*Submitted to the Annals of Statistics*

# PRECISE STATISTICAL ANALYSIS OF CLASSIFICATION ACCURACIES FOR ADVERSARIAL TRAINING

By Adel Javanmard[*,†] and Mahdi Soltanolkotabi[*,‡]

*University of Southern California*[*]

Despite the wide empirical success of modern machine learning algorithms and models in a multitude of applications, they are known to be highly susceptible to seemingly small indiscernible perturbations to the input data known as *adversarial attacks*. A variety of recent adversarial training procedures have been proposed to remedy this issue. Despite the success of such procedures at increasing accuracy on adversarially perturbed inputs or *robust accuracy*, these techniques often reduce accuracy on natural unperturbed inputs or *standard accuracy*. Complicating matters further, the effect and trend of adversarial training procedures on standard and robust accuracy is rather counter intuitive and radically dependent on a variety of factors including the perceived form of the perturbation during training, size/quality of data, model overparameterization, etc. In this paper we focus on binary classification problems where the data is generated according to the mixture of two Gaussians with general anisotropic covariance matrices and derive a precise characterization of the standard and robust accuracy for a class of minimax adversarially trained models. We consider a general norm-based adversarial model, where the adversary can add perturbations of bounded $\ell_p$ norm to each input data, for an arbitrary $p \geq 1$. Our comprehensive analysis allows us to theoretically explain several intriguing empirical phenomena and provide a precise understanding of the role of different problem parameters on standard and robust accuracies.

**1. Introduction** Over the past decade there has been a tremendous increase in the use of machine learning models, and deep learning in particular, in a myriad of domains spanning computer vision and speech recognition, to robotics, healthcare and e-commerce. Despite wide empirical success in these and related domains, these modern learning models are known to be

[†]Supported in part by Sloan Research Fellowship in Mathematics, Google Faculty Research Award, Adobe Data Science Research Award and the NSF CAREER Award DMS-1844481.

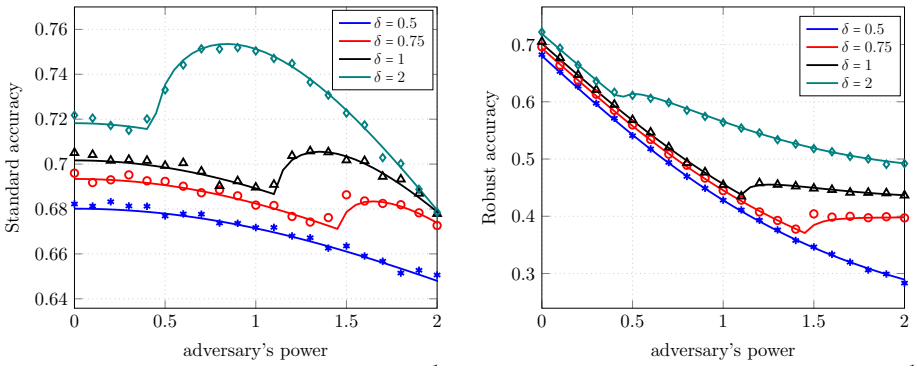
[‡]Supported by the Packard Fellowship in Science and Engineering, Sloan Research Fellowship in Mathematics, an NSF-CAREER under award #1846369, the Air Force Office of Scientific Research Young Investigator Program (AFOSR-YIP) under award #FA9550-18-1-0078, DARPA Learning with Less Labels (LwLL) and FastNICS programs, and NSF-CIF awards #1813877 and #2008443.

*Keywords and phrases:* precise high-dimensional asymptotics, adversarial training, binary classification

highly fragile and susceptible to *adversarial attacks*; even seemingly small imperceptible perturbations to the input data can significantly compromise their performance. As machine learning systems are increasingly being used in applications involving human subjects including healthcare and autonomous driving, such vulnerability can have catastrophic consequences. As a result there has been significant research over the past few years focused on proposing various *adversarial training* methods aimed at mitigating the effect of adversarial perturbations [20, 35, 42, 53, 68].

While adversarial training procedures have been successful in making machine learning models robust to adversarial attacks, their full effect on machine learning systems is not understood. Indeed, adversarial training procedures often behave in mysterious and somewhat counter intuitive ways. For instance, while they improve performance on adversarially perturbed inputs, this benefit often comes at the cost of decreasing accuracy on natural unperturbed inputs. This suggests that the two performance measures, *robust accuracy* –the accuracy on adversarially perturbed inputs– and the *standard accuracy* –accuracy on benign unperturbed inputs– may be fundamentally at conflict. Even more surprising, the performance of adversarial training procedure varies significantly in different settings. For instance, while adversarial trained models yield lower standard accuracy in comparison with non-adversarially trained counterparts, this behavior is completely reversed when there are very few training data with the standard accuracy of adversarially trained models outperforming that of non-adversarial models [65]. We refer the reader to Section 1.2 for a through discussion of recent empirical results that demonstrate how a variety of factors such as the adversary's power, the size of training data, and model over-parameterization affect the performance of adversarially trained models.

To clearly demonstrate the surprising and counterintuitive behavior of adversarially trained models, we plot the behavior of such an approach in Figure 1. We consider a simple binary classification problem with the data generated according to a mixture of two isotropic Gaussians and depict the performance of a commonly used adversarial training procedure. In particular, in this figure, we plot the standard and robust accuracy of an adversarially trained linear classifier for different values of the adversary's perceived power (measured in $\ell_\infty$ perturbations) and different sampling ratios (size of the training data divided by the number of parameters denoted by $\delta$). We would like to highlight the highly non-trivial behavior of the standard and robust accuracy curves with respect to the adversary's power and the sampling ratio. For instance, the standard accuracy first decreases, then increases and again decreases as a function of the adversary's power. Furthermore, the




(a) Standard accuracy (b) Robust accuracy

Fig 1: Depiction of standard and robust accuracies as a function of the adversary's power with $\ell_\infty$ ($p = \infty$) perturbation for different values of $\delta$ (ratio of the size of the training data to the number of parameters in the model). Solid curves are theoretical predictions and dots are the empirical results. We refer to Figure 5 and Section 5.2 for further details.

exact nature of this curve is highly reliant on the sampling ratio $\delta$. Similarly, for robust accuracy, we first observe a decreasing trend for all $\delta$, but after some threshold depending on $\delta$, robust accuracy increases and then decreases or stays constant. Even more surprising, as we will see in the forth-coming sections the behavior of these curves vary drastically for different forms of $\ell_p$ perturbations. This simple experiment clearly demonstrates the importance of having a precise theory for characterizing the rather nuanced performance of adversarial training procedures and demystify their behavior. Developing such a precise theoretical analysis is exactly the goal of this paper. Indeed, the solid curves in Figure 1 are based on our theoretical predictions!

1.1. *Contributions* In this paper we focus on binary classification problems where the data is generated according to the mixture of two Gaussians with general anisotropic covariance matrices and derive a precise characterization of the standard and robust accuracy for a class of minimax adversarially trained models. We consider a general norm-based adversarial model, where the adversary can add perturbations of bounded $\ell_p$ norm to each input data, for an arbitrary $p \geq 1$. We would like to emphasize that our theory provides a *precise characterization* of the performance of this class of adversarially trained models, rather than just upper bounds on the standard and robust

accuracies. Our analysis for such a broad setting allows us to capture several intriguing phenomena that we discuss next.

We show and theoretically prove an interesting phase transition phenomena holds for adversarial classification applied to the Gaussian mixture model. Specifically, we characterize a threshold $\delta_*$ for the ratio of size of training data to feature dimension, $\delta$ so that when $\delta < \delta_*$, the data is *robustly separable* with high probability, and for $\delta > \delta_*$ it is non-separable, with high probability. Here, *robust separability* is a generalization of the classical linear separability condition for data and roughly speaking means that there is a linear separator that correctly separates the two label classes with a positive margin that depends on the adversary's power. We precisely characterize the threshold $\delta_*$ in terms of various problem parameters including the mean and covariance of the mixture components, the adversary's power, and the $\ell_p$ perturbation norm. Interestingly, $\delta_*$ is related to the spherical width of a set defined in terms of the dual $\ell_q$ norm ($1/p + 1/q = 1$) conforming with classical notions of prior knowledge and complexity used in the compressive sensing literature.

Our precise theoretical characterization of standard and robust accuracies provides a precise understanding of the role that different problem parameters such as size/quality of the training data, feature covariates and means, model overparameterization $(1/\delta)$, and the adversary's perceived power have during training on these performance measures. Surprisingly, our analysis reveals that the effects of these factors very much depend on the choice of perturbations norm $\ell_p$. For example, in the robustly separable regime, we observe that for $p = 2$ adversarial training has no effect on standard accuracy, while for $p = 1$ and $p = \infty$ it hurts the standard accuracy. In the non-separable regime, we observe that for $p = 2$ adversarial training helps with improving the standard accuracy. However, for $p = \infty$ the adversarial training first improves the standard accuracy but as the training procedure hedges against stronger adversary, after some threshold on the adversary's power, we start to see a decrease in the standard accuracy of the resulting model. Interestingly, this threshold on the adversary's power varies with model overparameterization.

Lastly, a key ingredient of our analysis is a powerful extension of Gordon's Gaussian process inequality [21] known as the Convex Gaussian Minimax Theorem (CGMT) developed in [64] and further extended in [63, 12] for various learning settings. Using this technique we provide a precise prediction of the performance of adversarial training in terms of the optimal solutions to a convex-concave problem with a small number of scalar variables that can be easily solved by a low-dimensional gradient descent/ascent rather fast and accurately. In addition, this low-dimensional optimization problem can be significantly simplified for special cases of $p$ (see Section 5 for details). While

CGMT has been used to study the behavior of regularized M-estimators, using this framework for the broad class of minimax adversarially trained models studied in this paper (including general anisotropic covariance matrices and general choice of $\ell_p$ norm for adversarial perturbations) poses significant technical challenges. Specifically, the intrinsic differences between $\ell_p$ geometries and the interaction between the class means the feature covariance matrix in the model requires a rather intricate and technical analysis.

1.2. *Related work* We briefly discuss the related literature along two lines.

**Other models of adversarial perturbations.** Another popular model for adversarial attacks on the models is the so-called distribution shifts, wherein the adversary can shift the test data distribution, making it different from the training distribution. The adversary is assumed to have limited manipulative power in terms of the Wasserstein distance between the test and the training distributions [58, 52, 44]. The articles [2, 44] study the robust loss $L(\boldsymbol{\theta}; \varepsilon) = \sup_{\nu \in B_\varepsilon(\mu)} \mathbb{E}_\nu[\ell(\boldsymbol{z}, \boldsymbol{\theta})]$, where $B_\varepsilon(\mu)$ is the $\varepsilon$ ball around $\mu$ in the Wasserstein ($W_p$) distance for some $p \in [1, \infty)$, and the data $\boldsymbol{z} = (\boldsymbol{x}, y) \sim \mu$. A first order approximation of the robust loss $L(\boldsymbol{\theta}; \varepsilon)$ is given for small $\varepsilon$, in terms of a variation measure of the original loss $\ell$. Such characterization is used in [44] to investigate the tradeoff between the standard and robust accuracies for various learning problems. Note that these work are focused on the population loss ($n \to \infty$, with $d$ fixed). In comparison, in this paper we study norm bounded adversarial perturbations and work with empirical loss in asymptotic regime ($n, d \to \infty$, with $n/d = \delta$ fixed).

In adversarial training it is assumed that the modeler has access to clean (unperturbed) data and strives to construct a model that is resilient to potential adversarial perturbations of the test data. The article [36] considers a different adversarial setup in which an attacker can observe and modify all training data samples in an adversarial manner so as to maximize the estimation error caused by his attack. This work introduces the notion of adversarial influence function (AIF) to quantify the sensitivity of estimators to such adversarial attacks, and further derive the optimal estimator, among a certain class of estimator, that minimizes AIF.

**Standard accuracy and robust accuracy tradeoffs.** Several recent papers contain empirical results suggesting a potential trade-off between standard accuracy and robust accuracy. A few papers have started to shed light on the theoretical foundations of such tradeoffs [42, 57, 65, 54, 69, 30, 47, 14] often focusing on very specific models or settings. However, a comprehensive quantitative understanding of such tradeoffs is largely underdeveloped.

A central question we wish to address in this paper is whether there exists a fundamental conflict between robust accuracy and standard accuracy. We briefly mention a few papers that take a step towards addressing this question. In [65, 69], the authors provide examples of learning problems where no predictor can achieve both optimal standard accuracy and robust accuracy in the infinite data limit, pointing to such fundamental tradeoff. By contrast, [54] provides examples where there is no such tradeoff in the infinite data limit, in the sense that the optimal predictor performs well on both objectives, however a tradeoff is still observed with finite data. Despite this interesting progress a quantitive understanding of fundamental and algorithmic tradeoffs between standard and robust accuracies and how they are affected by various factors, such as overparameterization, adversary's power and the data model is still missing. Such a result requires novel perspectives and analytical tools to precisely characterize the behavior of robust and standard accuracies, which is one of the motivating factors behind our current paper.

More closely related to this paper, in [30] the current authors used the convex Gaussian minimax framework to provide a precise characterization of standard and robust accuracies for linear regression, studying the fundamental conflict between these objectives along with algorithmic tradeoffs for specific minimax estimators. For classification problems, a recent paper [14] focuses on characterizing the optimal $\ell_2$ and $\ell_\infty$ robust linear classifiers assuming access to the class means. This paper also studies some tradeoffs between standard and robust accuracies by contrasting this optimal robust classifier with the Bayes optimal classifier in a non-adversarial setting. This paper however does not directly study the tradeoffs of adversarial training procedures except for linear losses. A related publication [47] studies the generalization property of an adversarially trained model for classification on a Gaussian mixture model with a diagonal covariance matrix and a linear loss. In this setting, this work discusses the different effects that more training data can have on generalization based on the strength of the adversary. Using a linear loss in the above two classification papers is convenient as in this case the adversarially trained model admits a simple closed form representation. We also note that these two papers do not seem to focus on the high-dimensional regime where the number of training data grow in proportion to the number of parameters. In contrast, in this paper we focus on developing a comprehensive theory that provides a precise characterization of standard and robust accuracies and their tradeoffs in the high dimensional regime for a broad class of loss functions and covariance matrices. Such a comprehensive analysis allows us to better understand the role of the loss function in adversarial training. Indeed, as we demonstrate, the behavior of standard and robust accuracy for

nonlinear loss functions can be very different from linear losses. We also note that such a theoretical result requires much more intricate techniques as the adversarially trained model does not admit a simple closed form. Finally, we would like to note while in this paper we provide a precise understanding of the tradeoffs between standard and robust accuracies for commonly used adversarial training algorithms our work still does not address two tantalizing open questions: What is the optimal standard-robust accuracy tradeoff for a fixed ratio of sample size to dimension? Are there adversarial training approaches that achieve the optimal tradeoff between standard and robust accuracies universally over the range of adversary's power.

**2. Problem formulation** In this section we discuss the problem setting and formulation of this paper in greater detail. After adopting some notations, we describe the adversarial training for binary classification in Section 2.1. Next, we discuss the data model and asymptotic setting studied in this paper in Section 2.2. Finally, in Section 2.3 we formally define the standard and robust classification accuracies in this model.

**Notations.** For a vector $\boldsymbol{v} \in \mathbb{R}^d$, we write $\|\boldsymbol{v}\|_{\ell_p}$ for the standard $\ell_p$ norm of $\boldsymbol{v}$, i.e., $\|\boldsymbol{v}\|_{\ell_p} = (\sum_i |v_i|^p)^{1/p}$. For a matrix $\boldsymbol{\Sigma}$, $\|\boldsymbol{\Sigma}\|$ indicates the spectrum norm of $\boldsymbol{\Sigma}$. Throughout, we say a probabilistic event holds 'with high probability', when its probability converges to one as $n \to \infty$. In addition, for a sequence of random variables $\{X_n\}_{n\in\mathbb{N}}$ and a constant $c$ (independent of $n$) we write $\lim_{n\to\infty} X_n = c$, 'in probability' if $\forall \varepsilon > 0$ we have $\lim_{n\to\infty} \mathbb{P}(|X_n - c| > \varepsilon) = 0$.

2.1. *Adversarial training for binary classification* In binary classification we have access to a training data set of $n$ input-output pairs $\{(\boldsymbol{x}_i, y_i)\}_{i=1}^n$ with $\boldsymbol{x}_i \in \mathbb{R}^d$ representing the input features and $y_i \in \{-1, +1\}$ representing the binary class label associated to each data point. Throughout we assume the data points $(\boldsymbol{x}_i, y_i)$ are generated i.i.d. according to a distribution $\mathbb{P}$. To find a classifier that predicts the labels, one typically fits a function $f_{\boldsymbol{\theta}}$, parameterized by $\boldsymbol{\theta} \in \mathbb{R}^d$ to the training data via empirical risk minimization. In this paper we focus on linear classifiers of the form $f_{\boldsymbol{\theta}}(\boldsymbol{x}) = \langle \boldsymbol{x}, \boldsymbol{\theta} \rangle$ in which case the training problem takes the form

$$\widehat{\boldsymbol{\theta}} := \arg\min_{\boldsymbol{\theta}\in\mathbb{R}^d} \frac{1}{n}\sum_{i=1}^n \ell(y_i f_{\boldsymbol{\theta}}(\boldsymbol{x}_i)) = \arg\min_{\boldsymbol{\theta}\in\mathbb{R}^d} \frac{1}{n}\sum_{i=1}^n \ell(y_i \langle \boldsymbol{x}_i, \boldsymbol{\theta} \rangle)\,. \tag{2.1}$$

Here, $\ell$ is a loss and $\ell(y_i\langle \boldsymbol{x}_i, \boldsymbol{\theta}\rangle)$ approximately measuring the missclassification between the labels $y_i$ and the output of the model $\langle \boldsymbol{x}_i, \boldsymbol{\theta}\rangle$. Some common choices include logistic loss $\ell(t) = \log(1 + e^{-t})$, exponential loss $\ell(t) = e^{-t}$, and hinge loss $\ell(t) = \max(0, 1 - t)$. Once the parameter $\widehat{\boldsymbol{\theta}}$ is estimated one can

find the predicted label by simply calculating the sign of the model output $\widehat{y} = \mathrm{sgn}(f_{\widehat{\boldsymbol{\theta}}}(\boldsymbol{x})) = \mathrm{sgn}(\langle \boldsymbol{x}, \widehat{\boldsymbol{\theta}} \rangle)$.

Despite the widespread of empirical risk minimizers in supervised learning, these estimators are known to be highly vulnerable to even minute perturbations in the input features $\boldsymbol{x}_i$. In particular, it is known that even small, norm-bounded perturbations to the features that are imperceptible to the human eye, can lead to surprising miss-classification errors. These observations have spurred a surge of interest in adversarial training where the goal is to learn models that are robust against such adversarial perturbation. In this paper we focus on an adversarial training approach that is based on using a robust minimax loss [65, 42]. In our linear binary classification setting the robust minimax estimator takes the form

$$\widehat{\boldsymbol{\theta}}^{\varepsilon} := \arg\min_{\boldsymbol{\theta} \in \mathbb{R}^d} \frac{1}{n} \sum_{i=1}^{n} \max_{\|\boldsymbol{\delta}_i\|_{\ell_p} \le \varepsilon} \ell\left(y_i \langle \boldsymbol{x}_i + \boldsymbol{\delta}_i, \boldsymbol{\theta} \rangle\right). \tag{2.2}$$

The main intuition behind such an estimator is that although the learner has access to unperturbed training data, instead of fitting to that data she imitates potential adversarial perturbations to test data in the training data and aims to learn a model that performs well in the presence of such perturbations. One can also view this adversarial training approach as an implicit smoothing that tries to fit the same label $y_i$ to all the features in the $\varepsilon$-neighborhood of $\boldsymbol{x}_i$ simultaneously.

In this paper we focus on convex and decreasing losses such as the aforementioned logistic, exponential, and hinge losses. In such cases the inner maximization in (2.2) can be solved in closed form. In particular, the worst perturbation $\boldsymbol{\delta}_i$ in terms of loss value is given by $\boldsymbol{\delta}_i^* = \arg\min\{y_i \langle \boldsymbol{\delta}_i, \boldsymbol{\theta} \rangle : \|\boldsymbol{\delta}_i\|_{\ell_p} \le \varepsilon\}$, which by using Holder's inequality results in $y_i \langle \boldsymbol{\delta}_i^*, \boldsymbol{\theta} \rangle = -\varepsilon \|\boldsymbol{\theta}\|_{\ell_q}$. Therefore the adversarially trained model $\widehat{\boldsymbol{\theta}}^{\varepsilon}$ can be equivalently written as

$$\widehat{\boldsymbol{\theta}}^{\varepsilon} := \arg\min_{\boldsymbol{\theta} \in \mathbb{R}^d} \frac{1}{n} \sum_{i=1}^{n} \ell\left(y_i \langle \boldsymbol{x}_i, \boldsymbol{\theta} \rangle - \varepsilon \|\boldsymbol{\theta}\|_{\ell_q}\right). \tag{2.3}$$

2.2. *Data model and asymptotic setting* We consider supervised binary classification under a Gaussian Mixture data Model (GMM). Concretely, each data point belongs to one of two classes $\{\pm 1\}$ with corresponding probabilities $\pi_+$, $\pi_-$, so that $\pi_+ + \pi_- = 1$. Given the label $y_i \in \{-1, +1\}$ for data point $i$, the associated input/feature vectors $\boldsymbol{x}_i \in \mathbb{R}^d$ are generated independently according to the distribution $\boldsymbol{x}_i \sim \mathsf{N}(y_i \boldsymbol{\mu}, \boldsymbol{\Sigma})$, conditioned on $y_i$, where $\boldsymbol{\mu} \in \mathbb{R}^d$ and $\boldsymbol{\Sigma} \in \mathbb{R}^{d \times d}$. In other words the mean of feature vectors are $\pm \boldsymbol{\mu}$ depending on its class, and $\boldsymbol{\Sigma}$ is the covariance of features. We depict this mixture

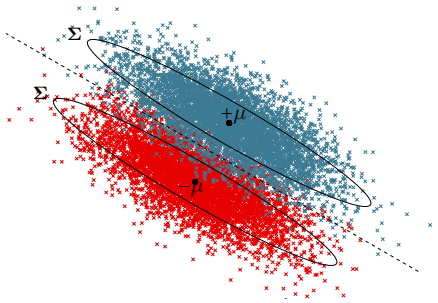



Fig 2: Depiction of the Mixture of Gaussian data model.

model in Figure 2. We next describe the asymptotic regime of interest and our assumptions in this paper.

**Assumption 1 (Asymptotic Setting)** *We focus on the following asymptotic regime:*

*(a) (Scaling of dimensions)* $n \to \infty$ *and* $\frac{n}{d} \to \delta \in (0, \infty)$.

*(b) (Scaling of signal to noise ratio) We have* $C_{\min} \le \frac{\|\boldsymbol{\mu}\|_{\ell_2}}{\|\boldsymbol{\Sigma}\|} \le C_{\max}$ *for some positive constants* $C_{\min}$ *and* $C_{\max}$*, which are independent of* $n$ *and* $d$.

*(c) (Scaling of adversary's power) We have* $\varepsilon = \varepsilon_0 \|\boldsymbol{\mu}\|_{\ell_p}$ *for a constant* $\varepsilon_0 \ge 0$ *which we refer to as adversary's normalized power.*

Assumption 1 (a) details our high-dimensional regime where the size of the training data $n$ and the dimension of the features $d$ grow proportionally with their ratio fixed at $\delta$. We would like to note that while we focus on this asymptotic regime our theoretical technique can also demonstrate very accurate concentration around this asymptotic behavior. Assumption 1 (b) demonstrates the scaling of the signal to noise ratio and ensures that the distance between the centers of the two components $2\|\boldsymbol{\mu}\|_{\ell_2}$ ('signal') is comparable to the projection of noise in any direction (noise). Finally, Assumption 1 details our scaling of the adversary's power. This scaling is justified as if the adversary could perturb data points $\boldsymbol{x}_i$ by $2\boldsymbol{\mu}$, she can flip the label of every data point, so that the leaner cannot do better than random guessing. Since the perturbations can be chosen arbitrary from an $\ell_p$ ball of radius $\varepsilon$, we require $\varepsilon$ to be comparable to $\|\boldsymbol{\mu}\|_{\ell_p}$.

2.3. *Standard and robust accuracies* Our goal is this paper is to precisely characterize performance of the estimator $\widehat{\boldsymbol{\theta}}^{\varepsilon}$ in terms of two accuracies and understand the interplay between them. The two accuracies are *standard accuracy* which is the accuracy on unperturbed test data, and *robust accuracy*

which is the accuracy on adversarially perturbed test data. More formally *standard accuracy* quantifies the accuracy of an estimator on an unperturbed test data that is generated from the same distribution as the training data:

$$\mathsf{SA}(\widehat{\boldsymbol{\theta}}) := \mathbb{P}\{\widehat{y} = y\}\,, \quad \text{where} \quad (\boldsymbol{x}, y) \sim \mathbb{P} \tag{2.4}$$

Our second accuracy, called *robust accuracy* quantifies robustness of an estimator to adversarial perturbations in the test data. Specifically,

$$\mathsf{RA}(\widehat{\boldsymbol{\theta}}) := \mathbb{E}\Big[\min_{\|\boldsymbol{\delta}\|_{\ell_p}\le\varepsilon} \mathbb{1}_{\{y\langle \boldsymbol{x}+\boldsymbol{\delta},\widehat{\boldsymbol{\theta}}\rangle \ge 0\}}\Big], \quad \text{where} \quad (\boldsymbol{x}, y) \sim \mathbb{P}. \tag{2.5}$$

We end this section by stating a lemma that characterizes $\mathsf{SA}(\widehat{\boldsymbol{\theta}})$ and $\mathsf{RA}(\widehat{\boldsymbol{\theta}})$ under the Gaussian mixture model. We defer the proof to Appendix E.1.

**Lemma 2.1** *Consider mixtures of Gaussian data model where $y_i \in \{-1, +1\}$ with corresponding probabilities $\pi_-, \pi_+$ and the feature vector distributed as $\boldsymbol{x} \sim \mathsf{N}(y\boldsymbol{\mu}, \boldsymbol{\Sigma})$, conditioned on $y$, where $\boldsymbol{\mu} \in \mathbb{R}^d$ and $\boldsymbol{\Sigma} \in \mathbb{R}^{d\times d}$. Then,*

$$\mathsf{SA}(\widehat{\boldsymbol{\theta}}) := \Phi\left(\frac{\langle \boldsymbol{\mu}, \widehat{\boldsymbol{\theta}}\rangle}{\|\boldsymbol{\Sigma}^{1/2}\widehat{\boldsymbol{\theta}}\|_{\ell_2}}\right), \tag{2.6}$$

$$\mathsf{RA}(\widehat{\boldsymbol{\theta}}) := \Phi\left(\frac{\langle \boldsymbol{\mu}, \widehat{\boldsymbol{\theta}}\rangle - \varepsilon\,\|\widehat{\boldsymbol{\theta}}\|_{\ell_q}}{\|\boldsymbol{\Sigma}^{1/2}\widehat{\boldsymbol{\theta}}\|_{\ell_2}}\right). \tag{2.7}$$

*Here, $\Phi(x) = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{x} e^{-\frac{t^2}{2}}\mathrm{d}t$ is the cdf of a standard Gaussian distribution and $q$ is such that $\frac{1}{p} + \frac{1}{q} = 1$.*

By Lemma 2.1, characterizing $\mathsf{SA}(\widehat{\boldsymbol{\theta}}^{\varepsilon})$ and $\mathsf{RA}(\widehat{\boldsymbol{\theta}}^{\varepsilon})$ amounts to characterizing $\langle \boldsymbol{\mu}, \widehat{\boldsymbol{\theta}}^{\varepsilon}\rangle$, $\|\boldsymbol{\Sigma}^{1/2}\widehat{\boldsymbol{\theta}}\|_{\ell_2}$, $\|\widehat{\boldsymbol{\theta}}^{\varepsilon}\|_{\ell_q}$, which constitutes the bulk of our analysis.

**3. Prelude: two regimes for adversarial training** Similar to normal classification, an interesting phenomena that arises in adversarial classification is that depending on the size of the training data there are two different regimes of operation: Robustly separable and non-separable. In the *robustly separable* regime there is a robust classifier that perfectly separates the training data, with a positive margin that depends on the adversary's power, while this is not possible in the non-separable case. We formally define this notion of robust separability below.

**Definition 3.1 (Robust linear separability)** *Given $\varepsilon > 0$ and $q \ge 1$, we call a training data $\{(\boldsymbol{x}_i, y_i)\}_{i=1}^n$, $(\varepsilon, q)$-separable if*

$$\exists \boldsymbol{\theta} \in \mathbb{R}^d : \quad \forall i \in [n],\ \ y_i\langle \boldsymbol{x}_i, \boldsymbol{\theta}\rangle - \varepsilon\,\|\boldsymbol{\theta}\|_{\ell_q} > 0. \tag{3.1}$$

We note that our notion of robust separability is closely related to the standard notion of separability by a linear classifier. In particular, using a simple rescaling argument[1] one can rewrite condition 3.1 as follows

$$\exists \boldsymbol{\theta},\ \|\boldsymbol{\theta}\|_{\ell_q} \le \frac{1}{\varepsilon}: \quad \forall i \in [n],\ y_i \langle \boldsymbol{x}_i, \boldsymbol{\theta} \rangle > 1. \tag{3.2}$$

Therefore, robust separability is akin to linear separability of the data but with a budget constraint on the $\ell_q$ norm of the coefficients of the classifier.

When the training data is $(\varepsilon, q)$-separable (with $\ell_q$ the dual norm of $\ell_p$), then the minimax estimator $\widehat{\boldsymbol{\theta}}^{\varepsilon}$ becomes unbounded and achieves zero adversarial training loss in (2.3). In other words, one can completely interpolate the data. This is due to the fact that if $\boldsymbol{\theta}$ is an $(\epsilon, q)$-separator, then $c\boldsymbol{\theta}$ with $c \to \infty$ leads to zero adversarial training loss and since the loss is nonnegative it is optimal. Although the norm of $\widehat{\boldsymbol{\theta}}^{\varepsilon}$ tends to infinity in the separable regime, what matters for our linear classifier is the direction of $\widehat{\boldsymbol{\theta}}^{\varepsilon}$. However, in this separable regime even the direction of the optimal solution $(\frac{\widehat{\boldsymbol{\theta}}^{\varepsilon}}{\|\widehat{\boldsymbol{\theta}}^{\varepsilon}\|_{\ell_2}})$ may not be unique. Even though there may be multiple optimal directions it is possible to show that the direction that gradient descent converges to is a specific maximum margin classifier. We formally state this result which is essentially a direct consequence of [41, 31] below.

**Proposition 3.2** *Consider the adversarial training loss*

$$\mathcal{L}(\boldsymbol{\theta}) := \frac{1}{n} \sum_{i=1}^{n} \ell\left( y_i \langle \boldsymbol{x}_i, \boldsymbol{\theta} \rangle - \varepsilon \|\boldsymbol{\theta}\|_{\ell_q} \right),$$

*with the loss $\ell(t)$ obeying certain technical assumptions[2] which are satisfied for common classification losses such as logistic, exponential, and hinge losses. Then, the gradient descent iterates*

$$\boldsymbol{\theta}_{\tau+1} = \boldsymbol{\theta}_{\tau} - \mu \nabla \mathcal{L}(\boldsymbol{\theta}_{\tau})$$

*with a sufficiently small step size $\mu$ obey*

$$\lim_{t \to \infty} \left\| \frac{\boldsymbol{\theta}_t}{\|\boldsymbol{\theta}_t\|_{\ell_2}} - \frac{\widetilde{\boldsymbol{\theta}}^{\varepsilon}}{\|\widetilde{\boldsymbol{\theta}}^{\varepsilon}\|_{\ell_2}} \right\|_{\ell_2} = 0, \tag{3.3}$$

[1] (3.1)⇒(3.2): Scaling by $\frac{1}{\varepsilon\|\boldsymbol{\theta}\|_{\ell_q}}$ we see that $y_i\langle x_i, \tilde{\boldsymbol{\theta}}\rangle > 1$ for $\tilde{\boldsymbol{\theta}} = \frac{\boldsymbol{\theta}}{\varepsilon\|\boldsymbol{\theta}\|_{\ell_q}}$, and $\|\tilde{\boldsymbol{\theta}}\|_{\ell_q} = \frac{1}{\varepsilon}$ by definition. (3.2)⇒(3.1): Letting $c = \frac{1}{\varepsilon\|\boldsymbol{\theta}\|_{\ell_q}} \ge 1$, we also have $y_i\langle x_i, c\boldsymbol{\theta}\rangle > 1$. Substituting for $c$ and rearranging the terms we get (3.1).

[2] See [41, Assumption S3 in Appendix F]. We list these assumptions in Appendix F for readers' convenience.

*where* $\widetilde{\boldsymbol{\theta}}^{\varepsilon}$ *is the solution to the following max-margin problem*

$$\widetilde{\boldsymbol{\theta}}^{\varepsilon} = \arg\min_{\boldsymbol{\theta}\in\mathbb{R}^d} \quad \|\boldsymbol{\theta}\|_{\ell_2}^2 \qquad \text{subject to} \quad y_i\langle \boldsymbol{x}_i, \boldsymbol{\theta}\rangle - \varepsilon \|\boldsymbol{\theta}\|_{\ell_q} \geq 1\,. \tag{3.4}$$

In the non-separable regime, as we show in the proof of Theorem 4.5 the minimizer $\widehat{\boldsymbol{\theta}}^{\varepsilon}$ is bounded. Moreover, the loss (2.3) is convex as it is pointwise maximum of a set of convex functions (see (2.2) and recall convexity of loss $\ell$). Therefore, a variety of iterative methods (including gradient descent) can be used to converge to a global minimizer of (2.2). Theorem 4.5 also shows that all global minimizers of (2.2) have the same standard and robust accuracy.

**4. Main results for isotropic features** In this section we present our main results. For the sake of exposition, in this section we state our results for the case where the features are isotropic (i.e. $\boldsymbol{\Sigma} = \boldsymbol{I}$). We discuss our more general results with anisotropic features in Section 6. In this paper, we establish a sharp phase-transition characterizing the separability of the training data generated according to a Gaussian mixture model. Specifically, in our asymptotic regime (see Section 2.1) we characterize a threshold $\delta_*$ such that for $\delta < \delta_*$ the data is $(\varepsilon, q)$-separable, with high probability, and for $\delta > \delta_*$ it is non-separable, with high probability. This phase transition for robust separability is discussed in Section 4.1. We also precisely characterize the standard accuracy $\mathsf{SA}(\widehat{\boldsymbol{\theta}}^{\varepsilon})$ and the robust accuracy $\mathsf{RA}(\widehat{\boldsymbol{\theta}}^{\varepsilon})$ of the point that gradient descent converges to in both the separable and non-separable data regimes which are the subject of Sections 4.2 and 4.3, respectively. We then discuss the implications of our main results for the special cases of $\ell_p$ perturbations with $p = 1$, $p = 2$, and $p = \infty$ in Section 5.

4.1. *Phase transition for robust data separability* In this section we discuss our results for characterizing the phase transition for $(\varepsilon, q)$-separability under the Gaussian mixtures model. As detailed earlier in Section 2.2, in our asymptotic setting the dimension of the mean vector $\boldsymbol{\mu}$ $(d)$ as well as the size of the training data $(n)$ grow to infinity in proportion with each other $n/d = \delta$. To state our main result we need a few technical assumptions on the limiting behavior of the mean vector. We begin with a simple assumption on the convergence of the Euclidean norm of the mean vector.

**Assumption 2 (Convergence of Euclidean norm of $\boldsymbol{\mu}$)** *We assume the Euclidean norm of the mean vector converges to a bounded quantity, that is* $\|\boldsymbol{\mu}\|_{\ell_2} \to V < \infty$, *as* $n \to \infty$ *and* $n/d \to \delta$.

We note that for the isotropic case, the boundedness condition in Assumption 2 is already implied by Assumption 1(b).

Naturally, the separability threshold depends on the mean vector and the adversary's power. For instance, intuitively, one expects the separability threshold to decrease as the adversary's power or the length of the mean vector increases. We also expect the direction of the mean vector $\frac{\boldsymbol{\mu}}{\|\boldsymbol{\mu}\|_{\ell_2}}$ to play a role. We capture these effects via the spherical width of a suitable set. Recall that the *spherical width* of a set $\mathcal{S} \subset \mathbb{R}^d$ is a measure of its complexity and is defined as

$$\omega_s(\mathcal{S}) = \mathbb{E}\Big[\sup_{\boldsymbol{z}\in\mathcal{S}} \boldsymbol{z}^T \boldsymbol{u}\Big],$$

where $\boldsymbol{u} \in \mathcal{S}^{d-1}$ is a vector chosen uniformly at random from the unit sphere. In particular, the appropriate set for characterizing the separability threshold takes the form

(4.1)
$$\mathcal{S}(\alpha,\theta,\varepsilon_0,\boldsymbol{\mu}) := \left\{\boldsymbol{z}\in\mathbb{R}^d : \quad \boldsymbol{z}^T\boldsymbol{\mu} = 0,\ \|\boldsymbol{z}\|_{\ell_2} \le \alpha,\ \left\|\boldsymbol{z} + \theta\frac{\boldsymbol{\mu}}{\|\boldsymbol{\mu}\|_{\ell_2}}\right\|_{\ell_q} \le \frac{1}{\varepsilon_0\|\boldsymbol{\mu}\|_{\ell_p}}\right\},$$

where $\varepsilon_0$ is the adversary's scaled power per Assumption 1(c). Next assumption focuses on the spherical width convergence in our asymptotic regime.

**Assumption 3 (Convergence of spherical width)** *We assume the following limit exists*

$$\omega(\alpha,\theta,\varepsilon_0) := \lim_{n\to\infty} \omega_s(\mathcal{S}(\alpha,\theta,\varepsilon_0,\boldsymbol{\mu}))\ . \tag{4.2}$$

As it will become clear later on in this section Assumptions 2 and 3 are trivially satisfied in various settings. With these assumptions in place we are ready to state our result precisely characterizing the separability threshold.

**Theorem 4.1** *Consider a data set generated i.i.d. according to an isotropic Gaussian mixture data model per Section 2.2 and suppose the mean vector $\boldsymbol{\mu}$ obeys Assumptions 2 and 3. Also define*

$$\delta_* := \min_{\alpha\ge 0,\theta} \frac{\omega(\alpha,\theta,\varepsilon_0)^2}{\mathbb{E}\left[\left(1 - V\theta + \sqrt{\alpha^2+\theta^2}g\right)_+^2\right]}, \tag{4.3}$$

*where the expectation is taken with respect to $g \sim \mathsf{N}(0,1)$. Then, under the asymptotic setting of Assumption 1, for $\delta < \delta_*$ the data are $(\varepsilon, q)$-separable*

*with high probability and for $\delta > \delta_*$, the data are non-separable, with high probability. Namely,*

$$\delta < \delta_* \Rightarrow \lim_{n\to\infty} \mathbb{P}(\textit{data is } (\varepsilon, q)\textit{-separable}) = 1\,,$$
$$\delta > \delta_* \Rightarrow \lim_{n\to\infty} \mathbb{P}(\textit{data is } (\varepsilon, q)\textit{-separable}) = 0\,.$$

Theorem 4.1 above precisely characterizes the separability threshold as a function of the adversary's power as well as properties of the mean vector. In particular since $\omega$ decreases with the increase in $\varepsilon_0$, this theorem indicates that the separability threshold decreases as the adversary's power increases. This of course conforms with our natural intuition and is consistent with characterization (3.2). To better understand the implications of Theorem 4.1 we now consider some special cases.

- **Example 1 (Non-adversarial setting).** Our first example focuses on the non-adversarial setting where $\varepsilon_0 = 0$. In this case the $\ell_q$ constraint in definition of $\mathcal{S}$, given by (4.1), is void and the set $\mathcal{S}$ becomes the intersection of $\ell_2$ ball of radius $\alpha$ with the hyperplane of dimension $d-1$ that is orthogonal to $\boldsymbol{\mu}$. Therefore $\omega(\alpha, \theta, \varepsilon_0) = \omega_s(\mathcal{S}) = \alpha$ and the separability threshold reduces to

$$\delta_* := \max_{\alpha\geq 0, \theta} \frac{\alpha^2}{\mathbb{E}\left[\left(1 - V\theta + \sqrt{\alpha^2+\theta^2}g\right)_+^2\right]}\,.$$

  By the change of variables $(\alpha, \frac{\theta}{\alpha}) \to (\alpha, \theta)$, it is straightforward to see that optimal $\alpha$ is at $+\infty$ and the separability condition reduces to

$$\delta_* := \left(\min_{\theta} \mathbb{E}\left[\left(-V\theta + \sqrt{1+\theta^2}g\right)_+^2\right]\right)^{-1}\,.$$

- **Example 2 ($\ell_2$ perturbation).** When $p = q = 2$, the set $\mathcal{S}$ becomes the intersection of $\ell_2$ ball of radius

$$R := \min\left(\alpha, \sqrt{\frac{1}{\varepsilon_0^2 \|\boldsymbol{\mu}\|_{\ell_2}^2} - \theta^2}\right)\,,$$

  with the hyperplane of dimension $d-1$ that is orthogonal to $\boldsymbol{\mu}$. Therefore $\omega(\alpha, \theta, \varepsilon_0) = \omega_s(\mathcal{S}) = \min\left(\alpha, \sqrt{\frac{1}{\varepsilon_0^2 V^2} - \theta^2}\right)$ and the separability threshold reduces to

$$\delta_* = \max_{\alpha\geq 0,\ \theta\leq \frac{1}{\varepsilon_0 V}} \frac{\min\left(\alpha^2, \frac{1}{\varepsilon_0^2 V^2} - \theta^2\right)}{\mathbb{E}\left[\left(1 - V\theta + \sqrt{\alpha^2+\theta^2}g\right)_+^2\right]}\,.$$

Note that the above ratio is decreasing in $\alpha$ over the range of $\alpha \geq \sqrt{\frac{1}{\varepsilon_0^2\|\boldsymbol{\mu}\|_{\ell_2}^2} - \theta^2}$. Therefore, the maximizer $\alpha$ should satisfy $\alpha \leq \sqrt{\frac{1}{\varepsilon_0^2\|\boldsymbol{\mu}\|_{\ell_2}^2} - \theta^2}$ and this further simplifies the expression for $\delta_*$ as follows

$$\delta_* = \max_{\alpha\geq 0,\, \alpha^2+\theta^2 \leq \frac{1}{\varepsilon_0^2 V^2}} \frac{\alpha^2}{\mathbb{E}\left[\left(1 - V\theta + \sqrt{\alpha^2+\theta^2}g\right)_+^2\right]}.$$

By the change of variable $(\alpha, \frac{\theta}{\alpha}) \to (\alpha, \theta)$, this can be written as:

$$\delta_* = \left(\min_{\alpha\geq 0,\theta,\, \alpha^2(1+\theta^2) \leq \frac{1}{\varepsilon_0^2 V^2}} \mathbb{E}\left[\left(\frac{1}{\alpha} - V\theta + \sqrt{1+\theta^2}g\right)_+^2\right]\right)^{-1}.$$

Since the inner function is decreasing in $\alpha$ it is minimized at $\alpha_* = \frac{1}{\varepsilon_0 V\sqrt{1+\theta^2}}$ which simplifies the separability threshold to the following:

$$\delta_* = \left(\min_{\theta} \mathbb{E}\left[\left((\varepsilon_0\sqrt{1+\theta^2} - \theta)V + \sqrt{1+\theta^2}g\right)_+^2\right]\right)^{-1}. \tag{4.4}$$

To the best of our knowledge, our paper is the first work that shows such a phase transition for robust separability in the adversarial setting. In the non-adversarial case, similar phase transitions have been shown for data separability (a.k.a interpolation threshold) [8, 49, 12]. More specifically, [8] derived separability threshold for a logistic link regression model. Similar phenomenon extends to other link functions, as characterized by [49], and also to Gaussian mixtures model [12]. Interestingly, our result specialized to the case where the adversary has no power (cf. Example 1) recovers the existing thresholds for Gaussian mixtures model.

We end this section by demonstrating that in addition to the examples above Assumption 3 holds for a fairly broad family of mean vectors. This is the subject of the next lemma. We defer the proof of this lemma to Appendix E.2.

**Assumption 4** *Suppose that the empirical distribution of the entries of $\sqrt{d}\boldsymbol{\mu}$ converges weakly to a distribution $\mathbb{P}_M$ on real line, with bounded $2^{nd}$ and $p^{th}$ moment ($\int x^2 \mathrm{d}\mathbb{P}_M(x) = \sigma_{M,2}^2 < \infty$, $\int |x|^p \mathrm{d}\mathbb{P}_M(x) = \sigma_{M,p}^p < \infty$).*

**Lemma 4.2** *Consider the asymptotic regime of $n \to \infty$ and $n/d \to \delta$, for some $\delta \in (0, \infty)$. Also, consider the function $J_q(\cdot;\cdot) : \mathbb{R} \times \mathbb{R}_{\geq 0} \mapsto \mathbb{R}_{\geq 0}$ defined by*

$$J_q(x;\lambda) = \min_u \frac{1}{2}(x-u)^2 + \lambda|u|^q. \tag{4.5}$$

*Then Assumption 4 implies Assumption 3 with*

$$\omega(\alpha,\theta,\varepsilon_0) = \min_{\lambda_0,\eta\geq 0,\nu} \sqrt{\delta}\left\{\frac{\nu^2}{2\eta}+\frac{1}{2\eta\delta}+\frac{\eta}{2}\alpha^2+\lambda_0(\varepsilon_0\sigma_{M,p})^{-q}\right\} - \eta\sqrt{\delta}\,\mathbb{E}\left[J_q\left(\frac{h}{\eta\sqrt{\delta}}-\left(\frac{\nu}{\eta}-\theta\right)\frac{M}{\sigma_{M,2}};\frac{\lambda_0}{\eta}\right)\right], \tag{4.6}$$

*where the expectation in the last line is taken with respect to the independent random variables $h\sim\mathsf{N}(0,1)$ and $M\sim\mathbb{P}_M$.*

4.2. *Precise characterization of SA and RA in the separable regime* In this section we precisely characterize the SA and RA of the classifier obtained as the limiting point of gradient descent on the loss (2.3) in the separable regime. As discussed in Proposition 3.2, the normalized iterations of gradient descent for the loss (2.3) converge to the max-margin classifier (F.2). Since $\mathsf{SA}(\boldsymbol{\theta})$ and $\mathsf{RA}(\boldsymbol{\theta})$ are only functions of the direction $\frac{\boldsymbol{\theta}}{\|\boldsymbol{\theta}\|_{\ell_2}}$, instead of studying the classifier obtained via GD iterations directly, we study the classification performance of the max-margin classifier.

Recall the function $J_q$ is given by (4.5), and define

$$\mathcal{J}(c_0,c_1;\lambda_0) = \mathbb{E}\left[J_q\left(\frac{c_0}{\sqrt{\delta}}h - c_1\frac{M}{\sigma_{M,2}};\lambda_0\sigma_{M,p}^q\right)\right], \tag{4.7}$$

where the expectation in the last line is taken with respect to the independent random variables $h\sim\mathsf{N}(0,1)$ and $M\sim\mathbb{P}_M$, per the setting of Assumption 4. Our characterization of $\mathsf{SA}$ and $\mathsf{RA}$ will be in terms of the function $\mathcal{J}$ as formalized in the next theorem.

**Theorem 4.3** *Consider a data set generated i.i.d. according to an isotropic Gaussian mixture data model per Section 2.2 and suppose the mean vector $\boldsymbol{\mu}$ obeys Assumptions 1 and 4. Also let $\widetilde{\boldsymbol{\theta}}^{\varepsilon}$ be the max margin solution per (F.2). If $\delta<\delta_*$, with $\delta_*$ given by (4.3), then in the asymptotic setting of Assumption 1 we have:*

*(a) The following convex-concave minimax scalar optimization has a bounded solution $(\alpha_*,\gamma_{0*},\theta_*,\beta_*,\lambda_{0*},\eta_*,\tilde{\eta}_*)$ with the minimization components $(\alpha_*,\gamma_{0*},\theta_*)$ unique:*

$$\min_{\alpha,\gamma_0\geq 0,\theta}\max_{\beta,\lambda_0,\eta\geq 0,\tilde{\eta}} \quad D_{\mathrm{s}}(\alpha,\gamma_0,\theta,\beta,\lambda_0,\eta,\tilde{\eta}), \quad \textit{where}$$

$$
\begin{aligned}
D_{\mathrm{s}}(\alpha,\gamma_0,\theta,\beta,\lambda_0,\eta,\tilde{\eta}) \;&= 2\left(1+\frac{\eta}{2\alpha}\right)^{-1}\mathcal{J}\left(\frac{\beta}{2},\frac{\tilde{\eta}}{2};\frac{\lambda_0}{q\gamma_0^{q-1}}\left(1+\frac{\eta}{2\alpha}\right)^{1-q}\right)\\
&\quad-\left(\frac{\beta^2}{\delta}+\tilde{\eta}^2\right)\frac{1}{4(1+\frac{\eta}{2\alpha})}-\frac{2\lambda_0}{q}\gamma_0-\frac{\eta\alpha}{2}-\tilde{\eta}\theta\\
&\quad+\beta\sqrt{\mathbb{E}\left[\left((1+\varepsilon_0\gamma_0-\theta\sigma_{M,2})+\alpha g\right)_+^2\right]},
\end{aligned}
\tag{4.8}
$$

*where the expectation in the last part is taken with respect to* $g\sim\mathsf{N}(0,1)$.

