2013).

[187] "Uci wine data set, web address : https://archive.ics.uci.edu/ml/datasets/wine."

[188] P. Kairouz, H. B. McMahan, B. Avent, A. Bellet, M. Bennis, A. N. Bhagoji, K. Bonawitz, Z. Charles, G. Cormode, R. Cummings, *et. al.*, *Advances and open problems in federated learning*, *Foundations and Trends® in Machine Learning* **14** (2021), no. 1–2 1–210.