IEEE Open Journal of Signal Processing

Received 28 September 2020; revised 21 December 2020; accepted 26 December 2020. Date of publication 14 January 2021; date of current version 19 March 2021. The review of this article was arranged by Associate Editor Karen Livescu.

Digital Object Identifier 10.1109/OJSP.2021.3051849

On the Precise Error Analysis of Support Vector Machines

ABLA KAMMOUN (Member, IEEE), AND MOHAMED-SLIM ALOUINI (Fellow, IEEE)

Computer, Electrical, and Mathematical Sciences and Engineering Division, KAUST, Thuwal 23955 6900, Makkah Province, Saudi Arabia

CORRESPONDING AUTHOR: ABLA KAMMOUN (e-mail: abla.kammoun@gmail.com)

This work was funded by the Office of Sponsored Research at KAUST.

This article has supplementary material provided by the authors available at https://doi.org/10.1109/OJSP.2021.3051849.

ABSTRACT This paper investigates the asymptotic behavior of the soft-margin and hard-margin support vector machine (SVM) classifiers for simultaneously high-dimensional and numerous data (large n and large p with $n/p \to \delta$) drawn from a Gaussian mixture distribution. Sharp predictions of the classification error rate of the hard-margin and soft-margin SVM are provided, as well as asymptotic limits of as such important parameters as the margin and the bias. As a further outcome, the analysis allows for the identification of the maximum number of training samples that the hard-margin SVM is able to separate. The precise nature of our results allows for an accurate performance comparison of the hard-margin and soft-margin SVM as well as a better understanding of the involved parameters (such as the number of measurements and the margin parameter) on the classification performance. Our analysis, confirmed by a set of numerical experiments, builds upon the convex Gaussian min-max Theorem, and extends its scope to new problems never studied before by this framework.

INDEX TERMS Performance analysis, statistical learning, support vector machines.

I. INTRODUCTION

With the advent of the era of big data, attention is now turned to modern classification problems that require to solve non-linear problems involving large and numerous data sets. Large margin classifiers constitute a typical example of these novel classification methods and include as particular cases support vector machines [1], logistic regression [2] and Adaboost [3]. The performance of these methods is known to be very sensitive to some design parameters, the setting of which is considered as a critical step, as inappropriate values can lead to severe degradation in the performance of the underlying classification technique. To properly set these design parameters, cross validation is the standard approach that has been adopted in the machine learning research. However, such an approach becomes rather computationally expensive in high dimensional settings, since it involves to design the classifier for each candidate value of the design parameters. Recently, a new technique based on large dimensional statistical analyses has emerged to assist in the design of a set of machine learning algorithms including kernel clustering techniques [4], classification [5], [6], and regression. It is based on determining sharp performance characterizations that can be assessed based on the foreknowledge of the data statistics or be approximated using training data. The advantages of this new technique are two-fold. First, it allows easy prediction of the performances for any set of design parameters, avoiding the prohibitively high computational complexity of the cross-validation approach and paving the way towards optimal setting of the design parameters. Second, it is more instrumental to gain a deep understanding of the performances with respect to the data statistics and the different underlying parameters. However, the application of this approach has been mainly concentrated on methods and algorithms in which the output possesses a closed-form expression, as algorithms involving implicit formulation are mathematically much less tractable.

Recently, a line of research work has emerged that studies the performance of high-dimensional regression problems involving non-smooth convex optimization methods. The approaches that have thus far been used can be classified into three main categories: a leave-one out approach proposed by El Karoui in [7], an approximate message passing based approach developed in [8] and finally the convex Gaussian

min-max theorem (CGMT) based approach initiated by Stojnic [9] and further developed by Thrampoulidis *et al* in [10]. As far as regression problems are concerned, the CGMT has been the basis of a unified framework that applies to a broad class of regression problems, requiring less assumptions on the structure of the objective function.

The present work focuses on the use of the CGMT for the asymptotic analysis of the popular support vector machines (SVM) [11]. An early conference version of this paper appears in [12] but is limited to the study of the hard-margin SVM and contains only sketch of the proofs. Previous works considering the analysis of the SVM have been based on either the replica method [13] or leave-one out based approaches [14]. However, these works were not shown rigorously with some intermediate results admitted without any proof. It should be noted that the optimization problem involved in SVM could not be written as an instance of the general high-dimensional regression problem considered in [10]. Moreover, it raises several new challenges towards the direct application of the CGMT. Compared to the works in [13] and [14], this work considered binary classification of isotropically distributed Gaussian data and studies the feasibility of the hard-margin SVM. As a major outcome, we prove that when the ratio between the number of samples and that of features is strictly less than 2, the hard-margin SVM is almost surely feasible irrespective of how close are the distribution of both classes. Additionally, we characterize the test error for both hard-margin and soft-margin SVM and illustrate how it is affected by the difference between mean vectors and the available number of training samples. Such an information can be leveraged in practice to acquire a fast estimation of the classification performances without resorting to cross-validation approaches. On a technical level, and as opposed to previous works pursuing the same line of research, the analysis is rigorous and is based on the CGMT framework. It develops new technical tools that can be leveraged in the future to extend the CGMT framework to more general settings.

The rest of the paper is organized as follows. Section II introduces the hard-margin and soft-margin SVM as well as the considered statistical model. Section III presents our main results along with some important implications. Numerical illustrations are provided in Section IV. Finally, Section V is devoted to the development of the technical proofs.

II. PROBLEM FORMULATION

Consider a set of training observations $\{(\mathbf{x}_i, y_i)\}_{i=1}^n$ where for each $\mathbf{x}_i \in \mathbb{R}^p$ a given input vector, $y_i = 1$ if $\mathbf{x}_i$ belongs to class $\mathcal{C}_1$ or $y_i = -1$ if $\mathbf{x}_i$ belongs to class $\mathcal{C}_0$. We assume that there are n_0 observations in class $\mathcal{C}_0$ and n_1 observations in class $\mathcal{C}_1$, both of them are drawn from Gaussian distribution with different means and common covariance matrix equal to $\sigma^2\mathbf{I}_p$. More specifically:

$$i \in \mathcal{C}_k \quad \Leftrightarrow \quad \mathbf{x}_i \sim \mathcal{N}(\boldsymbol{\mu}_k, \sigma^2\mathbf{I}_p) \ .$$

As suggested by several previous studies [15]–[17], the performance of a classifier shall depend on the difference between the mean vectors $\boldsymbol{\mu} \triangleq \boldsymbol{\mu}_1 - \boldsymbol{\mu}_0$ and the covariance matrix associated with each class, which is in our case equal to $\sigma^2\mathbf{I}_p$. Since the classification problem would not change upon a translation of all observations with the same vector, we will assume for technical reasons that $\boldsymbol{\mu}_0 = -\boldsymbol{\mu}$ and $\boldsymbol{\mu}_1 = \boldsymbol{\mu}$ without any loss of generality. This is because one can transform $\mathbf{x}_i$ into $\mathbf{x}_i - \frac{1}{2}(\boldsymbol{\mu}_1 + \boldsymbol{\mu}_2)$ without changing the classification performance[1].

A. HARD MARGIN SVM

Given a set of training data $\{(\mathbf{x}_i, y_i)\}_{i=1}^n$ that is linearly separable, hard-margin SVM seeks the affine plane that separates both classes with the maximum margin [1]. This amounts to solving the following optimization problem:

$$\Phi^{(n)} \triangleq \min_{\mathbf{w},b} \ \|\mathbf{w}\|_2^2$$
$$s.t. \ \forall i \in \{1,\ldots,n\}, \ \ y_i\left(\mathbf{w}^T\mathbf{x}_i + b\right) \geq 1 \ . \qquad (1)$$

Let $\hat{\mathbf{w}}_H$ and $\hat{b}_H$ solve the above problem, then the hard-margin classifier applied to an unseen observation $\mathbf{x}$ is given by $L_H(\mathbf{x}) = \text{sign}(\hat{\mathbf{w}}_H^T\mathbf{x} + \hat{b}_H)$.

B. SOFT MARGIN SVM

If the data are not linearly separable, the constraints of the hard-margin SVM can not all be satisfied together. As a result, the cost of the hard-margin optimization problem is infinite, since the minimum over an empty set is by convention ∞. Under such settings, one alternative is to use the soft-margin SVM which by construction tolerates that some training data are mis-classified but pays the cost of each misclassified observation by adding an upper bound on the number of the misclassified training observations. More formally, the soft-margin SVM is equivalent to solving the following optimization problem:

$$\tilde{\Phi} \triangleq \min_{\mathbf{w},b,\{\xi_i\}_{i=1}^n} \|\mathbf{w}\|_2^2 + \frac{\tilde{\tau}}{p}\sum_{i=1}^n \xi_i$$
$$s.t. \ \forall i \in \{1,\ldots,n\}, \ \ y_i\left(\mathbf{w}^T\mathbf{x}_i + b\right) \geq 1 - \xi_i, \ \ \xi_i \geq 0 \qquad (2)$$

where $\tilde{\tau}$ is a strictly positive scalar, set beforehand by the user, and aims to make a trade-off between maximizing the margin and minimizing the training error. In this respect, a small $\tilde{\tau}$ tends to put more emphasis on the margin while a larger $\tilde{\tau}$ penalize the training error. Let $\hat{\mathbf{w}}_S$ and $\hat{b}_S$ solve the above problem, then the soft-margin SVM classifier applied to an unseen observation $\mathbf{x}$ is given by $L_S(\mathbf{x}) = \text{sign}(\hat{\mathbf{w}}_S^T\mathbf{x} + \hat{b}_S)$.

III. MAIN RESULTS

The study of the statistical behavior of the hard-margin and soft-margin SVM is carried out under the following asymptotic regime:

Assumption A- 1: We shall assume the following

[1]The reader can easily see from our theoretical analysis that the same performances would be obtained if the empirical mean is subtracted from $\mathbf{x}_i$, i.e., $\mathbf{x}_i$ turned into $\mathbf{x}_i - \frac{n_0}{n}\boldsymbol{\mu}_0 - \frac{n_1}{n}\boldsymbol{\mu}_1$.

- n, n_0, n_1 and p grow to infinity with $\frac{n}{p} \to \delta$, $\frac{n_0}{n} \to \pi_0$ and $\frac{n_1}{n} \to \pi_1$.
- σ^2 is a fixed strictly positive scalar, while $\|\boldsymbol{\mu}\|_2 \to \mu$.
- The training samples $\mathbf{x}_1, \ldots, \mathbf{x}_n$ are independent. Moreover, for $k \in \{0, 1\}$, $\mathbf{x}_i \in \mathcal{C}_k$, if and only if $\mathbf{x}_i = y_k \boldsymbol{\mu} + \mathbf{z}_i$ with $\mathbf{z}_i \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_p)$ and $y_k = 1$ if $k = 1$ and $y_k = -1$ if $k = 0$.

A. HARD MARGIN SVM

In this section, we analyze the behavior of the hard-margin SVM under Assumption 1.

Theorem 1: Let $\eta^\star(\rho)$ be the unique solution in η to the following equation:

$$\eta = \frac{\pi_1 \mathbb{E}\left[\left(G - \frac{\rho\mu}{\sigma}\right)\mathbf{1}_{\{G \geq \frac{\rho\mu}{\sigma} + \eta\}}\right] + \pi_0 \mathbb{E}\left[\left(\frac{\rho\mu}{\sigma} - G\right)\mathbf{1}_{\{G \geq \frac{\rho\mu}{\sigma} - \eta\}}\right]}{\pi_1 \mathbb{P}\left[\mathbf{1}_{\{G \geq \frac{\rho\mu}{\sigma} + \eta\}}\right] + \pi_0 \mathbb{P}\left[\mathbf{1}_{\{G \geq \frac{\rho\mu}{\sigma} - \eta\}}\right]}. \tag{3}$$

where $G \sim \mathcal{N}(0, 1)$ denotes a standard normal random variable. Assume that:

$$\delta > \delta_\star \tag{4}$$

where $\delta_\star$ is given by:

$$= \left(\min_{0 \leq \rho \leq 1} \frac{1}{1 - \rho^2}\left(\pi_1 \mathbb{E}\left(\left(G - \frac{\rho\mu}{\sigma} - \eta^\star(\rho)\right)_+\right)^2 + \pi_0 \mathbb{E}\left(\left(G - \frac{\rho\mu}{\sigma} + \eta^\star(\rho)\right)_+\right)^2\right)\right)^{-1} \tag{5}$$

Then, under Assumption 1

$$\mathbb{P}\left[\Phi = \infty, n \text{ large enough}\right] = 1.$$

Proof: The proof is postponed to Section V-B. ■

Remark 1: At this point of the work, it is quite early to interpret the quantities $\eta(\rho)$ and ρ, the physical significance of which will appear clearly in the next theorem. However, for the reader convenience, we would like to mention that the proof shows clearly that asymptotically, the feasibility of the hard-margin SVM, i.e, finding the solutions $\hat{\mathbf{w}}_H$ and $\hat{b}_H$ to (1) asymptotically happens when it is possible to find the solutions of an equivalent optimization problem on the variables η and ρ. If the data are not linearly separable, the hard-margin SVM is not feasible and as such it is not possible to find $\hat{\mathbf{w}}_H$ and $\hat{b}_H$ that solve (1). Under this setting, the cost of an equivalent scalar optimization problem involving the scalar variables ρ and η becomes unbounded, a scenario that happens when $\delta > \delta_\star$. More details of this scenario are provided in Remark 3 subsequent to Theorem 2. Finally, it is worth mentioning that in case $\pi_0 = \pi_1$, it is easy to see that $\eta = 0$ is the unique solution to (3). Moreover, one can easily see that in this case (4) becomes:

$$\delta \geq \delta_\star \quad \text{with} \quad \delta_\star = \frac{1}{\inf_{t \in \mathbb{R}} \mathbb{E}\left(\left(\sqrt{1 + t^2}G - \frac{t\mu}{\sigma}\right)_+\right)^2} \tag{6}$$

where to obtain (6), we used the change of variable $t := \frac{\rho}{\sqrt{1-\rho^2}}$ Note that (6) is reminiscent of the condition derived by Candès *et al.* in [17], which provides a similar condition guaranteeing data separability. However, the work of [18] considered data drawn from the logistic model with mean zero and was based on different tools.

