Precise Tradeoffs in Adversarial Training for Linear Regression

Adel Javanmard[*], Mahdi Soltanolkotabi,[†] Hamed Hassani[‡]

February 25, 2020

Abstract

Despite breakthrough performance, modern learning models are known to be highly vulnerable to small adversarial perturbations in their inputs. While a wide variety of recent *adversarial training* methods have been effective at improving robustness to perturbed inputs (robust accuracy), often this benefit is accompanied by a decrease in accuracy on benign inputs (standard accuracy), leading to a tradeoff between often competing objectives. Complicating matters further, recent empirical evidence suggest that a variety of other factors (size and quality of training data, model size, etc.) affect this tradeoff in somewhat surprising ways. In this paper we provide a precise and comprehensive understanding of the role of adversarial training in the context of linear regression with Gaussian features. In particular, we characterize the fundamental tradeoff between the accuracies achievable by any algorithm regardless of computational power or size of the training data. Furthermore, we precisely characterize the standard/robust accuracy and the corresponding tradeoff achieved by a contemporary mini-max adversarial training approach in a high-dimensional regime where the number of data points and the parameters of the model grow in proportion to each other. Our theory for adversarial training algorithms also facilitates the rigorous study of how a variety of factors (size and quality of training data, model overparametrization etc.) affect the tradeoff between these two competing accuracies.

keywords. Tradeoffs in Adversarial Training, High-dimensional Statistics, Gaussian processes, Linear Regression

1 Introduction

Recent advances in machine learning and deep learning in particular, have led to trained models with breakthrough performance in a variety of applications spanning visual object classification to speech recognition and natural language processing. Despite wide empirical success, these modern learning models are known to be highly vulnerable to small adversarial perturbations to their inputs [BCM+13, SZS+14]. For instance, in the context of image classification even small perturbations of the image, which are imperceptible to a human, can lead to incorrect classification by these models. As these modern inferential techniques begin to be deployed in applications such as autonomous or recognition systems in which safety, reliability, and security are crucial, it is increasingly important to ensure trained models are robust against abrupt or adversarial perturbations to the input.

[*]Data Science and Operations Department, Marshall School of Business, University of Southern California, Los Angeles, CA

[†]Ming Hsieh Department of Electrical and Computer Engineering, University of Southern California, Los Angeles, CA

[‡]Department of Electrical and Systems Engineering, University of Pennsylvania, Philadelphia, PA

To mitigate the effect of adversarial perturbations, a wide variety of *adversarial training* methods have been developed [GSS15, KGB16, MMS+18, RSL18, WK18] which often involve augmenting the training loss so as to become more robust to input perturbations. While adversarial training methods have been rather successful at improving the accuracy of the trained model on adversarially perturbed inputs (*robust accuracy*), often this benefit comes at the cost of decreasing accuracy on natural unperturbed inputs (*standard accuracy*) [MMS+18]. Therefore, it is crucial to understand the tradeoff between robust and standard accuracy with adversarial training. Complicating matters further, recent empirical evidence suggest that a variety of other factors affect this tradeoff in somewhat surprising ways. For instance, experiments in [TSE+18] demonstrate that while adversarial training typically has a negative effect on standard accuracy, it outperforms non-adverserial training methods when there are only a few training samples. Perhaps surprisingly, the recent paper by [RXY+19] suggests that in some cases the tradeoff between standard and robust accuracy can be mitigated with additional unlabeled data. Towards demystifying these empirical phenomena, in this paper we aim to precisely characterize the role of adversarial training by focusing on the following key questions:

> *What is the fundamental tradeoff between robust and standard accuracies in both finite and infinite data limits? How can we algorithmically achieve this tradeoff and what is the role of adversarial training? What is the effect of the size/quality of the data on this tradeoff? How does the model size (e.g. overparametrization) change this tradeoff?*

A few recent papers have begun to answer some of these questions in specific settings [TSE+18, ZYJ+19, RXY+19]. See Section 4 for a detailed discussion. Despite this interesting recent progress, a comprehensive understanding of the role of adversarial training and how it precisely affects the aforementioned tradeoffs remains largely mysterious. In this paper we aim to provide a precise characterization of the role of adversarial training by focusing on the simple yet foundational problem of linear regression.

Contributions. We formally introduce the linear regression problem with adversarially perturbed inputs in Section 2 and address the questions above in this setting.

- We characterize the fundamental tradeoff between standard risk[1] (SR) and adversarial risk (AR) achievable by any algorithm regardless of the computational power and the size of the available training data (see Section 3.1). This is carried out by deriving the asymptotic expressions of standard and adversarial risks, and analysing the Pareto optimal points of a two dimensional region consisting of all the achievable $(\mathsf{SR}, \mathsf{AR})$ pairs. This analysis clearly demonstrates the existence of a non-trivial tradeoff between the two risks in linear regression as depicted in Figure 1.

- In Section 3.2, we turn our attention to modern adversarial training algorithms and provide a precise characterizition of the standard and adversarial risks achieved by them. This is carried out in a high-dimensional regime where the size of the training data n and the number of parameters p grow proportional to each other with their ratio $n/p \to \delta$ for fixed $\delta \in (0, +\infty)$. A key ingredient of our analysis is a powerful extension of a classical Gaussian process inequality [Gor88] known as the Convex Gaussian Minimax Theorem developed in [TOH15] and further extended in [TAH18, DKT19].

[1] Since we focus on a regression problem henceforth we focus on risk in lieu of accuracy.

- Our precise characterization of the standard and robust risks for adversarial training algorithms allows us to rigorously study a variety of phenomena. First, we study the tradeoffs between standard and adversarial risks for a contemporary adversarial training algorithm and show that as the limiting ratio $n/p \to \delta$ between the number of training data n and number of parameters p grows, the algorithmic tradeoff curve approaches the fundamental (Pareto-optimal) tradeoff curve. These findings are manifested empirically in Figure 1. We also characterize the effect of the size of the training data and model overparametrization (see Section 3.3). We argue analytically and empirically that in the overparametrized regime (i.e. when $\delta < 1$) adversarial training helps improve standard risk (compared to normal training). However, as the size of training data grows (i.e. δ becomes large) adversarial training effectively hurts standard risk. In short, adversarial training improves generalization in the overparametrized regime, but effectively hurts generalization in the sufficiently underparametrized regime. Finally, in Section 3.4 we demonstrate and prove the emergence of a phenomenon in adversarial training which is similar to the so-called double-descent phenomenon. When traditional training is used, the double-descent phenomena demonstrates that increasing the model complexity beyond a certain interpolation threshold always improves generalization. We show that the double-descent behavior continues to hold with adversarial training. However, for linear regression model considered in this paper, the global minimum of the risk is achieved under the interpolation threshold *whose value changes with* ε. Our theory also allows us to study how the adversarial training affects the interpolation threshold.

2 Problem formulation

In a typical supervised learning problem, we wish to fit a function $f_{\boldsymbol{\theta}}$, parameterized by $\boldsymbol{\theta} \in \mathbb{R}^p$ to a training data set of n input-output pairs $\{(\boldsymbol{x}_i, y_i)\}_{i=1}^n$ drawn i.i.d. from some common law $\mathcal{P}$. The fitting problem often consists of finding a parameter $\widehat{\boldsymbol{\theta}}$ that minimizes the empirical risk

$$\widehat{\boldsymbol{\theta}} \in \arg\min_{\boldsymbol{\theta}\in\mathbb{R}^p} \quad \frac{1}{n}\sum_{i=1}^n \ell\left(\boldsymbol{x}_i, y_i; \boldsymbol{\theta}\right) := \frac{1}{n}\sum_{i=1}^n \widetilde{\ell}(f_{\boldsymbol{\theta}}(\boldsymbol{x}_i), y_i)\,, \tag{2.1}$$

over the space of all parameters $\boldsymbol{\theta}$. The loss $\widetilde{\ell}(f_{\boldsymbol{\theta}}(\boldsymbol{x}), y)$ measures discrepancy between the output (or label) y and the prediction $f_{\boldsymbol{\theta}}(\boldsymbol{x})$. The goal is of course to learn models that perform well on the yet unseen test data that is also generated from the same distribution $\mathcal{P}$. In particular, the empirical risk above serves as a surrogate for the population risk (loss) $\mathbb{E}_{(\boldsymbol{x},y)\sim\mathcal{P}}[\ell(\boldsymbol{x}, y; \boldsymbol{\theta})]$.

In practice, many models trained by following this paradigm are often highly vulnerable to adversarial perturbations with many well documented examples in deep learning. This observation has given rise to a surge of interest in both, finding such perturbations (a.k.a adversarial attacks) and also learning models that are robust against such perturbations (a.k.a. adversarial training). A line of recent work [TSE+18, MMS+17] propose training approaches that demonstrate promising empirical performance against adversarial perturbations. Motivated by applications in image processing, these papers consider an adversarial attack model where for a predefined perturbation set $\mathcal{S}$, the adversary has the power of perturbing each data point $\boldsymbol{x}$ by adding an element of $\mathcal{S}$. Then an estimator $\widehat{\boldsymbol{\theta}}^{\mathcal{S}}$ is constructed by solving a saddle point problem that takes into account such manipulative power for the adversary:

$$\widehat{\boldsymbol{\theta}}^{\mathcal{S}} \in \arg\min_{\boldsymbol{\theta}\in\mathbb{R}^p} \max_{\boldsymbol{\delta}_i\in\mathcal{S}} \quad \frac{1}{n}\sum_{i=1}^n \ell\left(\boldsymbol{x}_i + \boldsymbol{\delta}_i, y_i; \boldsymbol{\theta}\right)\,. \tag{2.2}$$

To evaluate the performance of such an estimator, in this paper we consider two metrics of particular interest, *standard risk* and *adversarial risk*.

Standard risk. This is the expected prediction loss of an estimator $\widehat{\boldsymbol{\theta}}$ on an uncorrupted test data point that is generated from the same distribution as the training data. Namely,

$$\mathsf{SR}(\widehat{\boldsymbol{\theta}}) := \frac{1}{p}\,\mathbb{E}\left[\ell\left(\boldsymbol{x}, y; \widehat{\boldsymbol{\theta}}\right)\right] \quad \text{where} \quad (\boldsymbol{x}, y) \sim \mathcal{P}\,. \tag{2.3}$$

Adversarial risk. This is the expected prediction loss of an estimator $\widehat{\boldsymbol{\theta}}$ on an adversarially corrupted test data point according to the attack model (2.2). Namely,

$$\mathsf{AR}(\widehat{\boldsymbol{\theta}}) := \frac{1}{p}\,\mathbb{E}\Big[\max_{\boldsymbol{\delta}\in\mathcal{S}} \ell\left(\boldsymbol{x}+\boldsymbol{\delta}, y; \widehat{\boldsymbol{\theta}}\right)\Big] \quad \text{where} \quad (\boldsymbol{x}, y) \sim \mathcal{P}\,. \tag{2.4}$$

Stated differently, the adversarial risk measures how well the estimator $\widehat{\boldsymbol{\theta}}$ performs in predicting the true label when it is fed with an adversarially corrupted test data point. We note that the factor $1/p$ is the proper scaling so the risk has a finite limit under our asymptotic regime.

Focusing on linear regression, in this paper we aim to derive asymptotically exact characterizations of these two metrics and study the tradeoff achieved by the class of estimators $\widehat{\boldsymbol{\theta}}^{\mathcal{S}}$ of the form (2.2). These characterizations will also enable us to study the effect of various quantities (e.g. size and quality of the training data, model size, etc.) on the trade-off between statistical and adversarial risk. Specifically, we consider the linear regression model below.

Definition 2.1 (Linear Regression Setting)**.** *We consider standard Gaussian linear regression model with the training data consisting of n i.i.d pairs $(\boldsymbol{x}_i, y_i)$, with $\boldsymbol{x}_i \sim \mathsf{N}(0, \boldsymbol{I}_p)$ representing the features and $y_i \in \mathbb{R}$ the corresponding label given by* [2]

$$y_i = \langle \boldsymbol{x}_i, \boldsymbol{\theta}_0 \rangle + w_i \quad \textit{where} \quad w_i \sim \mathsf{N}(0, \sigma_0^2)\,. \tag{2.5}$$

We also focus on training linear models of the form $f_{\boldsymbol{\theta}}(\boldsymbol{x}) = \langle \boldsymbol{x}, \boldsymbol{\theta} \rangle$ via a quadratic loss $\ell(\boldsymbol{x}, y; \boldsymbol{\theta}) = \frac{1}{2}(y - \langle \boldsymbol{x}, \boldsymbol{\theta} \rangle)^2$ and consider perturbation sets of the form $\mathcal{S} := \{\boldsymbol{\delta} \in \mathbb{R}^p : \|\boldsymbol{\delta}\|_{\ell_2} \le \epsilon\}$ where ε is a measure of the adversary's power. To make the dependence on ε explicit in our notation, we replace $\widehat{\boldsymbol{\theta}}^{\mathcal{S}}$ for this choice of $\mathcal{S}$ by $\widehat{\boldsymbol{\theta}}^{\varepsilon}$. In this case (2.2) takes the form

$$\widehat{\boldsymbol{\theta}}^{\varepsilon} \in \arg\min_{\boldsymbol{\theta}\in\mathbb{R}^p} \max_{\|\boldsymbol{\delta}_i\|_{\ell_2} \le \varepsilon} \frac{1}{2n} \sum_{i=1}^{n} \left(y_i - \langle \boldsymbol{x}_i + \boldsymbol{\delta}_i, \boldsymbol{\theta} \rangle\right)^2 . \tag{2.6}$$

Next we formally introduce the asymptotic regime of interest in this paper.

Asymptotic regime. For a given sample size n, we define an *instance* of the standard Gaussian model by a tuple $(\boldsymbol{\theta}_0, p, \sigma_0)$, with $\boldsymbol{\theta}_0 \in \mathbb{R}^p$, $p \in \mathbb{N}$ and $\sigma_0 \in \mathbb{R}_{\ge 0}$. We consider sequence of instances of the Gaussian model indexed by the sample size n.

Definition 2.2. *The sequence of instances $\{\boldsymbol{\theta}_0(n), p(n), \sigma_0(n)\}_{n\in\mathbb{N}}$ indexed by n is called a converging sequence if:*

[2] We note that our analysis in Section 6 can be extended to general Gaussian linear regression where $\boldsymbol{x}_i \sim \mathsf{N}(0, \boldsymbol{\Sigma})$. This however requires more involved derivations that are not included in this version.

- *We have* $\frac{n}{p} \to \delta \in (0, \infty)$ *and* $\frac{\sigma_0^2(n)}{p} \to \sigma^2$ *as* $n \to \infty$.
- *Empirical second moment of the signal converges, i.e.,* $\frac{1}{p}\sum_{i=1}^{p} \theta_{0,i}(n)^2 \to V^2 < \infty$, *as* $n \to \infty$.

In summary, we have introduced the following notations and terms which will be used throughout the paper: the dimension p, number of training data points n, overparametrization parameter $\delta = n/p$, normalized noise power σ^2, normalized norm of the true model V^2, and the adversary's power ε.

3 Main Results

In this paper we wish to understand fundamental tradeoffs between standard and adversarial risks as well as what can be achieved by modern adversarial training approaches. In Section 3.1 we characterize the fundamental tradeoff between standard and adversarial risk achievable by any algorithm regardless of the computational power and the size of the available training data. Then in Section 3.2 we turn our attention to precisely characterizing the standard and adversarial accuracy tradeoffs achieved by modern adversarial training algorithms of the form (2.2). This is carried out in a high-dimensional regime where the size of the training data n and the number of parameters p grow proportional to each other with their ratio $n/p \to \delta$ for fixed $\delta \in (0, +\infty)$. Next, in Section 3.3 we focus on studying the role of that the size of the training data plays and how it affects the standard accuracy. Finally, in Section 3.4 we prove the emergence of a phenomena in adversarial training similar to the so-called double-descent phenomena without adversarial training.

3.1 Fundamental tradeoffs between standard and adversarial risk

Motivated by the conflict observed between standard and adversarial risk in modern adversarial training [MMS+18], we first wish to understand the fundamental tradeoffs that can be achieved between the two objectives. That is, the optimal tradeoff that can be achieved between standard and adversarial risk objectives for *any estimator* $\widehat{\boldsymbol{\theta}}$ even with access to infinite computational power and infinite training data. We discuss the tradeoffs achievable by specific algorithms with finite training data in the next section.

$(\mathsf{SR}, \mathsf{AR})$ Region and its Pareto Optimal Curve: As discussed previously in Section 2 for an estimator $\widehat{\boldsymbol{\theta}}$ we use $\mathsf{SR}(\widehat{\boldsymbol{\theta}})$ and $\mathsf{AR}(\widehat{\boldsymbol{\theta}})$ to denote the standard and adversarial risks achieved by $\widehat{\boldsymbol{\theta}}$. Thus, for any estimator $\widehat{\boldsymbol{\theta}}$ we obtain a point $(\mathsf{SR}(\widehat{\boldsymbol{\theta}}), \mathsf{AR}(\widehat{\boldsymbol{\theta}}))$ in the 2-d plane. We refer to the set of all such points, for all $\widehat{\boldsymbol{\theta}} \in \mathbb{R}^p$, as the $(\mathsf{SR}, \mathsf{AR})$ region. To obtain the optimal tradeoff between standard and adversarial risks we need to characterize the Pareto-optimal points of this region.[3]

In the linear regression setting of this paper the expressions of standard accuracy (2.3) and adversarial accuracy (2.4) are convex functions of $\boldsymbol{\theta}$. Therefore, using standard results in multi-objective optimization we can derive all the Pareto optimal points of the $(\mathsf{SR}, \mathsf{AR})$ region, by minimizing a weighted combination of these two accuracies for different weights λ.

$$\boldsymbol{\theta}^{\lambda} = \arg\min_{\boldsymbol{\theta}} \; \lambda \overbrace{\mathbb{E}\left\{(y - \langle \boldsymbol{x}, \boldsymbol{\theta}\rangle)^2\right\}}^{\text{standard risk}} + \overbrace{\mathbb{E}\left\{\max_{\|\boldsymbol{\delta}\|_{\ell_2} \le \varepsilon_{\text{test}}} (y - \langle \boldsymbol{x} + \boldsymbol{\delta}, \boldsymbol{\theta}\rangle)^2\right\}}^{\text{adversarial risk}} . \tag{3.1}$$

[3] Given a region $\mathcal{C} \in \mathbb{R}^2$, a point $(x, y) \in \mathcal{C}$ is Pareto optimal if there exists no other point $(x', y') \in \mathcal{C}$ s.t. $x' \le x$ and $y' \le y$.

The Pareto-optimal curve is then given by $\{(\mathsf{SR}(\boldsymbol{\theta}^\lambda), \mathsf{AR}(\boldsymbol{\theta}^\lambda) : \lambda \geq 0\}$.

Analytical Expression of the Optimal Tradeoffs: Before we proceed to calculate $\boldsymbol{\theta}^\lambda$, we derive the standard and adversarial risks ($\mathsf{SR}(\widehat{\boldsymbol{\theta}})$ and $\mathsf{AR}(\widehat{\boldsymbol{\theta}})$) as a functions of $\boldsymbol{\theta}_0$ and σ_0^2 in the Gaussian linear regression model. We defer the proof of this Lemma to Section 6.7.1.

Lemma 3.1. *Consider the linear regression setting of Definition 2.1. For a given estimator $\widehat{\boldsymbol{\theta}}$ the standard risk (2.3) is equal to*

$$\mathsf{SR}(\widehat{\boldsymbol{\theta}}) := \frac{1}{p}\,\mathbb{E}\left[(y - \langle \boldsymbol{x}, \widehat{\boldsymbol{\theta}}\rangle)^2\right] = \frac{\sigma_0^2}{p} + \frac{1}{p}\left\|\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0\right\|_{\ell_2}^2,$$

Furthermore, the adversarial risk (2.4) with a corruption level of $\varepsilon_{\text{test}}$ is equal to

$$\begin{aligned}\mathsf{AR}(\widehat{\boldsymbol{\theta}}) :=& \frac{1}{p}\,\mathbb{E}\left[\max_{\|\boldsymbol{\delta}\|_{\ell_2} \leq \varepsilon_{\text{test}}} (y - \langle \boldsymbol{x} + \boldsymbol{\delta}, \widehat{\boldsymbol{\theta}}\rangle)^2\right] \\ =& \frac{1}{p}\left(\sigma_0^2 + \left\|\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0\right\|_{\ell_2}^2 + \varepsilon_{\text{test}}^2 \left\|\widehat{\boldsymbol{\theta}}\right\|_{\ell_2}^2\right) + 2\sqrt{\frac{2}{\pi}}\frac{\varepsilon_{\text{test}}}{\sqrt{p}}\left\|\widehat{\boldsymbol{\theta}}\right\|_{\ell_2}\left(\frac{\sigma_0^2}{p} + \frac{1}{p}\left\|\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0\right\|_{\ell_2}^2\right)^{1/2}.\end{aligned}$$

With a precise expression of the standard and adversarial risk in hand our next theorem characterizes the solution $\boldsymbol{\theta}^\lambda$ of the optimization problem (3.1) which in conjunction with Lemma 3.1 determines the Pareto-optimal tradeoff curve. We defer the proof of this result to Section 6.7.2.

Proposition 3.2. *Under the linear regression setting of Definition 2.1, the solution $\boldsymbol{\theta}^\lambda$ of the optimization problem (3.1) is given by* $\boldsymbol{\theta}^\lambda = (1 + \gamma_0^\lambda)^{-1}\boldsymbol{\theta}_0$,

with γ_0^λ the fixed point of the following two equations:

$$\gamma_0^\lambda = \frac{\varepsilon_{\text{test}}^2 + \sqrt{\frac{2}{\pi}}\varepsilon_{\text{test}} A^\lambda}{1 + \lambda + \sqrt{\frac{2}{\pi}}\frac{\varepsilon_{\text{test}}}{A^\lambda}} \quad \text{and} \quad A^\lambda = \frac{1}{\|\boldsymbol{\theta}_0\|_{\ell_2}}\left((1+\gamma_0^\lambda)^2\sigma_0^2 + (\gamma_0^\lambda)^2\|\boldsymbol{\theta}_0\|_{\ell_2}^2\right)^{1/2}.$$

In Figure 1 we plot the Pareto optimal curve in the $(\mathsf{SR}, \mathsf{AR})$ plane in black for an instance where $\varepsilon_{\text{test}} = 0.5$ and the normalized norm of the true model and the noise power are both equal to one ($\sigma = V = 1$). This curve serves as a fundamental limit on the performance of any algorithm even with access to infinite data and computational power. This figure also contains algorithmic tradeoffs which we discuss in further detail in the next section. In particular, in the next section we precisely characterize the SR-AR tradeoff achieved by a specific adversarial training algorithm.

3.2 Algorithmic tradeoffs between standard and adversarial risks

Given the fundamental tradeoff of the previous section, the natural question that arises is whether it is possible to achieve this tradeoff algorithmically with only finite data and computational power? Specifically, what is the tradeoff achieved by common adversarial training algorithms? In this section we consider the class of estimators $\widehat{\boldsymbol{\theta}}^\varepsilon$ constructed through the saddle point problem (2.6) for various values ε at training i.e. $\{\widehat{\boldsymbol{\theta}}^\varepsilon : \varepsilon \geq 0\}$. We wish to precisely derive the tradeoff curve between the standard and the adversarial risks achieved by this class of estimators. We refer to such curve as *algorithmic* tradeoff curve since it corresponds to the specific class of saddle point estimators as opposed to the Pareto optimal trade off curves studied in Section 3.1 which serve as lowerbound for any estimator. To avoid any confusion about the tradeoffs discussed we would like to emphasize that:

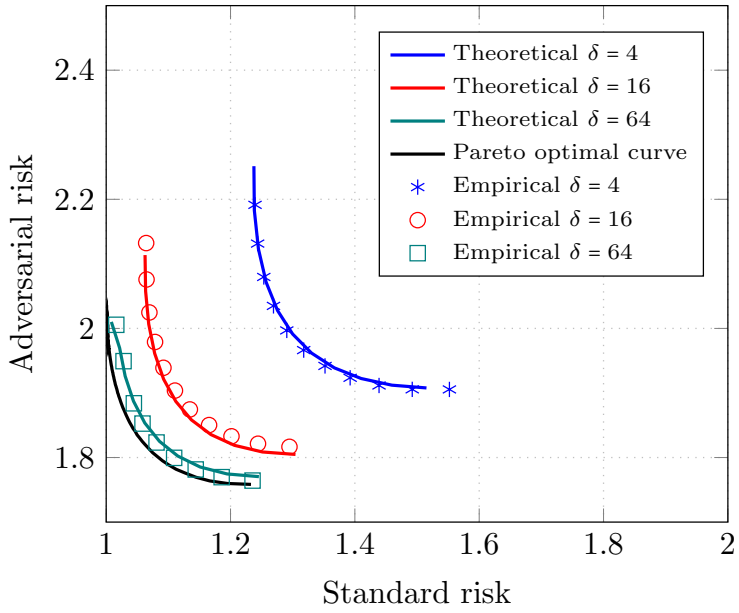


Figure 1: Pareto optimal curve along with algorithmic curves for several values of δ. As δ grows the algorithmic tradeoff curves approach the fundamental Pareto optimal curve. The dots correspond to the empirical data obtained by solving for the optimal solution $\widehat{\boldsymbol{\theta}}^{\varepsilon}$ of (2.6) using gradient descent and then computing $(\mathsf{SR}(\widehat{\boldsymbol{\theta}}^{\varepsilon}), \mathsf{AR}(\widehat{\boldsymbol{\theta}}^{\varepsilon}))$ from Lemma 3.1 with different values of ε. Here, $\sigma = 1$, $V = 1$, $p = 1000$, and $\varepsilon_{\text{test}} = 0.5$.

(i) In the *training* phase, we are *varying* the adversarial power ε, and accordingly, obtain a range of estimators $\widehat{\boldsymbol{\theta}}^{\varepsilon}$ by solving (2.6).

(ii) At *test* time, the adversarial power is fixed to a given value $\varepsilon_{\text{test}}$ and we will measure the (expected) standard and adversarial risks of the trained estimators $\widehat{\boldsymbol{\theta}}^{\varepsilon}$ with respect to the true adversarial power $\varepsilon_{\text{test}}$. By varying ε at training time, we expect to sweep a tradeoff between standard and adversarial risks, i.e. estimators $\widehat{\boldsymbol{\theta}}^{\varepsilon}$ with large ε should have a smaller adversarial risk but higher standard risk, and estimators with smaller ε should behave the opposite.

Analytical Expression of the Algorithmic Tradeoffs. Our goal for the rest of this section is to analytically derive the algorithmic tradeoffs in terms of the overparametrization parameter $n/p \to \delta \in (0, \infty)$ which represents the number of training data points per dimension. We focus on converging sequences of Gaussian model instances as described in Definition 2.2. Recall that By virtue of Lemma 3.1, in order to derive the asymptotic standard and adversarial risk of $\widehat{\boldsymbol{\theta}}^{\varepsilon}$, it suffices to obtain an exact characterization of the asymptotic error $\lim_{n\to\infty} \frac{1}{p} \left\| \widehat{\boldsymbol{\theta}}^{\varepsilon} - \boldsymbol{\theta}_0 \right\|_{\ell_2}^2$ and the asymptotic estimator norm $\lim_{n\to\infty} \frac{1}{p} \left\| \widehat{\boldsymbol{\theta}}^{\varepsilon} \right\|_{\ell_2}^2$. This is the subject of the next theorem formally proven in Section 6.8.1.