*(b) It holds in probability that*

$$
\lim_{n\to\infty}\frac{1}{\|\boldsymbol{\mu}\|_{\ell_2}}\langle\boldsymbol{\mu},\widetilde{\boldsymbol{\theta}}^{\varepsilon}\rangle=\theta_*\,,
\tag{4.9}
$$

$$
\lim_{n\to\infty}\left\|\widetilde{\boldsymbol{\theta}}^{\varepsilon}\right\|_{\ell_2}=\alpha_*\,,
\tag{4.10}
$$

$$
\lim_{n\to\infty}\|\boldsymbol{\mu}\|_{\ell_p}\left\|\widetilde{\boldsymbol{\theta}}^{\varepsilon}\right\|_{\ell_q}=\gamma_{0*}\,.
\tag{4.11}
$$

*(c) Furthermore, part part (b) combined with Lemma 2.1 imply the following limits hold in probability:*

$$
\lim_{n\to\infty}\mathsf{SA}(\widetilde{\boldsymbol{\theta}}^{\varepsilon})=\Phi\left(\sigma_{M,2}\frac{\theta_*}{\alpha_*}\right),
\tag{4.12}
$$

$$
\lim_{n\to\infty}\mathsf{RA}(\widetilde{\boldsymbol{\theta}}^{\varepsilon})=\Phi\left(-\frac{\varepsilon_0\gamma_{0*}}{\alpha_*}+\sigma_{M,2}\frac{\theta_*}{\alpha_*}\right).
\tag{4.13}
$$

Theorem 4.3 above provides us with a precise characterization of SA and RA and allows us to rigorously quantify the effect of adversary's manipulative power $\varepsilon_0$, mean vector $\boldsymbol{\mu}$, and scaling of dimensions $\delta$ on SA and RA. In particular, this theorem precisely characterizes the performance of the max margin classifier (and in turn the classifier GD converges to) in terms of the optimal solutions to a low-dimensional optimization problem, namely (4.8). It is worth noting that by part (b), $\theta_*$ is the asymptotic value of the projection of the estimator $\widetilde{\boldsymbol{\theta}}^{\varepsilon}$ along the direction of the class averages $\boldsymbol{\mu}$, and $\alpha_*$ represents the asymptotic value of the $\ell_2$ norm of the estimator. Therefore, the $\theta_*/\alpha_*$ term appearing in the SA and RA formulae corresponds to the correlation coefficient between the estimator $\widetilde{\boldsymbol{\theta}}^{\varepsilon}$ and the class averages $\boldsymbol{\mu}$.

While the optimization problem (4.8) may look quite complicated, we note that it is a convex-concave problem in a handful number of scalar variables and hence can be easily solved by a low-dimensional gradient descent/ascent rather fast and accurately. In addition, this low-dimensional optimization problem significantly simplifies for special cases of $p$. We discuss some of these cases, which are also of particular practical interest, in Sections 5.1 and 5.2.

4.3. *Precise characterization of SA and RA in non-separable regime* In this section we precisely characterize the SA and RA of the classifier obtained by running gradient descent on the loss (2.3) in the non-separable regime. Before we can state our main result we need the definition of the Moreau envelop.

**Definition 4.4 (Moreau envelope and expected Moreau envelope)** *The* Moreau envelope *or* Moreau-Yosida regularization *of a function $\ell$ is given by*

$$e_\ell(x;\mu) := \min_t \frac{1}{2\mu}(x-t)^2 + \ell(t)\,. \tag{4.14}$$

*We also define the* expected *Moreau envelope*

$$L(a,b,\mu) = \mathbb{E}[e_\ell(ag+b;\mu)]\,, \tag{4.15}$$

*where the expectation is taken with respect to $g \sim \mathsf{N}(0,1)$.*

We this definition in place we are now ready to state our main result in the non-separable regime.

**Theorem 4.5** *Consider a data set generated i.i.d. according to an isotropic Gaussian mixture data model per Section 2.2 and suppose the mean vector $\boldsymbol{\mu}$ obeys Assumption 4. Also let $\widehat{\boldsymbol{\theta}}^{\varepsilon}$ be the solution to optimization (2.3). If $\delta > \delta_*$, with $\delta_*$ given by (4.3), then in the asymptotic setting of Assumption 1 we have:*

*(a) The following convex-concave minimax scalar optimization has a bounded solution $(\theta_*, \alpha_*, \gamma_{0*}, \tau_{g*}, \beta_*, \tau_{h*})$ with the minimization components $(\alpha_*, \gamma_{0*}, \theta_*)$ unique:*

$$\min_{\theta, 0\le\alpha,\gamma_0,\tau_g} \max_{0\le\beta,\tau_h} D_{\rm ns}(\alpha,\gamma_0,\theta,\tau_g,\beta,\tau_h)$$

$$D_{\rm ns}(\alpha,\gamma_0,\theta,\tau_g,\beta,\tau_h) = \frac{\beta\tau_g}{2} + L\left(\sqrt{\alpha^2+\theta^2}, \sigma_{M,2}\theta - \varepsilon_0\gamma_0, \frac{\tau_g}{\beta}\right)$$

$$- \min_{\lambda_0\ge 0,\nu}\left[\frac{\alpha}{\tau_h}\left\{\frac{\beta^2}{2\delta} + \lambda_0\left(\frac{\gamma_0\tau_h}{\alpha}\right)^q + \frac{\nu^2}{2} - \mathcal{J}\left(\beta, \left(\frac{\tau_h\theta}{\alpha}+\nu\right); \lambda_0\right)\right\} + \frac{\alpha\tau_h}{2}\right]. \tag{4.16}$$

*(b) It holds in probability that*

$$\lim_{n\to\infty} \frac{1}{\|\boldsymbol{\mu}\|_{\ell_2}} \langle \boldsymbol{\mu}, \widehat{\boldsymbol{\theta}}^{\varepsilon}\rangle = \theta_*\,, \tag{4.17}$$

$$\lim_{n\to\infty} \left\|\mathsf{P}_{\boldsymbol{\mu}}^{\perp}\widehat{\boldsymbol{\theta}}^{\varepsilon}\right\|_{\ell_2} = \alpha_*\,, \tag{4.18}$$

$$\lim_{n\to\infty} \|\boldsymbol{\mu}\|_{\ell_p} \left\|\widehat{\boldsymbol{\theta}}^{\varepsilon}\right\|_{\ell_q} = \gamma_{0*}\,. \tag{4.19}$$

(c) *As a corollary of part (b) and Lemma 2.1, the following limits hold in probability:*

$$\lim_{n\to\infty} \mathsf{SA}(\widehat{\boldsymbol{\theta}}^{\varepsilon}) = \Phi\left(\frac{\sigma_{M,2}\theta_*}{\sqrt{\alpha_*^2+\theta_*^2}}\right), \tag{4.20}$$

$$\lim_{n\to\infty} \mathsf{RA}(\widehat{\boldsymbol{\theta}}^{\varepsilon}) = \Phi\left(\frac{-\varepsilon_0\gamma_{0*}+\sigma_{M,2}\theta_*}{\sqrt{\alpha_*^2+\theta_*^2}}\right). \tag{4.21}$$

It is worth noting that by part (b), $\theta_*$ is the asymptotic value of the projection of the estimator $\widehat{\boldsymbol{\theta}}^{\varepsilon}$ along the direction of the class averages $\boldsymbol{\mu}$. In addition,

$$\lim_{n\to\infty} \left\|\mathsf{P}_{\boldsymbol{\mu}}^{\perp}\widehat{\boldsymbol{\theta}}^{\varepsilon}\right\|_{\ell_2}^2 + \left\|\mathsf{P}_{\boldsymbol{\mu}}\widehat{\boldsymbol{\theta}}^{\varepsilon}\right\|_{\ell_2}^2 = \alpha_*^2+\theta_*^2$$

represents the asymptotic value of the squared $\ell_2$ norm of the estimator. Therefore, the $\theta_*/\sqrt{\alpha_*^2+\theta_*^2}$ term appearing in the SA and RA formulae corresponds to the correlation coefficient between the estimator $\widehat{\boldsymbol{\theta}}^{\varepsilon}$ and the class averages $\boldsymbol{\mu}$.

Theorem 4.5 complements the result of Theorem 4.3 by providing a precise characterization of SA and RA measures in the non-separable regime. In the remaining part of this section and also in the next section, we specialize our results to several specific choices of $p$ that are of particular practical interest.

**Remark 4.4** *In stating our results (Theorems 4.3 and 4.5), we are implicitly assuming the same variable $\varepsilon_0$ for both the perturbation level to the test data as well as the 'perceived' perturbation level used in the robust minimax estimator $\widehat{\boldsymbol{\theta}}^{\varepsilon}$. In principle, we can use different variable for the test perturbation level, say $\varepsilon_{0,\text{test}}$. The same results applies to this setting with minimal modifications; only in the* RA*formalue, cf. equations (4.13), (4.21) the variable $\varepsilon_0$ should be replaced by $\varepsilon_{0,\text{test}}$.*

**5. Results for special cases of $p$** In this section we discuss the implications of our main results for the special cases of $\ell_p$ perturbations with $p=2$ in Section 5.1, $p=\infty$ in Section 5.2, and $p=1$ in Section 5.3. We refer to Appendix D for the proofs of theorems and corollaries stated in this section.

5.1. *Results for $\ell_2$ perturbation* We begin with stating our results for $\ell_p$ perturbation with $p=2$. This result can be viewed as a corollary of Theorem 4.1, Theorem 4.3, and Theorem 4.5 specializing our main result for $p=2$.

**Corollary 5.1** *Consider a data set generated i.i.d. according to an isotropic Gaussian mixture data model per Section 2.2 and suppose the mean vector $\boldsymbol{\mu}$ obeys Assumptions 4. Then in the asymptotic setting of Assumption 1 we have:*

*(a) The separability threshold $\delta_*$ is given by*

$$\delta_* = \left(\min_{\theta} \ \mathbb{E}\left[\left((\varepsilon_0\sqrt{1+\theta^2}-\theta)V + \sqrt{1+\theta^2}g\right)_+^2\right]\right)^{-1} . \tag{5.1}$$

*(b) In the separable regime where $\delta < \delta_*$, the followings hold in probability for the max margin solution $\widetilde{\boldsymbol{\theta}}^{\varepsilon}$ (see (F.2)):*

$$\lim_{n\to\infty} \mathsf{SA}(\widetilde{\boldsymbol{\theta}}^{\varepsilon}) = \Phi\left(\sigma_{M,2}\frac{\theta_*}{\alpha_*}\right), \quad \lim_{n\to\infty} \mathsf{RA}(\widetilde{\boldsymbol{\theta}}^{\varepsilon}) = \Phi\left(\frac{\theta_*}{\alpha_*}\sigma_{M,2} - \varepsilon_0\sigma_{M,2}\right), \tag{5.2}$$

*where*

$$\alpha_* = \left(\tilde{\alpha}_*^{-1} - \varepsilon_0\sigma_{M,2}\right)^{-1}, \quad \theta_* = u_*\alpha_* . \tag{5.3}$$

*Here, $(\tilde{\alpha}_*, u_*)$ the solution to the following problem:*

$$\begin{aligned} &\min_{\tilde{\alpha}\geq 0, u} \ \tilde{\alpha}^2 \\ &\text{subject to} \quad 1 \geq u^2 + \delta\,\mathbb{E}\left[\left(\frac{1}{\tilde{\alpha}} - u\sigma_{M,2} + g\right)_+^2\right], \end{aligned} \tag{5.4}$$

*with expectation taken with respect to $g \sim \mathsf{N}(0,1)$.*

*(c) In the non-separable regime where $\delta > \delta_*$, the followings hold in probability for the optimal solution $\widehat{\boldsymbol{\theta}}^{\varepsilon}$ of (2.3):*

$$\lim_{n\to\infty} \mathsf{SA}(\widehat{\boldsymbol{\theta}}^{\varepsilon}) = \Phi\left(\frac{\sigma_{M,2}\theta_*}{\sqrt{\alpha_*^2+\theta_*^2}}\right), \tag{5.5}$$

$$\lim_{n\to\infty} \mathsf{RA}(\widehat{\boldsymbol{\theta}}^{\varepsilon}) = \Phi\left(\frac{\theta_*}{\sqrt{\alpha_*^2+\theta_*^2}}\sigma_{M,2} - \varepsilon_0\sigma_{M,2}\right). \tag{5.6}$$

*where $(\alpha_*, \theta_*, \beta_*)$ is the bounded solution of the following convex-concave minimax scalar optimization problem with the minimization components $(\alpha_*, \theta_*)$ unique:*

$$\begin{aligned} &\max_{0\leq\beta} \ \min_{\theta, 0\leq\alpha} \ D_{\mathrm{ns}}(\alpha,\theta,\beta) \\ &D(\alpha,\theta,\beta) = L\left(\sqrt{\alpha^2+\theta^2}, \sigma_{M,2}\theta - \varepsilon_0\sqrt{\alpha^2+\theta^2}, \frac{\alpha}{\beta\sqrt{\delta}}\right) - \frac{\alpha\beta}{2\sqrt{\delta}} . \end{aligned} \tag{5.7}$$

The corollary above precisely characterizes the behavior of the classifier that gradient descent converges to in terms of low-dimensional optimization problems ((5.4) in the separable regime and (5.7) in the non-separable regime).

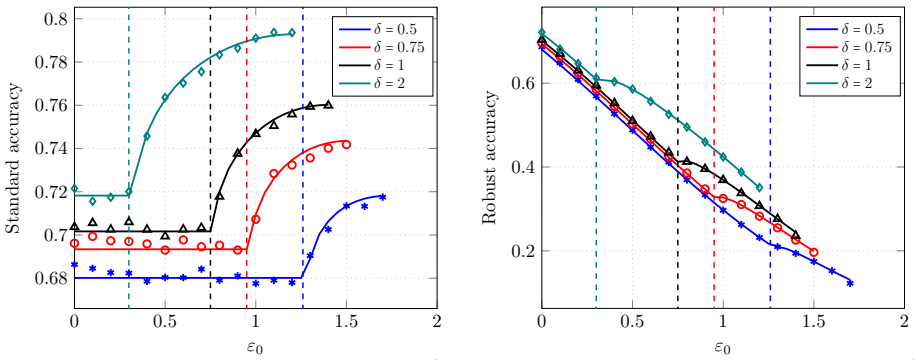
Recall that the term $\theta_*/\alpha_*$ in the separable regime and the term $\theta_*/\sqrt{\alpha_*^2+\theta_*^2}$ in the non-separable regime correspond to the correlation coefficient between the robust minimax estimator and the classes average $\boldsymbol{\mu}$. As we will see in Figure 4, the standard accuracy is decreasing in $1/\delta$, for any fixed $\varepsilon_0$, which equivalently indicates that the correlation between the estimator and $\boldsymbol{\mu}$ is monotone increasing in the sample-to-dimension ratio $\delta$.

As we will see in the coming sections, SA and RA curves have a highly non-trivial behavior which also strongly depend on the choice of $p$. This necessitate a rigorous theory (such as the above) that can precisely predict these curves. To better understand the implications and consequences of this result we focus on its various predictions. Specifically, we find the global optima of the two low-dimensional optimization problems via simple gradient descent/ascent and use it to calculate the corresponding SA and RA based on (5.2) and (5.5). We also verify these theoretical predictions with the performance of gradient descent on the loss (2.3) with a polyak/approximate polyak step size in the separable/non-separable regimes.[3]

We plot the theoretically predicted standard and robust accuracy versus the adversary's power $\varepsilon_0$ together with the corresponding empirical results in Figure 3 (a) and (b). The solid lines depict theoretical predictions with the dots representing the empirical performance of gradient descent with the algorithmic settings discussed above. The data set is generated according to a Gaussian Mixture Model per Section 2.2 with $\boldsymbol{\mu} \in \mathbb{R}^d$ consisting of i.i.d. $\mathcal{N}\left(0, \frac{1}{d}\right)$ entries with dimension $d = 400$. Each dot represents the average of 100 trials. These figures demonstrate that even for moderate dimension sizes our theoretical prediction is a near perfect match with the empirical performance of gradient descent. We note that when $\varepsilon_0$ is sufficiently large then the adversarially trained model $\widehat{\boldsymbol{\theta}}^{\varepsilon}$ becomes zero due to the large regularization in the argument of loss function in (2.3) and SA and RA measures are not defined. The curves are plotted up to that $\varepsilon_0$.

An intriguing observation of Corollary 5.1 is that in the separable regime in the case of $p = 2$, the standard accuracy does not depend on $\varepsilon_0$. In other words, adversarial training has no effect on the performance on benign unperturbed data. The robust accuracy, however is decreasing in $\varepsilon_0$. Figure 3 (a) and (b) also verify this predicted behavior and capture the effect of the adversary's power $\varepsilon_0$ on standard and robust accuracy. In the separable regime, SA is flat which implies that adversarial training has no effect on

[3]Specifically we run gradient descent iterations of the form $\boldsymbol{\theta}_{\tau+1} = \boldsymbol{\theta}_\tau - \alpha_\tau \nabla \mathcal{L}(\boldsymbol{\theta}_\tau)$ on (2.3) with a Polyak step size $\alpha_\tau = \frac{\mathcal{L}(\boldsymbol{\theta}_\tau)}{\|\nabla \mathcal{L}(\boldsymbol{\theta}_\tau)\|_{\ell_2}^2}$ in the separable regime and an approximate Polyak step size $\alpha_\tau = \frac{\mathcal{L}(\boldsymbol{\theta}_\tau) - \min_{0\le t\le \tau} \mathcal{L}(\boldsymbol{\theta}_t) + \frac{\gamma}{\tau}}{\|\nabla \mathcal{L}(\boldsymbol{\theta}_\tau)\|_{\ell_2}^2}$ in the non-separable regime.



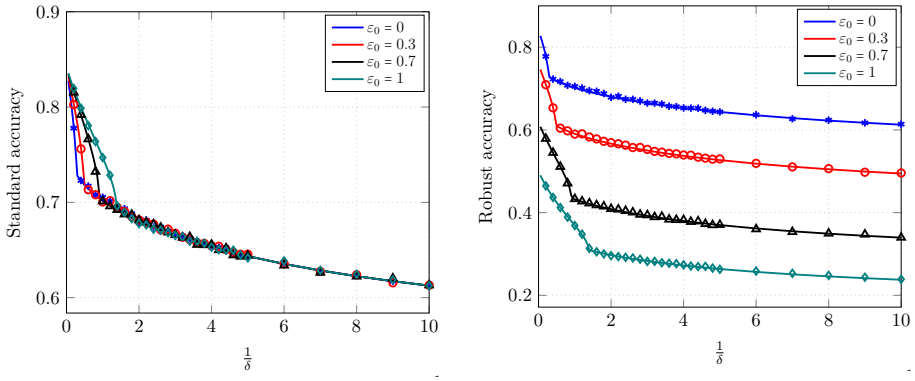

(a) Standard accuracy (b) Robust accuracy

Fig 3: Depiction of standard and robust accuracies as a function of the adversary's normalized power $\varepsilon_0$ with $\ell_2$ ($p = 2$) perturbation for different values of $\delta$. Solid curves are theoretical predictions and dots are results obtained based on gradient descent on the robust objective (2.3). The dashed lines depict the separability threshold for that $\delta$. Each dot represents the average of 100 trials. The data set is generated according to a Gaussian Mixture Model per Section 2.2 with $\boldsymbol{\mu} \in \mathbb{R}^d$ consisting of i.i.d. $\mathcal{N}\left(0, \frac{1}{d}\right)$ entries with dimension $d = 400$.

standard accuracy (or the generalization error on unperturbed data). However, adversarial training does affect RA because now the trained model is used to classify the adversarially perturbed test data.

In the non-separable regime, we observe that adversarial training helps with improving the standard accuracy! Further, such positive impact is observed for all choice of $\delta$ with a rather robust trend. Note that this behavior is significantly different from a regression setting where adversarial training first improves with the standard accuracy but then there is a turning point beyond which the standard accuracy will decrease as $\varepsilon_0$ grows. We refer to [30, Figure 3] and discussion therein for more details on a regression setting. Moreover, as depicted in Figure 3(b) we see that RA always declines as adversary gets more powerful (i.e., $\varepsilon_0$ grows) as expected.

Next in Figure 4, we plot SA and RA versus dimension-to-sample ratio $\frac{1}{\delta} = \frac{d}{n}$, which is a measure of model complexity, for several values of $\varepsilon_0$. It has been shown that the standard risk (which amounts to $1 - \mathsf{SA}$ in our setting) as a function of model complexity $\frac{1}{\delta}$ undergoes a double-descent behavior for various learning models [6, 5, 24]. Specifically, the risk depicts a




(a) Standard accuracy (b) Robust accuracy

Fig 4: Depiction of standard and robust accuracies as a function of dimension-to-sample ratio $\frac{1}{\delta} = \frac{d}{n}$, which is a measure of model complexity, for several values of $\varepsilon_0$ with $\ell_2$ ($p = 2$) perturbation, under a similar setting as in Figure 3.

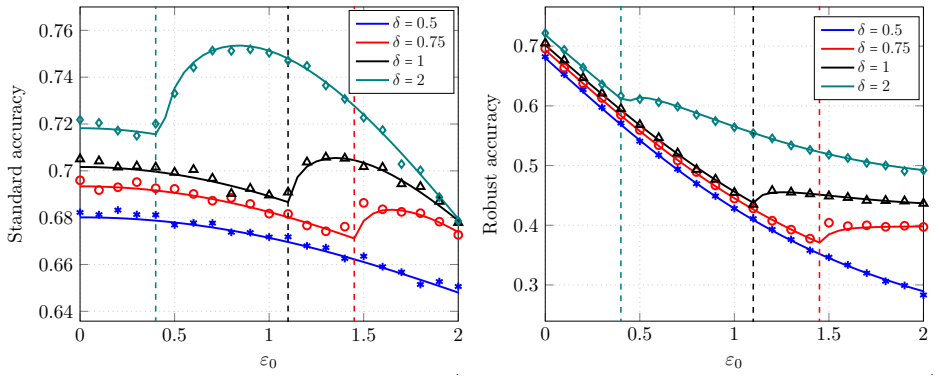
U-shape before the interpolation threshold (separability threshold in binary classification) and then starts to decline afterwards. Interestingly, for the current setting of experiments here we do not observe such double descent behavior and the standard accuracy always decreases as $\frac{1}{\delta}$ grows, albeit at different rates in the separable and non-separable regimes.[4]

5.2. *Results for $\ell_\infty$ perturbation* For the case of $\ell_\infty$ perturbation ($p = \infty$ and $q = 1$), Theorem 4.1, Theorem 4.3, and Theorem 4.5 do not substantially simplify. However, we can calculate the function $J_q$ defined by (4.5) in closed form. In this case $J_q$ becomes the Huber function given by

$$J_1(x, \lambda) = \begin{cases} \lambda|x| - \frac{\lambda^2}{2} & |x| \geq \lambda \\ \frac{x^2}{2} & |x| \leq \lambda \end{cases}$$

Using Theorems 4.1, 4.3, and 4.5 with this closed form for $J_q$, in Figure 5, we again depict our theoretical predictions for standard and robust accuracy as well as the empirical performance of gradient descent as a function of the

[4]It is worth noting that the double descent phenomenon has been observed for binary classification in a non-adversarial setting with model misspecification. In such a model the learner observes only a subset $S \subset [d]$ of size $p$ of the covariates with $d/n \to \zeta \geq 1$ and $p/n \to \kappa \in (0, \zeta]$ (see [12] for further details). Our theoretical analysis can in principle be used to analyze such a setting, however we do not pursue this direction in this paper.



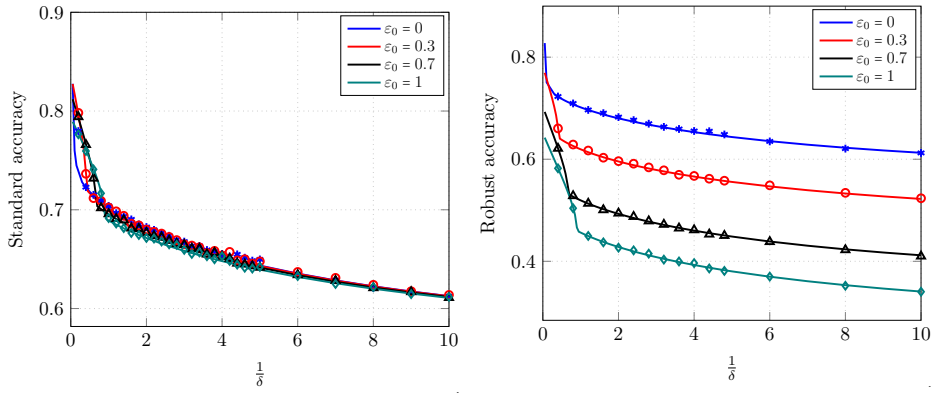

(a) Standard accuracy (b) Robust accuracy

Fig 5: Depiction of standard and robust accuracies as a function of $\varepsilon_0$ with $\ell_\infty$ ($p = \infty$) perturbation for different values of $\delta$, and under a similar setting as in Figure 3.

adversary's normalized power for various values of $\delta$. As in the $p = 2$ case our theoretical predictions is very accurate even for moderate dimensions $d$.

More specifically, Figure 5(a) depicts the standard accuracy (SA) versus the adversary's normalized power. Similar to our $p = 2$ results the data set is generated according to a Gaussian Mixture Model per Section 2.2 with $\boldsymbol{\mu} \in \mathbb{R}^d$ consisting of i.i.d. $\mathcal{N}\left(0, \frac{1}{d}\right)$ entries with dimension $d = 400$ and each data points represents the average of 100 trials. In the case of $p = \infty$ however, we do not use the scaling $\varepsilon = \varepsilon_0 \|\boldsymbol{\mu}\|_{\ell_\infty}$ as $\left\|\sqrt{d}\boldsymbol{\mu}\right\|_{\ell_\infty}$ grows with $\sqrt{\log d}$ and therefore violates Assumption 4. Instead we shall use a slightly different scaling of $\varepsilon = \frac{\varepsilon_0}{\sqrt{d}}$. In the separable regime, we see that adversarial training hurts the standard accuracy. However, in the non-separable regime, the standard accuracy starts increasing indicating that adversarial training is improving the standard accuracy. Furthermore, after some value of $\varepsilon_0$, which interestingly shifts with $\delta$, the standard accuracy starts to go down as $\varepsilon_0$ grows.[5] We note that this behavior is rather counterintuitive and very different from the $p = 2$ case, further highlighting the need for a precise theory that can predict such nuanced behavior. Figure 5(b) shows the robust accuracy RA versus $\varepsilon_0$ for various values of $\delta$. In the separable regime, we observe a similar trend for all $\delta$, namely RA decreases at an almost linear rate. In the non-separable regime though we have different trends depending on the value of $\delta$.

[5]Note that for $\delta = 0.5$, we are in the separable regime over the entire range $[0, \varepsilon_0]$.




(a) Standard accuracy (b) Robust accuracy

Fig 6: Depiction of standard and robust accuracies as a function of dimension-to-sample ratio $\frac{1}{\delta} = \frac{d}{n}$, which is a measure of model complexity, for several values of $\varepsilon_0$ with $\ell_\infty$ ($p = \infty$) perturbation, under a similar setting as in Figure 3.

Finally, in Figure 6 we depict the effect of overparameterization $\frac{1}{\delta}$ on SA and RA. We observe a similar pattern as in the case of $p = 2$. In particular, we do not observe a double descent behavior and the standard accuracy always decreases as $\frac{1}{\delta}$ grows, albeit at different rates in the separable and non-separable regimes.

5.3. *Results for $\ell_1$ perturbation* Our characterization of SA and RA given by Theorem 4.3, for separable regime, and by Theorem 4.5, for non-separable regime involve the function $\mathcal{J}$ defined by (4.7) which in turn depends on the function $J_q$ given by (4.5). However, $J_q$ is only defined for finite $q$ and therefore the case of $p = 1$, $q = \infty$ is not directly covered by our results in Section 4. That said, a very similar analysis can be used to characterize SA and RA in this case. We formalize our results for this case in the next theorem.

**Theorem 5.2** *Consider a data set generated i.i.d. according to an isotropic Gaussian mixture data model per Section 2.2 and suppose the mean vector $\boldsymbol{\mu}$ obeys Assumptions 4. Also define*

$$f(c_0, c_1; t_0) = \frac{1}{2}\,\mathbb{E}\left[\mathsf{ST}\left(\frac{c_0}{\sqrt{\delta}}h - c_1\frac{M}{\sigma_{M,2}}; \frac{t_0}{\sigma_{M,1}}\right)^2\right], \tag{5.8}$$

*where* $\mathsf{ST}(x;a) := sgn(x)\,(|x|-a)_+$ *is the soft-thresholding function. Then in the asymptotic setting of Assumption 1 we have:*

*(a) The separability threshold* $\delta_*$ *is given by*

$$\delta_* := \min_{\alpha\geq 0,\theta} \frac{\omega\left(\alpha,\theta,\varepsilon_0\right)^2}{\mathbb{E}\left[\left(1-V\theta+\sqrt{\alpha^2+\theta^2}g\right)_+^2\right]}$$

(5.9)

$$\text{with}\quad \omega\left(\alpha,\theta,\varepsilon_0\right) := \min_{\eta\geq 0,\nu} \sqrt{\delta}\left\{\frac{\nu^2}{2\eta}+\frac{1}{2\eta\delta}+\frac{\eta}{2}\alpha^2-\eta f\left(\frac{1}{\eta},\frac{\nu}{\eta}-\theta;\frac{1}{\varepsilon_0}\right)\right\}.$$

*(b) In the separable regime where* $\delta<\delta_*$*, the followings hold in probability for the max margin solution* $\widetilde{\boldsymbol{\theta}}^{\varepsilon}$ *(see (F.2)):*

$$\lim_{n\to\infty} \mathsf{SA}(\widetilde{\boldsymbol{\theta}}^{\varepsilon}) = \Phi\left(\sigma_{M,2}\frac{\theta_*}{\alpha_*}\right), \tag{5.10}$$

$$\lim_{n\to\infty} \mathsf{RA}(\widetilde{\boldsymbol{\theta}}^{\varepsilon}) = \Phi\left(-\frac{\varepsilon_0\gamma_{0*}}{\alpha_*}+\sigma_{M,2}\frac{\theta_*}{\alpha_*}\right). \tag{5.11}$$

*Here,* $(\alpha_*,\gamma_{0*},\theta_*)$ *are the unique minimization component of the following convex-concave minimax scalar optimization with bounded solution* $(\alpha_*,\gamma_{0*},\theta_*,\beta_*,\eta_*,\tilde{\eta}_*)$*.*

$$\min_{\alpha,\gamma_0\geq 0,\theta}\max_{\beta,\eta\geq 0,\tilde{\eta}} D_{\mathrm{s}}(\alpha,\gamma_0,\theta,\beta,\eta,\tilde{\eta}),\quad \textit{where}$$

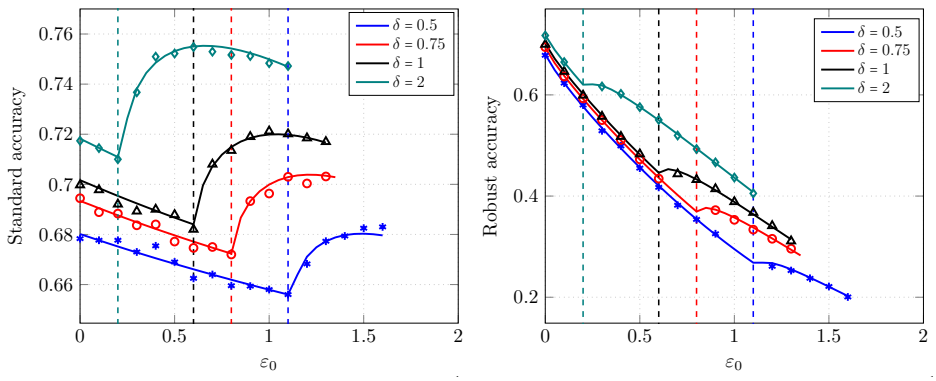
$$\begin{aligned} D_{\mathrm{s}}(\alpha,\gamma_0,\theta,\beta,\lambda_0,\eta,\tilde{\eta}) \;=\; \min_{\alpha,\gamma_0\geq 0,\theta}\max_{\beta,\eta\geq 0,\tilde{\eta}} \quad & \frac{1}{2(1+\frac{\eta}{2\alpha})} f\left(\beta,\tilde{\eta};2\gamma_0\left(1+\frac{\eta}{2\alpha}\right)\right) \\ & -\left(\frac{\beta^2}{\delta}+\tilde{\eta}^2\right)\frac{1}{4(1+\frac{\eta}{2\alpha})}-\frac{\eta\alpha}{2}-\tilde{\eta}\theta \\ & +\beta\sqrt{\mathbb{E}\left[\left((1+\varepsilon_0\gamma_0-\theta\sigma_{M,2})+\alpha g\right)_+^2\right]}, \end{aligned} \tag{5.12}$$

*with expectation taken with respect to* $g\sim\mathsf{N}(0,1)$*.*

*(c) In the non-separable regime where* $\delta>\delta_*$*, the followings hold in probability the optimal solution* $\widehat{\boldsymbol{\theta}}^{\varepsilon}$ *of (2.3):*

$$\lim_{n\to\infty} \mathsf{SA}(\widehat{\boldsymbol{\theta}}^{\varepsilon}) = \Phi\left(\frac{\sigma_{M,2}\theta_*}{\sqrt{\alpha_*^2+\theta_*^2}}\right), \tag{5.13}$$

$$\lim_{n\to\infty} \mathsf{RA}(\widehat{\boldsymbol{\theta}}^{\varepsilon}) = \Phi\left(\frac{-\varepsilon_0\gamma_{0*}+\sigma_{M,2}\theta_*}{\sqrt{\alpha_*^2+\theta_*^2}}\right). \tag{5.14}$$




(a) Standard accuracy (b) Robust accuracy

Fig 7: Depiction of standard and robust accuracies as a function of $\varepsilon_0$ with $\ell_1$ ($p = 1$) perturbation for different values of $\delta$, under a similar setting as in Figure 3.

*Here,* $(\alpha_*, \gamma_{0*}, \theta_*)$ *are the unique minimization components of the following convex-concave minimax scalar optimization with bounded solution* $(\theta_*, \alpha_*, \gamma_{0*}, \tau_{g*}, \beta_*, \tau_{h*})$.

$$\min_{\theta, 0\le\alpha,\gamma_0,\tau_g} \max_{0\le\beta,\tau_h} D_{\rm ns}(\alpha, \gamma_0, \theta, \tau_g, \beta, \tau_h)$$

$$D_{\rm ns}(\alpha, \gamma_0, \theta, \tau_g, \beta, \tau_h) = \frac{\beta\tau_g}{2} + L\left(\sqrt{\alpha^2+\theta^2}, \sigma_{M,2}\theta - \varepsilon_0\gamma_0, \frac{\tau_g}{\beta}\right) - \min_{\nu}\left[\frac{\alpha}{\tau_h}\left\{\frac{\beta^2}{2\delta} + \frac{\nu^2}{2} - f\left(\beta, \frac{\tau_h\theta}{\alpha} + \nu; \frac{\gamma_0\tau_h}{\alpha}\right)\right\} + \frac{\alpha\tau_h}{2}\right]. \tag{5.15}$$

In Figure 7, we again depict our theoretical predictions for standard and robust accuracy as well as the empirical performance of gradient descent as a function of the adversary's normalized power for various values of $\delta$. We note however that in this case we do not actually run gradient descent in our simulations as $p = 1$ corresponds to $q = +\infty$ and GD convergence is extremely slow since the gradient only has one non-zero entry. Therefore, for our empirical simulations we use CVX, a package for specifying and solving convex programs [22], in the non-separable regime which given the uniqueness of the global optima yields the same answer as GD. Similarly, in the separable regime we use (F.2) which based on Proposition 3.2 is the direction GD eventually converges to. We observe that as in the $p = 2$ and $p = +\infty$ cases our theoretical predictions are very accurate even for moderate dimensions $d$.

More specifically, Figure 7(a) depicts the standard accuracy (SA) versus the adversary's normalized power. Similar to our $p = +\infty$ results the data set is generated according to a Gaussian Mixture Model per Section 2.2 with $\boldsymbol{\mu} \in \mathbb{R}^d$ consisting of i.i.d. $\mathcal{N}\left(0, \frac{1}{d}\right)$ entries with dimension $d = 400$ and each data points represents the average of 100 trials. In the separable regime, we see that adversarial training hurts the standard accuracy. However, in the non-separable regime, the standard accuracy starts increasing indicating that adversarial training is improving the standard accuracy. Furthermore, after some value of $\varepsilon_0$, which interestingly shifts with $\delta$, the standard accuracy starts to go down as $\varepsilon_0$ grows.[6] We note that this behavior is rather counterintuitive and very different from the $p = 2$ case but somewhat similar to the $p = +\infty$ case. This again highlights the need for a precise theory that can predict such nuanced behavior. Figure 7(b) shows the robust accuracy RA versus $\varepsilon_0$ for various values of $\delta$. In the separable regime, we observe a similar trend for all $\delta$, namely RA decreases at an almost linear rate. In the non-separable regime though we have different trends depending on the value of $\delta$.

**6. Extension to anisotropic Gaussians** In this section we extend our results to Gaussian distributions with general covariance matrices that obey a certain spiked covariance assumption stated below.

**Assumption 5** *(Spiked covariance)* $\boldsymbol{\mu}$ *is an eigenvector of* $\boldsymbol{\Sigma}$ *with eigenvalue* $a^2$*, i.e,* $\boldsymbol{\Sigma}\boldsymbol{\mu} = a^2\boldsymbol{\mu}$.

Similar spiked covariance models have been used to model data in a number of statistical problems, including matrix denoising and structured learning [32, 16], sparse PCA [13], synchronization and clustering [28].

To extend our results in Section 4 to the anisotropic case we also need to generalize the definition of the set $\mathcal{S}$ as follows:

$$\mathcal{S}(\alpha, \theta, \varepsilon_0, \boldsymbol{\mu}) := \left\{ \boldsymbol{z} \in \mathbb{R}^d : \quad \boldsymbol{z}^T \widetilde{\boldsymbol{\mu}} = 0,\ \|\boldsymbol{z}\|_{\ell_2} = \alpha,\ \left\| \boldsymbol{\Sigma}^{-1/2}\boldsymbol{z} + \theta\widetilde{\boldsymbol{\mu}} \right\|_{\ell_q} \le \frac{1}{\varepsilon_0 \|\boldsymbol{\mu}\|_{\ell_p}} \right\}.$$

We are now ready to state our main results in the anisotropic case. We start by the separability threshold which generalizes Theorem 4.1.

**Theorem 6.1** *Consider a data set generated i.i.d. according to an anisotropic Gaussian mixture data model per Section 2.2 with a spiked covariance per Assumption 5. Also suppose the mean vector* $\boldsymbol{\mu}$ *and covariance matrix* $\boldsymbol{\Sigma}$ *obey*

[6] Note that for $\delta = 0.5$, we are in the separable regime over the entire range $[0, \varepsilon_0]$.

*Assumptions 2 and 3. Also define*

$$\delta_* := \min_{\alpha\geq 0,\theta} \frac{\omega\left(\alpha,\theta,\varepsilon_0\right)^2}{\mathbb{E}\left[\left(1-V\theta+\sqrt{\alpha^2+a^2\theta^2}g\right)_+^2\right]}, \tag{6.1}$$

*where the expectation is taken with respect to $g\sim\mathsf{N}(0,1)$. Then, under the asymptotic setting of Assumption 1, for $\delta<\delta_*$ the data are $(\varepsilon,q)$- separable with high probability and for $\delta>\delta_*$, the data are non-separable, with high probability. Namely,*

$$\begin{aligned}\delta<\delta_* &\Rightarrow \lim_{n\to\infty}\mathbb{P}(\textit{data is } (\varepsilon,q)\textit{-separable}) = 1\,,\\ \delta>\delta_* &\Rightarrow \lim_{n\to\infty}\mathbb{P}(\textit{data is } (\varepsilon,q)\textit{-separable}) = 0\,.\end{aligned}$$

Our next theorem precisely characterizes SA and RA in the separable regime and generalizes Theorem 4.3 to the anisotropic case. Before proceeding to state the theorem we need to establish some definitions and assumptions.

**Definition 6.2** *For a given matrix $\boldsymbol{A}\succeq 0$ and a function $f$, we define the* weighted Moreau envelope *of $f$ as follows:*

$$e_{f,\boldsymbol{A}}(\boldsymbol{x};\lambda) := \min_{\boldsymbol{v}}\frac{1}{2}\|\boldsymbol{x}-\boldsymbol{v}\|_{\boldsymbol{A}}^2+\lambda f(\boldsymbol{v})$$

*When $\boldsymbol{A}=\boldsymbol{I}$, we recover the (scaled) classical Moreau envelope. We denote by $e_{q,\boldsymbol{\Sigma}}$ the weighted Moreau envelope corresponding to $\|\cdot\|_{\ell_q}^q$ function.*

**Assumption 6** *For the sequence of instances $\{\boldsymbol{\Sigma}(n),\boldsymbol{\mu}(n),d(n)\}_{n\in\mathbb{N}}$ indexed by $n$, we assume that:*

*(a) The following (in probability) limit exists for any scalars $c_0,c_1,\lambda_0,\eta\in\mathbb{R}_+$:*

$$\mathsf{F}(c_0,c_1;b_0,b_1) := \lim_{n\to\infty} e_{q,\boldsymbol{I}+b_0\boldsymbol{\Sigma}}\left((\boldsymbol{I}+b_0\boldsymbol{\Sigma})^{-1}\left\{\frac{c_0}{2\sqrt{n}}\boldsymbol{\Sigma}^{1/2}\mathsf{P}_{\boldsymbol{\mu}}^{\perp}\boldsymbol{h}-\frac{c_1}{2}\widetilde{\boldsymbol{\mu}}\right\};b_1\|\boldsymbol{\mu}\|_{\ell_p}^q\right).$$

*(b) The empirical distribution of eigenvalues of $\boldsymbol{\Sigma}$ converges weakly to a distribution $\rho$ with Stieltjes transform $S_\rho(z):=\int\frac{\rho(t)}{z-t}\mathrm{d}t$.*

With these definitions and assumptions in place we are ready to state our result in the separable regime.

**Theorem 6.3** *Consider a data set generated i.i.d. according to an anisotropic Gaussian mixture data model per Section 2.2 with a spiked covariance per Assumption 5. Also suppose the mean vector $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$ obey*

*Assumptions 2, 3, and 6. Also let $\widetilde{\boldsymbol{\theta}}^{\varepsilon}$ be the max margin solution per (F.2). If $\delta < \delta_*$, with $\delta_*$ given by (4.3), then in the asymptotic setting of Assumption 1 we have:*

*(a) The following convex-concave minimax scalar optimization problem has bounded solution $(\alpha_*, \gamma_{0*}, \theta_*, \beta_*, \lambda_{0*}, \eta_*, \tilde{\eta}_*)$ with the minimization components $(\alpha_*, \gamma_{0*}, \theta_*)$ unique:*

$$\min_{\alpha,\gamma_0\geq 0,\theta}\ \max_{\beta,\lambda_0,\eta\geq 0,\tilde{\eta}} \quad D_{\mathrm{s}}(\alpha,\gamma_0,\theta,\beta,\lambda_0,\eta,\tilde{\eta}), \quad \text{where}$$

$$\begin{aligned} D_{\mathrm{s}}(\alpha,\gamma_0,\theta,\beta,\lambda_0,\eta,\tilde{\eta}) \ &= 2\mathsf{F}\left(\beta,\tilde{\eta};\frac{\eta}{2\alpha},\frac{\lambda_0}{q\gamma_0^{q-1}}\right) - \frac{\beta^2\alpha}{2\delta\eta}\left(1+\frac{2\alpha}{\eta}S_\rho\left(-\frac{2\alpha}{\eta}\right)\right) \\ &\quad - \frac{2\lambda_0}{q}\gamma_0 - \frac{\eta\alpha}{2} - \tilde{\eta}\theta \\ &\quad - \frac{\tilde{\eta}^2}{4(1+\frac{\eta}{2\alpha}a^2)} + \beta\sqrt{\mathbb{E}\left[\left((1+\varepsilon_0\gamma_0-\theta V)+\alpha g\right)_+^2\right]}, \end{aligned} \tag{6.2}$$

*with expectation in last part taken with respect to $g \sim \mathsf{N}(0,1)$.*

*(b) It holds in probability that*

$$\lim_{n\to\infty} \frac{1}{\|\boldsymbol{\mu}\|_{\ell_2}} \langle \boldsymbol{\mu}, \widetilde{\boldsymbol{\theta}}^{\varepsilon} \rangle = \theta_* \,, \tag{6.3}$$

$$\lim_{n\to\infty} \|\widetilde{\boldsymbol{\theta}}^{\varepsilon}\|_{\ell_2} = \alpha_* \,, \tag{6.4}$$

$$\lim_{n\to\infty} \|\boldsymbol{\mu}\|_{\ell_p} \|\widetilde{\boldsymbol{\theta}}^{\varepsilon}\|_{\ell_q} = \gamma_{0*} \,. \tag{6.5}$$

*(c) As a corollary of part (b) and Lemma 2.1, the following limits hold in probability:*

$$\lim_{n\to\infty} \mathsf{SA}(\widetilde{\boldsymbol{\theta}}^{\varepsilon}) = \Phi\left(V\frac{\theta_*}{\alpha_*}\right), \tag{6.6}$$

$$\lim_{n\to\infty} \mathsf{RA}(\widetilde{\boldsymbol{\theta}}^{\varepsilon}) = \Phi\left(-\frac{\varepsilon_0\gamma_{0*}}{\alpha_*} + V\frac{\theta_*}{\alpha_*}\right). \tag{6.7}$$

Next we turn our attention to characterizing SA and RA on the non-separable regime. To state result we need an additional assumption

**Assumption 7** *For the sequence of instances $\{\boldsymbol{\Sigma}(n), \boldsymbol{\mu}(n), p(n)\}_{n\in\mathbb{N}}$ indexed by $n$, we assume that the following (in probability) limit exists for any scalars $c_0, c_1 \in \mathbb{R}_+$ and $\lambda_0 \in \mathbb{R}$:*

$$\mathsf{E}(c_0,c_1;\lambda_0) := \lim_{n\to\infty} e_{q,\boldsymbol{\Sigma}}\left(\frac{c_0}{\sqrt{n}}\boldsymbol{\Sigma}^{-1/2}\boldsymbol{h} - c_1\widetilde{\boldsymbol{\mu}}; \lambda_0\|\boldsymbol{\mu}\|_{\ell_p}^q\right), \tag{6.8}$$

*where we recall $\widetilde{\boldsymbol{\mu}} = \boldsymbol{\mu}/\|\boldsymbol{\mu}\|_{\ell_2}$.*

Our next theorem generalizes Theorem 4.5 to anisotropic case.

**Theorem 6.4** *Consider a data set generated i.i.d. according to an anisotropic Gaussian mixture data model per Section 2.2 with a spiked covariance per Assumption 5. Also suppose the mean vector $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$ obey Assumptions 2, 3, and 7. Also let $\widehat{\boldsymbol{\theta}}^{\varepsilon}$ be the solution to optimization (2.3). If $\delta > \delta_*$, with $\delta_*$ given by (4.3), then in the asymptotic setting of Assumption 1 we have:*

*(a) The following convex-concave minimax scalar optimization problem has bounded solution $(\theta_*, \alpha_*, \gamma_{0*}, \tau_{g*}, \beta_*, \tau_{h*})$ with the minimization components $(\alpha_*, \gamma_{0*}, \theta_*)$ unique:*

$$\min_{\theta, 0\le\alpha,\gamma_0,\tau_g} \max_{0\le\beta,\tau_h} D_{\rm ns}(\alpha,\gamma_0,\theta,\tau_g,\beta,\tau_h)$$

$$D_{\rm ns}(\alpha,\gamma_0,\theta,\tau_g,\beta,\tau_h) = \frac{\beta\tau_g}{2} + L\left(\sqrt{\alpha^2+a^2\theta^2}, V\theta - \varepsilon_0\gamma_0, \frac{\tau_g}{\beta}\right)$$

$$- \min_{\lambda_0\ge 0,\nu}\left[\frac{\alpha}{\tau_h}\left\{\frac{\beta^2}{2\delta} + \lambda_0\left(\frac{\gamma_0\tau_h}{\alpha}\right)^q + \frac{\nu^2}{2} - \mathsf{E}\left(\beta, \left(\frac{\tau_h\theta}{\alpha}+\frac{\nu}{a}\right); \lambda_0\right)\right\} + \frac{\alpha\tau_h}{2}\right]. \tag{6.9}$$

*(b) It holds in probability that*

$$\lim_{n\to\infty} \frac{1}{\|\boldsymbol{\mu}\|_{\ell_2}} \langle \boldsymbol{\mu}, \widehat{\boldsymbol{\theta}}^{\varepsilon}\rangle = \theta_*\,, \tag{6.10}$$

$$\lim_{n\to\infty} \left\|\mathsf{P}_{\boldsymbol{\mu}}^{\perp}\widehat{\boldsymbol{\theta}}^{\varepsilon}\right\|_{\ell_2} = \alpha_*\,, \tag{6.11}$$

$$\lim_{n\to\infty} \|\boldsymbol{\mu}\|_{\ell_p} \left\|\widehat{\boldsymbol{\theta}}^{\varepsilon}\right\|_{\ell_q} = \gamma_{0*}\,. \tag{6.12}$$

*(c) As a corollary of part (b) and Lemma 2.1, the following limits hold in probability:*

$$\lim_{n\to\infty} \mathsf{SA}(\widehat{\boldsymbol{\theta}}^{\varepsilon}) = \Phi\left(\frac{V\theta_*}{\sqrt{\alpha_*^2+a^2\theta_*^2}}\right), \tag{6.13}$$

$$\lim_{n\to\infty} \mathsf{RA}(\widehat{\boldsymbol{\theta}}^{\varepsilon}) = \Phi\left(\frac{-\varepsilon_0\gamma_{0*}+V\theta_*}{\sqrt{\alpha_*^2+a^2\theta_*^2}}\right). \tag{6.14}$$

The results above generalize out results to the anisotropic case. The reader may of course be wondering when Assumptions 6 and 7 hold. This is the subject of the next Remark which we prove in Appendix B.4.