Remark 2: Theorem 1 establishes the failure of the hard-margin SVM to linearly separate the data when the ratio between the number of samples and that of features is almost surely strictly above a certain threshold. Equivalently, it can be used to have an idea of the minimum number of training samples that cannot be linearly separated without errors. Assuming that n and p are sufficiently large, if the number of training samples satisfies:

$$n \ > \ p(\delta_\star + \epsilon)$$

for some fixed $\epsilon > 0$, then the hard margin SVM fails to linearly separate the training samples. The above result does not tell, however, as to when the hard-margin SVM guarantees perfect separation of the training samples. This constitutes the objective of the following Theorem, which guarantees the almost sure feasibility of the hard-margin SVM when $\delta < \delta_\star$ and determines under this condition almost sure limits of the margin, the bias, and the angle between the solution vector $\hat{\mathbf{w}}_H$ and vector $\boldsymbol{\mu}$.

Theorem 2: Assume that:

$$\delta < \delta_\star$$

Let $\beta(q_0) := \min_{\substack{0 \leq \rho \leq 1 \\ \eta \in \mathbb{R}}} D_H(q_0, \rho, \eta)$ where D_H is defined in (7) shown at bottom of this page. Then, function β has a unique zero $q^\star_{0,H}$. Moreover, with probability 1, for n and p large enough,

$$\|\hat{\mathbf{w}}_H\| \to q^\star_{0,H}.$$

Let $\rho^\star_H$ and $\eta^\star_H$ be such that $(\rho^\star_H, \eta^\star_H) = \operatorname{argmin}_{\substack{0 \leq \rho \leq 1 \\ \eta \in \mathbb{R}}} D_H(q^\star_{0,H}, \rho, \eta)$, then, with probability 1,

$$\frac{\hat{\mathbf{w}}_H^T \boldsymbol{\mu}}{\|\hat{\mathbf{w}}_H\| \|\boldsymbol{\mu}\|} \to \rho^\star_H, \quad \text{and} \quad \hat{b}_H \to \eta^\star_H q^\star_{0,H} \sigma \ .$$

Proof: The proof is postponed to Section V-B. ■

Remark 3: In reference to our discussion after Theorem 1, and to give the reader the intuition behind the result of the above theorem, it followed from the proof that the feasibility

$$D_H(q_0, \rho, \eta) \triangleq \sqrt{\delta\pi_1 \mathbb{E}\left(\left(G + \frac{1}{q_0\sigma} - \frac{\rho\mu}{\sigma} - \eta\right)_+\right)^2 + \delta\pi_0 \mathbb{E}\left((G + \frac{1}{q_0\sigma} - \frac{\rho\mu}{\sigma} + \eta)_+\right)^2 - \sqrt{1 - \rho^2}} \tag{7}$$

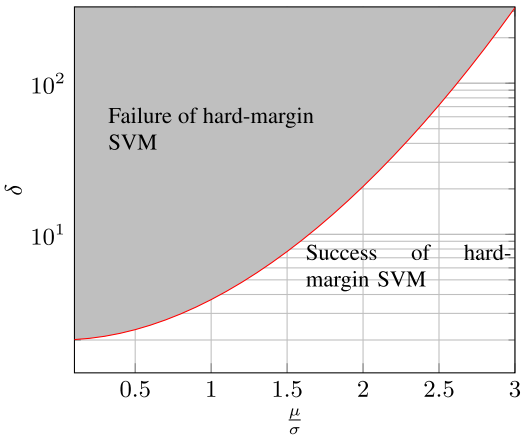


FIGURE 1. Theoretical predictions of the regions of failure and success of hard-margin SVM when $\pi_0 = \pi_1 = 0.5$.

of the hard-margin SVM is equivalent to showing that the cost of the following scalar optimization problem:

$$\inf_{q_0 \geq 0} q_0^2 + \sup_{m \geq 0} m \min_{\substack{-1 \leq \rho \leq 1 \\ \eta \in \mathbb{R}}} D_H(q_0, \rho, \eta)$$
$$= \inf_{q_0 \geq 0} q_0^2 + \sup_{m \geq 0} m q_0 \beta(q_0) \tag{8}$$

is bounded. As

$$q_0 \mapsto \min_{\substack{-1 \leq \rho \leq 1 \\ \eta \in \mathbb{R}}} D_H(q_0, \rho, \eta)$$

is decreasing with q_0 and grows like $\frac{1}{q_0}$ when $q_0 \downarrow 0$, for the optimal cost of the above optimization problem to be bounded, necessarily we must have $\beta(q_0) \leq 0$ for some $q_0 > 0$. Since function β is decreasing, such a condition holds if and only if $\lim_{q_0 \to \infty} \beta(q_0) < 0$, which can be easily seen to be equivalent to $\delta < \delta_\star$. Moreover, when $\delta < \delta_\star$, using the fact that β is decreasing, the minimizer of (8) corresponds to the unique zero of function β.

Remark 4: The combination of the results of Theorem 1 and Theorem 2 provides a complete picture of the behavior of the hard-margin-SVM. The importance of these results is that they allow us to predict the performance for any setting characterized by n, p, μ and σ, and constitutes thus a valuable alternative to cross-validation approaches. Particularly, it entails from these results that for the hard-margin SVM to lead to perfect linear separation of the training samples, the number of samples should be strictly less than:

$$n < \; p\,(\delta_\star - \epsilon)$$

for some $\epsilon > 0$. In case $\frac{n}{p} \to \delta_\star$, no conclusion can be drawn. This phase transition phenomenon is illustrated in Fig. 1 $\pi_0 = \pi_1 = 0.5$, which displays the failure and success regions with varying $\frac{\mu}{\sigma}$ and δ. Interestingly, it is noteworthy to mention that as the factor $\frac{\mu}{\sigma}$ increases, the capabilities of the hard-margin SVM get improved, the number of samples that can be linearly separated increasing significantly.

Remark 5: It is easy to see that $\delta_\star$ increases with $\frac{\mu}{\sigma}$. Hence $\delta_\star(\frac{\mu}{\sigma}) \geq \delta_\star(\frac{\mu}{\sigma} = 0)$. Consider the case in which $\frac{\mu}{\sigma} = 0$, and $\pi_0 = \pi_1 = \frac{1}{2}$, which describes the situation in which data from both classes follow the same distribution with mean 0 and covariance $\sigma^2 \mathbf{I}_p$. Simple calculations lead to that $\delta_\star(\frac{\mu}{\sigma} = 0) = 2$. So, when the data is uniformly sampled across the classes, the hard-margin SVM would be able to linearly separate the data when $\delta < 2$, yielding perfect classification of training data. Curiously, this holds even when the training data from both classes are drawn from the same distribution.

However, as will be shown below, in this situation, the test error will be equal to 0.5, since both classes are generated similarly.

Corollary 3: Let $L_H(\mathbf{x}) = \hat{\mathbf{w}}_H^T \mathbf{x} + \hat{b}_H$ be the hard-margin classifier, where $\hat{\mathbf{w}}_H$ and $\hat{b}_H$ are solutions to (1). Under the asymptotic regime defined in Assumption 1 and when $\delta < \delta_\star$, the classification error rate associated with class $\mathcal{C}_0$ and $\mathcal{C}_1$ converges to:

$$\mathbb{P}\left[L_H(\mathbf{x}) > 0 | \mathbf{x} \in \mathcal{C}_0\right] \to Q\left(\frac{\rho_H^\star \mu}{\sigma} - \eta_H^\star\right)$$
$$\mathbb{P}\left[L_H(\mathbf{x}) < 0 | \mathbf{x} \in \mathcal{C}_1\right] \to Q\left(\frac{\rho_H^\star \mu}{\sigma} + \eta_H^\star\right),$$

where $Q(x) = \frac{1}{\sqrt{2\pi}} \int_x^\infty \exp(-\frac{t^2}{2}) dt$. Let ε_H denote the test error of the hard-margin SVM. It thus converges to $\varepsilon_H^\star = \pi_0 Q(\frac{\rho_H^\star \mu}{\sigma} - \eta_H^\star) + \pi_1 Q(\frac{\rho_H^\star \mu}{\sigma} + \eta_H^\star)$.

Remark 6: In reference to Remark 5, it can be shown that when $\pi_0 = \pi_1 = 0.5$ and $\mu = 0$, the test error converges to 0.5 although the training error is zero. This is expected since data from both classes are drawn from the same distribution. Moreover, as seen in Corollary 3, the classification error rates depends on the bias through $\eta_H^\star$, and the angle between $\boldsymbol{\mu}$ and $\hat{\mathbf{w}}_H$, capitalized by the quantity $\rho_H^\star$. In our case, the optimal Bayes separating hyperplane has direction aligned with $\boldsymbol{\mu}$, hence $\rho_H^\star$ also represents the angle between the direction of SVM separating hyperplane and the Bayes optimal separating hyperplane. As $\|\hat{\mathbf{w}}_H\| \to q_{0,H}^\star$, the projection of $\hat{\mathbf{w}}_H$ on the space orthogonal to $\boldsymbol{\mu}$ has a norm that converges to $q_{0,H}^\star \sqrt{1 - (\rho_H^\star)^2}$. Finally, it is important to note that the classification error rate is not the same for both classes, unless $\pi_0 = \pi_1$ in which case it is easy to see that $\eta_H^\star = 0$. Moreover, if $\pi_1 > \pi_0$, it is easy to prove that $\eta_H^\star > 0$. Hence, it is the class with a higher number of training data that presents the lowest misclassification error rate.

Remark 7: The expressions provided in Theorem 2 and Corollary 3 can be used to qualitatively understand the limiting behaviors of $q_{0,H}^\star$ and $\rho_H^\star$ in the balanced training data ($\pi_1 = \pi_0 = 0.5$) when 1) δ is fixed and $\frac{\mu}{\sigma}$ tends to zero or infinity and when 2) $\delta \to 0$ or tends from below to $\delta_\star$.

Impact of low mean difference: Assume $\delta < 2$ is fixed and $\frac{\mu}{\sigma} \downarrow 0$. Then D_H can be approximated as:

$$D_H(q_0, \rho, 0) \underset{\frac{\mu}{\sigma} \to 0}{\sim} \sqrt{\delta} \sqrt{\mathbb{E}\left(G + \frac{1}{q_0 \sigma}\right)_+^2} - \sqrt{1 - \rho^2}$$

Hence, when $\frac{\mu}{\sigma} \downarrow 0$, $\rho_H^\star \to 0$ and $q_{0,H}^\star \to \tilde{q}_0$ where $\tilde{q}_0$ is the unique solution to:

$$\mathbb{E}\left(G + \frac{1}{q_0\sigma}\right)_+^2 = \frac{1}{\delta}.$$

Impact of high mean difference: Assume δ is fixed below $\delta_\star$ and $\frac{\mu}{\sigma} \to \infty$. Then, considering function $\tilde{\beta}(q_0) := \frac{\mu}{\sigma}\beta(q_0)$, one can easily see that:

$$\tilde{\beta} \underset{\frac{\mu}{\sigma}\to\infty}{\sim} \sqrt{\delta} \min_{0\le\rho\le 1} \sqrt{\mathbb{E}(G\frac{\mu}{\sigma} + \frac{1}{q_0\mu} - \rho)_+^2}$$
$$\sim \min_{0\le\rho\le 1} \sqrt{\delta}(\frac{1}{q_0\mu} - \rho)_+$$

Obviously for all q_0, the minimum in ρ of the asymptotic equivalent of $\tilde{\beta}$ is achieved when $\rho = 1$. Moreover, choosing $q_0 = \frac{1}{\mu}$ and $\rho = 1$, we obtain $\tilde{\beta} \sim 0$. It follows from this, that when $\frac{\mu}{\sigma} \to \infty$, $\rho_H^\star \to 1$ and $q_{0,H}^\star \sim \frac{1}{\mu}$.

Impact of small sample size Assume $\delta \downarrow 0$ while $\frac{\mu}{\sigma}$ is fixed. Then function β can be approximated as:

$$\beta \underset{\delta\downarrow 0}{\sim} \min_{0\le\rho\le 1} \frac{\sqrt{\delta}}{q_0\sigma} - \sqrt{1-\rho^2}$$

Hence, $\rho_H^\star \to 0$ and $q_{0,H}^\star \sim \frac{\sqrt{\delta}}{\sigma}$.

Impact of using the highest possible sample size In the limit when $\delta \to \delta_\star$, the set $\{(q_0, \rho) | D_H(q_0, \rho, 0) \le 0\}$ converges to $\{(q_0 = \infty, \overline{\rho}_H)\}$ where $\overline{\rho}_H$ is the unique ρ satisfying:

$$\sqrt{\delta_\star}\sqrt{\mathbb{E}\left(G - \frac{\rho\mu}{\sigma}\right)_+^2} - \sqrt{1-\rho^2} = 0.$$

Hence, when $\delta \to \delta_\star$, $q_{0,H}^\star \to \infty$ and $\rho_H^\star \to \overline{\rho}_H$.

As seen from Corollary 3, it is the alignment with $\boldsymbol{\mu}$ and not the margin that determines the performance. A bad alignment occurs when $\rho_H^\star \to 0$ and is associated with poor performances. This happens for instance when $\frac{\mu}{\sigma} \to 0$ or when the sample size is small ($\delta \to 0$). When $\frac{\mu}{\sigma} \to \infty$, $\hat{\mathbf{w}}_H$ tends to align perfectly with $\boldsymbol{\mu}$, which translates into perfect classification. On the contrary, the use of as many samples as allowed by the condition $\delta < \delta_\star$ is not associated with perfect alignment. A better performance is expected as compared with $\delta \to 0$ but far from perfect classification error rate.

B. SOFT MARGIN SVM

The following Theorem characterizes the asymptotic behavior of the solution to the soft-margin SVM under the asymptotic regime defined in Assumption 1.