Theorem 3.3. *Let $\{(\boldsymbol{\theta}_0(n), p(n), \sigma_0(n))\}_{n\in\mathbb{N}}$ be a converging sequence of instances of the standard Gaussian design model. Consider the linear regression model (2.5) and let $\widehat{\boldsymbol{\theta}}^{\varepsilon}$ be a solution of (2.6). If $\varepsilon, \delta > 0$ or $\varepsilon = 0$, $\delta > 1$, then*

(a) The following convex-concave minimax scalar optimization has a unique solution $(\alpha_*, \beta_*, \gamma_*, \tau_{h*}, \tau_{g*})$*:*

$$\max_{0\le\beta\le K_\beta} \sup_{\gamma,\tau_h\ge 0} \min_{0\le\alpha\le K_\alpha} \min_{\tau_g\ge 0} \quad D(\alpha,\beta,\gamma,\tau_h,\tau_g)\,, \quad \textit{where} \tag{3.2}$$

$$\begin{aligned} D(\alpha,\beta,\gamma,\tau_h,\tau_g) &:= \frac{\delta\beta}{2(\tau_g+\beta)}\left(\alpha^2+\sigma^2\right) \\ &+ \delta\mathbb{1}_{\left\{\frac{\gamma(\tau_g+\beta)}{\delta\varepsilon\beta\sqrt{\alpha^2+\sigma^2}}>\sqrt{\frac{2}{\pi}}\right\}}\frac{\beta^2(\alpha^2+\sigma^2)}{2\tau_g(\tau_g+\beta)}\left(\mathrm{erf}\left(\frac{\tau_*}{\sqrt{2}}\right)-\frac{\gamma(\tau_g+\beta)}{\delta\varepsilon\beta\sqrt{\alpha^2+\sigma^2}}\,\tau_*\right) \\ &- \frac{\alpha}{2\tau_h}(\gamma^2+\beta^2)+\gamma\sqrt{\frac{\alpha^2\beta^2}{\tau_h^2}+V^2}-\frac{\alpha\tau_h}{2}+\frac{\beta\tau_g}{2}\,, \end{aligned} \tag{3.3}$$

and τ_* *is the unique solution to*

$$\frac{\gamma(\tau_g+\beta)}{\delta\varepsilon\beta\sqrt{\alpha^2+\sigma^2}}-\frac{\beta}{\tau_g}\tau-\tau\cdot\mathrm{erf}\left(\frac{\tau}{\sqrt{2}}\right)-\sqrt{\frac{2}{\pi}}e^{-\frac{\tau^2}{2}}=0$$

(b) It holds in probability that $\lim_{n\to\infty}\frac{1}{p}\left\|\widehat{\boldsymbol{\theta}}^{\varepsilon}-\boldsymbol{\theta}_0\right\|_{\ell_2}^2=\alpha_*^2$.

(c) It holds in probability that

$$\lim_{n\to\infty}\frac{1}{\sqrt{p}}\left\|\widehat{\boldsymbol{\theta}}^{\varepsilon}\right\|_{\ell_2}=\frac{\beta_*\tau_*\sqrt{\alpha_*^2+\sigma^2}}{\varepsilon\tau_{g*}}\,. \tag{3.4}$$

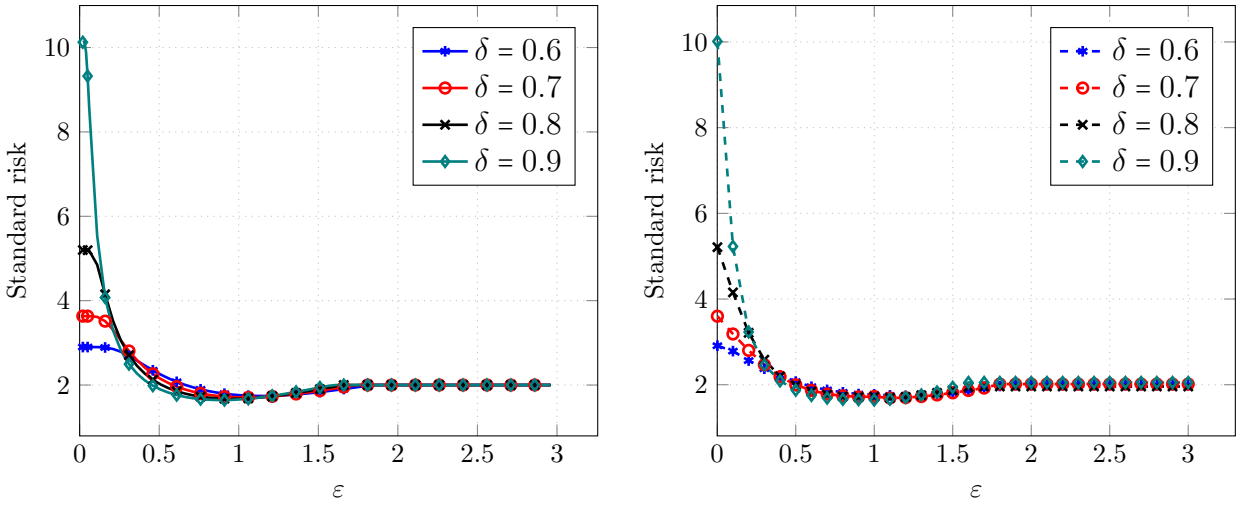
We note that the loss (2.6) and its optimal solution are a rather complicated and high-dimensional function of the features/label pairs $\{(\boldsymbol{x}_i,y_i)\}_{i=1}^n$. Nevertheless the Theorem above provides a precise characterization of its properties using a 5 dimensional convex-concave mini-max optimization problem! Such a precise characterization allows us to provide a precise understanding of the standard and adversarial accuracies. In particular, combining Theorem 3.3 (parts (b)-(c)) with Lemma 3.1 we can obtain the asymptotic values of $\mathsf{SR}(\widehat{\boldsymbol{\theta}}^{\varepsilon})$ and $\mathsf{AR}(\widehat{\boldsymbol{\theta}}^{\varepsilon})$, and derive the algorithmic tradeoff curve achieved by the class $\{\widehat{\boldsymbol{\theta}}^{\varepsilon}:\varepsilon\ge 0\}$ as ε varies (discussed in the next corollary proven in Section 6.8.2).

Corollary 3.4. *Let* $\{(\boldsymbol{\theta}_0(n),p(n),\sigma_0(n))\}_{n\in\mathbb{N}}$ *be a converging sequence of instances of the standard Gaussian design model. Consider the linear regression model (2.5) and let* $\widehat{\boldsymbol{\theta}}^{\varepsilon}$ *be a solution of (2.6). Further assume that* $\varepsilon,\delta>0$ *or* $\varepsilon=0$, $\delta>1$. *Also denote* $(\alpha_*,\beta_*,\gamma_*,\tau_{h*},\tau_{g*})$ *as the optimal solutions of the minimax optimization (6.22). Then, the following identities hold in probability:*

$$\lim_{n\to\infty}\mathsf{SR}(\widehat{\boldsymbol{\theta}}^{\varepsilon})=\sigma^2+\alpha_*^2\,, \tag{3.5}$$

$$\lim_{n\to\infty}\mathsf{AR}(\widehat{\boldsymbol{\theta}}^{\varepsilon})=\left(\sigma^2+\alpha_*^2+\varepsilon_{\mathrm{test}}^2(\alpha_*^2+\sigma^2)\left(\frac{\beta_*\tau_*}{\varepsilon\tau_{g*}}\right)^2\right)+2\sqrt{\frac{2}{\pi}}\frac{\varepsilon_{\mathrm{test}}\beta_*\tau_*}{\varepsilon\tau_{g*}}(\sigma^2+\alpha_*^2)\,. \tag{3.6}$$

The corollary above provides a precise characterization of the standard and adversarial accuracy achieved by the adversarial training algorithm consisting of running gradient descent on the saddle point problem (2.6). In Figure 1, we plot the algorithmic tradeoff curve for several values of δ as well as the empirical values obtained by running gradient descent. As we observe, our theoretical prediction and the empirical values are rather close match even for moderately large parameter values ($p=1000$). Such a precise characterization allows us to rigorously study a variety of phenomena. We



(a) Theoretical curves

(b) Empirical curves

Figure 2: Standard risk ($\mathsf{SR}(\widehat{\boldsymbol{\theta}}^{\varepsilon})$) versus ε for several values of $\delta < 1$. Left panel corresponds to the theoretical curve obtained by Theorem 3.3 (with $\sigma = 1$ and $V = 1$), and the right panel corresponds to the empirical results (with $\sigma = 1$ and $\theta_{0,i} \sim \mathsf{N}(0,1)$). The empirical results are averaged over 100 different realizations of noise and features. As δ grows to one, we observe a faster decay in the standard risk with respect to the adversarial power ε.

mention one such phenomena below and discuss others in the coming sections. The plots in Figure 1 clearly show that when δ grows the algorithmic tradeoff curve approaches the Pareto-optimal tradeoff curve. In other words, one can achieve optimal tradeoff of standard and adversarial risks by the specific class of estimators $\widehat{\boldsymbol{\theta}}^{\varepsilon}$ constructed by the saddle point problem (2.6). This observation is formally stated in the next theorem with the proof deferred to Section 6.8.3.

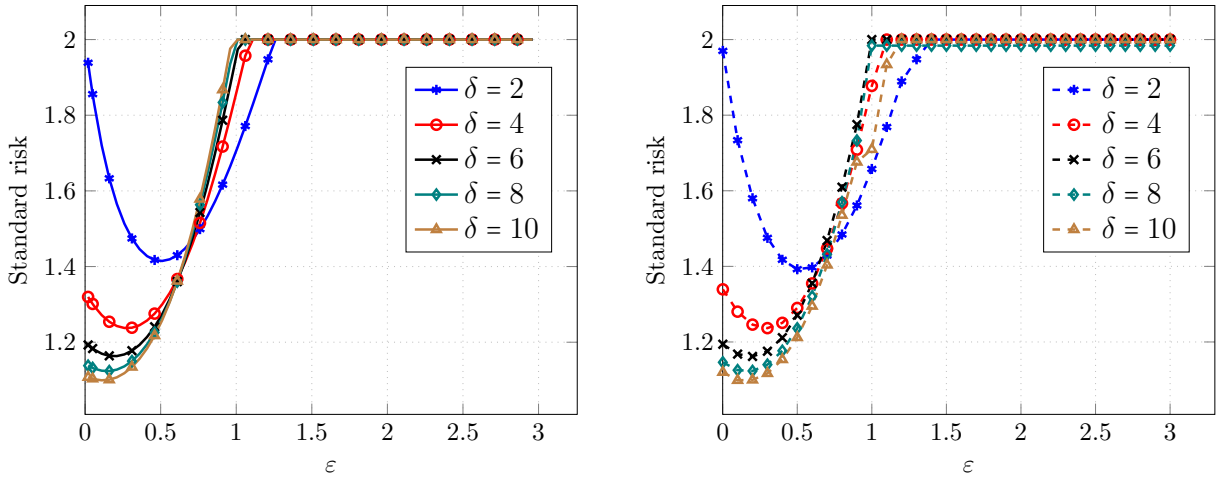
Theorem 3.5. *Let $\{(\boldsymbol{\theta}_0(n), p(n), \sigma_0(n))\}_{n\in\mathbb{N}}$ be a converging sequence of instances of the standard Gaussian design model. Consider the linear regression model (2.5), and let $\widehat{\boldsymbol{\theta}}^{\varepsilon}$ be a solution of (2.6) and $\boldsymbol{\theta}^{\lambda}$ the solution of (3.1). Then for any $\lambda \geq 0$ there exists $\varepsilon = \varepsilon(\sigma, V, \varepsilon_{\text{test}}, \lambda)$, such that*

$$\lim_{\delta\to\infty}\lim_{n\to\infty}\mathsf{SR}(\widehat{\boldsymbol{\theta}}^{\varepsilon}) = \lim_{p\to\infty}\mathsf{SR}(\boldsymbol{\theta}^{\lambda})\,, \qquad \lim_{\delta\to\infty}\lim_{n\to\infty}\mathsf{AR}(\widehat{\boldsymbol{\theta}}^{\varepsilon}) = \lim_{p\to\infty}\mathsf{AR}(\boldsymbol{\theta}^{\lambda})\,. \tag{3.7}$$

The theorem above formally proves that in the infinite data limit ($\delta \to +\infty$) one of the commonly used adversarial training algorithms achieves the optimal tradeoff between standard and robust accuracies.

3.3 The role of the size of the training data and overparameterization

As discussed, our precise understanding of the optimal solution of adversarial training allows us to precisely characterize the effect of various phenomena. In particular in this section we focus on the role of the size of the training data. We begin by considering the common scenario in modern learning where trained models often consist of more parameters than the training data set. In Figure 2-(a) we plot the standard risk, using Theorem 3.3 Part (b), versus ε for different values of



(a) Theoretical curves

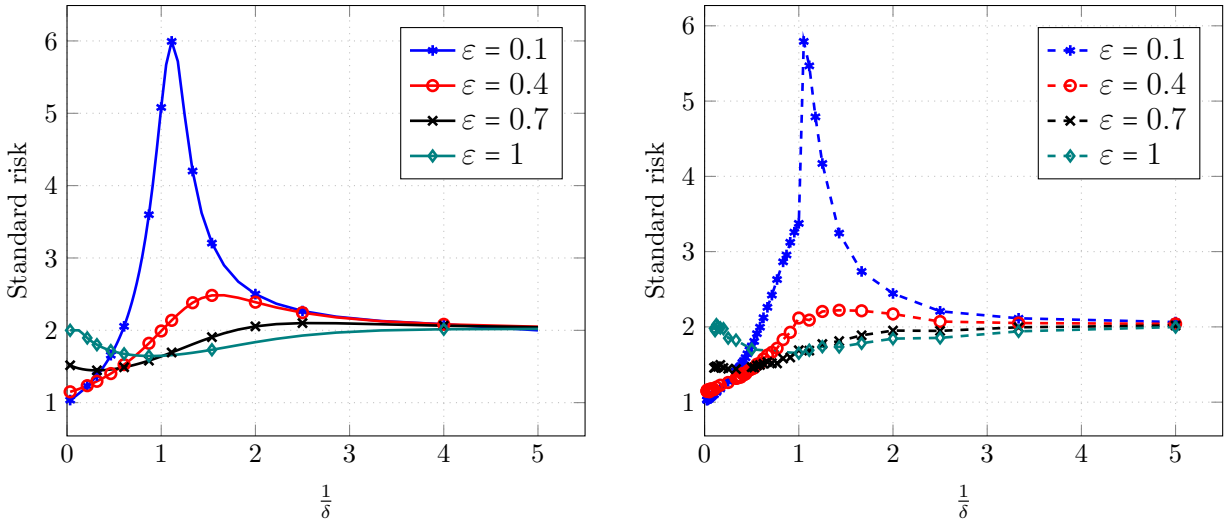
(b) Empirical curves

Figure 3: Standard risk ($\mathsf{SR}(\widehat{\boldsymbol{\theta}}^{\varepsilon})$) versus ε for several values of $\delta > 1$. Left panel corresponds to the theoretical curve obtained by Theorem 3.3 (with $\sigma = 1$ and $V = 1$), and the right panel corresponds to the empirical results (with $\sigma = 1$ and $\theta_{0,i} \sim \mathsf{N}(0, 1)$). The empirical results are averaged over 100 different realizations of noise and features. As δ grows, we observe a slower decay in the standard risk at small ε due to adversarial training. For $\delta = 10$, the standard risk has a small initial slope with respect to ε and then starts to increase rapidly. Put differently, with larger δ, the negative effect of adversarial training on the standard risk starts at smaller ε.

$\delta < 1$. As we observe for small to moderate values of ε, this curve is decreasing in ε, which implies that adversarial training helps with improving standard accuracy. The standard risk falls steeper as δ becomes closer to one. In Figure 3-(a) we observe a similar trend for $\delta > 1$. However, as δ grows larger than one, the positive effect of the adversarial training on the standard risk falters and we see a lower decline. When $\delta = 10$, the curve almost levels at $\varepsilon = 0$ and then starts to becomes increasing with ε. In other words, for larger δ we start to see that adversarial training has a negative effect on standard risk starting from smaller values of ε. Our theoretical prediction are in line with recent empirical observations of a similar flavor [TSE+18] observed in neural networks. Therefore, our theoretical results formally proves the emergence of such a behavior. We provide further insight into the emergence of this phenomena a long with some more rigorous theoretical guarantees in Appendix A.

3.4 Double-descent in adversarial training

When $\varepsilon = 0$, the estimator $\widehat{\boldsymbol{\theta}}^{\varepsilon}$ given by (2.6) reduces to the least-squares estimator. It is known that the plot of standard risk as a function of number of model complexity ($1/\delta = p/n$) exhibits a so-called 'double-descent' behavior [BMM18, BHMM18, HMRT19]. Namely, (1) up to the interpolation threshold $\delta = 1$ (beyond which the estimator achieves zero training error and the model interpolates the training data) the risk curve follows a U-shape; the risk first decreases as p increases because the model becomes less biased but then starts to increase because of the inflated variance of the



(a) Theoretical curves

(b) Empirical curves

Figure 4: Standard risk versus model complexity $1/\delta = p/n$. Left panel corresponds to the theoretical curve obtained by Theorem 3.3 (with $\sigma = 1$ and $V = 1$), and the right panel corresponds to the empirical results (with $\sigma = 1$ and $\theta_{0,i} \sim \mathsf{N}(0, 1)$). Here, we recover the double-descent behavior where the interpolation threshold shifts with ε.

estimator. (2) After the peak at the interpolation threshold, the risk decreases and essentially attains its global minimum at 'infinite' model complexity (extremely overparametrized regime).

The double-descent phenomenon is not limited to neural networks and have been empirically observed in a variety of models including random features and random forest models. Recently, analytical derivation of this phenomenon has been developed for least square regression and random features model [TSE+18, MM19]. For least square regression with Gaussian covariates, it is shown that the global minimum of the risk is achieved in the underparametrized setting $\delta > 1$ (unless miss-specified structures are assumed). Nonetheless, these work are focused on training with unperturbed features.

In Figure 4 (a), we plot the standard risk (theoretical predictions from Theorem 3.3) versus $1/\delta = p/n$, for several values of adversarial power ε. We also depict the empirical version of these curves in Figure Figure 4 (b). These plots demonstrate that the double-descent phenomena continues to hold even with adversarial training. Interestingly however the interpolation threshold changes with ε. For small ε, we observe double-descent behavior with the interpolation threshold $\delta \approx 1$. However, as ε increases the location of the peak shifts to higher values of $1/\delta$.

4 Further Related Work

The trade-off between standard and adversarial accuracy has been studied recently in [MMS+18, SST+18, TSE+18, RXY+19, ZYJ+19, PJ19]. An central question is whether standard and robust objectives are fundamentally at conflict? In other words, is there a predictor that can achieve both optimal standard accuracy and robust accuracy when the number of training data samples is sufficiently large? In this regard, [TSE+18, ZYJ+19] construct learning problems where the optimal robust accuracy is fundamentally at conflict with the standard accuracy, i.e. no predictor can achieve

both optimal standard accuracy and robust accuracy even in the infinite data limit. However, there are clearly many natural learning problems in which a predictor with optimal standard and high robust accuracy exists (hence the two objectives are not at conflict). An instance of such cases has been studied in [RXY+19] suggesting that the inconsistency between adversarial accuracy and standard accuracy may be due to insufficient number of training samples. In contrast, in this paper we have shown that a fundamental tradeoff exists between the two accuracies in linear regression even with limited samples.

Another line of work considers the tradeoff between standard and robust accuracy when the capacity of the learning model varies [Nak19, GCL+19]. In particular, [Nak19] provides classification problems where simple classifiers with high standard accuracy exist; but having high robust accuracy is possible through more complex classifiers. The notions of capacity and complexity in the presence of adversarially perturbed inputs (a.k.a. adversarially robust learnability) have also been studied in a series of interesting papers [BLPR19, CBM18, KL18, YRB19, MHS19]. In particular [MHS19] show that any hypothesis class with finite VC dimension is adversarially-robust PAC learnable in the ℓ_∞ metric using modified (improper) learning rules. Finally, let us point out that under specific high-dimensional data distributions (e.g. isotropic Gaussian), any classifier becomes highly vulnerable to adversarial ℓ_2 perturbations [GMF+18, MDM19, SHS+19]. Thus the adversarial error approaches 1 as the dimension grows. This phenomenon does not occur in our regression setting as the regression loss is smoothly varying as opposed to the classification error.

5 Sketch and roadmap of the proof

To be able to provide a precise characterization of the various tradeoffs we need to develop a precise understanding of the adversarial training objective

$$\min_{\boldsymbol{\theta}\in\mathbb{R}^p} \mathcal{L}(\boldsymbol{\theta}) := \min_{\boldsymbol{\theta}\in\mathbb{R}^p} \max_{\|\boldsymbol{\delta}_i\|_{\ell_2}\le\varepsilon} \frac{1}{2n}\sum_{i=1}^{n} \left(y_i - \langle \boldsymbol{x}_i + \boldsymbol{\delta}_i, \boldsymbol{\theta}\rangle\right)^2,$$

and its optimal solution $\widehat{\boldsymbol{\theta}}^{\varepsilon} \in \arg\min_{\boldsymbol{\theta}\in\mathbb{R}^p} \mathcal{L}(\boldsymbol{\theta})$. To achieve this we carry out the following steps.

Step I: Simplification of the loss (Section 6.2). The maximization objective is equal to the optimal value of a maximization problem and hence characterizing its properties directly is challenging. In the first step of our proof we show that one can in-fact solve this maximization problem and derive an expression for the loss in closed form. Specifically, we show

$$\mathcal{L}(\boldsymbol{\theta}) = \frac{1}{2n}\sum_{i=1}^{n}\left(|y_i - \langle \boldsymbol{x}_i, \boldsymbol{\theta}\rangle| + \varepsilon \|\boldsymbol{\theta}\|_{\ell_2}\right)^2 = \frac{1}{2n}\left\||\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\theta}| + \varepsilon \|\boldsymbol{\theta}\|_{\ell_2}\right\|_{\ell_2}^2. \tag{5.1}$$

The main intuition behind this derivation is that one can think of the min-max optimization problem above as a game between a learner and an adversary where the learner first chooses a parameter $\boldsymbol{\theta}$ and then the adversary changes each feature $\boldsymbol{x}_i$ given the label y_i and the learner's choice of $\boldsymbol{\theta}$. We show that the best choice for the adversary to maximize the error is to pick $\boldsymbol{\delta}_i$ in the direction of $\boldsymbol{\theta}$ with a magnitude of ε (maximum power of the adversary) and with the sign of the misfit on the i the training data point ($\mathrm{sgn}(\langle \boldsymbol{x}_i, \boldsymbol{\theta}\rangle - y_i)$). We formally prove this result by connecting it to the well-known trust region subproblem in optimization.

Step II: Reduction to an Auxiliary Optimization (AO) problem (Section 6.3). The loss (5.1), while significantly simplified, is still rather complicated and it is completely unclear how to precisely characterize its behavior and the quality of its optimal solution. In particular,

the dependence on the random data matrix $\boldsymbol{X}$ is still rather complex hindering statistical analysis even in an asymptotic setting. To bring the optimization problem into a form more amenable to precise asymptotic analysis we carry out a series of reformulations of the optimization problem. First, we rescale the loss. Next we consider a change of variable of the form $\boldsymbol{z} = \frac{1}{\sqrt{p}}(\boldsymbol{\theta} - \boldsymbol{\theta}_0)$ and add new variables by adding equality constraints. Finally, we use duality to cast the problem into a mini-max form. Combining these steps we arrive at the following equivalent Primal Optimization (PO) problem

$$\min_{\boldsymbol{z}\in\mathbb{R}^p, \boldsymbol{v}\in\mathbb{R}^n} \max_{\boldsymbol{u}\in\mathbb{R}^n} \quad \frac{1}{\sqrt{p}}\boldsymbol{u}^T\boldsymbol{X}\boldsymbol{z} - \frac{1}{\sqrt{p}}\boldsymbol{u}^T\boldsymbol{\omega} + \frac{1}{\sqrt{p}}\boldsymbol{u}^T\boldsymbol{v} + \ell(\boldsymbol{v};\boldsymbol{z}), \tag{5.2}$$

where $\boldsymbol{\omega} = \frac{\boldsymbol{w}}{\sqrt{p}} \in \mathbb{R}^n$ is a Gaussian vector with i.i.d. $\mathcal{N}(0,\sigma^2)$ entires and

$$\ell(\boldsymbol{v};\boldsymbol{z}) := \frac{1}{2p}\left(\|\boldsymbol{v}\|_{\ell_2}^2 + 2\frac{\varepsilon}{\sqrt{p}}\|\boldsymbol{v}\|_{\ell_1}\|\boldsymbol{\theta}_0 + \sqrt{p}\boldsymbol{z}\|_{\ell_2} + \frac{\varepsilon^2}{p}\|\boldsymbol{\theta}_0 + \sqrt{p}\boldsymbol{z}\|_{\ell_2}^2\right).$$

This equivalent form may be counter-intuitive as we started by simplifying a different mini-max optimization problem and we have now again introduced a new maximization! The main advantage of this new form is that it is in fact affine in the data matrix $\boldsymbol{X}$. This particular form allows us to use a powerful extension of a classical Gaussian process inequality due to [Gor88] known as Convex Gaussian Minimax Theorem (CGMT) [TOH15] which focuses on characterizing the asymptotic behavior of mini-max optimization problems that are affine in a Gaussian matrix $\boldsymbol{X}$. This result enables us to characterize the properties of (5.2) by studying the asymptotic behavior of the following, arguable simpler, *Auxiliary Optimization (AO)* problem instead

$$\min_{\boldsymbol{z}\in\mathbb{R}^p, \boldsymbol{v}} \max_{\boldsymbol{u}\in\mathbb{R}^n} \frac{1}{\sqrt{p}}\left(\|\boldsymbol{z}\|_{\ell_2}\boldsymbol{g}^T\boldsymbol{u} + \|\boldsymbol{u}\|_{\ell_2}\boldsymbol{h}^T\boldsymbol{z} - \boldsymbol{u}^T\boldsymbol{\omega} + \boldsymbol{u}^T\boldsymbol{v}\right) + \ell(\boldsymbol{v};\boldsymbol{z}). \tag{5.3}$$

We emphasize that the relationship between the above AO problem (5.3) and how it is exactly related to the PO problem (5.2) is much more intricate and technical. See Section 6.3 for details.

The CGMT framework has been recently used to derive precise characterization of the generalization error of the max-margin linear classifiers in overparametrized regime with separable data [DKT19, MRSY19]. Also [LS20] uses the CGMT framework to analyze max-ℓ_1-margin classifiers.

Step III: Scalarization of the Auxiliary Optimization (AO) problem (Section 6.4). In this step we further simplify the AO problem in (5.3). In particular we show the asymptotic behavior of the AO can be characterized rather precisely via the following scalar optimization problem involving five variables:

$$\max_{0\le\beta\le K_\beta} \sup_{\gamma,\tau_h\ge 0} \min_{0\le\alpha\le K_\alpha} \min_{\tau_g\ge 0} \quad D(\alpha,\beta,\gamma,\tau_h,\tau_g) \quad \text{where} \tag{5.4}$$

$$\begin{aligned} D(\alpha,\beta,\gamma,\tau_h,\tau_g) :=& \frac{\delta\beta}{2(\tau_g+\beta)}\left(\alpha^2+\sigma^2\right) \\ &+ \delta\mathbb{1}_{\left\{\gamma(\tau_g+\beta)>\sqrt{\frac{2}{\pi}}\delta\varepsilon\beta\sqrt{\alpha^2+\sigma^2}\right\}}\frac{\beta^2(\alpha^2+\sigma^2)}{2\tau_g(\tau_g+\beta)}\left(\operatorname{erf}\left(\frac{\tau_*}{\sqrt{2}}\right) - \frac{\gamma(\tau_g+\beta)}{\delta\varepsilon\beta\sqrt{\alpha^2+\sigma^2}}\tau_*\right) \\ &- \frac{\alpha}{2\tau_h}(\gamma^2+\beta^2) + \gamma\sqrt{\frac{\alpha^2\beta^2}{\tau_h^2}+V^2} - \frac{\alpha\tau_h}{2} + \frac{\beta\tau_g}{2} \end{aligned} \tag{5.5}$$

In particular a variety of conclusions can be derived based on the optimal solutions of the above optimization problem as we discuss in the next step. We note that while the expressions may look complicated we prove that this optimization problem is in fact convex in the minimization parameters (α, τ_g) and concave in the maximization parameters (β, γ, τ_h) so that its optimal solutions can be easily derived via a simple low-dimensional gradient descent rather quickly and accurately. We also note that this proof is quite intricate and involved, so it is not possible to give an intuitive sketch of the arguments here. We refer to Section 6.4 for details.