**Remark 6.1** *For the case of $\ell_2$ perturbation ($p = q = 2$), the following two conditions are sufficient for Assumption 6 and 7 to hold:*

(*i*) *The empirical distribution of the entries of* $\sqrt{d}\boldsymbol{\mu}$ *converges weakly to a distribution* $\mathbb{P}_M$ *on real line, with bounded second moment, i.e.* $\int x^2 \mathrm{d}\mathbb{P}_M(x) = \sigma_{M,2}^2 < \infty$.

(*ii*) *The empirical distribution of eigenvalues of* $\boldsymbol{\Sigma}$ *converges weakly to a distribution* $\rho$ *with Stieltjes transform* $S_\rho(z) := \int \frac{\rho(t)}{z-t}\mathrm{d}t$.

**7. Proof sketch and mathematical challenges** Our theoretical results on adversarial training for binary classification fits in the rapidly growing recent literature on developing *sharp* high-dimensional asymptotics of (possibly non-smooth) convex optimization-based estimators [18, 59, 4, 1, 60, 17, 51, 64, 34, 19, 15, 50, 48, 67, 9, 25, 7]. Most of this line of work focus on linear models and regression problems. It has been only recently that the literature witnessed a surge of interest in sharp analysis of a variety of methods tailored to binary classification models [26, 8, 61, 43, 33, 56, 62, 12, 49, 38, 46, 62, 40]. However, none of these papers study adversarial training and its impact on standard/robust accuracies.

On a technical level, our sharp analysis relies on the Convex Gaussian Min-max Theorem (CGMT) [64] (see also [60, 51, 50])), which is a powerful extension of the Gordon's Gaussian comparison inequality [21]. We refer to Section 7 for an overview of this framework and the mathematical challenges we encounter in applying it to our adversarial setting. We next present a proof sketch for deriving our main results which illustrates the key ideas.

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

$$\min_{\boldsymbol{\theta}\in\mathbb{R}^d} \mathcal{L}(\boldsymbol{\theta}) := \frac{1}{n}\sum_{i=1}^{n} \ell\left(y_i\langle \boldsymbol{x}_i, \boldsymbol{\theta}\rangle - \varepsilon \left\|\boldsymbol{\theta}\right\|_{\ell_q}\right), \tag{7.1}$$

and its optimal solution $\widehat{\boldsymbol{\theta}}^{\varepsilon} \in \arg\min_{\boldsymbol{\theta}\in\mathbb{R}^d} \mathcal{L}(\boldsymbol{\theta})$. Given the classification nature of the problem, as discussed earlier, we have to study this loss in the two different regimes of separable and non-separable as well as characterize the threshold of separability. In this section we wish to provide a brief overview of the steps of our proofs and some of the challenges. We focus our exposition on the non-separable case. While the details of the derivations for the separable case and the calculation of the separability threshold differ from the non-separable case the general steps are similar and therefore the steps below also provides a general road map for the proof of these results as well. Specifically, our proofs in the non-separable regime consists of the following steps:

**Step I: Reformulation of the loss.**
The loss (7.1), while significantly simplified due to the removal of the max function, is still rather complicated and precisely characterizing the behavior

and the quality of its optimal solution is still challenging. 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. Combining these reformulation steps we arrive at the following equivalent Primal Optimization (PO) problem

$$\min_{\boldsymbol{\theta},\boldsymbol{v}\in\mathbb{R}^n}\max_{\boldsymbol{u}\in\mathbb{R}^n}\frac{1}{n}\Big\{\boldsymbol{u}^\mathsf{T}\mathbf{1}\boldsymbol{\mu}^T\boldsymbol{\theta}+\boldsymbol{u}^\mathsf{T}\boldsymbol{D}_{\boldsymbol{y}}\boldsymbol{Z}\boldsymbol{\Sigma}^{1/2}\boldsymbol{\theta}-\boldsymbol{u}^\mathsf{T}\boldsymbol{v}\Big\}+\frac{1}{n}\sum_{i=1}^{n}\ell\left(v_i-\varepsilon\left\|\boldsymbol{\theta}\right\|_{\ell_q}\right) \tag{7.2}$$

**Step II: Reduction to an Auxiliary Optimization (AO) problem.** The equivalent form above 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 [21] known as Convex Gaussian Minimax Theorem (CGMT) [64] 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 (7.1) by studying the asymptotic behavior of the following, arguable simpler, *Auxiliary Optimization (AO)* problem instead

$$\begin{aligned}\min_{\boldsymbol{\theta},\boldsymbol{v}\in\mathbb{R}^n}\max_{\boldsymbol{u}\in\mathbb{R}^n}\frac{1}{n}\Big\{&\left\|\mathsf{P}_{\boldsymbol{\mu}}^{\perp}\boldsymbol{\Sigma}^{1/2}\boldsymbol{\theta}\right\|_{\ell_2}\boldsymbol{g}^T\boldsymbol{D}_{\boldsymbol{y}}\boldsymbol{u}+\left\|\boldsymbol{D}_{\boldsymbol{y}}\boldsymbol{u}\right\|_{\ell_2}\boldsymbol{h}^T\mathsf{P}_{\boldsymbol{\mu}}^{\perp}\boldsymbol{\Sigma}^{1/2}\boldsymbol{\theta}\\&+(\boldsymbol{u}^\mathsf{T}\boldsymbol{D}_{\boldsymbol{y}}\boldsymbol{z})(\widetilde{\boldsymbol{\mu}}^T\boldsymbol{\Sigma}^{1/2}\boldsymbol{\theta})+\boldsymbol{u}^\mathsf{T}\mathbf{1}\boldsymbol{\mu}^T\boldsymbol{\theta}-\boldsymbol{u}^\mathsf{T}\boldsymbol{v}\Big\}+\frac{1}{n}\sum_{i=1}^{n}\ell\left(v_i-\varepsilon\left\|\boldsymbol{\theta}\right\|_{\ell_q}\right),\end{aligned} \tag{7.3}$$

where $\boldsymbol{g}\sim\mathsf{N}(0,\boldsymbol{I}_n)$ and $\boldsymbol{h}\sim\mathsf{N}(0,\boldsymbol{I}_d)$, $\mathsf{P}_{\boldsymbol{\mu}}^{\perp}:=\boldsymbol{I}-\widetilde{\boldsymbol{\mu}}\widetilde{\boldsymbol{\mu}}^T$, and $\mathsf{P}_{\boldsymbol{\mu}}:=\widetilde{\boldsymbol{\mu}}\widetilde{\boldsymbol{\mu}}^T$.

We emphasize that the relationship between the above PO problem (7.2) and how it is exactly related to the AO problem (7.3) is more intricate and technical compared with classical CGMT and related work in the context of classification [56, 62]. In particular, prior work on binary classification such as [56, 62] via CGMT (which corresponds to the non-robust case i.e. $\varepsilon=0$) utilize the fact that (7.2) is rotationally invariant and hence one can assume $\boldsymbol{\mu}=\boldsymbol{e}_1$ without loss of generality. However, in the robust version (unless $p=q=2$) the direction of $\boldsymbol{\mu}$ plays a crucial role due to the regularization term $\frac{1}{n}\sum_{i=1}^{n}\ell\left(v_i-\varepsilon\left\|\boldsymbol{\theta}\right\|_{\ell_q}\right)$.

**Step III: Scalarization of the Auxiliary Optimization (AO) problem.**
In this step we further simplify the AO problem in (7.3). In particular we

show the asymptotic behavior of the AO can be characterized rather precisely via the scalar optimization problem

$$\min_{\theta,0\le\alpha,\gamma_0,\tau_g}\ \max_{0\le\beta,\tau_h}\ D_{\rm ns}(\alpha,\gamma_0,\theta,\tau_g,\beta,\tau_h) := \frac{\beta\tau_g}{2} + L\left(\sqrt{\alpha^2+a^2\theta^2}, V\theta-\varepsilon_0\gamma_0, \frac{\tau_g}{\beta}\right) \tag{7.4}$$
$$- \min_{\lambda_0\ge 0,\nu}\left[\frac{\alpha}{\tau_h}\left\{\frac{\beta^2}{2\delta}+\lambda_0\left(\frac{\gamma_0\tau_h}{\alpha}\right)^q+\frac{\nu^2}{2}-\mathsf{E}\left(\beta,\left(\frac{\tau_h\theta}{\alpha}+\frac{\nu}{a}\right);\lambda_0\right)\right\}+\frac{\alpha\tau_h}{2}\right].$$

More specifically, 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 this expression may look complicated we prove that this optimization problem is in fact convex in the minimization parameters $(\theta,\alpha,\tau_g)$ and concave in the maximization parameters $(\beta,\tau_h)$ 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 B for details. However, we briefly state a few mathematical challenges that is unique to simplifying (7.3). First, the AO (7.3) does not have a simple regularization whose scalarization reduces to a simple mean width calculation as in most simple CGMT uses. Instead the regularization has a complicated form $\frac{1}{n}\sum_{i=1}^n \ell\left(v_i-\varepsilon\|\boldsymbol{\theta}\|_{\ell_q}\right)$ which requires rather intricate and involved scalarization calculations. Second, this AO regularization term is not separable in $\boldsymbol{\theta}$ which significantly complicates the scalarization of the AO. Finally, we handle the case of more general covariance matrices where $\boldsymbol{\Sigma}\neq\boldsymbol{I}$.

**Step IV: Completing the proof of the theorems.**
Finally, we utilize the above scalar form to derive all of the different theorems and results. This is done by relating the quantities of interest in each theorem to the optimal solutions of (7.4). For instance, we show that $\lim_{n\to\infty}\mathsf{SA}(\widehat{\boldsymbol{\theta}}^\varepsilon) = \Phi\left(\frac{V\theta_*}{\sqrt{\alpha_*^2+a^2\theta_*^2}}\right)$ and $\lim_{n\to\infty}\mathsf{RA}(\widehat{\boldsymbol{\theta}}^\varepsilon) = \Phi\left(\frac{-\varepsilon_0\gamma_{0*}+V\theta_*}{\sqrt{\alpha_*^2+a^2\theta_*^2}}\right)$ where $\alpha_*$ and $\theta_*$ are the optimal solutions over $\alpha$ and $\theta$. These calculations/proofs are carried out in detail in Section B. Since each argument is different we do not provide a summary here and refer to the corresponding sections.

**8. Discussion** We conclude the paper by discussing some of the potential extensions and applications of our theory as well as comparison with more classical approaches to binary classification.

8.1. *Generalization to random features models* While our focus in this paper was on linear classifiers, these models are quite foundational and serve as

the basis for more complex models. For instance, one potential generalization of our results is to the class of random features models given by

$$\mathcal{F}_{\mathrm{RF}} := \left\{ f(\boldsymbol{x}; \boldsymbol{\theta}, \boldsymbol{W}) = \mathrm{sign}(\langle \boldsymbol{\theta}, \sigma(\boldsymbol{W}\boldsymbol{x}) \rangle) : \quad \boldsymbol{\theta} \in \mathbb{R}^N \right\},$$

where $\boldsymbol{x} \in \mathbb{R}^d$ represents the feature vector, $\boldsymbol{W} \in \mathbb{R}^{N \times d}$ is a random matrix whose rows are chosen uniformly at random from the unit sphere in $d$-dimension, and $\sigma : \mathbb{R} \mapsto \mathbb{R}$ is a nonlinear function (for a vector $\boldsymbol{v}$, $\sigma(\boldsymbol{v}) = (\sigma(v_1), \ldots, \sigma(v_m))$ is applied entry-wise). Random features model can also be described as a two-layer fully connected neural network with random first-layer weights fixed to $\boldsymbol{W}$ and not optimized, while the second layer weights are represented by vector $\boldsymbol{\theta}$ and are optimized over to minimize the loss of interest. The random features model was introduced by [55] for scaling kernel methods to large datasets and there has been a large body of work drawing connections between random features models, kernel methods and fully trained neural networks [11, 10, 27, 37].

An intriguing phenomenon, pointed out by [45, 49] from the analysis of random features model in non-adversarial contexts, is that the random features model has the same asymptotic behavior as a simpler noisy linear features model whose second order statistics match the nonlinear random features model, namely a linear model with noisy features $\boldsymbol{u} \in \mathbb{R}^N$ given by $\boldsymbol{u} = \eta_0 + \eta_1 \boldsymbol{W}\boldsymbol{x} + \eta_2 \boldsymbol{z}$, where $\boldsymbol{z}$ has i.i.d standard normal entries, independent of $\boldsymbol{W}$ and $\boldsymbol{x}$. Also, the constants $\eta_0, \eta_1, \eta_2$ depend on the activation function $\sigma(\cdot)$ and are chosen so that the two models have the same first and second moments. A promising direction is to establish a similar connection for an adversarial setting and use our theory (relied on CGMT framework) to analyze the equivalent noisy linear model, from which we obtain an asymptotic characterization for adversarial training under the random features model. Very recently and after this paper was posted, [23] has pursued a similar approach to precisely characterize the role of overparametrization on robust generalization of random features in a regression setting.

8.2. *Optimal $\varepsilon_0$ for the robust minimax estimator* An interesting application of our theory is to derive the optimal value $\varepsilon_0^{\mathrm{op}}$ (perceived perturbation level) in the robust minimax estimator (2.2), while fixing the adversary's (actual) perturbation level on test inputs to $\varepsilon_{0,\mathrm{test}}$. (See Remark 4.4 on how our theory applies to this setting.) The optimality here is with respect to maximizing the robust accuracy. Somewhat surprisingly $\varepsilon_0^{\mathrm{op}}$ is different than $\varepsilon_{0,\mathrm{test}}$ in general and depends on $\delta$ and the choice of perturbation norm $\ell_p$ in a non-trivial way (There is no one-fit-all solution and this highlights the importance of having a precise theory to understand the effect of adversarial

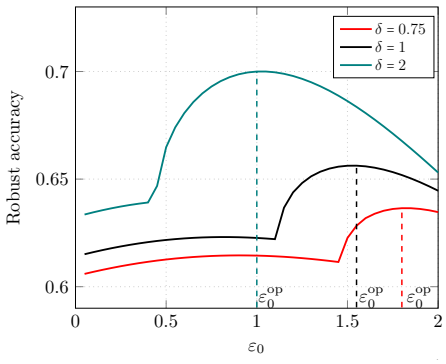



Fig 8: Robust accuracy curves versus $\varepsilon_0$ for different choices of $\delta$, and the perturbation norm $\ell_1$ ($p = 1$). The optimal choice of $\varepsilon_0$ for the robust minimax estimator decreases with $\delta$.

training which is the primary goal of the current work). For example, in the particular case of $\varepsilon_{0,\text{test}} = 0$, the question reduces to finding the value of $\varepsilon_0$ which maximizes standard accuracy. As we already discussed, the answer very much depends on $\delta$ and $p$. For $p = 2$, we observe that (cf. Figure 3(a)) adversarial training helps with improving the standard accuracy. However for $p = \infty$, $\varepsilon_0^{\text{op}}$ should be large enough so that the problem becomes non-separable and also its value decreases as $\delta$ increases (cf. Figure 5(a)). As another example, we consider the case of $\varepsilon_{0,\text{test}} = 0.3$ with $\ell_\infty$ perturbations. In Figure 8 we plot the robust accuracy versus $\varepsilon_0$, and the dashed vertical lines show the value of $\varepsilon_0^{\text{op}}$. As we see its value decreases by increasing $\delta$, however, its exact value requires a precise analysis.

8.3. *Comparison with Linear Discriminant Analysis (LDA)* A classical approach to binary classification under the Gaussian-mixture model is the Linear Discriminant Analysis. In comparing the robustness property of LDA and the robust minimax estimator studied in this paper, we cannot say one estimator always outperforms the others. To further discuss this point, we consider the Gaussian-mixture model with identity covariance $\boldsymbol{\Sigma} = \boldsymbol{I}$ and balanced classes. In this case, the LDA estimator reduces to $\widehat{\boldsymbol{\mu}}^{\text{LDA}} = \frac{1}{n}\sum_{i=1}^n y_i \boldsymbol{x}_i$ and the corresponding classification rule given by $\hat{y} = \text{sign}(\langle \boldsymbol{x}, \widehat{\boldsymbol{\mu}}^{\text{LDA}} \rangle)$. In the supplementary [29] (Section A), we compare the robust accuracy of LDA estimator with that of the robust minimax estimator $\widehat{\boldsymbol{\theta}}^{\varepsilon}$ for some choices of $p$. As we will discuss, the depending on $p$ and the adversary's power $\varepsilon_0$, one can outperform the other.

SUPPLEMENTARY MATERIAL

**Supplement to: "Precise Statistical Analysis of Classification Accuracies for Adversarial Training"**

(). Due to space constraints, proofs of theorems and some of the technical details are provided in the Supplementary Material [29].

## References


[1] D. Amelunxen, M. Lotz, M. B. McCoy, and J. A. Tropp. Living on the edge: A geometric theory of phase transitions in convex optimization. *arXiv preprint arXiv:1303.6672*, 2013.

[2] D. Bartl, S. Drapeau, J. Obloj, and J. Wiesel. Robust uncertainty sensitivity analysis. *arXiv preprint arXiv:2006.12022*, 2020.

[3] M. Bayati and A. Montanari. The dynamics of message passing on dense graphs, with applications to compressed sensing. *IEEE Transactions on Information Theory*, 57(2):764–785, 2011.

[4] M. Bayati and A. Montanari. The lasso risk for gaussian matrices. *Information Theory, IEEE Transactions on*, 58(4):1997–2017, 2012.

[5] M. Belkin, D. Hsu, S. Ma, and S. Mandal. Reconciling modern machine learning and the bias-variance trade-off. *arXiv preprint arXiv:1812.11118*, 2018.

[6] M. Belkin, S. Ma, and S. Mandal. To understand deep learning we need to understand kernel learning. In *International Conference on Machine Learning*, pages 541–549, 2018.

[7] Z. Bu, J. Klusowski, C. Rush, and W. Su. Algorithmic analysis and statistical estimation of slope via approximate message passing. In *Advances in Neural Information Processing Systems*, pages 9361–9371, 2019.

[8] E. J. Candès and P. Sur. The phase transition for the existence of the maximum likelihood estimate in high-dimensional logistic regression. *The Annals of Statistics*, 48(1):27–42, 2020.

[9] M. Celentano and A. Montanari. Fundamental barriers to high-dimensional regression with convex penalties. *arXiv preprint arXiv:1903.10603*, 2019.

[10] A. Daniely. Sgd learns the conjugate kernel class of the network. In *Advances in Neural Information Processing Systems*, pages 2422–2430, 2017.

[11] A. Daniely, R. Frostig, and Y. Singer. Toward deeper understanding of neural networks: The power of initialization and a dual view on expressivity. In *Advances In Neural Information Processing Systems*, pages 2253–2261, 2016.

[12] Z. Deng, A. Kammoun, and C. Thrampoulidis. A model of double descent for high-dimensional binary linear classification. *arXiv preprint arXiv:1911.05822*, 2019.

[13] Y. Deshpande and A. Montanari. Sparse pca via covariance thresholding. *Advances in Neural Information Processing Systems*, 27:334–342, 2014.

[14] E. Dobriban, H. Hassani, D. Hong, and A. Robey. Provable tradeoffs in adversarially robust classification. *arXiv preprint arXiv:2006.05161*, 2020.

[15] D. Donoho and A. Montanari. High dimensional robust m-estimation: Asymptotic variance via approximate message passing. *Probability Theory and Related Fields*, 166(3-4):935–969, 2016.

[16] D. L. Donoho, M. Gavish, and I. M. Johnstone. Optimal shrinkage of eigenvalues in the spiked covariance model. *Annals of statistics*, 46(4):1742, 2018.

[17] D. L. Donoho, A. Javanmard, and A. Montanari. Information-theoretically optimal

compressed sensing via spatial coupling and approximate message passing. *IEEE transactions on information theory*, 59(11):7434–7464, 2013.

[18] D. L. Donoho, A. Maleki, and A. Montanari. The noise-sensitivity phase transition in compressed sensing. *Information Theory, IEEE Transactions on*, 57(10):6920–6941, 2011.

[19] N. El Karoui. On the impact of predictor geometry on the performance on high-dimensional ridge-regularized generalized robust regression estimators. *Probability Theory and Related Fields*, 170(1-2):95–175, 2018.

[20] I. J. Goodfellow, J. Shlens, and C. Szegedy. Explaining and harnessing adversarial examples. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015.

[21] Y. Gordon. On milman's inequality and random subspaces which escape through a mesh in $r^n$. In *Geometric aspects of functional analysis*, pages 84–106. Springer, 1988.

[22] M. Grant, S. Boyd, and Y. Ye. Cvx: Matlab software for disciplined convex programming, 2008.

[23] H. Hassani and A. Javanmard. The curse of overparametrization in adversarial training: Precise analysis of robust generalization for random features regression. *arXiv preprint arXiv:2201.05149*, 2022.

[24] T. Hastie, A. Montanari, S. Rosset, and R. J. Tibshirani. Surprises in high-dimensional ridgeless least squares interpolation. *arXiv preprint arXiv:1903.08560*, 2019.

[25] H. Hu and Y. M. Lu. Asymptotics and optimal designs of slope for sparse linear regression. *arXiv preprint arXiv:1903.11582*, 2019.

[26] H. Huang. Asymptotic behavior of support vector machine for spiked population model. *The Journal of Machine Learning Research*, 18(1):1472–1492, 2017.

[27] A. Jacot, F. Gabriel, and C. Hongler. Neural tangent kernel: Convergence and generalization in neural networks. In *Advances in neural information processing systems*, pages 8571–8580, 2018.

[28] A. Javanmard, A. Montanari, and F. Ricci-Tersenghi. Phase transitions in semidefinite relaxations. *Proceedings of the National Academy of Sciences*, 113(16):E2218–E2223, 2016.

[29] A. Javanmard and M. Soltanolkotabi. Supplementary material to "precise statistical analysis of classification accuracies for adversarial training". 2020.

[30] A. Javanmard, M. Soltanolkotabi, and H. Hassani. Precise tradeoffs in adversarial training for linear regression. volume 125 of *Proceedings of Machine Learning Research, Conference of Learning Theory (COLT)*, pages 2034–2078. PMLR, 09–12 Jul 2020.

[31] Z. Ji and M. Telgarsky. Risk and parameter convergence of logistic regression. *arXiv preprint arXiv:1803.07300*, 2018.

[32] I. M. Johnstone. On the distribution of the largest eigenvalue in principal components analysis. *Annals of statistics*, pages 295–327, 2001.

[33] A. Kammoun and M.-S. Alouini. On the precise error analysis of support vector machines. *Submitted to IEEE Transactions on information theory*, 2019.

[34] N. E. Karoui. Asymptotic behavior of unregularized and ridge-regularized high-dimensional robust regression estimators: rigorous results. *arXiv preprint arXiv:1311.2445*, 2013.

[35] A. Kurakin, I. Goodfellow, and S. Bengio. Adversarial machine learning at scale. *arXiv preprint arXiv:1611.01236*, 2016.

[36] L. Lai and E. Bayraktar. On the adversarial robustness of robust estimators. *IEEE Transactions on Information Theory*, 66(8):5097–5109, 2020.

[37] Y. Li and Y. Liang. Learning overparameterized neural networks via stochastic gradient descent on structured data. *NeurIPS*, 2018.

[38] T. Liang and P. Sur. A precise high-dimensional asymptotic theory for boosting and min-l1-norm interpolated classifiers. *arXiv preprint arXiv:2002.01586*, 2020.

[39] F. Liese and K.-J. Miescke. Statistical decision theory: Estimation, testing, and selection. In *Springer Science & Business Media*, 2008.

[40] P. Lolas. Regularization in high-dimensional regression and classification via random matrix theory. *arXiv preprint arXiv:2003.13723*, 2020.

[41] K. Lyu and J. Li. Gradient descent maximizes the margin of homogeneous neural networks. *arXiv preprint arXiv:1906.05890*, 2019.

[42] A. Madry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu. Towards deep learning models resistant to adversarial attacks. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*, 2018.

[43] X. Mai, Z. Liao, and R. Couillet. A large scale analysis of logistic regression: Asymptotic performance and new insights. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 3357–3361. IEEE, 2019.

[44] M. Mehrabi, A. Javanmard, R. A. Rossi, A. Rao, and T. Mai. Fundamental tradeoffs in distributionally adversarial training. volume 139 of *Proceedings of the 38th International Conference on Machine Learning*, pages 7544–7554. PMLR, 2021.

[45] S. Mei and A. Montanari. The generalization error of random features regression: Precise asymptotics and double descent curve. *arXiv preprint arXiv:1908.05355*, 2019.

[46] F. Mignacco, F. Krzakala, Y. M. Lu, and L. Zdeborová. The role of regularization in classification of high-dimensional noisy gaussian mixture. *arXiv preprint arXiv:2002.11544*, 2020.

[47] Y. Min, L. Chen, and A. Karbasi. The curious case of adversarially robust models: More data can help, double descend, or hurt generalization. *arXiv preprint arXiv:2002.11080*, 2020.

[48] L. Miolane and A. Montanari. The distribution of the lasso: Uniform control over sparse balls and adaptive parameter tuning. *arXiv preprint arXiv:1811.01212*, 2018.

[49] A. Montanari, F. Ruan, Y. Sohn, and J. Yan. The generalization error of max-margin linear classifiers: High-dimensional asymptotics in the overparametrized regime. *arXiv preprint arXiv:1911.01544*, 2019.

[50] S. Oymak, B. Recht, and M. Soltanolkotabi. Sharp time–data tradeoffs for linear inverse problems. *IEEE Transactions on Information Theory*, 64(6):4129–4158, 2017.

[51] S. Oymak, C. Thrampoulidis, and B. Hassibi. The squared-error of generalized lasso: A precise analysis. *arXiv preprint arXiv:1311.0830*, 2013.

[52] M. S. Pydi and V. Jog. Adversarial risk via optimal transport and optimal couplings. In *International Conference on Machine Learning*, pages 7814–7823. PMLR, 2020.

[53] A. Raghunathan, J. Steinhardt, and P. Liang. Certified defenses against adversarial examples. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*, 2018.

[54] A. Raghunathan, S. M. Xie, F. Yang, J. C. Duchi, and P. Liang. Adversarial training can hurt generalization. *arXiv preprint arXiv:1906.06032*, 2019.

[55] A. Rahimi and B. Recht. Random features for large-scale kernel machines. *Advances in neural information processing systems*, 20:1177–1184, 2007.

[56] F. Salehi, E. Abbasi, and B. Hassibi. The impact of regularization on high-dimensional logistic regression. *arXiv preprint arXiv:1906.03761*, 2019.

[57] L. Schmidt, S. Santurkar, D. Tsipras, K. Talwar, and A. Madry. Adversarially robust generalization requires more data. In *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, 3-8 December 2018, Montréal, Canada*, pages 5019–5031, 2018.

[58] M. Staib and S. Jegelka. Distributionally robust deep learning as a generalization of adversarial training. In *NIPS workshop on Machine Learning and Computer Security*, 2017.

[59] M. Stojnic. Various thresholds for $\ell_1$-optimization in compressed sensing. *arXiv preprint arXiv:0907.3666*, 2009.

[60] M. Stojnic. A framework to characterize performance of lasso algorithms. *arXiv preprint arXiv:1303.7291*, 2013.

[61] P. Sur and E. J. Candès. A modern maximum-likelihood theory for high-dimensional logistic regression. *Proceedings of the National Academy of Sciences*, 116(29):14516–14525, 2019.

[62] H. Taheri, R. Pedarsani, and C. Thrampoulidis. Sharp asymptotics and optimal performance for inference in binary models. *arXiv preprint arXiv:2002.07284*, 2020.

[63] C. Thrampoulidis, E. Abbasi, and B. Hassibi. Precise error analysis of regularized $m$-estimators in high dimensions. *IEEE Transactions on Information Theory*, 64(8):5592–5628, 2018.

[64] C. Thrampoulidis, S. Oymak, and B. Hassibi. Regularized linear regression: A precise analysis of the estimation error. In *Proceedings of The 28th Conference on Learning Theory*, pages 1683–1709, 2015.

[65] D. Tsipras, S. Santurkar, L. Engstrom, A. Turner, and A. Madry. Robustness may be at odds with accuracy. *arXiv preprint arXiv:1805.12152*, 2018.

[66] R. Vershynin. *High-dimensional probability: An introduction with applications in data science*, volume 47. Cambridge University Press, 2018.

[67] S. Wang, H. Weng, and A. Maleki. Does slope outperform bridge regression? *arXiv preprint arXiv:1909.09345*, 2019.

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

[69] H. Zhang, Y. Yu, J. Jiao, E. P. Xing, L. E. Ghaoui, and M. I. Jordan. Theoretically principled trade-off between robustness and accuracy. In *Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA*, pages 7472–7482, 2019.

*Submitted to the Annals of Statistics*

# SUPPLEMENTARY MATERIAL TO "PRECISE STATISTICAL ANALYSIS OF CLASSIFICATION ACCURACIES FOR ADVERSARIAL TRAINING"

BY ADEL JAVANMARD[§] AND MAHDI SOLTANOLKOTABI[§]

*University of Southern California*[§]

## APPENDIX A: COMPARISON WITH LINEAR DISCRIMINANT ANALYSIS (LDA)

A classical approach to binary classification under the Gaussian-mixture model is the Linear Discriminant Analysis. In comparing the robustness property of LDA and the robust minimax estimator studied in this paper, we cannot say one estimator always outperforms the others. To further discuss this point, we consider the Gaussian-mixture model with identity covariance $\boldsymbol{\Sigma} = \boldsymbol{I}$ and balanced classes. In this case, the LDA estimator reduces to $\widehat{\boldsymbol{\mu}}^{\rm LDA} = \frac{1}{n}\sum_{i=1}^{n} y_i \boldsymbol{x}_i$ and the corresponding classification rule given by $\hat{y} = \mathrm{sign}(\langle \boldsymbol{x}, \widehat{\boldsymbol{\mu}}^{\rm LDA}\rangle)$. Under the Gaussian-mixture model we have $\boldsymbol{x} = y\boldsymbol{\mu} + \boldsymbol{z}$ with $\boldsymbol{z} \sim \mathsf{N}(0, \boldsymbol{I})$. Therefore,

$$\widehat{\boldsymbol{\mu}}^{\rm LDA} = \frac{1}{n}\sum_{i=1}^{n} y_i(y_i\boldsymbol{\mu} + \boldsymbol{z}_i) = \boldsymbol{\mu} + \frac{1}{n}\sum_{i=1}^{n} y_i \boldsymbol{z}_i = \boldsymbol{\mu} + \tilde{\boldsymbol{z}}\,, \quad \tilde{\boldsymbol{z}} \sim \mathsf{N}(\boldsymbol{0}, \frac{1}{n}\boldsymbol{I})$$

For simplicity we assume that the class averages $\boldsymbol{\mu}$ is generated as $\boldsymbol{\mu} \sim (\boldsymbol{0}, \frac{1}{d}\boldsymbol{I})$, similar to the setting considered in the numerical experiments. In asymptotic regime of $n \to \infty$ and $n/d \to \delta$, we have that in probability:

$$\begin{aligned}
\lim_{n\to\infty} \langle \boldsymbol{\mu}, \widehat{\boldsymbol{\mu}}^{\rm LDA}\rangle &= \lim_{n\to\infty} \|\boldsymbol{\mu}\|_{\ell_2}^2 = 1\,,\\
\lim_{n\to\infty} d^{1/2-1/q} \left\|\widehat{\boldsymbol{\mu}}^{\rm LDA}\right\|_{\ell_q} &= \lim_{n\to\infty} d^{1/2-1/q} \|\boldsymbol{\mu} + \tilde{\boldsymbol{z}}\|_{\ell_q}\\
&= \lim_{n\to\infty} d^{1/2-1/q} \left(\frac{1}{d} + \frac{1}{n}\right)^{1/2} d^{1/q} C_q = \left(1 + \frac{1}{\delta}\right)^{1/2} C_q\,,
\end{aligned}$$

where in the first equation we used the fact that $\langle \boldsymbol{\mu}, \tilde{\boldsymbol{z}}\rangle \sim \mathsf{N}(0, \frac{1}{n}\|\boldsymbol{\mu}\|_{\ell_2}^2)$ has vanishing variance as $n \to \infty$. In the second inequality, $C_q$ is the $q$-th moment of standard normal distribution. Recall that $\varepsilon = \varepsilon_0 \|\boldsymbol{\mu}\|_{\ell_p}$ with $1/p + 1/q = 1$, and also $\|\boldsymbol{\mu}\|_{\ell_p} \to d^{1/p-1/2} C_p = d^{1/2-1/q} C_p$. Using these identities along with

the characterization of standard and robust accuracies given by Lemma 2.1 of the paper, we arrive at

$$\lim_{n\to\infty} \mathsf{SA}(\widehat{\boldsymbol{\mu}}^{\mathrm{LDA}}) = \Phi\left(\sqrt{\frac{\delta}{1+\delta}}\right),$$
$$\lim_{n\to\infty} \mathsf{RA}(\widehat{\boldsymbol{\mu}}^{\mathrm{LDA}}) = \Phi\left(\sqrt{\frac{\delta}{1+\delta}} - \varepsilon_0 C_q C_p\right). \tag{A.1}$$

We next compare the robust accuracy of LDA estimator with that of the robust minimax estimator $\widehat{\boldsymbol{\theta}}^{\varepsilon}$ for some choices of $p$. As we will discuss, the depending on $p$ and the adversary's power $\varepsilon_0$, one can outperform the other.

- $(p = q = 2)$. Figure 9(a) compares $\mathsf{RA}(\widehat{\boldsymbol{\mu}}^{\mathrm{LDA}})$ with $\mathsf{RA}(\widehat{\boldsymbol{\theta}}^{\varepsilon})$ versus $\varepsilon_0$ for several values of $\delta$. Here, the solid lines correspond to the robust minimax estimator and the dashed lines correspond to the LDA estimator. Figure 9(b) compares $\mathsf{RA}(\widehat{\boldsymbol{\mu}}^{\mathrm{LDA}})$ with $\mathsf{RA}(\widehat{\boldsymbol{\theta}}^{\varepsilon})$ versus $1/\delta$ for various choices of $\varepsilon_0$. As we see for the case of $p = 2$, the LDA has better robust accuracy and it is mostly very close to that of the robust estimator.
- $(p = \infty,\ q = 1)$. Similar to the setting of experiments in Section 5.2, here we consider the scaling $\varepsilon = \varepsilon_0/\sqrt{d}$. Figure 10 (a) compares the robust accuracies versus $\varepsilon_0$ for several values of $\delta$. As we see for any $\delta$, there exists $\varepsilon_0^*(\delta)$ above which the robust minimax outperforms the LDA. Figure 10(b) compares the robust accuracies versus $1/\delta$ for several values of $\varepsilon_0$. Rewording the above observation, for any $\varepsilon_0$ there exists $\delta^*(\varepsilon_0)$ below which the robust minimax outperforms the LDA estimator.
- $(p = 1,\ q = \infty)$. Similar to the setting of experiments in Section 5.3, we have $\varepsilon = \varepsilon_0 \|\boldsymbol{\mu}\|_{\ell_p} = \sqrt{\frac{2}{\pi}}\frac{\varepsilon_0}{\sqrt{d}}$. Invoking equations (A.1), we have $\lim_{n\to\infty} \mathsf{RA}(\widehat{\boldsymbol{\mu}}^{\mathrm{LDA}}) = 0$ because $C_q = \sqrt{2\log d} \to \infty$. However, as we see in Figure 7, the robust minimax estimator $\widehat{\boldsymbol{\theta}}^{\varepsilon}$ achieves non-trivial positive robust accuracies and hence outperforms LDA.

## APPENDIX B: PROOFS FOR ANISOTROPIC GAUSSIAN MODEL (SECTION 6)

**B.1. Proof of Theorem 6.1** As discussed the $(\varepsilon, q)$-separability condition can alternatively be written as (3.2), which we repeat here:

$$\exists \boldsymbol{\theta},\ \|\boldsymbol{\theta}\|_{\ell_q} \le \frac{1}{\varepsilon} :\quad \forall i \in [n],\ y_i\langle \boldsymbol{x}_i, \boldsymbol{\theta}\rangle > 1. \tag{B.1}$$

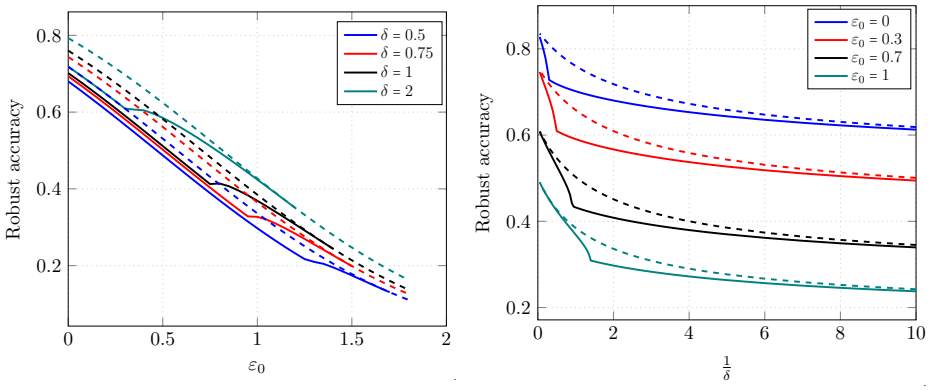



Fig 9: Robust accuracies for the LDA estimator and the robust minimax estimator versus the adversary's power with $\ell_2$ ($p = 2$) perturbations for different values of $\delta$. Solid curves correspond to the robust minimax estimator and the dashed curves correspond to the LDA estimator.

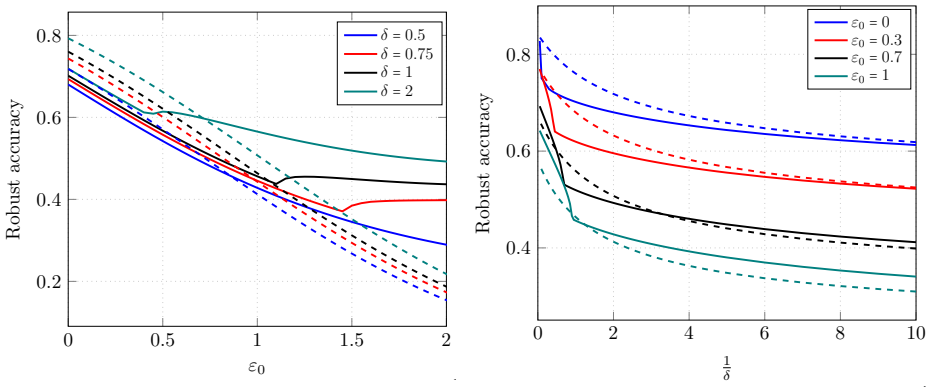



Fig 10: Robust accuracies for LDA estimator and the robust minimax estimator versus the adversary's power with $\ell_\infty$ ($p = \infty$) perturbations for different values of $\delta$. Solid curves correspond to the robust minimax estimator and the dashed curves correspond to the LDA estimator.

To find the separability threshold we consider the following feasibility problem

$$\min_{\boldsymbol{\theta}\in\mathbb{R}^d} 0 \quad \text{subject to } y_i\langle \boldsymbol{x}_i,\boldsymbol{\theta}\rangle > 1, \quad \|\boldsymbol{\theta}\|_{\ell_q} \le \frac{1}{\varepsilon}\,. \tag{B.2}$$

Clearly this is a convex optimization problem since $q \ge 1$. Writing the partial Lagrangian for the above problem with $u_i/n$ as dual coefficients, this is equivalent to

$$\min_{\boldsymbol{\theta}} \max_{u_i\ge 0} \quad \frac{1}{n}\sum_{i=1}^{n} u_i\left(1 - y_i\langle x_i,\boldsymbol{\theta}\rangle\right) \quad \text{subject to} \quad \|\boldsymbol{\theta}\|_{\ell_q} \le \frac{1}{\varepsilon}\,. \tag{B.3}$$

Under our Gaussian Mixture data model, we can substitute for $\boldsymbol{X} = \boldsymbol{y}\boldsymbol{\mu}^T + \boldsymbol{Z}\boldsymbol{\Sigma}^{1/2}$, which results in

$$\min_{\boldsymbol{\theta}} \max_{u_i\ge 0} \quad \frac{1}{n}\boldsymbol{u}^T\left(\mathbf{1}\left(1-\boldsymbol{\mu}^T\boldsymbol{\theta}\right) - \boldsymbol{D}_{\boldsymbol{y}}\boldsymbol{Z}\boldsymbol{\Sigma}^{1/2}\boldsymbol{\theta}\right) \quad \text{subject to} \quad \|\boldsymbol{\theta}\|_{\ell_q} \le \frac{1}{\varepsilon}\,. \tag{B.4}$$

The above dual problem has a finite optimal value if and only if the data is $(\varepsilon, q)$-separable. So we aim at finding the largest $\delta$ such that the above problem has still a finite optimal value. (Recall that $\frac{n}{d} \to \delta$.)

**Reduction to an auxiliary optimization problem via CGMT.** Note that $y_i = \pm 1$ are independent of $\boldsymbol{Z}$. In addition, the objective function in (B.4) is affine in the standard Gaussian matrix $\boldsymbol{Z}$ and the rest of the terms form a convex-concave function in $\boldsymbol{\theta}$, $\boldsymbol{u}$. Due to this particular form we are able to apply a powerful extension of a classical Gaussian process inequality due to Gordon [21] known as Convex Gaussian Minimax Theorem (CGMT) [64], and is discussed in the proof sketch in Section 7. The CGMT framework provides a principled machinery to characterize the asymptotic behavior of certain minimax optimization problems that are affine in a Gaussian matrix $\boldsymbol{X}$.

As discussed in the CGMT framework in Section 7, we require minimization/maximization to be over compact sets. The vector $\boldsymbol{\theta}$ already lies in the $\ell_q$ ball of radius $1/\varepsilon$ by constraint. In addition, since $u_i \ge 0$, and we are focused on the regime that (B.4) has finite optimal value, the optimal values of $u_i$ should all be finite as well.

We are now ready to applying the CGMT framework. The corresponding Auxiliary Optimization (AO) reads as

$$\min_{\boldsymbol{\theta}} \max_{\boldsymbol{u}\ge 0} \quad \frac{1}{n}\Big\{(\boldsymbol{u}^T\mathbf{1})\left(1-\boldsymbol{\mu}^T\boldsymbol{\theta}\right) + \left\|\boldsymbol{\Sigma}^{1/2}\boldsymbol{\theta}\right\|_{\ell_2}\boldsymbol{g}^T\boldsymbol{u} + \|\boldsymbol{u}\|_{\ell_2}\boldsymbol{h}^T\boldsymbol{\Sigma}^{1/2}\boldsymbol{\theta}\Big\}\,,$$
$$\text{subject to} \quad \|\boldsymbol{\theta}\|_{\ell_q} \le \frac{1}{\varepsilon}\,, \tag{B.5}$$

where $\boldsymbol{g} \sim \mathsf{N}(0, \boldsymbol{I}_n)$ and $\boldsymbol{h} \sim \mathsf{N}(0, \boldsymbol{I}_d)$. Fixing $\beta := \frac{\|\boldsymbol{u}\|_{\ell_2}}{\sqrt{n}}$ and optimizing over $\boldsymbol{u}$ on the non-negative orthant we get

$$\begin{aligned} \min_{\boldsymbol{\theta}} \max_{\beta \geq 0} \quad & \frac{\beta}{\sqrt{n}} \boldsymbol{h}^T \boldsymbol{\Sigma}^{1/2} \boldsymbol{\theta} + \frac{\beta}{\sqrt{n}} \left\| \left( \left(1 - \boldsymbol{\mu}^T \boldsymbol{\theta}\right) \mathbf{1} + \left\| \boldsymbol{\Sigma}^{1/2} \boldsymbol{\theta} \right\|_{\ell_2} \boldsymbol{g} \right)_+ \right\|_{\ell_2}, \\ \text{subject to} \quad & \|\boldsymbol{\theta}\|_{\ell_q} \leq \frac{1}{\varepsilon}. \end{aligned} \tag{B.6}$$

For data to be separable the above dual optimization should take finite optimal value and therefore the coefficient of $\beta$ should be non-positive. As such the problem is separable if and only if the optimal value of the following problem is non-positive:

$$\begin{aligned} \min_{\boldsymbol{\theta}} \frac{1}{\sqrt{n}} \boldsymbol{h}^T \boldsymbol{\Sigma}^{1/2} \boldsymbol{\theta} + \frac{1}{\sqrt{n}} \left\| \left( \left(1 - \boldsymbol{\mu}^T \boldsymbol{\theta}\right) \mathbf{1} + \left\| \boldsymbol{\Sigma}^{1/2} \boldsymbol{\theta} \right\|_{\ell_2} \boldsymbol{g} \right)_+ \right\|_{\ell_2} \leq 0, \\ \text{subject to} \quad \|\boldsymbol{\theta}\|_{\ell_q} \leq \frac{1}{\varepsilon}. \end{aligned} \tag{B.7}$$

Consider the decomposition $\boldsymbol{\theta} = \boldsymbol{\theta}_\perp + \theta \widetilde{\boldsymbol{\mu}}$ with $\boldsymbol{\theta}_\perp = \mathsf{P}_{\boldsymbol{\mu}}^{\perp} \boldsymbol{\theta}$. Note that

$$\begin{aligned} \frac{1}{\sqrt{n}} \boldsymbol{h}^T \boldsymbol{\Sigma}^{1/2} \boldsymbol{\theta} &= \frac{1}{\sqrt{n}} \boldsymbol{h}^T \boldsymbol{\Sigma}^{1/2} \boldsymbol{\theta}_\perp + \frac{1}{\sqrt{n}} \theta \boldsymbol{h}^T \boldsymbol{\Sigma}^{1/2} \widetilde{\boldsymbol{\mu}} \\ &= \frac{1}{\sqrt{n}} \boldsymbol{h}^T \boldsymbol{\Sigma}^{1/2} \boldsymbol{\theta}_\perp + \frac{1}{\sqrt{n}} a \theta \boldsymbol{h}^T \widetilde{\boldsymbol{\mu}} \end{aligned}$$

Since $\boldsymbol{h}^T \widetilde{\boldsymbol{\mu}} \sim \mathsf{N}(0, 1)$ and $\theta$ is bounded the contribution of the second term is negligible in the large sample limit $n \to \infty$. This along with the symmetry of the distribution of $\boldsymbol{h}$ bring us to

$$\begin{aligned} \min_{\alpha \geq 0, \theta, \boldsymbol{\theta}} \quad & -\frac{1}{\sqrt{n}} \boldsymbol{h}^T \boldsymbol{\Sigma}^{1/2} \boldsymbol{\theta}_\perp + \frac{1}{\sqrt{n}} \left\| \left( \left(1 - \|\boldsymbol{\mu}\|_{\ell_2} \theta\right) \mathbf{1} + \sqrt{\alpha^2 + a^2 \theta^2}\, \boldsymbol{g} \right)_+ \right\|_{\ell_2} \\ \text{subject to} \quad & \|\boldsymbol{\theta}\|_{\ell_q} \leq \frac{1}{\varepsilon}, \quad \left\| \boldsymbol{\Sigma}^{1/2} \boldsymbol{\theta}_\perp \right\|_{\ell_2} = \alpha, \quad \widetilde{\boldsymbol{\mu}}^T \boldsymbol{\theta} = \theta \end{aligned} \tag{B.8}$$

**Scalarization of the auxiliary optimization problem.** To continue recall the definition of set $\mathcal{S}$ given by

$$\mathcal{S}(\alpha, \theta, \varepsilon_0, \boldsymbol{\mu}) := \left\{ \boldsymbol{z} \in \mathbb{R}^d : \quad \boldsymbol{z}^T \widetilde{\boldsymbol{\mu}} = 0,\ \|\boldsymbol{z}\|_{\ell_2} = \alpha,\ \left\| \boldsymbol{\Sigma}^{-1/2} \boldsymbol{z} + \theta \widetilde{\boldsymbol{\mu}} \right\|_{\ell_q} \leq \frac{1}{\varepsilon_0 \|\boldsymbol{\mu}\|_{\ell_p}} \right\}.$$

Recall that $\varepsilon = \varepsilon_0 \|\boldsymbol{\mu}\|_{\ell_p}$ and so the optimization problem (B.8) above can be rewritten in the form

$$\min_{\alpha \geq 0, \theta} \min_{\boldsymbol{z} \in \mathcal{S}(\alpha, \theta, \varepsilon_0, \boldsymbol{\mu})} \quad -\frac{1}{\sqrt{n}} \boldsymbol{h}^T \boldsymbol{z} + \frac{1}{\sqrt{n}} \left\| \left( \left(1 - \|\boldsymbol{\mu}\|_{\ell_2} \theta\right) \mathbf{1} + \sqrt{\alpha^2 + a^2 \theta^2}\, \boldsymbol{g} \right)_+ \right\|_{\ell_2} \tag{B.9}$$

Recall the spherical width of a set $\mathcal{S} \subset \mathbb{R}^d$ defined as

$$\omega_s(\mathcal{S}) = \mathbb{E}\Big[\sup_{\boldsymbol{z}\in\mathcal{S}} \boldsymbol{z}^T\boldsymbol{u}\Big],$$

where $\boldsymbol{u} \in \mathcal{S}^{d-1}$ is a vector chosen at random from the unit sphere. Using this definition and the fact that $\min z = -\max -z$ we have

$$\min_{\boldsymbol{z}\in\mathcal{S}(\alpha,\theta,\varepsilon_0)} -\frac{1}{\sqrt{n}}\boldsymbol{h}^T\boldsymbol{z} = -\frac{1}{\sqrt{n/d}} \sup_{\boldsymbol{z}\in\mathcal{S}(\alpha,\theta,\varepsilon_0)} \frac{1}{\sqrt{d}}\boldsymbol{h}^T\boldsymbol{z} \to -\frac{1}{\sqrt{\delta}}\omega(\alpha,\theta,\varepsilon_0)\,,$$

in probability, where in the last line we use the fact, for $\mathcal{S} \in \mathbb{S}^{d-1}$, the function $f(\boldsymbol{u}) = \sup_{\boldsymbol{z}\in\mathcal{S}} \boldsymbol{z}^T\boldsymbol{u}$ is Lipschitz. Therefore, using the concentration of Lipschitz functions of Gaussian random vectors (see e.g. [66, Theorem 5.2.2]), $f(\boldsymbol{u})$ concentrates around its mean $\mathbb{E} f(\boldsymbol{u}) = \omega_s(\mathcal{S}(\alpha,\theta,\varepsilon_0,\boldsymbol{\mu}))$. More precisely,

$$\mathbb{P}\left\{\Big|\sup_{\boldsymbol{z}\in\mathcal{S}} \frac{1}{\sqrt{d}}\boldsymbol{h}^T\boldsymbol{z} - \omega_s(\mathcal{S}(\alpha,\theta,\varepsilon_0,\boldsymbol{\mu}))\Big|\right\} \le 2e^{-cdt^2}\,,$$

for an absolute constant $c > 0$ and for every $t \ge 0$. Therefore, by invoking the assumption on the convergence of spherical width, cf. Assumption 3, we arrive at

$$\lim_{d\to\infty} \mathbb{P}\left\{\Big|\sup_{\boldsymbol{z}\in\mathcal{S}} \frac{1}{\sqrt{d}}\boldsymbol{h}^T\boldsymbol{z} - \omega(\alpha,\theta,\varepsilon_0)\Big| \ge \eta\right\} = 0\,, \quad \forall \eta > 0\,.$$

Therefore, $\sup_{\boldsymbol{z}\in\mathcal{S}(\alpha,\theta,\varepsilon_0)} \frac{1}{\sqrt{d}}\boldsymbol{h}^T\boldsymbol{z} \to \omega(\alpha,\theta,\varepsilon_0)$, in probability.