Theorem 4: Let the map $R_{\tilde{\tau}}(x, \rho, \eta, \xi) : \mathbb{R}_{>0} \times [0, 1] \times \mathbb{R} \times \mathbb{R}_{>0} \to \mathbb{R}$ be defined in (9), shown at bottom of this page. Define $\mathcal{D}_{S,\tilde{\tau}}(q_0, \rho, \eta, \xi)$:$\mathbb{R}_{>0} \times [0, 1] \times \mathbb{R} \times \mathbb{R}_{>0} \to \mathbb{R}$ as:

$$\mathcal{D}_{S,\tilde{\tau}}(q_0, \rho, \eta, \xi) := q_0^2 + q_0 R_{\tilde{\tau}}\left(\frac{1}{q_0}, \rho, \eta, \xi\right).$$

Then, the following convex-concave minimax optimization problem

$$\inf_{q_0>0} \inf_{\eta\in\mathbb{R}} \min_{0\le\rho\le 1} \sup_{\xi>0} \mathcal{D}_{S,\tilde{\tau}}(q_0, \rho, \eta, \xi) \tag{10}$$

admits a unique solution $(q_{0,S}^\star, \rho_S^\star, \eta_S^\star)$. Moreover, with probability 1, the following convergences hold true:

$$\|\hat{\mathbf{w}}_S\|_2 \xrightarrow{\text{a.s.}} q_{0,S}^\star, \quad \frac{\hat{\mathbf{w}}_S^T\boldsymbol{\mu}}{\|\hat{\mathbf{w}}_S\|_2\|\boldsymbol{\mu}\|_2} \xrightarrow{\text{a.s.}} \rho_S^\star$$
$$\text{and } \hat{b}_S \xrightarrow{\text{a.s.}} \eta_S^\star q_{0,S}^\star \sigma.$$

Proof: See Section V-C in the technical report [18]. ■

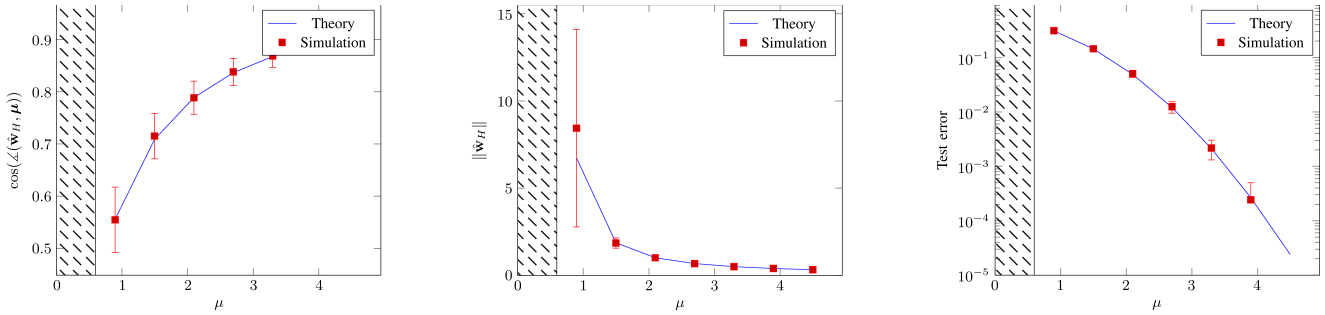
Corollary 5 (Misclassification error rate): Let $L_S(\mathbf{x}) = \hat{\mathbf{w}}_S^T\mathbf{x} + \hat{b}_S$ be the soft margin SVM classifier, where $\hat{\mathbf{w}}_S$ and $\hat{b}_S$ are solutions to (2). Under the asymptotic regime defined in Assumption 1, the classification error rate of the soft-margin SVM classifier associated with class $\mathcal{C}_0$ and class $\mathcal{C}_1$ converge to:

$$\mathbb{P}\left[\hat{L}_S(\mathbf{x}) > 0 | \mathbf{x} \in \mathcal{C}_0\right] \to Q\left(\frac{\rho_S^\star\mu}{\sigma} - \eta_S^\star\right)$$
$$\mathbb{P}\left[\hat{L}_S(\mathbf{x}) < 0 | \mathbf{x} \in \mathcal{C}_1\right] \to Q\left(\frac{\rho_S^\star\mu}{\sigma} + \eta_S^\star\right).$$

where $\rho_S^\star$, $\eta_S^\star$ and $q_{0,S}^\star$ are the unique solutions to (10). Let ε_S denote the test error of the soft-margin SVM. It thus converges to $\varepsilon_S^\star = \pi_0 Q(\frac{\rho_S^\star\mu}{\sigma} - \eta_S^\star) + \pi_1 Q(\frac{\rho_S^\star\mu}{\sigma} + \eta_S^\star)$.

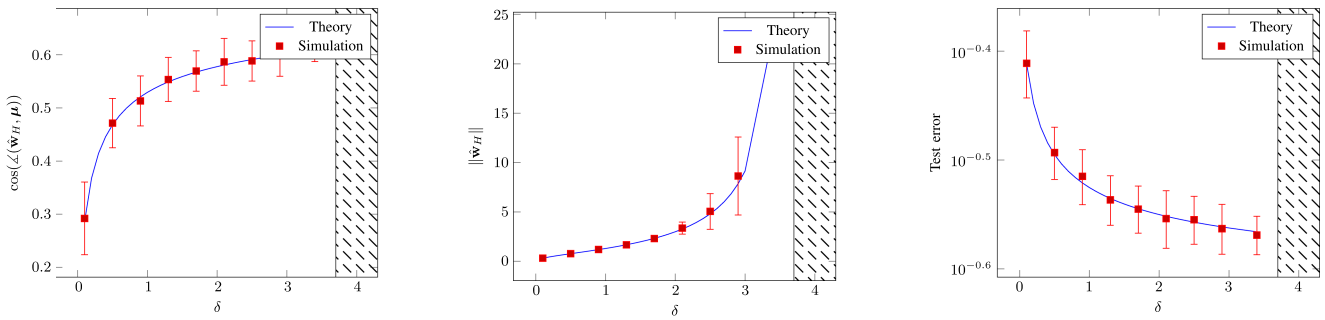
Remark 8: First, it is worth mentioning that the expressions obtained for the soft-margin SVM can be numerically approximated using a coordinate descent algorithm. Second, similar to the hard margin SVM, in case of balanced classes ($\pi_0 = \pi_1 = 0.5$), it is easy to see that $\eta^\star = 0$. This confirms the intuition according to which, for the symmetric case ($\boldsymbol{\mu}_1 = -\boldsymbol{\mu}_2$), it is best to separate the data with a hyperplane crossing the origin. Again it is easy to see that if $\pi_1 > \pi_0$, $\eta^\star > 0$, showing

$$R_{\tilde{\tau}}(x, \rho, \eta, \xi) := \tilde{\tau}\pi_1\delta\mathbb{E}\left[\left(G + \frac{x}{\sigma} - \frac{\rho\mu}{\sigma} - \eta - \frac{\tilde{\tau}}{2\xi}\right)\mathbf{1}_{\{G \ge \frac{\tilde{\tau}}{\xi} + \eta + \frac{\rho\mu}{\sigma} - \frac{x}{\sigma}\}}\right] + \tilde{\tau}\pi_0\delta\mathbb{E}\left[\left(G + \frac{x}{\sigma} - \frac{\rho\mu}{\sigma} + \eta - \frac{\tilde{\tau}}{2\xi}\right)\mathbf{1}_{\{G \ge \frac{\tilde{\tau}}{\xi} - \eta + \frac{\rho\mu}{\sigma} - \frac{x}{\sigma}\}}\right]$$
$$+ \frac{\xi\pi_1\delta}{2}\mathbb{E}\left[\left(\left(G + \frac{x}{\sigma} - \frac{\rho\mu}{\sigma} - \eta\right)_+\right)^2\mathbf{1}_{\{G \le -\frac{x}{\sigma} + \frac{\rho\mu}{\sigma} + \eta + \frac{\tilde{\tau}}{\xi}\}}\right] + \frac{\xi\pi_0\delta}{2}\mathbb{E}\left[\left(\left(G + \frac{x}{\sigma} - \frac{\rho\mu}{\sigma} + \eta\right)_+\right)^2\mathbf{1}_{\{G \le -\frac{x}{\sigma} + \frac{\rho\mu}{\sigma} - \eta + \frac{\tilde{\tau}}{\xi}\}}\right]$$
$$- \frac{\xi}{2}(1 - \rho^2) \tag{9}$$



(a) Angle between $\boldsymbol{\mu}$ and $\hat{\mathbf{w}}_H$ (b) $\|\hat{\mathbf{w}}_H\|$ (c) Test error

FIGURE 2. Effect of μ on hard-margin SVM performances, when $\delta = 2.5$, $\pi_0 = \pi_1 = 0.5$, $p = 100$ and $\sigma = 1$. The solid blue line corresponds to $\rho_H^\star$, $q_{0,H}^\star$ and $\varepsilon^\star$ as defined in Theorem 2 and Corollary 3, while the squares and bars represent the mean and standard deviation of $\cos(\angle(\mu, \hat{\mathbf{w}}_H))$, $\|\hat{\mathbf{w}}_H\|_2$ and ε based on 100 simulated data sets.



(a) Angle between $\boldsymbol{\mu}$ and $\hat{\mathbf{w}}_H$ (b) $\|\hat{\mathbf{w}}_H\|$ (c) Test error

FIGURE 3. Effect of δ on hard-margin SVM performances, when $\mu = 1$, $\pi_0 = \pi_1 = 0.5$, $p = 200$ and $\sigma = 1$. The solid blue line corresponds to $\rho_H^\star$, $q_0^\star$ and $\varepsilon^\star$ as defined in Theorem 2 and Corollary 3, while the squares and bars represent the mean and standard deviation of $\cos(\angle(\mu, \hat{\mathbf{w}}_H))$, $\|\hat{\mathbf{w}}_H\|_2$ and ε based on 100 simulated data sets.

that the class with more training data is the one that presents the best misclassification performance.

IV. NUMERICAL RESULTS

In this section, we validate through a set of numerical results the accuracy of our theoretical finding. All the data are drawn from a Gaussian mixture model as defined in Assumption 1.

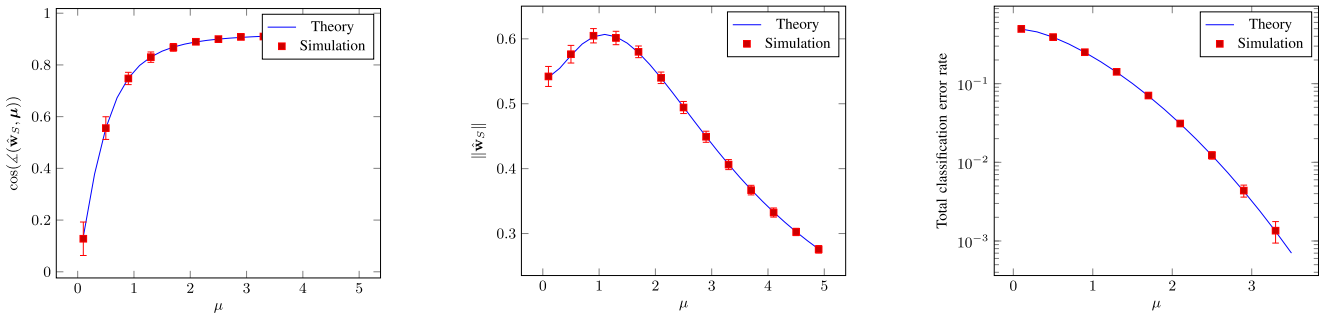
A. HARD MARGIN SVM

Fig. 2 illustrates the impact of μ on the angle between the optimal Bayes separating hyperplane, (in our case aligned with $\boldsymbol{\mu}$) and $\hat{\mathbf{w}}_H$ the separating hyperplane of hard-margin SVM, as well as on the margin and the classification error rate. As can be seen, the alignment with $\boldsymbol{\mu}$ improves rapidly with μ when μ is small before saturating for large values of μ. Moreover, when μ is small, the inverse of the margin which is proportional to the norm of $\|\hat{\mathbf{w}}_H\|$ reaches very high values, being in the limit of feasibility of the hard-margin SVM. Finally, as can be seen, the classification error rate decreases considerably as μ increases. Fig. 3 describes the impact of δ. We note that the use of more training samples tends to improve the alignment and at the same time to decrease the margin. This does not imply a reduction in the classification error performance. On the contrary, the better alignment results in a higher classification performance, despite the decrease of the margin value. For the sake of illustration, we represent hatched zones in Fig. 2 and Fig. 3 to refer to regions in which the hard-margin SVM is unfeasible. We note that the norm of $\hat{\mathbf{w}}$ increases when approaching to these unfeasibility regions.

B. SOFT MARGIN SVM

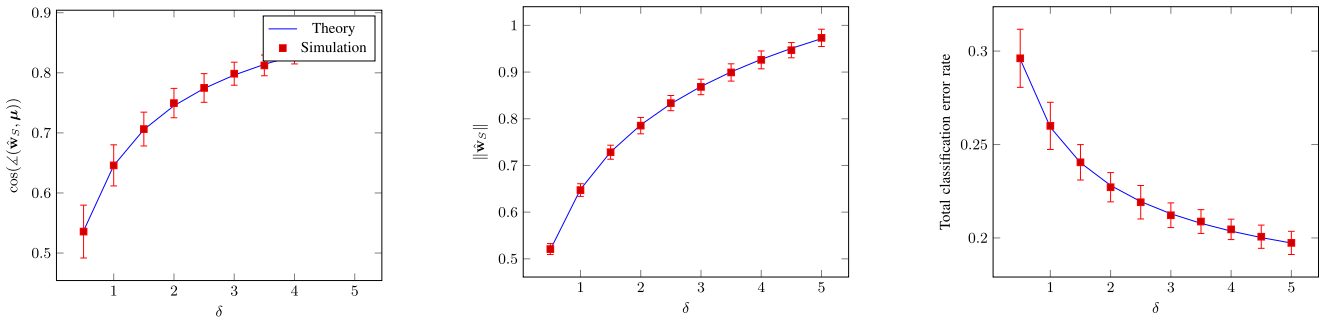
Fig. 4 investigates the impact of μ on the angle between the optimal Bayes separating hyperplane aligned with $\boldsymbol{\mu}$ and $\hat{\mathbf{w}}_S$ the separating hyperplane of SVM, as well as on the inverse of the margin and the classification error rate. It shows that the alignment significantly improves as μ increases fast when $\mu < 2$. The increase then becomes less important for high μ. We also note that curiously the margin tends to decrease in the range of small μ. This can be explained by the fact that in this region the alignment with the mean vector $\boldsymbol{\mu}$ is weak, causing the margin to decrease when μ is small. In the region of large μ, the margin increases rapidly ($\|\hat{\mathbf{w}}_S\|_2$ decreases).