Step IV: Completing the proof of the theorems (Sections 6.7 and 6.8).
Finally, we utilize the above scalar form to derive all of the different theorems and results stated in Section 3. This is done by relating the quantities of interest in each theorem to the optimal solutions of (5.4). For instance, we show that $\lim_{n\to\infty} \frac{1}{p}\left\|\widehat{\boldsymbol{\theta}}^{\varepsilon} - \boldsymbol{\theta}_0\right\|_{\ell_2}^2 = \alpha_*^2$ with α_* the optimal solution over α. These calculations/proofs are carried out in detail in Sections 6.7 and 6.8. Since each argument is different we do not provide a summary here and refer to the corresponding sections.

6 Proofs

6.1 Notations

We define the data matrix $\boldsymbol{X} \in \mathbb{R}^{n\times p}$ with the rows consisting of the training data features $\boldsymbol{x}_1, \boldsymbol{x}_2, \dots, \boldsymbol{x}_n$. For a convex function $f : \mathbb{R}^m \to \mathbb{R}$, we denote its its Fenchel conjugate by $f^*(\boldsymbol{y}) = \sup_{\boldsymbol{x}} \boldsymbol{y}^T\boldsymbol{x} - f(\boldsymbol{x})$. We also define the Moreau envelope function of f at $\boldsymbol{x}$ with parameter τ as

$$e_f(\boldsymbol{x};\tau) \equiv \min_{\boldsymbol{v}} \frac{1}{2\tau}\left\|\boldsymbol{x}-\boldsymbol{v}\right\|_{\ell_2}^2 + f(\boldsymbol{v})\,.$$

6.2 Simplification of the loss

As discussed earlier in this section we wish to derive a closed form for the loss

$$\mathcal{L}(\boldsymbol{\theta}) \coloneqq \max_{\|\boldsymbol{\delta}_i\|_{\ell_2}\le\varepsilon} \frac{1}{2n}\sum_{i=1}^{n}\left(y_i - \langle \boldsymbol{x}_i + \boldsymbol{\delta}_i, \boldsymbol{\theta}\rangle\right)^2 \tag{6.1}$$

and in particular show that

$$\mathcal{L}(\boldsymbol{\theta}) = \frac{1}{2n}\sum_{i=1}^{n}\left(|y_i - \langle \boldsymbol{x}_i, \boldsymbol{\theta}\rangle| + \varepsilon\left\|\boldsymbol{\theta}\right\|_{\ell_2}\right)^2 = \frac{1}{2n}\left\||\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\theta}| + \varepsilon\left\|\boldsymbol{\theta}\right\|_{\ell_2}\right\|_{\ell_2}^2 \tag{6.2}$$

To this aim first note that the maximization in (6.1) decouples over i so that we can write

$$\mathcal{L}(\boldsymbol{\theta}) \coloneqq \frac{1}{2n}\sum_{i=1}^{n}\max_{\|\boldsymbol{\delta}_i\|_{\ell_2}\le\varepsilon}\left(y_i - \langle \boldsymbol{x}_i + \boldsymbol{\delta}_i, \boldsymbol{\theta}\rangle\right)^2$$

To continue further define $\widetilde{y}_i \coloneqq y_i - \langle \boldsymbol{x}_i, \boldsymbol{\theta}\rangle$. By expanding the square the optimization over $\boldsymbol{\delta}_i$ can be rewritten in the form

$$\min_{\|\boldsymbol{\delta}_i\|_{\ell_2}\le\varepsilon} \; -\frac{1}{2}\widetilde{y}_i^2 + \widetilde{y}_i\langle\boldsymbol{\theta},\boldsymbol{\delta}_i\rangle - \frac{1}{2}\langle\boldsymbol{\theta},\boldsymbol{\delta}_i\rangle^2\,.$$

Note that this is trust-region subproblem and $\boldsymbol{\delta}_i$ is a solution if and only if $\left\|\boldsymbol{\delta}_i\right\|_{\ell_2} \le \varepsilon$ and there exists $\lambda_i \ge 0$ such that

1. $(-\boldsymbol{\theta}\boldsymbol{\theta}^\mathsf{T} + \lambda_i \boldsymbol{I})\boldsymbol{\delta}_i = -\widetilde{y}_i \boldsymbol{\theta}$.
2. $-\boldsymbol{\theta}\boldsymbol{\theta}^\mathsf{T} + \lambda_i \boldsymbol{I} \succeq \boldsymbol{0}$ (or equivalently $\lambda_i \geq \|\boldsymbol{\theta}\|_{\ell_2}^2$)
3. $\lambda_i(\varepsilon - \|\boldsymbol{\delta}_i\|_{\ell_2}) = 0$.

Since by (2), $\lambda_i > 0$, condition (3) reduces to $\|\boldsymbol{\delta}_i\|_{\ell_2} = \varepsilon$. Also from (1), we have

$$\begin{aligned}
\boldsymbol{\delta}_i &= -\widetilde{y}_i(-\boldsymbol{\theta}\boldsymbol{\theta}^\mathsf{T} + \lambda_i \boldsymbol{I})^{-1}\boldsymbol{\theta} \\
&= -\lambda_i^{-1}\widetilde{y}_i \left(\boldsymbol{I} + \frac{\boldsymbol{\theta}\boldsymbol{\theta}^\mathsf{T}}{\lambda_i - \|\boldsymbol{\theta}\|_{\ell_2}^2} \right) \boldsymbol{\theta} \\
&= -\lambda_i^{-1}\widetilde{y}_i \boldsymbol{\theta} \frac{\lambda_i}{\lambda_i - \|\boldsymbol{\theta}\|_{\ell_2}^2} \\
&= -\widetilde{y}_i \boldsymbol{\theta} \frac{1}{\lambda_i - \|\boldsymbol{\theta}\|_{\ell_2}^2} \,. \qquad (6.3)
\end{aligned}$$

Using the fact that $\|\boldsymbol{\delta}_i\|_{\ell_2} = \varepsilon$ in the latter identity we thus conclude that $\lambda_i = (1/\varepsilon)\,\|\boldsymbol{\theta}\|_{\ell_2}\,|\widetilde{y}_i| + \|\boldsymbol{\theta}\|_{\ell_2}^2$. Substituting for λ_i in (6.3) we obtain

$$\boldsymbol{\delta}_i = -\frac{\widetilde{y}_i}{|\widetilde{y}_i|}\frac{\boldsymbol{\theta}\varepsilon}{\|\boldsymbol{\theta}\|_{\ell_2}} = -\varepsilon \mathrm{sgn}(y_i - \langle \boldsymbol{x}_i, \boldsymbol{\theta}\rangle)\frac{\boldsymbol{\theta}}{\|\boldsymbol{\theta}\|_{\ell_2}} \,.$$

Substituting the latter into (6.1) we arrive at (6.2) to complete our simplification of the loss.

6.3 Reduction to an auxiliary optimization problem via CGMT

We are interested in characterizing the properties of the optimal paramter $\widehat{\boldsymbol{\theta}}^\varepsilon$ and thus it shall be convenient to work with a scaled version of the loss (6.2). This scaling of course does not affect the optimal solution $\widehat{\boldsymbol{\theta}}^\varepsilon$. Thus hence forth we focus on the following objective

$$\widehat{\boldsymbol{\theta}}^\varepsilon = \arg\min_{\boldsymbol{\theta}\in\mathbb{R}^p} \frac{1}{2p^2}\sum_{i=1}^n \left(|y_i - \langle \boldsymbol{x}_i, \boldsymbol{\theta}\rangle| + \varepsilon \|\boldsymbol{\theta}\|_{\ell_2} \right)^2 . \qquad (6.4)$$

To continue further it is convenient to consider a change of variable of the form $\boldsymbol{z} = \frac{1}{\sqrt{p}}(\boldsymbol{\theta} - \boldsymbol{\theta}_0)$ and note that

$$y_i - \langle \boldsymbol{x}_i, \boldsymbol{\theta}\rangle = w_i + \langle \boldsymbol{x}_i, \boldsymbol{\theta}_0 - \boldsymbol{\theta}\rangle = w_i - \sqrt{p}\langle \boldsymbol{x}_i, \boldsymbol{z}\rangle.$$

Define $\ell(v;\boldsymbol{\theta}) := \frac{1}{2}\left(|v| + \varepsilon\|\boldsymbol{\theta}\|_{\ell_2}\right)^2$ and note that with this change of variable we have that $\widehat{\boldsymbol{z}}^\varepsilon = \frac{1}{\sqrt{p}}(\widehat{\boldsymbol{\theta}}^\varepsilon - \boldsymbol{\theta}_0)$ is given by

$$\widehat{\boldsymbol{z}}^\varepsilon = \arg\min_{\boldsymbol{z}} \frac{1}{p^2}\sum_{i=1}^n \ell\left(w_i - \sqrt{p}\langle \boldsymbol{x}_i, \boldsymbol{z}\rangle; \boldsymbol{\theta}_0 + \sqrt{p}\boldsymbol{z}\right).$$

Equivalently we can rewrite this optimization problem in the form

$$\min_{\boldsymbol{z}\in\mathbb{R}^p, \boldsymbol{v}\in\mathbb{R}^n} \frac{1}{p^2}\sum_{i=1}^n \ell\left(\sqrt{p}v_i; \boldsymbol{\theta}_0 + \sqrt{p}\boldsymbol{z}\right) \quad \text{subject to} \quad \sqrt{p}\boldsymbol{v} = \boldsymbol{w} - \sqrt{p}\boldsymbol{X}\boldsymbol{z}. \qquad (6.5)$$

We note that the scaling of $\boldsymbol{v}$ is arbitrary but serves the purpose of simplifying the exposition later on. The loss above is still rather complicated and it is unclear how to study and characterize the properties of its optimal solution in an asymptotic regime where the size of the training data and the number of parameters grow in proportion with each other. To study this loss in an asymptotic fashion we first cast it as a different mini-max optimization using duality. In particular by associating a dual variable $\frac{\boldsymbol{u}}{p}$ with the equality constraint, we obtain

$$\begin{aligned}\min_{\boldsymbol{z}\in\mathbb{R}^p,\boldsymbol{v}\in\mathbb{R}^n}\max_{\boldsymbol{u}\in\mathbb{R}^n}\frac{1}{p}\Big\{\boldsymbol{u}^T(\sqrt{p}\boldsymbol{X})\boldsymbol{z}-\boldsymbol{u}^T\boldsymbol{w}+\sqrt{p}\boldsymbol{u}^T\boldsymbol{v}\Big\}+\frac{1}{p^2}\sum_{i=1}^n\ell\left(\sqrt{p}v_i;\boldsymbol{\theta}_0+\sqrt{p}\boldsymbol{z}\right)\\ =\min_{\boldsymbol{z}\in\mathbb{R}^p,\boldsymbol{v}\in\mathbb{R}^n}\max_{\boldsymbol{u}\in\mathbb{R}^n}\frac{1}{p}\Big\{\boldsymbol{u}^T(\sqrt{p}\boldsymbol{X})\boldsymbol{z}-\boldsymbol{u}^T\boldsymbol{w}+\sqrt{p}\boldsymbol{u}^T\boldsymbol{v}\Big\}+\ell(\boldsymbol{v};\boldsymbol{z})\end{aligned} \tag{6.6}$$

where

$$\ell(\boldsymbol{v};\boldsymbol{z}):=\frac{1}{p^2}\sum_{i=1}^n\ell\left(\sqrt{p}v_i;\boldsymbol{\theta}_0+\sqrt{p}\boldsymbol{z}\right)=\frac{1}{2p}\left(\|\boldsymbol{v}\|_{\ell_2}^2+2\frac{\varepsilon}{\sqrt{p}}\|\boldsymbol{v}\|_{\ell_1}\|\boldsymbol{\theta}_0+\sqrt{p}\boldsymbol{z}\|_{\ell_2}+\frac{\varepsilon^2}{p}\|\boldsymbol{\theta}_0+\sqrt{p}\boldsymbol{z}\|_{\ell_2}^2\right)$$

At first this may be counter-intuitive as we started by simplifying a different mini-max optimization problem and now we are again introducing a new maximization! The main advantage of this new form is that (6.6) is in fact affine in the matrix. This particular form allows us to use a powerful extension of a classical Gaussian process inequality due to Gordon [Gor88] known as Convex Gaussian Minimax Theorem (CGMT) [TOH15] which focuses on characterizing the asymptotic behavior of mini-max optimization problems that are affine in a Gaussian matrix $\boldsymbol{X}$. Formally, the CGMT framework shows that a problem of the form

$$\min_{\boldsymbol{z}\in\mathcal{S}_{\boldsymbol{z}}}\max_{\boldsymbol{u}\in\mathcal{S}_{\boldsymbol{u}}}\quad\boldsymbol{u}^T\boldsymbol{X}\boldsymbol{z}+\psi(\boldsymbol{z},\boldsymbol{u}) \tag{6.7}$$

with $\boldsymbol{X}$ a matrix with $\mathcal{N}(0,1)$ entries can be replaced asymptotically with

$$\min_{\boldsymbol{z}\in\mathcal{S}_{\boldsymbol{z}}}\max_{\boldsymbol{u}\in\mathcal{S}_{\boldsymbol{u}}}\quad\|\boldsymbol{z}\|_{\ell_2}\boldsymbol{g}^T\boldsymbol{u}+\|\boldsymbol{u}\|_{\ell_2}\boldsymbol{h}^T\boldsymbol{z}+\psi(\boldsymbol{z},\boldsymbol{u}) \tag{6.8}$$

where $\boldsymbol{g}$ and $\boldsymbol{h}$ are independent Gaussian vectors with i.i.d. $\mathcal{N}(0,1)$ entries and $\psi(\boldsymbol{z},\boldsymbol{u})$ is convex in $\boldsymbol{z}$ and concave in $\boldsymbol{u}$. In the above $\mathcal{S}_{\boldsymbol{z}}$ and $\mathcal{S}_{\boldsymbol{u}}$ are compact sets. We refer to [TOH15, Theorem 3] for precise statements. Following [TOH15] we shall refer to problems of the form (6.7) and (6.8) as the Primal Problem (PO) and the Auxiliary Problem (AO).

As evident from the above to be able to apply CGMT, requires the minimization/maximization to be over compact sets. To avoid this technical issue one can introduce "artificial" boundedness constraint so that they do not change the optimal solution. More specifically, we can add constraints of the form $\mathcal{S}_{\boldsymbol{z}}=\{\boldsymbol{z}|\ \ \|\boldsymbol{z}\|_{\ell_2}\le K_\alpha\}$ and $\mathcal{S}_{\boldsymbol{u}}=\{\boldsymbol{u}:\|\boldsymbol{u}\|_{\ell_2}\le K_\beta\}$ for sufficiently large constants K_α and K_β without changing the optimal solution of (6.6) in a precise asymptotic sense. See Appendix B for precise statements and proofs. This allows us to replace (6.6) with

$$\min_{\boldsymbol{z}\in\mathcal{S}_{\boldsymbol{z}},\boldsymbol{v}\in\mathbb{R}^n}\max_{\boldsymbol{u}\in\mathcal{S}_{\boldsymbol{u}}}\quad\frac{1}{\sqrt{p}}\boldsymbol{u}^T\boldsymbol{X}\boldsymbol{z}-\frac{1}{\sqrt{p}}\boldsymbol{u}^T\boldsymbol{\omega}+\frac{1}{\sqrt{p}}\boldsymbol{u}^T\boldsymbol{v}+\ell(\boldsymbol{v};\boldsymbol{z}), \tag{6.9}$$

where $\boldsymbol{\omega}=\frac{\boldsymbol{w}}{\sqrt{p}}\in\mathbb{R}^n$ is a Gaussian vector with i.i.d. $\mathcal{N}(0,\sigma^2)$ entires.

With these compact constraints in place we can now apply the CGMT result. To this aim note that this optimization is in the desired form of a Primary Optimization (PO): it has a bilinear term $\boldsymbol{u}^T\boldsymbol{X}\boldsymbol{z}$ plus a function

$$\psi(\boldsymbol{z},\boldsymbol{u}) = \min_{\boldsymbol{v}\in\mathbb{R}^n} \frac{1}{\sqrt{p}}\left(-\boldsymbol{u}^T\boldsymbol{\omega} + \boldsymbol{u}^T\boldsymbol{v}\right) + \ell(\boldsymbol{v};\boldsymbol{z})$$

which is convex in $\boldsymbol{z}$[4] and concave in $\boldsymbol{u}$. The corresponding Auxiliary Optimization (AO) thus takes the form

$$\begin{aligned}\min_{\boldsymbol{z}\in\mathcal{S}_{\boldsymbol{z}}}\max_{\boldsymbol{u}\in\mathcal{S}_{\boldsymbol{u}}} \frac{1}{\sqrt{p}}\left(\|\boldsymbol{z}\|_{\ell_2}\boldsymbol{g}^T\boldsymbol{u} + \|\boldsymbol{u}\|_{\ell_2}\boldsymbol{h}^T\boldsymbol{z}\right) + \min_{\boldsymbol{v}\in\mathbb{R}^n} \frac{1}{\sqrt{p}}\left(-\boldsymbol{u}^T\boldsymbol{\omega} + \boldsymbol{u}^T\boldsymbol{v}\right) + \ell(\boldsymbol{v};\boldsymbol{z})\\ = \min_{\boldsymbol{z}\in\mathcal{S}_{\boldsymbol{z}},\boldsymbol{v}}\max_{\boldsymbol{u}\in\mathcal{S}_{\boldsymbol{u}}} \frac{1}{\sqrt{p}}\left(\|\boldsymbol{z}\|_{\ell_2}\boldsymbol{g}^T\boldsymbol{u} + \|\boldsymbol{u}\|_{\ell_2}\boldsymbol{h}^T\boldsymbol{z} - \boldsymbol{u}^T\boldsymbol{\omega} + \boldsymbol{u}^T\boldsymbol{v}\right) + \ell(\boldsymbol{v};\boldsymbol{z}).\end{aligned} \tag{6.10}$$

This completes the derivation of the AO.

6.4 Scalarization of the auxilary optimization problem

In this section we continue our proof by significantly simplifying the AO problem. In particular we show that the behavior of the AO and hence the PO can be completely characterized by (6.23). This is arguably the most intricate part of our proofs.

We begin simplifying the AO by maximizing over $\boldsymbol{u}$. To this aim we decompose the optimization problem over $\mathcal{S}_{\boldsymbol{u}}$ in terms of its direction and radius. Specifically, $\boldsymbol{u} = \beta\widetilde{\boldsymbol{u}}$ with $\widetilde{\boldsymbol{u}} \in \mathbb{S}^{n-1}$ and $0 \le \beta \le K_\beta$. Using this decomposition we have

$$\begin{aligned}\max_{\boldsymbol{u}\in\mathcal{S}_{\boldsymbol{u}}} \frac{1}{\sqrt{p}}\left(\|\boldsymbol{z}\|_{\ell_2}\boldsymbol{g}^T\boldsymbol{u} + \|\boldsymbol{u}\|_{\ell_2}\boldsymbol{h}^T\boldsymbol{z} - \boldsymbol{u}^T\boldsymbol{\omega} + \boldsymbol{u}^T\boldsymbol{v}\right)\\ = \max_{0\le\beta\le K_\beta}\ \max_{\boldsymbol{u}\in\mathbb{S}^{n-1}} \frac{1}{\sqrt{p}}\left(\|\boldsymbol{z}\|_{\ell_2}\boldsymbol{g}^T\boldsymbol{u} + \|\boldsymbol{u}\|_{\ell_2}\boldsymbol{h}^T\boldsymbol{z} - \boldsymbol{u}^T\boldsymbol{\omega} + \boldsymbol{u}^T\boldsymbol{v}\right)\\ = \max_{0\le\beta\le K_\beta}\ \max_{\boldsymbol{u}\in\mathbb{S}^{n-1}} \frac{1}{\sqrt{p}}\boldsymbol{u}^T\left(\|\boldsymbol{z}\|_{\ell_2}\boldsymbol{g} - \boldsymbol{\omega} + \boldsymbol{v}\right) + \frac{\beta}{\sqrt{p}}\boldsymbol{h}^T\boldsymbol{z}\\ = \max_{0\le\beta\le K_\beta} \frac{1}{\sqrt{p}}\left\|\|\boldsymbol{z}\|_{\ell_2}\boldsymbol{g} - \boldsymbol{\omega} + \boldsymbol{v}\right\|_{\ell_2} + \frac{\beta}{\sqrt{p}}\boldsymbol{h}^T\boldsymbol{z}.\end{aligned}$$

Plugging the latter into (6.10) the AO reduces to

$$\min_{\boldsymbol{z}\in\mathcal{S}_{\boldsymbol{z}},\boldsymbol{v}}\max_{0\le\beta\le K_\beta} \frac{1}{\sqrt{p}}\left\|\|\boldsymbol{z}\|_{\ell_2}\boldsymbol{g} - \boldsymbol{\omega} + \boldsymbol{v}\right\|_{\ell_2} + \frac{\beta}{\sqrt{p}}\boldsymbol{h}^T\boldsymbol{z} + \ell(\boldsymbol{v};\boldsymbol{z}).$$

We hope to eventually simplify the minimization over $\boldsymbol{v}$ and $\boldsymbol{z}$ also. For this minimization to become easier in our later calculation we proceed by writing $\ell(\boldsymbol{v};\boldsymbol{z})$ in terms of its conjugate with respect to $\boldsymbol{z}$. That is,

$$\ell(\boldsymbol{v};\boldsymbol{z}) = \sup_{\boldsymbol{q}} \boldsymbol{q}^T\boldsymbol{z} - \widetilde{\ell}(\boldsymbol{v};\boldsymbol{q})$$

[4]Note that the prior to the minimization over $\boldsymbol{v}$ the problem is trivially jointly convex in $(\boldsymbol{z},\boldsymbol{v})$ and partial minimization preserves convexity.

where $\widetilde{\ell}(\boldsymbol{v};\boldsymbol{q})$ is the conjugate of ℓ with respect to $\boldsymbol{z}$. The logic behind this is that AO with then simplify to

$$\min_{\boldsymbol{z}\in\mathcal{S}_{\boldsymbol{z}},\boldsymbol{v}} \max_{0\le\beta\le K_\beta,\boldsymbol{q}} \frac{\beta}{\sqrt{p}} \left\| \|\boldsymbol{z}\|_{\ell_2} \boldsymbol{g} - \boldsymbol{\omega} + \boldsymbol{v} \right\|_{\ell_2} + \frac{\beta}{\sqrt{p}} \boldsymbol{h}^T \boldsymbol{z} + \boldsymbol{q}^T \boldsymbol{z} - \widetilde{\ell}(\boldsymbol{v};\boldsymbol{q}). \tag{6.11}$$

To proceed it would be convenient to flip the order of minimum and maximum in the above. However, for this to be allowed the mini-max problem typically has to be convex/concave in the min/max parameters (e.g. via the celebrated Sion's min-max Theorem [S+58]). It is not clear that the above objective has this form so that the flipping of the order of the min and max is justified. However, since the original PO problem is convex/concave in the min/max parameters one can justify such a flipping of the min and max in the AO based on the PO. We note that this is justified for asymptotic calculations and refer to [TAH15, Appendix A.2.4] for precise details on this derivation. Thus, we will instead consider the following problem as the (AO) which is asymptotically equivalent to (6.11)

$$\max_{0\le\beta\le K_\beta,\boldsymbol{q}} \min_{\boldsymbol{z}\in\mathcal{S}_{\boldsymbol{z}},\boldsymbol{v}} \frac{\beta}{\sqrt{p}} \left\| \|\boldsymbol{z}\|_{\ell_2} \boldsymbol{g} - \boldsymbol{\omega} + \boldsymbol{v} \right\|_{\ell_2} + \frac{\beta}{\sqrt{p}} \boldsymbol{h}^T \boldsymbol{z} + \boldsymbol{q}^T \boldsymbol{z} - \widetilde{\ell}(\boldsymbol{v};\boldsymbol{q}).$$

To simplify further we now optimize over the direction and norm of $\boldsymbol{z}$ ($\|\boldsymbol{z}\|_{\ell_2} = \alpha$) to arrive at

$$\max_{0\le\beta\le K_\beta,\boldsymbol{q}} \min_{0\le\alpha\le K_\alpha,\boldsymbol{v}} \frac{\beta}{\sqrt{p}} \|\alpha\boldsymbol{g} - \boldsymbol{\omega} + \boldsymbol{v}\|_{\ell_2} - \alpha \left\| \frac{\beta}{\sqrt{p}} \boldsymbol{h} + \boldsymbol{q} \right\|_{\ell_2} - \widetilde{\ell}(\boldsymbol{v};\boldsymbol{q}). \tag{6.12}$$

Next note that $-\widetilde{\ell}(\boldsymbol{v};\boldsymbol{q})$ is convex in $\boldsymbol{v}$. To see this first note that

$$\widetilde{\ell}(\boldsymbol{v};\boldsymbol{q}) = \sup_{\boldsymbol{z}} \boldsymbol{q}^T \boldsymbol{z} - \ell(\boldsymbol{v};\boldsymbol{z}).$$

Also since ℓ is jointly convex in $(\boldsymbol{v},\boldsymbol{z})$, then $-\ell(\boldsymbol{v};\boldsymbol{z})$ is jointly concave in $(\boldsymbol{v},\boldsymbol{z})$. Also $\boldsymbol{q}^T\boldsymbol{z}$ is jointly concave in $(\boldsymbol{v},\boldsymbol{z})$. Therefore, $\boldsymbol{q}^T\boldsymbol{z} - \ell(\boldsymbol{v};\boldsymbol{z})$ is jointly concave in $(\boldsymbol{v},\boldsymbol{z})$ and based on the partial maximization rule we can conclude that $\widetilde{\ell}(\boldsymbol{v};\boldsymbol{q})$ should be concave in $\boldsymbol{v}$ which in turn implies $-\widetilde{\ell}(\boldsymbol{v};\boldsymbol{q})$ is convex in $\boldsymbol{v}$. The other terms are also trivially jointly convex in $\alpha,\boldsymbol{v}$ so that overall the objective is jointly convex in $\alpha,\boldsymbol{v}$. The objective above is also trivially jointly concave in $\beta,\boldsymbol{q}$. Thus based on Sion's min-max Theorem [S+58]) we could change the order of the mins and maxs as we please. This allows us to reorder $\max_{\boldsymbol{q}}$ and $\min_{\alpha,\boldsymbol{v}}$ to arrive at

$$\max_{0\le\beta\le K_\beta} \min_{0\le\alpha\le K_\alpha,\boldsymbol{v}} \max_{\boldsymbol{q}} \frac{\beta}{\sqrt{p}} \|\alpha\boldsymbol{g} - \boldsymbol{\omega} + \boldsymbol{v}\|_{\ell_2} - \alpha \left\| \frac{\beta}{\sqrt{p}} \boldsymbol{h} + \boldsymbol{q} \right\|_{\ell_2} - \widetilde{\ell}(\boldsymbol{v};\boldsymbol{q}) \tag{6.13}$$

To proceed, we first compute $\widetilde{\ell}(\boldsymbol{v};\boldsymbol{q})$ in the Lemma below with the proof deferred to Appendix C.1.