Furthermore, $\|\boldsymbol{\mu}\|_{\ell_2} \to V$ by Assumption 2 and since $\boldsymbol{g} \sim \mathsf{N}(0,\boldsymbol{I}_n)$ by applying the Weak Law of Large Numbers we have

$$\frac{1}{\sqrt{n}}\Big\|\Big(\big(1-\|\boldsymbol{\mu}\|_{\ell_2}\,\theta\big)\mathbf{1} + \sqrt{\alpha^2+a^2\theta^2}\,\boldsymbol{g}\Big)_+\Big\|_{\ell_2} \to \sqrt{\mathbb{E}\Big[\Big(1 - V\theta + \sqrt{\alpha^2+a^2\theta^2}\,g\Big)_+^2\Big]}$$

Thus the objective function in the optimization problem (B.9) converges pointwise to

$$\min_{\alpha\ge 0,\theta} -\frac{1}{\sqrt{\delta}}\omega(\alpha,\theta,\varepsilon_0) + \sqrt{\mathbb{E}\Big[\Big(1 - V\theta + \sqrt{\alpha^2+a^2\theta^2}\,g\Big)_+^2\Big]} \tag{B.10}$$

Also the problem (B.9) is convex as a function of $(\alpha,\theta,\boldsymbol{z})$ and since partial maximization preserves convexity, the objective of (B.9) (after minimization over $\boldsymbol{z}$) is a convex function of $(\alpha,\theta)$. We can thus apply the convexity lemma [63, Lemma B.2] to conclude that the minimum value of (B.9) over

$\alpha \geq 0, \theta$ also converges to that of (B.10). Therefore, we conclude that data is $(\varepsilon, q)$-separable if and only if the optimal value in (B.10) is finite. Rearranging the terms gives us that (B.10) has a finite optimal value if and only if

$$\delta < \delta_*, \quad \text{with } \delta_* := \min_{\alpha \geq 0, \theta} \frac{\omega\left(\alpha, \theta, \varepsilon_0\right)^2}{\mathbb{E}\left[\left(1 - V\theta + \sqrt{\alpha^2 + \theta^2} g\right)_+^2\right]} . \tag{B.11}$$

This completes the proof of Theorem 6.1.

**B.2. Proof of Theorem 6.3** We prove Theorem 6.3 using the Convex Gaussian Minimax Theorem (CGMT) as outlined in Section 7. The max-margin problem (F.2) can be equivalently written as

$$\begin{aligned} (\widetilde{\boldsymbol{\theta}}^{\varepsilon}, \widetilde{\gamma}) = \arg \min_{\boldsymbol{\theta}, \gamma \geq 0} \quad & \|\boldsymbol{\theta}\|_{\ell_2}^2 \\ \text{subject to } \; & y_i \langle \boldsymbol{x}_i, \boldsymbol{\theta} \rangle - \varepsilon \gamma \geq 1, \quad \gamma \geq \|\boldsymbol{\theta}\|_{\ell_q} \end{aligned} \tag{B.12}$$

Now note that writing the Lagrangian for the max-margin problem with $u_i/n$ and $2\lambda$ as dual coefficients, this is equivalent to

$$\min_{\boldsymbol{\theta}, \gamma \geq 0} \max_{u_i, \lambda \geq 0} \quad \|\boldsymbol{\theta}\|_{\ell_2}^2 + \frac{1}{n} \sum_{i=1}^{n} u_i \left(1 + \varepsilon \gamma - y_i \langle x_i, \boldsymbol{\theta} \rangle\right) + 2\lambda \left(\|\boldsymbol{\theta}\|_{\ell_q} - \gamma\right) . \tag{B.13}$$

We next substitute for $\boldsymbol{X} = \boldsymbol{y}\boldsymbol{\mu}^T + \boldsymbol{Z}\boldsymbol{\Sigma}^{1/2}$ based on the Gaussian mixtures model to arrive at

$$\min_{\boldsymbol{\theta}, \gamma \geq 0} \max_{u_i \geq 0, \lambda \geq 0} \quad \|\boldsymbol{\theta}\|_{\ell_2}^2 + \frac{1}{n} \left(\boldsymbol{u}^T \mathbf{1} + \varepsilon \gamma \boldsymbol{u}^T \mathbf{1} - \boldsymbol{u}^T \boldsymbol{D}_{\boldsymbol{y}} \boldsymbol{Z} \boldsymbol{\Sigma}^{1/2} \boldsymbol{\theta} - \boldsymbol{u}^T \mathbf{1} \boldsymbol{\mu}^T \boldsymbol{\theta}\right) + 2\lambda \left(\|\boldsymbol{\theta}\|_{\ell_q} - \gamma\right) . \tag{B.14}$$

The advantage of the Lagrangian form in (B.14) is that it is a minimax problem and the objective is an affine function of the standard Gaussian matrix $\boldsymbol{Z}$. Therefore, we can deploy the Convex Gaussian Minimax Theorem (CGMT) [64], described in Section 7, to characterize asymptotic values of certain functions of this optimization solution, in a high probability sense.

To recall, the CGMT framework shows that a problem of the form

$$\min_{\boldsymbol{\theta} \in \mathcal{S}_{\boldsymbol{\theta}}} \max_{\boldsymbol{u} \in \mathcal{S}_{\boldsymbol{u}}} \quad \boldsymbol{u}^T \boldsymbol{Z} \boldsymbol{\theta} + \psi(\boldsymbol{\theta}, \boldsymbol{u}) \tag{B.15}$$

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

$$\min_{\boldsymbol{\theta} \in \mathcal{S}_{\boldsymbol{\theta}}} \max_{\boldsymbol{u} \in \mathcal{S}_{\boldsymbol{u}}} \quad \|\boldsymbol{\theta}\|_{\ell_2} \boldsymbol{g}^T \boldsymbol{u} + \|\boldsymbol{u}\|_{\ell_2} \boldsymbol{h}^T \boldsymbol{\theta} + \psi(\boldsymbol{\theta}, \boldsymbol{u}) \tag{B.16}$$

where $\boldsymbol{g}$ and $\boldsymbol{h}$ are independent Gaussian vectors with i.i.d. $\mathcal{N}(0,1)$ entries and $\psi(\boldsymbol{\theta},\boldsymbol{u})$ is convex in $\boldsymbol{\theta}$ and concave in $\boldsymbol{u}$. Specifically, the optimal value and corresponding solution of (B.15) converge in probability to the optimal value and the corresponding solution of (B.16). In the above $\mathcal{S}_{\boldsymbol{\theta}}$ and $\mathcal{S}_{\boldsymbol{u}}$ are compact sets. We refer to [64, Theorem 3] for precise statements. As explained in the proof sketch in 7, we follow [64] in referring to problems of the form (B.15) and (B.16) as the Primal Problem (PO) and the Auxiliary Problem (AO).

Note that in order to apply CGMT, we need the minimization/maximization to be over compact sets. This technical issue can be avoided by introducing "artificial" boundedness constraints on the optimization variables that they do not change the optimal solution. Concretely, we can add constraints of the form $\mathcal{S}_{\boldsymbol{\theta}} = \{\boldsymbol{\theta} : \ \|\boldsymbol{\theta}\|_{\ell_q} \le K_{\boldsymbol{\theta}}\}$ and $\mathcal{S}_{\boldsymbol{u}} = \{\boldsymbol{u} : 0 \le u_i,\ \frac{1}{n}\mathbf{1}^T\boldsymbol{u} \le K_{\boldsymbol{u}}\}$ for sufficiently large constants $K_{\boldsymbol{\theta}}$, $K_{\boldsymbol{u}}$ without changing the optimal solution of (B.14) in a precise asymptotic sense. We refer to Appendix E.3.1 for precise statements and proofs. This allows us to replace (B.14) with

(B.17)
$$\min_{\boldsymbol{\theta}\in\mathcal{S}_{\boldsymbol{\theta}},\gamma\ge 0}\ \max_{\boldsymbol{u}\in\mathcal{S}_{\boldsymbol{u}},\lambda\ge 0}\ \|\boldsymbol{\theta}\|_{\ell_2}^2 + \frac{1}{n}\left(\boldsymbol{u}^T\mathbf{1} + \varepsilon\gamma\boldsymbol{u}^T\mathbf{1} - \boldsymbol{u}^T\boldsymbol{D}_{\boldsymbol{y}}\boldsymbol{Z}\boldsymbol{\Sigma}^{1/2}\boldsymbol{\theta} - \boldsymbol{u}^T\mathbf{1}\boldsymbol{\mu}^T\boldsymbol{\theta}\right) + 2\lambda\left(\|\boldsymbol{\theta}\|_{\ell_q} - \gamma\right).$$

**Reduction to an auxiliary optimization problem via CGMT.** With these compact constraints in place we can now apply the CGMT result to obtain the auxiliary optimization (AO) problem.

We proceed by defining the projection matrices

$$\mathsf{P}_{\boldsymbol{\mu}}^{\perp} := \boldsymbol{I} - \widetilde{\boldsymbol{\mu}}\widetilde{\boldsymbol{\mu}}^T, \quad \mathsf{P}_{\boldsymbol{\mu}} := \widetilde{\boldsymbol{\mu}}\widetilde{\boldsymbol{\mu}}^T$$

and rewrite $\boldsymbol{Z}\boldsymbol{\Sigma}^{1/2} = \boldsymbol{Z}\left(\mathsf{P}_{\boldsymbol{\mu}} + \mathsf{P}_{\boldsymbol{\mu}}^{\perp}\right)\boldsymbol{\Sigma}^{1/2}$. Since $\boldsymbol{Z}\mathsf{P}_{\boldsymbol{\mu}}$ and $\boldsymbol{Z}\mathsf{P}_{\boldsymbol{\mu}}^{\perp}$ are independent from each other the latter has the same distribution as

$$\boldsymbol{Z}\boldsymbol{\Sigma}^{1/2} \sim \boldsymbol{z}\left(\boldsymbol{\Sigma}^{1/2}\widetilde{\boldsymbol{\mu}}\right)^T + \boldsymbol{Z}\mathsf{P}_{\boldsymbol{\mu}}^{\perp}\boldsymbol{\Sigma}^{1/2}$$

where $\boldsymbol{z} \sim \mathcal{N}(0, \boldsymbol{I}_n)$ and is independent from the matrix $\boldsymbol{Z}$. This brings us to the following representation

$$\min_{\boldsymbol{\theta},\gamma\ge 0}\ \max_{\boldsymbol{u}\ge 0,\lambda\ge 0}\ \|\boldsymbol{\theta}\|_{\ell_2}^2 + 2\lambda\left(\|\boldsymbol{\theta}\|_{\ell_q} - \gamma\right) + \frac{1}{n}\Big\{\boldsymbol{u}^T\mathbf{1} + \varepsilon\gamma\boldsymbol{u}^T\mathbf{1} - \boldsymbol{u}^T\mathbf{1}\boldsymbol{\mu}^T\boldsymbol{\theta} - (\boldsymbol{u}^{\mathsf{T}}\boldsymbol{D}_{\boldsymbol{y}}\boldsymbol{z})(\widetilde{\boldsymbol{\mu}}^T\boldsymbol{\Sigma}^{1/2}\boldsymbol{\theta}) - \boldsymbol{u}^{\mathsf{T}}\boldsymbol{D}_{\boldsymbol{y}}\boldsymbol{Z}\mathsf{P}_{\boldsymbol{\mu}}^{\perp}\boldsymbol{\Sigma}^{1/2}\boldsymbol{\theta}\Big\}$$

Since $y_i = \pm 1$ are independent of $\boldsymbol{Z}$, by applying CGMT framework, the AO reads as

$$\min_{\boldsymbol{\theta},\gamma\geq 0}\max_{\boldsymbol{u}\geq 0,\lambda\geq 0} \quad \|\boldsymbol{\theta}\|_{\ell_2}^2 + 2\lambda\Big(\|\boldsymbol{\theta}\|_{\ell_q} - \gamma\Big) + \frac{1}{n}\Big\{\boldsymbol{u}^T\mathbf{1} + \varepsilon\gamma\boldsymbol{u}^T\mathbf{1} - \boldsymbol{u}^T\mathbf{1}\boldsymbol{\mu}^T\boldsymbol{\theta} + (\boldsymbol{u}^T\boldsymbol{z})(\widetilde{\boldsymbol{\mu}}^T\boldsymbol{\Sigma}^{1/2}\boldsymbol{\theta}) + \left\|\mathsf{P}_{\boldsymbol{\mu}}^{\perp}\boldsymbol{\Sigma}^{1/2}\boldsymbol{\theta}\right\|_{\ell_2}\boldsymbol{g}^T\boldsymbol{u} + \|\boldsymbol{u}\|_{\ell_2}\boldsymbol{h}^T\mathsf{P}_{\boldsymbol{\mu}}^{\perp}\boldsymbol{\Sigma}^{1/2}\boldsymbol{\theta}\Big\}$$

Fixing $\beta := \frac{\|\boldsymbol{u}\|_{\ell_2}}{\sqrt{n}}$ and optimizing over $\boldsymbol{u}$ on the non-negative orthant we get

$$\min_{\boldsymbol{\theta},\gamma\geq 0}\max_{\beta\geq 0,\lambda\geq 0} \quad \|\boldsymbol{\theta}\|_{\ell_2}^2 + 2\lambda\Big(\|\boldsymbol{\theta}\|_{\ell_q} - \gamma\Big) + \frac{\beta}{\sqrt{n}}\boldsymbol{h}^T\mathsf{P}_{\boldsymbol{\mu}}^{\perp}\boldsymbol{\Sigma}^{1/2}\boldsymbol{\theta} + \frac{\beta}{\sqrt{n}}\left\|\left(\left(1+\varepsilon\gamma - \boldsymbol{\mu}^T\boldsymbol{\theta}\right)\mathbf{1} + (\widetilde{\boldsymbol{\mu}}^T\boldsymbol{\Sigma}^{1/2}\boldsymbol{\theta})\boldsymbol{z} + \left\|\mathsf{P}_{\boldsymbol{\mu}}^{\perp}\boldsymbol{\Sigma}^{1/2}\boldsymbol{\theta}\right\|_{\ell_2}\boldsymbol{g}\right)_+\right\|_{\ell_2} \tag{B.18}$$

Since $\boldsymbol{z}, \boldsymbol{g} \sim \mathsf{N}(0, \boldsymbol{I}_n)$ are independent, by applying the Weak Law of Large Numbers we have

$$\begin{aligned}
&\frac{1}{\sqrt{n}}\left\|\left(\left(1+\varepsilon\gamma - \boldsymbol{\mu}^T\boldsymbol{\theta}\right)\mathbf{1} + (\widetilde{\boldsymbol{\mu}}^T\boldsymbol{\Sigma}^{1/2}\boldsymbol{\theta})\boldsymbol{z} + \left\|\mathsf{P}_{\boldsymbol{\mu}}^{\perp}\boldsymbol{\Sigma}^{1/2}\boldsymbol{\theta}\right\|_{\ell_2}\boldsymbol{g}\right)_+\right\|_{\ell_2} \\
&\qquad \to \left(\mathbb{E}\Big[\left(\left(1+\varepsilon\gamma - \boldsymbol{\mu}^T\boldsymbol{\theta}\right) + \sqrt{(\widetilde{\boldsymbol{\mu}}^T\boldsymbol{\Sigma}^{1/2}\boldsymbol{\theta})^2 + \left\|\mathsf{P}_{\boldsymbol{\mu}}^{\perp}\boldsymbol{\Sigma}^{1/2}\boldsymbol{\theta}\right\|_{\ell_2}^2}\, g\right)_+^2\Big]\right)^{\frac{1}{2}} \\
&\qquad = \left(\mathbb{E}\left[\left(\left(1+\varepsilon\gamma - \boldsymbol{\mu}^T\boldsymbol{\theta}\right) + \left\|\boldsymbol{\Sigma}^{\frac{1}{2}}\boldsymbol{\theta}\right\|_{\ell_2} g\right)_+^2\right]\right)^{\frac{1}{2}}
\end{aligned}$$

Thus we arrive at

$$\min_{\boldsymbol{\theta},\gamma\geq 0}\max_{\beta\geq 0,\lambda\geq 0} \quad \|\boldsymbol{\theta}\|_{\ell_2}^2 + 2\lambda\Big(\|\boldsymbol{\theta}\|_{\ell_q} - \gamma\Big) + \frac{\beta}{\sqrt{n}}\boldsymbol{h}^T\mathsf{P}_{\boldsymbol{\mu}}^{\perp}\boldsymbol{\Sigma}^{1/2}\boldsymbol{\theta} + \beta\sqrt{\mathbb{E}\left[\left(\left(1+\varepsilon\gamma - \boldsymbol{\mu}^T\boldsymbol{\theta}\right) + \left\|\boldsymbol{\Sigma}^{\frac{1}{2}}\boldsymbol{\theta}\right\|_{\ell_2} g\right)_+^2\right]}. \tag{B.19}$$

We note that for $a \geq 0$,

$$\mathbb{E}[ag+b]_+^2 = \frac{a^2+b^2}{2}\left(1 + \operatorname{erf}\left(\frac{b}{\sqrt{2}a}\right)\right) + \frac{ab}{\sqrt{2\pi}}e^{-\frac{b^2}{2a^2}}.$$

and its derivative with respect to $a$ is given by $2a(1+\operatorname{erf}(\frac{b}{\sqrt{2}a})) > 0$ which implies that the function is increasing in $a > 0$. Therefore the optimization

(B.19) can be equivalently written as

$$\min_{\boldsymbol{\theta},\gamma,\alpha\geq 0}\ \max_{\beta\geq 0,\lambda\geq 0}\quad \|\boldsymbol{\theta}\|_{\ell_2}^2+2\lambda\Big(\|\boldsymbol{\theta}\|_{\ell_q}-\gamma\Big)+\frac{\beta}{\sqrt{n}}\boldsymbol{h}^T\mathsf{P}_{\boldsymbol{\mu}}^{\perp}\boldsymbol{\Sigma}^{1/2}\boldsymbol{\theta}+\beta\sqrt{\mathbb{E}\left[\left((1+\varepsilon\gamma-\boldsymbol{\mu}^T\boldsymbol{\theta})+\alpha g\right)_+^2\right]}$$

(B.20)

$$\text{subject to}\quad \left\|\boldsymbol{\Sigma}^{\frac{1}{2}}\boldsymbol{\theta}\right\|_{\ell_2}\leq\alpha.$$

Note that the above is trivially jointly convex in $(\boldsymbol{\theta},\gamma,\alpha)$ and jointly concave in $(\beta,\lambda)$. We fix the parallel component of $\boldsymbol{\theta}$ on $\boldsymbol{\mu}$ to $\theta$, namely $\theta=\widetilde{\boldsymbol{\mu}}^T\boldsymbol{\theta}$. We next optimize over $\boldsymbol{\theta}$ while fixing $\theta$.

$$\min_{\theta,\boldsymbol{\theta},\gamma\geq 0,\alpha\geq 0}\ \max_{\beta\geq 0,\lambda\geq 0}\quad \|\boldsymbol{\theta}\|_{\ell_2}^2+2\lambda\Big(\|\boldsymbol{\theta}\|_{\ell_q}-\gamma\Big)+\frac{\beta}{\sqrt{n}}\boldsymbol{h}^T\mathsf{P}_{\boldsymbol{\mu}}^{\perp}\boldsymbol{\Sigma}^{1/2}\boldsymbol{\theta}+\beta\sqrt{\mathbb{E}\left[\left(\left(1+\varepsilon\gamma-\theta\,\|\boldsymbol{\mu}\|_{\ell_2}\right)+\alpha g\right)_+^2\right]}$$

(B.21)

$$\text{subject to}\quad \left\|\boldsymbol{\Sigma}^{\frac{1}{2}}\boldsymbol{\theta}\right\|_{\ell_2}\leq\alpha,\quad \widetilde{\boldsymbol{\mu}}^T\boldsymbol{\theta}=\theta$$

Bringing the constraints into the objective via Lagrange multipliers we obtain

$$\min_{\theta,\boldsymbol{\theta},\gamma\geq 0,\alpha\geq 0}\ \max_{\beta,\lambda,\eta\geq 0,\widetilde{\eta}}\quad \|\boldsymbol{\theta}\|_{\ell_2}^2+2\lambda\Big(\|\boldsymbol{\theta}\|_{\ell_q}-\gamma\Big)+\frac{\beta}{\sqrt{n}}\boldsymbol{h}^T\mathsf{P}_{\boldsymbol{\mu}}^{\perp}\boldsymbol{\Sigma}^{1/2}\boldsymbol{\theta}+\beta\sqrt{\mathbb{E}\left[\left(\left(1+\varepsilon\gamma-\theta\,\|\boldsymbol{\mu}\|_{\ell_2}\right)+\alpha g\right)_+^2\right]}$$

(B.22)

$$+\eta\left(\left\|\boldsymbol{\Sigma}^{\frac{1}{2}}\boldsymbol{\theta}\right\|_{\ell_2}-\alpha\right)+\widetilde{\eta}\left(\widetilde{\boldsymbol{\mu}}^T\boldsymbol{\theta}-\theta\right)$$

Next note that $\left\|\boldsymbol{\Sigma}^{\frac{1}{2}}\boldsymbol{\theta}\right\|_{\ell_2}=\min_{\tau\geq 0}\frac{\left\|\boldsymbol{\Sigma}^{\frac{1}{2}}\boldsymbol{\theta}\right\|_{\ell_2}^2}{2\tau}+\frac{\tau}{2}$ and $\|\boldsymbol{\theta}\|_{\ell_q}=\min_{t\geq 0}\frac{\|\boldsymbol{\theta}\|_{\ell_q}^q}{qt^{q-1}}+\frac{q-1}{q}t$

Thus, above reduces to

$$\min_{\theta,\boldsymbol{\theta},\gamma\geq 0,\alpha\geq 0}\ \max_{\beta,\lambda,\eta\geq 0,\widetilde{\eta}}\ \min_{\tau\geq 0,t\geq 0}\quad \|\boldsymbol{\theta}\|_{\ell_2}^2+\frac{2\lambda}{qt^{q-1}}\|\boldsymbol{\theta}\|_{\ell_q}^q+2\lambda\frac{q-1}{q}t-2\lambda\gamma+\frac{\beta}{\sqrt{n}}\boldsymbol{h}^T\mathsf{P}_{\boldsymbol{\mu}}^{\perp}\boldsymbol{\Sigma}^{1/2}\boldsymbol{\theta}$$
$$+\beta\sqrt{\mathbb{E}\left[\left(\left(1+\varepsilon\gamma-\theta\,\|\boldsymbol{\mu}\|_{\ell_2}\right)+\alpha g\right)_+^2\right]}$$

(B.23)

$$+\frac{\eta}{2\tau}\left\|\boldsymbol{\Sigma}^{\frac{1}{2}}\boldsymbol{\theta}\right\|_{\ell_2}^2+\frac{\eta\tau}{2}-\eta\alpha+\widetilde{\eta}\left(\widetilde{\boldsymbol{\mu}}^T\boldsymbol{\theta}-\theta\right)$$

To continue note that $\frac{\|\boldsymbol{\theta}\|_{\ell_q}^q}{t^{q-1}}=t\left\|\frac{\boldsymbol{\theta}}{t}\right\|_{\ell_q}^q$ and $\frac{\left\|\boldsymbol{\Sigma}^{\frac{1}{2}}\boldsymbol{\theta}\right\|_{\ell_2}^2}{\tau}=\tau\left\|\boldsymbol{\Sigma}^{\frac{1}{2}}\frac{\boldsymbol{\theta}}{\tau}\right\|_{\ell_2}^2$ and thus using the fact that the perspective of a convex function is convex both are

jointly convex with respect to $(\boldsymbol{\theta}, t)$ and $(\boldsymbol{\theta}, \tau)$. Thus the objective above is jointly convex in $(\boldsymbol{\theta}, \theta, \gamma, \alpha, t, \tau)$ and jointly concave in $(\beta, \lambda, \eta, \widetilde{\eta})$. Due to this convexity/concavity with respect to the minimization/maximization parameters we can change the order of min and max. We thus proceed by optimizing over $\boldsymbol{\theta}$. The optimization over $\boldsymbol{\theta}$ takes the form

$$\min_{\boldsymbol{\theta}} \quad \|\boldsymbol{\theta}\|_{\ell_2}^2 + \frac{2\lambda}{qt^{q-1}} \|\boldsymbol{\theta}\|_{\ell_q}^q + \frac{\eta}{2\tau} \left\|\boldsymbol{\Sigma}^{\frac{1}{2}}\boldsymbol{\theta}\right\|_{\ell_2}^2 + \frac{\beta}{\sqrt{n}} \boldsymbol{h}^T \mathsf{P}_{\boldsymbol{\mu}}^{\perp} \boldsymbol{\Sigma}^{1/2} \boldsymbol{\theta} + \widetilde{\eta} \widetilde{\boldsymbol{\mu}}^T \boldsymbol{\theta} \tag{B.24}$$

By completing the square the objective can be alternatively written as

$$\begin{aligned}
&\boldsymbol{\theta}^T \left(\boldsymbol{I} + \frac{\eta}{2\tau}\boldsymbol{\Sigma}\right)\boldsymbol{\theta} + \frac{2\lambda}{qt^{q-1}} \|\boldsymbol{\theta}\|_{\ell_q}^q + \frac{\beta}{\sqrt{n}} \boldsymbol{h}^T \mathsf{P}_{\boldsymbol{\mu}}^{\perp} \boldsymbol{\Sigma}^{1/2} \boldsymbol{\theta} + \widetilde{\eta} \widetilde{\boldsymbol{\mu}}^T \boldsymbol{\theta} \\
&= \left\| \left(\boldsymbol{I} + \frac{\eta}{2\tau}\boldsymbol{\Sigma}\right)^{1/2} \boldsymbol{\theta} + \frac{\beta}{2\sqrt{n}} \left(\boldsymbol{I} + \frac{\eta}{2\tau}\boldsymbol{\Sigma}\right)^{-1/2} \boldsymbol{\Sigma}^{1/2} \mathsf{P}_{\boldsymbol{\mu}}^{\perp} \boldsymbol{h} + \left(\boldsymbol{I} + \frac{\eta}{2\tau}\boldsymbol{\Sigma}\right)^{-1/2} \frac{\tilde{\eta}}{2} \widetilde{\boldsymbol{\mu}} \right\|_{\ell_2}^2 \\
&+ \frac{2\lambda}{qt^{q-1}} \|\boldsymbol{\theta}\|_{\ell_q}^q - \frac{\beta^2}{4n} \boldsymbol{h}^T \mathsf{P}_{\boldsymbol{\mu}}^{\perp} \boldsymbol{\Sigma}^{1/2} \left(\boldsymbol{I} + \frac{\eta}{2\tau}\boldsymbol{\Sigma}\right)^{-1} \boldsymbol{\Sigma}^{1/2} \mathsf{P}_{\boldsymbol{\mu}}^{\perp} \boldsymbol{h} \\
&- \frac{\tilde{\eta}^2}{4} \widetilde{\boldsymbol{\mu}}^T \left(\boldsymbol{I} + \frac{\eta}{2\tau}\boldsymbol{\Sigma}\right)^{-1} \widetilde{\boldsymbol{\mu}} - \frac{\beta\tilde{\eta}}{2\sqrt{n}} \widetilde{\boldsymbol{\mu}}^T \left(\boldsymbol{I} + \frac{\eta}{2\tau}\boldsymbol{\Sigma}\right)^{-1} \boldsymbol{\Sigma}^{1/2} \mathsf{P}_{\boldsymbol{\mu}}^{\perp} \boldsymbol{h} .
\end{aligned} \tag{B.25}$$

Since $\boldsymbol{\Sigma}\widetilde{\boldsymbol{\mu}} = a^2 \widetilde{\boldsymbol{\mu}}$ we have

$$\widetilde{\boldsymbol{\mu}}^T \left(\boldsymbol{I} + \frac{\eta}{2\tau}\boldsymbol{\Sigma}\right)^{-1} \boldsymbol{\Sigma}^{1/2} \mathsf{P}_{\boldsymbol{\mu}}^{\perp} \boldsymbol{h} = 0, \quad \widetilde{\boldsymbol{\mu}}^T \left(\boldsymbol{I} + \frac{\eta}{2\tau}\boldsymbol{\Sigma}\right)^{-1} \widetilde{\boldsymbol{\mu}} = \frac{1}{(1 + \frac{\eta}{2\tau}a^2)} .$$

We consider a singular value decomposition $\boldsymbol{\Sigma} = \boldsymbol{U}\boldsymbol{S}\boldsymbol{U}^T$ with $\boldsymbol{S} = \mathrm{diag}(s_1, \ldots, s_d)$, and the first column of $\boldsymbol{U}$ being $\widetilde{\boldsymbol{\mu}}$ and $s_1 = a^2$ (Recall that $\widetilde{\boldsymbol{\mu}}$ is a singular value of $\boldsymbol{\Sigma}$ with eigenvalue $a^2$.) Then,

$$\begin{aligned}
\frac{1}{n} \boldsymbol{h}^T \mathsf{P}_{\boldsymbol{\mu}}^{\perp} \boldsymbol{\Sigma}^{1/2} \left(\boldsymbol{I} + \frac{\eta}{2\tau}\boldsymbol{\Sigma}\right)^{-1} \boldsymbol{\Sigma}^{1/2} \mathsf{P}_{\boldsymbol{\mu}}^{\perp} \boldsymbol{h} &= \frac{1}{n} \boldsymbol{h}^T \mathsf{P}_{\boldsymbol{\mu}}^{\perp} \boldsymbol{U} \left(\boldsymbol{I} + \frac{\eta}{2\tau}\boldsymbol{S}\right)^{-1} \boldsymbol{S}\boldsymbol{U}^T \mathsf{P}_{\boldsymbol{\mu}}^{\perp} \boldsymbol{h} \\
&= \frac{1}{\delta d} \sum_{i=2}^{d} \frac{s_i}{1 + \frac{\eta}{2\tau} s_i} h_i^2 \\
&\overset{P}{\Rightarrow} \frac{1}{\delta d} \sum_{i=1}^{d} \frac{s_i}{1 + \frac{\eta}{2\tau} s_i} \\
&= \frac{2\tau}{\delta d \eta} \sum_{i=1}^{d} \left(1 - \frac{1}{\frac{\eta}{2\tau}(s_i + \frac{2\tau}{\eta})}\right) \\
&= \frac{2\tau}{\delta\eta} \left(1 + \frac{2\tau}{\eta} S_\rho\left(-\frac{2\tau}{\eta}\right)\right)
\end{aligned}$$

with $S_\rho(z) := \int \frac{\rho(t)}{z-t}\mathrm{d}t$ the Stieltjes transform of the spectrum of $\boldsymbol{\Sigma}$.

Using the above identities (B.25) reduces to

$$\left\|\left(\boldsymbol{I}+\frac{\eta}{2\tau}\boldsymbol{\Sigma}\right)^{1/2}\boldsymbol{\theta}+\frac{\beta}{2\sqrt{n}}\left(\boldsymbol{I}+\frac{\eta}{2\tau}\boldsymbol{\Sigma}\right)^{-1/2}\boldsymbol{\Sigma}^{1/2}\mathsf{P}_{\boldsymbol{\mu}}^{\perp}\boldsymbol{h}+\left(\boldsymbol{I}+\frac{\eta}{2\tau}\boldsymbol{\Sigma}\right)^{-1/2}\frac{\tilde{\eta}}{2}\widetilde{\boldsymbol{\mu}}\right\|_{\ell_2}^2$$
$$+\frac{2\lambda}{qt^{q-1}}\|\boldsymbol{\theta}\|_{\ell_q}^q-\frac{\tau\beta^2}{2\delta\eta}\left(1+\frac{2\tau}{\eta}S_\rho\left(-\frac{2\tau}{\eta}\right)\right)-\frac{\tilde{\eta}^2}{4(1+\frac{\eta}{2\tau}a^2)}\,.$$

We then write the minimum value over $\boldsymbol{\theta}$ in terms of the weighted Moreau envelope, given by Definition 6.2.

$$\min_{\boldsymbol{\theta}}\left\|\left(\boldsymbol{I}+\frac{\eta}{2\tau}\boldsymbol{\Sigma}\right)^{1/2}\boldsymbol{\theta}+\frac{\beta}{2\sqrt{n}}\left(\boldsymbol{I}+\frac{\eta}{2\tau}\boldsymbol{\Sigma}\right)^{-1/2}\boldsymbol{\Sigma}^{1/2}\mathsf{P}_{\boldsymbol{\mu}}^{\perp}\boldsymbol{h}+\left(\boldsymbol{I}+\frac{\eta}{2\tau}\boldsymbol{\Sigma}\right)^{-1/2}\frac{\tilde{\eta}}{2}\widetilde{\boldsymbol{\mu}}\right\|_{\ell_2}^2+\frac{2\lambda}{qt^{q-1}}\|\boldsymbol{\theta}\|_{\ell_q}^q$$

(B.26)

$$=2e_{q,\boldsymbol{I}+\frac{\eta}{2\tau}\boldsymbol{\Sigma}}\left(\left(\boldsymbol{I}+\frac{\eta}{2\tau}\boldsymbol{\Sigma}\right)^{-1}\left\{\frac{\beta}{2\sqrt{n}}\boldsymbol{\Sigma}^{1/2}\mathsf{P}_{\boldsymbol{\mu}}^{\perp}\boldsymbol{h}-\frac{\tilde{\eta}}{2}\widetilde{\boldsymbol{\mu}}\right\};\frac{\lambda}{qt^{q-1}}\right),$$

where we used symmetry of the distribution of $\boldsymbol{h}$.

Putting all pieces together in(B.25) we get

(B.27)

$$\min_{\boldsymbol{\theta}}\|\boldsymbol{\theta}\|_{\ell_2}^2+\frac{2\lambda}{qt^{q-1}}\|\boldsymbol{\theta}\|_{\ell_q}^q+\frac{\eta}{2\tau}\left\|\boldsymbol{\Sigma}^{\frac{1}{2}}\boldsymbol{\theta}\right\|_{\ell_2}^2+\frac{\beta}{\sqrt{n}}\boldsymbol{h}^T\mathsf{P}_{\boldsymbol{\mu}}^{\perp}\boldsymbol{\Sigma}^{1/2}\boldsymbol{\theta}+\widetilde{\eta}\widetilde{\boldsymbol{\mu}}^T\boldsymbol{\theta}$$
$$=2e_{q,\boldsymbol{I}+\frac{\eta}{2\tau}\boldsymbol{\Sigma}}\left(\left(\boldsymbol{I}+\frac{\eta}{2\tau}\boldsymbol{\Sigma}\right)^{-1}\left\{\frac{\beta}{2\sqrt{n}}\boldsymbol{\Sigma}^{1/2}\mathsf{P}_{\boldsymbol{\mu}}^{\perp}\boldsymbol{h}-\frac{\tilde{\eta}}{2}\widetilde{\boldsymbol{\mu}}\right\};\frac{\lambda}{qt^{q-1}}\right)-\frac{\beta^2\tau}{2\delta\eta}\left(1+\frac{2\tau}{\eta}S_\rho\left(-\frac{2\tau}{\eta}\right)\right)$$
$$-\frac{\tilde{\eta}^2}{4(1+\frac{\eta}{2\tau}a^2)}\,.$$

Using (B.27) in (B.23), the AO problem reduces to

$$\min_{\gamma\geq 0,\theta}\max_{\beta,\lambda,\eta\geq 0,\tilde{\eta}}\min_{\tau\geq 0,t\geq 0}\quad 2e_{q,\boldsymbol{I}+\frac{\eta}{2\tau}\boldsymbol{\Sigma}}\left(\left(\boldsymbol{I}+\frac{\eta}{2\tau}\boldsymbol{\Sigma}\right)^{-1}\left\{\frac{\beta}{2\sqrt{n}}\boldsymbol{\Sigma}^{1/2}\mathsf{P}_{\boldsymbol{\mu}}^{\perp}\boldsymbol{h}-\frac{\tilde{\eta}}{2}\widetilde{\boldsymbol{\mu}}\right\};\frac{\lambda}{qt^{q-1}}\right)$$
$$-\frac{\beta^2\tau}{2\delta\eta}\left(1+\frac{2\tau}{\eta}S_\rho\left(-\frac{2\tau}{\eta}\right)\right)$$
$$-\frac{\tilde{\eta}^2}{4(1+\frac{\eta}{2\tau}a^2)}+\beta\sqrt{\mathbb{E}\left[\left(\left(1+\varepsilon\gamma-\theta\|\boldsymbol{\mu}\|_{\ell_2}\right)+\alpha g\right)_+^2\right]}$$
$$+2\lambda\frac{q-1}{q}t-2\lambda\gamma+\frac{\eta\tau}{2}-\eta\alpha-\tilde{\eta}\theta \qquad \text{(B.28)}$$

**Scalarization of the auxiliary optimization problem.** We proceed by defining $\lambda_0 := \frac{\lambda}{\|\boldsymbol{\mu}\|_{\ell_p}}$, $\gamma_0 := \gamma \|\boldsymbol{\mu}\|_{\ell_p}$ and $t_0 := t \|\boldsymbol{\mu}\|_{\ell_p}$. Under Assumptions 2 and 6, the asymptotic auxiliary optimization (AO) problem becomes

$$\begin{aligned}\min_{\alpha,\gamma_0\geq 0,\theta}\max_{\beta,\lambda_0,\eta\geq 0,\tilde{\eta}}\min_{\tau\geq 0,t_0\geq 0} \quad & 2\mathsf{F}\left(\beta,\tilde{\eta};\frac{\eta}{2\tau},\frac{\lambda_0}{qt_0^{q-1}}\right)-\frac{\beta^2\tau}{2\delta\eta}\left(1+\frac{2\tau}{\eta}S_\rho\left(-\frac{2\tau}{\eta}\right)\right)\\ & +2\lambda_0\frac{q-1}{q}t_0-2\lambda_0\gamma_0+\frac{\eta\tau}{2}-\eta\alpha-\tilde{\eta}\theta\\ & -\frac{\tilde{\eta}^2}{4(1+\frac{\eta}{2\tau}a^2)}+\beta\sqrt{\mathbb{E}\left[\left((1+\varepsilon_0\gamma_0-\theta V)+\alpha g\right)_+^2\right]}\end{aligned} \tag{B.29}$$

Here we used the relation $\varepsilon = \varepsilon_0 \|\boldsymbol{\mu}\|_{\ell_p}$.

We next solve for some of the variables in the AO problem by writing the KKT conditions.

1. Define

$$f\left(\frac{\eta}{\tau}\right) := 2\mathsf{F}\left(\beta,\tilde{\eta};\frac{\eta}{2\tau},\frac{\lambda_0}{qt_0^{q-1}}\right)-\frac{\beta^2\tau}{2\delta\eta}\left(1+\frac{2\tau}{\eta}S_\rho\left(-\frac{2\tau}{\eta}\right)\right)-\frac{\tilde{\eta}^2}{4(1+\frac{\eta}{2\tau}a^2)},$$

where we only made the dependence on $\frac{\eta}{\tau}$ explicit in the notation $f\left(\frac{\eta}{\tau}\right)$. Setting derivative with respect to $\eta$ to zero, we obtain

$$\frac{1}{\tau}f'\left(\frac{\eta}{\tau}\right)+\frac{\tau}{2}-\alpha=0\,. \tag{B.30}$$

Setting derivative with respect to $\tau$ to zero, we obtain

$$-\frac{\eta}{\tau^2}f'\left(\frac{\eta}{\tau}\right)+\frac{\eta}{2}=0\,. \tag{B.31}$$

Combining (B.30) and (B.31), we get $\eta(1-\frac{\alpha}{\tau})=0$. So either $\alpha=\tau$ or $\eta=0$. If $\eta=0$, then it is clear that the terms involving $\tau$ in the AO problem would vanish and therefore the value of $\tau$ does not matter. So in this case, we can as well assume $\tau=\alpha$. This simplifies the AO problem by replacing for $\tau$:

$$\begin{aligned}\min_{\alpha,\gamma_0\geq 0,\theta}\max_{\beta,\lambda_0,\eta\geq 0,\tilde{\eta}}\min_{t_0\geq 0} \quad & 2\mathsf{F}\left(\beta,\tilde{\eta};\frac{\eta}{2\alpha},\frac{\lambda_0}{qt_0^{q-1}}\right)-\frac{\beta^2\alpha}{2\delta\eta}\left(1+\frac{2\alpha}{\eta}S_\rho\left(-\frac{2\alpha}{\eta}\right)\right)\\ & +2\lambda_0\frac{q-1}{q}t_0-2\lambda_0\gamma_0-\frac{\eta\alpha}{2}-\tilde{\eta}\theta\\ & -\frac{\tilde{\eta}^2}{4(1+\frac{\eta}{2\alpha}a^2)}+\beta\sqrt{\mathbb{E}\left[\left((1+\varepsilon_0\gamma_0-\theta V)+\alpha g\right)_+^2\right]}.\end{aligned} \tag{B.32}$$

2. Setting derivative with respect to $\lambda_0$ to zero, we get

$$2\mathsf{F}_4'\left(\beta,\tilde{\eta};\frac{\eta}{2\alpha},\frac{\lambda_0}{qt_0^{q-1}}\right)\frac{1}{qt_0^{q-1}}+2\frac{q-1}{q}t_0-2\gamma_0=0\,, \tag{B.33}$$

where $\mathsf{F}_4'$ denotes the derivative of function $\mathsf{F}$ with respect to its forth argument. Also, by setting derivative with respect to $t_0$ to zero we get

$$2\mathsf{F}_4'\left(\beta,\tilde{\eta};\frac{\eta}{2\alpha},\frac{\lambda_0}{qt_0^{q-1}}\right)\lambda_0\frac{1-q}{q}t_0^{-q}+2\lambda_0\frac{q-1}{q}=0\,. \tag{B.34}$$

Combining (B.33) and (B.34) implies that

$$2\lambda_0(q-1)\left(\frac{\gamma_0}{t_0}-1\right)=0\,. \tag{B.35}$$

Therefore either $\gamma_0=t_0$ or $\lambda_0=0$ or $q=1$. If $\lambda=0$ or $q=1$ then the terms involving $t_0$ in (B.32) vanish and hence we can assume $t_0=\gamma_0$ in this cases as well. Replacing $t_0$ with $\gamma_0$ in (B.32) we obtain

$$\begin{aligned}\min_{\alpha,\gamma_0\geq 0,\theta}\max_{\beta,\lambda_0,\eta\geq 0,\tilde{\eta}}\quad & 2\mathsf{F}\left(\beta,\tilde{\eta};\frac{\eta}{2\alpha},\frac{\lambda_0}{q\gamma_0^{q-1}}\right)-\frac{\beta^2\alpha}{2\delta\eta}\left(1+\frac{2\alpha}{\eta}S_\rho\left(-\frac{2\alpha}{\eta}\right)\right)-\frac{2\lambda_0}{q}\gamma_0-\frac{\eta\alpha}{2}-\tilde{\eta}\theta\\ & -\frac{\tilde{\eta}^2}{4(1+\frac{\eta}{2\alpha}a^2)}+\beta\sqrt{\mathbb{E}\left[\left((1+\varepsilon_0\gamma_0-\theta V)+\alpha g\right)_+^2\right]}\,,\end{aligned} \tag{B.36}$$

which is the expression for $D_{\mathrm{s}}(\alpha,\gamma_0,\theta,\beta,\lambda_0,\eta,\tilde{\eta})$ given by (6.2).

**Uniqueness and boundedness of the solution to AO problem.** Note that since $\delta\leq\delta_*$, by using Theorem 6.1, we are in the separable regime and therefore optimization (F.2) is feasible with high probability and admits a bounded solution. This implies that the PO problem (B.17) has bounded solution and since AO and PO problems are asymptotically equivalent this implies that the AO problem (B.36) has bounded solution.

To show the uniqueness of the solution of (B.36), note that as we argued throughout the proof, its objective function $D_{\mathrm{s}}$ is jointly strictly convex in $(\alpha,\gamma_0,\theta)$ and jointly concave in $(\beta,\lambda_0,\eta,\tilde{\eta})$. Therefor, $\max_{\beta,\lambda_0,\eta,\tilde{\eta}}D_{\mathrm{s}}(\alpha,\gamma_0,\theta,\beta,\lambda_0,\eta,\tilde{\eta})$ is strictly convex in $(\alpha,\gamma_0,\theta)$. This follows from the fact that if a function $f(\boldsymbol{x},\boldsymbol{y})$ is strictly convex in $\boldsymbol{x}$, then $\max_{\boldsymbol{y}}f(\boldsymbol{x},\boldsymbol{y})$ is also strictly convex in $\boldsymbol{x}$ and therefore has a unique minimizer $(\alpha_*,\gamma_{0*},\theta_*)$.

Part (b) of the theorem follows readily from our definition of parameters $\alpha$, $\theta$ and $\gamma$.

Part (c) also follows from combining Lemma 2.1 with part (b) of the theorem.