Fig. 5 investigates the impact of the number of samples on the classification performances. As expected, as more training data are used, a better alignment with the mean vector $\boldsymbol{\mu}$ is noted. However, this results also in a decrease in the margin which does not hopefully translates into a loss in classification



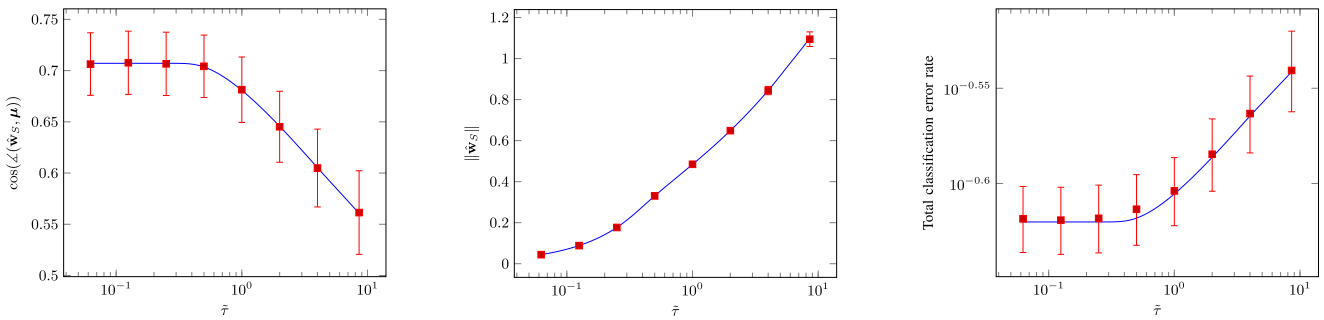
(a) Angle between $\boldsymbol{\mu}$ and $\hat{\mathbf{w}}_S$ (b) $\|\hat{\mathbf{w}}_S\|$ (c) Total classification error rate

FIGURE 4. Effect of μ when when $\tilde{\tau} = 2$, $\delta = 2$, $p = 200$, $\sigma = 1$, $\pi_1 = \pi_0 = 0.5$. The solid blue line corresponds to $\rho_S^\star$, $q_{0,S}^\star$ and $\varepsilon^\star$ as defined in Theorem 4, while the squares and bars represent the mean and standard deviation of $\cos(\angle(\mu, \hat{\mathbf{w}}_S))$, $\|\hat{\mathbf{w}}_S\|_2$ and ε based on 100 simulated data sets.



(a) Angle between $\boldsymbol{\mu}$ and $\hat{\mathbf{w}}_S$ (b) $\|\hat{\mathbf{w}}_S\|$ (c) Total classification error rate

FIGURE 5. Effect of δ when $\tilde{\tau} = 2$, $\mu = 1$, $p = 200$, $\sigma = 1$, $\pi_1 = \pi_0 = 0.5$. The solid blue line corresponds to $\rho_S^\star$, $q_{0,S}^\star$ and $\varepsilon^\star$ as defined in Theorem 4, while the squares and bars represent the mean and standard deviation of $\cos(\angle(\mu, \hat{\mathbf{w}}_S))$, $\|\hat{\mathbf{w}}_S\|_2$ and ε_S based on 100 simulated data sets.



(a) Angle between $\boldsymbol{\mu}$ and $\hat{\mathbf{w}}_S$ (b) $\|\hat{\mathbf{w}}_S\|$ (c) Total classification error rate

FIGURE 6. Effect of $\tilde{\tau}$ when $\mu = 1$, $p = 200$, $\delta = 1$, $\sigma = 1$ and $\pi_0 = \pi_1 = 0.5$. The solid blue line corresponds to $\rho_S^\star$, $q_{0,S}^\star$ and $\varepsilon_S^\star$ as defined in Theorem 4, while the squares and bars represent the mean and standard deviation of $\cos(\angle(\mu, \hat{\mathbf{w}}_S))$, $\|\hat{\mathbf{w}}_S\|_2$ and ε_S based on 2000 simulated data sets. (a) Angle between μ and $\hat{\mathbf{w}}_S$. (b) $\|\hat{\mathbf{w}}\|_S$. (c) Total classification error rate.

performances, these latter being determined by only how good is the alignment with $\boldsymbol{\mu}$.

Finally, we investigate in Fig. 6 the impact of $\tilde{\tau}$ on the performances. As seen, the alignment with $\boldsymbol{\mu}$ and the margin decreases significantly when τ is greater than a certain threshold value, suggesting to use smallest values of $\tilde{\tau}$. Such an observation is in agreement with the simulations of [13], where it was suggested to use the threshold value since using too tiny values for $\tilde{\tau}$ is known to pose numerical difficulties in solving the optimization problem.

V. TECHNICAL PROOFS

A. CGMT FRAMEWORK

Our technical proofs builds upon the CGMT framework, rooted in the works of Stojnic [9] and further mathematically formulated in the works of Thrampoulidis *et al.* in [10]

and [19]. The CGMT can be regarded as a generalization of a classical Gaussian comparison dating back to the early works of Gordon in 1988 [20]. This inequality provides a high-probability lower bound of the optimal cost function of any optimization problem that can be written in the following form:

$$\Phi^{(n)}(\mathbf{G}) := \min_{\mathbf{w}\in\mathcal{S}_\mathbf{w}} \max_{\mathbf{u}\in\mathcal{S}_\mathbf{u}} \mathbf{u}^T\mathbf{G}\mathbf{w} + \psi(\mathbf{w},\mathbf{u}), \tag{11}$$

where $\mathbf{G}\in\mathbb{R}^{n\times p}$ is a standard Gaussian matrix, $\mathcal{S}_\mathbf{w}$ and $\mathcal{S}_\mathbf{u}$ are two compact sets in $\mathbb{R}^p$ and $\mathbb{R}^n$ and $\psi : \mathbb{R}^p \times \mathbb{R}^n \to \mathbb{R}$ is continuous on $\mathcal{S}_\mathbf{w} \times \mathcal{S}_\mathbf{u}$, possibly random but independent of $\mathbf{G}$. The optimization problem in (11) is identified as a primary optimization problem (PO), the asymptotic behavior of which cannot be directly studied in general, due to the coupling between vectors $\mathbf{w}$ and $\mathbf{u}$ in the bilinear term. To this end, based on Gaussian comparison inequalities [21], we associate with it the following optimization problem

$$\phi^{(n)}(\mathbf{g},\mathbf{h}) := \min_{\mathbf{w}\in\mathcal{S}_\mathbf{w}} \max_{\mathbf{u}\in\mathcal{S}_\mathbf{u}} \|\mathbf{w}\|_2\mathbf{g}^T\mathbf{u} + \|\mathbf{u}\|_2\mathbf{h}^T\mathbf{w} + \psi(\mathbf{w},\mathbf{u}), \tag{12}$$

where $\mathbf{g}\in\mathbb{R}^n$ and $\mathbf{h}\in\mathbb{R}^p$ are standard Gaussian vectors. According to Gordon's comparison inequality, for any $c\in\mathbb{R}$, it holds that:

$$\mathbb{P}\left[\Phi^{(n)}(\mathbf{G}) < c\right] \le 2\mathbb{P}\left[\phi^{(n)}(\mathbf{g},\mathbf{h}) < c\right]. \tag{13}$$

Particularly, if a high-probability lower bound of the (AO) can be found, then by (13), this lower bound translates also into a high-probability lower bound of the (PO). The importance of this result lies in that so far it does not require any assumption on the convexity of function ψ or the sets $\mathcal{S}_\mathbf{w}$ and $\mathcal{S}_\mathbf{u}$, and most importantly it allows to relate the (PO) to a seemingly unrelated (AO) problem which presents the advantage of being in general much easier to analyze than the (PO) problem, as the bilinear term is now decoupled into two independent quantities involving respectively vectors $\mathbf{g}$ and $\mathbf{h}$. Combining the Gordon's original result with convexity, it was shown that this result can be strenghened to a more precise characterization of the asymptotic behavior of the (PO), [9], [22]. Particularly, if the sets $\mathcal{S}_\mathbf{w}$ and $\mathcal{S}_\mathbf{u}$ are additionally convex and ψ is convex-concave on $\mathcal{S}_\mathbf{w} \times \mathcal{S}_\mathbf{u}$, then, for any $\kappa\in\mathbb{R}$, and $t > 0$,

$$\mathbb{P}\left[\left|\Phi^{(n)}(\mathbf{G}) - \kappa\right| > t\right] \le 2\mathbb{P}\left[\left|\phi^{(n)}(\mathbf{g},\mathbf{h}) - \kappa\right| > t\right].$$

A direct consequence of this inequality is that if the optimal cost of the (AO) problem converges to ν then the optimal cost of the (PO) converges also to the same constant. More formally, if for some $\nu\in\mathbb{R}$, the optimal cost of the (AO) concentrates around ν in the sense that:

$$\mathbb{P}\left[\left|\phi^{(n)}(\mathbf{g},\mathbf{h}) - \nu\right| > t\right] \underset{n\to\infty}{\to} 0, \tag{14}$$

then similarly, the optimal cost of the (PO) satisfies:

$$\mathbb{P}\left[\left|\Phi^{(n)}(\mathbf{G}) - \nu\right| > t\right] \underset{n\to\infty}{\to} 0, \tag{15}$$

However, in most cases, the ultimate goal is not to characterize the optimal cost of the PO but rather a functional of the minimizer of $\Phi^{(n)}(\mathbf{G})$ which we denote by $\mathbf{w}_\Phi$. Although not directly obvious, this can be related to evaluation of the optimal cost of the (AO) as shown in the following Theorem.

Theorem 6: (CGMT, [23]) Let $\mathcal{S}$ be an arbitrary open subset of $\mathcal{S}_\mathbf{w}$ and $\mathcal{S}^c = \mathcal{S}_\mathbf{w}\backslash\mathcal{S}$. Denote $\phi^{(n)}_{\mathcal{S}^c}(\mathbf{g},\mathbf{h})$ the optimal cost of (11) when the optimization is constrained over $\mathbf{w}\in\mathcal{S}^c$. Suppose there exists constants $\overline{\phi}$ and $\overline{\phi}_{\mathcal{S}_c}$ such that (i) $\phi^{(n)}(\mathbf{g},\mathbf{h}) \to \overline{\phi}$ in probability, (ii) $\phi^{(n)}_{\mathcal{S}^c}(\mathbf{g},\mathbf{h}) > \overline{\phi}_{\mathcal{S}^c}$ with probability approaching 1, (iii) $\overline{\phi} < \overline{\phi}_{\mathcal{S}^c}$. Then, $\lim_{n\to\infty}\mathbb{P}[\mathbf{w}_\Phi\in\mathcal{S}] = 1$, where $\mathbf{w}_\Phi$ is a minimizer of (11).

Remark 9: Theorem 6 can be used to characterize a set in which lies the minimizer of (11) with probability approaching 1. The main ingredient is to compare the asymptotic limit of the AO optimal costs on the set of interest and its complementary. Note that Theorem 6 requires the asymptotic statements to hold in probability and not almost surely. In practice, as will be shown next, it is often the case that all asymptotic results of the (AO) hold in the almost sure sense. However, this cannot be straighforwardly leveraged to obtain results of the (PO) holding almost surely. The reason lies in that Gordon's lemma involves probability inequalities which directly establishes asymptotic results in probability. Using a converse version of the Borel Cantelli Lemma, we show that it is possible to transfer almost sure convergence results of the (AO) to that of the (PO), which allows us to obtain a stronger version for the CGMT.

Theorem 7: Let $\mathcal{S}$ be an arbitrary open subset of $\mathcal{S}_\mathbf{w}$ and $\mathcal{S}^c = \mathcal{S}_\mathbf{w}\backslash\mathcal{S}$. Denote $\phi^{(n)}_{\mathcal{S}^c}(\mathbf{g},\mathbf{h})$ the optimal cost of (11) when the optimization is constrained over $\mathbf{w}\in\mathcal{S}^c$. Suppose there exists constants $\overline{\phi}$ and $\overline{\phi}_{\mathcal{S}_c}$ such that (i) $\phi^{(n)}(\mathbf{g},\mathbf{h}) \to \overline{\phi}$ almost surely, (ii) $\phi^{(n)}_{\mathcal{S}^c}(\mathbf{g},\mathbf{h}) > \overline{\phi}_{\mathcal{S}^c}$ almost surely, (iii) $\overline{\phi} < \overline{\phi}_{\mathcal{S}^c}$. Then, $\mathbb{P}[\mathbf{w}_\Phi\in\mathcal{S}, \text{i.o.}] = 1$, where $\mathbf{w}_\Phi$ is a minimizer of (11).