Lemma 6.1. *The conjugate of*

$$\ell(\boldsymbol{v};\boldsymbol{z}) := \frac{1}{2p}\left(\|\boldsymbol{v}\|_{\ell_2}^2 + 2\frac{\varepsilon}{\sqrt{p}} \|\boldsymbol{v}\|_{\ell_1} \|\boldsymbol{\theta}_0 + \sqrt{p}\boldsymbol{z}\|_{\ell_2} + \frac{\varepsilon^2}{p} \|\boldsymbol{\theta}_0 + \sqrt{p}\boldsymbol{z}\|_{\ell_2}^2 \right)$$

with respect to the variable $\boldsymbol{z}$ is given by

$$\widetilde{\ell}(\boldsymbol{v};\boldsymbol{q}) := \sup_{\boldsymbol{z}} \boldsymbol{q}^T\boldsymbol{z} - \ell(\boldsymbol{v};\boldsymbol{z}) = -\frac{1}{\sqrt{p}}\boldsymbol{q}^T\boldsymbol{\theta}_0 + \frac{1}{2\delta p^2}\left(\frac{p}{\varepsilon}\|\boldsymbol{q}\|_{\ell_2} - \|\boldsymbol{v}\|_{\ell_1} \right)_+^2 - \frac{1}{2p}\|\boldsymbol{v}\|_{\ell_2}^2 .$$

Using this characterization of $\widetilde{\ell}(\boldsymbol{v};\boldsymbol{q})$ we arrive at the following representation of AO problem

$$\min_{0\le\alpha\le K_\alpha,\boldsymbol{v}}\max_{0\le\beta\le K_\beta}\max_{\boldsymbol{q}}\ \frac{\beta}{\sqrt{p}}\left\|\alpha\boldsymbol{g}-\boldsymbol{\omega}+\boldsymbol{v}\right\|_{\ell_2}-\alpha\left\|\frac{\beta}{\sqrt{p}}\boldsymbol{h}+\boldsymbol{q}\right\|_{\ell_2}+\frac{1}{\sqrt{p}}\boldsymbol{q}^T\boldsymbol{\theta}_0 \\ -\frac{1}{2\delta p^2}\left(\frac{p}{\varepsilon}\|\boldsymbol{q}\|_{\ell_2}-\|\boldsymbol{v}\|_{\ell_1}\right)_+^2+\frac{1}{2p}\|\boldsymbol{v}\|_{\ell_2}^2 \tag{6.14}$$

To simplify further we next focus on the maximization over $\boldsymbol{q}$ or equivalently the following minimization problem

$$\min_{\boldsymbol{q}}\quad \alpha\left\|\frac{\beta}{\sqrt{p}}\boldsymbol{h}+\boldsymbol{q}\right\|_{\ell_2}+\frac{1}{2\delta p^2}\left(\frac{p}{\varepsilon}\|\boldsymbol{q}\|_{\ell_2}-\|\boldsymbol{v}\|_{\ell_1}\right)_+^2-\frac{1}{\sqrt{p}}\boldsymbol{q}^T\boldsymbol{\theta}_0$$

$$\min_{\boldsymbol{q}}\ \inf_{\tau_h\ge 0}\ \frac{\alpha}{2\tau_h}\left\|\frac{\beta}{\sqrt{p}}\boldsymbol{h}+\boldsymbol{q}\right\|_{\ell_2}^2+\frac{\alpha\tau_h}{2}+\frac{1}{2\delta p^2}\left(\frac{p}{\varepsilon}\|\boldsymbol{q}\|_{\ell_2}-\|\boldsymbol{v}\|_{\ell_1}\right)_+^2-\frac{1}{\sqrt{p}}\boldsymbol{q}^T\boldsymbol{\theta}_0$$

$$\min_{\boldsymbol{q}}\ \inf_{\tau_h\ge 0}\ \frac{\alpha}{2\tau_h}\|\boldsymbol{q}\|_{\ell_2}^2+\frac{\alpha\beta^2}{2p\tau_h}\|\boldsymbol{h}\|_{\ell_2}^2+\frac{\alpha\beta}{\tau_h\sqrt{p}}\boldsymbol{h}^T\boldsymbol{q}+\frac{\alpha\tau_h}{2}+\frac{1}{2\delta p^2}\left(\frac{p}{\varepsilon}\|\boldsymbol{q}\|_{\ell_2}-\|\boldsymbol{v}\|_{\ell_1}\right)_+^2-\frac{1}{\sqrt{p}}\boldsymbol{q}^T\boldsymbol{\theta}_0$$

The above is a linear function of $\boldsymbol{q}$ plus a term depending on $\|\boldsymbol{q}\|_{\ell_2}$. So fixing $\|\boldsymbol{q}\|_{\ell_2}=\gamma\ge 0$ the optimal $\boldsymbol{q}$ is given by $\boldsymbol{q}=-\gamma\frac{\frac{\alpha\beta}{\tau_h}\boldsymbol{h}-\boldsymbol{\theta}_0}{\left\|\frac{\alpha\beta}{\tau_h}\boldsymbol{h}-\boldsymbol{\theta}_0\right\|_{\ell_2}}$, which simplifies the above to

$$\inf_{\tau_h,\gamma\ge 0}\quad \frac{\alpha}{2\tau_h}\gamma^2+\frac{\alpha\beta^2}{2p\tau_h}\|\boldsymbol{h}\|_{\ell_2}^2-\frac{\gamma}{\sqrt{p}}\left\|\frac{\alpha\beta}{\tau_h}\boldsymbol{h}-\boldsymbol{\theta}_0\right\|_{\ell_2}+\frac{\alpha\tau_h}{2}+\frac{1}{2\delta p^2}\left(\frac{p}{\varepsilon}\gamma-\|\boldsymbol{v}\|_{\ell_1}\right)_+^2$$

Plugging the latter into (6.14) the AO reduces to

$$\min_{0\le\alpha\le K_\alpha,\boldsymbol{v}}\max_{0\le\beta\le K_\beta}\sup_{\gamma,\tau_h\ge 0}\ \frac{\beta}{\sqrt{p}}\left\|\alpha\boldsymbol{g}-\boldsymbol{\omega}+\boldsymbol{v}\right\|_{\ell_2}+\frac{1}{2p}\|\boldsymbol{v}\|_{\ell_2}^2 \\ -\frac{\alpha}{2\tau_h}\gamma^2-\frac{\alpha\beta^2}{2p\tau_h}\|\boldsymbol{h}\|_{\ell_2}^2+\frac{\gamma}{\sqrt{p}}\left\|\frac{\alpha\beta}{\tau_h}\boldsymbol{h}-\boldsymbol{\theta}_0\right\|_{\ell_2}-\frac{\alpha\tau_h}{2}-\frac{1}{2\delta p^2}\left(\frac{p}{\varepsilon}\gamma-\|\boldsymbol{v}\|_{\ell_1}\right)_+^2 \tag{6.15}$$

To continue we state a lemma with the proof deferred to Appendix C.2

Lemma 6.2. *The function*

$$f(\gamma,\beta,\tau_h):=\gamma^2+\frac{\beta^2}{p}\|\boldsymbol{h}\|_{\ell_2}^2-2\frac{\gamma}{\sqrt{p}}\left\|\beta\boldsymbol{h}-\frac{\boldsymbol{\theta}_0}{\alpha}\right\|_{\ell_2}$$

is jointly convex in the parameters (γ,β,τ_h).

Using this lemma we can trivially conclude that the objective (6.15) is jointly concave in (γ,β,τ_h). Also note that $\widetilde{\ell}$ is concave in $\boldsymbol{v}$ and hence $-\widetilde{\ell}$ is convex in $\boldsymbol{v}$. This implies that the objective in (6.14) is jointly convex in $(\alpha,\boldsymbol{v})$. Since maximization (with respect to the direction of $\boldsymbol{q}$) preserves convexity therefore (6.15) is trivially jointly convex in $(\alpha,\boldsymbol{v})$. Therefore, we can flip the order of

min and max in (6.15) (again using Sion's min-max Theorem) to arrive at

$$\max_{0\le\beta\le K_\beta}\sup_{\gamma,\tau_h\ge 0}\min_{0\le\alpha\le K_\alpha}\min_{\boldsymbol{v}}\ \frac{\beta}{\sqrt{p}}\left\|\alpha\boldsymbol{g}-\boldsymbol{\omega}+\boldsymbol{v}\right\|_{\ell_2}+\frac{1}{2p}\left\|\boldsymbol{v}\right\|_{\ell_2}^2 \\ -\frac{\alpha}{2\tau_h}\gamma^2-\frac{\alpha\beta^2}{2p\tau_h}\left\|\boldsymbol{h}\right\|_{\ell_2}^2+\frac{\gamma}{\sqrt{p}}\left\|\frac{\alpha\beta}{\tau_h}\boldsymbol{h}-\boldsymbol{\theta_0}\right\|_{\ell_2}-\frac{\alpha\tau_h}{2}-\frac{1}{2\delta p^2}\left(\frac{p}{\varepsilon}\gamma-\left\|\boldsymbol{v}\right\|_{\ell_1}\right)_+^2 \tag{6.16}$$

We now focus on minimization over $\boldsymbol{v}$. To this aim note that

$$\begin{aligned}
&\min_{\boldsymbol{v}}\ \frac{\beta}{\sqrt{p}}\left\|\alpha\boldsymbol{g}-\boldsymbol{\omega}+\boldsymbol{v}\right\|_{\ell_2}+\frac{1}{2p}\left\|\boldsymbol{v}\right\|_{\ell_2}^2-\frac{1}{2\delta p^2}\left(\frac{p}{\varepsilon}\gamma-\left\|\boldsymbol{v}\right\|_{\ell_1}\right)_+^2 \\
&\min_{\tau_g\ge 0,\boldsymbol{v}}\ \frac{\beta}{2\tau_g p}\left\|\alpha\boldsymbol{g}-\boldsymbol{\omega}+\boldsymbol{v}\right\|_{\ell_2}^2+\frac{\beta\tau_g}{2}+\frac{1}{2p}\left\|\boldsymbol{v}\right\|_{\ell_2}^2-\frac{1}{2\delta p^2}\left(\frac{p}{\varepsilon}\gamma-\left\|\boldsymbol{v}\right\|_{\ell_1}\right)_+^2 \\
&\min_{\tau_g\ge 0,\boldsymbol{v}}\ \frac{\beta}{2\tau_g p}\left\|\alpha\boldsymbol{g}-\boldsymbol{\omega}+\boldsymbol{v}\right\|_{\ell_2}^2+\frac{\beta\tau_g}{2}+\frac{1}{2p}\left\|\boldsymbol{v}\right\|_{\ell_2}^2-\frac{1}{2\delta p^2}\left(\frac{p}{\varepsilon}\gamma-\left\|\boldsymbol{v}\right\|_{\ell_1}\right)_+^2
\end{aligned} \tag{6.17}$$

Recall the definition of the Moreau envelope function of a function f at a point $\boldsymbol{x}$ with parameter μ,

$$e_f(\boldsymbol{x};\mu)\equiv\min_{\boldsymbol{v}}\frac{1}{2\mu}\left\|\boldsymbol{x}-\boldsymbol{v}\right\|_{\ell_2}^2+f(\boldsymbol{v})\,.$$

and define

$$f(\boldsymbol{v};\gamma)\equiv\frac{1}{2}\left\|\boldsymbol{v}\right\|_{\ell_2}^2-\frac{1}{2\delta p}(\frac{p}{\varepsilon}\gamma-\left\|\boldsymbol{v}\right\|_{\ell_1})_+^2\,, \tag{6.18}$$

Note that $f(\boldsymbol{v};\gamma)$ is convex in $\boldsymbol{v}$ (since $-\widetilde{\ell}(\boldsymbol{v};\boldsymbol{q})$ was convex in $\boldsymbol{v}$). Thus, (6.17) can be rewritten in the more compact form

$$\min_{\tau_g\ge 0}\ \frac{1}{p}e_f\left(\boldsymbol{\omega}-\alpha\boldsymbol{g};\frac{\tau_g}{\beta}\right)+\frac{\beta\tau_g}{2} \tag{6.19}$$

In our next lemma we compute e_f. We defer the proof to Appendix C.3.

Lemma 6.3. *Consider the function f given by (6.18). Then,*

$$e_f(\boldsymbol{x};\mu)=\frac{1}{2(\mu+1)}\left\|\boldsymbol{x}\right\|_{\ell_2}^2+\min_{\tau\ge 0}G(\boldsymbol{x};\mu,\tau)$$

where

$$G(\boldsymbol{x};\mu,\tau)=\frac{1}{2\mu(\mu+1)}\left\|\boldsymbol{x}-\mathsf{ST}(\boldsymbol{x};\tau)\right\|_{\ell_2}^2-\frac{1}{2n}\left(\frac{p}{\varepsilon}\gamma-\frac{1}{1+\mu}\left\|\mathsf{ST}(\boldsymbol{x};\tau)\right\|_{\ell_1}\right)_+^2.$$

Furthermore, $e_f(\boldsymbol{x};\tau)$ is strictly convex in $\boldsymbol{x}$.

Plugging Lemma 6.3 into (A.13) we have

$$\frac{1}{p}e_f\left(\alpha\boldsymbol{g}-\boldsymbol{\omega};\frac{\tau_g}{\beta}\right)+\frac{\beta\tau_g}{2}=\frac{\beta}{2(\tau_g+\beta)}\frac{1}{p}\left\|\alpha\boldsymbol{g}-\boldsymbol{\omega}\right\|_{\ell_2}^2+\min_{\tau\ge 0}\frac{1}{p}G\left(\alpha\boldsymbol{g}-\boldsymbol{\omega};\frac{\tau_g}{\beta},\tau\right)+\frac{\beta\tau_g}{2}$$

Plugging this in (6.16) the AO problem reduces to

$$\max_{0\le\beta\le K_\beta}\sup_{\gamma,\tau_h\ge 0}\min_{0\le\alpha\le K_\alpha}\min_{\tau_g\ge 0}\min_{\tau\ge 0}\quad \frac{\beta}{2(\tau_g+\beta)}\frac{1}{p}\left\|\alpha\boldsymbol{g}-\boldsymbol{\omega}\right\|_{\ell_2}^2+\frac{1}{p}G\left(\alpha\boldsymbol{g}-\boldsymbol{\omega};\frac{\tau_g}{\beta},\tau\right)+\frac{\beta\tau_g}{2}$$
$$-\frac{\alpha}{2\tau_h}\gamma^2-\frac{\alpha\beta^2}{2p\tau_h}\left\|\boldsymbol{h}\right\|_{\ell_2}^2+\frac{\gamma}{\sqrt{p}}\left\|\frac{\alpha\beta}{\tau_h}\boldsymbol{h}-\boldsymbol{\theta_0}\right\|_{\ell_2}-\frac{\alpha\tau_h}{2}\tag{6.20}$$

We note that since the problem (6.17) was jointly convex in $(\boldsymbol{v},\alpha,\tau_g)$ and (6.16) jointly concave in (β,γ,τ_h) and partial minimization preserves convexity we thus conclude that the objective is jointly convex in (α,τ_g) and jointly concave in (β,γ,τ_h) (after the minimization over $\tau\ge 0$ has been carried out). Note that trivially in an asymptotic regime

$$\frac{\left\|\boldsymbol{h}\right\|_{\ell_2}^2}{p}\to 1\quad\text{and}\quad\frac{\left\|\alpha\boldsymbol{g}-\boldsymbol{\omega}\right\|_{\ell_2}^2}{n}\to\left(\alpha^2+\sigma^2\right)$$

Also using concentration of Lipschitz functions of Gaussian we have

$$\frac{1}{\sqrt{p}}\left\|\frac{\alpha\beta}{\tau_h}\boldsymbol{h}-\boldsymbol{\theta_0}\right\|_{\ell_2}\to\frac{1}{\sqrt{p}}\sqrt{\mathbb{E}\left[\left\|\frac{\alpha\beta}{\tau_h}\boldsymbol{h}-\boldsymbol{\theta_0}\right\|_{\ell_2}^2\right]}\to\sqrt{\frac{\alpha^2\beta^2}{\tau_h^2}+V^2}.$$

Plugging all of the above in (6.20) we arrive at

$$\max_{0\le\beta\le K_\beta}\sup_{\gamma,\tau_h\ge 0}\min_{0\le\alpha\le K_\alpha}\min_{\tau_g\ge 0}\min_{\tau\ge 0}\quad \frac{\delta\beta}{2(\tau_g+\beta)}\left(\alpha^2+\sigma^2\right)+\frac{1}{p}G\left(\alpha\boldsymbol{g}-\boldsymbol{\omega};\frac{\tau_g}{\beta},\tau\right)+\frac{\beta\tau_g}{2}$$
$$-\frac{\alpha}{2\tau_h}\gamma^2-\frac{\alpha\beta^2}{2\tau_h}+\gamma\sqrt{\frac{\alpha^2\beta^2}{\tau_h^2}+V^2}-\frac{\alpha\tau_h}{2}\tag{6.21}$$

To simplify further we also need an asymptotic characterization of $\frac{1}{p}G\left(\alpha\boldsymbol{g}-\boldsymbol{\omega};\frac{\tau_g}{\beta},\tau\right)$. To this aim we prove the following lemma with the proof deferred to Appendix C.4.

Lemma 6.4. *Let $\boldsymbol{w}\in\mathbb{R}^n$ be a Gaussian random vector distributed as $\mathcal{N}(\boldsymbol{0},\omega^2\boldsymbol{I}_n)$. Also assume*

$$G(\boldsymbol{w};\mu,\tau):=\frac{1}{2\mu(\mu+1)}\left\|\boldsymbol{w}-\mathsf{ST}(\boldsymbol{w};\tau)\right\|_{\ell_2}^2-\frac{1}{2n}\left(\frac{p}{\varepsilon}\gamma-\frac{1}{1+\mu}\left\|\mathsf{ST}(\boldsymbol{w};\tau)\right\|_{\ell_1}\right)_+^2.$$

Then

$$\lim_{n\to\infty}\frac{1}{n}G(\boldsymbol{w};\mu,\tau)=\frac{\omega^2}{2\mu(\mu+1)}\left(\left(1-\sqrt{\frac{2}{\pi}}\frac{\tau}{\omega}e^{-\frac{\tau^2}{2\omega^2}}\right)+\left(\frac{\tau^2}{\omega^2}-1\right)\operatorname{erfc}\left(\frac{1}{\sqrt{2}}\frac{\tau}{\omega}\right)\right)$$
$$-\frac{\omega^2}{2(\mu+1)^2}\left(\frac{\gamma(\mu+1)}{\delta\varepsilon\omega}+\frac{\tau}{\omega}\cdot\operatorname{erfc}\left(\frac{1}{\sqrt{2}}\frac{\tau}{\omega}\right)-\sqrt{\frac{2}{\pi}}e^{-\frac{\tau^2}{2\omega^2}}\right)_+^2.$$

Furthermore,

$$\min_{\tau\ge 0}\lim_{n\to\infty}\frac{1}{n}G(\boldsymbol{w};\mu,\tau)$$
$$=\begin{cases}0 & \text{if } \gamma(\mu+1)\le\sqrt{\frac{2}{\pi}}\delta\varepsilon\omega\\ \frac{\omega^2}{2\mu(\mu+1)}\left(\operatorname{erf}\left(\frac{\tau^*\left(\frac{\gamma(\mu+1)}{\delta\varepsilon\omega},\mu\right)}{\sqrt{2}}\right)-\frac{\gamma(\mu+1)}{\delta\varepsilon\omega}\tau^*\left(\frac{\gamma(\mu+1)}{\delta\varepsilon\omega},\mu\right)\right) & \text{if } \gamma(\mu+1)>\sqrt{\frac{2}{\pi}}\delta\varepsilon\omega\end{cases}$$

where $\tau^*(a,\mu)$ is the unique solution to

$$a - \frac{1}{\mu}\tau - \tau \cdot \operatorname{erf}\left(\frac{\tau}{\sqrt{2}}\right) - \sqrt{\frac{2}{\pi}} e^{-\frac{\tau^2}{2}} = 0$$

Plugging the above lemma in (6.21) we arrive at

$$\max_{0\le\beta\le K_\beta} \sup_{\gamma,\tau_h\ge 0} \min_{0\le\alpha\le K_\alpha} \min_{\tau_g\ge 0} \quad D(\alpha,\beta,\gamma,\tau_h,\tau_g)\,, \tag{6.22}$$

where

$$\begin{aligned} D(\alpha,\beta,\gamma,\tau_h,\tau_g) =& \frac{\delta\beta}{2(\tau_g+\beta)}\left(\alpha^2+\sigma^2\right) \\ &+ \delta \mathbb{1}_{\left\{\gamma(\tau_g+\beta)>\sqrt{\frac{2}{\pi}}\delta\varepsilon\beta\sqrt{\alpha^2+\sigma^2}\right\}} \frac{\beta^2(\alpha^2+\sigma^2)}{2\tau_g(\tau_g+\beta)}\left(\operatorname{erf}\left(\frac{\tau_*}{\sqrt{2}}\right) - \frac{\gamma(\tau_g+\beta)}{\delta\varepsilon\beta\sqrt{\alpha^2+\sigma^2}}\tau_*\right) \\ &- \frac{\alpha}{2\tau_h}(\gamma^2+\beta^2) + \gamma\sqrt{\frac{\alpha^2\beta^2}{\tau_h^2}+V^2} - \frac{\alpha\tau_h}{2} + \frac{\beta\tau_g}{2} \end{aligned} \tag{6.23}$$

and τ_* is the unique solution to

$$\frac{\gamma(\tau_g+\beta)}{\delta\varepsilon\beta\sqrt{\alpha^2+\sigma^2}} - \frac{\beta}{\tau_g}\tau - \tau\cdot\operatorname{erf}\left(\frac{\tau}{\sqrt{2}}\right) - \sqrt{\frac{2}{\pi}}e^{-\frac{\tau^2}{2}} = 0 \tag{6.24}$$

This completes the scalarization of the AO.

Remarks 6.5. (Convergence analysis). *In above we showed the point wise convergence of the objective function in (6.20) to function D given by (6.23). However, what is required in this framework, is (local) uniform convergence so we get that the minimax solution of the objective function in (6.20) also converges to the minimax solution of the AO problem (6.23). This can be shown by following similar arguments as in [TAH18, Lemma A.5] that is essentially based on a result known as "convexity lemma" in the literature (see e.g. [LM08, Lemma 7.75]) by which point wise convergence of convex functions implies uniform convergence in compact subsets.*

6.6 Uniqueness of the solution of the AO problem

As we discussed after Equation (6.18), the function $f(\boldsymbol{v};\gamma)$ is convex in $\boldsymbol{v}$. Furthermore, we wrote (6.17) (part of the objective that depends on $\boldsymbol{v}$) in terms of the Moreau envelope $\frac{1}{p}e_f(\boldsymbol{\omega}-\alpha\boldsymbol{g};\frac{\tau_g}{\beta})$ and as $n\to\infty$, its limit goes to the *expected Moreau envelope*. Now by using the result of [TAH18, Lemma 4.4] the expected Moreau envelope of a convex function is *strictly* convex (without requiring any strong or strict convexity assumption on the function itself). Therefore, the convexity-concavity property discussed after (6.20) is preserved after taking the limit and the AO objective $D(\alpha,\beta,\gamma,\tau_h,\tau_g)$ is jointly strictly convex in (α,τ_g) and jointly concave in (β,γ,τ_h).

We next note that $\sup_{\beta,\gamma,\tau_h} D(\alpha,\beta,\gamma,\tau_h,\tau_g)$ is strictly convex in (α,τ_g). This follows from the fact that if $f(\boldsymbol{x},\boldsymbol{y})$ is strictly convex in $\boldsymbol{x}$, then $\sup_{\boldsymbol{y}} f(\boldsymbol{x},\boldsymbol{y})$ is also strictly convex in $\boldsymbol{x}$. We next use [TAH18, Lemma C.5] to conclude that $\inf_{\tau_g}\sup_{\beta,\gamma,\tau_h} D(\alpha,\beta,\gamma,\tau_h,\tau_g)$ is strictly convex in $\alpha>0$.Therefore, its minimizer over $\alpha\ge 0$ is unique, which completes the proof.

6.7 Proofs for fundamental tradeoffs

6.7.1 Proof of Lemma 3.1

We have

$$\mathsf{SR}(\widehat{\boldsymbol{\theta}}) := \frac{1}{p}\,\mathbb{E}\left[(y - \langle \boldsymbol{x}, \widehat{\boldsymbol{\theta}}\rangle)^2\right] = \frac{1}{p}\,\mathbb{E}\left[(w - \langle \boldsymbol{x}, \widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0\rangle)^2\right] = \frac{\sigma_0^2}{p} + \frac{1}{p}\left\|\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0\right\|_{\ell_2}^2. \tag{6.25}$$

To characterize $\mathsf{AR}(\widehat{\boldsymbol{\theta}})$, note that by following a similar argument as in Section 6.2, the solution of the problem

$$\max_{\|\boldsymbol{\delta}\|_{\ell_2} \le \varepsilon_{\text{test}}} (y - \langle \boldsymbol{x} + \boldsymbol{\delta}, \widehat{\boldsymbol{\theta}}\rangle)^2$$

is given by

$$\boldsymbol{\delta}_i = -\varepsilon_{\text{test}} \operatorname{sgn}(y - \langle \boldsymbol{x}, \widehat{\boldsymbol{\theta}}\rangle) \frac{\widehat{\boldsymbol{\theta}}}{\|\widehat{\boldsymbol{\theta}}\|_{\ell_2}}.$$

Therefore the adversarial risk can be written as

$$\mathsf{AR}(\widehat{\boldsymbol{\theta}}) = \frac{1}{p}\,\mathbb{E}\left[\left(|y - \langle \boldsymbol{x}, \widehat{\boldsymbol{\theta}}\rangle| + \varepsilon_{\text{test}} \|\widehat{\boldsymbol{\theta}}\|_{\ell_2}\right)^2\right] \tag{6.26}$$

By substituting for $y = \langle \boldsymbol{x}, \boldsymbol{\theta}_0\rangle + w$ and expanding the terms, we get

$$\begin{aligned}
&\mathbb{E}\left[\left(|y - \langle \boldsymbol{x}, \widehat{\boldsymbol{\theta}}\rangle| + \varepsilon_{\text{test}} \|\widehat{\boldsymbol{\theta}}\|_{\ell_2}\right)^2\right] \\
&= \mathbb{E}[\langle \boldsymbol{x}, \boldsymbol{\theta}_0 - \widehat{\boldsymbol{\theta}}\rangle^2] + \mathbb{E}[w^2] + \varepsilon_{\text{test}}^2 \|\widehat{\boldsymbol{\theta}}\|_{\ell_2}^2 + 2\varepsilon_{\text{test}} \|\widehat{\boldsymbol{\theta}}\|_{\ell_2}\, \mathbb{E}\left[|\langle \boldsymbol{x}, \widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0\rangle + w|\right] \\
&= \|\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0\|_{\ell_2}^2 + \sigma_0^2 + \varepsilon_{\text{test}}^2 \|\widehat{\boldsymbol{\theta}}\|_{\ell_2}^2 + 2\sqrt{\frac{2}{\pi}}\varepsilon_{\text{test}} \|\widehat{\boldsymbol{\theta}}\|_{\ell_2} \left(\sigma_0^2 + \|\boldsymbol{\theta}_0 - \widehat{\boldsymbol{\theta}}\|_{\ell_2}^2\right)^{1/2},
\end{aligned} \tag{6.27}$$

where in the first line we used the fact that $\widehat{\boldsymbol{\theta}}$ is independent of $\boldsymbol{x}$ and w (the test data and the corresponding response) and in the second line we used that $\langle \boldsymbol{x}, \widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0\rangle + w \sim \mathsf{N}(0, \sigma^2 + \|\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0\|_{\ell_2}^2)$ since $\boldsymbol{x} \sim \mathsf{N}(0, \boldsymbol{I}_p)$. This completes the proof.