**B.3. Proof of Theorem 6.4** The goal of this theorem is to derive precise asymptotic behavior for the adversarially trained model $\widehat{\boldsymbol{\theta}}^{\varepsilon}$ given by

$$\widehat{\boldsymbol{\theta}}^{\varepsilon} = \arg\min_{\boldsymbol{\theta}\in\mathbb{R}^d} \frac{1}{n}\sum_{i=1}^{n} \ell\left(y_i\langle \boldsymbol{x}_i, \boldsymbol{\theta}\rangle - \varepsilon \left\|\boldsymbol{\theta}\right\|_{\ell_q}\right). \tag{B.37}$$

Letting $v_i := y_i\langle \boldsymbol{x}_i, \boldsymbol{\theta}\rangle$, this optimization can be equivalently written as

$$\min_{\boldsymbol{\theta},\boldsymbol{v}\in\mathbb{R}^n} \frac{1}{2p}\sum_{i=1}^{n} \ell\left(v_i - \varepsilon \left\|\boldsymbol{\theta}\right\|_{\ell_q}\right) \quad \text{subject to } \boldsymbol{v} = \boldsymbol{D}_{\boldsymbol{y}}\boldsymbol{X}\boldsymbol{\theta},$$

with $\boldsymbol{D}_{\boldsymbol{y}} = \mathrm{diag}(y_1, \ldots, y_n)$. Therefore, by writing the Lagrangian by $\boldsymbol{u}/n$ as the dual variable for the equality constraint, we arrive at

$$\min_{\boldsymbol{\theta},\boldsymbol{v}\in\mathbb{R}^n} \max_{\boldsymbol{u}\in\mathbb{R}^n} \frac{1}{n}\left\{\boldsymbol{u}^{\mathsf{T}}\boldsymbol{D}_{\boldsymbol{y}}\boldsymbol{X}\boldsymbol{\theta} - \boldsymbol{u}^{\mathsf{T}}\boldsymbol{v}\right\} + \frac{1}{n}\sum_{i=1}^{n} \ell\left(v_i - \varepsilon \left\|\boldsymbol{\theta}\right\|_{\ell_q}\right)$$

We next substitute for $\boldsymbol{X} = \boldsymbol{y}\boldsymbol{\mu}^T + \boldsymbol{Z}\boldsymbol{\Sigma}^{1/2}$, under the Gaussian mixtures model, which gives us

$$\min_{\boldsymbol{\theta},\boldsymbol{v}\in\mathbb{R}^n} \max_{\boldsymbol{u}\in\mathbb{R}^n} \frac{1}{n}\left\{\boldsymbol{u}^{\mathsf{T}}\mathbf{1}\boldsymbol{\mu}^T\boldsymbol{\theta} + \boldsymbol{u}^{\mathsf{T}}\boldsymbol{D}_{\boldsymbol{y}}\boldsymbol{Z}\boldsymbol{\Sigma}^{1/2}\boldsymbol{\theta} - \boldsymbol{u}^{\mathsf{T}}\boldsymbol{v}\right\} + \frac{1}{n}\sum_{i=1}^{n} \ell\left(v_i - \varepsilon \left\|\boldsymbol{\theta}\right\|_{\ell_q}\right) \tag{B.38}$$

Note that by the above Lagrangian is in a minimax problem in the form of $\min_{\boldsymbol{\theta}} \max_{\boldsymbol{u}}\ \boldsymbol{u}^T\boldsymbol{Z}\boldsymbol{\theta} + \psi(\boldsymbol{\theta}, \boldsymbol{u})$, with $\boldsymbol{Z}$ standard Gaussian matrix and $\psi(\boldsymbol{\theta}, \boldsymbol{u})$ is convex in the minimization variable $\boldsymbol{\theta}$ and concave in the maximization variable $\boldsymbol{u}$. This form allows us to apply the CGMT framework as outlined in Section 7 and similar to the proof of Theorem 6.3. But in order to do that, we need the minimization/maximization to be over compact sets. Similar to the proof of Theorem 6.3 we cope with this technical issue by introducing artificial boundedness constraints on the optimization variables that they do not change the optimal solution. Specifically, we can add constraints of the form $\mathcal{S}_{\boldsymbol{\theta}} = \{\boldsymbol{\theta} :\ \left\|\boldsymbol{\theta}\right\|_{\ell_q} \le K_{\boldsymbol{\theta}}\}$ and $\mathcal{S}_{\boldsymbol{u}} = \{\boldsymbol{u} :\ \left\|\boldsymbol{u}\right\|_{\infty} \le K_{\boldsymbol{u}}\}$ for sufficiently large constants $K_{\boldsymbol{\theta}}, K_{\boldsymbol{u}}$, without changing the optimal solution of (B.38). We refer to Appendix E.3.1 for precise statements and proofs. This allows us to

replace (B.38) with

$$\min_{\boldsymbol{\theta}\in\mathcal{S}_{\boldsymbol{\theta}},\boldsymbol{v}\in\mathbb{R}^n}\max_{\boldsymbol{u}\in\boldsymbol{\mathcal{S}}_{\boldsymbol{u}}}\frac{1}{n}\Big\{\boldsymbol{u}^{\mathsf{T}}\mathbf{1}\boldsymbol{\mu}^T\boldsymbol{\theta}+\boldsymbol{u}^{\mathsf{T}}\boldsymbol{D}_{\boldsymbol{y}}\boldsymbol{Z}\boldsymbol{\Sigma}^{1/2}\boldsymbol{\theta}-\boldsymbol{u}^{\mathsf{T}}\boldsymbol{v}\Big\}+\frac{1}{n}\sum_{i=1}^{n}\ell\left(v_i-\varepsilon\left\|\boldsymbol{\theta}\right\|_{\ell_q}\right). \tag{B.39}$$

B.3.1. *Reduction to an auxiliary optimization problem via CGMT* Next we define the projection matrices

$$\mathsf{P}_{\boldsymbol{\mu}}^{\perp}:=\boldsymbol{I}-\widetilde{\boldsymbol{\mu}}\widetilde{\boldsymbol{\mu}}^T,\quad \mathsf{P}_{\boldsymbol{\mu}}:=\widetilde{\boldsymbol{\mu}}\widetilde{\boldsymbol{\mu}}^T$$

and rewrite $\boldsymbol{Z}\boldsymbol{\Sigma}^{1/2}=\boldsymbol{Z}\left(\mathsf{P}_{\boldsymbol{\mu}}+\mathsf{P}_{\boldsymbol{\mu}}^{\perp}\right)\boldsymbol{\Sigma}^{1/2}$. Since $\boldsymbol{Z}\mathsf{P}_{\boldsymbol{\mu}}$ and $\boldsymbol{Z}\mathsf{P}_{\boldsymbol{\mu}}^{\perp}$ are independent from each other the latter has the same distribution as

$$\boldsymbol{Z}\boldsymbol{\Sigma}^{1/2}\sim\boldsymbol{z}\left(\boldsymbol{\Sigma}^{1/2}\widetilde{\boldsymbol{\mu}}\right)^T+\boldsymbol{Z}\mathsf{P}_{\boldsymbol{\mu}}^{\perp}\boldsymbol{\Sigma}^{1/2}. \tag{B.40}$$

where $\boldsymbol{z}\sim\mathcal{N}(0,\boldsymbol{I}_n)$ and is independent from the matrix $\boldsymbol{Z}$. Thus the above optimization problem is equivalent to

$$\min_{\boldsymbol{\theta}\in\mathcal{S}_{\boldsymbol{\theta}},\boldsymbol{v}}\max_{\boldsymbol{u}\in\mathcal{S}_{\boldsymbol{u}}}\frac{1}{n}\Big\{\boldsymbol{u}^{\mathsf{T}}\mathbf{1}\boldsymbol{\mu}^T\boldsymbol{\theta}+(\boldsymbol{u}^{\mathsf{T}}\boldsymbol{D}_{\boldsymbol{y}}\boldsymbol{z})(\widetilde{\boldsymbol{\mu}}^T\boldsymbol{\Sigma}^{1/2}\boldsymbol{\theta})+\boldsymbol{u}^{\mathsf{T}}\boldsymbol{D}_{\boldsymbol{y}}\boldsymbol{Z}\mathsf{P}_{\boldsymbol{\mu}}^{\perp}\boldsymbol{\Sigma}^{1/2}\boldsymbol{\theta}-\boldsymbol{u}^{\mathsf{T}}\boldsymbol{v}\Big\}+\frac{1}{n}\sum_{i=1}^{n}\ell\left(v_i-\varepsilon\left\|\boldsymbol{\theta}\right\|_{\ell_q}\right). \tag{B.41}$$

Using CGMT and the corresponding AO takes the form

$$\begin{aligned}\min_{\boldsymbol{\theta}\in\mathcal{S}_{\boldsymbol{\theta}},\boldsymbol{v}}\max_{\boldsymbol{u}\in\mathcal{S}_{\boldsymbol{u}}}\frac{1}{n}\Big\{&\left\|\mathsf{P}_{\boldsymbol{\mu}}^{\perp}\boldsymbol{\Sigma}^{1/2}\boldsymbol{\theta}\right\|_{\ell_2}\boldsymbol{g}^T\boldsymbol{D}_{\boldsymbol{y}}\boldsymbol{u}+\left\|\boldsymbol{D}_{\boldsymbol{y}}\boldsymbol{u}\right\|_{\ell_2}\boldsymbol{h}^T\mathsf{P}_{\boldsymbol{\mu}}^{\perp}\boldsymbol{\Sigma}^{1/2}\boldsymbol{\theta}+(\boldsymbol{u}^{\mathsf{T}}\boldsymbol{D}_{\boldsymbol{y}}\boldsymbol{z})(\widetilde{\boldsymbol{\mu}}^T\boldsymbol{\Sigma}^{1/2}\boldsymbol{\theta})\\&+\boldsymbol{u}^{\mathsf{T}}\mathbf{1}\boldsymbol{\mu}^T\boldsymbol{\theta}-\boldsymbol{u}^{\mathsf{T}}\boldsymbol{v}\Big\}+\frac{1}{n}\sum_{i=1}^{n}\ell\left(v_i-\varepsilon\left\|\boldsymbol{\theta}\right\|_{\ell_q}\right),\end{aligned} \tag{B.42}$$

where $\boldsymbol{g}\sim\mathsf{N}(0,\boldsymbol{I}_n)$ and $\boldsymbol{h}\sim\mathsf{N}(0,\boldsymbol{I}_d)$.

Given $y_i=\pm1$ are independent of $\boldsymbol{Z}$ and hence $\boldsymbol{g}$, we have $\boldsymbol{D}_{\boldsymbol{y}}\boldsymbol{g},\boldsymbol{D}_{\boldsymbol{y}}\boldsymbol{z}\sim\mathsf{N}(0,\boldsymbol{I}_n)$ and $\left\|\boldsymbol{D}_{\boldsymbol{y}}\boldsymbol{u}\right\|_{\ell_2}=\left\|\boldsymbol{u}\right\|_{\ell_2}$. This results in

$$\begin{aligned}\min_{\boldsymbol{\theta}\in\mathcal{S}_{\boldsymbol{\theta}},\boldsymbol{v}}\max_{\boldsymbol{u}\in\mathcal{S}_{\boldsymbol{u}}}\frac{1}{n}\Big\{&\left\|\mathsf{P}_{\boldsymbol{\mu}}^{\perp}\boldsymbol{\Sigma}^{1/2}\boldsymbol{\theta}\right\|_{\ell_2}\boldsymbol{g}^T\boldsymbol{u}+\left\|\boldsymbol{u}\right\|_{\ell_2}\boldsymbol{h}^T\mathsf{P}_{\boldsymbol{\mu}}^{\perp}\boldsymbol{\Sigma}^{1/2}\boldsymbol{\theta}+(\boldsymbol{u}^{\mathsf{T}}\boldsymbol{z})(\widetilde{\boldsymbol{\mu}}^T\boldsymbol{\Sigma}^{1/2}\boldsymbol{\theta})+\boldsymbol{u}^{\mathsf{T}}\mathbf{1}\boldsymbol{\mu}^T\boldsymbol{\theta}-\boldsymbol{u}^{\mathsf{T}}\boldsymbol{v}\Big\}\\&+\frac{1}{n}\sum_{i=1}^{n}\ell\left(v_i-\varepsilon\left\|\boldsymbol{\theta}\right\|_{\ell_q}\right).\end{aligned} \tag{B.43}$$

Letting $\beta := \frac{1}{\sqrt{n}} \|\boldsymbol{u}\|_{\ell_2}$ and optimizing over direction of $\boldsymbol{u}$, we get

$$\max_{\boldsymbol{u}\in\mathcal{S}_{\boldsymbol{u}}} \frac{1}{n}\left(\left\|\mathsf{P}_{\boldsymbol{\mu}}^{\perp}\boldsymbol{\Sigma}^{1/2}\boldsymbol{\theta}\right\|_{\ell_2}\boldsymbol{g}^T\boldsymbol{u} + \|\boldsymbol{u}\|_{\ell_2}\boldsymbol{h}^T\mathsf{P}_{\boldsymbol{\mu}}^{\perp}\boldsymbol{\Sigma}^{1/2}\boldsymbol{\theta} + (\boldsymbol{u}^{\mathsf{T}}\boldsymbol{z})(\widetilde{\boldsymbol{\mu}}^T\boldsymbol{\Sigma}^{1/2}\boldsymbol{\theta}) + \boldsymbol{u}^{\mathsf{T}}\mathbf{1}\boldsymbol{\mu}^T\boldsymbol{\theta} - \boldsymbol{u}^{\mathsf{T}}\boldsymbol{v}\right)$$

(B.44)

$$= \max_{0\le\beta\le K} \frac{\beta}{\sqrt{n}}\left\|\left\|\mathsf{P}_{\boldsymbol{\mu}}^{\perp}\boldsymbol{\Sigma}^{1/2}\boldsymbol{\theta}\right\|_{\ell_2}\boldsymbol{g} + \widetilde{\boldsymbol{\mu}}^T\boldsymbol{\Sigma}^{1/2}\boldsymbol{\theta}\boldsymbol{z} + \mathbf{1}\boldsymbol{\mu}^T\boldsymbol{\theta} - \boldsymbol{v}\right\|_{\ell_2} + \frac{\beta}{\sqrt{n}}\boldsymbol{h}^T\mathsf{P}_{\boldsymbol{\mu}}^{\perp}\boldsymbol{\Sigma}^{1/2}\boldsymbol{\theta}\,,$$

where $K := \max_{\boldsymbol{u}\in\mathcal{S}_{\boldsymbol{u}}} \frac{1}{\sqrt{n}}\|\boldsymbol{u}\|_{\ell_2} < K_{\boldsymbol{u}}$ by definition of $\mathcal{S}_{\boldsymbol{u}}$.

Plugging the latter into AO becomes

(B.45)

$$\min_{\boldsymbol{\theta}\in\mathcal{S}_{\boldsymbol{\theta}},\boldsymbol{v}} \max_{0\le\beta\le K} \frac{\beta}{\sqrt{n}}\left\|\left\|\mathsf{P}_{\boldsymbol{\mu}}^{\perp}\boldsymbol{\Sigma}^{1/2}\boldsymbol{\theta}\right\|_{\ell_2}\boldsymbol{g} + \widetilde{\boldsymbol{\mu}}^T\boldsymbol{\Sigma}^{1/2}\boldsymbol{\theta}\boldsymbol{z} + \mathbf{1}\boldsymbol{\mu}^T\boldsymbol{\theta} - \boldsymbol{v}\right\|_{\ell_2} + \frac{\beta}{\sqrt{n}}\boldsymbol{h}^T\mathsf{P}_{\boldsymbol{\mu}}^{\perp}\boldsymbol{\Sigma}^{1/2}\boldsymbol{\theta} + \frac{1}{n}\sum_{i=1}^{n}\ell\left(v_i - \varepsilon\|\boldsymbol{\theta}\|_{\ell_q}\right).$$

We hereafter use the shorthand

$$\ell(\boldsymbol{v},\boldsymbol{\theta}) = \frac{1}{n}\sum_{i=1}^{n}\ell\left(v_i - \varepsilon\|\boldsymbol{\theta}\|_{\ell_q}\right),$$

for simplicity of notation. For the minimization, with respect to $\boldsymbol{\theta}$ and then $\boldsymbol{v}$, to become easier in our later calculation we proceed by writing $\ell(\boldsymbol{v},\boldsymbol{\theta})$ in terms of its conjugate with respect to $\boldsymbol{\theta}$. That is,

$$\ell(\boldsymbol{v},\boldsymbol{\theta}) = \sup_{\boldsymbol{w}} \boldsymbol{w}^T\boldsymbol{\theta} - \widetilde{\ell}(\boldsymbol{v},\boldsymbol{w})$$

where $\widetilde{\ell}(\boldsymbol{v},\boldsymbol{w})$ is the conjugate of $\ell$ with respect to $\boldsymbol{\theta}$. The logic behind this is that AO will then simplify to

$$\min_{\boldsymbol{\theta}\in\mathcal{S}_{\boldsymbol{\theta}},\boldsymbol{v}} \max_{0\le\beta\le K,\boldsymbol{w}} \frac{\beta}{\sqrt{n}}\left\|\left\|\mathsf{P}_{\boldsymbol{\mu}}^{\perp}\boldsymbol{\Sigma}^{1/2}\boldsymbol{\theta}\right\|_{\ell_2}\boldsymbol{g} + \widetilde{\boldsymbol{\mu}}^T\boldsymbol{\Sigma}^{1/2}\boldsymbol{\theta}\boldsymbol{z} + \mathbf{1}\boldsymbol{\mu}^T\boldsymbol{\theta} - \boldsymbol{v}\right\|_{\ell_2} + \frac{\beta}{\sqrt{n}}\boldsymbol{h}^T\mathsf{P}_{\boldsymbol{\mu}}^{\perp}\boldsymbol{\Sigma}^{1/2}\boldsymbol{\theta} + \boldsymbol{w}^T\boldsymbol{\theta} - \widetilde{\ell}(\boldsymbol{v},\boldsymbol{w})$$

which after flipping (allowed based on the correct form of convexity/concavity of PO) becomes

(B.46)

$$\max_{0\le\beta\le K,\boldsymbol{w}} \min_{\boldsymbol{\theta}\in\mathcal{S}_{\boldsymbol{\theta}},\boldsymbol{v}} \frac{\beta}{\sqrt{n}}\left\|\left\|\mathsf{P}_{\boldsymbol{\mu}}^{\perp}\boldsymbol{\Sigma}^{1/2}\boldsymbol{\theta}\right\|_{\ell_2}\boldsymbol{g} + \widetilde{\boldsymbol{\mu}}^T\boldsymbol{\Sigma}^{1/2}\boldsymbol{\theta}\boldsymbol{z} + \mathbf{1}\boldsymbol{\mu}^T\boldsymbol{\theta} - \boldsymbol{v}\right\|_{\ell_2} + \frac{\beta}{\sqrt{n}}\boldsymbol{h}^T\mathsf{P}_{\boldsymbol{\mu}}^{\perp}\boldsymbol{\Sigma}^{1/2}\boldsymbol{\theta} + \boldsymbol{w}^T\boldsymbol{\theta} - \widetilde{\ell}(\boldsymbol{v},\boldsymbol{w}).$$

We define the parallel and perpendicular components of $\boldsymbol{\theta}$ along vector $\boldsymbol{\mu}$ as follows:

$$\boldsymbol{\theta}_{\perp} = \mathsf{P}_{\boldsymbol{\mu}}^{\perp}\boldsymbol{\theta}, \quad \theta := \widetilde{\boldsymbol{\mu}}^T\boldsymbol{\theta}, \quad \mathsf{P}_{\boldsymbol{\mu}}\boldsymbol{\theta} = \theta\widetilde{\boldsymbol{\mu}}\,. \tag{B.47}$$

Given that $\widetilde{\boldsymbol{\mu}}$ is an eigenvector of $\boldsymbol{\Sigma}$, cf. Assumption 5, we have $\mathsf{P}_{\boldsymbol{\mu}}\boldsymbol{\Sigma}^{1/2}\mathsf{P}_{\boldsymbol{\mu}}^{\perp} = 0$ and therefore

$$\mathsf{P}_{\boldsymbol{\mu}}^{\perp}\boldsymbol{\Sigma}^{1/2}\boldsymbol{\theta} = \mathsf{P}_{\boldsymbol{\mu}}^{\perp}\boldsymbol{\Sigma}^{1/2}\left(\mathsf{P}_{\boldsymbol{\mu}} + \mathsf{P}_{\boldsymbol{\mu}}^{\perp}\right)\boldsymbol{\theta} = \mathsf{P}_{\boldsymbol{\mu}}^{\perp}\boldsymbol{\Sigma}^{1/2}\mathsf{P}_{\boldsymbol{\mu}}^{\perp}\boldsymbol{\theta} = \mathsf{P}_{\boldsymbol{\mu}}^{\perp}\boldsymbol{\Sigma}^{1/2}\mathsf{P}_{\boldsymbol{\mu}}^{\perp}\boldsymbol{\theta}_{\perp}\,.$$

Similarly, since $\boldsymbol{\Sigma}^{1/2}\widetilde{\boldsymbol{\mu}} = a\widetilde{\boldsymbol{\mu}}$.

$$\widetilde{\boldsymbol{\mu}}^{T}\boldsymbol{\Sigma}^{1/2}\boldsymbol{\theta} = a\widetilde{\boldsymbol{\mu}}^{T}\boldsymbol{\theta} = a\theta.$$

Rewriting the AO problem, we get

$$\begin{aligned}\max_{0\le\beta\le K,\boldsymbol{w}}\min_{\boldsymbol{\theta}\in\mathcal{S}_{\boldsymbol{\theta}},\boldsymbol{v}}\ &\frac{\beta}{\sqrt{n}}\left\|\left\|\mathsf{P}_{\boldsymbol{\mu}}^{\perp}\boldsymbol{\Sigma}^{1/2}\mathsf{P}_{\boldsymbol{\mu}}^{\perp}\boldsymbol{\theta}_{\perp}\right\|_{\ell_2}\boldsymbol{g} + a\theta\boldsymbol{z} + \mathbf{1}\left\|\boldsymbol{\mu}\right\|_{\ell_2}\theta - \boldsymbol{v}\right\|_{\ell_2}\\ &+\frac{\beta}{\sqrt{n}}\boldsymbol{h}^{T}\mathsf{P}_{\boldsymbol{\mu}}^{\perp}\boldsymbol{\Sigma}^{1/2}\mathsf{P}_{\boldsymbol{\mu}}^{\perp}\boldsymbol{\theta}_{\perp} + \boldsymbol{w}^{T}\mathsf{P}_{\boldsymbol{\mu}}^{\perp}\boldsymbol{\theta}_{\perp} + \boldsymbol{w}^{T}\widetilde{\boldsymbol{\mu}}\theta - \widetilde{\ell}(\boldsymbol{v},\boldsymbol{w}).\end{aligned}\tag{B.48}$$

We can rewrite this as

$$\begin{aligned}\max_{0\le\beta\le K,\boldsymbol{w}}\min_{\boldsymbol{\theta}\in\mathcal{S}_{\boldsymbol{\theta}},\boldsymbol{v}}\ &\frac{\beta}{\sqrt{n}}\left\|\left\|\mathsf{P}_{\boldsymbol{\mu}}^{\perp}\boldsymbol{\Sigma}^{1/2}\mathsf{P}_{\boldsymbol{\mu}}^{\perp}\boldsymbol{\theta}_{\perp}\right\|_{\ell_2}\boldsymbol{g} + a\theta\boldsymbol{z} + \mathbf{1}\left\|\boldsymbol{\mu}\right\|_{\ell_2}\theta - \boldsymbol{v}\right\|_{\ell_2}\\ &+\frac{\beta}{\sqrt{n}}\left(\mathsf{P}_{\boldsymbol{\mu}}^{\perp}\boldsymbol{h}\right)^{T}\mathsf{P}_{\boldsymbol{\mu}}^{\perp}\boldsymbol{\Sigma}^{1/2}\mathsf{P}_{\boldsymbol{\mu}}^{\perp}\boldsymbol{\theta}_{\perp} + \boldsymbol{w}^{T}\mathsf{P}_{\boldsymbol{\mu}}^{\perp}\boldsymbol{\Sigma}^{-1/2}\mathsf{P}_{\boldsymbol{\mu}}^{\perp}\left(\mathsf{P}_{\boldsymbol{\mu}}^{\perp}\boldsymbol{\Sigma}^{1/2}\mathsf{P}_{\boldsymbol{\mu}}^{\perp}\right)\boldsymbol{\theta}_{\perp} + \boldsymbol{w}^{T}\widetilde{\boldsymbol{\mu}}\theta - \widetilde{\ell}(\boldsymbol{v},\boldsymbol{w}).\end{aligned}\tag{B.49}$$

Here we used the assumption that $\widetilde{\boldsymbol{\mu}}$ is an eigenvector of $\boldsymbol{\Sigma}$ which in turn implies that

$$\mathsf{P}_{\boldsymbol{\mu}}^{\perp}\boldsymbol{\Sigma}^{-1/2}\mathsf{P}_{\boldsymbol{\mu}}^{\perp}\left(\mathsf{P}_{\boldsymbol{\mu}}^{\perp}\boldsymbol{\Sigma}^{1/2}\mathsf{P}_{\boldsymbol{\mu}}^{\perp}\right) = \mathsf{P}_{\boldsymbol{\mu}}^{\perp}\boldsymbol{\Sigma}^{-1/2}\mathsf{P}_{\boldsymbol{\mu}}^{\perp}\boldsymbol{\Sigma}^{1/2}\mathsf{P}_{\boldsymbol{\mu}}^{\perp} = \mathsf{P}_{\boldsymbol{\mu}}^{\perp}\boldsymbol{\Sigma}^{-1/2}\left(\mathsf{P}_{\boldsymbol{\mu}} + \mathsf{P}_{\boldsymbol{\mu}}^{\perp}\right)\boldsymbol{\Sigma}^{1/2}\mathsf{P}_{\boldsymbol{\mu}}^{\perp} = \mathsf{P}_{\boldsymbol{\mu}}^{\perp}.\tag{B.50}$$

We next optimize over $\boldsymbol{\theta}$ using lemma below and its proof is deferred to Appendix E.4.

**Lemma B.1** *For a given vector $\boldsymbol{r}$ and $\alpha \ge 0$ consider the following optimization*

$$\textit{min}_{\boldsymbol{\theta}\in\mathbb{R}^p}\quad \langle\mathsf{P}_{\boldsymbol{\mu}}^{\perp}\boldsymbol{r}, \left(\mathsf{P}_{\boldsymbol{\mu}}^{\perp}\boldsymbol{\Sigma}^{1/2}\mathsf{P}_{\boldsymbol{\mu}}^{\perp}\right)\boldsymbol{\theta}_{\perp}\rangle\tag{B.51}$$

$$\textit{subject to}\ \left\|\left(\mathsf{P}_{\boldsymbol{\mu}}^{\perp}\boldsymbol{\Sigma}^{1/2}\mathsf{P}_{\boldsymbol{\mu}}^{\perp}\right)\boldsymbol{\theta}_{\perp}\right\|_{\ell_2} = \alpha\tag{B.52}$$

*Under the assumption that $\mathsf{P}_{\boldsymbol{\mu}}\boldsymbol{\Sigma}^{1/2}\mathsf{P}_{\boldsymbol{\mu}}^{\perp} = 0$, the optimal value of this optimization is given by $-\alpha\left\|\mathsf{P}_{\boldsymbol{\mu}}^{\perp}\boldsymbol{r}\right\|_{\ell_2}$.*

Now note that

$$|\theta| = |\widetilde{\boldsymbol{\mu}}^T \boldsymbol{\theta}| \le \|\boldsymbol{\theta}\|_{\ell_2} ,$$
$$\alpha = \left\| \left( \mathsf{P}_{\boldsymbol{\mu}}^{\perp} \boldsymbol{\Sigma}^{1/2} \mathsf{P}_{\boldsymbol{\mu}}^{\perp} \right) \boldsymbol{\theta}_{\perp} \right\|_{\ell_2} = \left\| \mathsf{P}_{\boldsymbol{\mu}}^{\perp} \boldsymbol{\Sigma}^{1/2} \boldsymbol{\theta} \right\|_{\ell_2} \le \left\| \boldsymbol{\Sigma}^{1/2} \boldsymbol{\theta} \right\|_{\ell_2} \le C_{\max}^{1/2} \|\boldsymbol{\theta}\|_{\ell_2}$$

where in the second line we used Assumption 1$(b)$,$(d)$. Since $\boldsymbol{\theta} \in \mathcal{S}_{\boldsymbol{\theta}}$ a bounded set, we can choose $K' > 0$ large enough so that $0 \le |\theta|, \alpha \le K'$ and hence so do the optimization over this bounded range. That said, we use Lemma B.1 with $\boldsymbol{r} = \frac{\beta}{\sqrt{n}} \boldsymbol{h} + \boldsymbol{\Sigma}^{-1/2} \mathsf{P}_{\boldsymbol{\mu}}^{\perp} \boldsymbol{w}$, to simplify the AO problem as follows:

$$\max_{0 \le \beta \le K, \boldsymbol{w}} \min_{0 \le \alpha, |\theta| \le K', \boldsymbol{v}} \frac{\beta}{\sqrt{n}} \left\| \alpha \boldsymbol{g} + a\theta \boldsymbol{z} + \mathbf{1} \|\boldsymbol{\mu}\|_{\ell_2} \theta - \boldsymbol{v} \right\|_{\ell_2} - \alpha \left\| \frac{\beta}{\sqrt{n}} \mathsf{P}_{\boldsymbol{\mu}}^{\perp} \boldsymbol{h} + \mathsf{P}_{\boldsymbol{\mu}}^{\perp} \boldsymbol{\Sigma}^{-1/2} \mathsf{P}_{\boldsymbol{\mu}}^{\perp} \boldsymbol{w} \right\|_{\ell_2} + \boldsymbol{w}^T \widetilde{\boldsymbol{\mu}} \theta - \widetilde{\ell}(\boldsymbol{v}, \boldsymbol{w}) \tag{B.53}$$

To continue we shall calculate the conjugate function $\widetilde{\ell}$. This is the subject of the next lemma and we refer to Appendix E.5 for its proof.

**Lemma B.2** *The conjugate of the function*

$$\ell(\boldsymbol{v}, \boldsymbol{\theta}) = \frac{1}{n} \sum_{i=1}^{n} \ell \left( v_i - \varepsilon \|\boldsymbol{\theta}\|_{\ell_q} \right)$$

*with respect to $\boldsymbol{\theta}$ is equal to*

$$\widetilde{\ell}(\boldsymbol{v}, \boldsymbol{w}) = \sup_{\gamma \ge 0} \quad \gamma \|\boldsymbol{w}\|_{\ell_p} - \frac{1}{n} \sum_{i=1}^{n} \ell \left( v_i - \varepsilon \gamma \right) .$$

Using the above lemma we have

$$-\widetilde{\ell}(\boldsymbol{v}, \boldsymbol{w}) = -\left( \sup_{\gamma \ge 0} \quad \gamma \|\boldsymbol{w}\|_{\ell_p} - \frac{1}{n} \sum_{i=1}^{n} \ell \left( v_i - \varepsilon \gamma \right) \right) = \inf_{\gamma \ge 0} \quad -\gamma \|\boldsymbol{w}\|_{\ell_p} + \frac{1}{n} \sum_{i=1}^{n} \ell \left( v_i - \varepsilon \gamma \right) .$$

Plugging this into (B.53) we arrive at

$$\max_{0 \le \beta \le K, \boldsymbol{w}} \min_{0 \le \alpha, |\theta| \le K', \boldsymbol{v}, 0 \le \gamma} \frac{\beta}{\sqrt{n}} \left\| \alpha \boldsymbol{g} + a\theta \boldsymbol{z} + \mathbf{1} \|\boldsymbol{\mu}\|_{\ell_2} \theta - \boldsymbol{v} \right\|_{\ell_2} - \alpha \left\| \frac{\beta}{\sqrt{n}} \mathsf{P}_{\boldsymbol{\mu}}^{\perp} \boldsymbol{h} + \mathsf{P}_{\boldsymbol{\mu}}^{\perp} \boldsymbol{\Sigma}^{-1/2} \mathsf{P}_{\boldsymbol{\mu}}^{\perp} \boldsymbol{w} \right\|_{\ell_2} + \boldsymbol{w}^T \widetilde{\boldsymbol{\mu}} \theta - \gamma \|\boldsymbol{w}\|_{\ell_p} + \frac{1}{n} \sum_{i=1}^{n} \ell \left( v_i - \varepsilon \gamma \right) \tag{B.54}$$

Note that when $p \ge 1$ the objective is jointly concave in $(\boldsymbol{w}, \beta)$ and jointly convex in $\alpha, \theta, \boldsymbol{v}$ and therefore we can switch the orders of min and max.

We next focus on optimization over $\boldsymbol{v}$. Using the observation that for all $x \in \mathbb{R}$, $\min_{\tau \geq 0} \frac{\tau}{2} + \frac{x^2}{2\tau} = x$ we write

$$\begin{aligned}
&\min_{\boldsymbol{v}} \frac{\beta}{\sqrt{n}} \left\| \alpha \boldsymbol{g} + a\theta \boldsymbol{z} + \mathbf{1} \left\| \boldsymbol{\mu} \right\|_{\ell_2} \theta - \boldsymbol{v} \right\|_{\ell_2} + \frac{1}{n} \sum_{i=1}^{n} \ell \left( v_i - \varepsilon\gamma \right) \\
&= \min_{\boldsymbol{v}} \inf_{\tau_g \geq 0} \frac{\beta}{2\tau_g n} \left\| \alpha \boldsymbol{g} + a\theta \boldsymbol{z} + \mathbf{1} \left\| \boldsymbol{\mu} \right\|_{\ell_2} \theta - \boldsymbol{v} \right\|_{\ell_2}^2 + \frac{\beta\tau_g}{2} + \frac{1}{n} \sum_{i=1}^{n} \ell \left( v_i - \varepsilon\gamma \right) \\
&= \min_{\boldsymbol{v}} \inf_{\tau_g \geq 0} \frac{\beta}{2\tau_g n} \sum_{i=1}^{n} \left( \alpha g_i + a\theta z_i + \left\| \boldsymbol{\mu} \right\|_{\ell_2} \theta - v_i \right)^2 + \frac{\beta\tau_g}{2} + \frac{1}{n} \sum_{i=1}^{n} \ell \left( v_i - \varepsilon\gamma \right) \\
&= \min_{\widetilde{v}_i} \inf_{\tau_g \geq 0} \frac{\beta}{2\tau_g n} \sum_{i=1}^{n} \left( \alpha g_i + a\theta z_i + \left\| \boldsymbol{\mu} \right\|_{\ell_2} \theta - \widetilde{v}_i - \varepsilon\gamma \right)^2 + \frac{\beta\tau_g}{2} + \frac{1}{n} \sum_{i=1}^{n} \ell \left( \widetilde{v}_i \right)
\end{aligned} \tag{B.55}$$

As a result (B.54) can be rewritten as

$$\begin{aligned}
\max_{0 \leq \beta \leq K, \boldsymbol{w}} \min_{0 \leq \alpha, |\theta| \leq K', \widetilde{\boldsymbol{v}}, 0 \leq \gamma} \inf_{\tau_g \geq 0} &\frac{\beta}{2\tau_g n} \left\| \alpha \boldsymbol{g} + a\theta \boldsymbol{z} + \mathbf{1} \left\| \boldsymbol{\mu} \right\|_{\ell_2} \theta - \widetilde{\boldsymbol{v}} - \varepsilon\gamma \mathbf{1} \right\|_{\ell_2}^2 + \frac{\beta\tau_g}{2} \\
&- \alpha \left\| \frac{\beta}{\sqrt{n}} \mathsf{P}_{\boldsymbol{\mu}}^{\perp} \boldsymbol{h} + \mathsf{P}_{\boldsymbol{\mu}}^{\perp} \boldsymbol{\Sigma}^{-1/2} \mathsf{P}_{\boldsymbol{\mu}}^{\perp} \boldsymbol{w} \right\|_{\ell_2} + \boldsymbol{w}^T \widetilde{\boldsymbol{\mu}} \theta - \gamma \left\| \boldsymbol{w} \right\|_{\ell_p} + \frac{1}{n} \sum_{i=1}^{n} \ell \left( \widetilde{v}_i \right)
\end{aligned} \tag{B.56}$$

We note that since the quadratic over linear function is jointly convex the above loss is jointly convex in the parameters $(\alpha, \gamma, \theta, \tau_g, \widetilde{\boldsymbol{v}})$. Also for $p \geq 1$ the $\|\cdot\|_{\ell_p}$ is convex and thus the objective is also jointly concave in $(\beta, \boldsymbol{w})$.

We recall the definition of the Moreau envelope of function $\ell$ at a point $x$ with parameter $\mu$, that is given by

$$e_\ell(x; \mu) := \min_t \frac{1}{2\mu} (x - t)^2 + \ell(t) \,. \tag{B.57}$$

We can now rewrite equation (B.56) in terms of Moreau envelope of the loss function $\ell$.

$$\begin{aligned}
&\min_{\widetilde{\boldsymbol{v}}} \inf_{\tau_g \geq 0} \frac{\beta}{2\tau_g n} \left\| \alpha \boldsymbol{g} + a\theta \boldsymbol{z} + \mathbf{1} \left\| \boldsymbol{\mu} \right\|_{\ell_2} \theta - \widetilde{\boldsymbol{v}} - \varepsilon\gamma \mathbf{1} \right\|_{\ell_2}^2 + \frac{\beta\tau_g}{2} + \frac{1}{n} \sum_{i=1}^{n} \ell \left( \widetilde{v}_i \right) \\
&\qquad = \inf_{\tau_g \geq 0} \frac{\beta\tau_g}{2} + \frac{1}{n} \sum_{i=1}^{n} e_\ell \left( \alpha g_i + a\theta z_i + \left\| \boldsymbol{\mu} \right\|_{\ell_2} \theta - \varepsilon\gamma; \frac{\tau_g}{\beta} \right)
\end{aligned} \tag{B.58}$$

Thus (B.56) can be rewritten in the form

$$\begin{aligned}
\max_{0 \leq \beta \leq K, \boldsymbol{w}} \min_{0 \leq \alpha, |\theta| \leq K', 0 \leq \gamma} \inf_{\tau_g \geq 0} &\frac{\beta\tau_g}{2} + \frac{1}{n} \sum_{i=1}^{n} e_\ell \left( \alpha g_i + a\theta z_i + \left\| \boldsymbol{\mu} \right\|_{\ell_2} \theta - \varepsilon\gamma; \frac{\tau_g}{\beta} \right) \\
&- \alpha \left\| \frac{\beta}{\sqrt{n}} \mathsf{P}_{\boldsymbol{\mu}}^{\perp} \boldsymbol{h} + \mathsf{P}_{\boldsymbol{\mu}}^{\perp} \boldsymbol{\Sigma}^{-1/2} \mathsf{P}_{\boldsymbol{\mu}}^{\perp} \boldsymbol{w} \right\|_{\ell_2} + \boldsymbol{w}^T \widetilde{\boldsymbol{\mu}} \theta - \gamma \left\| \boldsymbol{w} \right\|_{\ell_p}
\end{aligned} \tag{B.59}$$

Note that since (B.56) is jointly convex in $(\alpha, \gamma, \theta, \tau_g, \widetilde{\boldsymbol{v}})$ and jointly concave in $(\beta, \boldsymbol{w})$ and partial minimization preserves convexity thus (B.59) is jointly convex in $(\alpha, \gamma, \theta, \tau_g)$ and jointly concave in $(\beta, \boldsymbol{w})$.

B.3.2. *Scalarization of the auxiliary optimization problem* The auxiliary problem (B.59) is in terms of high-dimensional vectors $\boldsymbol{g}, \boldsymbol{z}, \boldsymbol{h}, \boldsymbol{w}, \boldsymbol{\mu}$. We turn this problem into a scalar optimization by taking the pointwise limit of its objective and then showing that such convergence indeed holds in a uniform sense and therefore the minimax value also converges to that of the limit objective.

Note that by definition of the Moreau envelope, for all $x$ and $\mu$ we have

$$e_\ell(x;\mu) \le \frac{1}{2\mu}(x-x)^2 + \ell(x) = \ell(x) = \log(1+e^{-x}) \le \log 2 + |x|\,.$$

Hence,

$$\mathbb{E}\left[e_\ell\left(\alpha g + a\theta z + \|\boldsymbol{\mu}\|_{\ell_2}\theta - \varepsilon\gamma; \frac{\tau_g}{\beta}\right)\right] \le \log 2 + \mathbb{E}[|\alpha g + a\theta z + \|\boldsymbol{\mu}\|_{\ell_2}\theta - \varepsilon\gamma|] < \infty\,,$$

for any finite value of $\alpha$, $\theta$ and $\gamma$. Therefore by an application of the Weak Law of Large Numbers, we have that

$$\begin{aligned}\frac{1}{n}\sum_{i=1}^{n} e_\ell\left(\alpha g_i + a\theta z_i + \|\boldsymbol{\mu}\|_{\ell_2}\theta - \varepsilon\gamma; \frac{\tau_g}{\beta}\right) &\to \mathbb{E}\left[e_\ell\left(\alpha g + a\theta z + \|\boldsymbol{\mu}\|_{\ell_2}\theta - \varepsilon\gamma; \frac{\tau_g}{\beta}\right)\right] \\ &= \mathbb{E}\left[e_\ell\left(\sqrt{\alpha^2 + a^2\theta^2}\, g + \|\boldsymbol{\mu}\|_{\ell_2}\theta - \varepsilon\gamma; \frac{\tau_g}{\beta}\right)\right].\end{aligned}$$

We define the expected Moreau envelope $L(a,b,\mu) = \mathbb{E}[e_\ell(ag+b;\mu)]$, where the expectation is taken with respect to independent standard normal variable $g$.

This simplifies the AO problem as

$$\begin{aligned}\max_{0\le\beta\le K, \boldsymbol{w}} \min_{0\le\alpha,|\theta|\le K', 0\le\gamma,\tau_g} \; & \frac{\beta\tau_g}{2} + L\left(\sqrt{\alpha^2+a^2\theta^2}, \|\boldsymbol{\mu}\|_{\ell_2}\theta - \varepsilon\gamma, \frac{\tau_g}{\beta}\right) \\ & - \alpha\left\|\frac{\beta}{\sqrt{n}}\mathsf{P}_{\boldsymbol{\mu}}^{\perp}\boldsymbol{h} + \mathsf{P}_{\boldsymbol{\mu}}^{\perp}\boldsymbol{\Sigma}^{-1/2}\mathsf{P}_{\boldsymbol{\mu}}^{\perp}\boldsymbol{w}\right\|_{\ell_2} + \boldsymbol{w}^T\widetilde{\boldsymbol{\mu}}\theta - \gamma\|\boldsymbol{w}\|_{\ell_p}\,. \end{aligned} \tag{B.60}$$

We note that since (B.59) is jointly convex in $(\alpha, \gamma, \theta, \tau_g)$ and jointly concave in $(\beta, \boldsymbol{w})$ and expectation preserves convexity thus the objective in (B.60) is also jointly convex in $(\alpha, \gamma, \theta, \tau_g)$ and jointly concave in $(\beta, \boldsymbol{w})$ and

by Sinov's theorem we can flip the maximization over $\boldsymbol{w}$ and the minimization to arrive at

$$\max_{0\le\beta\le K} \min_{0\le\alpha,|\theta|\le K',0<\gamma,\tau_g} \frac{\beta\tau_g}{2} + L\left(\sqrt{\alpha^2+a^2\theta^2}, \|\boldsymbol{\mu}\|_{\ell_2}\theta - \varepsilon\gamma, \frac{\tau_g}{\beta}\right) - \min_{\boldsymbol{w}} \left\{\alpha \left\| \frac{\beta}{\sqrt{n}}\mathsf{P}_{\boldsymbol{\mu}}^{\perp}\boldsymbol{h} + \mathsf{P}_{\boldsymbol{\mu}}^{\perp}\boldsymbol{\Sigma}^{-1/2}\mathsf{P}_{\boldsymbol{\mu}}^{\perp}\boldsymbol{w}\right\|_{\ell_2} - \boldsymbol{w}^T\widetilde{\boldsymbol{\mu}}\theta + \gamma\|\boldsymbol{w}\|_{\ell_p}\right\}. \tag{B.61}$$

By our asymptotic setting (cf. Definition 1, part (c)), $\varepsilon = \varepsilon_0\|\boldsymbol{\mu}\|_{\ell_p}$ for a constant $\varepsilon_0$. We let $\gamma_0 := \gamma\|\boldsymbol{\mu}\|_{\ell_p}$ and rewriting (B.61) in terms of $\gamma_0$ in lieu of $\gamma$ we arrive at

$$\max_{0\le\beta\le K} \min_{0\le\alpha,|\theta|\le K',0<\gamma_0<K'',0<\tau_g} \frac{\beta\tau_g}{2} + L\left(\sqrt{\alpha^2+a^2\theta^2}, \|\boldsymbol{\mu}\|_{\ell_2}\theta - \varepsilon_0\gamma_0, \frac{\tau_g}{\beta}\right) - \min_{\boldsymbol{w}} \left\{\alpha \left\| \frac{\beta}{\sqrt{n}}\mathsf{P}_{\boldsymbol{\mu}}^{\perp}\boldsymbol{h} + \mathsf{P}_{\boldsymbol{\mu}}^{\perp}\boldsymbol{\Sigma}^{-1/2}\mathsf{P}_{\boldsymbol{\mu}}^{\perp}\boldsymbol{w}\right\|_{\ell_2} - \boldsymbol{w}^T\widetilde{\boldsymbol{\mu}}\theta + \frac{\gamma_0}{\|\boldsymbol{\mu}\|_{\ell_p}}\|\boldsymbol{w}\|_{\ell_p}\right\}. \tag{B.62}$$

- **Optimization over $\boldsymbol{w}$.** Continuing with optimization over $\boldsymbol{w}$ we have

$$\begin{aligned}
&\min_{\boldsymbol{w}} \quad \alpha \left\| \frac{\beta}{\sqrt{n}}\mathsf{P}_{\boldsymbol{\mu}}^{\perp}\boldsymbol{h} + \mathsf{P}_{\boldsymbol{\mu}}^{\perp}\boldsymbol{\Sigma}^{-1/2}\mathsf{P}_{\boldsymbol{\mu}}^{\perp}\boldsymbol{w}\right\|_{\ell_2} - \boldsymbol{w}^T\widetilde{\boldsymbol{\mu}}\theta + \frac{\gamma_0}{\|\boldsymbol{\mu}\|_{\ell_p}}\|\boldsymbol{w}\|_{\ell_p} \\
&= \min_{\boldsymbol{w},\tau_h\ge 0} \quad \frac{\alpha}{2\tau_h}\left\| \frac{\beta}{\sqrt{n}}\mathsf{P}_{\boldsymbol{\mu}}^{\perp}\boldsymbol{h} + \mathsf{P}_{\boldsymbol{\mu}}^{\perp}\boldsymbol{\Sigma}^{-1/2}\mathsf{P}_{\boldsymbol{\mu}}^{\perp}\boldsymbol{w}\right\|_{\ell_2}^2 + \frac{\alpha\tau_h}{2} - \boldsymbol{w}^T\widetilde{\boldsymbol{\mu}}\theta + \frac{\gamma_0}{\|\boldsymbol{\mu}\|_{\ell_p}}\|\boldsymbol{w}\|_{\ell_p} \\
&= \min_{\boldsymbol{w},\tau_h\ge 0} \quad \frac{\alpha}{2\tau_h}\left\| \frac{\beta}{\sqrt{n}}\mathsf{P}_{\boldsymbol{\mu}}^{\perp}\boldsymbol{h} + \mathsf{P}_{\boldsymbol{\mu}}^{\perp}\boldsymbol{\Sigma}^{-1/2}\boldsymbol{w}\right\|_{\ell_2}^2 + \frac{\alpha\tau_h}{2} - \boldsymbol{w}^T\widetilde{\boldsymbol{\mu}}\theta + \frac{\gamma_0}{\|\boldsymbol{\mu}\|_{\ell_p}}\|\boldsymbol{w}\|_{\ell_p},
\end{aligned} \tag{B.63}$$

  where in the last step we used that $\mathsf{P}_{\boldsymbol{\mu}}^{\perp}\boldsymbol{\Sigma}^{-1/2}\mathsf{P}_{\boldsymbol{\mu}} = 0$, which follows from Assumption 5. Note that the above loss is jointly convex in $(\boldsymbol{w},\tau_h)$. So that continuing from (B.61) the overall objective is jointly convex in $(\alpha,\gamma,\theta,\tau_g)$ and jointly concave in $(\beta,\boldsymbol{w},\tau_h)$.
  Let $\tilde{\boldsymbol{w}} := \boldsymbol{\Sigma}^{-1/2}\boldsymbol{w}$. The optimization over $\boldsymbol{w}$ can be written as

$$\min_{\tilde{\boldsymbol{w}}} \quad \frac{1}{2}\left\| \frac{\beta}{\sqrt{n}}\mathsf{P}_{\boldsymbol{\mu}}^{\perp}\boldsymbol{h} + \mathsf{P}_{\boldsymbol{\mu}}^{\perp}\tilde{\boldsymbol{w}}\right\|_{\ell_2}^2 + f(\tilde{\boldsymbol{w}}), \tag{B.64}$$

  where

$$f(\tilde{\boldsymbol{w}}) := -\langle\tilde{\boldsymbol{w}}, \boldsymbol{\Sigma}^{1/2}\widetilde{\boldsymbol{\mu}}\rangle\frac{\theta\tau_h}{\alpha} + \frac{\tau_h}{\alpha}\frac{\gamma_0}{\|\boldsymbol{\mu}\|_{\ell_p}}\left\|\boldsymbol{\Sigma}^{1/2}\tilde{\boldsymbol{w}}\right\|_{\ell_p}.$$

Let $\tilde{\boldsymbol{w}}^*$ be the optimal solution. Then,

$$-\mathsf{P}_{\boldsymbol{\mu}}^{\perp}\left(\frac{\beta}{\sqrt{n}}\boldsymbol{h}+\tilde{\boldsymbol{w}}^*\right)\in\partial f(\tilde{\boldsymbol{w}}^*)\,. \tag{B.65}$$

By the conjugate subgradient theorem, this implies that

$$\tilde{\boldsymbol{w}}^*\in\partial f^*\left(-\mathsf{P}_{\boldsymbol{\mu}}^{\perp}\left(\frac{\beta}{\sqrt{n}}\boldsymbol{h}+\tilde{\boldsymbol{w}}^*\right)\right).$$

Let $\boldsymbol{t}^*:=\frac{\beta}{\sqrt{n}}\boldsymbol{h}+\tilde{\boldsymbol{w}}^*$, then writing the above equation in terms of $t$,

$$\boldsymbol{t}^*-\frac{\beta}{\sqrt{n}}\boldsymbol{h}\in\partial f^*\left(-\mathsf{P}_{\boldsymbol{\mu}}^{\perp}\boldsymbol{t}^*\right). \tag{B.66}$$

Therefore,

$$-\mathsf{P}_{\boldsymbol{\mu}}^{\perp}\left(\boldsymbol{t}^*-\frac{\beta}{\sqrt{n}}\boldsymbol{h}\right)\in-\mathsf{P}_{\boldsymbol{\mu}}^{\perp}\partial f^*\left(-\mathsf{P}_{\boldsymbol{\mu}}^{\perp}\boldsymbol{t}^*\right). \tag{B.67}$$

Equation (B.67) is equivalent to saying that

$$\boldsymbol{t}^*\in\arg\min_{\boldsymbol{t}}\frac{1}{2}\left\|\mathsf{P}_{\boldsymbol{\mu}}^{\perp}\left(\frac{\beta}{\sqrt{n}}\boldsymbol{h}-\boldsymbol{t}\right)\right\|_{\ell_2}^2+f^*\left(-\mathsf{P}_{\boldsymbol{\mu}}^{\perp}\boldsymbol{t}\right). \tag{B.68}$$