Proof: Let $\eta = \frac{\overline{\phi}_{\mathcal{S}^c} - \overline{\phi}}{3} > 0$. Then, $\overline{\phi}_{\mathcal{S}^c} - \eta = \overline{\phi} + 2\eta$. The event $\mathcal{G}_n := \{\phi^{(n)}(\mathbf{g},\mathbf{h}) \ge \overline{\phi} + \eta\}$ does not occur infinitely often, hence,

$$\mathbb{P}\left[\phi^{(n)}(\mathbf{g},\mathbf{h}) \ge \overline{\phi} + \eta, i.o.\right] = 0.$$

Since $\mathcal{G}_n$ are independent, each event being generated by independent vectors $\mathbf{g}$ and $\mathbf{h}$, the converse of Borel-Cantelli Lemma implies that $\sum_{n=1}^{\infty}\mathbb{P}(\mathcal{G}_n) < \infty$. Similarly, we can prove that $\mathcal{R}_n := \{\phi^{(n)}_{\mathcal{S}^c}(\mathbf{g},\mathbf{h}) \le \overline{\phi}_{\mathcal{S}^c} - \eta\}$ satisfy $\sum_{n=1}^{\infty}\mathbb{P}[\mathcal{R}_n] < \infty$. Let $\Phi^{(n)}_{\mathcal{S}^c}(\mathbf{G})$ be the optimal cost of the PO problem when the minimization is constrained over $\mathbf{w}\in\mathcal{S}^c$. Consider now the event

$$\mathcal{K}_n = \left\{\Phi^{(n)}_{\mathcal{S}^c}(\mathbf{G}) \ge \overline{\phi}_{\mathcal{S}^c} - \eta,\ \Phi(\mathbf{G})^{(n)} \le \overline{\phi} + \eta\right\}$$

In this event, we have $\Phi^{(n)}_{\mathcal{S}^c}(\mathbf{G}) \ge \overline{\phi}_{\mathcal{S}^c} - \eta = \overline{\phi} + 2\eta$. Hence, $\Phi^{(n)}_{\mathcal{S}^c}(\mathbf{G}) > \overline{\Phi}(\mathbf{G})$, which implies that $\mathbf{w}_\Phi\in\mathcal{S}$. As a consequence,

$$\mathbb{P}\left[\mathbf{w}_\Phi\notin\mathcal{S}\right] \le \mathbb{P}(\mathcal{K}_n^c)$$

where $\mathcal{K}_n^c$ is the complementary event of $\mathcal{K}_n$. From the union bound,

$$\mathbb{P}\left[\mathcal{K}_n^c\right] \leq \mathbb{P}[\mathcal{R}_n] + \mathbb{P}[\mathcal{G}_n].$$

Hence,

$$\sum_{n=1}^{\infty} \left[\mathcal{K}_n^c\right] < \infty,$$

which proves that $\mathcal{K}_n^c$ and thus $\{\mathbf{w}_\Phi \in \mathcal{S}\}$ do not occur infinitely often. ■

Remark 10: In practice, to satisfy (i) and (ii) in Theorem 6 or Theorem 7, one can prove that $\phi^{(n)}(\mathbf{g}, \mathbf{h})$ converges to $\overline{\phi}$, while $\phi_{\mathcal{S}^c}^{(n)}(\mathbf{g}, \mathbf{h})$ is lower-bounded by a quantity that converges to $\overline{\phi}_{\mathcal{S}^c}$ with

$$\overline{\phi}_{\mathcal{S}^c} > \overline{\phi}. \tag{16}$$

Moreover, it is often the case that $\overline{\phi}$ and $\overline{\phi}_{\mathcal{S}^c}$ represent optimal costs of the same asymptotic objective function involving the same optimization variables, but with the variables of the latter being constrained to be away of the solution of the former. Under this setting, showing that their associated optimization problem possesses a unique solution is sufficient to prove (16).

The main advantage of the CGMT as a technical tool is that it leads to a unified approach for handling the performance analysis of solutions to high-dimensional optimization problems. When the optimization sets follow the assumptions of Theorem 7, the proof based on the CGMT proceeds in general into the following steps:

- Identification of the (PO) and its associated (AO)
- Simplification of the (AO): In practice this step involves reducing the (AO) into a scalar optimization problem.
- Asymptotic analysis of the (AO): This step involves proving that the (AO) converges to the optimal costs of a certain asymptotic optimization problem involving only scalar variables.
- Proof of the uniqueness of the solution to the asymptotic optimization problem associated with the (AO). This will allow us to satisfy the requirements (i), (ii) and (iii) in Theorem 6 and Theorem 7.

However, one major difficulty towards applying the CGMT in practice is related to the assumptions of compactness that does not always hold as well as the possible unfeasibility of the optimization problem. This is for instance the case of the hard-margin SVM considered in the present work. To overcome these technical issues, we approximate the original (PO) by a sequence of "bounded" (PO) problems, each satisfying the compactness conditions of the CGMT. We associate with this sequence of (PO) a sequence of (AO) and analyze their asymptotic behaviors. On each problem of the (PO) sequence, we apply Theorem 7 to analyze its asymptotic behavior. Based on this analysis, we develop a principled machinery that allows us to analyze the behavior of the unbounded original (PO).

B. HARD MARGIN SVM: PROOF OF THEOREM 1

We develop in this section the proof of Theorem 1 establishing the phase transition in the behavior of the hard-margin SVM. The proof relies on the CGMT framework and proceeds into the following steps:

Identification of the PO problem. The max-margin solution is obtained by solving the following optimization problem

$$\min_{\mathbf{w},b} \max_{\substack{\tilde{u}_i \geq 0 \\ i=1,\ldots,n}} \mathbf{w}^T\mathbf{w} + \sum_{i=1}^{n} \tilde{u}_i \left(1 - y_i \mathbf{w}^T \left(y_i \boldsymbol{\mu} + \sigma \mathbf{z}_i\right) - y_i b\right).$$

Let $\tilde{\mathbf{u}} = [\tilde{u}_1, \ldots, \tilde{u}_n]^T$ and $\mathbf{j}_i$ be the vector indexing the observations belonging to class $\mathcal{C}_i$. Let $\mathbf{Z} = [-y_1\mathbf{z}_1, \ldots, -y_n\mathbf{z}_n]^T$. We need thus to solve the following optimization problem

$$\min_{\mathbf{w},b} \max_{\tilde{\mathbf{u}} \geq 0} \mathbf{w}^T\mathbf{w} + \mathbf{1}^T\tilde{\mathbf{u}} - \mathbf{w}^T\boldsymbol{\mu}\mathbf{1}^T\tilde{\mathbf{u}} - (\mathbf{j}_1^T - \mathbf{j}_0^T)\tilde{\mathbf{u}}\, b + \sigma\tilde{\mathbf{u}}^T\mathbf{Z}\mathbf{w}.$$

Performing the change of variable $\mathbf{u} = \sigma\sqrt{p}\tilde{\mathbf{u}}$ leads to the following primary optimization problem

$$\Phi^{(n)} \triangleq \min_{\mathbf{w},b} \max_{\mathbf{u} \geq 0} \frac{1}{\sqrt{p}}\mathbf{u}^T\mathbf{Z}\mathbf{w} + \psi(\mathbf{w}, \mathbf{u}). \tag{17}$$

with $\psi(\mathbf{w}, \mathbf{u}) \triangleq \mathbf{w}^T\mathbf{w} + \frac{1}{\sqrt{p}\sigma}\mathbf{1}^T\mathbf{u} - \frac{1}{\sqrt{p}\sigma}\mathbf{w}^T\boldsymbol{\mu}\mathbf{1}^T\mathbf{u} - \frac{1}{\sqrt{p}\sigma}(\mathbf{j}_1^T - \mathbf{j}_0^T)\mathbf{u}\, b$.

Construction of a Sequence of primary optimization problems satisfying the CGMT constraints.

The CGMT requires the feasibility sets of the optimization variables to be compact. This constitutes a major technical difficulty precluding the direct use of the standard CGMT framework. Obviously, this is not satisfied since the feasibility set associated with $\mathbf{w}$ and $\mathbf{u}$ are not compact. To solve this issue, we write $\Phi^{(n)}$ as:

$$\Phi^{(n)} = \inf_{r,B \geq 0} \sup_{\theta \geq 0} \min_{\substack{\mathbf{w} \in \mathbb{R}^p \\ \|\mathbf{w}\|_2 \leq r \\ |b| \leq B}} \max_{\substack{\mathbf{u} \geq 0 \\ \|\mathbf{u}\|_2 \leq \theta}} \frac{1}{\sqrt{p}}\mathbf{u}^T\mathbf{Z}\mathbf{w} + \psi(\mathbf{w}, \mathbf{u}) \tag{18}$$

and denote by $\Phi_{r,B,\theta}^{(n)}$ the following optimization problems:

$$\Phi_{r,B,\theta}^{(n)} = \min_{\substack{\mathbf{w} \in \mathbb{R}^p \\ \|\mathbf{w}\|_2 \leq r \\ |b| \leq B}} \max_{\substack{\mathbf{u} \geq 0 \\ \|\mathbf{u}\|_2 \leq \theta}} \frac{1}{\sqrt{p}}\mathbf{u}^T\mathbf{Z}\mathbf{w} + \psi(\mathbf{w}, \mathbf{u}) \tag{19}$$

Moreover, we also denote by $\Phi_{r,B}$ the following optimization problem:

$$\Phi_{r,B}^{(n)} = \sup_{\theta \geq 0} \min_{\substack{\mathbf{w} \in \mathbb{R}^p \\ \|\mathbf{w}\|_2 \leq r \\ |b| \leq B}} \max_{\substack{\mathbf{u} \geq 0 \\ \|\mathbf{u}\|_2 \leq \theta}} \frac{1}{\sqrt{p}}\mathbf{u}^T\mathbf{Z}\mathbf{w} + \psi(\mathbf{w}, \mathbf{u}) \tag{20}$$

We identified thus a family of primary problems indexed by (r, B, θ), each of which admits the desired format and satisfies the compactness conditions required by the CGMT theorem. Particularly, we can easily distinguish the the bilinear form $\frac{1}{\sqrt{p}}\mathbf{u}^T\mathbf{Z}\mathbf{w}$ and the function $\psi(\mathbf{w}, \mathbf{u})$ which is convex in $\mathbf{w}$ and linear thus concave in $\mathbf{u}$.

Identification of the associated sequence of AO problems.

We associate thus with each one of them the following auxiliary optimization (AO) problem which can be written as:

$$\phi_{r,B,\theta}^{(n)} = \min_{\substack{\mathbf{w}\in\mathbb{R}^p \\ \|\mathbf{w}\|_2\le r \\ |b|\le B}} \max_{\substack{\mathbf{u}\ge 0 \\ \|\mathbf{u}\|_2\le\theta}} \frac{1}{\sqrt{p}}\|\mathbf{w}\|_2\mathbf{g}^T\mathbf{u} - \frac{1}{\sqrt{p}}\|\mathbf{u}\|_2\mathbf{h}^T\mathbf{w} + \psi(\mathbf{w},\mathbf{u})$$

Now that we have identified the (AO) problems, we wish to solve them and infer their asymptotic behavior. To this end, we proceed in two steps. First, we simplify the (AO) problems by reducing them to problems that involve only optimization over a few number of scalars. In doing so, the asymptotic behavior of the AO problems is much simplified and is carried out in the second step. We will see later how the study of the asymptotic behavior of the AO problems sequences allow us to infer the behavior of the original PO in (18)

Simplification of the AO problems. One major step towards the simplification of the (AO) problems is to reduce them to problems that involve only few scalar optimization parameters. Obviously, the objective function of the AO lends itself to this kind of simplification, vector $\mathbf{w}$ appearing only through its norm or its scalar product $\mathbf{w}^T\boldsymbol{\mu}$ and $\mathbf{w}^T\mathbf{h}$. In light of this observation, we decompose $\mathbf{w}$ as:

$$\mathbf{w} = \alpha_1\frac{\boldsymbol{\mu}}{\|\boldsymbol{\mu}\|_2} + \alpha_2\mathbf{w}_\perp,$$

where $\mathbf{w}_\perp$ is a unit norm vector orthogonal to $\boldsymbol{\mu}$. With these notations at hand, we write the (AO) as:

$$\begin{aligned}\phi_{r,B,\theta}^{(n)} = \min_{\substack{\mathbf{w}\in\mathbb{R}^p \\ \|\mathbf{w}\|_2\le r\|\mathbf{u}\|_2\le\theta \\ |b|\le B}} \max_{\substack{\mathbf{u}\ge 0 \\ \|\mathbf{u}\|_2\le\theta}} &\frac{1}{\sqrt{p}}\sqrt{\alpha_1^2+\alpha_2^2}\mathbf{g}^T\mathbf{u} \\ &- \frac{1}{\sqrt{p}}\|\mathbf{u}\|_2\left(\alpha_1\frac{\mathbf{h}^T\boldsymbol{\mu}}{\|\boldsymbol{\mu}\|_2} + \alpha_2\mathbf{h}^T\mathbf{w}_\perp\right) + \alpha_1^2+\alpha_2^2 \\ &+ \frac{1}{\sqrt{p}\sigma}\mathbf{1}^T\mathbf{u} - \alpha_1\|\boldsymbol{\mu}\|_2\frac{1}{\sqrt{p}\sigma}\mathbf{1}^T\mathbf{u} \\ &- \frac{1}{\sqrt{p}\sigma}(\mathbf{j}_1^T - \mathbf{j}_0^T)\mathbf{u}b.\end{aligned}$$

We will now prove that optimizing over $\mathbf{w}$ reduces to optimizing over the set of scalars (α_1, α_2). Note here, that flipping the min-max is not permitted since the objective function is not convex in $\mathbf{w}$ and concave in $\mathbf{u}$ and as a consequence, the use of the Sion's min max theorem is not allowed. One however is tempted to replace $\mathbf{w}_\perp$ by $\text{sign}(\alpha_2)\frac{\mathbf{h}_\perp}{\|\mathbf{h}_\perp\|_2}$, where $\mathbf{h}_\perp$ is the orthogonal projection of $\mathbf{h}$ onto the subspace orthogonal to $\boldsymbol{\mu}$, since this would minimize the objective function for any $\mathbf{u}$. This property, that the vector $\mathbf{w}_\perp = \text{sign}(\alpha_2)\frac{\mathbf{h}_\perp}{\|\mathbf{h}_\perp\|_2}$ minimizes the objective function for any $\mathbf{u}$ allows us, using Lemma 8 proven in the Appendix, to show that $\phi_{r,B,\theta}$ is also given by:

$$\begin{aligned}\phi_{r,B,\theta}^{(n)} = \min_{\substack{\alpha_1,\alpha_2\in\mathbb{R} \\ \alpha_1^2+\alpha_2^2\le r^2 \\ |b|\le B}} \max_{\substack{\mathbf{u}\ge 0 \\ \|\mathbf{u}\|_2\le\theta}} &\frac{1}{\sqrt{p}}\sqrt{\alpha_1^2+\alpha_2^2}\mathbf{g}^T\mathbf{u} \\ &- \frac{1}{\sqrt{p}}\|\mathbf{u}\|_2\left(\alpha_1\frac{\mathbf{h}^T\boldsymbol{\mu}}{\|\boldsymbol{\mu}\|_2} + |\alpha_2|\|\mathbf{h}_\perp\|_2\right) + \alpha_1^2+\alpha_2^2 + \frac{1}{\sqrt{p}\sigma}\mathbf{1}^T\mathbf{u} \\ &- \alpha_1\|\boldsymbol{\mu}\|_2\frac{1}{\sqrt{p}\sigma}\mathbf{1}^T\mathbf{u} - \frac{1}{\sqrt{p}\sigma}(\mathbf{j}_1^T - \mathbf{j}_0^T)\mathbf{u}b.\end{aligned}$$

obtained by replacing $\mathbf{w}_\perp$ by $\text{sign}(\alpha_2)\frac{\mathbf{h}}{\|\mathbf{h}\|_2}$. The (AO) problems are thus simplified in that the minimization over $\mathbf{w}$ is reduced to minimizing over the scalars α_1 and α_2. In the sequel, it is convenient to perform the optimization over $q_0 = \sqrt{\alpha_1^2+\alpha_2^2}$ and α_1. With this notation at hand, $\phi_{r,B,\theta}$ is simplified in (21) shown at bottom of next page, where (21a) follows from decomposing the maximization over $\mathbf{u}$ into the maximization over its direction and its magnitude, (21b) is obtained by applying lemma 11 and (21d) is derived by performing the change of variable $\rho = \frac{\alpha_1}{q_0}$.