6.7.2 Proof of Proposition 3.2

By definition,

$$\boldsymbol{\theta}^\lambda = \arg\min_{\boldsymbol{\theta}} \;\; \lambda \mathsf{SR}(\boldsymbol{\theta}) + \mathsf{AR}(\boldsymbol{\theta})$$

Substituting for $\mathsf{SR}(\boldsymbol{\theta})$ and $\mathsf{AR}(\boldsymbol{\theta})$ from Lemma 6.7.1 and scaling the objective by a factor p, we get

$$\boldsymbol{\theta}^\lambda = \arg\min_{\boldsymbol{\theta}} \;\; (1+\lambda)\left(\sigma_0^2 + \|\boldsymbol{\theta} - \boldsymbol{\theta}_0\|_{\ell_2}^2\right) + \varepsilon_{\text{test}}^2 \|\boldsymbol{\theta}\|_{\ell_2}^2 + 2\sqrt{\frac{2}{\pi}}\varepsilon_{\text{test}} \|\boldsymbol{\theta}\|_{\ell_2} \left(\sigma_0^2 + \|\boldsymbol{\theta}_0 - \boldsymbol{\theta}\|_{\ell_2}^2\right)^{1/2}$$

Now by setting the derivative to zero we arrive at the following identity for $\boldsymbol{\theta}^\lambda$:

$$(1+\lambda)(\boldsymbol{\theta}^\lambda - \boldsymbol{\theta}_0) + \varepsilon_{\text{test}}^2 \boldsymbol{\theta}^\lambda + \sqrt{\frac{2}{\pi}}\varepsilon_{\text{test}} \left(\frac{\boldsymbol{\theta}^\lambda}{\|\boldsymbol{\theta}^\lambda\|} \left(\sigma_0^2 + \|\boldsymbol{\theta}_0 - \boldsymbol{\theta}\|_{\ell_2}^2\right)^{1/2} + \frac{\|\boldsymbol{\theta}^\lambda\|}{\left(\sigma_0^2 + \|\boldsymbol{\theta}_0 - \boldsymbol{\theta}\|_{\ell_2}^2\right)^{1/2}} (\boldsymbol{\theta}^\lambda - \boldsymbol{\theta}_0)\right) = 0. \tag{6.28}$$

Adopting the shorthand

$$A^\lambda := \frac{1}{\|\boldsymbol{\theta}^\lambda\|}\left(\sigma_0^2 + \left\|\boldsymbol{\theta}_0 - \boldsymbol{\theta}^\lambda\right\|_{\ell_2}^2\right)^{1/2}$$

and rearranging the terms term we get

$$\left[\left(1+\lambda+\sqrt{\frac{2}{\pi}}\frac{\varepsilon_{\text{test}}}{A^\lambda}\right)+\left(\varepsilon_{\text{test}}^2+\sqrt{\frac{2}{\pi}}\varepsilon_{\text{test}}A^\lambda\right)\boldsymbol{I}\right]\boldsymbol{\theta}^\lambda = \left(1+\lambda+\sqrt{\frac{2}{\pi}}\frac{\varepsilon_{\text{test}}}{A}\right)\boldsymbol{\theta}_0\,.$$

The above equation can be written as

$$\boldsymbol{\theta}^\lambda = (1+\gamma_0^\lambda)^{-1}\boldsymbol{\theta}_0\,,$$

with

$$\gamma_0 := \frac{\varepsilon_{\text{test}}^2+\sqrt{\frac{2}{\pi}}\varepsilon_{\text{test}}A^\lambda}{1+\lambda+\sqrt{\frac{2}{\pi}}\frac{\varepsilon_{\text{test}}}{A}}\,,$$

which is the desired claim. The proof is complete by noting that

$$A^\lambda := \frac{1}{\|\boldsymbol{\theta}^\lambda\|_{\ell_2}}\left(\sigma_0^2+\|\boldsymbol{\theta}_0-\boldsymbol{\theta}\|_{\ell_2}^2\right)^{1/2} = \frac{1}{\|\boldsymbol{\theta}_0\|_{\ell_2}}\left((1+\gamma_0^\lambda)^2\sigma_0^2+(\gamma_0^\lambda)^2\|\boldsymbol{\theta}_0\|_{\ell_2}^2\right)^{1/2}.$$

6.8 Proofs for algorithmic tradeoffs

6.8.1 Proof of Theorem 3.3

We have already prove part (a) in the previous sections. Part (b) is also trivial from (6.23) as

$$\lim_{n\to\infty}\frac{1}{p}\left\|\widehat{\boldsymbol{\theta}}^\varepsilon-\boldsymbol{\theta}_0\right\|_{\ell_2}^2 = \lim_{n\to\infty}\|\widehat{z}^\varepsilon\|_{\ell_2}^2 = \alpha_*^2.$$

We thus turn our attention to part (c) and discuss how to calculate $\frac{\|\widehat{\boldsymbol{\theta}}^\varepsilon\|_{\ell_2}}{\sqrt{p}}$ asymptotically. As discussed earlier using a change of variable of the form $\boldsymbol{\theta} = \boldsymbol{\theta}_0 + \sqrt{p}\boldsymbol{z}$ the optimization problem can be written in the form

$$\min_{\boldsymbol{z}\in\mathcal{S}_{\boldsymbol{z}},\boldsymbol{v}}\max_{\boldsymbol{u}\in\mathcal{S}_{\boldsymbol{u}}}\frac{1}{\sqrt{p}}\left(\boldsymbol{u}^T\boldsymbol{X}\boldsymbol{z}-\boldsymbol{u}^T\boldsymbol{\omega}+\boldsymbol{u}^T\boldsymbol{v}\right)+\ell(\boldsymbol{v};\boldsymbol{z})$$

where

$$\ell(\boldsymbol{v};\boldsymbol{z}) := \frac{1}{2p}\left(\|\boldsymbol{v}\|_{\ell_2}^2+2\frac{\varepsilon}{\sqrt{p}}\|\boldsymbol{v}\|_{\ell_1}\|\boldsymbol{\theta}_0+\sqrt{p}\boldsymbol{z}\|_{\ell_2}+\frac{\varepsilon^2}{p}\|\boldsymbol{\theta}_0+\sqrt{p}\boldsymbol{z}\|_{\ell_2}^2\right)$$

As in the previous argument on calculating $\left\|\widehat{\boldsymbol{\theta}}^\varepsilon-\boldsymbol{\theta}_0\right\|_{\ell_2}$ asymptotically via the AO we proceed by writing $\ell(\boldsymbol{v};\boldsymbol{z})$ in terms of its conjugate with respect to $\boldsymbol{z}$. That is,

$$\ell(\boldsymbol{v};\boldsymbol{z}) = \sup_{\boldsymbol{q}}\boldsymbol{q}^T\boldsymbol{z}-\widetilde{\ell}(\boldsymbol{v};\boldsymbol{q})$$

As discussed in Section 6.3 the conjugate function takes the form

$$\widetilde{\ell}(\boldsymbol{v};\boldsymbol{q}) = -\frac{1}{\sqrt{p}}\boldsymbol{q}^T\boldsymbol{\theta}_0 + \frac{1}{2\delta p^2}\left(\frac{p}{\varepsilon}\left\|\boldsymbol{q}\right\|_{\ell_2} - \left\|\boldsymbol{v}\right\|_{\ell_1}\right)_+^2 - \frac{1}{2p}\left\|\boldsymbol{v}\right\|_{\ell_2}^2$$

and the AO problem can therefore be written as (same as (6.12))

$$\min_{0\le\alpha\le K_\alpha,\boldsymbol{v}}\max_{0\le\beta\le K_\beta}\max_{\boldsymbol{q}}\ \frac{\beta}{\sqrt{p}}\left\|\alpha\boldsymbol{g}-\boldsymbol{\omega}+\boldsymbol{v}\right\|_{\ell_2} - \alpha\left\|\frac{\beta}{\sqrt{p}}\boldsymbol{h}+\boldsymbol{q}\right\|_{\ell_2} - \widetilde{\ell}(\boldsymbol{v};\boldsymbol{q})$$

Our key observation is that the same AO can be used to calculate $\frac{\|\widehat{\boldsymbol{\theta}}^{\varepsilon}\|_{\ell_2}}{\sqrt{p}}$. To make this precise we show how to write $\left\|\widehat{\boldsymbol{\theta}}^{\varepsilon}\right\|_{\ell_2}$ in terms of functions of the $\boldsymbol{q}$ and $\boldsymbol{v}$ that maximizes the AO. To this aim note that $\widehat{\boldsymbol{z}} = \frac{1}{\sqrt{p}}\left(\widehat{\boldsymbol{\theta}}^{\varepsilon} - \boldsymbol{\theta}_0\right)$ obeys

$$\begin{aligned}
\widehat{\boldsymbol{z}} &= \arg\max_{\boldsymbol{z}}\ \boldsymbol{q}^T\boldsymbol{z} - \ell(\boldsymbol{v};\boldsymbol{z}) \\
&= \arg\max_{\boldsymbol{z}}\ \boldsymbol{q}^T\boldsymbol{z} - \frac{1}{2p}\sum_{i=1}^{n}\left(|v_i| + \frac{\varepsilon}{\sqrt{p}}\left\|\boldsymbol{\theta}_0+\sqrt{p}\boldsymbol{z}\right\|_{\ell_2}\right)^2 \\
&= \arg\max_{\boldsymbol{z}}\ \boldsymbol{q}^T\boldsymbol{z} - \frac{1}{2p}\left(\left\|\boldsymbol{v}\right\|_{\ell_2}^2 + \frac{2\varepsilon}{\sqrt{p}}\left\|\boldsymbol{v}\right\|_{\ell_1}\left\|\boldsymbol{\theta}_0+\sqrt{p}\boldsymbol{z}\right\|_{\ell_2} + \delta\varepsilon^2\left\|\boldsymbol{\theta}_0+\sqrt{p}\boldsymbol{z}\right\|_{\ell_2}^2\right)
\end{aligned}$$

Setting derivative w.r.t $\boldsymbol{z}$ to zero we arrive at

$$\boldsymbol{q} - \frac{\varepsilon}{p^{3/2}}\left\|\boldsymbol{v}\right\|_{\ell_1}\frac{\boldsymbol{\theta}_0+\sqrt{p}\widehat{\boldsymbol{z}}}{\left\|\boldsymbol{\theta}_0+\sqrt{p}\widehat{\boldsymbol{z}}\right\|_{\ell_2}}\sqrt{p} - \frac{\delta\varepsilon^2}{p}(\boldsymbol{\theta}_0+\sqrt{p}\widehat{\boldsymbol{z}})\sqrt{p} = 0 \tag{6.29}$$

Therefore

$$\boldsymbol{\theta}_0+\sqrt{p}\widehat{\boldsymbol{z}} = \left(\frac{\varepsilon\left\|\boldsymbol{v}\right\|_{\ell_1}}{p\left\|\boldsymbol{\theta}_0+\sqrt{p}\widehat{\boldsymbol{z}}\right\|_{\ell_2}} + \frac{\delta\varepsilon^2}{\sqrt{p}}\right)^{-1}\boldsymbol{q}.$$

Thus taking Euclidean norm of both sides of the identity we have

$$\left\|\boldsymbol{\theta}_0+\sqrt{p}\widehat{\boldsymbol{z}}\right\|_{\ell_2}\left(\frac{\varepsilon\left\|\boldsymbol{v}\right\|_{\ell_1}}{p\left\|\boldsymbol{\theta}_0+\sqrt{p}\widehat{\boldsymbol{z}}\right\|_{\ell_2}} + \frac{\delta\varepsilon^2}{\sqrt{p}}\right) = \left\|\boldsymbol{q}\right\|_{\ell_2} \Rightarrow$$

$$\left\|\boldsymbol{\theta}_0+\sqrt{p}\widehat{\boldsymbol{z}}\right\|_{\ell_2} = \frac{\left\|\boldsymbol{q}\right\|_{\ell_2} - \frac{\varepsilon\|\boldsymbol{v}\|_{\ell_1}}{p}}{\frac{\delta\varepsilon^2}{\sqrt{p}}} = \frac{\sqrt{p}}{\delta\varepsilon^2}\left\|\boldsymbol{q}\right\|_{\ell_2} - \frac{1}{\delta\varepsilon\sqrt{p}}\left\|\boldsymbol{v}\right\|_{\ell_1}.$$

The latter holds as long as $\frac{\sqrt{p}}{\delta\varepsilon^2}\left\|\boldsymbol{q}\right\|_{\ell_2} \ge \frac{1}{\delta\varepsilon\sqrt{p}}\left\|\boldsymbol{v}\right\|_{\ell_1}$. When $\frac{\sqrt{p}}{\delta\varepsilon^2}\left\|\boldsymbol{q}\right\|_{\ell_2} < \frac{1}{\delta\varepsilon\sqrt{p}}\left\|\boldsymbol{v}\right\|_{\ell_1}$ it is easy to verify that that the objective value is smaller than or equal to $-\frac{\boldsymbol{q}^T\boldsymbol{\theta}_0}{\sqrt{p}} - \frac{1}{2p}\left\|\boldsymbol{v}\right\|_{\ell_2}^2$ and therefore $\widehat{\boldsymbol{z}} = -\boldsymbol{\theta}_0/\sqrt{p}$ which in turn implies that $\left\|\widehat{\boldsymbol{\theta}}^{\varepsilon}\right\|_{\ell_2} = \left\|\boldsymbol{\theta}_0+\sqrt{p}\widehat{\boldsymbol{z}}\right\|_{\ell_2} = 0$. We thus have

$$\frac{1}{\sqrt{p}}\left\|\widehat{\boldsymbol{\theta}}^{\varepsilon}\right\|_{\ell_2} = \frac{1}{\sqrt{p}}\left\|\sqrt{p}\widehat{\boldsymbol{z}}+\boldsymbol{\theta}_0\right\|_{\ell_2} = \frac{1}{\sqrt{p}}\left(\frac{\sqrt{p}}{\delta\varepsilon^2}\left\|\boldsymbol{q}\right\|_{\ell_2} - \frac{1}{\delta\varepsilon\sqrt{p}}\left\|\boldsymbol{v}\right\|_{\ell_1}\right)_+ = \frac{1}{\delta\varepsilon p}\left(\frac{p}{\varepsilon}\left\|\boldsymbol{q}\right\|_{\ell_2} - \left\|\boldsymbol{v}\right\|_{\ell_1}\right)_+. \tag{6.30}$$

So to get the asymptotic value of $\frac{1}{\sqrt{p}}\left\|\widehat{\boldsymbol{\theta}}^{\varepsilon}\right\|_{\ell_2}$ we can simply look at $\frac{1}{\delta\varepsilon p}\left(\frac{p}{\varepsilon}\gamma-\|\boldsymbol{v}\|_{\ell_1}\right)_+$ with $\boldsymbol{v}$ and $\gamma=\|\boldsymbol{q}\|_{\ell_2}$ the optimal solutions of the AO. Note that based on the argument in Lemma 6.4 for this optimal solution of $\boldsymbol{v}$ we have

$$\lim_{n\to+\infty}\frac{1}{n^2}\left(\frac{p}{\varepsilon}\gamma-\|\boldsymbol{v}\|_{\ell_1}\right)^2=\frac{\omega^2}{(\mu+1)^2}\left(\frac{\gamma(\mu+1)}{\delta\varepsilon\omega}+\tau^*\cdot\operatorname{erfc}\left(\frac{1}{\sqrt{2}}\tau^*\right)-\sqrt{\frac{2}{\pi}}e^{-\frac{(\tau^*)^2}{2}}\right)_+^2 \tag{6.31}$$

with $\omega=\sqrt{\alpha^2+\sigma^2}$, $\mu=\frac{\tau_g}{\beta}$, $\tau^*:=\tau^*\left(\frac{\gamma(\mu+1)}{\delta\varepsilon\omega},\mu\right)$ and $\tau^*(a,\mu)$ is the unique solution to

$$a-\frac{\mu+1}{\mu}\tau+\tau\cdot\operatorname{erfc}\left(\frac{\tau}{\sqrt{2}}\right)-\sqrt{\frac{2}{\pi}}e^{-\frac{\tau^2}{2}}=0 \tag{6.32}$$

Therefore, squaring (6.30) and plugging in (6.31) we conclude that

$$\begin{aligned}
\lim_{p\to\infty}\frac{1}{p}\left\|\widehat{\boldsymbol{\theta}}^{\varepsilon}\right\|_{\ell_2}^2&=\lim_{p\to\infty}\frac{1}{\delta^2\varepsilon^2p^2}\left(\frac{p}{\varepsilon}\gamma-\|\boldsymbol{v}\|_{\ell_1}\right)_+^2\\
&=\frac{1}{\varepsilon^2}\cdot\lim_{n\to\infty}\frac{1}{n^2}\left(\frac{p}{\varepsilon}\gamma-\|\boldsymbol{v}\|_{\ell_1}\right)_+^2\\
&=\frac{\omega^2}{\varepsilon^2(\mu+1)^2}\left(\frac{\gamma(\mu+1)}{\delta\varepsilon\omega}+\tau^*\cdot\operatorname{erfc}\left(\frac{1}{\sqrt{2}}\tau^*\right)-\sqrt{\frac{2}{\pi}}e^{-\frac{(\tau^*)^2}{2}}\right)_+^2\\
&=\frac{\omega^2}{\varepsilon^2(\mu+1)^2}\left(\frac{\gamma(\mu+1)}{\delta\varepsilon\omega}+\frac{\mu+1}{\mu}\tau^*-\frac{\gamma(\mu+1)}{\delta\varepsilon\omega}\right)_+^2\\
&=\frac{\omega^2}{\varepsilon^2(\mu+1)^2}\left(\frac{\mu+1}{\mu}\tau^*\right)_+^2\\
&=\frac{\omega^2\tau_*^2}{\varepsilon^2\mu^2}\\
&=\frac{(\alpha_*^2+\sigma^2)\tau_*^2}{\varepsilon^2\mu^2}.
\end{aligned}$$

6.8.2 Proof of Corollary 3.4

The result follows readily from Lemma 3.1 along with Theorem 3.3 (Parts (b) and (c)).

6.8.3 Proof of Theorem 3.5

We start by analyzing $\lim_{p\to\infty}\mathsf{SR}(\boldsymbol{\theta}^\lambda)$ and $\lim_{p\to\infty}\mathsf{AR}(\boldsymbol{\theta}^\lambda)$. Using Lemma 3.1, we have

$$\begin{aligned}
\lim_{p\to\infty}\mathsf{SR}(\boldsymbol{\theta}^\lambda)&=\sigma^2+\lim_{p\to\infty}\frac{1}{p}\left\|\boldsymbol{\theta}^\lambda-\boldsymbol{\theta}_0\right\|_{\ell_2}^2\\
&=\sigma^2+\lim_{p\to\infty}\frac{1}{p}\left\|\boldsymbol{\theta}_0\right\|_{\ell_2}^2\left(\frac{\gamma_0^\lambda}{1+\gamma_0^\lambda}\right)^2\\
&=\sigma^2+\left(\frac{\gamma_0^\lambda V}{1+\gamma_0^\lambda}\right)^2.
\end{aligned} \tag{6.33}$$

Likewise,

$$\lim_{p\to\infty} \mathsf{AR}(\boldsymbol{\theta}^\lambda) = \sigma^2 + V^2\left(\frac{\gamma_0^\lambda}{1+\gamma_0^\lambda}\right)^2 + \varepsilon_{\text{test}}^2 \frac{V^2}{(1+\gamma_0^\lambda)^2} + 2\sqrt{\frac{2}{\pi}}\frac{\varepsilon_{\text{test}} V}{1+\gamma_0^\lambda}\left(\sigma^2 + \left(\frac{\gamma_0^\lambda V}{1+\gamma_0^\lambda}\right)^2\right)^{1/2}, \tag{6.34}$$

with γ_0^λ the fixed point of the following two equations:

$$\gamma_0^\lambda = \frac{\varepsilon_{\text{test}}^2 + \sqrt{\frac{2}{\pi}}\varepsilon_{\text{test}} A^\lambda}{1+\lambda+\sqrt{\frac{2}{\pi}}\frac{\varepsilon_{\text{test}}}{A^\lambda}}, \quad A^\lambda = \frac{1}{V}\left((1+\gamma_0^\lambda)^2\sigma^2 + (\gamma_0^\lambda)^2 V^2\right)^{1/2}. \tag{6.35}$$

We next analyze $\lim_{\delta\to\infty}\lim_{n\to\infty}\mathsf{SR}(\widehat{\boldsymbol{\theta}}^\varepsilon)$ and $\lim_{\delta\to\infty}\lim_{n\to\infty}\mathsf{AR}(\widehat{\boldsymbol{\theta}}^\varepsilon)$. By using Corollary 3.4, we have

$$\lim_{\delta\to\infty}\lim_{n\to\infty}\mathsf{SR}(\widehat{\boldsymbol{\theta}}^\varepsilon) = \lim_{\delta\to\infty}\left(\sigma^2+\alpha_*^2\right), \tag{6.36}$$

$$\lim_{\delta\to\infty}\lim_{n\to\infty}\mathsf{AR}(\widehat{\boldsymbol{\theta}}^\varepsilon) = \lim_{\delta\to\infty}\left\{\sigma^2+\alpha_*^2+\varepsilon_{\text{test}}^2(\alpha_*^2+\sigma^2)\left(\frac{\beta_*\tau_*}{\varepsilon\tau_{g*}}\right)^2 + 2\sqrt{\frac{2}{\pi}}\frac{\varepsilon_{\text{test}}\beta_*\tau_*}{\varepsilon\tau_{g*}}(\sigma^2+\alpha_*^2)\right\}. \tag{6.37}$$

Therefore, we need to study the solution of the convex-concave minimax optimization (6.23) at the limits $\delta\to\infty$. It is straightforward to see that as $\delta\to\infty$, the indicator in (6.23) is active and hence it reduces to

$$\begin{aligned} D(\alpha,\beta,\gamma,\tau_h,\tau_g) =& \frac{\delta\beta}{2(\tau_g+\beta)}\left(\alpha^2+\sigma^2\right) + \frac{\delta\beta^2(\alpha^2+\sigma^2)}{2\tau_g(\tau_g+\beta)}\operatorname{erf}\left(\frac{\tau_*}{\sqrt{2}}\right) \\ &- \frac{\alpha}{2\tau_h}\gamma^2 + \gamma\left(\sqrt{\frac{\alpha^2\beta^2}{\tau_h^2}+V^2} - \frac{\beta\tau_*\sqrt{\alpha^2+\sigma^2}}{\varepsilon\tau_g}\right) \\ &- \frac{\alpha}{2\tau_h}\beta^2 - \frac{\alpha\tau_h}{2} + \frac{\beta\tau_g}{2}. \end{aligned} \tag{6.38}$$

Solving for γ, we obtain

$$\gamma_* = \frac{\tau_h}{\alpha}\left(\sqrt{\frac{\alpha^2\beta^2}{\tau_h^2}+V^2} - \frac{\beta\tau_*\sqrt{\alpha^2+\sigma^2}}{\varepsilon\tau_g}\right).$$

Since $\gamma(\tau_g+\beta) > \sqrt{\frac{2}{\pi}}\delta\varepsilon\beta\sqrt{\alpha^2+\sigma^2}$, we have $\gamma\to\infty$ as $\delta\to\infty$, and by the above equation for γ_*, we obtain that $\tau_h\to\infty$. Therefore,

$$\gamma_* \to \frac{\tau_h}{\alpha}\left(V - \frac{\beta\tau_*\sqrt{\alpha^2+\sigma^2}}{\varepsilon\tau_g}\right). \tag{6.39}$$

In addition, $\tau_*\to 0$ as $\delta\to\infty$. Writing the Taylor expansion of the characteristic equation of τ_* as per (6.24), we get

$$\frac{\gamma(\tau_g+\beta)}{\beta\delta\varepsilon\sqrt{\alpha^2+\sigma^2}} = \sqrt{\frac{2}{\pi}} + \frac{\beta\tau_*}{\tau_g} + O(\tau_*^2). \tag{6.40}$$

We adopt the shorthands $\omega := \sqrt{\alpha^2+\sigma^2}$ and $\mu := \frac{\tau_g}{\beta}$. Combining (6.40) with (6.39) yields

$$\frac{\tau_h(\mu+1)}{\alpha\delta\varepsilon\omega}\left(V-\frac{\tau_*\omega}{\varepsilon\mu}\right)=\sqrt{\frac{2}{\pi}}+\frac{\tau_*}{\mu}+O(\tau_*^2)\,.$$

Writing the objective D given by (6.38) in terms of ω, μ, η and after substituting for γ_* we arrive at

$$\begin{aligned}D=&\frac{\delta\omega^2}{2(\mu+1)}+\frac{\delta\omega^2}{2\mu(\mu+1)}\mathrm{erf}\left(\frac{\tau_*}{\sqrt{2}}\right)\\&+\frac{\tau_h}{2\alpha}\left(V-\frac{\tau_*\omega}{\varepsilon\mu}\right)^2-\frac{\alpha}{2\tau_h}\beta^2-\frac{\alpha\tau_h}{2}+\frac{\beta^2\mu}{2}\,.\end{aligned}\tag{6.41}$$

Since $\delta,\tau_h\to\infty$, keeping only the dominant terms results in

$$D=\frac{\delta\omega^2}{2(\mu+1)}+\frac{\delta\omega^2}{2\mu(\mu+1)}\mathrm{erf}\left(\frac{\tau_*}{\sqrt{2}}\right)+\frac{\tau_h}{2\alpha}\left(V-\frac{\tau_*\omega}{\varepsilon\mu}\right)^2-\frac{\alpha\tau_h}{2}\,,\tag{6.42}$$

and by keeping only terms of $O(\tau_*^2)$ we have

$$D=\frac{\delta\omega^2}{2(\mu+1)}\left(1+\sqrt{\frac{2}{\pi}}\frac{\tau_*}{\mu}\right)+\frac{\tau_h}{2\alpha}\left(V-\frac{\tau_*\omega}{\varepsilon\mu}\right)^2-\frac{\alpha\tau_h}{2}\,.\tag{6.43}$$

Setting the derivative of D, with respect to τ_h, to zero, we get

$$\alpha=V-\frac{\tau_*\omega}{\varepsilon\mu}\,.\tag{6.44}$$

We next set the derivative of D, with respect to α, to zero, which implies

$$\frac{\delta\alpha}{\mu+1}\left(1+\sqrt{\frac{2}{\pi}}\frac{\tau_*}{\mu}\right)-\frac{\tau_h}{2\alpha^2}\left(V-\frac{\tau_*\omega}{\varepsilon\mu}\right)^2-\frac{\tau_h}{\alpha}\left(V-\frac{\tau_*\omega}{\varepsilon\mu}\right)\frac{\tau_*}{\varepsilon\mu}\frac{\alpha}{\omega}-\frac{\tau_h}{2}=0\,.$$

Plugging in for α from (6.44) we obtain

$$\alpha=\varepsilon\omega\frac{\sqrt{\frac{2}{\pi}}+\frac{\tau_*}{\mu}}{1+\sqrt{\frac{2}{\pi}}\frac{\tau_*}{\mu}}\left(1+\frac{\tau_*\alpha}{\varepsilon\mu\omega}\right)\,.\tag{6.45}$$

Defining $A^\varepsilon := \frac{\varepsilon\mu}{\tau_*}$ and $\gamma_0^\varepsilon := \frac{\varepsilon\mu V}{\tau_*\omega}-1$, the above two equations (6.44), (6.45) imply that

$$\alpha=V-\frac{\omega}{A^\varepsilon}=\frac{\gamma_0^\varepsilon V}{1+\gamma_0^\varepsilon},\tag{6.46}$$

$$\frac{\alpha\varepsilon}{\omega}=\varepsilon^2\frac{1+\sqrt{\frac{2}{\pi}}\frac{\mu}{\tau_*}}{\frac{\mu}{\tau_*}+\sqrt{\frac{2}{\pi}}}\left(1+\frac{\tau_*\alpha}{\varepsilon\mu\omega}\right)=\frac{\varepsilon^2+\sqrt{\frac{2}{\pi}}\varepsilon A^\varepsilon}{\frac{A^\varepsilon}{\varepsilon}+\sqrt{\frac{2}{\pi}}}\left(1+\frac{\tau_*\alpha}{\varepsilon\mu\omega}\right).\tag{6.47}$$

From (6.46) we obtain

$$\frac{1}{V}\left((1+\gamma_0^\varepsilon)^2\sigma^2+(\gamma_0^\varepsilon)^2V^2\right)^{1/2}=\frac{1+\gamma_0^\varepsilon}{V}\omega=A^\varepsilon.\tag{6.48}$$

In addition, from (6.46) and (6.47) we have

$$\gamma_0^\varepsilon = \frac{VA^\varepsilon}{\omega} - 1 = \frac{A^\varepsilon \alpha}{\omega} = \frac{\varepsilon^2 + \sqrt{\frac{2}{\pi}}\varepsilon A^\varepsilon}{1 + \sqrt{\frac{2}{\pi}}\frac{\varepsilon}{A^\varepsilon}} \left(1 + \frac{\gamma_0^\varepsilon}{(A^\varepsilon)^2}\right). \tag{6.49}$$

Combining equations (6.48) and (6.49), we have that γ_0^ε is the fixed point of the following two equations:

$$\gamma_0^\varepsilon = \frac{\varepsilon^2 + \sqrt{\frac{2}{\pi}}\varepsilon A^\varepsilon}{1 - (\frac{\varepsilon}{A^\varepsilon})^2}, \quad A^\varepsilon = \frac{1}{V}\left((1+\gamma_0^\varepsilon)^2\sigma^2 + (\gamma_0^\varepsilon)^2 V^2\right)^{1/2}. \tag{6.50}$$

Now consider a fixed $\lambda \geq 0$ and let $\gamma_0^\lambda, A^\lambda$ be defined by (6.35). Comparing equations (6.33) and (6.34) with (6.36) and (6.37), we see that in order to prove the statement, it suffices to find corresponding $\varepsilon \geq 0$ such that $\gamma_0^\varepsilon = \gamma_0^\lambda$ (Note that the statement $\gamma_0^\varepsilon = \gamma_0^\lambda$ implies that $A^\varepsilon = A^\lambda$ as well). Such value of ε is hence found from the following equation (which equates $\gamma_0^\varepsilon = \gamma_0^\lambda$ and $A^\lambda = A^\varepsilon$):

$$\frac{\varepsilon_{\text{test}}^2 + \sqrt{\frac{2}{\pi}}\varepsilon_{\text{test}} A^\lambda}{1 + \lambda + \sqrt{\frac{2}{\pi}}\frac{\varepsilon_{\text{test}}}{A^\lambda}} = \frac{\varepsilon^2 + \sqrt{\frac{2}{\pi}}\varepsilon A^\lambda}{1 - (\frac{\varepsilon}{A^\lambda})^2}.$$

Rearranging terms, we reach to:

$$\varepsilon^2\left(1 + \lambda + \sqrt{\frac{2}{\pi}}\frac{\varepsilon_{\text{test}}}{A^\lambda} + \left(\frac{\varepsilon_{\text{test}}}{A^\lambda}\right)^2 + \sqrt{\frac{2}{\pi}}\frac{\varepsilon_{\text{test}}}{A^\lambda}\right) + \varepsilon\sqrt{\frac{2}{\pi}}\left(A^\lambda(1+\lambda) + \sqrt{\frac{2}{\pi}}\varepsilon_{\text{test}}\right) - \left(\varepsilon_{\text{test}}^2 + \sqrt{\frac{2}{\pi}}\varepsilon_{\text{test}} A^\lambda\right) = 0.$$

The thesis now follows by noting that the above equation is a quadratic form in ε and has always a positive solution, which gives the value of ε in terms of λ.