**Lemma B.3** *For function $f:\mathbb{R}^p\mapsto\mathbb{R}$ given by*

$$f(\tilde{\boldsymbol{w}}):=-\langle\tilde{\boldsymbol{w}},\boldsymbol{\Sigma}^{1/2}\widetilde{\boldsymbol{\mu}}\rangle\frac{\theta\tau_h}{\alpha}+\frac{\tau_h}{\alpha}\frac{\gamma_0}{\|\boldsymbol{\mu}\|_{\ell_p}}\left\|\boldsymbol{\Sigma}^{1/2}\tilde{\boldsymbol{w}}\right\|_{\ell_p},$$

*its convex conjugate reads as*

$$f^*(\boldsymbol{u})=\mathbb{1}_S(\boldsymbol{u}),\quad S:=\left\{\boldsymbol{u}:\quad\left\|\boldsymbol{\Sigma}^{-1/2}\boldsymbol{u}+\frac{\tau_h\theta}{\alpha}\widetilde{\boldsymbol{\mu}}\right\|_{\ell_q}\le\frac{\gamma\tau_h}{\alpha}\right\},\quad\mathbb{1}_S(\boldsymbol{u})=\begin{cases}0 & \text{if } \boldsymbol{u}\in S\\ \infty & \text{if } \boldsymbol{u}\notin S\end{cases}$$

The proof of Lemma B.3 is delegated to Appendix E.6.
Define $\mathcal{B}:=\{\boldsymbol{\mu}\}^{\perp}\cap-\mathcal{S}$. Then (B.68) implies that

$$\mathsf{P}_{\boldsymbol{\mu}}^{\perp}\boldsymbol{t}^*=\mathsf{P}_{\mathcal{B}}\left(\mathsf{P}_{\boldsymbol{\mu}}^{\perp}\left(\frac{\beta}{\sqrt{n}}\boldsymbol{h}\right)\right). \tag{B.69}$$

**Lemma B.4** *For a convex set $\mathcal{S}$ and $\mathcal{B}:=\{\boldsymbol{\mu}\}^{\perp}\cap\mathcal{S}$, we have $\mathsf{P}_{\mathcal{B}}\mathsf{P}_{\boldsymbol{\mu}}^{\perp}=\mathsf{P}_{\mathcal{B}}$.*

We refer to Appendix E.7 for the proof of Lemma B.4.
Using Lemma B.4 and (B.69) we obtain

$$\mathsf{P}_{\boldsymbol{\mu}}^{\perp}\boldsymbol{t}^{*} = \mathsf{P}_{\mathcal{B}}\left(\frac{\beta}{\sqrt{n}}\boldsymbol{h}\right).$$

Recalling definition of $\boldsymbol{t}^{*}$ this implies

$$\mathsf{P}_{\boldsymbol{\mu}}^{\perp}\tilde{\boldsymbol{w}}^{*} = \mathsf{P}_{\mathcal{B}}\left(\frac{\beta}{\sqrt{n}}\boldsymbol{h}\right) - \frac{\beta}{\sqrt{n}}\mathsf{P}_{\boldsymbol{\mu}}^{\perp}\boldsymbol{h}\,. \tag{B.70}$$

Now note that for $p>1$, $\nabla\|\boldsymbol{w}\|_{\ell_p} = \frac{1}{\|\boldsymbol{w}\|_{\ell_p}^{p-1}}[|w_1|^{p-1}\mathrm{sign}(w_1),\dots,|w_p|^{p-1}\mathrm{sign}(w_p)]^{\mathsf{T}}$.
Therefore in this case $\langle\nabla\|\boldsymbol{w}\|_{\ell_p},\boldsymbol{w}\rangle = \|\boldsymbol{w}\|_{\ell_p}$. Similarly, for $p=1$ for any $\boldsymbol{s}\in\partial\|\boldsymbol{w}\|_{\ell_p}$ we have $\langle\boldsymbol{s},\boldsymbol{w}\rangle = \|\boldsymbol{w}\|_{\ell_p}$. Therefore, for all $p\geq 1$ for any $\boldsymbol{s}\in\partial\|\boldsymbol{w}\|_{\ell_p}$ we have $\langle\boldsymbol{s},\boldsymbol{w}\rangle = \|\boldsymbol{w}\|_{\ell_p}$.
Therefore, for the defined function $f$ and any $\boldsymbol{s}\in\partial f(\tilde{\boldsymbol{w}})$ there is a vector $\widetilde{\boldsymbol{s}}\in\partial\|\boldsymbol{x}\|_{\ell_p}\big|_{\boldsymbol{x}=\boldsymbol{\Sigma}^{1/2}\tilde{\boldsymbol{w}}}$ such that

$$\begin{aligned}
\langle\boldsymbol{s},\tilde{\boldsymbol{w}}\rangle &= \left\langle -\boldsymbol{\Sigma}^{1/2}\widetilde{\boldsymbol{\mu}}\frac{\theta\tau_h}{\alpha} + \frac{\tau_h}{\alpha}\frac{\gamma_0}{\|\boldsymbol{\mu}\|_{\ell_p}}\boldsymbol{\Sigma}^{1/2}\widetilde{\boldsymbol{s}},\tilde{\boldsymbol{w}}\right\rangle\\
&= -\langle\tilde{\boldsymbol{w}},\boldsymbol{\Sigma}^{1/2}\widetilde{\boldsymbol{\mu}}\rangle\frac{\theta\tau_h}{\alpha} + \frac{\tau_h}{\alpha}\frac{\gamma_0}{\|\boldsymbol{\mu}\|_{\ell_p}}\left\langle\widetilde{\boldsymbol{s}},\boldsymbol{\Sigma}^{1/2}\tilde{\boldsymbol{w}}\right\rangle\\
&= -\langle\tilde{\boldsymbol{w}},\boldsymbol{\Sigma}^{1/2}\widetilde{\boldsymbol{\mu}}\rangle\frac{\theta\tau_h}{\alpha} + \frac{\tau_h}{\alpha}\frac{\gamma_0}{\|\boldsymbol{\mu}\|_{\ell_p}}\left\|\boldsymbol{\Sigma}^{1/2}\tilde{\boldsymbol{w}}\right\|_{\ell_p}\\
&= f(\tilde{\boldsymbol{w}})\,.
\end{aligned} \tag{B.71}$$

Therefore by invoking (B.65)

$$\begin{aligned}
f(\tilde{\boldsymbol{w}}^{*}) &= \left\langle -\mathsf{P}_{\boldsymbol{\mu}}^{\perp}\left(\frac{\beta}{\sqrt{n}}\boldsymbol{h}+\tilde{\boldsymbol{w}}^{*}\right),\tilde{\boldsymbol{w}}^{*}\right\rangle\\
&= \left\langle -\frac{\beta}{\sqrt{n}}\mathsf{P}_{\boldsymbol{\mu}}^{\perp}\boldsymbol{h} - \mathsf{P}_{\boldsymbol{\mu}}^{\perp}\tilde{\boldsymbol{w}}^{*},\mathsf{P}_{\boldsymbol{\mu}}^{\perp}\tilde{\boldsymbol{w}}^{*}\right\rangle\\
&= \left\langle -\mathsf{P}_{\mathcal{B}}\left(\frac{\beta}{\sqrt{n}}\boldsymbol{h}\right),\mathsf{P}_{\mathcal{B}}\left(\frac{\beta}{\sqrt{n}}\boldsymbol{h}\right) - \frac{\beta}{\sqrt{n}}\mathsf{P}_{\boldsymbol{\mu}}^{\perp}\boldsymbol{h}\right\rangle\\
&= -\left\|\mathsf{P}_{\mathcal{B}}\left(\frac{\beta}{\sqrt{n}}\boldsymbol{h}\right)\right\|_{\ell_2}^{2} + \frac{\beta}{\sqrt{n}}\left\langle\mathsf{P}_{\mathcal{B}}\left(\frac{\beta}{\sqrt{n}}\boldsymbol{h}\right),\mathsf{P}_{\boldsymbol{\mu}}^{\perp}\boldsymbol{h}\right\rangle.
\end{aligned} \tag{B.72}$$

Putting things together, the optimal value of objective (B.64) over $\boldsymbol{w}$ is given by

$$\begin{aligned}
&\min_{\tilde{\boldsymbol{w}}} \quad \frac{1}{2}\left\|\frac{\beta}{\sqrt{n}}\mathsf{P}_{\boldsymbol{\mu}}^{\perp}\boldsymbol{h}+\mathsf{P}_{\boldsymbol{\mu}}^{\perp}\tilde{\boldsymbol{w}}\right\|_{\ell_2}^2+f(\tilde{\boldsymbol{w}})\\
&=\frac{1}{2}\left\|\frac{\beta}{\sqrt{n}}\mathsf{P}_{\boldsymbol{\mu}}^{\perp}\boldsymbol{h}+\mathsf{P}_{\boldsymbol{\mu}}^{\perp}\tilde{\boldsymbol{w}}^*\right\|_{\ell_2}^2+f(\tilde{\boldsymbol{w}}^*)\\
&=\frac{1}{2}\left\|\mathsf{P}_{\mathcal{B}}\left(\frac{\beta}{\sqrt{n}}\boldsymbol{h}\right)\right\|_{\ell_2}^2-\left\|\mathsf{P}_{\mathcal{B}}\left(\frac{\beta}{\sqrt{n}}\boldsymbol{h}\right)\right\|_{\ell_2}^2+\frac{\beta}{\sqrt{n}}\left\langle\mathsf{P}_{\mathcal{B}}\left(\frac{\beta}{\sqrt{n}}\boldsymbol{h}\right),\mathsf{P}_{\boldsymbol{\mu}}^{\perp}\boldsymbol{h}\right\rangle\\
&=-\frac{1}{2}\left\|\mathsf{P}_{\mathcal{B}}\left(\frac{\beta}{\sqrt{n}}\boldsymbol{h}\right)\right\|_{\ell_2}^2+\frac{\beta}{\sqrt{n}}\left\langle\mathsf{P}_{\mathcal{B}}\left(\frac{\beta}{\sqrt{n}}\boldsymbol{h}\right),\mathsf{P}_{\boldsymbol{\mu}}^{\perp}\boldsymbol{h}\right\rangle\\
&=\frac{\beta^2}{2n}\left\|\mathsf{P}_{\boldsymbol{\mu}}^{\perp}\boldsymbol{h}\right\|_{\ell_2}^2-\frac{1}{2}\left\|\mathsf{P}_{\mathcal{B}}\left(\frac{\beta}{\sqrt{n}}\boldsymbol{h}\right)-\frac{\beta}{\sqrt{n}}\mathsf{P}_{\boldsymbol{\mu}}^{\perp}\boldsymbol{h}\right\|_{\ell_2}^2\\
&=\frac{\beta^2}{2n}\left\|\mathsf{P}_{\boldsymbol{\mu}}^{\perp}\boldsymbol{h}\right\|_{\ell_2}^2-\frac{1}{2}\left\|\mathsf{P}_{\mathcal{B}}\left(\frac{\beta}{\sqrt{n}}\boldsymbol{h}\right)-\frac{\beta}{\sqrt{n}}\boldsymbol{h}\right\|_{\ell_2}^2+\frac{\beta^2}{2n}\left\|\mathsf{P}_{\boldsymbol{\mu}}\boldsymbol{h}\right\|_{\ell_2}^2\\
&=\frac{\beta^2}{2n}\left\|\boldsymbol{h}\right\|_{\ell_2}^2-\frac{1}{2}\left\|\mathsf{P}_{\mathcal{B}}\left(\frac{\beta}{\sqrt{n}}\boldsymbol{h}\right)-\frac{\beta}{\sqrt{n}}\boldsymbol{h}\right\|_{\ell_2}^2. \qquad \text{(B.73)}
\end{aligned}$$

Following the argument after (B.63) since the objective is jointly convex in $(\alpha,\gamma,\theta,\tau_g)$ and jointly concave in $(\beta,\boldsymbol{w},\tau_h)$ and partial maximization preserves concavity after plugging the above the objective is jointly convex in $(\alpha,\gamma,\theta,\tau_g)$ and jointly concave in $(\beta,\tau_h)$.

- **On projection $\mathsf{P}_{\mathcal{B}}$.** As part of our scalarization process of the auxiliary optimization problem, in the next lemma we provide an alternative characterization of the distance $\|\mathsf{P}_{\mathcal{B}}(\boldsymbol{h})-\boldsymbol{h}\|_{\ell_2}$, and refer to Appendix E.8 for its proof.

**Lemma B.5** *Recall the set $\mathcal{B}:=\{\boldsymbol{\mu}\}^{\perp}\cap-\mathcal{S}$, where $\mathcal{S}$ is given by*

$$S:=\left\{\boldsymbol{u}:\quad\left\|\boldsymbol{\Sigma}^{-1/2}\boldsymbol{u}+\frac{\tau_h\theta}{\alpha}\widetilde{\boldsymbol{\mu}}\right\|_{\ell_q}\le\frac{\tau_h}{\alpha}\frac{\gamma_0}{\|\boldsymbol{\mu}\|_{\ell_p}}\right\}.$$

*Also, suppose that $\boldsymbol{\Sigma}^{1/2}\widetilde{\boldsymbol{\mu}}=a\widetilde{\boldsymbol{\mu}}$. Then, for any vector $\boldsymbol{h}$ the following holds:*

$$\frac{1}{2}\left\|\mathsf{P}_{\mathcal{B}}\left(\boldsymbol{h}\right)-\boldsymbol{h}\right\|_{\ell_2}^2=\sup_{\lambda\ge0,\nu}\ e_{q,\boldsymbol{\Sigma}}\left(\boldsymbol{\Sigma}^{-1/2}\boldsymbol{h}-\left(\frac{\tau_h\theta}{\alpha}+\frac{\nu}{a}\right)\widetilde{\boldsymbol{\mu}};\lambda\right)-\lambda\left(\frac{\gamma_0}{\|\boldsymbol{\mu}\|_{\ell_p}}\frac{\tau_h}{\alpha}\right)^q+\nu\widetilde{\boldsymbol{\mu}}^T\boldsymbol{h}-\frac{\nu^2}{2} \qquad \text{(B.74)}$$

Using equation (B.73) along with Lemma B.5 we have

$$\min_{\tilde{\boldsymbol{w}}} \quad \frac{1}{2}\left\|\frac{\beta}{\sqrt{n}}\mathsf{P}_{\boldsymbol{\mu}}^{\perp}\boldsymbol{h}+\mathsf{P}_{\boldsymbol{\mu}}^{\perp}\tilde{\boldsymbol{w}}\right\|_{\ell_2}^2+f(\tilde{\boldsymbol{w}})$$

(B.75)

$$=\inf_{\lambda\geq 0,\nu}\ \frac{\beta^2}{2n}\left\|\boldsymbol{h}\right\|_{\ell_2}^2+\lambda\left(\frac{\tau_h}{\alpha}\frac{\gamma_0}{\left\|\boldsymbol{\mu}\right\|_{\ell_p}}\right)^q-\frac{\nu\beta}{\sqrt{n}}\widetilde{\boldsymbol{\mu}}^T\boldsymbol{h}+\frac{\nu^2}{2}-e_{q,\boldsymbol{\Sigma}}\left(\frac{\beta}{\sqrt{n}}\boldsymbol{\Sigma}^{-1/2}\boldsymbol{h}-\left(\frac{\tau_h\theta}{\alpha}+\frac{\nu}{a}\right)\widetilde{\boldsymbol{\mu}};\lambda\right)$$

Recalling equation (B.63) we have

$$\begin{aligned}
&\min_{\boldsymbol{w}}\quad \alpha\left\|\frac{\beta}{\sqrt{n}}\mathsf{P}_{\boldsymbol{\mu}}^{\perp}\boldsymbol{h}+\mathsf{P}_{\boldsymbol{\mu}}^{\perp}\boldsymbol{\Sigma}^{-1/2}\mathsf{P}_{\boldsymbol{\mu}}^{\perp}\boldsymbol{w}\right\|_{\ell_2}-\boldsymbol{w}^T\widetilde{\boldsymbol{\mu}}\theta+\frac{\gamma_0}{\left\|\boldsymbol{\mu}\right\|_{\ell_p}}\left\|\boldsymbol{w}\right\|_{\ell_p}\\
&=\min_{\tau_h,\lambda\geq 0,\nu}\frac{\alpha}{\tau_h}\left\{\frac{\beta^2}{2n}\left\|\boldsymbol{h}\right\|_{\ell_2}^2+\lambda\left(\frac{\tau_h}{\alpha}\frac{\gamma_0}{\left\|\boldsymbol{\mu}\right\|_{\ell_p}}\right)^q-\frac{\nu\beta}{\sqrt{n}}\widetilde{\boldsymbol{\mu}}^T\boldsymbol{h}+\frac{\nu^2}{2}-e_{q,\boldsymbol{\Sigma}}\left(\frac{\beta}{\sqrt{n}}\boldsymbol{\Sigma}^{-1/2}\boldsymbol{h}-\left(\frac{\tau_h\theta}{\alpha}+\frac{\nu}{a}\right)\widetilde{\boldsymbol{\mu}};\lambda\right)\right\}\\
&\qquad+\frac{\alpha\tau_h}{2}\\
&=\min_{\tau_h,\lambda\geq 0,\nu}\frac{\alpha}{\tau_h}\left\{\frac{\beta^2}{2n}\left\|\boldsymbol{h}\right\|_{\ell_2}^2+\lambda_0\left(\frac{\tau_h\gamma_0}{\alpha}\right)^q-\frac{\nu\beta}{\sqrt{n}}\widetilde{\boldsymbol{\mu}}^T\boldsymbol{h}+\frac{\nu^2}{2}-e_{q,\boldsymbol{\Sigma}}\left(\frac{\beta}{\sqrt{n}}\boldsymbol{\Sigma}^{-1/2}\boldsymbol{h}-\left(\frac{\tau_h\theta}{\alpha}+\frac{\nu}{a}\right)\widetilde{\boldsymbol{\mu}};\lambda_0\left\|\boldsymbol{\mu}\right\|_{\ell_p}^q\right)\right\}
\end{aligned}$$

(B.76)

$$+\frac{\alpha\tau_h}{2}$$

where we used the reparameterization $\lambda_0 := \frac{\lambda}{\left\|\boldsymbol{\mu}\right\|_{\ell_p}^q}$. Next we use Assumption 7 to take the limit of the above expression as $n\to\infty$. By definition of function $\mathsf{E}$ we have

$$\lim_{n\to\infty} e_{q,\boldsymbol{\Sigma}}\left(\frac{\beta}{\sqrt{n}}\boldsymbol{\Sigma}^{-1/2}\boldsymbol{h}-\left(\frac{\tau_h\theta}{\alpha}+\frac{\nu}{a}\right)\widetilde{\boldsymbol{\mu}};\lambda_0\left\|\boldsymbol{\mu}\right\|_{\ell_p}^q\right)=\mathsf{E}\left(\beta,\left(\frac{\tau_h\theta}{\alpha}+\frac{\nu}{a}\right);\lambda_0\right).$$

Also, since $\boldsymbol{h}\sim\mathsf{N}(0,\boldsymbol{I}_d)$ we have

$$\lim_{n\to\infty}\frac{1}{n}\left\|\boldsymbol{h}\right\|_{\ell_2}^2=\frac{1}{\delta},\qquad \lim_{n\to\infty}\frac{1}{\sqrt{n}}\widetilde{\boldsymbol{\mu}}^T\boldsymbol{h}=0.$$

Using the above two equations in (B.76) we have

$$\lim_{n\to\infty}\min_{\boldsymbol{w}}\left\{\alpha\left\|\frac{\beta}{\sqrt{n}}\mathsf{P}_{\boldsymbol{\mu}}^{\perp}\boldsymbol{h}+\mathsf{P}_{\boldsymbol{\mu}}^{\perp}\boldsymbol{\Sigma}^{-1/2}\mathsf{P}_{\boldsymbol{\mu}}^{\perp}\boldsymbol{w}\right\|_{\ell_2}-\boldsymbol{w}^T\widetilde{\boldsymbol{\mu}}\theta+\frac{\gamma_0}{\left\|\boldsymbol{\mu}\right\|_{\ell_p}}\left\|\boldsymbol{w}\right\|_{\ell_p}\right\}$$

(B.77)

$$=\frac{\alpha}{\tau_h}\left\{\frac{\beta^2}{2\delta}+\lambda_0\left(\frac{\gamma_0\tau_h}{\alpha}\right)^q+\frac{\nu^2}{2}-\mathsf{E}\left(\beta,\left(\frac{\tau_h\theta}{\alpha}+\frac{\nu}{a}\right);\lambda_0\right)\right\}+\frac{\alpha\tau_h}{2}.$$

Finally, incorporating the above equation in (B.62) and using Assumption 2, the AO problem simplifies to:

$$\max_{0\le\beta\le K}\ \min_{0\le\alpha,|\theta|\le K',0<\gamma_0<K'',0<\tau_g}\ \frac{\beta\tau_g}{2}+L\left(\sqrt{\alpha^2+a^2\theta^2},V\theta-\varepsilon_0\gamma_0,\frac{\tau_g}{\beta}\right)$$

(B.78)

$$-\min_{\tau_h,\lambda_0\ge0,\nu}\left[\frac{\alpha}{\tau_h}\left\{\frac{\beta^2}{2\delta}+\lambda_0\left(\frac{\gamma_0\tau_h}{\alpha}\right)^q+\frac{\nu^2}{2}-\mathsf{E}\left(\beta,\left(\frac{\tau_h\theta}{\alpha}+\frac{\nu}{a}\right);\lambda_0\right)\right\}+\frac{\alpha\tau_h}{2}\right]$$

Now recall the argument after (B.73) that the objective is jointly convex in $(\alpha,\gamma,\theta,\tau_g)$ and jointly concave in $(\beta,\tau_h)$. We used Lemma B.5 to provide alternative characterization for quantity $\|\mathsf{P}_{\mathcal{B}}(\boldsymbol{h})-\boldsymbol{h}\|_{\ell_2}^2$, which led into introducing the new variables $\lambda_0,\nu$. Therefore, the objective (B.78), after maximization over $\lambda_0,\nu$, is jointly convex in $(\alpha,\gamma_0,\theta,\tau_g)$ and jointly concave in $(\beta,\tau_h)$. Because of that we can interchange the order of minimization and minimization over using Sion's minimax theorem to get the following.

$$\min_{\theta,0\le\alpha,\gamma_0,\tau_g}\ \max_{0\le\beta,\tau_h}\ D_{\text{ns}}(\alpha,\gamma_0,\theta,\tau_g,\beta,\tau_h)$$

$$D_{\text{ns}}(\alpha,\gamma_0,\theta,\tau_g,\beta,\tau_h)=\frac{\beta\tau_g}{2}+L\left(\sqrt{\alpha^2+a^2\theta^2},V\theta-\varepsilon_0\gamma_0,\frac{\tau_g}{\beta}\right)$$

(B.79)

$$-\min_{\lambda_0\ge0,\nu}\left[\frac{\alpha}{\tau_h}\left\{\frac{\beta^2}{2\delta}+\lambda_0\left(\frac{\gamma_0\tau_h}{\alpha}\right)^q+\frac{\nu^2}{2}-\mathsf{E}\left(\beta,\left(\frac{\tau_h\theta}{\alpha}+\frac{\nu}{a}\right);\lambda_0\right)\right\}+\frac{\alpha\tau_h}{2}\right].$$

B.3.3. *Uniform convergence of the auxiliary problem to its scalarized version* We showed that the auxiliary optimization objective converges pointwise to the function $D_{\text{ns}}$ given by (B.79). However, we are interested in the minimax optimal solution of the auxiliary problem and need to have convergence of optimal points to the minimax solution of $D$. What is required for this aim is (local) uniform convergence of the auxiliary objective to function $D$. This can be shown by following similar arguments as in [63, Lemma A.5] that is essentially based on a result known as "convexity lemma" in the literature (see e.g. [39, Lemma 7.75]) by which pointwise convergence of convex functions implies uniform convergence in compact subsets.

B.3.4. *Uniqueness of the solution of the AO problem* First note that since the loss $\ell(t)$ is a convex function and $\frac{1}{2\mu}(x-t)^2$ is jointly convex in $(x,t,\mu)$, then $\frac{1}{2\mu}(x-t)^2+\ell(t)$ is jointly convex in $(x,t,\mu)$. Given that partial minimization preserves convexity, the Moreau envelope $e_\ell(x;\mu)$ is jointly convex in $(x,\mu)$. In addition, by using the result of [63, Lemma 4.4] the

expected Moreau envelope of a convex function is jointly "strictly" convex (indeed this holds without requiring any strong or strict convexity assumption on the function itself). An application of this result to our case implies that $L(a,b,\mu)$ is jointly strictly convex in $\mathbb{R}_{\geq 0}\times\mathbb{R}\times\mathbb{R}_{\geq 0}$.

In addition, as we argued before the function $D_{\rm ns}$ given by (B.79) is jointly convex in $(\alpha,\gamma_0,\theta,\tau_g)$ and jointly concave in $(\beta,\tau_h)$. Hence, using strict convexity of $L(a,b,\mu)$, the function $D_{\rm ns}$ is indeed jointly "strictly" convex in $(\alpha,\gamma_0,\theta,\tau_g)$ and jointly concave in $(\beta,\tau_h)$.

As the next step, we note that $\max_{\beta,\tau_h} D(\alpha,\gamma_0,\theta,\tau_g,\beta,\tau_h)$ is strictly convex in $(\alpha,\gamma_0,\theta,\tau_g)$. This follows from the fact that if a function $f(\boldsymbol{x},\boldsymbol{y})$ is strictly convex in $\boldsymbol{x}$, then $\max_{\boldsymbol{y}} f(\boldsymbol{x},\boldsymbol{y})$ is also strictly convex in $\boldsymbol{x}$. Moreover, by using the result of [63, Lemma C.5] we have that $\inf_{\tau_g>0}\max_{\beta,\tau_h} D(\alpha,\gamma_0,\theta,\tau_g,\beta,\tau_h)$ is strictly convex in $(\alpha,\gamma_0,\theta)$ and therefore has a unique minimizer $(\alpha_*,\gamma_{0*},\theta_*)$. This concludes the part (a) of the theorem and the given scalar minimax optimization to characterize the limiting behavior of parameter of interest $\alpha,\gamma_0,\theta$.

Part (b) of the theorem follows readily from our definition of parameters $\alpha$, $\theta$ and $\gamma$. Part (c) of the theorem also follows from combining Lemma 2.1 with part (b) of the theorem.

This completes the proof of Theorem 6.4.

**B.4. Proof of Remark 6.1** We start by establishing an explicit expression for the weighted Moreau envelope $e_{q,\boldsymbol{\Sigma}}$ for case of $p=q=2$.

**Lemma B.6** *We have*

$$e_{2,\boldsymbol{\Sigma}}(\boldsymbol{x};\lambda)=\lambda\left\|(\boldsymbol{\Sigma}+2\lambda\boldsymbol{I})^{-1/2}\boldsymbol{\Sigma}^{1/2}\boldsymbol{x}\right\|_{\ell_2}^2$$

The proof of Lemma B.6 is given in Appendix E.9.

Suppose that items $(i),(ii)$ in the statement of the remark are satisfied. We then prove that Assumption 6 and 7 hold.

**Proof** [Verification of Assumption 6] To check Assumption 6 for $p=q=2$, we use Lemma B.6 to get

$$\lim_{n\to\infty} e_{2,\boldsymbol{I}+b_0\boldsymbol{\Sigma}}\left((\boldsymbol{I}+b_0\boldsymbol{\Sigma})^{-1}\left\{\frac{c_0}{2\sqrt{n}}\boldsymbol{\Sigma}^{1/2}\mathsf{P}_{\boldsymbol{\mu}}^{\perp}\boldsymbol{h}-\frac{c_1}{2}\widetilde{\boldsymbol{\mu}}\right\};b_1\|\boldsymbol{\mu}\|_{\ell_2}^2\right)$$

(B.80)

$$=\lim_{n\to\infty} b_1\|\boldsymbol{\mu}\|_{\ell_2}^2\left\|\left((1+2b_1\|\boldsymbol{\mu}\|_{\ell_2}^2)\boldsymbol{I}+b_0\boldsymbol{\Sigma}\right)^{-1/2}(\boldsymbol{I}+b_0\boldsymbol{\Sigma})^{-1/2}\left(\frac{c_0}{2\sqrt{n}}\boldsymbol{\Sigma}^{1/2}\mathsf{P}_{\boldsymbol{\mu}}^{\perp}\boldsymbol{h}-\frac{c_1}{2}\widetilde{\boldsymbol{\mu}}\right)\right\|_{\ell_2}^2$$

Consider a singular value decomposition $\boldsymbol{\Sigma}=\boldsymbol{U}\boldsymbol{S}\boldsymbol{U}^T$ with $\boldsymbol{S}=\mathrm{diag}(s_1,\dots,s_d)$, and the first column of $\boldsymbol{U}$ being $\widetilde{\boldsymbol{\mu}}$ and $s_1=a^2$ (Recall that $\widetilde{\boldsymbol{\mu}}$ is a singular

value of $\boldsymbol{\Sigma}$ with eigenvalue $a^2$.) Also let $\tilde{\boldsymbol{h}} \coloneqq \boldsymbol{U}^T\boldsymbol{h} \sim \mathsf{N}(0, \boldsymbol{I}_d)$. Continuing from (B.80) we write

$$\lim_{n\to\infty} e_{2,\boldsymbol{I}+b_0\boldsymbol{\Sigma}}\left((\boldsymbol{I}+b_0\boldsymbol{\Sigma})^{-1}\left\{\frac{c_0}{2\sqrt{n}}\boldsymbol{\Sigma}^{1/2}\mathsf{P}_{\boldsymbol{\mu}}^{\perp}\boldsymbol{h}-\frac{c_1}{2}\widetilde{\boldsymbol{\mu}}\right\}; b_1\left\|\boldsymbol{\mu}\right\|_{\ell_2}^2\right)$$

(B.81)

$$= \lim_{n\to\infty}\frac{1}{n}b_1\left\|\boldsymbol{\mu}\right\|_{\ell_2}^2\left\|\boldsymbol{U}\left((1+2b_1\left\|\boldsymbol{\mu}\right\|_{\ell_2}^2)\boldsymbol{I}+b_0\boldsymbol{S}\right)^{-1/2}(\boldsymbol{I}+b_0\boldsymbol{S})^{-1/2}\boldsymbol{U}^T\left(\frac{c_0}{2}\boldsymbol{\Sigma}^{1/2}\mathsf{P}_{\boldsymbol{\mu}}^{\perp}\boldsymbol{h}-\frac{c_1}{2}\sqrt{n}\widetilde{\boldsymbol{\mu}}\right)\right\|_{\ell_2}^2$$

Write $\boldsymbol{U} = [\widetilde{\boldsymbol{\mu}}, \tilde{\boldsymbol{U}}]$ and $\tilde{\boldsymbol{S}} = \mathrm{diag}(s_2, \ldots, s_d)$. In addition, define $\tilde{\boldsymbol{h}} \coloneqq \tilde{\boldsymbol{U}}^T\boldsymbol{h} \sim \mathsf{N}(0, \boldsymbol{I}_{d-1})$. We then have

$$\boldsymbol{U}^T\boldsymbol{\Sigma}^{1/2}\mathsf{P}_{\boldsymbol{\mu}}^{\perp}\boldsymbol{h} = \begin{pmatrix}0\\ \tilde{\boldsymbol{S}}^{1/2}\tilde{\boldsymbol{h}}\end{pmatrix}, \quad \boldsymbol{U}^T\widetilde{\boldsymbol{\mu}} = \boldsymbol{e}_1\,, \quad \lim_{n\to\infty}\left\|\boldsymbol{\mu}\right\|_{\ell_2} = \sigma_{M,2}\,,$$

in probability, with the last limit following from item $(i)$ in the statement Remark 6.1. Using the above identities in (B.81) we get

$$\lim_{n\to\infty} e_{2,\boldsymbol{I}+b_0\boldsymbol{\Sigma}}\left((\boldsymbol{I}+b_0\boldsymbol{\Sigma})^{-1}\left\{\frac{c_0}{2\sqrt{n}}\boldsymbol{\Sigma}^{1/2}\mathsf{P}_{\boldsymbol{\mu}}^{\perp}\boldsymbol{h}-\frac{c_1}{2}\widetilde{\boldsymbol{\mu}}\right\}; b_1\left\|\boldsymbol{\mu}\right\|_{\ell_2}^2\right)$$

$$= \lim_{n\to\infty}\frac{b_1\sigma_{M,2}^2}{n}\left\{\frac{c_1^2 n}{4}\cdot\frac{1}{(1+2b_1\sigma_{M,2}^2+b_0a^2)(1+b_0a^2)}+\frac{c_0^2}{4}\sum_{i=2}^{d}\frac{s_i\tilde{h}_i^2}{(1+b_0s_i)(1+2b_1\sigma_{M,2}^2+b_0s_i)}\right\}$$

(B.82)

$$= b_1\sigma_{M,2}^2\left\{\frac{c_1^2}{4(1+b_0a^2)(1+2b_1\sigma_{M,2}^2+b_0a^2)}+\frac{c_0^2}{4\delta}\lim_{d\to\infty}\frac{1}{d}\sum_{i=2}^{d}\frac{s_i\tilde{h}_i^2}{(1+b_0s_i)(1+2b_1\sigma_{M,2}^2+b_0s_i)}\right\}$$

Define $\nu_i \coloneqq s_i(1+b_0s_i)^{-1}(1+2b_1\sigma_{M,2}^2+b_0s_i)^{-1}$. Then the last sum reads as $\frac{1}{d}\sum_{i=2}^{d}\nu_i\tilde{h}_i^2$. Recall that $\tilde{\boldsymbol{h}} \sim \mathsf{N}(0, \boldsymbol{I}_{d-1})$. Therefore, by applying the Kolmogorov's criterion of SLLN the above limit exists (almost surely and so in probability as well) provided that $\frac{1}{d^2}\sum_{i=2}^{d}\nu_i^2\mathrm{Var}(\tilde{h}_i^2) < \infty$. We note that $Var(\tilde{h}_i^2) = 2$ and since $\nu_i \geq 0$, we have

$$\frac{1}{d^2}\sum_{i=2}^{d}\nu_i^2 \leq \left(\frac{1}{d}\sum_{i=2}^{d}\nu_i\right)^2.$$

Hence it suffices to show that $\frac{1}{d}\sum_{i=2}^{d}\nu_i < \infty$. Now by item $(ii)$ of Remark 6.1, the empirical distribution of eigenvalues of $\boldsymbol{\Sigma}$ converges weakly to a distribu-

tion $\rho$ with Stieltjes transform $S_\rho(z) := \int \frac{\rho(t)}{z-t}\mathrm{d}t$. We write

$$\begin{aligned}
&\frac{s_i}{(1+b_0 s_i)(1+2b_1\sigma_{M,2}^2+b_0 s_i)} \\
&= \frac{1}{2b_0 b_1\sigma_{M,2}^2}\left\{-\frac{1}{1+b_0 s_i}+\frac{1+2b_1\sigma_{M,2}^2}{1+2b_1\sigma_{M,2}^2+b_0 s_i}\right\} \\
&= \frac{1}{2b_0^2 b_1\sigma_{M,2}^2}\left\{-\frac{1}{\frac{1}{b_0}+s_i}+\frac{1+2b_1\sigma_{M,2}^2}{\frac{1+2b_1\sigma_{M,2}^2}{b_0}+s_i}\right\} \qquad \text{(B.83)}
\end{aligned}$$

Therefore,

$$\begin{aligned}
\lim_{d\to\infty}\frac{1}{d}\sum_{i=2}^{d}\nu_i &= \lim_{d\to\infty}\frac{1}{d}\sum_{i=2}^{d}\frac{s_i}{(1+b_0 s_i)(1+2b_1\sigma_{M,2}^2+b_0 s_i)} \\
&= \frac{1}{2b_0^2 b_1\sigma_{M,2}^2}\lim_{d\to\infty}\frac{1}{d}\sum_{i=2}^{d}\left\{-\frac{1}{\frac{1}{b_0}+s_i}+\frac{1+2b_1\sigma_{M,2}^2}{\frac{1+2b_1\sigma_{M,2}^2}{b_0}+s_i}\right\} \\
&= \frac{1}{2b_0^2 b_1\sigma_{M,2}^2}\left\{S_\rho\left(-\frac{1}{b_0}\right)-(1+2b_1\sigma_{M,2}^2)S_\rho\left(-\frac{1+2b_1\sigma_{M,2}^2}{b_0}\right)\right\}. \qquad \text{(B.84)}
\end{aligned}$$

It is worth noting that although the sum is over $2 \le i \le d$, the term for $i = 1$ is $O(1/d)$ and is negligible in the limit. Therefore, we can include that in our calculation above. By using Equation (B.84) in (B.82) we get that Assumption 6 holds with

$$\begin{aligned}
\mathsf{F}(c_0,c_1;b_0,b_1) =& \frac{b_1\sigma_{M,2}^2 c_1^2}{4(1+b_0 a^2)(1+2b_1\sigma_{M,2}^2+b_0 a^2)} \\
&+\frac{b_1\sigma_{M,2}^2 c_0^2}{8\delta b_0^2 b_1\sigma_{M,2}^2}\left\{S_\rho\left(-\frac{1}{b_0}\right)-(1+2b_1\sigma_{M,2}^2)S_\rho\left(-\frac{1+2b_1\sigma_{M,2}^2}{b_0}\right)\right\}. \qquad \text{(B.85)}
\end{aligned}$$

■

**Proof** [Verification of Assumption 7] To check Assumption 7 we use Lemma B.6 and write

$$\begin{aligned}
&\lim_{n\to\infty} e_{2,\boldsymbol{\Sigma}}\left(\frac{c_0}{\sqrt{n}}\boldsymbol{\Sigma}^{-1/2}\boldsymbol{h}-c_1\widetilde{\boldsymbol{\mu}};\lambda_0\|\boldsymbol{\mu}\|_{\ell_2}^2\right) \\
&= \lim_{n\to\infty}\lambda_0\|\boldsymbol{\mu}\|_{\ell_2}^2\left\|(\boldsymbol{\Sigma}+2\lambda_0\|\boldsymbol{\mu}\|_{\ell_2}^2\boldsymbol{I})^{-1/2}\boldsymbol{\Sigma}^{1/2}\left(\frac{c_0}{\sqrt{n}}\boldsymbol{\Sigma}^{-1/2}\boldsymbol{h}-c_1\widetilde{\boldsymbol{\mu}}\right)\right\|_{\ell_2}^2 \\
&= \lim_{n\to\infty}\lambda_0\|\boldsymbol{\mu}\|_{\ell_2}^2\left\|(\boldsymbol{\Sigma}+2\lambda_0\|\boldsymbol{\mu}\|_{\ell_2}^2\boldsymbol{I})^{-1/2}\left(\frac{c_0}{\sqrt{n}}\boldsymbol{h}-c_1 a\widetilde{\boldsymbol{\mu}}\right)\right\|_{\ell_2}^2 \qquad \text{(B.86)}
\end{aligned}$$

Consider a singular value decomposition $\boldsymbol{\Sigma} = \boldsymbol{U}\boldsymbol{S}\boldsymbol{U}^T$ with $\boldsymbol{S} = \mathrm{diag}(s_1, \ldots, s_d)$, and the first column of $\boldsymbol{U}$ being $\widetilde{\boldsymbol{\mu}}$ and $s_1 = a^2$ (Recall that $\widetilde{\boldsymbol{\mu}}$ is a singular value of $\boldsymbol{\Sigma}$ with eigenvalue $a^2$.) Also let $\tilde{\boldsymbol{h}} := \boldsymbol{U}^T\boldsymbol{h} \sim \mathsf{N}(0, \boldsymbol{I}_d)$. Continuing from (B.86) we write

$$\begin{aligned}
&\lim_{n\to\infty} e_{2,\boldsymbol{\Sigma}}\left(\frac{c_0}{\sqrt{n}}\boldsymbol{\Sigma}^{-1/2}\boldsymbol{h} - c_1\widetilde{\boldsymbol{\mu}}; \lambda_0 \|\boldsymbol{\mu}\|_{\ell_2}^2\right)\\
&= \lim_{n\to\infty} \frac{1}{n}\lambda_0 \|\boldsymbol{\mu}\|_{\ell_2}^2 \left\|\boldsymbol{U}(\boldsymbol{S} + 2\lambda_0 \|\boldsymbol{\mu}\|_{\ell_2}^2 \boldsymbol{I})^{-1/2}\left(c_0\tilde{\boldsymbol{h}} - c_1\sqrt{n}a\boldsymbol{e}_1\right)\right\|_{\ell_2}^2\\
&= \lim_{n\to\infty} \frac{\lambda_0\sigma_{M,2}^2}{n}\left\{\frac{(c_0\tilde{h}_1 - c_1\sqrt{n}a)^2}{a^2 + 2\lambda_0\sigma_{M,2}^2} + \sum_{i=2}^{d}\frac{c_0^2\tilde{h}_i^2}{s_i + 2\lambda_0\sigma_{M,2}^2}\right\}\\
&= \lambda_0\sigma_{M,2}^2\left\{\frac{c_1^2a^2}{a^2 + 2\lambda_0\sigma_{M,2}^2} + \frac{1}{\delta}\lim_{d\to\infty}\frac{1}{d}\sum_{i=2}^{d}\frac{c_0^2\tilde{h}_i^2}{s_i + 2\lambda_0\sigma_{M,2}^2}\right\}. \qquad \text{(B.87)}
\end{aligned}$$

By applying the Kolmogorov's criterion of SLLN the above limit exists (almost surely and so in probability as well) provided that $\frac{1}{d^2}\sum_{i=2}^{d}\frac{1}{(s_i+2\lambda_0\sigma_{M,2}^2)^2} < \infty$. Note that since $\lambda_0, s_i \geq 0$, we have

$$\frac{1}{d^2}\sum_{i=2}^{d}\frac{1}{(s_i + 2\lambda_0\sigma_{M,2}^2)^2} \leq \frac{1}{d^2}\sum_{i=2}^{d}\frac{1}{4\lambda_0^2\sigma_{M,2}^4} \to 0\,.$$

By using the LLN we obtain that the summation in (B.87) converges (almost surely) to its expectation. Now recalling item $(ii)$ in Remark 6.1, we know that the empirical distribution of eigenvalues of $\boldsymbol{\Sigma}$ converges weakly to a distribution $\rho$ with Stieltjes transform $S_\rho(z) := \int \frac{\rho(t)}{z-t}\mathrm{d}t$, and therefore we have

$$\begin{aligned}
\mathsf{E}(c_0, c_1; \lambda_0) &:= \lim_{n\to\infty} e_{2,\boldsymbol{\Sigma}}\left(\frac{c_0}{\sqrt{n}}\boldsymbol{\Sigma}^{-1/2}\boldsymbol{h} - c_1\widetilde{\boldsymbol{\mu}}; \lambda_0 \|\boldsymbol{\mu}\|_{\ell_2}^2\right)\\
&= \lambda_0\sigma_{M,2}^2\left\{\frac{c_1^2a^2}{a^2 + 2\lambda_0\sigma_{M,2}^2} - \frac{c_0^2}{\delta}S_\rho(-2\lambda_0\sigma_{M,2}^2)\right\}. \qquad \text{(B.88)}
\end{aligned}$$

■

## APPENDIX C: PROOFS FOR ISOTROPIC GAUSSIAN MODEL (SECTION 4)

This section is devoted to the proof of our theorems for the isotropic Gaussian model. We discuss how these theorems can be derived as special cases of our results for the anisotropic model, after some algebraic simplifications.

The claim of Theorem 4.1 on the separability threshold is an immediate corollary of Theorem 6.1, with $\mathbf{\Sigma} = \boldsymbol{I}_{p\times p}$ and $a = 1$. We next move to the two other theorems on precise characterization of standard and robust accuracy in the separable and non-separable regimes.

**C.1. Proof of Theorem 4.3** Suppose that Assumption 4 in the statement of Theorem 4.3 holds. We first show that this assumption implies Assumption 6, required by Theorem 6.3, in case of $\mathbf{\Sigma} = \boldsymbol{I}$ and then show how Theorem 4.3 can be derived as a special case of Theorem 6.3.

To prove Assumption 6(b) for isotropic case, we use the following two properties of the weighted Moreau envelop that holds for all $q \geq 0$:

$$e_{q,\alpha\boldsymbol{I}}(\boldsymbol{x};\lambda) = \alpha e_{q,\boldsymbol{I}}\left(\boldsymbol{x},\frac{\lambda}{\alpha}\right), \tag{C.1}$$

$$\frac{1}{b^2}e_{q,\boldsymbol{I}}\left(b\boldsymbol{x};\frac{\lambda}{b^{q-2}}\right) = e_{q,\boldsymbol{I}}(\boldsymbol{x};\lambda). \tag{C.2}$$

Combining the above two identities we get

$$e_{q,\alpha\boldsymbol{I}}(\alpha^{-1}\boldsymbol{x};\lambda) = \alpha e_{q,\boldsymbol{I}}\left(\frac{\boldsymbol{x}}{\alpha};\frac{\lambda}{\alpha}\right) = \frac{\alpha}{b^2}e_{q,\boldsymbol{I}}\left(b\frac{\boldsymbol{x}}{\alpha};\frac{\lambda}{\alpha b^{q-2}}\right).$$

Using the above identity with $\alpha = 1 + b_0$ and $b = \alpha\sqrt{d}$ we have

$$\begin{aligned} &e_{q,(1+b_0)\boldsymbol{I}}\left((1+b_0)^{-1}\left\{\frac{c_0}{2\sqrt{n}}\mathsf{P}_{\boldsymbol{\mu}}^{\perp}\boldsymbol{h} - \frac{c_1}{2}\widetilde{\boldsymbol{\mu}}\right\}; b_1\|\boldsymbol{\mu}\|_{\ell_p}^q\right) \\ &= \frac{1}{(1+b_0)d}e_{q,\boldsymbol{I}}\left(\frac{\sqrt{d}c_0}{2\sqrt{n}}\mathsf{P}_{\boldsymbol{\mu}}^{\perp}\boldsymbol{h} - \frac{c_1\sqrt{d}}{2}\widetilde{\boldsymbol{\mu}}; \frac{b_1\|\boldsymbol{\mu}\|_{\ell_p}^q}{(1+b_0)^{q-1}d^{\frac{q}{2}-1}}\right) \end{aligned} \tag{C.3}$$

We next proceed to take the limit of the above expression as $n \to \infty$. By Assumption 4 we have

$$\|\boldsymbol{\mu}\|_{\ell_p}^q \to \sigma_{M,p}^q d^{\frac{q}{2}-1}, \quad \|\boldsymbol{\mu}\|_{\ell_2} \to \sigma_{M,2}, \tag{C.4}$$

with high probability. Also by Assumption 1, we have $n/d \to \delta$. Therefore,

$$\begin{aligned} &\lim_{n\to\infty} e_{q,(1+b_0)\boldsymbol{I}}\left((1+b_0)^{-1}\left\{\frac{c_0}{2\sqrt{n}}\mathsf{P}_{\boldsymbol{\mu}}^{\perp}\boldsymbol{h} - \frac{c_1}{2}\widetilde{\boldsymbol{\mu}}\right\}; b_1\|\boldsymbol{\mu}\|_{\ell_p}^q\right) \\ &= \lim_{n\to\infty}\frac{1}{(1+b_0)d}e_{q,\boldsymbol{I}}\left(\frac{\sqrt{d}c_0}{2\sqrt{n}}\mathsf{P}_{\boldsymbol{\mu}}^{\perp}\boldsymbol{h} - \frac{c_1\sqrt{d}}{2}\widetilde{\boldsymbol{\mu}}; \frac{b_1\|\boldsymbol{\mu}\|_{\ell_p}^q}{(1+b_0)^{q-1}d^{\frac{q}{2}-1}}\right) \\ &= \lim_{n\to\infty}\frac{1}{(1+b_0)d}e_{q,\boldsymbol{I}}\left(\frac{\sqrt{d}c_0}{2\sqrt{n}}\mathsf{P}_{\boldsymbol{\mu}}^{\perp}\boldsymbol{h} - \frac{c_1}{2}\frac{\sqrt{d}\boldsymbol{\mu}}{\|\boldsymbol{\mu}\|_{\ell_2}}; \frac{b_1\|\boldsymbol{\mu}\|_{\ell_p}^q}{(1+b_0)^{q-1}d^{\frac{q}{2}-1}}\right) \\ &= \lim_{n\to\infty}\frac{1}{(1+b_0)d}e_{q,\boldsymbol{I}}\left(\frac{c_0}{2\sqrt{\delta}}\mathsf{P}_{\boldsymbol{\mu}}^{\perp}\boldsymbol{h} - \frac{c_1}{2}\frac{\sqrt{d}\boldsymbol{\mu}}{\sigma_{M,2}}; b_1(1+b_0)^{1-q}\sigma_{M,p}^q\right) \end{aligned} \tag{C.5}$$

We next note that by definition of the weighted Moreau envelop we have $e_{q,\boldsymbol{I}}(\boldsymbol{x};\lambda) = \sum_{i=1}^{d} J_q(x_i;\lambda)$. Also, by Assumption 7 the empirical distribution of entries of $\sqrt{d}\boldsymbol{\mu}$ converges weakly to distribution $\mathbb{P}_M$. Therefore, continuing from (C.5) we can write

$$\begin{aligned}
&\lim_{n\to\infty} e_{q,(1+b_0)\boldsymbol{I}}\left((1+b_0)^{-1}\left\{\frac{c_0}{2\sqrt{n}}\mathsf{P}_{\boldsymbol{\mu}}^{\perp}\boldsymbol{h}-\frac{c_1}{2}\widetilde{\boldsymbol{\mu}}\right\};b_1\left\|\boldsymbol{\mu}\right\|_{\ell_p}^q\right)\\
&\overset{(a)}{=}\lim_{n\to\infty} e_{q,(1+b_0)\boldsymbol{I}}\left((1+b_0)^{-1}\left\{\frac{c_0}{2\sqrt{n}}\boldsymbol{h}-\frac{c_1}{2}\widetilde{\boldsymbol{\mu}}\right\};b_1\left\|\boldsymbol{\mu}\right\|_{\ell_p}^q\right)\\
&=\lim_{n\to\infty}\frac{1}{(1+b_0)d}\sum_{i=1}^{d} J_q\left(\frac{c_0 h_i}{2\sqrt{\delta}}-\frac{c_1}{2}\frac{\sqrt{d}\mu_i}{\sigma_{M,2}};b_1(1+b_0)^{1-q}\sigma_{M,p}^q\right)
\end{aligned}\tag{C.6}$$

$$\overset{(b)}{=}\frac{1}{1+b_0}\mathbb{E}\left[J_q\left(\frac{c_0 h}{2\sqrt{\delta}}-\frac{c_1 M}{2\sigma_{M,2}};b_1(1+b_0)^{1-q}\sigma_{M,p}^q\right)\right]=(1+b_0)^{-1}\mathcal{J}\left(\frac{c_0}{2},\frac{c_1}{2};b_1(1+b_0)^{1-q}\right),$$

where the expectation in $(b)$ is taken with respect to the independent random variables $h\sim\mathsf{N}(0,1)$ and $M\sim\mathbb{P}_M$. The last equality follows by definition of function $\mathcal{J}$ given by (4.7) and by deploying Assumption 4.