The above simplification of the auxiliary problem follows through a deterministic analysis that does not involve any asymptotic approximations. Contrary to the original writing of the AO, this new simplification is more handy towards understanding its asymptotic behavior since it involves only optimizatin over scalar variables. The next step will involve the study of the asymptotic behavior of the (AO) problems.

Asymptotic Behavior of the (AO) problems (Proof of Theorem 1). A well-known fact is that the hard-margin SVM does not always lead to a finite solution, but it is not clear as to when this happens. In the following, we identify through a careful analysis of the sequence of AO problems the condition that guarantees the almost sure feasibility of the hard-margin SVM. Particularly, we will prove that if the condition in Theorem 1 holds true, with probability 1, the hard-margin SVM leads to infinite solution for sufficiently large dimensions n and p. The key idea of the proof relies on showing that the cost of the sequence of AO problems increase at least linearly with θ, that is:

$$\phi_{r,B,\theta}^{(n)} \ge u_n\theta, \tag{22}$$

where u_n is a certain sequence independent of the parameters r, B and θ. We can easily see that if (22) holds, then tending θ to infinity the cost of the sequence of AO problems will grow to infinity when θ grows to infinity. We will prove later how this property translates into the infeasibility of the original PO problem. Before tackling the proof, we shall provide the intuition behind the linear increase of the AO cost with θ. Recall that in this step, we place ourselves in the scenario wherein the SVM is not feasible, and as such its cost is infinite. Due to the difficulty of handling unbounded results, we approximate the dual problem associated with the SVM by a sequence of bounded PO problems, in which the optimization variables and the Lagrangian coefficients are optimized over

an increasing sequence of balls. The constraints of the SVM problem could not be satisfied for all $\mathbf{w}$ in $\mathbb{R}^n$ and thus for all $\mathbf{w}$ constrained on the optimization set of each PO bounded problem. This results in an increase of the cost of each PO (and hence of the AO) with θ, the radius of the θ-balls on which are constrained the lagrangian vector $\mathbf{u}$.

To begin with, we define for fixed $q_0 \in \mathbb{R}_+$, $\rho \in [-1, 1]$ and $\eta \in \mathbb{R}$ function $\hat{D}_H(q_0, \rho, \eta)$ in (23), shown at bottom of this page. We can thus lower-bound $\phi^{(n)}_{r,B,\theta}$ as:

$$\phi^{(n)}_{r,B,\theta} \geq \min_{\substack{0\leq q_0\leq r\\ -1\leq\rho\leq 1\\ \eta\in\mathbb{R}}} q_0^2 + \theta q_0 \left(\hat{D}_H(q_0,\rho,\eta)\right)_+ . \tag{24}$$

where $(a)_+ = \max(a, 0)$. Function $h_n : q_0 \mapsto \hat{D}_H(q_0, \rho, \eta)$ is decreasing in q_0. We may thus find a lower-bound for it by taking its limit as $q_0 \to \infty$. However this would not be helpful, since after replacing this function by this lower-bound, and optimizing over q_0, we find that $\tilde{\phi}_{r,B,\theta} \geq 0$, a fact that does not carry a lot of information. To solve this problem, we need to consider the cases when q_0 is in the vicinity of zero, and when q_0 is sufficiently far away from zero. When q_0 is very close to zero, in a sense that will be defined, we may expect $h_n(q_0) \geq \frac{C}{q_0}$, and hence $q_0 h_n(q_0) \geq C$. This will allow us to prove the sought-for scaling behaviour with respect to θ of $\phi_{r,B,\theta}$ when q_0 is in the vicinity of zero. One can easily see that if $0 \leq q_0 \leq q_U \triangleq \frac{1}{2\sigma \max_{1\leq i\leq n}|g_i|+2\|\boldsymbol{\mu}\|_2}$, then

$$\phi^{(n)}_{r,B,\theta} = \min_{\substack{0\leq q_0\leq r\\ |\alpha_1|\leq q_0\\ |b|\leq B}} \max_{\substack{\mathbf{u}\geq 0\\ \|\mathbf{u}\|_2\leq\theta}} \frac{1}{\sqrt{p}} q_0 \mathbf{g}^T\mathbf{u} - \frac{1}{\sqrt{p}}\|\mathbf{u}\|_2 \left(\alpha_1 \frac{\mathbf{h}^T\boldsymbol{\mu}}{\|\boldsymbol{\mu}\|_2} + \sqrt{q_0^2-\alpha_1^2}\|\mathbf{h}_\perp\|_2\right) + \alpha_1^2 + \alpha_2^2 + \frac{1}{\sqrt{p}\sigma}\mathbf{1}^T\mathbf{u} - \alpha_1\|\boldsymbol{\mu}\|_2\frac{1}{\sqrt{p}\sigma}\mathbf{1}^T\mathbf{u}$$

$$- \frac{1}{\sqrt{p}\sigma}(\mathbf{j}_1^T - \mathbf{j}_0^T)\mathbf{u}b, \tag{21a}$$

$$= \min_{\substack{q_0\leq r\\ |\alpha_1|\leq q_0\\ |b|\leq B}} \max_{\theta\geq m\geq 0} \max_{\|\mathbf{u}\|_2=m} \mathbf{u}^T\left(q_0\mathbf{g} + \frac{1}{\sqrt{p}\sigma}\mathbf{1} - \alpha_1\|\mathbf{u}\|_2\frac{1}{\sqrt{p}\sigma}\mathbf{1} - \frac{1}{\sqrt{p}\sigma}(\mathbf{j}_1-\mathbf{j}_0)b\right) + q_0^2$$

$$- m\left(\frac{1}{\sqrt{p}}\alpha_1\frac{\mathbf{h}^T\boldsymbol{\mu}}{\|\boldsymbol{\mu}\|_2} + \sqrt{q_0^2-\alpha_1^2}\frac{1}{\sqrt{p}}\|\mathbf{h}_\perp\|_2\right) \tag{21b}$$

$$= \min_{\substack{0\leq q_0\leq r\\ |\alpha_1|\leq q_0\\ |b|\leq B}} \max_{\theta\geq m\geq 0} q_0^2 + m\sqrt{\frac{1}{p}\sum_{i\in\mathcal{C}_1}\left(q_0 g_i + \frac{1-\alpha_1\|\boldsymbol{\mu}\|_2 - b}{\sigma}\right)_+^2 + \frac{1}{p}\sum_{i\in\mathcal{C}_0}\left(q_0 g_i + \frac{1-\alpha_1\|\boldsymbol{\mu}\|_2 + b}{\sigma}\right)_+^2}$$

$$- m\left(\frac{1}{\sqrt{p}}\alpha_1\frac{\mathbf{h}^T\boldsymbol{\mu}}{\|\boldsymbol{\mu}\|_2} + \sqrt{q_0^2-\alpha_1^2}\frac{1}{\sqrt{p}}\|\mathbf{h}_\perp\|_2\right) \tag{21c}$$

$$= \min_{\substack{0\leq q_0\leq r\\ -1\leq\rho\leq 1\\ |b|\leq B}} q_0^2 + \theta q_0\left(\sqrt{\frac{1}{p}\sum_{i\in\mathcal{C}_1}\left(g_i + \frac{1}{q_0\sigma} - \frac{\rho\|\boldsymbol{\mu}\|_2}{\sigma} - \frac{b}{q_0\sigma}\right)_+^2 + \frac{1}{p}\sum_{i\in\mathcal{C}_0}\left(g_i + \frac{1}{q_0\sigma} - \rho\frac{\|\boldsymbol{\mu}\|_2}{\sigma} + \frac{b}{\sigma q_0}\right)_+^2}\right.$$

$$\left. - \left(\frac{1}{\sqrt{p}}\rho\frac{\mathbf{h}^T\boldsymbol{\mu}}{\|\boldsymbol{\mu}\|_2} + \sqrt{1-\rho^2}\frac{1}{\sqrt{p}}\|\mathbf{h}_\perp\|_2\right)\right)_+ \tag{21d}$$

$$= \min_{\substack{0\leq q_0\leq r\\ |\alpha_1|\leq q_0\\ |b|\leq B}} q_0^2 + \theta q_0\left(\hat{D}_H(q_0,\rho,\frac{b}{\sigma q_0})\right)_+ \tag{21e}$$

$$\hat{D}_H : (q_0,\rho,\eta) \mapsto \sqrt{\frac{1}{p}\sum_{i\in\mathcal{C}_1}\left(g_i + \frac{1}{q_0\sigma} - \frac{\rho\|\boldsymbol{\mu}\|_2}{\sigma} - \eta\right)_+^2 + \frac{1}{p}\sum_{i\in\mathcal{C}_0}\left(g_i + \frac{1}{q_0\sigma} - \rho\frac{\|\boldsymbol{\mu}\|_2}{\sigma} + \eta\right)_+^2}$$
$$- \left(\frac{1}{\sqrt{p}}\rho\frac{|\mathbf{h}^T\boldsymbol{\mu}|}{\|\boldsymbol{\mu}\|_2} + \sqrt{1-\rho^2}\frac{1}{\sqrt{p}}\|\mathbf{h}_\perp\|_2\right) \tag{23}$$

$g_i + \frac{1}{q_0\sigma} - \frac{\rho\|\boldsymbol{\mu}_2\|}{\sigma} \geq \frac{1}{2q_0\sigma}$, thereby implying that:

$$\min_{\substack{0\leq q_0\leq q_U\\ \eta\in\mathbb{R}\\ |\rho|\leq 1}} \theta q_0 \left(\hat{D}_H(q_0,\rho,\eta)\right) \tag{25}$$

$$\geq \min_{\substack{0\leq q_0\leq q_U\\ \eta\in\mathbb{R}\\ |\rho|\leq 1}} \theta q_0 \left(\sqrt{\frac{n_1}{p}\left(\frac{1}{2q_0\sigma}-\eta\right)_+^2 + \frac{n_0}{p}\left(\frac{1}{2q_0\sigma}+\eta\right)_+^2}\right. \tag{26}$$

$$\left. - \frac{1}{\sqrt{p}}\frac{|\mathbf{h}^T\boldsymbol{\mu}|}{\|\boldsymbol{\mu}\|_2} - \frac{1}{\sqrt{p}}\|\mathbf{h}_\perp\|_2\right)_+$$

$$\overset{(a)}{\geq} \theta\frac{1}{\sigma}\sqrt{\frac{n_1 n_0}{np}} - \theta q_U\left(\frac{1}{\sqrt{p}}\frac{|\mathbf{h}^T\boldsymbol{\mu}|}{\|\boldsymbol{\mu}\|_2} + \frac{1}{\sqrt{p}}\|\mathbf{h}_\perp\|\right) \ . \tag{27}$$

where (a) follows from performing the optimization over $\eta \in \mathbb{R}$. Since $q_U \leq \frac{1}{2\sigma|\max_{1\leq i\leq n} g_i| + 2\|\boldsymbol{\mu}\|_2}$, and $|\frac{\max_{1\leq i\leq n} g_i}{\sqrt{2\log n}}| \xrightarrow{\text{a.s.}} 1$, q_U converges to 0 almost surely. Hence, with probability 1 as n and p are sufficiently large, $q_U \leq \frac{1}{2\sigma}\sqrt{\frac{n_1 n_0}{np}}$. We have thus proved that with probability 1, for large n and p,

$$\min_{\substack{0\leq q_0\leq q_U\\ \eta\in\mathbb{R}\\ |\rho|\leq 1}} \theta q_0 \left(\hat{D}_H(q_0,\rho,\eta)\right) \geq \theta\frac{1}{2\sigma}\sqrt{\frac{n_1 n_0}{np}} \ . \tag{28}$$