Acknowledgements

A. Javanmard is partially supported by a Google Faculty Research Award and the NSF CAREER Award DMS-1844481. M. Soltanolkotabi is supported by the Packard Fellowship in Science and Engineering, a Sloan Research Fellowship in Mathematics, an NSF-CAREER under award #1846369, the Air Force Office of Scientific Research Young Investigator Program (AFOSR-YIP) under award #FA9550 – 18 – 1 – 0078, Darpa Learning with Less Labels (LwLL) program, an NSF-CIF award #1813877, and a Google faculty research award. This work was done in part while M.S. was visiting the Simons Institute for the Theory of Computing. The research of H. Hassani is supported by NSF HDR TRIPODS award 1934876, NSF award CPS-1837253, NSF award CIF-1910056, and NSF CAREER award CIF-1943064.

References

[BCM+13] Battista Biggio, Igino Corona, Davide Maiorca, Blaine Nelson, Nedim Šrndić, Pavel Laskov, Giorgio Giacinto, and Fabio Roli, *Evasion attacks against machine learning at test time*, Joint European conference on machine learning and knowledge discovery in databases, Springer, 2013, pp. 387–402. 1

[BHMM18] Mikhail Belkin, Daniel Hsu, Siyuan Ma, and Soumik Mandal, *Reconciling modern machine learning and the bias-variance trade-off*, arXiv preprint arXiv:1812.11118 (2018). 10, 33

[BLPR19] Sébastien Bubeck, Yin Tat Lee, Eric Price, and Ilya P. Razenshteyn, *Adversarial examples from computational constraints*, Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA, 2019, pp. 831–840. 12

[BMM18] Mikhail Belkin, Siyuan Ma, and Soumik Mandal, *To understand deep learning we need to understand kernel learning*, International Conference on Machine Learning, 2018, pp. 541–549. 10, 33

[CBM18] Daniel Cullina, Arjun Nitin Bhagoji, and Prateek Mittal, *Pac-learning in the presence of adversaries*, Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, 3-8 December 2018, Montréal, Canada, 2018, pp. 228–239. 12

[DKT19] Zeyu Deng, Abla Kammoun, and Christos Thrampoulidis, *A model of double descent for high-dimensional binary linear classification*, arXiv preprint arXiv:1911.05822 (2019). 2, 13

[GCL+19] Ruiqi Gao, Tianle Cai, Haochuan Li, Cho-Jui Hsieh, Liwei Wang, and Jason D. Lee, *Convergence of adversarial training in overparametrized neural networks*, Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, 8-14 December 2019, Vancouver, BC, Canada, 2019, pp. 13009–13020. 12

[GMF+18] Justin Gilmer, Luke Metz, Fartash Faghri, Samuel S Schoenholz, Maithra Raghu, Martin Wattenberg, and Ian Goodfellow, *Adversarial spheres*, arXiv preprint arXiv:1801.02774 (2018). 12

[Gor88] Yehoram Gordon, *On milman's inequality and random subspaces which escape through a mesh in* $\mathbb{R}^n$, Geometric aspects of functional analysis, Springer, 1988, pp. 84–106. 2, 13, 16

[GSS15] Ian J. Goodfellow, Jonathon Shlens, and Christian Szegedy, *Explaining and harnessing adversarial examples*, 3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings, 2015. 2

[HMRT19] Trevor Hastie, Andrea Montanari, Saharon Rosset, and Ryan J Tibshirani, *Surprises in high-dimensional ridgeless least squares interpolation*, arXiv preprint arXiv:1903.08560 (2019). 10, 33

[KGB16] Alexey Kurakin, Ian Goodfellow, and Samy Bengio, *Adversarial machine learning at scale*, arXiv preprint arXiv:1611.01236 (2016). 2

[KL18] Justin Khim and Po-Ling Loh, *Adversarial risk bounds for binary classification via function transformation*, CoRR **abs/1810.09519** (2018). 12

[LM08] Friedrich Liese and Klaus-J. Miescke, *Statistical decision theory: Estimation, testing, and selection*, Springer Science & Business Media, 2008. 22

[LS20] Tengyuan Liang and Pragya Sur, *A precise high-dimensional asymptotic theory for boosting and min-l1-norm interpolated classifiers*, arXiv preprint arXiv:2002.01586 (2020). 13

[MDM19] Saeed Mahloujifar, Dimitrios I Diochnos, and Mohammad Mahmoody, *The curse of concentration in robust learning: Evasion and poisoning attacks from concentration of measure*, Proceedings of the AAAI Conference on Artificial Intelligence, vol. 33, 2019, pp. 4536–4543. 12

[MHS19] Omar Montasser, Steve Hanneke, and Nathan Srebro, *VC classes are adversarially robustly learnable, but only improperly*, Conference on Learning Theory, COLT 2019, 25-28 June 2019, Phoenix, AZ, USA, 2019, pp. 2512–2530. 12

[MM19] Song Mei and Andrea Montanari, *The generalization error of random features regression: Precise asymptotics and double descent curve*, arXiv preprint arXiv:1908.05355 (2019). 11

[MMS+17] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu, *Towards deep learning models resistant to adversarial attacks*, arXiv preprint arXiv:1706.06083 (2017). 3

[MMS+18] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu, *Towards deep learning models resistant to adversarial attacks*, 6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings, 2018. 2, 5, 11

[MRSY19] Andrea Montanari, Feng Ruan, Youngtak Sohn, and Jun Yan, *The generalization error of max-margin linear classifiers: High-dimensional asymptotics in the overparametrized regime*, arXiv preprint arXiv:1911.01544 (2019). 13

[Nak19] Preetum Nakkiran, *Adversarial robustness may be at odds with simplicity*, arXiv preprint arXiv:1901.00532 (2019). 12

[PJ19] Muni Sreenivas Pydi and Varun Jog, *Adversarial risk via optimal transport and optimal couplings*, arXiv preprint arXiv:1912.02794 (2019). 11

[RSL18] Aditi Raghunathan, Jacob Steinhardt, and Percy Liang, *Certified defenses against adversarial examples*, 6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings, 2018. 2

[RXY+19] Aditi Raghunathan, Sang Michael Xie, Fanny Yang, John C Duchi, and Percy Liang, *Adversarial training can hurt generalization*, arXiv preprint arXiv:1906.06032 (2019). 2, 11, 12

[S+58] Maurice Sion et al., *On general minimax theorems.*, Pacific Journal of mathematics **8** (1958), no. 1, 171–176. 18

[SHS+19] Ali Shafahi, W. Ronny Huang, Christoph Studer, Soheil Feizi, and Tom Goldstein, *Are adversarial examples inevitable?*, 7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019, 2019. 12

[SST+18] Ludwig Schmidt, Shibani Santurkar, Dimitris Tsipras, Kunal Talwar, and Aleksander Madry, *Adversarially robust generalization requires more data*, Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, 3-8 December 2018, Montréal, Canada, 2018, pp. 5019–5031. 11

[SZS+14] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian J Goodfellow, and Rob Fergus, *Intriguing properties of neural networks. iclr, abs/1312.6199, 2014*, 2014. 1

[TAH15] Christos Thrampoulidis, Ehsan Abbasi, and Babak Hassibi, *Precise high-dimensional error analysis of regularized m-estimators*, 2015 53rd Annual Allerton Conference on Communication, Control, and Computing (Allerton), IEEE, 2015, pp. 410–417. 18, 43

[TAH18] ———, *Precise error analysis of regularized m-estimators in high dimensions*, IEEE Transactions on Information Theory **64** (2018), no. 8, 5592–5628. 2, 22, 38

[TOH15] Christos Thrampoulidis, Samet Oymak, and Babak Hassibi, *Regularized linear regression: A precise analysis of the estimation error*, Conference on Learning Theory, 2015, pp. 1683–1709. 2, 13, 16

[TSE+18] Dimitris Tsipras, Shibani Santurkar, Logan Engstrom, Alexander Turner, and Aleksander Madry, *Robustness may be at odds with accuracy*, arXiv preprint arXiv:1805.12152 (2018). 2, 3, 10, 11

[WK18] Eric Wong and J. Zico Kolter, *Provable defenses against adversarial examples via the convex outer adversarial polytope*, Proceedings of the 35th International Conference on Machine Learning, ICML 2018, Stockholmsmässan, Stockholm, Sweden, July 10-15, 2018, 2018, pp. 5283–5292. 2

[YRB19] Dong Yin, Kannan Ramchandran, and Peter L. Bartlett, *Rademacher complexity for adversarially robust generalization*, Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA, 2019, pp. 7085–7094. 12

[ZYJ+19] Hongyang Zhang, Yaodong Yu, Jiantao Jiao, Eric P. Xing, Laurent El Ghaoui, and Michael I. Jordan, *Theoretically principled trade-off between robustness and accuracy*, Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA, 2019, pp. 7472–7482. 2, 11

A Further insights and guarantees into the effect of the size of the training data

To provide further insight into the role of the size of the training data on adversarial training we note that we have already shown in our proofs (See Section 6.2 and equation (6.2)) that the inner

maximization in the saddle point problem (2.6) has a closed form solution and the estimator $\widehat{\boldsymbol{\theta}}^{\varepsilon}$ can be equivalently defined by

$$\widehat{\boldsymbol{\theta}}^{\varepsilon} \in \arg\min_{\boldsymbol{\theta}\in\mathbb{R}^p} \frac{1}{2p}\sum_{i=1}^{n}\left(|y_i - \langle \boldsymbol{x}_i, \boldsymbol{\theta}\rangle| + \varepsilon \|\boldsymbol{\theta}\|_{\ell_2}\right)^2 . \tag{A.1}$$

Therefore for linear regression, adversarial training by the saddle point optimization (2.2) amounts to a *regularized estimator*. When $\delta < 1$, we are in the overparametrized regime and regularization helps with standard accuracy. In particular, when $\delta \to 1$, the condition number of the covariate matrix diverges (a.k.a interpolation threshold [BMM18, BHMM18, HMRT19]) and the role of regularization becomes crucial, without which the standard risk would diverge. This is reflected in Figure 2 in that the standard risk diverges at $\varepsilon = 0$ as $\delta \to 1$, and also the statistical risk plummets quickly with ε ; See also Proposition A.1 below.

Nonetheless, in the $\delta > 1$ regime the effect of regularization starts to weaken. To see why, note that as δ grows, the ratio of sample size n to the dimension p increases, and the reduction in the variance of the estimator due to regularization becomes comparative to the increase in the bias caused by this term. As a result the overall positive effect of regularization on standard risk lessens and we see in Figure 3, the negative slope at $\varepsilon = 0$ decreases as δ increases. In addition, at large δ, the standard risk will start to quickly becomes increasing with ε. In other words, for larger δ, the negative effect of adversarial training on standard risk starts to emerge at smaller values of ε. (For example at $\delta = 10$, this effect kicks in at $\varepsilon = 0.15$.)

Our next proposition describes the standard risk at small values of ε.

Proposition A.1. *Under the assumptions of Theorem 3.3 and for $\delta \geq 1$ and $\varepsilon \leq 1$, we have*

$$\lim_{n\to\infty} \mathsf{SR}(\widehat{\boldsymbol{\theta}}^{\varepsilon}) = \frac{\delta\sigma^2}{\delta-1} - 2\sqrt{\frac{2}{\pi}}\frac{\sigma^3\delta^{3/2}}{(\delta-1)^2}\cdot\frac{1}{\sqrt{\sigma^2+V^2(\delta-1)}}\varepsilon + O(\varepsilon^2) . \tag{A.2}$$

As a result of Proposition A.1, for ε small and $\delta \geq 1$: (i) standard risk α_* falls with ε at vicinity of $\varepsilon = 0$ (ii) the risk falls slower at larger δ (iii) as $\delta \to 1$, the slope diverges and the risk plummets rapidly. These observations corroborates our justification and insights provided above.

We finish this appendix by the proof of Proposition A.1.

Proof of Proposition A.1. Define $\boldsymbol{x} = (\alpha, \beta, \tau_h, \tau_g, \gamma)$. We can write the objective of the convex-concave minimax problem (6.22) as

$$D(\alpha,\beta,\tau_h,\tau_g,\gamma) = \bar{D}(\alpha,\beta,\tau_h,\tau_g,\gamma) + \mathbb{1}_{\left\{\frac{\gamma(\tau_g+\beta)}{\delta\varepsilon\beta\sqrt{\alpha^2+\sigma^2}} > \sqrt{\frac{2}{\pi}}\right\}}\widetilde{D}(\alpha,\beta,\tau_h,\tau_g,\gamma) ,$$

where $\bar{D}$ does not depend on ε. It is easy to see that when $\varepsilon = 0$, then $\gamma = 0$. Otherwise $\tau^* = \infty$ and $\widetilde{D} = -\infty$ which implies that the maximum of D over γ is achieved at $\gamma = 0$. Therefore at $\varepsilon = 0$, we get

$$D = \bar{D} = \frac{\delta\beta}{2(\tau_g+\beta)}(\alpha^2+\sigma^2) - \frac{\alpha}{2\tau_h}\beta^2 - \frac{\alpha\tau_h}{2} + \frac{\beta\tau_g}{2} .$$

The stationary point is given by $(\tau_g+\beta)^2 = \delta(\alpha^2+\sigma^2)$, $\tau_h = \beta$ and $\delta\alpha = \tau_g + \beta$, $\tau_g = \alpha$ (derivative with respect to β). Putting things together we have

$$\alpha^2 = \frac{\sigma^2}{\delta-1}, \quad \tau_g = \alpha = \frac{\sigma}{\sqrt{\delta-1}}, \quad \tau_h = \beta = \sigma\sqrt{\delta-1}, \quad \gamma = 0 . \tag{A.3}$$

We next study the behavior of the convex-concave minimax problem 6.22 at infinitesimal ε. Rewriting the expressions for $\bar{D}$ and $\widetilde{D}$, we have

$$\begin{aligned}\bar{D} &= \frac{\delta\beta}{2(\tau_g+\beta)}\left(\alpha^2+\sigma^2\right) - \frac{\alpha}{2\tau_h}(\gamma^2+\beta^2) + \gamma\sqrt{\frac{\alpha^2\beta^2}{\tau_h^2}+V^2} - \frac{\alpha\tau_h}{2} + \frac{\beta\tau_g}{2}\,,\\ \widetilde{D} &= \frac{\delta\beta^2(\alpha^2+\sigma^2)}{2\tau_g(\tau_g+\beta)}\left(\operatorname{erf}\left(\frac{\tau^*}{\sqrt{2}}\right) - \frac{\gamma(\tau_g+\beta)}{\delta\varepsilon\beta\sqrt{\alpha^2+\sigma^2}}\tau^*\right).\end{aligned} \tag{A.4}$$

Let $\gamma_0 := \sqrt{\frac{2}{\pi}}\frac{\delta\varepsilon\beta\sqrt{\alpha^2+\sigma^2}}{\tau_g+\beta}$. If $\gamma \geq \gamma_0$, then D is a quadratic function of γ with the peak location at

$$\gamma_1 := \sqrt{\beta^2+\frac{\tau_h^2}{\alpha^2}V^2} - \frac{\tau_h\beta\sqrt{\alpha^2+\sigma^2}}{2\alpha\varepsilon\tau_g}\tau_*\,.$$

If $\gamma < \gamma_0$, then $D = \bar{D}$ is quadratic in γ with the peak location at

$$\gamma_2 := \sqrt{\beta^2+\frac{\tau_h^2}{\alpha^2}V^2}\,.$$

Therefore, to find the optimal γ we need to consider three different cases, giving us

$$\gamma_* = \begin{cases}\gamma_1 & \text{if } \gamma_0 \leq \gamma_1 \leq \gamma_2\,,\\ \gamma_0 & \text{if } \gamma_1 \leq \gamma_0 \leq \gamma_2\,,\\ \gamma_2 & \text{if } \gamma_1 \leq \gamma_2 \leq \gamma_0\,.\end{cases} \tag{A.5}$$

As $\varepsilon \to 0$, we have $\gamma_0 \to 0$. However, using (A.3) we get $\gamma_2 \to \sqrt{\sigma^2(\delta-1)+(\delta-1)^2V^2} > 0$. By continuity, at infinitesimal ε we get $\gamma_0 < \gamma_2$. Hence, in (A.5) only the first two cases may happen. Suppose that the first case occurs. Then, $0 \leq \gamma_0 \leq \gamma_1$ and by definition of γ_1 we obtain that $\tau_* = O(\varepsilon)$. Invoking the characterization equation of τ_* as per (6.24), we get

$$\frac{\gamma_*(\tau_g+\beta)}{\delta\varepsilon\beta\sqrt{\alpha^2+\sigma^2}} = \sqrt{\frac{2}{\pi}} + O(\varepsilon)\,, \quad \tau_* = O(\varepsilon)\,. \tag{A.6}$$

If the second case in (A.5) happens, we have $\gamma_* = \gamma_0 = \sqrt{\frac{2}{\pi}}\frac{\delta\varepsilon\beta\sqrt{\alpha^2+\sigma^2}}{\tau_g+\beta}$ and $\tau_* = 0$. So this case is subsumed in (A.6) and henceforth we can proceed with (A.6).

By Taylor expansion of the erf function we have

$$\operatorname{erf}\left(\frac{\tau^*}{\sqrt{2}}\right) = \sqrt{\frac{2}{\pi}}\tau_* + O(\tau_*^3)\,,$$

which implies that $\widetilde{D} = O(\tau_*^3) = O(\varepsilon^3)$. Separating $O(\varepsilon^2)$ terms from the lower order terms we get

$$D(\alpha,\beta,\tau_g,\tau_h) = D_0(\alpha,\beta,\tau_g,\tau_h) + \varepsilon D_1(\alpha,\beta,\tau_g,\tau_h) + O(\varepsilon^2)\,, \tag{A.7}$$

$$D_0(\alpha,\beta,\tau_g,\tau_h) = \frac{\delta\beta}{2(\tau_g+\beta)}\left(\alpha^2+\sigma^2\right) - \frac{\alpha}{2\tau_h}\beta^2 - \frac{\alpha\tau_h}{2} + \frac{\beta\tau_g}{2}\,,$$

$$D_1(\alpha,\beta,\tau_g,\tau_h) = \sqrt{\frac{2}{\pi}}\frac{\delta\beta\sqrt{\alpha^2+\sigma^2}}{\tau_g+\beta}\sqrt{\frac{\alpha^2\beta^2}{\tau_h^2}+V^2}\,. \tag{A.8}$$

Letting $\boldsymbol{x} = (\alpha, \beta, \tau_g, \tau_h)$, we then have

$$\nabla D(\boldsymbol{x}) = \nabla D_0(\boldsymbol{x}) + \varepsilon \nabla D_1(\boldsymbol{x}) + O(\varepsilon^2)\,. \tag{A.9}$$

To get the stationary points, we need to solve for $\nabla D(\boldsymbol{x}) = 0$. However, to find the solution up to $O(\varepsilon)$ term we can instead solve for $\nabla D_0(\boldsymbol{x}) + \varepsilon \nabla D_1(\boldsymbol{x}) = 0$. To see why, suppose that $\nabla D(\boldsymbol{x}_*) = 0$ and write $\boldsymbol{x}_* = \boldsymbol{x}_0 + \varepsilon \boldsymbol{x}_1 + O(\varepsilon^2)$. Hence,

$$\begin{aligned} \mathbf{0} = \nabla D(\boldsymbol{x}_*) &= \nabla D_0(\boldsymbol{x}_*) + \varepsilon \nabla D_1(\boldsymbol{x}_*) + O(\varepsilon^2) \\ &= \nabla D_0(\boldsymbol{x}_0) + \varepsilon(\nabla^2 D_0(\boldsymbol{x}_0)\boldsymbol{x}_1 + \nabla D_1(\boldsymbol{x}_0)) + O(\varepsilon^2)\,. \end{aligned} \tag{A.10}$$

This implies that $\boldsymbol{x}_0$ and $\boldsymbol{x}_1$ should satisfy

$$\nabla D_0(\boldsymbol{x}_0) = 0 \quad \text{and} \quad \nabla^2 D_0(\boldsymbol{x}_0)\boldsymbol{x}_1 + \nabla D_1(\boldsymbol{x}_0) = 0\,. \tag{A.11}$$

Likewise, let $\tilde{\boldsymbol{x}}_*$ be the solution of $\nabla D_0(\boldsymbol{x}) + \varepsilon \nabla D_1(\boldsymbol{x}) = 0$ and write $\tilde{\boldsymbol{x}}_* = \tilde{\boldsymbol{x}}_0 + \varepsilon \tilde{\boldsymbol{x}}_1 + O(\varepsilon^2)$. Then following similar arguments, we get

$$\nabla D_0(\tilde{\boldsymbol{x}}_0) = 0 \quad \text{and} \quad \nabla^2 D_0(\tilde{\boldsymbol{x}}_0)\tilde{\boldsymbol{x}}_1 + \nabla D_1(\tilde{\boldsymbol{x}}_0) = 0\,. \tag{A.12}$$

Comparing equations (A.11) and (A.12), we see that $\boldsymbol{x}_0 = \tilde{\boldsymbol{x}}_0$ and $\boldsymbol{x}_1 = \tilde{\boldsymbol{x}}_1$. Therefore, to find the stationary point $\boldsymbol{x}_*$ up to $O(\varepsilon)$ terms, we can neglect $O(\varepsilon^2)$ term in (A.9).