Here, $(a)$ follows by writing

$$\frac{c_0}{2\sqrt{\delta}}\mathsf{P}_{\boldsymbol{\mu}}^{\perp}\boldsymbol{h}-\frac{c_1\sqrt{d}\boldsymbol{\mu}}{2\sigma_{M,2}}=\frac{c_0}{2\sqrt{\delta}}\boldsymbol{h}-\left(\frac{c_1\sqrt{d}}{2\sigma_{M,2}}+\frac{c_0}{2\sqrt{\delta}\left\|\boldsymbol{\mu}\right\|_{\ell_2}}\boldsymbol{h}^T\widetilde{\boldsymbol{\mu}}\right)\boldsymbol{\mu}$$

and noting that $\widetilde{\boldsymbol{\mu}}^T\boldsymbol{h}\sim\mathsf{N}(0,1)$ since $\left\|\widetilde{\boldsymbol{\mu}}\right\|_{\ell_2}=1$, and $\left\|\boldsymbol{\mu}\right\|_{\ell_2}\to\sigma_{M,2}$ which implies that the last term in the right-hand side is dominated by the second term therein that is of order $\sqrt{d}$.

The chain of equalities in (C.6) shows that Assumption 6(a) is satisfied by $F(c_0,c_1;b_0,b_1)=\mathcal{J}\left(\frac{c_0}{2},\frac{c_1}{2};b_1(1+b_0)^{1-q}\right)$ for the isotropic model.

Assumption 6(b) also clearly holds for isotropic model ($\boldsymbol{\Sigma}=\boldsymbol{I}$) with $S_\rho(z)=\frac{1}{z-1}$.

Now that Assumption 6 holds we can use the result of Theorem 6.3 for the special case of $\boldsymbol{\Sigma}=\boldsymbol{I}$. As we showed above for this case, we have the following identities

$$F(c_0,c_1;b_0,b_1)=\mathcal{J}\left(\frac{c_0}{2},\frac{c_1}{2};b_1(1+b_0)^{1-q}\right),\quad S_\rho(z)=\frac{1}{z-1}.\tag{C.7}$$

Now by using these identities in the AO problem (6.2) and after some simple algebraic manipulation we obtain the AO problem (4.8).

**C.2. Proof of Theorem 4.5** We prove Theorem 4.5 as a special case of Theorem 6.4. We first show that in the isotropic case, Assumption 4 implies Assumption 7, required by Theorem 6.4.

Note that by Assumption 4 we have

$$\|\boldsymbol{\mu}\|_{\ell_p}^q \to \sigma_{M,p}^q d^{\frac{q}{2}-1}, \quad \|\boldsymbol{\mu}\|_{\ell_2} \to \sigma_{M,2}\,, \tag{C.8}$$

with high probability.

We then write

$$\begin{aligned}
\lim_{n\to\infty} e_{q,\boldsymbol{I}}\left(\frac{c_0}{\sqrt{n}}\boldsymbol{h} - c_1\widetilde{\boldsymbol{\mu}}; \lambda_0\|\boldsymbol{\mu}\|_{\ell_p}^q\right) &\overset{(a)}{=} \lim_{n\to\infty}\frac{1}{d}e_{q,\boldsymbol{I}}\left(\frac{c_0\sqrt{d}}{\sqrt{n}}\boldsymbol{h} - c_1\sqrt{d}\widetilde{\boldsymbol{\mu}}; \frac{\lambda_0\|\boldsymbol{\mu}\|_{\ell_p}^q}{d^{\frac{q}{2}-1}}\right)\\
&= \lim_{n\to\infty}\frac{1}{d}e_{q,\boldsymbol{I}}\left(\frac{c_0\sqrt{d}}{\sqrt{n}}\boldsymbol{h} - c_1\frac{\sqrt{d}\boldsymbol{\mu}}{\|\boldsymbol{\mu}\|_{\ell_2}}; \frac{\lambda_0\|\boldsymbol{\mu}\|_{\ell_p}^q}{d^{\frac{q}{2}-1}}\right)\\
&\overset{(b)}{=} \lim_{n\to\infty}\frac{1}{d}e_{q,\boldsymbol{I}}\left(\frac{c_0}{\sqrt{\delta}}\boldsymbol{h} - c_1\frac{\sqrt{d}\boldsymbol{\mu}}{\sigma_{M,2}}; \lambda_0\sigma_{M,p}^q\right)\\
&\overset{(c)}{=} \lim_{n\to\infty}\frac{1}{d}\sum_{i=1}^d J_q\left(\frac{c_0 h_i}{\sqrt{\delta}} - c_1\frac{\sqrt{d}\mu_i}{\sigma_{M,2}}; \lambda_0\sigma_{M,p}^q\right)\\
&\overset{(d)}{=} \mathbb{E}\left[J_q\left(\frac{c_0}{\sqrt{\delta}}h - c_1\frac{M}{\sigma_{M,2}}; \lambda_0\sigma_{M,p}^q\right)\right]\\
&= \mathcal{J}\left(c_0, c_1; \lambda_0\right)\,.
\end{aligned} \tag{C.9}$$

Here $(a)$ follows from (C.2) with $b = \sqrt{d}$; $(b)$ follows from (C.8); $(c)$ holds due to the identity $e_{q,\boldsymbol{I}}(\boldsymbol{x};\lambda) = \sum_{i=1}^d J_q(x_i;\lambda)$, which follows readily from the definition of weighted Moreau envelop $e_{q,\boldsymbol{I}}$ and the function $J_q$ given by (4.5). Finally, $(d)$ holds due to Assumption 4. The series of equalities (C.9) implies that Assumption 7 holds in isotropic case with

$$\mathsf{E}(c_0, c_1; \lambda_0) = \mathcal{J}\left(c_0, c_1; \lambda_0\right)\,. \tag{C.10}$$

Having Assumption 6 in place, we can specialize the result of Theorem 6.4 to isotropic model. Substituting for $\mathsf{E}(c_0, c_1; \lambda_0)$ from (C.10) in the AO problem (6.4) yields the AO problem (4.16).

## APPENDIX D: PROOFS FOR SPECIAL CASES OF $P$ (SECTION 5)

**D.1. Proof of Corollary 5.1** Part (a) is already proved in Example 2, cf. (4.4).

**Proof** [Part (b)] We start by an explicit characterization of $\mathcal{J}$ function for case of $p = q = 2$.

**Lemma D.1** *Recall function $\mathcal{J}(c_0, c_1; \lambda_0)$ given by*

$$\mathcal{J}(c_0, c_1; \lambda_0) = \mathbb{E}\left[ J_q\left( \frac{c_0}{\sqrt{\delta}} h - c_1 \frac{M}{\sigma_{M,2}}; \lambda_0 \sigma_{M,p}^q \right) \right], \tag{D.1}$$

*Then the following identity holds for case of $p = q = 2$:*

$$\mathcal{J}(c_0, c_1; \lambda_0) = \frac{\lambda}{\alpha^2 + 2\lambda\alpha} \|\boldsymbol{x}\|_{\ell_2}^2$$

**Proof** It is straightforward to see that

$$J_2(x; \lambda) = \frac{\lambda}{1 + 2\lambda} x^2 .$$

Therefore,

$$\mathcal{J}(c_0, c_1; \lambda_0) = \frac{\lambda_0 \sigma_{M,2}^2}{1 + 2\lambda_0 \sigma_{M,2}^2} \mathbb{E}\left[ \left( \frac{c_0}{\sqrt{\delta}} h - c_1 \frac{M}{\sigma_{M,2}} \right)^2 \right] = \frac{\lambda_0 \sigma_{M,2}^2}{1 + 2\lambda_0 \sigma_{M,2}^2} \left( \frac{c_0^2}{\delta} + c_1^2 \right).$$

■

Using Lemma D.1 in AO problem (4.8) for $q = 2$, we have

$$\begin{aligned}
D_{\mathrm{s}}(\alpha, \gamma_0, \theta, \beta, \lambda_0, \eta, \tilde{\eta}) &= 2\left(1 + \frac{\eta}{2\alpha}\right)^{-1} \mathcal{J}\left( \frac{\beta}{2}, \frac{\tilde{\eta}}{2}; \frac{\lambda_0}{2\gamma_0}\left(1 + \frac{\eta}{2\alpha}\right)^{-1} \right) \\
&\quad - \left( \frac{\beta^2}{\delta} + \tilde{\eta}^2 \right) \frac{1}{4(1 + \frac{\eta}{2\alpha})} - \lambda_0\gamma_0 - \frac{\eta\alpha}{2} - \tilde{\eta}\theta \\
&\quad + \beta \sqrt{ \mathbb{E}\left[ \left( (1 + \varepsilon_0\gamma_0 - \theta\sigma_{M,2}) + \alpha g \right)_+^2 \right] } \\
&= \frac{\frac{\lambda_0}{4\gamma_0}\sigma_{M,2}^2}{(1 + \frac{\eta}{2\alpha})^2 + (1 + \frac{\eta}{2\alpha})\frac{\lambda_0}{\gamma_0}\sigma_{M,2}^2} \left( \frac{\beta^2}{\delta} + \tilde{\eta}^2 \right) \\
&\quad - \left( \frac{\beta^2}{\delta} + \tilde{\eta}^2 \right) \frac{1}{4(1 + \frac{\eta}{2\alpha})} - \lambda_0\gamma_0 - \frac{\eta\alpha}{2} - \tilde{\eta}\theta \\
&\quad + \beta \sqrt{ \mathbb{E}\left[ \left( (1 + \varepsilon_0\gamma_0 - \theta\sigma_{M,2}) + \alpha g \right)_+^2 \right] } \\
&= - \left( \frac{\beta^2}{\delta} + \tilde{\eta}^2 \right) \frac{1}{4(1 + \frac{\eta}{2\alpha} + \frac{\lambda_0}{\gamma_0}\sigma_{M,2}^2)} - \lambda_0\gamma_0 - \frac{\eta\alpha}{2} - \tilde{\eta}\theta \\
&\quad + \beta \sqrt{ \mathbb{E}\left[ \left( (1 + \varepsilon_0\gamma_0 - \theta\sigma_{M,2}) + \alpha g \right)_+^2 \right] }.
\end{aligned} \tag{D.2}$$

Setting $\frac{\partial D_s}{\partial \beta}$ to zero we conclude that

$$\widehat{\beta} = 2\delta\left(1+\frac{\eta}{2\alpha}+\frac{\lambda_0}{\gamma_0}\sigma_{M,2}^2\right)\sqrt{\mathbb{E}\left[\left((1+\varepsilon_0\gamma_0-\theta\sigma_{M,2})+\alpha g\right)_+^2\right]}.$$

Thus the AO problem reduces to

$$\begin{aligned}\min_{\alpha,\gamma_0\geq 0,\theta}\max_{\lambda_0,\eta\geq 0,\tilde{\eta}} & -\tilde{\eta}^2\frac{1}{4(1+\frac{\eta}{2\alpha}+\frac{\lambda_0}{\gamma_0}\sigma_{M,2}^2)}-\lambda_0\gamma_0-\frac{\eta\alpha}{2}-\tilde{\eta}\theta\\ &+\delta\left(1+\frac{\eta}{2\alpha}+\frac{\lambda_0}{\gamma_0}\sigma_{M,2}^2\right)\mathbb{E}\left[\left((1+\varepsilon_0\gamma_0-\theta\sigma_{M,2})+\alpha g\right)_+^2\right].\end{aligned}\tag{D.3}$$

Setting the derivative with respect to $\widetilde{\eta}$ to zero we arrive at

$$\widehat{\widetilde{\eta}} = -2\theta\left(1+\frac{\eta}{2\alpha}+\frac{\lambda_0}{\gamma_0}\sigma_{M,2}^2\right),$$

which further simplifies the AO problem to

$$\begin{aligned}\min_{\alpha,\gamma_0\geq 0,\theta}\max_{\lambda_0,\eta\geq 0} & -\lambda_0\gamma_0-\frac{\eta\alpha}{2}+\theta^2\left(1+\frac{\eta}{2\alpha}+\frac{\lambda_0}{\gamma_0}\sigma_{M,2}^2\right)\\ &+\delta\left(1+\frac{\eta}{2\alpha}+\frac{\lambda_0}{\gamma_0}\sigma_{M,2}^2\right)\mathbb{E}\left[\left((1+\varepsilon_0\gamma_0-\theta\sigma_{M,2})+\alpha g\right)_+^2\right].\end{aligned}\tag{D.4}$$

Note that if $\alpha^2 < \theta^2+\delta\,\mathbb{E}\left[\left((1+\varepsilon_0\gamma_0-\theta\sigma_{M,2})+\alpha g\right)_+^2\right]$ then the maximum over $\eta$ is $+\infty$. Furthermore, when $\alpha^2 \geq \theta^2+\delta\,\mathbb{E}\left[\left((1+\varepsilon_0\gamma_0-\theta\sigma_{M,2})+\alpha g\right)_+^2\right]$ then the optimal $\eta = 0$. Thus the above AO is equivalent to

$$\begin{aligned}\min_{\alpha,\gamma_0\geq 0,\theta}\max_{\lambda_0\geq 0} & -\lambda_0\gamma_0+\theta^2\left(1+\frac{\lambda_0}{\gamma_0}\sigma_{M,2}^2\right)\\ &+\delta\left(1+\frac{\lambda_0}{\gamma_0}\sigma_{M,2}^2\right)\mathbb{E}\left[\left((1+\varepsilon_0\gamma_0-\theta\sigma_{M,2})+\alpha g\right)_+^2\right]\\ \text{subject to}\quad & \alpha^2\geq\theta^2+\delta\,\mathbb{E}\left[\left((1+\varepsilon_0\gamma_0-\theta\sigma_{M,2})+\alpha g\right)_+^2\right]\end{aligned}\tag{D.5}$$

Using a similar argument for optimization over $\lambda_0$, it is straightforward to see that the above optimization is equivalent to

$$\begin{aligned}
&\min_{\alpha,\gamma_0\geq 0,\theta} \theta^2 + \delta\,\mathbb{E}\left[\left((1+\varepsilon_0\gamma_0-\theta\sigma_{M,2})+\alpha g\right)_+^2\right] \\
&\text{subject to} \quad \alpha^2 \geq \theta^2 + \delta\,\mathbb{E}\left[\left((1+\varepsilon_0\gamma_0-\theta\sigma_{M,2})+\alpha g\right)_+^2\right], \\
&\text{and} \quad \frac{\gamma_0^2}{\sigma_{M,2}^2} \geq \theta^2 + \delta\,\mathbb{E}\left[\left((1+\varepsilon_0\gamma_0-\theta\sigma_{M,2})+\alpha g\right)_+^2\right].
\end{aligned} \tag{D.6}$$

Since the objective function is increasing in $\alpha$ and $\gamma$, then the optimal $\alpha$ and $\gamma$ should make the inequality constraints equality and therefore $\gamma_0 = \alpha\sigma_{M,2}$. This brings us to the following problem:

$$\begin{aligned}
&\min_{\alpha\geq 0,u} \alpha^2 \\
&\text{subject to} \quad \alpha^2 \geq \theta^2 + \delta\,\mathbb{E}\left[\left(1+(\varepsilon_0\alpha-\theta)\sigma_{M,2}+\alpha g\right)_+^2\right].
\end{aligned} \tag{D.7}$$

By change of variable $u = \frac{\theta}{\alpha}$ we have

$$\begin{aligned}
&\min_{\alpha\geq 0,u} \alpha^2 \\
&\text{subject to} \quad 1 \geq u^2 + \delta\,\mathbb{E}\left[\left(\frac{1}{\alpha}+(\varepsilon_0-u)\sigma_{M,2}+g\right)_+^2\right]
\end{aligned} \tag{D.8}$$

By another change of variable $\tilde{\alpha} = \left(\frac{1}{\alpha}+\varepsilon_0\sigma_{M,2}\right)^{-1}$ we have

$$\begin{aligned}
&\min_{\frac{1}{\varepsilon_0\sigma_{M,2}}\geq\tilde{\alpha}\geq 0,u} \left(\frac{1}{\tilde{\alpha}}-\varepsilon_0\sigma_{M,2}\right)^{-2} \\
&\text{subject to} \quad 1 \geq u^2 + \delta\,\mathbb{E}\left[\left(\frac{1}{\tilde{\alpha}}-u\sigma_{M,2}+g\right)_+^2\right]
\end{aligned} \tag{D.9}$$

Since objective is increasing in $\tilde{\alpha}$ this is equivalent to

$$\begin{aligned}
&\min_{\frac{1}{\varepsilon_0\sigma_{M,2}}\geq\tilde{\alpha}\geq 0,u} \tilde{\alpha}^2 \\
&\text{subject to} \quad 1 \geq u^2 + \delta\,\mathbb{E}\left[\left(\frac{1}{\tilde{\alpha}}-u\sigma_{M,2}+g\right)_+^2\right]
\end{aligned} \tag{D.10}$$

Note that we can drop the constraint $\frac{1}{\varepsilon_0\sigma_{M,2}} \geq \tilde{\alpha}$ because for $\frac{1}{\varepsilon_0\sigma_{M,2}} = \tilde{\alpha}$ one can already find $u$ that satisfies the inequality constraint. As such the optimal

$\tilde{\alpha}$ should be less than $\frac{1}{\varepsilon_0\sigma_{M,2}}$. To see why, by letting $u = \frac{\theta}{\sqrt{1+\theta^2}}$ with $\theta$ the minimizer in separability condition (4.4) we have

$$\begin{aligned} u^2 + \delta\,\mathbb{E}\left[\left(\frac{1}{\tilde{\alpha}} - u\sigma_{M,2} + g\right)_+^2\right] &= \frac{\theta^2}{1+\theta^2} + \delta\,\mathbb{E}\left[\left((\varepsilon_0 - \frac{\theta}{\sqrt{1+\theta^2}})\sigma_{M,2} + g\right)_+^2\right] \\ &= \frac{\theta^2}{1+\theta^2} + \frac{\delta}{1+\theta^2}\,\mathbb{E}\left[\left((\sqrt{1+\theta^2}\varepsilon_0 - \theta)\sigma_{M,2} + \sqrt{1+\theta^2}g\right)_+^2\right] \\ &\le \frac{\theta^2}{1+\theta^2} + \frac{1}{1+\theta^2} = 1\,. \end{aligned}$$

This brings us to the following AO problem:

$$\begin{aligned} &\min_{\tilde{\alpha}\ge 0, u} \tilde{\alpha}^2 \\ &\text{subject to} \quad 1 \ge u^2 + \delta\,\mathbb{E}\left[\left(\frac{1}{\tilde{\alpha}} - u\sigma_{M,2} + g\right)_+^2\right] \end{aligned} \tag{D.11}$$

Denoting by $\tilde{\alpha}_*$ the solution of the above problem, it is clear that by our change of variable we have

$$\alpha_* = \left(\tilde{\alpha}_*^{-1} - \varepsilon_0\sigma_{M,2}\right)^{-1}, \quad \theta_* = u_*\alpha_*, \quad \gamma_{0*} = \alpha_*\sigma_{M,2}\,. \tag{D.12}$$

This concludes the proof of part (b). ∎

**Proof** [Part (c)] We focus on part of the AO problem (4.16) that involves the variables $\lambda_0, \nu, \tau_h$ and specialize it to the case of $q = 2$:

$$\min_{\lambda_0\ge 0,\nu}\left[\frac{\alpha}{\tau_h}\left\{\frac{\beta^2}{2\delta} + \lambda_0\left(\frac{\gamma_0\tau_h}{\alpha}\right)^2 + \frac{\nu^2}{2} - \mathcal{J}\left(\beta, \left(\frac{\tau_h\theta}{\alpha} + \nu\right); \lambda_0\right)\right\} + \frac{\alpha\tau_h}{2}\right].$$

We next plug in for $\mathcal{J}(c_0, c_1; \lambda_0)$ from Lemma D.1 which results in

$$\min_{\lambda_0\ge 0,\nu}\left[\frac{\alpha}{\tau_h}\left\{\frac{\beta^2}{2\delta} + \lambda_0\left(\frac{\gamma_0\tau_h}{\alpha}\right)^2 + \frac{\nu^2}{2} - \frac{\lambda_0\sigma_{M,2}^2}{1+2\lambda_0\sigma_{M,2}^2}\left(\frac{\beta^2}{\delta} + \left(\frac{\tau_h\theta}{\alpha} + \nu\right)^2\right)\right\} + \frac{\alpha\tau_h}{2}\right].$$

Writing the first order optimality for $\lambda_0$, $\nu$, $\tau_h$ we get a set of equations that admits a solution only if $\gamma_0 = \sigma_{M,2}\sqrt{\alpha^2+\theta^2}$. Then,

$$\nu = 2\lambda_0\sigma_{M,2}^2\frac{\tau_h\theta}{\alpha}, \quad \tau_h = \frac{1}{1+2\lambda_0\sigma_{M,2}^2}\frac{\beta}{\sqrt{\delta}}\,.$$

In this case, the value of $\lambda_0$ does not matter and the above part of the AO simplifies to $\alpha\beta/\sqrt{\delta}$.

This simplifies the AO problem (4.16) to

$$\max_{0\le\beta}\ \min_{\theta,0\le\alpha,\tau_g}\ \frac{\beta\tau_g}{2}+L\left(\sqrt{\alpha^2+\theta^2},\sigma_{M,2}\theta-\varepsilon_0\sqrt{\alpha^2+\theta^2},\frac{\tau_g}{\beta}\right)-\frac{\alpha\beta}{\sqrt{\delta}}\,. \tag{D.13}$$

We next further simplifies the AO problem by solving for $\tau_g$. We use the shorthand $L_3'(a,b;\mu)=\frac{\partial L}{\partial\mu}L(a,b;\mu)$ to denote the derivative of the expected Moreau envelop with respect to its third argument. Writing the first order optimality condition for $\beta$ and $\tau_g$ in optimization (D.13), we get

$$\begin{aligned}&\frac{\beta}{2}+\frac{1}{\beta}L_3'\left(\sqrt{\alpha^2+\theta^2},\sigma_{M,2}\theta-\varepsilon_0\sqrt{\alpha^2+\theta^2},\frac{\tau_g}{\beta}\right)=0\,,\\ &\frac{\tau_g}{2}-\frac{\tau_g}{\beta^2}L_3'\left(\sqrt{\alpha^2+\theta^2},\sigma_{M,2}\theta-\varepsilon_0\sqrt{\alpha^2+\theta^2},\frac{\tau_g}{\beta}\right)-\frac{\alpha}{\sqrt{\delta}}=0\,.\end{aligned} \tag{D.14}$$

Combining the above two equations, we obtain $\tau_g=\frac{\alpha}{\sqrt{\delta}}$. Substituting for $\tau_g$ in (D.13), the AO problem for case of $q=2$ simplifies to

$$\max_{0\le\beta}\ \min_{\theta,0\le\alpha}\ L\left(\sqrt{\alpha^2+\theta^2},\sigma_{M,2}\theta-\varepsilon_0\sqrt{\alpha^2+\theta^2},\frac{\alpha}{\beta\sqrt{\delta}}\right)-\frac{\alpha\beta}{2\sqrt{\delta}}\,. \tag{D.15}$$

This completes the proof of part (c). ∎

**D.2. Proof of Theorem 5.2** **Proof** [Part (a)] The first part of the theorem is on precise characterization of the separability threshold. We use the result of Theorem 4.1 that holds for any choice of $(p,q)$, in particular $(p=1,q=\infty)$, and relates the separability threshold to the spherical width. What is remaining to prove is the characterization of the spherical width given by (5.9). To this end, we follow a similar argument as in Lemma 4.2. However, since $\|\cdot\|_{\ell_q}^q$ is not well defined for $q=\infty$ (recall that $\ell_q$ is the dual norm of $\ell_p$ and $p=1$), it requires a slightly different analysis. Specifically, in the Lagrangian we write the constraint $\|\boldsymbol{u}\|_{\ell_\infty}\le\frac{1}{\varepsilon_0\|\boldsymbol{\mu}\|_{\ell_\infty}}$ as the term $\|\boldsymbol{u}\|_{\ell_\infty}-\frac{1}{\varepsilon_0\|\boldsymbol{\mu}\|_{\ell_p}}$, as compared to the case of finite $q$ where we raised the both sides to power $q$ to use the separability property of function $\|\cdot\|_{\ell_q}^q$. Then, by following a similar derivation as in (E.4), we obtain

$$\begin{aligned}&\min_{\boldsymbol{z}\in\mathcal{S}(\alpha,\theta,\varepsilon_0,\boldsymbol{\mu})}\ -\frac{1}{\sqrt{n}}\boldsymbol{h}^T\boldsymbol{z}\\ &=\sup_{\lambda,\eta\ge0,\nu}\ \eta J_\infty\left(\frac{\boldsymbol{h}}{\eta\sqrt{n}}-\left(\frac{\nu}{\eta}-\theta\right)\widetilde{\boldsymbol{\mu}};\frac{\lambda}{\eta}\right)-\frac{\nu^2}{2\eta}-\frac{1}{2\eta n}\|\boldsymbol{h}\|_{\ell_2}^2+\frac{\nu}{\eta\sqrt{n}}\widetilde{\boldsymbol{\mu}}^T\boldsymbol{h}-\frac{\eta}{2}\alpha^2-\frac{\lambda}{\varepsilon_0\|\boldsymbol{\mu}\|_{\ell_1}}\,,\end{aligned} \tag{D.16}$$

where $J_\infty(\boldsymbol{x};\lambda)$ is defined as

$$J_\infty(\boldsymbol{x};\lambda) = \min_{\boldsymbol{v}} \frac{1}{2}\|\boldsymbol{x}-\boldsymbol{v}\|_{\ell_2}^2 + \lambda\|\boldsymbol{v}\|_{\ell_\infty}\,. \tag{D.17}$$

Our next step is scalarization of the optimization (D.16) in the large sample limit (as $n\to\infty$), and a challenge along this way is that the function $J(\boldsymbol{x};\lambda)$ is not a separable function over the entries of $\boldsymbol{x}$. To cope with this problem, we propose an alternative representation of this function that involves an additional variable $t_0$.

We write

$$\begin{aligned} J_\infty(\boldsymbol{x};\lambda) &= \min_{\boldsymbol{v}} \frac{1}{2}\|\boldsymbol{x}-\boldsymbol{v}\|_{\ell_2}^2 + \lambda\|\boldsymbol{v}\|_{\ell_\infty} \\ &= \min_{\boldsymbol{v},t\ge 0} \frac{1}{2}\|\boldsymbol{x}-\boldsymbol{v}\|_{\ell_2}^2 + \lambda t \quad \text{subject to} \quad \|\boldsymbol{v}\|_{\ell_\infty}\le t \\ &= \min_{t\ge 0} \frac{1}{2}\|\mathsf{ST}(\boldsymbol{x};t)\|_{\ell_2}^2 + \lambda t\,. \end{aligned} \tag{D.18}$$

Let $t_0 = \|\boldsymbol{\mu}\|_{\ell_1} t$ and $\lambda_0 = \frac{\lambda}{\|\boldsymbol{\mu}\|_{\ell_1}}$. Similar to the trick of 'artificial' boundedness that we used in applying the CGMT framework (e.g., cf. explanation after (B.16) and Appendix E.3), we continue by the ansatz that the optimal value of $\lambda_0$ and $t_0$ remain bounded as $n\to\infty$. After we take the limit of the Lagrangian to obtain a scalar auxiliary optimization (AO) problem, this ansatz is verified by the boundedness of solutions of the AO problem.

For $\boldsymbol{x} = \frac{c_0}{\sqrt{n}}\boldsymbol{h} - c_1\widetilde{\boldsymbol{\mu}}$ we have

$$\begin{aligned} &\lim_{n\to\infty} \frac{1}{2}\|\mathsf{ST}(\boldsymbol{x};t)\|_{\ell_2}^2 + \lambda t \\ &= \lim_{n\to\infty} \frac{1}{2}\sum_{i=1}^d \mathsf{ST}\left(x_i; \frac{t_0}{\|\boldsymbol{\mu}\|_{\ell_1}}\right)^2 + \lambda_0 t_0 \\ &= \lim_{n\to\infty} \frac{1}{2}\sum_{i=1}^d \mathsf{ST}\left(\frac{c_0}{\sqrt{n}}h_i - c_1\widetilde{\boldsymbol{\mu}}_i; \frac{t_0}{\|\boldsymbol{\mu}\|_{\ell_1}}\right)^2 + \lambda_0 t_0 \\ &\overset{(a)}{=} \lim_{n\to\infty} \frac{1}{2d}\sum_{i=1}^d \mathsf{ST}\left(\frac{c_0}{\sqrt{\delta}}h_i - c_1\frac{\sqrt{d}\mu_i}{\sigma_{M,2}}; \frac{t_0}{\sigma_{M,1}}\right)^2 + \lambda_0 t_0 \\ &= \frac{1}{2}\mathbb{E}\left[\mathsf{ST}\left(\frac{c_0}{\sqrt{\delta}}h - c_1\frac{M}{\sigma_{M,2}}; \frac{t_0}{\sigma_{M,1}}\right)^2\right] + \lambda_0 t_0\,, \\ &= f(c_0,c_1;t_0) + \lambda_0 t_0\,, \end{aligned} \tag{D.19}$$

with high probability. In $(a)$ we used the fact that as $n\to\infty$, we have $\|\boldsymbol{\mu}\|_{\ell_1}\to\sqrt{d}\sigma_{M,1}$ along with the identity $\frac{1}{a^2}\mathsf{ST}(a\boldsymbol{x};a\lambda) = \mathsf{ST}(\boldsymbol{x};\lambda)$.

Note that this is a pointwise convergence. However, the left hand side is a convex function of $t$ and it's minimizer satisfies $t \le \|\boldsymbol{x}\|_\infty \le c_0 + c_1$ and so belongs to a compact set. Therefore, by applying the convexity lemma, see e.g, [39, Lemma 7.75], [63, Lemma B1], we can change the order of limit and minimization, and get that for $\boldsymbol{x} = \frac{c_0}{\sqrt{n}}\boldsymbol{h} - c_1\widetilde{\boldsymbol{\mu}}$ ,

$$\begin{aligned} &\lim_{n\to\infty} J_\infty(\boldsymbol{x};\lambda) \\ &= \lim_{n\to\infty} \min_{t\ge 0} \left\{ \frac{1}{2}\left\|\mathsf{ST}(\boldsymbol{x};t)\right\|_{\ell_2}^2 + \lambda t \right\} \\ &= \min_{t_0\ge 0} \left\{ f(c_0, c_1; t_0) + \lambda_0 t_0 \right\} , \end{aligned} \tag{D.20}$$

in probability. In addition, as $n \to \infty$ we have

$$\frac{1}{n}\|\boldsymbol{h}\|_{\ell_2}^2 \to \frac{1}{\delta}, \quad \frac{1}{\sqrt{n}}\widetilde{\boldsymbol{\mu}}^T\boldsymbol{h} \to 0\,, \quad \|\boldsymbol{\mu}\|_{\ell_1} \to \sigma_{M,1}\sqrt{d}\,, \tag{D.21}$$

with high probability.

Using the above limits, we see that the objective function (D.16) converges pointwise to the following function:

$$\min_{t_0\ge 0} \eta\left\{ f\left(\frac{1}{\eta}, \frac{\nu}{\eta} - \theta; t_0\right) + \frac{\lambda_0}{\eta} t_0 \right\} - \frac{\nu^2}{2\eta} - \frac{1}{2\eta\delta} - \frac{\eta}{2}\alpha^2 - \frac{\lambda_0}{\varepsilon_0} \tag{D.22}$$

Note that (E.4) is the dual optimization and hence is a concave problem. We apply the convexity lemma [63, Lemma B.2] to conclude that the objective value in (E.4) also converges to the supremum of function (E.8) over $\lambda_0, \eta \ge 0, \nu$. Therefore the solution of optimization (D.16) converges to the solution of the following optimization problem:

$$\sup_{\lambda_0,\eta\ge 0,\nu} \min_{t_0\ge 0} \eta f\left(\frac{1}{\eta}, \frac{\nu}{\eta} - \theta; t_0\right) + \lambda_0 t_0 - \frac{\nu^2}{2\eta} - \frac{1}{2\eta\delta} - \frac{\eta}{2}\alpha^2 - \frac{\lambda_0}{\varepsilon_0} \tag{D.23}$$

Note that the above objective is linear in $\lambda_0$. Therefore the optimal $t_0^*$ should satisfy $t_0^* \le \frac{1}{\varepsilon_0}$. Otherwise $\lambda_0^* = \infty$ which makes the above max-min value unbounded, and this is a contradiction because the above problem involves minimization over $t_0$ and it is easy to see that by choosing $t_0 = 0$ the optimal objective value over $\{\lambda_0, \eta \ge 0, \nu\}$ becomes zero.

Therefore, we can assume $t_0 \le \frac{1}{\varepsilon_0}$ which yields $\lambda_0^*(t_0 - \frac{1}{\varepsilon_0}) = 0$. This simplifies the problem (D.23) to

$$\sup_{\eta\ge 0,\nu} \min_{0\le t_0\le \frac{1}{\varepsilon_0}} \eta f\left(\frac{1}{\eta}, \frac{\nu}{\eta} - \theta; t_0\right) - \frac{\nu^2}{2\eta} - \frac{1}{2\eta\delta} - \frac{\eta}{2}\alpha^2 \tag{D.24}$$

Since $f(c_0, c_1; t)$ is decreasing in $t$, the optimal value $t_0^*$ is given by $t_0^* = \frac{1}{\varepsilon_0}$ and the problem is further simplified and along with (D.16) implies that

$$\begin{aligned}
&\min_{\boldsymbol{z}\in\mathcal{S}(\alpha,\theta,\varepsilon_0,\boldsymbol{\mu})} \; -\frac{1}{\sqrt{n}}\boldsymbol{h}^T\boldsymbol{z} \\
&= \sup_{\eta\geq 0,\nu} \; \eta f\left(\frac{1}{\eta}, \frac{\nu}{\eta} - \theta; \frac{1}{\varepsilon_0}\right) - \frac{\nu^2}{2\eta} - \frac{1}{2\eta\delta} - \frac{\eta}{2}\alpha^2 \qquad \text{(D.25)}
\end{aligned}$$

Now similar to the proof of Lemma 4.2 we use Equation (E.3) to write

$$\begin{aligned}
&\omega(\alpha, \theta, \varepsilon_0) \\
&= \lim_{n\to\infty} \omega_s\left(\mathcal{S}(\alpha, \theta, \varepsilon_0, \boldsymbol{\mu})\right) \\
&= -\sup_{\eta\geq 0,\nu} \; \sqrt{\delta}\left\{\eta f\left(\frac{1}{\eta}, \frac{\nu}{\eta} - \theta; \frac{1}{\varepsilon_0}\right) - \frac{\nu^2}{2\eta} - \frac{1}{2\eta\delta} - \frac{\eta}{2}\alpha^2\right\} \\
&= \min_{\eta\geq 0,\nu} \; \sqrt{\delta}\left\{\frac{\nu^2}{2\eta} + \frac{1}{2\eta\delta} + \frac{\eta}{2}\alpha^2 - \eta f\left(\frac{1}{\eta}, \frac{\nu}{\eta} - \theta; \frac{1}{\varepsilon_0}\right)\right\}. \qquad \text{(D.26)}
\end{aligned}$$

This completes the proof. ∎

**Proof** [Part (b)] The proof of this parts proceeds along the same lines of Theorem 4.3 for the special case of $p = 1$, $q = \infty$. However, it requires a slightly different treatment as in part (a) because the function $\|\cdot\|_{\ell_q}^q$ and therefore $J_q$ given by (4.5) are not well defined in this case.

Here we only highlight the modifications that are needed to the proof of Theorem 4.3 to apply it for case of $q = \infty$.

We proceed the exact same derivation that yields (B.22), repeated here for convenience:

$$\begin{aligned}
&\min_{\theta,\boldsymbol{\theta},\gamma\geq 0,\alpha\geq 0} \; \max_{\beta,\lambda,\eta\geq 0,\widetilde{\eta}} \; \|\boldsymbol{\theta}\|_{\ell_2}^2 + 2\lambda\left(\|\boldsymbol{\theta}\|_{\ell_q} - \gamma\right) + \frac{\beta}{\sqrt{n}}\boldsymbol{h}^T\mathsf{P}_{\boldsymbol{\mu}}^{\perp}\boldsymbol{\Sigma}^{1/2}\boldsymbol{\theta} \\
&\quad + \beta\sqrt{\mathbb{E}\left[\left(\left(1 + \varepsilon\gamma - \theta\,\|\boldsymbol{\mu}\|_{\ell_2}\right) + \alpha g\right)_+^2\right]} + \eta\left(\left\|\boldsymbol{\Sigma}^{\frac{1}{2}}\boldsymbol{\theta}\right\|_{\ell_2} - \alpha\right) + \widetilde{\eta}\left(\widetilde{\boldsymbol{\mu}}^T\boldsymbol{\theta} - \theta\right) \qquad \text{(D.27)}
\end{aligned}$$

We substitute for $\left\|\boldsymbol{\Sigma}^{\frac{1}{2}}\boldsymbol{\theta}\right\|_{\ell_2}$ using the identity $\left\|\boldsymbol{\Sigma}^{\frac{1}{2}}\boldsymbol{\theta}\right\|_{\ell_2} = \min_{\tau\geq 0} \frac{\left\|\boldsymbol{\Sigma}^{\frac{1}{2}}\boldsymbol{\theta}\right\|_{\ell_2}^2}{2\tau} + \frac{\tau}{2}$

to get

$$\min_{\theta,\boldsymbol{\theta},\gamma\geq 0,\alpha\geq 0}\ \max_{\beta,\lambda,\eta\geq 0,\widetilde{\eta}}\ \min_{\tau\geq 0}\quad \|\boldsymbol{\theta}\|_{\ell_2}^2+2\lambda\|\boldsymbol{\theta}\|_{\ell_\infty}-2\lambda\gamma+\frac{\beta}{\sqrt{n}}\boldsymbol{h}^T\mathsf{P}_{\boldsymbol{\mu}}^{\perp}\boldsymbol{\Sigma}^{1/2}\boldsymbol{\theta}$$
$$+\beta\sqrt{\mathbb{E}\left[\left(\left(1+\varepsilon\gamma-\theta\|\boldsymbol{\mu}\|_{\ell_2}\right)+\alpha g\right)_+^2\right]}$$
$$+\frac{\eta}{2\tau}\left\|\boldsymbol{\Sigma}^{\frac{1}{2}}\boldsymbol{\theta}\right\|_{\ell_2}^2+\frac{\eta\tau}{2}-\eta\alpha+\widetilde{\eta}\left(\widetilde{\boldsymbol{\mu}}^T\boldsymbol{\theta}-\theta\right). \tag{D.28}$$

Specializing it to $\boldsymbol{\Sigma}=\boldsymbol{I}$ and $q=\infty$, the optimization over $\boldsymbol{\theta}$ takes the form

$$\min_{\boldsymbol{\theta}}\quad \left(1+\frac{\eta}{2\tau}\right)\|\boldsymbol{\theta}\|_{\ell_2}^2+2\lambda\|\boldsymbol{\theta}\|_{\ell_\infty}+\frac{\beta}{\sqrt{n}}\boldsymbol{h}^T\mathsf{P}_{\boldsymbol{\mu}}^{\perp}\boldsymbol{\theta}+\widetilde{\eta}\widetilde{\boldsymbol{\mu}}^T\boldsymbol{\theta}\,. \tag{D.29}$$

Note that

$$\begin{aligned}\frac{\beta}{\sqrt{n}}\,\mathsf{P}_{\boldsymbol{\mu}}^{\perp}\boldsymbol{h}-\frac{\tilde{\eta}\boldsymbol{\mu}}{\|\boldsymbol{\mu}\|_{\ell_2}}&=\frac{\beta}{\sqrt{n}}\,\boldsymbol{h}-\frac{\tilde{\eta}\boldsymbol{\mu}}{\|\boldsymbol{\mu}\|_{\ell_2}}-\frac{\beta}{\sqrt{n}}\,\mathsf{P}_{\boldsymbol{\mu}}\boldsymbol{h}\\&=\frac{\beta}{\sqrt{n}}\,\boldsymbol{h}-\left(\frac{\tilde{\eta}}{\|\boldsymbol{\mu}\|_{\ell_2}}+\frac{\beta}{\sqrt{n}\,\|\boldsymbol{\mu}\|_{\ell_2}}\boldsymbol{h}^T\widetilde{\boldsymbol{\mu}}\right)\boldsymbol{\mu}\,,\end{aligned}$$

where $\widetilde{\boldsymbol{\mu}}^T\boldsymbol{h}\sim\mathsf{N}(0,1)$ since $\|\widetilde{\boldsymbol{\mu}}\|_{\ell_2}=1$, and $\|\boldsymbol{\mu}\|_{\ell_2}\to\sigma_{M,2}$ which implies that the last term in the right-hand side is dominated by the second term. Therefore in the asymptotic regime $n\to\infty$, we can equivalently work replace $\mathsf{P}_{\boldsymbol{\mu}}^{\perp}\boldsymbol{h}$ by $\boldsymbol{h}$ and by using the symmetry of the Gaussian distribution work with

$$\min_{\boldsymbol{\theta}}\quad \left(1+\frac{\eta}{2\tau}\right)\|\boldsymbol{\theta}\|_{\ell_2}^2+2\lambda\|\boldsymbol{\theta}\|_{\ell_\infty}-\frac{\beta}{\sqrt{n}}\boldsymbol{h}^T\boldsymbol{\theta}+\widetilde{\eta}\widetilde{\boldsymbol{\mu}}^T\boldsymbol{\theta}\,. \tag{D.30}$$

We let $\boldsymbol{x}=\frac{\beta}{\sqrt{n}}\boldsymbol{h}-\tilde{\eta}\widetilde{\boldsymbol{\mu}}$ and consider the change of variable $\boldsymbol{u}:=2(1+\frac{\eta}{2\tau})\boldsymbol{\theta}$ and write

$$\begin{aligned}&\min_{\boldsymbol{\theta}}\quad \left(1+\frac{\eta}{2\tau}\right)\|\boldsymbol{\theta}\|_{\ell_2}^2+2\lambda\|\boldsymbol{\theta}\|_{\ell_\infty}-\frac{\beta}{\sqrt{n}}\boldsymbol{h}^T\boldsymbol{\theta}+\tilde{\eta}\widetilde{\boldsymbol{\mu}}^T\boldsymbol{\theta}\\&=\min_{\boldsymbol{u}}\frac{1}{2}\left(1+\frac{\eta}{2\tau}\right)^{-1}\left\{\frac{1}{2}\|\boldsymbol{u}\|_{\ell_2}^2+2\lambda\|\boldsymbol{u}\|_{\ell_\infty}-\boldsymbol{x}^T\boldsymbol{u}\right\}\\&=\min_{\boldsymbol{u}}\frac{1}{2}\left(1+\frac{\eta}{2\tau}\right)^{-1}\left\{\frac{1}{2}\|\boldsymbol{u}-\boldsymbol{x}\|_{\ell_2}^2+2\lambda\|\boldsymbol{u}\|_{\ell_\infty}-\frac{\|\boldsymbol{x}\|_{\ell_2}^2}{2}\right\}\\&=\frac{1}{2}\left(1+\frac{\eta}{2\tau}\right)^{-1}J_\infty(\boldsymbol{x};2\lambda)-\frac{1}{4}\left(1+\frac{\eta}{2\tau}\right)^{-1}\|\boldsymbol{x}\|_{\ell_2}^2\end{aligned} \tag{D.31}$$

Using (D.31) and substituting for $\boldsymbol{x}$ in (D.27), our AO problem becomes

$$\min_{\theta,\gamma\geq 0,\alpha\geq 0}\ \max_{\beta,\lambda,\eta\geq 0,\widetilde{\eta}}\ \min_{\tau\geq 0}\ \frac{1}{2}\left(1+\frac{\eta}{2\tau}\right)^{-1} J_\infty\left(\frac{\beta}{\sqrt{n}}\boldsymbol{h}-\tilde{\eta}\widetilde{\boldsymbol{\mu}};2\lambda\right)-\frac{1}{4}\left(1+\frac{\eta}{2\tau}\right)^{-1}\left\|\frac{\beta}{\sqrt{n}}\boldsymbol{h}-\tilde{\eta}\widetilde{\boldsymbol{\mu}}\right\|_{\ell_2}^2$$

$$-2\lambda\gamma+\beta\sqrt{\mathbb{E}\left[\left(\left(1+\varepsilon\gamma-\theta\left\|\boldsymbol{\mu}\right\|_{\ell_2}\right)+\alpha g\right)_+^2\right]}+\frac{\eta\tau}{2}-\eta\alpha-\widetilde{\eta}\theta \tag{D.32}$$

Our next step is to scalarize the AO problem by taking the asymptotic limit of the objective.

We have

$$\lim_{n\to\infty}\left\|\frac{\beta}{\sqrt{n}}\boldsymbol{h}-\tilde{\eta}\widetilde{\boldsymbol{\mu}}\right\|_{\ell_2}^2=\frac{\beta^2}{\delta}+\tilde{\eta}^2\,,\quad \lim_{n\to\infty}\left\|\boldsymbol{\mu}\right\|_{\ell_2}=\sigma_{M,2}\,.$$

in probability. Also by using (D.20) we have

$$\lim_{n\to\infty} J_\infty\left(\frac{\beta}{\sqrt{n}}\boldsymbol{h}-\tilde{\eta}\widetilde{\boldsymbol{\mu}};2\lambda\right)=\min_{t_0\geq 0}\left\{f(\beta,\tilde{\eta};t_0)+2\lambda_0 t_0\right\}\,,$$

with $\lambda_0=\frac{\lambda}{\left\|\boldsymbol{\mu}\right\|_{\ell_1}}$. Using these limits in the AO problem (D.32) we obtain the following scalar AO problem

$$\min_{\theta,\gamma\geq 0,\alpha\geq 0}\ \max_{\beta,\lambda,\eta\geq 0,\widetilde{\eta}}\ \min_{\tau,t_0\geq 0}\ \frac{1}{2}\left(1+\frac{\eta}{2\tau}\right)^{-1}\left(f(\beta,\tilde{\eta};t_0)+2\lambda_0 t_0\right)-\frac{1}{4}\left(1+\frac{\eta}{2\tau}\right)^{-1}\left(\frac{\beta^2}{\delta}+\tilde{\eta}^2\right)$$

$$-2\lambda_0\gamma_0+\beta\sqrt{\mathbb{E}\left[\left(\left(1+\varepsilon_0\gamma_0-\theta\sigma_{M,2}\right)+\alpha g\right)_+^2\right]}+\frac{\eta\tau}{2}-\eta\alpha-\widetilde{\eta}\theta\,, \tag{D.33}$$

where we recall our notation $\varepsilon_0=\frac{\varepsilon}{\left\|\boldsymbol{\mu}\right\|_{\ell_1}}$ and $\gamma_0=\gamma\left\|\boldsymbol{\mu}\right\|_{\ell_1}$.