We will now consider the optimization of $\phi^{(n)}_{r,B,\theta}$ when $q_U \leq q_0 \leq r$.[2] Using the fact that function h_n is decreasing in q_0, we obtain:

$$\min_{\substack{q_U\leq q_0\leq r\\ |\rho|\leq 1\\ \eta\in\mathbb{R}}} \theta q_0 \left(\hat{D}_H(q_0,\rho,\eta)\right)_+ \geq \min_{\substack{q_U\leq q_0\leq r\\ |\rho|\leq 1\\ \eta\in\mathbb{R}}} \theta q_0 \left(\ell_n(\rho,\eta)\right)_+ \tag{29}$$

where $\ell_n : [-1,1]\times\mathbb{R}$ with

$$\ell_n(\rho,\eta) = \sqrt{\frac{1}{p}\sum_{i\in\mathcal{C}_1}\left(g_i - \frac{\rho\|\boldsymbol{\mu}\|}{\sigma} - \eta\right)_+^2 + \frac{1}{p}\sum_{i\in\mathcal{C}_0}\left(g_i - \frac{\rho\|\boldsymbol{\mu}\|}{\sigma} + \eta\right)_+^2}$$

$$- \frac{1}{\sqrt{p}}\frac{\mathbf{h}^T\boldsymbol{\mu}}{\|\boldsymbol{\mu}\|} - \sqrt{\frac{1-\rho^2}{p}}\|\mathbf{h}_\perp\|. \tag{30}$$

It is easy to see that ℓ_n is jointly convex function in its arguments (ρ,η) and converges almost surely to

$$\overline{\ell} : (\rho,\eta) \mapsto \sqrt{\delta\pi_1\mathbb{E}\left(G - \frac{\rho\mu}{\sigma} - \eta\right)_+^2 + \delta\pi_0\mathbb{E}\left(G - \frac{\rho\mu}{\sigma} + \eta\right)_+^2} - \sqrt{1-\rho^2}$$

where $G \sim \mathcal{N}(0,1)$. Since $\lim_{\eta\to\infty}\overline{\ell}(\rho,\eta) = \infty$, using Lemma 11 and Lemma 10 in [10], we obtain:

$$\min_{\eta\in\mathbb{R}} \ell_n(\rho,\eta) \xrightarrow{\text{a.s.}} \min_{\eta\in\mathbb{R}} \overline{\ell}(\rho,\eta) \ .$$

Moreover, for $\rho \in [-1,1]$, expressing the first order conditions with respect to η, the optimum $\eta^\star(\rho)$ is the solution to the following equation:

$$\eta = \frac{\frac{\pi_1}{\sqrt{2\pi}}\int_{\frac{\rho\mu}{\sigma}+\eta}^{\infty}(x - \frac{\rho\mu}{\sigma})Dx + \frac{\pi_0}{\sqrt{2\pi}}\int_{\frac{\rho\mu}{\sigma}-\eta}^{\infty}(\frac{\rho\mu}{\sigma} - x)Dx}{\frac{\pi_1}{\sqrt{2\pi}}\int_{\frac{\rho\mu}{\sigma}+\eta}^{\infty}Dx + \frac{\pi_0}{\sqrt{2\pi}}\int_{\frac{\rho\mu}{\sigma}-\eta}^{\infty}Dx} \tag{31}$$

where $Dx = e^{-x^2/2}dx$. It is easy to see that the solution of (31) is unique. This is because function

$$\eta :\mapsto \eta\left(\pi_1\int_{\frac{\rho\mu}{\sigma}+\eta}^{\infty}Dx + \pi_0\int_{\frac{\rho\mu}{\sigma}-\eta}^{\infty}Dx\right) - \pi_1\int_{\frac{\rho\mu}{\sigma}+\eta}^{\infty}(x - \frac{\rho\mu}{\sigma})Dx - \pi_0\int_{\frac{\rho\mu}{\sigma}-\eta}^{\infty}\left(\frac{\rho\mu}{\sigma} - x\right)Dx \tag{32}$$

is decreasing with limits ∞ and $-\infty$ when $\eta\to-\infty$ and $\eta\to\infty$ respectively. Using Lemma 10 in the Appendix, Function $\rho\mapsto\min_{\eta\in\mathbb{R}}\ell_n(\eta,\rho)$ is convex in ρ. Since the convergence of convex functions is uniform over compacts, from Theorem 2.1 in [24], we have:

$$\min_{\substack{\eta\in\mathbb{R}\\ -1\leq\rho\leq 1}} \ell_n(\rho,\eta) \xrightarrow{\text{a.s.}} \min_{\substack{\eta\in\mathbb{R}\\ -1\leq\rho\leq 1}} \overline{\ell}(\rho,\eta) \ . \tag{33}$$

If Condition (5) is satisfied, $\overline{\ell} \triangleq \min_{\substack{\eta\in\mathbb{R}\\ -1\leq\rho\leq 1}} \overline{\ell}(\rho,\eta) > 0$, which implies that for all $\epsilon > 0$, and sufficiently large n and p we have with probability 1,

$$\min_{\substack{q_U\leq q_0\leq r\\ |\rho|\leq 1}} \theta q_0\hat{D}_H(q_0,\rho,\eta) \geq \theta q_U\left(\overline{\ell} - \epsilon\right) \ . \tag{34}$$

Taking $\epsilon = \frac{\overline{\ell}}{2}$ and combining (27) and (34) leads to:

$$\min_{\substack{q_U\leq q_0\leq r\\ |\rho|\leq 1}} \theta q_0\hat{D}_H(q_0,\rho,\eta) \geq \theta q_U\frac{\overline{\ell}}{2} \tag{35}$$

almost surely for enough large n and p. Combining (28) and (35), yields

$$\phi^{(n)}_{r,B,\theta} \geq \theta\min\left(\frac{1}{2\sigma}\sqrt{\frac{n_1 n_0}{np}}, q_U\frac{\overline{\ell}}{2}\right)$$

As q_U converges almost surely to zero, for sufficiently large n and p, $\min(\frac{1}{2\sigma}\sqrt{\frac{n_1 n_0}{np}}, q_U\frac{\overline{\ell}}{2}) = q_U\frac{\overline{\ell}}{2}$. Hence for sufficiently large n and p, we obtain

$$\phi^{(n)}_{r,B,\theta} \geq \theta q_U\frac{\overline{\ell}}{2} \tag{36}$$

which establishes (22). With the above inequality (36) at hand, we are now ready to establish Theorem 1. First, we shall bring to the reader's attention that the order of magnitude of n and p

[2]Without loss of generality, we assume that $r \geq q_U$.

above which (35) holds is independent of r, B and θ. It entails from this that the set:

$$\mathcal{E} \triangleq \left\{ \cup_{k=1}^{\infty} \cap_{m=1}^{\infty} \left\{ \phi_{k,k,m}^{(n)} \geq m q_U \frac{\overline{\ell}}{2} \right\} \right. \\ \left. \text{for } n \text{ and } p \text{ sufficiently large} \right\} \tag{37}$$

verifies $\mathbb{P}[\mathcal{E}] = 1$. Let us now consider the optimal value $\Phi^{(n)}$ of the primary optimization problem and illustrate how the characterization of the auxiliary problem allows to ensure that under the setting of Theorem 1, $\Phi = \infty$ for n and p sufficiently large. One way to prove this is to show that for all $x > 0$, $\mathbb{P}[\Phi^{(n)} \leq x, \quad \text{for} \quad n, p \quad \text{sufficiently large}] = 0$. From (18), if $\Phi^{(n)} \neq \infty$, for $\epsilon > 0$ sufficiently small, there exists $k \in \mathbb{N}$ such that $\Phi^{(n)} \geq \Phi_{k,k}^{(n)} - \epsilon$. Hence,

$$\mathbb{P}\left[\left\{\Phi^{(n)} \leq x\right\}\right] \leq \mathbb{P}\left[\cup_{k=1}^{\infty} \left\{\Phi_{k,k}^{(n)} \leq x + \epsilon\right\}\right] \\ \leq \mathbb{P}\left[\cup_{k=1}^{\infty} \left\{\cap_{m=1}^{\infty} \left\{\Phi_{k,k,m}^{(n)} \leq x + \epsilon\right\}\right\}\right],$$

For $m \in \mathbb{N}^{\star}$, the events $\mathcal{E}_k = \{\cap_{m=1}^{\infty}\{\Phi_{k,k,m}^{(n)} \leq (x+\epsilon)\}\}$ forms an increasing sequence of events, thus:

$$\mathbb{P}\left[\cup_{k=1}^{\infty} \mathcal{E}_k\right] = \lim_{k \to \infty} \mathbb{P}\left[\mathcal{E}_k\right].$$

Similarly, as $\Phi_{k,k,m}^{(n)} \geq \Phi_{k,k,m-1}^{(n)}$, for $k \in \mathbb{N}^{\star}$, the sequence of events, $\mathcal{E}_{k,m} = \{\Phi_{k,k,m}^{(n)} \leq x\}$ is decreasing, thus:

$$\mathbb{P}\left[\cap_{m=1}^{\infty} \mathcal{E}_{k,m}\right] = \lim_{m \to \infty} \mathbb{P}\left[\mathcal{E}_{k,m}\right]$$

We thus obtain:

$$\mathbb{P}\left[\Phi^{(n)} \leq x\right] \leq \lim_{k \to \infty} \lim_{m \to \infty} \mathbb{P}\left[\left\{\Phi_{k,k,m}^{(n)} \leq x + \epsilon\right\}\right]$$

From the CGMT theorem, we have:

$$\mathbb{P}\left[\Phi_{k,k,m}^{(n)} \leq x + \epsilon\right] \leq 2\mathbb{P}\left[\phi_{k,k,m}^{(n)} \leq x + \epsilon\right].$$

Hence,

$$\lim_{k \to \infty} \lim_{m \to \infty} \mathbb{P}\left[\Phi_{k,k,m}^{(n)} \leq x + \epsilon\right] \leq \lim_{k \to \infty} \lim_{m \to \infty} 2\mathbb{P} \\ \times \left[\phi_{k,k,m}^{(n)} \leq x + \epsilon\right] \tag{38}$$

$$= 2\mathbb{P}\left[\cup_{k=1}^{\infty} \left\{\cap_{m=1}^{\infty} \{\phi_{k,k,m}^{(n)} \leq x + \epsilon\}\right\}\right] \tag{39}$$

Using the fact that $\mathbb{P}[\mathcal{E}] = 1$ with $\mathcal{E}$ given by (37), the event $A_n = \{\cup_{k=1}^{\infty}\{\cap_{m=1}^{\infty}\{\phi_{k,k,m}^{(n)} \leq x + \epsilon\}\}\}$ does not occur infinitely often, or in other words $\mathbb{P}[A_n, i.o] = 0$. Since (A_n) are independent, each event being generated by independent vectors $\mathbf{g}$ and $\mathbf{h}$ in $\mathbb{R}^{n \times 1}$ and $\mathbb{R}^{p \times 1}$, the converse of Borel-Cantelli lemma implies that $\sum_{n=1}^{\infty} \mathbb{P}(A_n) < \infty$. Therefore,

$$\sum_{n=1}^{\infty} \mathbb{P}\left[\Phi^{(n)} \leq x\right] < \infty \ .$$

Using Borel-Cantelli Lemma, we deduce that for any x,

$$\mathbb{P}\left[\Phi^{(n)} \leq x, \text{i.o}\right] = 0 \ .$$

This implies that $\{\Phi^{(n)} = \infty\}$ occurs infinitely often.

C. ASYMPTOTIC BEHAVIOR OF THE (AO) PROBLEMS (PROOF OF THEOREM 2)

Proof of the uniqueness of $q_{0,H}^{\star}$, the zero of function β and the minimizers of $D_H(q_{0,H}^{\star}, \rho_H^{\star}, \eta_H^{\star})$. To begin with, we check first that function β has a unique zero $q_{0,H}^{\star}$. Towards this end, note that the $\eta^{\star}(\rho, q_0)$ minimizing $D_H(q_0, \rho, \eta)$ for fixed q_0 and ρ should be solution to (40) shown at bottom of this page, in η. Such an equation admits a unique solution because function

$$\eta \mapsto \eta \left[\pi_1 \int_{\frac{\rho\mu}{\sigma} - \frac{1}{q_0} + \eta}^{\infty} Dx + \pi_0 \int_{\frac{\rho\mu}{\sigma} - \frac{1}{q_0} - \eta}^{\infty} Dx\right] \\ - \left[\pi_1 \int_{\frac{\rho\mu}{\sigma} - \frac{1}{q_0} + \eta}^{\infty} \left(x - \frac{\rho\mu}{\sigma} + \frac{1}{q_0}\right) Dx \right. \\ \left. - \pi_0 \int_{\frac{\rho\mu}{\sigma} - \frac{1}{q_0\sigma} - \eta}^{\infty} \left(x - \frac{\rho\mu}{\sigma} + \frac{1}{q_0\sigma}\right) Dx\right]$$

is an increasing function with limits $-\infty$ and ∞ when $\eta \to -\infty$ and $\eta \to \infty$. Moreover, $(\rho, q_0) \mapsto \eta^{\star}(\rho, q_0)$ is a continuous function. From the Maximum Theorem [25, Theorem 9.17], function $q_0 \mapsto \min_{-1 \leq \rho \leq 1} D_H(q_0, \rho, \eta^{\star}(\rho, q_0))$ is continuous. It tends to ∞ as $q_0 \to 0^+$ and to

$$\min_{-1 \leq \rho \leq 1} \left(\delta\pi_1 \mathbb{E}\left[\left(\left(G - \frac{\rho\mu}{\sigma} - \eta^{\star}(\rho)\right)_+\right)^2\right] \right. \\ \left. + \delta\pi_0 \mathbb{E}\left[\left(\left(G - \frac{\rho\mu}{\sigma} + \eta^{\star}(\rho)\right)_+\right)^2\right]\right)^{\frac{1}{2}} - \sqrt{1-\rho^2} < 0 \tag{41}$$

when q tends to ∞. There exists thus $q_{0,H}^{\star}$ such that $\min_{\substack{-1 \leq \rho \leq 1 \\ \eta \in \mathbb{R}}} D_H(q_{0,H}^{\star}, \rho, \eta) = 0$. We will prove that necessarily such a $q_{0,H}^{\star}$ is unique. Assume that there exists two solutions $q_{01}^{\star}$ and $q_{02}^{\star}$ such that $\min_{\substack{-1 \leq \rho \leq 1 \\ \eta \in \mathbb{R}}} D_H(q_{01}^{\star}, \rho, \eta) = \min_{\substack{-1 \leq \rho \leq 1 \\ \eta \in \mathbb{R}}} D_H(q_{02}^{\star}, \rho, \eta) = 0$. Let $(\rho_1^{\star}, \rho_2^{\star})$ and $(\eta_1^{\star}(\rho_1^{\star}, q_{01}^{\star}), \eta_2^{\star}(\rho_2, q_{02}^{\star}))$ such that

$$\eta = \frac{\pi_1 \mathbb{E}\left[(G - \frac{\rho\mu}{\sigma} + \frac{1}{q_0})\mathbf{1}_{\{G \geq \frac{\rho\mu}{\sigma} - \frac{1}{q_0} + \eta\}}\right] - \pi_0 \mathbb{E}\left[(G - \frac{\rho\mu}{\sigma} + \frac{1}{q_0})\mathbf{1}_{\{G \geq \frac{\rho\mu}{\sigma} - \frac{1}{q_0} - \eta\}}\right]}{\pi_1 \mathbb{P}\left[G \geq \frac{\rho\mu}{\sigma} - \frac{1}{q_0} + \eta\right] + \pi_0 \mathbb{P}\left[G \geq \frac{\rho\mu}{\sigma} - \frac{1}{q_0} - \eta\right]} \tag{40}$$