We proceed by computing the stationary points of $D_0(\boldsymbol{x}) + \varepsilon D_1(\boldsymbol{x})$. Writing KKT conditions with respect to α, β, τ_g, τ_h we have

$$\begin{aligned} &\frac{\sqrt{\frac{2}{\pi}}\alpha\beta\delta\varepsilon\sqrt{\frac{\alpha^2\beta^2}{\tau_h^2}+V^2}}{\sqrt{\alpha^2+\sigma^2}(\beta+\tau_g)} + \frac{\sqrt{\frac{2}{\pi}}\alpha\beta^3\delta\varepsilon\sqrt{\alpha^2+\sigma^2}}{\tau_h^2(\beta+\tau_g)\sqrt{\frac{\alpha^2\beta^2}{\tau_h^2}+V^2}} + \frac{\alpha\beta\delta}{\beta+\tau_g} - \frac{\beta^2}{2\tau_h} - \frac{\tau_h}{2} = 0\,, \\ &\frac{\sqrt{2}\alpha^2\beta^2\delta\varepsilon\sqrt{\alpha^2+\sigma^2}}{\sqrt{\pi}\tau_h^2(\beta+\tau_g)\sqrt{\frac{\alpha^2\beta^2}{\tau_h^2}+V^2}} - \frac{\sqrt{\frac{2}{\pi}}\beta\delta\varepsilon\sqrt{\alpha^2+\sigma^2}\sqrt{\frac{\alpha^2\beta^2}{\tau_h^2}+V^2}}{(\beta+\tau_g)^2} \\ &\qquad + \frac{\sqrt{\frac{2}{\pi}}\delta\varepsilon\sqrt{\alpha^2+\sigma^2}\sqrt{\frac{\alpha^2\beta^2}{\tau_h^2}+V^2}}{\beta+\tau_g} - \frac{\beta\delta\left(\alpha^2+\sigma^2\right)}{2(\beta+\tau_g)^2} + \frac{\delta\left(\alpha^2+\sigma^2\right)}{2(\beta+\tau_g)} - \frac{\alpha\beta}{\tau_h} + \frac{\tau_g}{2} = 0\,, \\ &- \frac{\sqrt{\frac{2}{\pi}}\beta\delta\varepsilon\sqrt{\alpha^2+\sigma^2}\sqrt{\frac{\alpha^2\beta^2}{\tau_h^2}+V^2}}{(\beta+\tau_g)^2} - \frac{\beta\delta\left(\alpha^2+\sigma^2\right)}{2(\beta+\tau_g)^2} + \frac{\beta}{2} = 0\,, \\ &- \frac{\sqrt{\frac{2}{\pi}}\alpha^2\beta^3\delta\varepsilon\sqrt{\alpha^2+\sigma^2}}{\tau_h^3(\beta+\tau_g)\sqrt{\frac{\alpha^2\beta^2}{\tau_h^2}+V^2}} + \frac{\alpha\beta^2}{2\tau_h^2} - \frac{\alpha}{2} = 0\,. \end{aligned}$$

Second equation can be simplified using other equations as

$$\begin{aligned} &\frac{\alpha\beta}{2\tau_h} - \frac{\alpha\tau_h}{2\beta} - \frac{\beta}{2} - \frac{\delta(\alpha^2+\sigma^2)}{2(\beta+\tau_g)} + \frac{\beta+\tau_g}{2} + \frac{\delta\left(\alpha^2+\sigma^2\right)}{2(\beta+\tau_g)} - \frac{\alpha\beta}{\tau_h} + \frac{\tau_g}{2} = 0\,. \\ &\to -\frac{\alpha\beta}{2\tau_h} + \tau_g - \frac{\alpha\tau_h}{2\beta} = 0\,. \end{aligned}$$

The first equation also simplifies to

$$
\begin{aligned}
&-\frac{\alpha\beta\delta}{2(\beta+\tau_g)}+\frac{\alpha(\beta+\tau_g)\beta}{2(\alpha^2+\sigma^2)}+\frac{\beta^2}{2\tau_h}-\frac{\tau_h}{2}+\frac{\alpha\beta\delta}{\beta+\tau_g}-\frac{\beta^2}{2\tau_h}-\frac{\tau_h}{2}=0\,,\\
&\to\frac{\alpha\beta\delta}{2(\beta+\tau_g)}+\frac{\alpha(\beta+\tau_g)\beta}{2(\alpha^2+\sigma^2)}-\tau_h=0\,.
\end{aligned}
$$

Define $\eta=\beta/\tau_h>1$ (since $\varepsilon>0$). The second equation gives $\tau_g=\alpha/2(\eta+1/\eta)$. While this becomes useful in finding optimal τ_g it does not matter with our goal of finding α as everywhere τ_g appears in form $\beta+\tau_g$. The first equation though gives

$$
\frac{\delta}{\beta+\tau_g}+\frac{\beta+\tau_g}{(\alpha^2+\sigma^2)}=\frac{2}{\alpha\eta}\,. \tag{A.13}
$$

The third equation gives

$$
2\sqrt{\tfrac{2}{\pi}}\delta\varepsilon\sqrt{\alpha^2+\sigma^2}\sqrt{\alpha^2\eta^2+V^2}+\delta(\alpha^2+\sigma^2)=(\beta+\tau_g)^2\,. \tag{A.14}
$$

The fourth equation gives

$$
\frac{2\sqrt{\frac{2}{\pi}}\alpha\eta^3\delta\varepsilon\sqrt{\alpha^2+\sigma^2}}{(\eta^2-1)\sqrt{\alpha^2\eta^2+V^2}}=\beta+\tau_g\,. \tag{A.15}
$$

Continuing from (A.13) we get

$$
\sqrt{\frac{2}{\pi}}\varepsilon\sqrt{\alpha^2\eta^2+V^2}+\sqrt{\alpha^2+\sigma^2}=(\alpha^2+\sigma^2)\frac{2\sqrt{\frac{2}{\pi}}\eta^2\varepsilon}{(\eta^2-1)\sqrt{\alpha^2\eta^2+V^2}}\,. \tag{A.16}
$$

Simplifying this equation,

$$
\sqrt{\frac{2}{\pi}}\varepsilon\sqrt{\alpha^2\eta^2+V^2}\,(\eta^2-1)+\sqrt{\alpha^2+\sigma^2}(\eta^2-1)=(\alpha^2+\sigma^2)\frac{2\sqrt{\frac{2}{\pi}}\eta^2\varepsilon}{\sqrt{\alpha^2\eta^2+V^2}}\,. \tag{A.17}
$$

We now proceed by taking derivatives of both equations implicitly with respect to ε and evaluate them at

$$
\tau_g^*=\alpha^*=\frac{\sigma}{\sqrt{\delta-1}},\quad \tau_h^*=\beta^*=\sigma\sqrt{\delta-1}\,,\quad \gamma^*=0,\quad \text{and}\quad \varepsilon=0\,.
$$

Note that the derivative of the first equation yields

$$
\frac{\mathrm{d}}{\mathrm{d}\varepsilon}\left(\sqrt{\alpha^2+\sigma^2}(\eta^2-1)+\varepsilon\left(\sqrt{\tfrac{2}{\pi}}\sqrt{\alpha^2\eta^2+V^2}(\eta^2-1)-(\alpha^2+\sigma^2)\frac{2\sqrt{\frac{2}{\pi}}\eta^2}{\sqrt{\alpha^2\eta^2+V^2}}\right)\right)=0\,.
$$

Thus

$$
\begin{aligned}
&\frac{\mathrm{d}}{\mathrm{d}\varepsilon}\left(\sqrt{\alpha^2+\sigma^2}(\eta^2-1)\right)+\left(\sqrt{\tfrac{2}{\pi}}\sqrt{\alpha^2\eta^2+V^2}(\eta^2-1)-(\alpha^2+\sigma^2)\frac{2\sqrt{\frac{2}{\pi}}\eta^2}{\sqrt{\alpha^2\eta^2+V^2}}\right)+\\
&\varepsilon\frac{\mathrm{d}}{\mathrm{d}\varepsilon}\left(\sqrt{\tfrac{2}{\pi}}\sqrt{\alpha^2\eta^2+V^2}(\eta^2-1)-(\alpha^2+\sigma^2)\frac{2\sqrt{\frac{2}{\pi}}\eta^2}{\sqrt{\alpha^2\eta^2+V^2}}\right)=0\,.
\end{aligned}
$$

Setting $\varepsilon = 0$ in the above yields

$$\frac{\mathrm{d}}{\mathrm{d}\varepsilon}\left(\sqrt{\alpha^2+\sigma^2}(\eta^2-1)\right)+\left(\sqrt{\tfrac{2}{\pi}}\sqrt{\alpha^2\eta^2+V^2}(\eta^2-1)-(\alpha^2+\sigma^2)\frac{2\sqrt{\frac{2}{\pi}}\eta^2}{\sqrt{\alpha^2\eta^2+V^2}}\right)=0\,.$$

Thus

$$\frac{\alpha^*}{\sqrt{(\alpha^*)^2+\sigma^2}}\frac{\mathrm{d}\alpha}{\mathrm{d}\varepsilon}(\eta_*^2-1)+2\eta_*\sqrt{\alpha_*^2+\sigma^2}\frac{\mathrm{d}\eta}{\mathrm{d}\varepsilon}=(\alpha_*^2+\sigma^2)\frac{2\sqrt{\frac{2}{\pi}}\eta_*^2}{\sqrt{\alpha_*^2\eta_*^2+V^2}}-\sqrt{\tfrac{2}{\pi}}\sqrt{\alpha_*^2\eta_*^2+V^2}(\eta_*^2-1)\,.$$

Setting $\eta_* = 1$ this simplifies to

$$\frac{\mathrm{d}\eta}{\mathrm{d}\varepsilon}=\sqrt{(\alpha^*)^2+\sigma^2}\frac{\sqrt{\frac{2}{\pi}}}{\sqrt{(\alpha^*)^2+V^2}}=\sigma\sqrt{\frac{2\delta}{\pi}}\frac{1}{\sqrt{\sigma^2+V^2(\delta-1)}}\,.$$

In addition, from (A.13)

$$\left(-\frac{\delta}{(\beta_*+\tau_{g*})^2}+\frac{1}{\alpha_*^2+\sigma^2}\right)\frac{\mathrm{d}}{\mathrm{d}\varepsilon}(\beta+\tau_g)-\frac{\beta_*+\tau_{g*}}{(\alpha_*^2+\sigma^2)^2}2\alpha_*\frac{\mathrm{d}\alpha}{\mathrm{d}\varepsilon}=-\frac{2}{\alpha_*\eta_*^2}\frac{\mathrm{d}\eta}{\mathrm{d}\varepsilon}-\frac{2}{\alpha_*^2\eta_*}\frac{\mathrm{d}\alpha}{\mathrm{d}\varepsilon}\,.$$

Plugging in for $\beta_*, \tau_{g*}, \alpha_*$ the coefficient of $\frac{\mathrm{d}}{\mathrm{d}\varepsilon}(\beta+\tau_g)$ vanishes and we arrive at

$$\frac{\frac{\sigma\delta}{\sqrt{\delta-1}}}{\left(\frac{\sigma^2\delta}{\delta-1}\right)^2}\frac{\sigma}{\sqrt{\delta-1}}\frac{\mathrm{d}\alpha}{\mathrm{d}\varepsilon}=\frac{\sqrt{\delta-1}}{\sigma}\frac{\mathrm{d}\eta}{\mathrm{d}\varepsilon}+\frac{\delta-1}{\sigma^2}\frac{\mathrm{d}\alpha}{\mathrm{d}\varepsilon}\,.$$

Rearranging the terms, we obtain

$$\frac{\mathrm{d}\alpha}{\mathrm{d}\varepsilon}=-\frac{\sigma\delta}{(\delta-1)^{3/2}}\frac{\mathrm{d}\eta}{\mathrm{d}\varepsilon}=-\sqrt{\frac{2}{\pi}}\sigma^2\left(\frac{\delta}{\delta-1}\right)^{3/2}\frac{1}{\sqrt{\sigma^2+V^2(\delta-1)}}\,.$$

Now, invoking the definition of statistical risk we have

$$\begin{aligned}\mathsf{SR}(\widehat{\boldsymbol{\theta}}^{\varepsilon})&=\mathsf{SR}(\widehat{\boldsymbol{\theta}}^{0})+\frac{\mathrm{d}}{\mathrm{d}\varepsilon}\mathsf{SR}(\widehat{\boldsymbol{\theta}}^{\varepsilon})\Big|_{\varepsilon=0}\varepsilon+O(\varepsilon^2)\\&=\sigma^2+\alpha_*^2+2\alpha_*\frac{\mathrm{d}\alpha}{\mathrm{d}\varepsilon}\Big|_{\varepsilon=0}+O(\varepsilon^2)\\&=\frac{\sigma^2\delta}{\delta-1}-\sqrt{\frac{2}{\pi}}\frac{\sigma^3\delta^{3/2}}{(\delta-1)^2}\cdot\frac{1}{\sqrt{\sigma^2+V^2(\delta-1)}}+O(\varepsilon^2)\,.\end{aligned}\tag{A.18}$$

The proof is complete. □

B Proofs that the minimization and maximization primal problems can be restricted to a compact set

In this section we demonstrate how the minimization and maximization problems can be restricted to compacts sets. We start with the restriction on $\boldsymbol{z}$. To this aim recall that that one of the main goals

of Theorem 3.3 is to characterize the distance of the optimal solution $\widehat{\boldsymbol{\theta}}^{\varepsilon}$ to $\boldsymbol{\theta}_0$ i.e. $\frac{\|\widehat{\boldsymbol{\theta}}^{\varepsilon}-\boldsymbol{\theta}_0\|_{\ell_2}}{\sqrt{p}} = \|\widehat{\boldsymbol{z}}^{\varepsilon}\|_{\ell_2}$ asymptotically and in particular to show $\|\boldsymbol{z}\|_{\ell_2} \to \alpha_*$ as $n \to \infty$, in probability, for some α_* to be determined. Now define the set $\mathcal{S}_{\boldsymbol{z}} = \{\boldsymbol{z} | \ \ \|\boldsymbol{z}\|_{\ell_2} \leq K_\alpha\}$ with $K_\alpha = \alpha_* + \zeta$ for a constant $\zeta > 0$ and consider the optimization problem

$$\min_{\boldsymbol{z}\in\mathcal{S}_{\boldsymbol{z}},\boldsymbol{v}\in\mathbb{R}^n} \max_{\boldsymbol{u}\in\mathbb{R}^n} \frac{1}{\sqrt{p}}\left(\boldsymbol{u}^T\boldsymbol{X}\boldsymbol{z} - \boldsymbol{u}^T\boldsymbol{\omega} + \boldsymbol{u}^T\boldsymbol{v}\right) + \ell(\boldsymbol{v};\boldsymbol{z}) \tag{B.1}$$

with $\boldsymbol{\omega} = \boldsymbol{w}/\sqrt{p}$. Based on the CGMT framework this optimization problem is equivalent to (6.6) in an asymptotic fashion in the sense that if the Euclidean norm of the optimum solution to the above converges asymptotically to a value α_* in probability as $n \to +\infty$ then $\|\widehat{\boldsymbol{z}}\|_{\ell_2}$ also converges to the same value ($\|\widehat{\boldsymbol{z}}\| \to \alpha_*$) in probability. See [TAH18, Theorem A.1] for a formal argument.

The optimization problem above is still not in a form where CGMT can be applied as there are no compact restriction on $\boldsymbol{u}$. This is the subject of the next lemma.

Lemma B.1. *The optimal solution $\boldsymbol{u}^*$ of (B.1) satisfies $\|\boldsymbol{u}^*\|_{\ell_2} \leq K_\beta$ for a sufficiently large constant $K_\beta > 0$ with probability at least $1 - 2e^{-cn}$.*

Proof. Writing the KKT conditions for (B.1) we have

$$\begin{aligned}
&\boldsymbol{X}\boldsymbol{z} - \frac{1}{\sqrt{p}}\boldsymbol{w} + \boldsymbol{v} = 0\\
&u_i = -\sqrt{p}[\nabla_{\boldsymbol{v}}\ell(\boldsymbol{v};\boldsymbol{z})]_i = -\frac{1}{\sqrt{p}}\left(v_i + \frac{\varepsilon}{p}\cdot\operatorname{sgn}(v_i)\,\|\boldsymbol{\theta}_0 + \sqrt{p}\boldsymbol{z}\|_{\ell_2}\right)
\end{aligned}$$

From the first equation we have that $\boldsymbol{v} = \frac{\boldsymbol{w}}{\sqrt{p}} - \boldsymbol{X}\boldsymbol{z}$. Thus,

$$\begin{aligned}
\|\boldsymbol{v}\|_{\ell_2} \leq& \frac{1}{\sqrt{p}}\|\boldsymbol{w}\|_{\ell_2} + \|\boldsymbol{X}\boldsymbol{z}\|_{\ell_2}\\
\leq& \frac{1}{\sqrt{p}}\|\boldsymbol{w}\|_{\ell_2} + \|\boldsymbol{X}\|\,\|\boldsymbol{z}\|_{\ell_2}\\
\overset{(a)}{\leq}& C\sqrt{n}\sigma + C\left(\sqrt{p}+\sqrt{n}\right)\|\boldsymbol{z}\|_{\ell_2}\\
\overset{(b)}{\leq}& C\sqrt{n}\sigma + C\left(\sqrt{p}+\sqrt{n}\right)K_\alpha
\end{aligned}$$

holds with probability at least $1 - 2e^{-cn}$. Here, (a) follows from well known bounds on the Euclidean norm of a Gaussian vector and the spectral norm of a Gaussian matrix and (b) follows from the fact that $\|\boldsymbol{z}\|_{\ell_2} \leq K_\alpha$. We thus have $\|\boldsymbol{v}\|_{\ell_2} \leq C_2\left(\sqrt{p}+\sqrt{n}\right)$, with high probability. Now using the second equation we have

$$\begin{aligned}
\|\boldsymbol{u}\|_{\ell_2} \leq& \frac{\|\boldsymbol{v}\|_{\ell_2}}{\sqrt{p}} + \frac{\varepsilon\sqrt{\delta}}{\sqrt{p}}\|\boldsymbol{\theta}_0 + \sqrt{p}\boldsymbol{z}\|_{\ell_2}\\
\leq& C\sigma\sqrt{\delta} + C(1+\sqrt{\delta})K_\alpha + \frac{\varepsilon\sqrt{\delta}}{\sqrt{p}}\|\boldsymbol{\theta}_0\|_{\ell_2} + \varepsilon\sqrt{\delta}\,\|\boldsymbol{z}\|_{\ell_2}\\
\leq& C\sigma\sqrt{\delta} + C(1+\sqrt{\delta})K_\alpha + \varepsilon\sqrt{\delta}\widetilde{C} + \varepsilon\sqrt{\delta}K_\alpha\\
\leq& K_\beta\,,
\end{aligned}$$

for some bounded constant K_β. In the penultimum step we used the fact that $\frac{\|\boldsymbol{\theta}_0\|_{\ell_2}}{\sqrt{p}}$ is bounded and $\|\boldsymbol{z}\|_{\ell_2} \leq K_\alpha$. □

C Proofs for scalarization of Auxilary Optimization (AO)

C.1 Proof of Lemma 6.1

We restate the lemma for the convenience of the reader.

Lemma C.1. *[Restatement of Lemma 6.1] The conjugate of*

$$\ell(\boldsymbol{v};\boldsymbol{z}) \coloneqq \frac{1}{2p}\left(\|\boldsymbol{v}\|_{\ell_2}^2 + 2\frac{\varepsilon}{\sqrt{p}}\|\boldsymbol{v}\|_{\ell_1}\|\boldsymbol{\theta}_0+\sqrt{p}\boldsymbol{z}\|_{\ell_2} + \frac{\varepsilon^2}{p}\|\boldsymbol{\theta}_0+\sqrt{p}\boldsymbol{z}\|_{\ell_2}^2\right)$$

with respect to the variable $\boldsymbol{z}$ is given by

$$\widetilde{\ell}(\boldsymbol{v};\boldsymbol{q}) \coloneqq \sup_{\boldsymbol{z}} \boldsymbol{q}^T\boldsymbol{z} - \ell(\boldsymbol{v};\boldsymbol{z}) = -\frac{1}{\sqrt{p}}\boldsymbol{q}^T\boldsymbol{\theta}_0 + \frac{1}{2\delta p^2}\left(\frac{p}{\varepsilon}\|\boldsymbol{q}\|_{\ell_2} - \|\boldsymbol{v}\|_{\ell_1}\right)_+^2 - \frac{1}{2p}\|\boldsymbol{v}\|_{\ell_2}^2 .$$

Proof. We begin by calculating the conjugate of a slightly simpler function

$$\bar{\ell}(\boldsymbol{v};\boldsymbol{\theta}) \coloneqq \frac{1}{2p}\sum_{i=1}^n\left(|v_i| + \frac{\varepsilon}{\sqrt{p}}\|\boldsymbol{\theta}\|_{\ell_2}\right)^2 .$$

We have

$$\begin{aligned}
\bar{\ell}^*(\boldsymbol{v};\boldsymbol{q}) &= \sup_{\boldsymbol{\theta}} \boldsymbol{q}^T\boldsymbol{\theta} - \bar{\ell}(\boldsymbol{v};\boldsymbol{\theta}) \\
&= \sup_{\boldsymbol{\theta}} \boldsymbol{q}^T\boldsymbol{\theta} - \frac{1}{2p}\sum_{i=1}^n\left(|v_i| + \frac{\varepsilon}{\sqrt{p}}\|\boldsymbol{\theta}\|_{\ell_2}\right)^2 \\
&= \sup_{\boldsymbol{\theta}}\sup_{\xi\geq 0} \boldsymbol{q}^T\boldsymbol{\theta} - \frac{1}{2p}\left(\|\boldsymbol{v}\|_{\ell_2}^2 + \frac{2\varepsilon}{\sqrt{p}}\|\boldsymbol{v}\|_{\ell_1}\left(\frac{\|\boldsymbol{\theta}\|_{\ell_2}^2}{2\xi} + \frac{\xi}{2}\right) + \delta\varepsilon^2\|\boldsymbol{\theta}\|_{\ell_2}^2\right) \\
&= \sup_{\xi\geq 0}\sup_{\boldsymbol{\theta}} \boldsymbol{q}^T\boldsymbol{\theta} - \frac{1}{2p}\left(\|\boldsymbol{v}\|_{\ell_2}^2 + \frac{2\varepsilon}{\sqrt{p}}\|\boldsymbol{v}\|_{\ell_1}\left(\frac{\|\boldsymbol{\theta}\|_{\ell_2}^2}{2\xi} + \frac{\xi}{2}\right) + \delta\varepsilon^2\|\boldsymbol{\theta}\|_{\ell_2}^2\right)
\end{aligned}$$

Setting derivative w.r.t $\boldsymbol{\theta}$ to zero, we get

$$\boldsymbol{q} - \frac{\varepsilon\|\boldsymbol{v}\|_{\ell_1}}{p^{3/2}\xi}\boldsymbol{\theta} - \frac{\delta\varepsilon^2}{p}\boldsymbol{\theta} = 0 \quad \Rightarrow \quad \boldsymbol{\theta} = \left(\frac{\varepsilon\|\boldsymbol{v}\|_{\ell_1}}{p^{3/2}\xi} + \frac{\delta\varepsilon^2}{p}\right)^{-1}\boldsymbol{q}$$

Setting the derivative with respect to ξ to zero we conclude that $\xi = \|\boldsymbol{\theta}\|_{\ell_2}$. Plugging the latter into above we conclude that

$$\boldsymbol{\theta} = \left(\frac{\varepsilon\|\boldsymbol{v}\|_{\ell_1}}{p^{3/2}\|\boldsymbol{\theta}\|_{\ell_2}} + \frac{\delta\varepsilon^2}{p}\right)^{-1}\boldsymbol{q}$$

Taking the Euclidean norm from both sides we conclude that

$$\|\boldsymbol{\theta}\|_{\ell_2}\left(\frac{\varepsilon\|\boldsymbol{v}\|_{\ell_1}}{p^{3/2}\|\boldsymbol{\theta}\|_{\ell_2}} + \frac{\delta\varepsilon^2}{p}\right) = \|\boldsymbol{q}\|_{\ell_2} \quad \Rightarrow \quad \|\boldsymbol{\theta}\|_{\ell_2} = \frac{\|\boldsymbol{q}\|_{\ell_2} - \frac{\varepsilon\|\boldsymbol{v}\|_{\ell_1}}{p^{\frac{3}{2}}}}{\frac{\delta\varepsilon^2}{p}} = \frac{p}{\delta\varepsilon^2}\|\boldsymbol{q}\|_{\ell_2} - \frac{1}{\delta\varepsilon\sqrt{p}}\|\boldsymbol{v}\|_{\ell_1}$$

If $\frac{p}{\delta\varepsilon^2}\|\boldsymbol{q}\|_{\ell_2} - \frac{1}{\delta\varepsilon\sqrt{p}}\|\boldsymbol{v}\|_{\ell_1} < 0$ then it is easy to verify that the objective is less than $-\frac{1}{2p}\|\boldsymbol{v}\|_{\ell_2}^2$ and hence the optimal is given by $\boldsymbol{\theta} = 0$.

Thus

$$\boldsymbol{\theta} = \left(\frac{p}{\delta\varepsilon^2}\|\boldsymbol{q}\|_{\ell_2} - \frac{1}{\delta\varepsilon\sqrt{p}}\|\boldsymbol{v}\|_{\ell_1}\right)\frac{\boldsymbol{q}}{\|\boldsymbol{q}\|_{\ell_2}} = \left(\frac{p}{\delta\varepsilon^2} - \frac{1}{\delta\varepsilon\sqrt{p}}\frac{\|\boldsymbol{v}\|_{\ell_1}}{\|\boldsymbol{q}\|_{\ell_2}}\right)\boldsymbol{q}$$

Substituting for $\boldsymbol{\theta}$ we have

$$\begin{aligned}
\bar{\ell}^*(\boldsymbol{v};\boldsymbol{q}) =& \left(\frac{p}{\delta\varepsilon^2} - \frac{1}{\delta\varepsilon\sqrt{p}}\frac{\|\boldsymbol{v}\|_{\ell_1}}{\|\boldsymbol{q}\|_{\ell_2}}\right)\|\boldsymbol{q}\|_{\ell_2}^2 \\
&- \frac{1}{2p}\left(\|\boldsymbol{v}\|_{\ell_2}^2 + \frac{2\varepsilon}{\sqrt{p}}\|\boldsymbol{v}\|_{\ell_1}\left(\frac{p}{\delta\varepsilon^2}\|\boldsymbol{q}\|_{\ell_2} - \frac{1}{\delta\varepsilon\sqrt{p}}\|\boldsymbol{v}\|_{\ell_1}\right) + \delta\varepsilon^2\left(\frac{p}{\delta\varepsilon^2}\|\boldsymbol{q}\|_{\ell_2} - \frac{1}{\delta\varepsilon\sqrt{p}}\|\boldsymbol{v}\|_{\ell_1}\right)^2\right) \\
=& \frac{p}{2\delta\varepsilon^2}\|\boldsymbol{q}\|_{\ell_2}^2 - \frac{1}{\delta\varepsilon\sqrt{p}}\|\boldsymbol{v}\|_{\ell_1}\|\boldsymbol{q}\|_{\ell_2} + \frac{1}{2\delta p^2}\|\boldsymbol{v}\|_{\ell_1}^2 - \frac{1}{2p}\|\boldsymbol{v}\|_{\ell_2}^2 \\
=& \frac{1}{2\delta p^2}\left(\frac{p^{\frac{3}{2}}}{\varepsilon}\|\boldsymbol{q}\|_{\ell_2} - \|\boldsymbol{v}\|_{\ell_1}\right)^2 - \frac{1}{2p}\|\boldsymbol{v}\|_{\ell_2}^2
\end{aligned}$$

if $\|\boldsymbol{v}\|_{\ell_1} \le \frac{p^{\frac{3}{2}}}{\varepsilon}\|\boldsymbol{q}\|_{\ell_2}$. Otherwise,

$$\bar{\ell}^*(\boldsymbol{v};\boldsymbol{q}) = -\frac{1}{2p}\|\boldsymbol{v}\|_{\ell_2}^2 .$$

We can put the two cases together using the notation $z_+ = \max(z, 0)$.

$$\bar{\ell}^*(\boldsymbol{v};\boldsymbol{q}) = \frac{1}{2\delta p^2}\left(\frac{p^{\frac{3}{2}}}{\varepsilon}\|\boldsymbol{q}\|_{\ell_2} - \|\boldsymbol{v}\|_{\ell_1}\right)_+^2 - \frac{1}{2p}\|\boldsymbol{v}\|_{\ell_2}^2 .$$

Now to calculate the conjugate of $\ell(\boldsymbol{v};\boldsymbol{z})$ note that

$$\ell(\boldsymbol{v};\boldsymbol{z}) = \bar{\ell}\left(\boldsymbol{v};\boldsymbol{\theta}_0 + \sqrt{p}\boldsymbol{z}\right)$$

To continue note that if we have $f(\boldsymbol{x}) = g(\boldsymbol{A}\boldsymbol{x} + \boldsymbol{x}_0)$ the conjugate is given by

$$f^*(\boldsymbol{y}) = -\langle \boldsymbol{A}^{-1}\boldsymbol{x}_0, \boldsymbol{y}\rangle + g^*\left(\boldsymbol{A}^{-T}\boldsymbol{y}\right)$$

Thus using above with $\boldsymbol{x}_0 = \boldsymbol{\theta}_0$ and $\boldsymbol{A} = \sqrt{p}$ we arrive at

$$\begin{aligned}
\widetilde{\ell}(\boldsymbol{v};\boldsymbol{q}) =& -\frac{1}{\sqrt{p}}\langle\boldsymbol{\theta}_0, \boldsymbol{q}\rangle + \bar{\ell}^*\left(\boldsymbol{v}; \frac{1}{\sqrt{p}}\boldsymbol{q}\right) \\
=& -\frac{1}{\sqrt{p}}\boldsymbol{q}^T\boldsymbol{\theta}_0 + \frac{1}{2\delta p^2}\left(\frac{p}{\varepsilon}\|\boldsymbol{q}\|_{\ell_2} - \|\boldsymbol{v}\|_{\ell_1}\right)_+^2 - \frac{1}{2p}\|\boldsymbol{v}\|_{\ell_2}^2 ,
\end{aligned}$$

concluding the proof. □

C.2 Proof of Lemma 6.2

Lemma C.2 (Restatement of Lemma 6.2)**.** *The function*

$$f(\gamma,\beta,\tau_h) := \gamma^2 + \frac{\beta^2}{p}\|\boldsymbol{h}\|_{\ell_2}^2 - 2\frac{\gamma}{\sqrt{p}}\left\|\beta\boldsymbol{h} - \frac{\boldsymbol{\theta_0}}{\alpha}\right\|_{\ell_2}$$

is jointly convex in the parameters (γ,β,τ_h)*.*

Proof.