Now note that objective function (D.33) is linear in $\lambda_0$ and therefore the optimal $t_0^*$ should satisfy $t_0^*\leq 2\gamma_0\left(1+\frac{\eta}{2\tau}\right)$, otherwise $\lambda_0^*=\infty$ which makes the above max-min value unbounded. As such, we also have $\lambda_0^*=0$ which further simplifies the problem as follows:

$$\min_{\theta,\gamma\geq 0,\alpha\geq 0}\ \max_{\beta,\eta\geq 0,\widetilde{\eta}}\ \min_{0\leq\tau,0\leq t_0\leq 2\gamma_0(1+\frac{\eta}{2\tau})}\ \frac{1}{2}\left(1+\frac{\eta}{2\tau}\right)^{-1} f(\beta,\tilde{\eta};t_0)-\frac{1}{4}\left(1+\frac{\eta}{2\tau}\right)^{-1}\left(\frac{\beta^2}{\delta}+\tilde{\eta}^2\right)$$

$$+\beta\sqrt{\mathbb{E}\left[\left(\left(1+\varepsilon_0\gamma_0-\theta\sigma_{M,2}\right)+\alpha g\right)_+^2\right]}+\frac{\eta\tau}{2}-\eta\alpha-\widetilde{\eta}\theta\,, \tag{D.34}$$

Since $f(c_0, c_1; t_0)$ is decreasing in $t_0$, the optimal value of $t_0$ is given by $t_0^\star = 2\gamma_0(1+\frac{\eta}{2\tau})$ which results in the following AO problem:

$$\min_{\theta,\gamma\geq 0,\alpha\geq 0}\ \max_{\beta,\eta\geq 0,\widetilde{\eta}}\ \min_{\tau\geq 0}\ \frac{1}{2}\left(1+\frac{\eta}{2\tau}\right)^{-1} f\left(\beta,\tilde{\eta};2\gamma_0(1+\frac{\eta}{2\tau})\right)-\frac{1}{4}\left(1+\frac{\eta}{2\tau}\right)^{-1}\left(\frac{\beta^2}{\delta}+\tilde{\eta}^2\right)$$
$$+\beta\sqrt{\mathbb{E}\left[\left((1+\varepsilon_0\gamma_0-\theta\sigma_{M,2})+\alpha g\right)_+^2\right]}+\frac{\eta\tau}{2}-\eta\alpha-\widetilde{\eta}\theta\,. \tag{D.35}$$

Our final step of simplification is to solve for $\tau$. To this end, we define the function

$$R\left(\frac{\eta}{\tau}\right) := \frac{1}{2}\left(1+\frac{\eta}{2\tau}\right)^{-1} f\left(\beta,\tilde{\eta};2\gamma_0(1+\frac{\eta}{2\tau})\right)-\frac{1}{4}\left(1+\frac{\eta}{2\tau}\right)^{-1}\left(\frac{\beta^2}{\delta}+\tilde{\eta}^2\right),$$

where we make the dependence on $\frac{\eta}{\tau}$ explicit in the notation. Setting derivative of the AO objective with respect to $\eta$, to zero we obtain

$$\frac{1}{\tau}R'\left(\frac{\eta}{\tau}\right)+\frac{\tau}{2}-\alpha=0\,.$$

Setting derivative with respect to $\tau$ to zero gives

$$-\frac{\eta}{\tau^2}R'\left(\frac{\eta}{\tau}\right)+\frac{\eta}{2}=0\,.$$

Combining the above two optimality condition implies that $\eta(1-\frac{\alpha}{\tau})=0$. So either $\alpha=\tau$ or $\eta=0$. If $\eta=0$, then it is clear that the terms involving $\tau$ in the AO problem would vanish and therefore the value of $\tau$ does not matter. So in this case, we can as well assume $\tau=\alpha$. Substituting for $\tau$ the AO problem further simplifies to

$$\min_{\theta,\gamma\geq 0,\alpha\geq 0}\ \max_{\beta,\eta\geq 0,\widetilde{\eta}}\ \frac{1}{2}\left(1+\frac{\eta}{2\alpha}\right)^{-1} f\left(\beta,\tilde{\eta};2\gamma_0(1+\frac{\eta}{2\alpha})\right)-\frac{1}{4}\left(1+\frac{\eta}{2\alpha}\right)^{-1}\left(\frac{\beta^2}{\delta}+\tilde{\eta}^2\right)$$
$$+\beta\sqrt{\mathbb{E}\left[\left((1+\varepsilon_0\gamma_0-\theta\sigma_{M,2})+\alpha g\right)_+^2\right]}-\frac{\eta\tau}{2}-\widetilde{\eta}\theta\,. \tag{D.36}$$

This concludes the proof of part (b). ∎

**Proof** [Part (c)] The proof of part (c) follows along the same lines of the proof of Theorem 6.4 (and Theorem 4.5 for isotropic case). But similar to previous parts, we need to make slight modifications to the proof.

Note that in our derivation of the AO problem (6.9), we replaced the constraint $\|\boldsymbol{u}\|_{\ell_q} \le \frac{\tau_h}{\alpha}\frac{\gamma_0}{\|\boldsymbol{\mu}\|_{\ell_p}}$ with the equivalent constraint $\|\boldsymbol{u}\|_{\ell_q}^q \le (\frac{\tau_h}{\alpha}\frac{\gamma_0}{\|\boldsymbol{\mu}\|_{\ell_p}})^q$, see (E.11) for more details. The benefit of this alternative representation is that it results in the Moreau-envelope $e_{q,\boldsymbol{\Sigma}}$ of the $\|\cdot\|_{\ell_q}^q$ function, see (E.15), which is separable over the samples. As a result, in the isotropic case the expected Moreau envelope reduces to the expected of the one-dimensional function $J_q$, given by (4.5), (C.9).

However, for $q = \infty$ the function $\|\cdot\|_{\ell_q}^q$ is not well-defined and requires a slightly different treatment. In this case we stay with the original constraint $\|\boldsymbol{u}\|_{\ell_q} \le \frac{\tau_h}{\alpha}\frac{\gamma_0}{\|\boldsymbol{\mu}\|_{\ell_p}}$. Proceeding along the same derivations of AO problem (4.16), it is straightforward to see that this results in the following AO problem for the non-separable regime:

$$\max_{0\le\beta,\tau_h} \min_{\theta,0\le\alpha,\gamma_0,\tau_g} D_{\text{ns}}(\alpha,\gamma_0,\theta,\tau_g,\beta,\tau_h)$$

$$D_{\text{ns}}(\alpha,\gamma_0,\theta,\tau_g,\beta,\tau_h) = \frac{\beta\tau_g}{2} + L\left(\sqrt{\alpha^2+\theta^2}, \sigma_{M,2}\theta - \varepsilon_0\gamma_0, \frac{\tau_g}{\beta}\right)$$

$$- \min_{\lambda_0\ge 0,\nu}\left[\frac{\alpha}{\tau_h}\left\{\frac{\beta^2}{2\delta} + \lambda_0\frac{\gamma_0\tau_h}{\alpha} + \frac{\nu^2}{2} - \tilde{\mathsf{E}}\left(\beta, \frac{\tau_h\theta}{\alpha} + \nu; \lambda_0\right)\right\} + \frac{\alpha\tau_h}{2}\right] \tag{D.37}$$

with

$$\tilde{\mathsf{E}}(c_0,c_1;\lambda_0) := \lim_{n\to\infty} J_\infty\left(\frac{c_0}{\sqrt{n}}\boldsymbol{h} - c_1\widetilde{\boldsymbol{\mu}}; \lambda_0\|\boldsymbol{\mu}\|_{\ell_1}\right), \tag{D.38}$$

and $J_\infty(\boldsymbol{x};\lambda)$ given by (D.17). Using (D.20), we have

$$\tilde{\mathsf{E}}(c_0,c_1;\lambda_0) = \min_{t_0\ge 0}\{f(c_0,c_1;t_0) + \lambda_0 t_0\}, \tag{D.39}$$

Substituting for $\tilde{\mathsf{E}}$ function in the AO problem (D.37) results in

$$\max_{0\le\beta,\tau_h} \min_{\theta,0\le\alpha,\gamma_0,\tau_g} D_{\text{ns}}(\alpha,\gamma_0,\theta,\tau_g,\beta,\tau_h)$$

$$D_{\text{ns}}(\alpha,\gamma_0,\theta,\tau_g,\beta,\tau_h) = \frac{\beta\tau_g}{2} + L\left(\sqrt{\alpha^2+\theta^2}, \sigma_{M,2}\theta - \varepsilon_0\gamma_0, \frac{\tau_g}{\beta}\right)$$

$$- \min_{\lambda_0\ge 0,\nu}\sup_{t_0\ge 0}\left[\frac{\alpha}{\tau_h}\left\{\frac{\beta^2}{2\delta} + \lambda_0\left(\frac{\gamma_0\tau_h}{\alpha} - t_0\right) + \frac{\nu^2}{2} - f\left(\beta, \frac{\tau_h\theta}{\alpha} + \nu; t_0\right)\right\} + \frac{\alpha\tau_h}{2}\right] \tag{D.40}$$

Note that the above objective is linear in $\lambda_0$. Clearly, the optimal value $t_0^*$ should satisfy $t_0^* \le \frac{\gamma_0\tau_h}{\alpha}$; otherwise $\lambda_0^* = \infty$ which makes the objective value

unbounded. For $t_0 \le \frac{\gamma_0\tau_h}{\alpha}$, we have $\lambda_0^* = 0$. Therefore, $t_0$ only appears in the term $f(\beta, \frac{\tau_h\theta}{\alpha} + \nu; t_0)$. Given that $f(c_0, c_1; t_0)$ is decreasing in $t_0$, we have $t_0^* = \frac{\gamma_0\tau_h}{\alpha}$.

We substitute for $t_0^*$ in the AO problem to obtain

$$\max_{0\le\beta,\tau_h} \min_{\theta,0\le\alpha,\gamma_0,\tau_g} D_{\rm ns}(\alpha,\gamma_0,\theta,\tau_g,\beta,\tau_h)$$

$$\begin{aligned} D_{\rm ns}(\alpha,\gamma_0,\theta,\tau_g,\beta,\tau_h) &= \frac{\beta\tau_g}{2} + L\left(\sqrt{\alpha^2+\theta^2}, \sigma_{M,2}\theta - \varepsilon_0\gamma_0, \frac{\tau_g}{\beta}\right) \\ &\quad - \min_{\nu}\left[\frac{\alpha}{\tau_h}\left\{\frac{\beta^2}{2\delta} + \frac{\nu^2}{2} - f\left(\beta, \frac{\tau_h\theta}{\alpha} + \nu; \frac{\gamma_0\tau_h}{\alpha}\right)\right\} + \frac{\alpha\tau_h}{2}\right]. \end{aligned} \tag{D.41}$$

This completes the proof of part (c). ∎

## APPENDIX E: PROOF OF TECHNICAL LEMMAS

**E.1. Proof of Lemma 2.1** By definition, we have

$$\begin{aligned} \mathsf{SA}(\widehat{\boldsymbol{\theta}}) &:= \mathbb{E}[\mathbb{1}(\hat{y} = y)] = \mathbb{P}(y\langle \boldsymbol{x}, \widehat{\boldsymbol{\theta}}\rangle > 0) \\ &= \mathbb{P}\left(y\langle y\boldsymbol{\mu} + \Sigma^{1/2}\boldsymbol{z}, \widehat{\boldsymbol{\theta}}\rangle > 0\right) \\ &= \mathbb{P}\left(\langle \boldsymbol{\mu} + \Sigma^{1/2}\boldsymbol{z}, \widehat{\boldsymbol{\theta}}\rangle > 0\right) \\ &= \mathbb{P}\left(\langle \boldsymbol{\mu}, \widehat{\boldsymbol{\theta}}\rangle + \left\|\Sigma^{1/2}\widehat{\boldsymbol{\theta}}\right\|_{\ell_2} Z > 0\right) \\ &= \Phi\left(\frac{\langle \boldsymbol{\mu}, \widehat{\boldsymbol{\theta}}\rangle}{\left\|\boldsymbol{\Sigma}^{1/2}\widehat{\boldsymbol{\theta}}\right\|_{\ell_2}}\right), \end{aligned} \tag{E.1}$$

where $\boldsymbol{z} \sim \mathsf{N}(0, \boldsymbol{I}_d)$ and $Z \sim \mathsf{N}(0,1)$.

Likewise for the adversarial risk we have

$$\begin{aligned}
\mathsf{RA}(\widehat{\boldsymbol{\theta}}) &:= \mathbb{E}\left[\min_{\|\boldsymbol{\delta}\|_{\ell_p}\le\varepsilon} \mathbb{1}(y\langle \boldsymbol{x}+\boldsymbol{\delta},\widehat{\boldsymbol{\theta}}\rangle \ge 0)\right] \\
&\overset{(a)}{=} \mathbb{E}\left[\mathbb{1}(y\langle \boldsymbol{x},\widehat{\boldsymbol{\theta}}\rangle - \varepsilon\left\|\widehat{\boldsymbol{\theta}}\right\|_{\ell_q} \ge 0)\right] \\
&= \mathbb{P}\left(y\langle \boldsymbol{x},\widehat{\boldsymbol{\theta}}\rangle - \varepsilon\left\|\widehat{\boldsymbol{\theta}}\right\|_{\ell_q} \ge 0\right) \\
&= \mathbb{P}\left(y\langle y\boldsymbol{\mu}+\boldsymbol{\Sigma}^{1/2}\boldsymbol{z},\widehat{\boldsymbol{\theta}}\rangle - \varepsilon\left\|\widehat{\boldsymbol{\theta}}\right\|_{\ell_q} \ge 0\right) \\
&\overset{(b)}{=} \mathbb{P}\left(\langle \boldsymbol{\mu},\widehat{\boldsymbol{\theta}}\rangle + \left\|\boldsymbol{\Sigma}^{1/2}\widehat{\boldsymbol{\theta}}\right\|_{\ell_2} Z - \varepsilon\left\|\widehat{\boldsymbol{\theta}}\right\|_{\ell_q} \ge 0\right) \\
&= \Phi\left(\frac{\langle \boldsymbol{\mu},\widehat{\boldsymbol{\theta}}\rangle - \varepsilon\left\|\widehat{\boldsymbol{\theta}}\right\|_{\ell_q}}{\left\|\boldsymbol{\Sigma}^{1/2}\widehat{\boldsymbol{\theta}}\right\|_{\ell_2}}\right), \qquad \text{(E.2)}
\end{aligned}$$

where $(a)$ we used that $\langle \delta,\widehat{\boldsymbol{\theta}}\rangle \ge -\|\boldsymbol{\delta}\|_{\ell_p}\left\|\widehat{\boldsymbol{\theta}}\right\|_{\ell_q} \ge -\varepsilon\left\|\widehat{\boldsymbol{\theta}}\right\|_{\ell_q}$, using Hölder inequality (with $\frac{1}{p}+\frac{1}{q}=1$) and that $\|\boldsymbol{\delta}\|_{\ell_p}\le\varepsilon$, with equality achieving for some $\boldsymbol{\delta}$ in this set. In $(b)$, we used the symmetry of Gaussian distribution.

**E.2. Proof of Lemma 4.2** We first note that

$$\min_{\boldsymbol{z}\in\mathcal{S}(\alpha,\theta,\varepsilon_0,\boldsymbol{\mu})} -\frac{1}{\sqrt{n}}\boldsymbol{h}^T\boldsymbol{z} = -\frac{1}{\sqrt{n/d}}\sup_{\boldsymbol{z}\in\mathcal{S}(\alpha,\theta,\varepsilon_0,\boldsymbol{\mu})} \frac{1}{\sqrt{d}}\boldsymbol{h}^T\boldsymbol{z} \to -\frac{1}{\sqrt{\delta}}\omega(\alpha,\theta,\varepsilon_0)\,, \qquad \text{(E.3)}$$

in probability, using the fact that $\boldsymbol{h}/\sqrt{d}$ is asymptotically uniform on the unit sphere, and for $\mathcal{S}\in\mathbb{S}^{d-1}$ the function $f(\boldsymbol{u}) = \sup_{\boldsymbol{z}\in\mathcal{S}}\boldsymbol{z}^T\boldsymbol{u}$ is Lipschitz. Therefore, using the concentration of Lipschitz functions on the sphere (see e.g. [66, Theorem 5.2.2]), $f(\boldsymbol{u})$ concentrates around its mean $\mathbb{E}\,f(\boldsymbol{u}) = \omega_s(\mathcal{S}(\alpha,\theta,\varepsilon_0,\boldsymbol{\mu}))$. More precisely,

$$\mathbb{P}\left\{\left|\sup_{\boldsymbol{z}\in\mathcal{S}}\frac{1}{\sqrt{d}}\boldsymbol{h}^T\boldsymbol{z} - \omega_s(\mathcal{S}(\alpha,\theta,\varepsilon_0,\boldsymbol{\mu}))\right|\right\} \le 2e^{-cdt^2}\,,$$

for an absolute constant $c>0$ and for every $t\ge 0$. Therefore, by invoking the assumption on the convergence of spherical width, cf. Assumption 3, we arrive at

$$\lim_{d\to\infty}\mathbb{P}\left\{\left|\sup_{\boldsymbol{z}\in\mathcal{S}}\frac{1}{\sqrt{d}}\boldsymbol{h}^T\boldsymbol{z} - \omega(\alpha,\theta,\varepsilon_0)\right| \ge \eta\right\} = 0\,, \quad \forall\eta>0\,.$$

Therefore, $\sup_{\boldsymbol{z}\in\mathcal{S}(\alpha,\theta,\varepsilon_0)} \ \frac{1}{\sqrt{d}}\boldsymbol{h}^T\boldsymbol{z} \to \omega(\alpha,\theta,\varepsilon_0)$, in probability.

To evaluate the left hand side, we form the Lagrangian corresponding to the set $\mathcal{S}$. Let $\widetilde{\boldsymbol{\mu}} \coloneqq \frac{\boldsymbol{\mu}}{\|\boldsymbol{\mu}\|_{\ell_2}}$ and consider the change of variable $\boldsymbol{u} \coloneqq \boldsymbol{z} + \theta\widetilde{\boldsymbol{\mu}}$. We then have

$$
\begin{aligned}
&\min_{\boldsymbol{z}\in\mathcal{S}(\alpha,\theta,\varepsilon_0,\boldsymbol{\mu})} \ -\frac{1}{\sqrt{n}}\boldsymbol{h}^T\boldsymbol{z} \\
&= \sup_{\lambda,\eta\geq 0,\nu} \min_{\boldsymbol{u}} \ -\frac{1}{\sqrt{n}}\boldsymbol{h}^T(\boldsymbol{u}-\theta\widetilde{\boldsymbol{\mu}}) + \lambda\left(\|\boldsymbol{u}\|_{\ell_q}^q - \left(\frac{1}{\varepsilon_0\|\boldsymbol{\mu}\|_{\ell_p}}\right)^q\right) + \frac{\eta}{2}\left(\|\boldsymbol{u}-\theta\widetilde{\boldsymbol{\mu}}\|_{\ell_2}^2 - \alpha^2\right) + \nu\widetilde{\boldsymbol{\mu}}^T(\boldsymbol{u}-\theta\widetilde{\boldsymbol{\mu}}) \\
&= \sup_{\lambda,\eta\geq 0,\nu} \min_{\boldsymbol{u}} \ \frac{\eta}{2}\left\|\boldsymbol{u}-\theta\widetilde{\boldsymbol{\mu}}+\frac{\nu}{\eta}\widetilde{\boldsymbol{\mu}}-\frac{\boldsymbol{h}}{\eta\sqrt{n}}\right\|_{\ell_2}^2 + \lambda\|\boldsymbol{u}\|_{\ell_q}^q - \frac{1}{2\eta}\left\|\nu\widetilde{\boldsymbol{\mu}}-\frac{\boldsymbol{h}}{\sqrt{n}}\right\|_{\ell_2}^2 - \frac{\eta}{2}\alpha^2 - \frac{\lambda}{(\varepsilon_0\|\boldsymbol{\mu}\|_{\ell_p})^q} \\
&= \sup_{\lambda,\eta\geq 0,\nu} \min_{\boldsymbol{u}} \ \eta\left[\frac{1}{2}\left\|\boldsymbol{u}+\left(\frac{\nu}{\eta}-\theta\right)\widetilde{\boldsymbol{\mu}}-\frac{\boldsymbol{h}}{\eta\sqrt{n}}\right\|_{\ell_2}^2 + \frac{\lambda}{\eta}\|\boldsymbol{u}\|_{\ell_q}^q\right] - \frac{1}{2\eta}\left\|\nu\widetilde{\boldsymbol{\mu}}-\frac{\boldsymbol{h}}{\sqrt{n}}\right\|_{\ell_2}^2 - \frac{\eta}{2}\alpha^2 - \frac{\lambda}{(\varepsilon_0\|\boldsymbol{\mu}\|_{\ell_p})^q} \\
&= \sup_{\lambda,\eta\geq 0,\nu} \ \eta\sum_{i=1}^d J_q\left(\frac{h_i}{\eta\sqrt{n}} - \left(\frac{\nu}{\eta}-\theta\right)\tilde{\mu}_i;\frac{\lambda}{\eta}\right) - \frac{1}{2\eta}\left\|\nu\widetilde{\boldsymbol{\mu}}-\frac{\boldsymbol{h}}{\sqrt{n}}\right\|_{\ell_2}^2 - \frac{\eta}{2}\alpha^2 - \frac{\lambda}{(\varepsilon_0\|\boldsymbol{\mu}\|_{\ell_p})^q} \qquad \text{(E.4)} \\
&= \sup_{\lambda,\eta\geq 0,\nu} \ \eta\sum_{i=1}^d J_q\left(\frac{h_i}{\eta\sqrt{n}} - \left(\frac{\nu}{\eta}-\theta\right)\tilde{\mu}_i;\frac{\lambda}{\eta}\right) - \frac{\nu^2}{2\eta} - \frac{1}{2\eta n}\|\boldsymbol{h}\|_{\ell_2}^2 + \frac{\nu}{\eta\sqrt{n}}\widetilde{\boldsymbol{\mu}}^T\boldsymbol{h} - \frac{\eta}{2}\alpha^2 - \frac{\lambda}{(\varepsilon_0\|\boldsymbol{\mu}\|_{\ell_p})^q}
\end{aligned}
$$

Recall that $\boldsymbol{h}\sim\mathsf{N}(0,\boldsymbol{I}_d)$. As $n\to\infty$ and $n/d\to\delta$, we have

$$
\frac{1}{n}\|\boldsymbol{h}\|_{\ell_2}^2 \to \frac{1}{\delta}, \quad \frac{1}{\sqrt{n}}\widetilde{\boldsymbol{\mu}}^T\boldsymbol{h}\to 0\,, \qquad \text{(E.5)}
$$

in probability. In addition,

$$
\|\boldsymbol{\mu}\|_{\ell_p} \to \sigma_{M,p}d^{\frac{1}{p}-\frac{1}{2}} = \sigma_{M,p}d^{\frac{1}{2}-\frac{1}{q}}\,, \qquad \text{(E.6)}
$$

in probability. Using the identity $J_q(x;\lambda) = c^2 J_q(x/c;\lambda c^{q-2})$ and letting $\lambda_0 \coloneqq \lambda d^{1-\frac{q}{2}}$ we have

$$
\begin{aligned}
\sum_{i=1}^d J_q\left(\frac{h_i}{\eta\sqrt{n}} - \left(\frac{\nu}{\eta}-\theta\right)\tilde{\mu}_i;\frac{\lambda}{\eta}\right) &= \frac{1}{d}\sum_{i=1}^d J_q\left(\frac{\sqrt{d}h_i}{\eta\sqrt{n}} - \left(\frac{\nu}{\eta}-\theta\right)\frac{\sqrt{d}\mu_i}{\sigma_{M,2}};\frac{\lambda}{\eta}d^{1-q/2}\right) \\
&= \frac{1}{d}\sum_{i=1}^d J_q\left(\frac{h_i}{\eta\sqrt{\delta}} - \left(\frac{\nu}{\eta}-\theta\right)\frac{\sqrt{d}\mu_i}{\sigma_{M,2}};\frac{\lambda_0}{\eta}\right).
\end{aligned}
$$

Since $J_q(x;\lambda)\leq\frac{1}{2}x^2$, the function $J_q$ is pseudo-lipschitz of order 2 and by an

application of [3, Lemma 5], we have

(E.7)
$$\lim_{n\to\infty}\frac{1}{d}\sum_{i=1}^{d} J_q\left(\frac{h_i}{\eta\sqrt{\delta}}-\left(\frac{\nu}{\eta}-\theta\right)\frac{\sqrt{d}\mu_i}{\sigma_{M,2}};\frac{\lambda_0}{\eta}\right)=\mathbb{E}\left[J_q\left(\frac{h}{\eta\sqrt{\delta}}-\left(\frac{\nu}{\eta}-\theta\right)\frac{M}{\sigma_{M,2}};\frac{\lambda_0}{\eta}\right)\right],$$

almost surely, where the expectation in the last line is taken with respect to the independent random variables $h\sim \mathsf{N}(0,1)$ and $M\sim\mathbb{P}_M$.

Using the above limits, we see that the objective function (E.4) converges pointwise to the following function:

$$\eta\,\mathbb{E}\left[J_q\left(\frac{h}{\eta\sqrt{\delta}}-\left(\frac{\nu}{\eta}-\theta\right)\frac{M}{\sigma_{M,2}};\frac{\lambda_0}{\eta}\right)\right]-\frac{\nu^2}{2\eta}-\frac{1}{2\eta\delta}-\frac{\eta}{2}\alpha^2-\lambda_0(\varepsilon_0\sigma_{M,p})^{-q} \tag{E.8}$$

Note that (E.4) is the dual optimization and hence is a concave problem. We apply the convexity lemma [63, Lemma B.2] to conclude that the objective value in (E.4) also converges to the supremum of function (E.8) over $\lambda_0,\eta\geq 0,\nu$.

Using this observation along with Equation (E.3) and (E.4) we obtain

$$\begin{aligned}
&\omega(\alpha,\theta,\varepsilon_0)\\
&=\lim_{n\to\infty}\omega_s\left(\mathcal{S}(\alpha,\theta,\varepsilon_0,\boldsymbol{\mu})\right)\\
&=-\sup_{\lambda_0,\eta\geq 0,\nu}\ \eta\sqrt{\delta}\,\mathbb{E}\left[J_q\left(\frac{h}{\eta\sqrt{\delta}}-\left(\frac{\nu}{\eta}-\theta\right)\frac{M}{\sigma_{M,2}};\frac{\lambda_0}{\eta}\right)\right]-\sqrt{\delta}\left\{\frac{\nu^2}{2\eta}+\frac{1}{2\eta\delta}+\frac{\eta}{2}\alpha^2+\lambda_0(\varepsilon_0\sigma_{M,p})^{-q}\right\}\\
&=\min_{\lambda_0,\eta\geq 0,\nu}\ \sqrt{\delta}\left\{\frac{\nu^2}{2\eta}+\frac{1}{2\eta\delta}+\frac{\eta}{2}\alpha^2+\lambda_0(\varepsilon_0\sigma_{M,p})^{-q}\right\}-\eta\sqrt{\delta}\,\mathbb{E}\left[J_q\left(\frac{h}{\eta\sqrt{\delta}}-\left(\frac{\nu}{\eta}-\theta\right)\frac{M}{\sigma_{M,2}};\frac{\lambda_0}{\eta}\right)\right],
\end{aligned} \tag{E.9}$$

which completes the proof.

**E.3. 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.

E.3.1. *Bounded domains in optimization* (B.17) We start with the restriction on $\boldsymbol{\theta}$. Note that one of the claims of Theorem 6.3, part (b), is to show that $\|\widetilde{\boldsymbol{\theta}}^{\varepsilon}\|_{\ell_2}\to\alpha_*$ as $n\to\infty$, in probability, for some $\alpha_*$ by the solution of minimax problem (6.2). We define $\mathcal{S}_{\boldsymbol{\theta}}=\{\boldsymbol{\theta}:\ \|\boldsymbol{\theta}\|_{\ell_2}\leq K_\alpha\}$ with $K_\alpha=\alpha_*+\xi$

for a constant $\xi > 0$. We start by the ansatz that $\boldsymbol{\theta} \in \mathcal{S}_{\boldsymbol{\theta}}$ and add this 'artificial constraint' in the minimax optimization. In addition, by the stationary condition for $\gamma$ in optimization (B.17) we have

$$\frac{1}{n}\varepsilon \boldsymbol{u}^T \mathbf{1} = 2\lambda\,.$$

Therefore

$$\frac{1}{n}\boldsymbol{u}^T\mathbf{1} = \frac{2\lambda}{\varepsilon} = \frac{2\lambda}{\varepsilon_0 \left\|\boldsymbol{\mu}\right\|_{\ell_p}}\,.$$

Let $\lambda_0 := \frac{\lambda}{\|\boldsymbol{\mu}\|_{\ell_p}}$. Assuming the ansatz that $\lambda_0 = O(1)$, we also use this 'artificial constraint' in the minimax optimization.

With these compact constraints in place, we then deploy the CGMT framework to prove Theorem 6.3. This theorem implies that $\left\|\widetilde{\boldsymbol{\theta}}^{\varepsilon}\right\|_{\ell_2} \to \alpha_*$ as $n \to \infty$ and so our initial ansatz on the boundedness of $\boldsymbol{\theta}$ is verified. Further, as it can be seen from the proof of Theorem 6.3 (see the line following Equation (B.28)), we have $\lambda_0 = \frac{\lambda}{\|\boldsymbol{\mu}\|_{\ell_p}} \to \lambda_{0*}$ as $n \to \infty$, in probability, for some $\lambda_{0*}$ that is determined by the solution of minimax problem (6.2). This also verifies our ansatz that $\lambda_0 = O(1)$, which in turn implies that $\boldsymbol{u} \in \mathcal{S}_{\boldsymbol{u}} = \{\boldsymbol{u} : \ 0 \le u_i,\ \frac{1}{n}\mathbf{1}^T\boldsymbol{u} \le K_{\boldsymbol{u}}\}$ for some sufficiently large constant $K_{\boldsymbol{u}} > 0$.

E.3.2. *Bounded domains in optimization* (B.38) Similar to previous subsection, we start by the ansatz that $\boldsymbol{\theta} \in \mathcal{S}_{\boldsymbol{\theta}}$ where $\mathcal{S}_{\boldsymbol{\theta}} = \{\boldsymbol{\theta} : \ \|\boldsymbol{\theta}\|_{\ell_2} \le K_\alpha\}$ with $K_\alpha = \alpha_* + \xi$ for a constant $\xi > 0$, and add this 'artificial constraint' in the minimax optimization (B.38). Also by stationarity condition for $\boldsymbol{v}$ in (B.38) we have $u_i = \ell'\left(v_i - \varepsilon \left\|\boldsymbol{\theta}\right\|_{\ell_q}\right)$ and hence $|u_i| \le K_{\boldsymbol{u}}$ for some large enough constant $K_{\boldsymbol{u}} > 0$, using our assumption on the loss function $\ell$.

**E.4. Proof of Lemma B.1** First note that by Cauchy–Schwarz inequality we have

$$\langle \mathsf{P}_{\boldsymbol{\mu}}^{\perp}\boldsymbol{r}, \left(\mathsf{P}_{\boldsymbol{\mu}}^{\perp}\boldsymbol{\Sigma}^{1/2}\mathsf{P}_{\boldsymbol{\mu}}^{\perp}\right)\boldsymbol{\theta}_{\perp}\rangle \ge -\left\|\mathsf{P}_{\boldsymbol{\mu}}^{\perp}\boldsymbol{r}\right\|_{\ell_2}\left\|\left(\mathsf{P}_{\boldsymbol{\mu}}^{\perp}\boldsymbol{\Sigma}^{1/2}\mathsf{P}_{\boldsymbol{\mu}}^{\perp}\right)\boldsymbol{\theta}_{\perp}\right\|_{\ell_2} = -\alpha\left\|\mathsf{P}_{\boldsymbol{\mu}}^{\perp}\boldsymbol{r}\right\|_{\ell_2}\,.$$

To achieve equality, note that similar to (B.50) we have

$$\left(\mathsf{P}_{\boldsymbol{\mu}}^{\perp}\boldsymbol{\Sigma}^{1/2}\mathsf{P}_{\boldsymbol{\mu}}^{\perp}\right)\left(\mathsf{P}_{\boldsymbol{\mu}}^{\perp}\boldsymbol{\Sigma}^{-1/2}\mathsf{P}_{\boldsymbol{\mu}}^{\perp}\right)\boldsymbol{r} = \mathsf{P}_{\boldsymbol{\mu}}^{\perp}\boldsymbol{r}\,.$$

Therefore equality is achieved by choosing $\boldsymbol{\theta} = \lambda\boldsymbol{\Sigma}^{-1/2}\mathsf{P}_{\boldsymbol{\mu}}^{\perp}\boldsymbol{r}$ with $\lambda = \frac{\alpha}{\left\|\mathsf{P}_{\boldsymbol{\mu}}^{\perp}\boldsymbol{r}\right\|_{\ell_2}}$.

**E.5. Proof of Lemma B.2** By definition of the conjugate function we have

$$\begin{aligned}\widetilde{\ell}(\boldsymbol{v},\boldsymbol{w}) &= \sup_{\boldsymbol{\theta}} \ \boldsymbol{w}^T\boldsymbol{\theta} - \ell(\boldsymbol{v},\boldsymbol{\theta}) \\ &= \sup_{\boldsymbol{\theta}} \ \boldsymbol{w}^T\boldsymbol{\theta} - \frac{1}{n}\sum_{i=1}^{n}\ell\left(v_i - \varepsilon\,\|\boldsymbol{\theta}\|_{\ell_q}\right)\end{aligned}$$

Now assume $\boldsymbol{\theta} = \gamma\boldsymbol{u}$ with $\|\boldsymbol{u}\|_{\ell_q} = 1$. We thus have,

$$\begin{aligned}\widetilde{\ell}(\boldsymbol{v},\boldsymbol{w}) &= \sup_{\boldsymbol{u}:\|\boldsymbol{u}\|_{\ell_q}=1,\gamma} \ \gamma\boldsymbol{w}^T\boldsymbol{u} - \frac{1}{n}\sum_{i=1}^{n}\ell\left(v_i - \varepsilon\gamma\right) \\ &= \sup_{\gamma\geq 0} \ \gamma\left(\sup_{\boldsymbol{u}:\|\boldsymbol{u}\|_{\ell_q}=1} \boldsymbol{w}^T\boldsymbol{u}\right) - \frac{1}{n}\sum_{i=1}^{n}\ell\left(v_i - \varepsilon\gamma\right) \\ &= \sup_{\gamma\geq 0} \ \gamma\,\|\boldsymbol{w}\|_{\ell_p} - \frac{1}{n}\sum_{i=1}^{n}\ell\left(v_i - \varepsilon\gamma\right)\ .\end{aligned}$$

**E.6. Proof of Lemma B.3** By definition

$$\begin{aligned}f^*(\boldsymbol{u}) &:= \sup_{\tilde{\boldsymbol{w}}} \ \langle\boldsymbol{u},\tilde{\boldsymbol{w}}\rangle - f(\tilde{\boldsymbol{w}}) \\ &= \sup_{\tilde{\boldsymbol{w}}} \ \langle\boldsymbol{u},\tilde{\boldsymbol{w}}\rangle + \langle\tilde{\boldsymbol{w}},\boldsymbol{\Sigma}^{1/2}\widetilde{\boldsymbol{\mu}}\rangle\frac{\theta\tau_h}{\alpha} - \frac{\tau_h}{\alpha}\frac{\gamma_0}{\|\boldsymbol{\mu}\|_{\ell_p}}\left\|\boldsymbol{\Sigma}^{1/2}\tilde{\boldsymbol{w}}\right\|_{\ell_p} \\ &= \sup_{\tilde{\boldsymbol{w}}} \ \left\langle\boldsymbol{u} + \boldsymbol{\Sigma}^{1/2}\widetilde{\boldsymbol{\mu}}\frac{\theta\tau_h}{\alpha},\tilde{\boldsymbol{w}}\right\rangle - \frac{\tau_h}{\alpha}\frac{\gamma_0}{\|\boldsymbol{\mu}\|_{\ell_p}}\left\|\boldsymbol{\Sigma}^{1/2}\tilde{\boldsymbol{w}}\right\|_{\ell_p} \\ &= \sup_{\tilde{\boldsymbol{w}}} \ \left\langle\boldsymbol{\Sigma}^{-1/2}\boldsymbol{u} + \widetilde{\boldsymbol{\mu}}\frac{\theta\tau_h}{\alpha},\boldsymbol{\Sigma}^{1/2}\tilde{\boldsymbol{w}}\right\rangle - \frac{\tau_h}{\alpha}\frac{\gamma_0}{\|\boldsymbol{\mu}\|_{\ell_p}}\left\|\boldsymbol{\Sigma}^{1/2}\tilde{\boldsymbol{w}}\right\|_{\ell_p}\end{aligned}$$

By Hölder's inequality,

$$\left\langle\boldsymbol{\Sigma}^{-1/2}\boldsymbol{u} + \widetilde{\boldsymbol{\mu}}\frac{\theta\tau_h}{\alpha},\boldsymbol{\Sigma}^{1/2}\tilde{\boldsymbol{w}}\right\rangle \leq \left\|\boldsymbol{\Sigma}^{-1/2}\boldsymbol{u} + \widetilde{\boldsymbol{\mu}}\frac{\theta\tau_h}{\alpha}\right\|_{\ell_q}\left\|\boldsymbol{\Sigma}^{1/2}\tilde{\boldsymbol{w}}\right\|_{\ell_p}$$

Therefore, if $\boldsymbol{u}\in S$ then the supremum is achieved by choosing $\tilde{\boldsymbol{w}} = 0$. If $\boldsymbol{u}\notin S$, by scaling $\tilde{\boldsymbol{w}}$ the supremum would be $+\infty$.

**E.7. Proof of Lemma B.4** Fix arbitrary $\boldsymbol{u}$. By definition,

$$\begin{aligned}
\mathsf{P}_{\mathcal{B}}(\boldsymbol{u}) &:= \arg\min_{\boldsymbol{z}\in\mathcal{B}} \|\boldsymbol{u}-\boldsymbol{z}\|_{\ell_2} \\
&= \arg\min_{\boldsymbol{z}\in\mathcal{B}} \left\|\mathsf{P}_{\boldsymbol{\mu}}^{\perp}(\boldsymbol{u}-\boldsymbol{z})\right\|_{\ell_2}^2 + \|\mathsf{P}_{\boldsymbol{\mu}}(\boldsymbol{u}-\boldsymbol{z})\|_{\ell_2}^2 \\
&\overset{(a)}{=} \arg\min_{\boldsymbol{z}\in\mathcal{B}} \left\|\mathsf{P}_{\boldsymbol{\mu}}^{\perp}\boldsymbol{u}-\boldsymbol{z}\right\|_{\ell_2}^2 + \|\mathsf{P}_{\boldsymbol{\mu}}\boldsymbol{u}\|_{\ell_2}^2 \\
&= \arg\min_{\boldsymbol{z}\in\mathcal{B}} \left\|\mathsf{P}_{\boldsymbol{\mu}}^{\perp}\boldsymbol{u}-\boldsymbol{z}\right\|_{\ell_2}^2 \\
&= \mathsf{P}_{\mathcal{B}}\mathsf{P}_{\boldsymbol{\mu}}^{\perp}\boldsymbol{u}\,, \qquad \text{(E.10)}
\end{aligned}$$

where step $(a)$ follows from that fact that $\boldsymbol{z}\in\mathcal{B}$ and hence $\boldsymbol{z}=\mathsf{P}_{\boldsymbol{\mu}}^{\perp}\boldsymbol{z}$.

**E.8. Proof of Lemma B.5** By definition of the set $\mathcal{B}$, the value of $\|\mathsf{P}_{\mathcal{B}}(\boldsymbol{h})-\boldsymbol{h}\|_{\ell_2}^2$ is given by the optimal objective value of the following optimization:

$$\begin{aligned}
&\text{minimize}_{\boldsymbol{z}} \quad \|\boldsymbol{z}-\boldsymbol{h}\|_{\ell_2} \\
&\text{subject to} \quad \left\|\boldsymbol{\Sigma}^{-1/2}\boldsymbol{z}-\frac{\tau_h\theta}{\alpha}\widetilde{\boldsymbol{\mu}}\right\|_{\ell_q} \le \frac{\tau_h}{\alpha}\frac{\gamma_0}{\|\boldsymbol{\mu}\|_{\ell_p}}\,, \quad \widetilde{\boldsymbol{\mu}}^T\boldsymbol{z}=0\,. \qquad \text{(E.11)}
\end{aligned}$$

By the change of variable $\boldsymbol{u}:=\boldsymbol{\Sigma}^{-1/2}\boldsymbol{z}-\frac{\tau_h\theta}{\alpha}\widetilde{\boldsymbol{\mu}}$ and forming the Lagrangian, the optimal value of (E.11) is equal to the optimal value of the following problem:

$$\sup_{\lambda\ge0,\nu}\ \min_{\boldsymbol{u}}\ \frac{1}{2}\left\|\boldsymbol{\Sigma}^{1/2}\left(\boldsymbol{u}+\frac{\tau_h\theta}{\alpha}\widetilde{\boldsymbol{\mu}}\right)-\boldsymbol{h}\right\|_{\ell_2}^2+\lambda\left(\|\boldsymbol{u}\|_{\ell_q}^q-\left(\frac{\tau_h}{\alpha}\frac{\gamma_0}{\|\boldsymbol{\mu}\|_{\ell_p}}\right)^q\right)+\nu\widetilde{\boldsymbol{\mu}}^T\boldsymbol{\Sigma}^{1/2}\left(\boldsymbol{u}+\frac{\tau_h\theta}{\alpha}\widetilde{\boldsymbol{\mu}}\right). \qquad \text{(E.12)}$$

Rearranging the terms we get the next alternative representation

$$\sup_{\lambda\ge0,\nu}\ \min_{\boldsymbol{u}}\ \frac{1}{2}\left\|\boldsymbol{\Sigma}^{1/2}\left(\boldsymbol{u}+\frac{\tau_h\theta}{\alpha}\widetilde{\boldsymbol{\mu}}\right)-\boldsymbol{h}+\nu\frac{\boldsymbol{\mu}}{\|\boldsymbol{\mu}\|_{\ell_2}}\right\|_{\ell_2}^2+\lambda\left(\|\boldsymbol{u}\|_{\ell_q}^q-\left(\frac{\tau_h}{\alpha}\frac{\gamma_0}{\|\boldsymbol{\mu}\|_{\ell_p}}\right)^q\right)+\nu\widetilde{\boldsymbol{\mu}}^T\boldsymbol{h}-\frac{\nu^2}{2}\,, \qquad \text{(E.13)}$$

Now adopting the notation $\|\boldsymbol{v}\|_{\boldsymbol{\Sigma}}^2:=\boldsymbol{v}^T\boldsymbol{\Sigma}\boldsymbol{v}$ and invoking the assumption $\boldsymbol{\Sigma}^{1/2}\widetilde{\boldsymbol{\mu}}=a\widetilde{\boldsymbol{\mu}}$, we rewrite the optimization as follows:

$$\sup_{\lambda\ge0,\nu}\ \min_{\boldsymbol{u}}\ \frac{1}{2}\left\|\boldsymbol{u}+\left(\frac{\tau_h\theta}{\alpha}+\frac{\nu}{a}\right)\widetilde{\boldsymbol{\mu}}-\boldsymbol{\Sigma}^{-1/2}\boldsymbol{h}\right\|_{\boldsymbol{\Sigma}}^2+\lambda\left(\|\boldsymbol{u}\|_{\ell_q}^q-\left(\frac{\tau_h}{\alpha}\frac{\gamma_0}{\|\boldsymbol{\mu}\|_{\ell_p}}\right)^q\right)+\nu\widetilde{\boldsymbol{\mu}}^T\boldsymbol{h}-\frac{\nu^2}{2}\,, \qquad \text{(E.14)}$$

Rearranging the terms further we obtain

$$\sup_{\lambda\geq 0,\nu}\ \min_{\boldsymbol{u}}\left[\frac{1}{2}\left\|\boldsymbol{u}+\left(\frac{\tau_h\theta}{\alpha}+\frac{\nu}{a}\right)\widetilde{\boldsymbol{\mu}}-\boldsymbol{\Sigma}^{-1/2}\boldsymbol{h}\right\|_{\boldsymbol{\Sigma}}^2+\lambda\left\|\boldsymbol{u}\right\|_{\ell_q}^q\right]-\lambda\left(\frac{\tau_h}{\alpha}\frac{\gamma_0}{\left\|\boldsymbol{\mu}\right\|_{\ell_p}}\right)^q+\nu\widetilde{\boldsymbol{\mu}}^T\boldsymbol{h}-\frac{\nu^2}{2}$$

(E.15)

$$=\sup_{\lambda\geq 0,\nu}\ e_{q,\boldsymbol{\Sigma}}\left(\boldsymbol{\Sigma}^{-1/2}\boldsymbol{h}-\left(\frac{\tau_h\theta}{\alpha}+\frac{\nu}{a}\right)\widetilde{\boldsymbol{\mu}};\lambda\right)-\lambda\left(\frac{\tau_h}{\alpha}\frac{\gamma_0}{\left\|\boldsymbol{\mu}\right\|_{\ell_p}}\right)^q+\nu\widetilde{\boldsymbol{\mu}}^T\boldsymbol{h}-\frac{\nu^2}{2}\,.$$

This concludes the proof.

**E.9. Proof of Lemma B.6** We recall the definition of weighted Moreau envelope

$$e_{2,\boldsymbol{\Sigma}}(\boldsymbol{x};\lambda)=\min_{\boldsymbol{v}}\frac{1}{2}\left\|\boldsymbol{x}-\boldsymbol{v}\right\|_{\boldsymbol{\Sigma}}^2+\lambda\left\|\boldsymbol{v}\right\|_{\ell_2}^2\,. \tag{E.16}$$

Setting derivative to zero we get

$$-\boldsymbol{\Sigma}(\boldsymbol{x}-\boldsymbol{v}^*)+2\lambda\boldsymbol{v}^*=0\,,$$

which implies that $\boldsymbol{v}_*=(\boldsymbol{\Sigma}+2\lambda\boldsymbol{I})^{-1}\boldsymbol{\Sigma}\boldsymbol{x}$. Now consider a singular value decomposition $\boldsymbol{\Sigma}=\boldsymbol{U}\boldsymbol{S}\boldsymbol{U}^T$. Then, $\boldsymbol{v}_*=\boldsymbol{U}(\boldsymbol{S}+2\lambda\boldsymbol{I})^{-1}\boldsymbol{S}\boldsymbol{U}^T\boldsymbol{x}$. Substituting for $\boldsymbol{v}_*$ in (E.16) we obtain

$$\begin{aligned}
e_{2,\boldsymbol{\Sigma}}(\boldsymbol{x};\lambda)&=2\lambda^2\left\|\boldsymbol{U}(\boldsymbol{S}+2\lambda\boldsymbol{I})^{-1}\boldsymbol{S}^{1/2}\boldsymbol{U}^T\boldsymbol{x}\right\|_{\ell_2}^2+\lambda\left\|\boldsymbol{U}(\boldsymbol{S}+2\lambda\boldsymbol{I})^{-1}\boldsymbol{S}\boldsymbol{U}^T\boldsymbol{x}\right\|_{\ell_2}^2\\
&=\lambda\boldsymbol{x}^T\boldsymbol{U}^T(\boldsymbol{S}+2\lambda\boldsymbol{I})^{-1}\boldsymbol{S}\boldsymbol{U}\boldsymbol{x}\\
&=\lambda\left\|\boldsymbol{U}(\boldsymbol{S}+2\lambda\boldsymbol{I})^{-1/2}\boldsymbol{S}^{1/2}\boldsymbol{U}\boldsymbol{x}\right\|_{\ell_2}^2\\
&=\lambda\left\|(\boldsymbol{\Sigma}+2\lambda\boldsymbol{I})^{-1/2}\boldsymbol{\Sigma}^{1/2}\boldsymbol{x}\right\|_{\ell_2}^2
\end{aligned}$$

which yields the desired result.

## APPENDIX F: PROPOSITION 3.2 (AN EXTENDED STATEMENT)

Consider the adversarial training loss

$$\mathcal{L}(\boldsymbol{\theta}):=\frac{1}{n}\sum_{i=1}^n\ell\left(y_i\langle\boldsymbol{x}_i,\boldsymbol{\theta}\rangle-\varepsilon\left\|\boldsymbol{\theta}\right\|_{\ell_q}\right),$$

where the loss $\ell(t)$ can be expressed as $\ell(t)=e^{-f(q)}$ obeying the following technical assumptions:

- $f : \mathbb{R} \to \mathbb{R}$ is $C^2$-smooth.
- $f'(q) > 0$ for all $q \in \mathbb{R}$.
- There exists $b_f \geq 0$ such that $qf'(q)$ is non-decreasing for $q \in (b_f, \infty)$ and $qf'(q) \to \infty$ as $q \to \infty$.
- Let $g : [f(b_f), \infty) \to [b_f, \infty)$ be the inverse function of $f$ on the domain $[b_f, \infty)$. There exists $p \geq 0$ such that for all $x > f(b_f)$, $y > b_f$,

$$\left|\frac{g''(x)}{g'(x)}\right| \leq \frac{p}{x}, \quad \left|\frac{f''(y)}{f'(y)}\right| \leq \frac{p}{y}.$$

(It can be verified that the above assumptions are satisfied by exponential loss and logistic loss.) Then, the gradient descent iterates

$$\boldsymbol{\theta}_{\tau+1} = \boldsymbol{\theta}_\tau - \mu \nabla \mathcal{L}(\boldsymbol{\theta}_\tau)$$

with a sufficiently small step size $\mu$ obey

$$\lim_{t\to\infty} \left\| \frac{\boldsymbol{\theta}_t}{\|\boldsymbol{\theta}_t\|_{\ell_2}} - \frac{\widetilde{\boldsymbol{\theta}}^{\varepsilon}}{\|\widetilde{\boldsymbol{\theta}}^{\varepsilon}\|_{\ell_2}} \right\|_{\ell_2} = 0, \tag{F.1}$$

where $\widetilde{\boldsymbol{\theta}}^{\varepsilon}$ is the solution to the following max-margin problem

$$\begin{aligned} \widetilde{\boldsymbol{\theta}}^{\varepsilon} = \arg\min_{\boldsymbol{\theta}\in\mathbb{R}^d} \quad & \|\boldsymbol{\theta}\|_{\ell_2}^2 \\ \text{subject to} \quad & y_i \langle \boldsymbol{x}_i, \boldsymbol{\theta} \rangle - \varepsilon \|\boldsymbol{\theta}\|_{\ell_q} \geq 1. \end{aligned} \tag{F.2}$$

University of Southern California
Data Science and Operations Department
Marshall School of Business
Los Angeles, CA 90089
E-mail: ajavanma@usc.edu

University of Southern California
Department of Electrical and Computer Engineering
Viterbi School of Engineering
Los Angeles, CA 90089
E-mail: soltanol@usc.edu