$$\gamma^2 + \frac{\beta^2}{p}\|\boldsymbol{h}\|_{\ell_2}^2 - 2\frac{\gamma}{\sqrt{p}}\left\|\beta\boldsymbol{h} - \frac{\boldsymbol{\theta_0}}{\alpha}\right\|_{\ell_2} = \gamma^2 + \frac{\beta^2}{p}\|\boldsymbol{h}\|_{\ell_2}^2 - 2\frac{\gamma}{\sqrt{p}}\sqrt{\beta^2\|\boldsymbol{h}\|_{\ell_2}^2 + \frac{1}{\alpha^2}\|\boldsymbol{\theta}_0\|_{\ell_2}^2 - \frac{2}{\alpha}\beta\boldsymbol{h}^T\boldsymbol{\theta}_0}$$

with the Hessian with respect to (γ,β) equal to

$$\begin{bmatrix} 2 & -\frac{1}{\sqrt{p}}\frac{2\beta\|\boldsymbol{h}\|_{\ell_2}^2 - \frac{2}{\alpha}\boldsymbol{h}^T\boldsymbol{\theta}_0}{\sqrt{\beta^2\|\boldsymbol{h}\|_{\ell_2}^2 + \frac{1}{\alpha^2}\|\boldsymbol{\theta}_0\|_{\ell_2}^2 - \frac{2}{\alpha}\beta\boldsymbol{h}^T\boldsymbol{\theta}_0}} \\ -\frac{1}{\sqrt{p}}\frac{2\beta\|\boldsymbol{h}\|_{\ell_2}^2 - \frac{2}{\alpha}\boldsymbol{h}^T\boldsymbol{\theta}_0}{\sqrt{\beta^2\|\boldsymbol{h}\|_{\ell_2}^2 + \frac{1}{\alpha^2}\|\boldsymbol{\theta}_0\|_{\ell_2}^2 - \frac{2}{\alpha}\beta\boldsymbol{h}^T\boldsymbol{\theta}_0}} & 2\frac{\|\boldsymbol{h}\|_{\ell_2}^2}{p} \end{bmatrix}$$

The determinant is equal to

$$\begin{aligned}\frac{4}{p}\left(\|\boldsymbol{h}\|_{\ell_2}^2 - \frac{\left(\beta\|\boldsymbol{h}\|_{\ell_2}^2 - \frac{\boldsymbol{h}^T\boldsymbol{\theta}_0}{\alpha}\right)^2}{\beta^2\|\boldsymbol{h}\|_{\ell_2}^2 + \frac{1}{\alpha^2}\|\boldsymbol{\theta}_0\|_{\ell_2}^2 - \frac{2}{\alpha}\beta\boldsymbol{h}^T\boldsymbol{\theta}_0}\right) \\ = \frac{4}{p\alpha^2}\frac{1}{\beta^2\|\boldsymbol{h}\|_{\ell_2}^2 + \frac{1}{\alpha^2}\|\boldsymbol{\theta}_0\|_{\ell_2}^2 - \frac{2}{\alpha}\beta\boldsymbol{h}^T\boldsymbol{\theta}_0}\left(\|\boldsymbol{h}\|_{\ell_2}^2\|\boldsymbol{\theta}_0\|_{\ell_2}^2 - \left(\boldsymbol{h}^T\boldsymbol{\theta}_0\right)^2\right) \\ \geq 0\end{aligned}$$

Thus

$$\frac{\alpha}{2}\gamma^2 + \frac{\alpha}{2}\frac{\beta^2}{p}\|\boldsymbol{h}\|_{\ell_2}^2 - \frac{\gamma}{\sqrt{p}}\|\alpha\beta\boldsymbol{h} - \boldsymbol{\theta}_0\|_{\ell_2} = \frac{\alpha}{2}\left(\gamma^2 + \frac{\beta^2}{p}\|\boldsymbol{h}\|_{\ell_2}^2 - 2\frac{\gamma}{\sqrt{p}}\left\|\beta\boldsymbol{h} - \frac{\boldsymbol{\theta_0}}{\alpha}\right\|_{\ell_2}\right)$$

is jointly convex in (γ,β). Therefore the perspective function

$$\tau_h\left(\frac{\alpha}{2}\left(\frac{\gamma}{\tau_h}\right)^2 + \frac{\alpha}{2}\frac{\beta^2}{p\tau_h^2}\|\boldsymbol{h}\|_{\ell_2}^2 - \frac{\gamma}{\tau_h\sqrt{p}}\left\|\alpha\frac{\beta}{\tau_h}\boldsymbol{h} - \boldsymbol{\theta}_0\right\|_{\ell_2}\right) = \frac{\alpha}{2\tau_h}\gamma^2 + \frac{\alpha\beta^2}{2p\tau_h}\|\boldsymbol{h}\|_{\ell_2}^2 - \frac{\gamma}{\sqrt{p}}\left\|\frac{\alpha\beta}{\tau_h}\boldsymbol{h} - \boldsymbol{\theta_0}\right\|_{\ell_2}$$

is jointly convex in (γ,β,τ_h). □

C.3 Proof of Lemma 6.3

We begin by stating and proving the following lemma.

Lemma C.3. *The value of the following problem (with $\lambda > 1$)*

$$\min_{\boldsymbol{v}\in\mathbb{R}^n} \quad \frac{\lambda}{2}\|\boldsymbol{x}-\boldsymbol{v}\|_{\ell_2}^2 - \frac{1}{2n}\left(\gamma - \|\boldsymbol{v}\|_{\ell_1}\right)_+^2$$

is given by

$$\min_{\tau\geq 0} \quad \frac{\lambda}{2}\|\boldsymbol{x}-\mathsf{ST}(\boldsymbol{x};\tau)\|_{\ell_2}^2 - \frac{1}{2n}\left(\gamma - \|\mathsf{ST}(\boldsymbol{x};\tau)\|_{\ell_1}\right)_+^2$$

where $\mathsf{ST}(\boldsymbol{x};\tau)$ is the soft-thresholding function.

Notably the lemma above transforms the first optimization (on vector $\boldsymbol{v}$) to an optimization over scalar τ.

Proof. We consider two case:
Case I: $\|\boldsymbol{x}\|_{\ell_1} > \gamma$
In this case the optimal value is achieved by $\boldsymbol{v} = \boldsymbol{x}$ resulting in an objective value of zero. We shall proceed by contradiction and assume $\boldsymbol{v} = \boldsymbol{x}$ is not an optimal solution. First note that under this contradictory assumption we must have $\|\boldsymbol{v}\|_{\ell_1} < \gamma$ as otherwise the $(\cdot)_+$ term would be inactive and the objective value would be greater than or equal to zero in which case $\boldsymbol{v} = \boldsymbol{x}$ would achieve the optimum value negating the contradictory assumption. We thus focus on the case that $\|\boldsymbol{v}\|_{\ell_1} < \gamma$. To reach a contradiction in this case note that we have

$$\begin{aligned}
&\frac{\lambda}{2}\|\boldsymbol{x}-\boldsymbol{v}\|_{\ell_2}^2 - \frac{1}{2n}\left(\gamma - \|\boldsymbol{v}\|_{\ell_1}\right)_+^2 \\
&\geq \frac{\lambda}{2}\|\boldsymbol{x}-\boldsymbol{v}\|_{\ell_2}^2 - \frac{1}{2n}\left(\|\boldsymbol{x}\|_{\ell_1} - \|\boldsymbol{v}\|_{\ell_1}\right)^2 \\
&\geq \frac{\lambda}{2}\|\boldsymbol{x}-\boldsymbol{v}\|_{\ell_2}^2 - \frac{1}{2n}\|\boldsymbol{x}-\boldsymbol{v}\|_{\ell_1}^2 \\
&> \frac{1}{2}\|\boldsymbol{x}-\boldsymbol{v}\|_{\ell_2}^2 - \frac{1}{2n}\|\boldsymbol{x}-\boldsymbol{v}\|_{\ell_1}^2 \\
&\geq 0,
\end{aligned}$$

where in the penultimum Since $\mathsf{ST}(\boldsymbol{x};0) = \boldsymbol{x}$ and we showed that it is the optimal $\boldsymbol{v}$, the claim holds in this case. Namely, the minimizer is achieved at a point in $\{\mathsf{ST}(\boldsymbol{x};\tau) : \tau \geq 0\}$.
Case II: $\|\boldsymbol{x}\|_{\ell_1} \leq \gamma$
Since $\|\boldsymbol{v}\|_{\ell_1}$ is invariant with respect to the sign of its entries, it is clear that at the solution $\boldsymbol{v}$, we must have $\operatorname{sign}(\boldsymbol{v}) = \operatorname{sign}(\boldsymbol{x})$. Moreover, without loss of generality we can assume $\|\boldsymbol{v}\|_{\ell_1} \leq \gamma$ as otherwise similar to the previous case $\boldsymbol{v} = \boldsymbol{x}$ would be a solution and the minimizer is achieved at a point in $\{\mathsf{ST}(\boldsymbol{x};\tau) : \tau \geq 0\}$. At the optimal solution we must have[5]

$$\mathbf{0} \in \lambda(\boldsymbol{v}-\boldsymbol{x}) - \frac{1}{n}(\|\boldsymbol{v}\|_{\ell_1} - \gamma)\partial\|\boldsymbol{v}\|_{\ell_1}$$

As we argued previously at an optimal solution we must have $\operatorname{sign}(\boldsymbol{v}) = \operatorname{sign}(\boldsymbol{x})$ and thus $\partial\|\boldsymbol{v}\|_{\ell_1} = \partial\|\boldsymbol{x}\|_{\ell_1}$ rearranging the terms gives

$$\boldsymbol{v} \in \boldsymbol{x} + \frac{1}{\lambda n}(\|\boldsymbol{v}\|_{\ell_1} - \gamma)\partial\|\boldsymbol{v}\|_{\ell_1} = \boldsymbol{x} - \frac{1}{\lambda n}(\gamma - \|\boldsymbol{v}\|_{\ell_1})\partial\|\boldsymbol{x}\|_{\ell_1}$$

Thus $\boldsymbol{v} = \mathsf{ST}(\boldsymbol{x};\tau)$ for $\tau = \frac{1}{\lambda n}(\gamma - \|\boldsymbol{v}\|_{\ell_1}) \geq 0$ and the claim follows. □

[5] We note that since $\lambda > 1$ the objective $\frac{\lambda}{2}\|\boldsymbol{x}-\boldsymbol{v}\|_{\ell_2}^2 - \frac{1}{2n}\left(\gamma - \|\boldsymbol{v}\|_{\ell_1}\right)_+^2$ is convex and thus optimality is given by zero being a sub-gradient.

With the lemma above in place we turn our attention to completing the proof of Lemma 6.3. To this aim note that since $f(\boldsymbol{v})$ is convex and $\|\boldsymbol{x}-\boldsymbol{v}\|_{\ell_2}^2$ is strictly convex, then $\frac{1}{2\mu}\|\boldsymbol{x}-\boldsymbol{v}\|_{\ell_2}^2 + f(\boldsymbol{v})$ is jointly striclty convex in $(\boldsymbol{x},\boldsymbol{v})$. Since partial minimization preserves convexity, $e_f(\boldsymbol{x};\mu)$ is strictly convex in $\boldsymbol{x}$ (also see [TAH15, Lemma C.5]).

We write the Moreau envelope as

$$\begin{aligned} e_f(\boldsymbol{x};\mu) &= \min_{\boldsymbol{v}} \frac{1}{2\mu}\|\boldsymbol{x}-\boldsymbol{v}\|_{\ell_2}^2 + \frac{1}{2}\|\boldsymbol{v}\|_{\ell_2}^2 - \frac{1}{2\delta p}\left(\frac{p}{\varepsilon}\gamma - \|\boldsymbol{v}\|_{\ell_1}\right)_+^2 \\ &= \min_{\boldsymbol{v}} \frac{1}{2}\left(\frac{1}{\mu}+1\right)\left\|\boldsymbol{v}-\frac{\boldsymbol{x}}{\mu+1}\right\|_{\ell_2}^2 + \frac{1}{2(\mu+1)}\|\boldsymbol{x}\|_{\ell_2}^2 - \frac{1}{2\delta p}\left(\frac{p}{\varepsilon}\gamma - \|\boldsymbol{v}\|_{\ell_1}\right)_+^2 \end{aligned}$$

Using Lemma C.3 with $\lambda = \frac{1+\mu}{\mu} > 1$, we arrive at

$$\begin{aligned} e_f(\boldsymbol{x};\mu) = \frac{1}{2(\mu+1)}\|\boldsymbol{x}\|_{\ell_2}^2 + \min_{\tau\geq 0}\frac{1}{2\mu(\mu+1)}\|\boldsymbol{x}-\mathsf{ST}(\boldsymbol{x};\tau(\mu+1))\|_{\ell_2}^2 \\ - \frac{1}{2n}\left(\frac{p}{\varepsilon}\gamma - \frac{1}{1+\mu}\|\mathsf{ST}(\boldsymbol{x};\tau(\mu+1))\|_{\ell_1}\right)_+^2. \end{aligned} \tag{C.1}$$

The result follows by a change of variable $\tau(\mu+1)\to\tau$.

C.4 Proof of Lemma 6.4

We begin by restating the lemma for the convenience of the reader.

Lemma C.4 (Restatement of Lemma 6.4)**.** *Let $\boldsymbol{w}\in\mathbb{R}^n$ be a Gaussian random vector distributed as $\mathcal{N}(\boldsymbol{0},\omega^2\boldsymbol{I}_n)$. Also assume*

$$G(\boldsymbol{w};\mu,\tau) := \frac{1}{2\mu(\mu+1)}\|\boldsymbol{w}-\mathsf{ST}(\boldsymbol{w};\tau)\|_{\ell_2}^2 - \frac{1}{2n}\left(\frac{p}{\varepsilon}\gamma - \frac{1}{1+\mu}\|\mathsf{ST}(\boldsymbol{w};\tau)\|_{\ell_1}\right)_+^2.$$

Then

$$\begin{aligned} \lim_{n\to\infty}\frac{1}{n}G(\boldsymbol{w};\mu,\tau) =& \frac{\omega^2}{2\mu(\mu+1)}\left(\left(1-\sqrt{\frac{2}{\pi}}\frac{\tau}{\omega}e^{-\frac{\tau^2}{2\omega^2}}\right)+\left(\frac{\tau^2}{\omega^2}-1\right)erfc\left(\frac{1}{\sqrt{2}}\frac{\tau}{\omega}\right)\right) \\ &- \frac{\omega^2}{2(\mu+1)^2}\left(\frac{\gamma(\mu+1)}{\delta\varepsilon\omega}+\frac{\tau}{\omega}\cdot erfc\left(\frac{1}{\sqrt{2}}\frac{\tau}{\omega}\right)-\sqrt{\frac{2}{\pi}}e^{-\frac{\tau^2}{2\omega^2}}\right)_+^2. \end{aligned}$$

Furthermore,

$$\begin{aligned} &\min_{\tau\geq 0}\lim_{n\to\infty}\frac{1}{n}G(\boldsymbol{w};\mu,\tau) \\ &\quad= \begin{cases} 0 & \text{if } \gamma(\mu+1)\leq\sqrt{\frac{2}{\pi}}\delta\varepsilon\omega \\ \frac{\omega^2}{2\mu(\mu+1)}\left(1-\operatorname{erfc}\left(\frac{\tau^*\left(\frac{\gamma(\mu+1)}{\delta\varepsilon\omega},\mu\right)}{\sqrt{2}}\right)-\frac{\gamma(\mu+1)}{\delta\varepsilon\omega}\tau^*\left(\frac{\gamma(\mu+1)}{\delta\varepsilon\omega},\mu\right)\right) & \text{if } \gamma(\mu+1)>\sqrt{\frac{2}{\pi}}\delta\varepsilon\omega \end{cases} \end{aligned}$$

where $\tau^*(a,\mu)$ is the unique solution to

$$a - \frac{\mu+1}{\mu}\tau + \tau\cdot\operatorname{erfc}\left(\frac{\tau}{\sqrt{2}}\right) - \sqrt{\frac{2}{\pi}}e^{-\frac{\tau^2}{2}} = 0 \tag{C.2}$$

Alternatively using the fact that erf = 1 − erfc *we can rewrite this in the form*

$$\min_{\tau\geq 0}\lim_{n\to\infty}\frac{1}{n}G(\boldsymbol{w};\mu,\tau)$$
$$=\begin{cases}0 & \text{if } \gamma(\mu+1)\leq\sqrt{\frac{2}{\pi}}\delta\varepsilon\omega\\ \frac{\omega^2}{2\mu(\mu+1)}\left(\operatorname{erf}\left(\frac{\tau^*\left(\frac{\gamma(\mu+1)}{\delta\varepsilon\omega},\mu\right)}{\sqrt{2}}\right)-\frac{\gamma(\mu+1)}{\delta\varepsilon\omega}\tau^*\left(\frac{\gamma(\mu+1)}{\delta\varepsilon\omega},\mu\right)\right) & \text{if } \gamma(\mu+1)>\sqrt{\frac{2}{\pi}}\delta\varepsilon\omega\end{cases}$$

where $\tau^*(a,\mu)$ *is the unique solution to*

$$a-\frac{1}{\mu}\tau-\tau\cdot\operatorname{erf}\left(\frac{\tau}{\sqrt{2}}\right)-\sqrt{\frac{2}{\pi}}e^{-\frac{\tau^2}{2}}=0$$

Proof. First note that by law-of large numbers we have

$$\begin{aligned}\lim_{n\to\infty}\frac{1}{n}\left\|\boldsymbol{w}-\mathsf{ST}(\boldsymbol{w};\tau)\right\|_{\ell_2}^2 &= \mathbb{E}_{g\sim\mathcal{N}(0,1)}\left[\left(\omega g-\mathsf{ST}(\omega g;\tau)\right)^2\right]\\ &=\omega^2\,\mathbb{E}_{g\sim\mathcal{N}(0,1)}\left[\left(g-\mathsf{ST}\left(g;\frac{\tau}{\omega}\right)\right)^2\right]\\ &=\omega^2\left(\frac{2}{\sqrt{2\pi}}\int_{+\frac{\tau}{\omega}}^{+\infty}\frac{\tau^2}{\omega^2}e^{-\frac{x^2}{2}}dx+\frac{1}{\sqrt{2\pi}}\int_{-\frac{\tau}{\omega}}^{+\frac{\tau}{\omega}}x^2e^{-\frac{x^2}{2}}dx\right)\\ &=\omega^2\left(\left(1-\sqrt{\frac{2}{\pi}}\frac{\tau}{\omega}e^{-\frac{\tau^2}{2\omega^2}}\right)+\left(\frac{\tau^2}{\omega^2}-1\right)\operatorname{erfc}\left(\frac{1}{\sqrt{2}}\frac{\tau}{\omega}\right)\right)\\ &=\omega\left(\omega-\sqrt{\frac{2}{\pi}}\tau e^{-\frac{\tau^2}{2\omega^2}}\right)+\left(\tau^2-\omega^2\right)\operatorname{erfc}\left(\frac{1}{\sqrt{2}}\frac{\tau}{\omega}\right)\end{aligned}$$

Next note that

$$\begin{aligned}\lim_{n\to\infty}\frac{1}{n}\left\|\mathsf{ST}(\boldsymbol{w};\tau)\right\|_{\ell_1} &= \mathbb{E}_{g\sim\mathcal{N}(0,1)}\left[\left|\mathsf{ST}(\omega g;\tau)\right|\right]\\ &=\omega\,\mathbb{E}_{g\sim\mathcal{N}(0,1)}\left[\left|\mathsf{ST}\left(g;\frac{\tau}{\omega}\right)\right|\right]\\ &=\frac{\omega}{\sqrt{2\pi}}\left(\int_{+\frac{\tau}{\omega}}^{+\infty}\left(x-\frac{\tau}{\omega}\right)e^{-\frac{x^2}{2}}dx-\int_{-\infty}^{-\frac{\tau}{\omega}}\left(x+\frac{\tau}{\omega}\right)e^{-\frac{x^2}{2}}dx\right)\\ &=\sqrt{\frac{2}{\pi}}\omega\left(\int_{+\frac{\tau}{\omega}}^{+\infty}\left(x-\frac{\tau}{\omega}\right)e^{-\frac{x^2}{2}}dx\right)\\ &=\sqrt{\frac{2}{\pi}}\omega\cdot e^{-\frac{\tau^2}{2\omega^2}}-\tau\cdot\operatorname{erfc}\left(\frac{1}{\sqrt{2}}\frac{\tau}{\omega}\right)\end{aligned}$$

Therefore,

$$\begin{aligned}\lim_{n\to\infty}\quad\frac{1}{2n^2}\left(\frac{p}{\varepsilon}\gamma-\frac{1}{1+\mu}\left\|\mathsf{ST}(\boldsymbol{w};\tau)\right\|_{\ell_1}\right)_+^2 &= \lim_{n\to\infty}\quad\frac{1}{2}\left(\frac{\gamma}{\delta\varepsilon}-\frac{1}{1+\mu}\frac{\left\|\mathsf{ST}(\boldsymbol{w};\tau)\right\|_{\ell_1}}{n}\right)_+^2\\ &=\frac{1}{2}\left(\frac{\gamma}{\delta\varepsilon}+\frac{\tau}{1+\mu}\operatorname{erfc}\left(\frac{1}{\sqrt{2}}\frac{\tau}{\omega}\right)-\sqrt{\frac{2}{\pi}}\frac{\omega}{1+\mu}e^{-\frac{\tau^2}{2\omega^2}}\right)_+^2\end{aligned}$$

The proof of the first identity follows by combining the two summands.

To prove the second identity note that using a change of variable $\tau/\omega \to \tau$

$$\begin{aligned}\min_{\tau\geq 0} \quad \lim_{n\to\infty} \frac{1}{n} G(\boldsymbol{w};\mu,\tau\omega) \\ &= \frac{\omega^2}{2(\mu+1)^2}\cdot \min_{\tau\geq 0} \quad \frac{\mu+1}{\mu}\left(\left(1-\sqrt{\frac{2}{\pi}}\tau e^{-\frac{\tau^2}{2}}\right)+\left(\tau^2-1\right)\operatorname{erfc}\left(\frac{\tau}{\sqrt{2}}\right)\right) \\ &\quad -\left(\frac{\gamma(\mu+1)}{\delta\varepsilon\omega}+\tau\cdot\operatorname{erfc}\left(\frac{\tau}{\sqrt{2}}\right)-\sqrt{\frac{2}{\pi}}e^{-\frac{\tau^2}{2}}\right)_+^2\end{aligned}$$

To continue note that if only the first term is active the derivative is given by

$$2\frac{\mu+1}{\mu}\tau\operatorname{erfc}\left(\frac{\tau}{\sqrt{2}}\right)\geq 0$$

and when both terms are active the derivative is given by

$$\begin{aligned}&2\tau\frac{\mu+1}{\mu}\operatorname{erfc}\left(\frac{\tau}{\sqrt{2}}\right)-2\operatorname{erfc}\left(\frac{\tau}{\sqrt{2}}\right)\left(\frac{\gamma(\mu+1)}{\delta\varepsilon\omega}+\tau\cdot\operatorname{erfc}\left(\frac{\tau}{\sqrt{2}}\right)-\sqrt{\frac{2}{\pi}}e^{-\frac{\tau^2}{2}}\right) \\ &\quad =-2\operatorname{erfc}\left(\frac{\tau}{\sqrt{2}}\right)\left((\mu+1)\left(\frac{\gamma}{\delta\varepsilon\omega}-\frac{\tau}{\mu}\right)+\tau\cdot\operatorname{erfc}\left(\frac{\tau}{\sqrt{2}}\right)-\sqrt{\frac{2}{\pi}}e^{-\frac{\tau^2}{2}}\right)\end{aligned}$$

We note that the function $(\mu+1)\left(\frac{\gamma}{\delta\varepsilon\omega}-\frac{\tau}{\mu}\right)+\tau\cdot\operatorname{erfc}\left(\frac{\tau}{\sqrt{2}}\right)-\sqrt{\frac{2}{\pi}}e^{-\frac{\tau^2}{2}}$ is always decreasing when $\tau\geq 0$ and its value at $\tau=0$ is given by $\frac{\gamma(\mu+1)}{\delta\varepsilon\omega}-\sqrt{\frac{2}{\pi}}$. To continue further consider two cases.

Case I: $\gamma(\mu+1)\leq\sqrt{\frac{2}{\pi}}\delta\varepsilon\omega$:

In this case the function is always increasing in $\tau\in[0,+\infty)$ and thus the minimum is achieved at $\tau=0$ with the corresponding optimal value given by

$$-\frac{\omega^2}{2(\mu+1)^2}\left(\frac{\gamma(\mu+1)}{\delta\varepsilon\omega}-\sqrt{\frac{2}{\pi}}\right)_+^2=0$$

Case II: $\gamma(\mu+1)>\sqrt{\frac{2}{\pi}}\delta\varepsilon\omega$:

In this case the function is decreasing at the beginning and then increases. Therefore, the minimum is achieved at a point where

$$(\mu+1)\left(\frac{\gamma}{\delta\varepsilon\omega}-\frac{\tau}{\mu}\right)+\tau\cdot\operatorname{erfc}\left(\frac{\tau}{\sqrt{2}}\right)-\sqrt{\frac{2}{\pi}}e^{-\frac{\tau^2}{2}}=0$$

Note that at such a point we have

$$\frac{\gamma(\mu+1)}{\delta\varepsilon\omega}+\tau\cdot\operatorname{erfc}\left(\frac{\tau}{\sqrt{2}}\right)-\sqrt{\frac{2}{\pi}}e^{-\frac{\tau^2}{2}}=\frac{\mu+1}{\mu}\tau$$

and

$$\left(1-\sqrt{\frac{2}{\pi}}\tau e^{-\frac{\tau^2}{2}}\right)+\left(\tau^2-1\right)\operatorname{erfc}\left(\frac{\tau}{\sqrt{2}}\right)=\tau^2\cdot\operatorname{erfc}\left(\frac{\tau}{\sqrt{2}}\right)-\tau\sqrt{\frac{2}{\pi}}e^{-\frac{\tau^2}{2}}+1-\operatorname{erfc}\left(\frac{\tau}{\sqrt{2}}\right)$$
$$=\frac{\mu+1}{\mu}\tau^2-\frac{\gamma(\mu+1)}{\delta\varepsilon\omega}\tau+1-\operatorname{erfc}\left(\frac{\tau}{\sqrt{2}}\right)$$

Thus

$$\frac{\mu+1}{\mu}\left(\left(1-\sqrt{\frac{2}{\pi}}\tau e^{-\frac{\tau^2}{2}}\right)+\left(\tau^2-1\right)\operatorname{erfc}\left(\frac{\tau}{\sqrt{2}}\right)\right)-\left(\frac{\gamma(\mu+1)}{\delta\varepsilon\omega}+\tau\cdot\operatorname{erfc}\left(\frac{\tau}{\sqrt{2}}\right)-\sqrt{\frac{2}{\pi}}e^{-\frac{\tau^2}{2}}\right)_+^2$$
$$=\frac{(\mu+1)^2}{\mu^2}\tau^2-\frac{\gamma(\mu+1)^2}{\delta\varepsilon\omega\mu}\tau+\frac{\mu+1}{\mu}\left(1-\operatorname{erfc}\left(\frac{\tau}{\sqrt{2}}\right)\right)-\frac{(\mu+1)^2}{\mu^2}\tau^2$$
$$=\frac{\mu+1}{\mu}\left(1-\operatorname{erfc}\left(\frac{\tau}{\sqrt{2}}\right)-\frac{\gamma(\mu+1)}{\delta\varepsilon\omega}\tau\right)$$

□