$\min_{\substack{-1\le\rho\le 1\\ \eta\in\mathbb{R}}} D_H(q_{01}^\star,\rho,\eta) = D_H(q_{01}^\star,\rho_1^\star,\eta_1^\star(\rho_1^\star,q_{01}^\star))$ and $\min_{\substack{-1\le\rho\le 1\\ \eta\in\mathbb{R}}} D_H(q_{02}^\star,\rho,\eta) = D_H(q_{02}^\star,\rho_2^\star,\eta_2^\star(\rho_2^\star,q_{02}^\star))$.
Hence,

$$\begin{aligned} 0 &= D_H(q_{02}^\star,\rho_2^\star,\eta^\star(q_{02}^\star,\rho_2^\star)) \\ &= D_H(q_{01}^\star,\rho_1^\star,\eta^\star(q_{01}^\star,\rho_1^\star)) \quad (42)\\ &\le D_H(q_{01}^\star,\rho_2^\star,\eta^\star(q_{02}^\star,\rho_2^\star)) \quad (43)\end{aligned}$$

Since for any $\eta\in\mathbb{R}$ and $\rho\in[-1,1]$, $q\mapsto D_H(q,\rho,\eta)$ is decreasing, $q_{01}^\star\ge q_{02}^\star$. The same reasoning leads also to $q_{01}^\star\le q_{02}^\star$. Hence $q_{01}^\star=q_{02}^\star$. We will prove now that there exists unique $\rho_H^\star$ and $\eta^\star(\rho_H^\star,q_{0,H}^\star)$ such that:

$$\min_{\substack{-1\le\rho\le 1\\ \eta\in\mathbb{R}}} D_H(q_{0,H}^\star,\rho,\eta) = D_H(q_{0,H}^\star,\ \eta^\star(\rho_H^\star,q_{0,H}^\star)) = 0.$$

Function

$$\varphi:(\rho,\eta)\mapsto \left(\delta\pi_1\mathbb{E}\left[\left(\left(G-\frac{\rho\mu}{\sigma}+\frac{1}{\rho q_{0,H}^\star}-\eta\right)_+\right)^2\right] +\delta\pi_0\mathbb{E}\left[\left(\left(G-\frac{\rho\mu}{\sigma}+\frac{1}{\rho q_{0,H}^\star}+\eta\right)_+\right)^2\right]\right)^{\frac{1}{2}}, \quad (44)$$

is jointly convex in its arguments. Hence, $\rho\mapsto\min_{\eta\in\mathbb{R}}\varphi(\rho,\eta)$ is convex in $[-1,1]$. As $\rho\mapsto-\sqrt{1-\rho^2}$ is strictly convex in $[-1,1]$, then $\rho\mapsto\min_{\eta\in\mathbb{R}} D_H(q_{0,H}^\star,\rho,\eta)$ is strictly convex in $[-1,1]$. Assume that there exists $\rho_H^\star$ and $\tilde\rho^\star$ in $[-1,1]$ such that:

$$\begin{aligned}\min_{\eta\in\mathbb{R}} D_H(q_{0,H}^\star,\rho_H^\star,\eta) &= \min_{\eta\in\mathbb{R}} D_H(q_{0,H}^\star,\tilde\rho^\star,\eta) \quad (45)\\ &= \min_{\substack{-1\le\rho\le 1\\ \eta\in\mathbb{R}}} D_H(q_{0,H}^\star,\rho,\eta)=0. \quad (46)\end{aligned}$$

Let $\lambda\in(0,1)$. Assume $\rho_H^\star\neq\tilde\rho^\star$. Then

$$\begin{aligned}&\min_{\eta\in\mathbb{R}} D_H(q_{0,H}^\star,\lambda\rho_H^\star+(1-\lambda)\tilde\rho^\star,\eta) \quad (47)\\ &< \lambda\min_{\eta\in\mathbb{R}} D_H(q_{0,H}^\star,\rho_H^\star,\eta)+(1-\lambda)\min_{\eta\in\mathbb{R}} D_H(q_{0,H}^\star,\tilde\rho^\star,\eta)=0 \quad (48)\end{aligned}$$

We obtain thus a contradiction, since $0=\min_{\substack{\eta\in\mathbb{R}\\ \rho\in[-1,1]}} D_H(q_{0,H}^\star,\rho,\eta)$. Hence the uniqueness of the minimizer $\rho_H^\star$. Combining all the above results shows the uniqueness of $q_{0,H}^\star,\eta^\star$ and $\rho_H^\star$. The uniqueness of $q_{0,H}^\star$ and that of the minimizers of function $(\rho,\eta)\mapsto D_H(q_{0,H}^\star,\rho,\eta)$ will allow us to satisfy the requirements (i), (ii) and (iii) of Theorem 7 when applied to elements of the sequence of (PO) problems. This will be made more clear in the sequel.

Proof of feasibility of the (PO) problem. To begin with, we prove that if $\delta<\delta_\star$, then, for there exists a positive constant C such that:

$$\mathbb{P}\left[\Phi^{(n)}>C,\ \text{i.o}\right]=0. \quad (49)$$

For that, it suffices to check that for k_0 sufficiently large and for any small $1>\epsilon>0$,

$$\mathbb{P}\left[\cup_{m=1}^\infty\left\{\phi_{k_0,k_0,m}^{(n)}\ge C+1-\epsilon\right\},\text{i.o.}\right]=0. \quad (50)$$

Indeed, assume that (50) is satisfied and let us prove that it implies 49. As $\Phi^{(n)}\le\Phi_{k_0,k_0}^{(n)}$,

$$\mathbb{P}\left[\Phi^{(n)}>C\right]\le\mathbb{P}\left[\Phi_{k_0,k_0}^{(n)}>C\right]$$

Recall the fact that $\Phi_{k_0,k_0}^{(n)}=\lim_{m\to\infty}\Phi_{k_0,k_0,m}^{(n)}$. Then, if $\Phi_{k_0,k_0}^{(n)}\neq\infty$, then for any $\epsilon>0$, there exists m_0 sufficiently large such that for any $m\ge m_0$, $\Phi^{(n)}\le\Phi_{k_0,k_0,m}^{(n)}+\epsilon$. In case $\Phi_{k_0,k_0}^{(n)}=\infty$, then necessarily $\Phi_{k_0,k_0,m}^{(n)}\ge C+1-\epsilon$ for sufficiently large m. So in both cases, $\Phi_{k_0,k_0}=\infty$ or $\Phi_{k_0,k_0}\neq\infty$, it holds true that

$$\Phi^{(n)}>C\Rightarrow\Phi_{k_0,k_0,m}^{(n)}\ge C+1-\epsilon\ \text{ for } m \text{ sufficiently large}$$

Hence,

$$\begin{aligned}\mathbb{P}\left[\Phi^{(n)}>C\right] &\le \mathbb{P}\left[\cup_{m=1}^\infty\left\{\Phi_{k_0,k_0,m}^{(n)}\ge C+1-\epsilon\right\}\right] \quad (51)\\ &= \lim_{m\to\infty}\mathbb{P}\left[\left\{\Phi_{k_0,k_0,m}^{(n)}\ge C+1-\epsilon\right\}\right] \quad (52)\\ &\le \lim_{m\to\infty}2\mathbb{P}\left[\left\{\phi_{k_0,k_0,m}^{(n)}\ge C+1-\epsilon\right\}\right] \quad (53)\\ &= 2\mathbb{P}\left[\cup_{m=1}^\infty\left\{\phi_{k_0,k_0,m}^{(n)}\ge C+1-\epsilon\right\}\right] \quad (54)\end{aligned}$$

In a similar way as previously, based on the converse of Borel-Cantelli lemma, we deduce that $\sum_{n=1}^\infty\mathbb{P}[\cup_{m=1}^\infty\{\phi_{k_0,k_0,m}^{(n)}\ge C+1-\epsilon\}]<\infty$ and hence $\sum_{n=1}^\infty\mathbb{P}[\Phi^{(n)}>C]<\infty$, which implies (49).

Proof of (50) To show the boundedness of the optimal PO cost, it suffices thus to establish (50). Towards this end, let Q be such that:

$$\min_{\substack{\eta\in\mathbb{R}\\ \rho\in[-1,1]}} D_H(Q,\rho,\eta)<0.$$

Note that such a Q exists since as already shown in (41), $\lim_{q\to\infty}\min_{\substack{\eta\in\mathbb{R}\\ \rho\in[-1,1]}} D_H(q,\rho,\eta)<0$. Let $\rho^\star(Q)$ and $\eta^\star(Q)$ be such that $D_H(Q,\rho^\star(Q),\eta^\star(Q))=\min_{\substack{\eta\in\mathbb{R}\\ \rho\in[-1,1]}} D_H(q,\rho,\eta)$. Function $Q\mapsto\hat D_H(Q,\rho^\star(Q),\eta^\star(Q))$ converges point-wise to $D_H(Q,\rho^\star(Q),\eta^\star(Q))<0$. Hence,

$$\hat D_H(Q,\rho^\star(Q),\eta^\star(Q))\xrightarrow{\text{a.s.}}D_H(Q,\rho^\star(Q),\eta^\star(Q))<0.$$

For n and p sufficiently large, we thus have:

$$\hat D_H(Q,\rho^\star(Q),\eta^\star(Q))<0. \quad (55)$$

Now, take integer k greater than $\max(\lceil Q \rceil, \eta^\star(Q)Q\sigma)$. Recalling that:

$$\phi_{k,k,m}^{(n)} = \min_{\substack{0\le q\le k\\ |\rho|\le 1\\ |b|\le k}} q^2 + m.q\left(\hat{D}_H\left(q,\rho,\frac{b}{q\sigma}\right)\right)_+ \tag{56}$$

$$\le Q^2 + m.Q(\hat{D}_H(Q,\rho^\star(Q),\eta^\star(Q)))_+ \tag{57}$$

From (55), we thus have for n and p sufficiently large, $\phi_{k,k,m}^{(n)} \le Q^2$, thereby establishing (50).

From the convergence of the AOs cost back to the convergence of the PO cost. To prove theorem 2, it suffices to establish the following convergences for any $\epsilon > 0$ sufficiently small and some integer k_0 sufficiently large:

$$\mathbb{P}\left[\cup_{k=k_0}^{\infty} \lim_{m\to\infty} \phi_{k,k,m} \le (q_{0,H}^\star)^2 - \epsilon, \text{i.o.}\right] = 0, \tag{58}$$

$$\mathbb{P}\left[\lim_{m\to\infty} \phi_{k,k,m} \ge (q_{0,H}^\star)^2 + \epsilon, \text{i.o.}\right] = 0, \quad \forall k \ge k_0. \tag{59}$$

Indeed, assume that (58) and (59) hold true, and let us prove that they translate into

$$\mathbb{P}\left[\Phi^{(n)} \le (q_{0,H}^\star)^2 - \epsilon, \text{ i.o.}\right] = 0 \tag{60}$$

$$\mathbb{P}\left[\Phi^{(n)} \ge (q_{0,H}^\star)^2 + \epsilon, \text{ i.o.}\right] = 0 \tag{61}$$

the combination of both of which leads to $\lim_{n\to\infty} \Phi^{(n)} \to (q_{0,H}^\star)^2$ almost surely.

Proof of (60). For $\epsilon > 0$ sufficiently small, there exists $k \in \mathbb{N}$ such that $\Phi^{(n)} \ge \Phi_{k,k}^{(n)} - \frac{\epsilon}{2}$. Hence,

$$\mathbb{P}\left[\Phi^{(n)} \le (q_{0,H}^\star)^2 - \epsilon\right] \le \mathbb{P}\left[\cup_{k=1}^{\infty}\left\{\Phi_{k,k}^{(n)} \le (q_{0,H}^\star)^2 - \frac{\epsilon}{2}\right\}\right]$$

$$\le \mathbb{P}\left[\cup_{k=1}^{\infty}\cap_{m=1}^{\infty}\left\{\Phi_{k,k,m}^{(n)} \le (q_{0,H}^\star)^2 - \frac{\epsilon}{2}\right\}\right]$$

$$\overset{(a)}{=} \lim_{k\to\infty}\lim_{m\to\infty} \mathbb{P}\left[\left\{\Phi_{k,k,m}^{(n)} \le (q_{0,H}^\star)^2 - \frac{\epsilon}{2}\right\}\right]$$

where (a) follows from the fact that the sequence of events $\{\cap_{m=1}^{\infty}\{\Phi_{k,k,m}^{(n)} \le (q_{0,H}^\star)^2 - \frac{\epsilon}{2}\}\}_{k\in\mathbb{N}^\star}$ forms an increasing sequence of events, while $\{\Phi_{k,k,m}^{(n)} \le (q_{0,H}^\star)^2 - \frac{\epsilon}{2}\}_{m\in\mathbb{N}^\star}$ forms a decreasing sequence of events. Using the CGMT Theorem, we have:

$$\mathbb{P}\left[\left\{\Phi_{k,k,m}^{(n)} \le (q_{0,H}^\star)^2 - \frac{\epsilon}{2}\right\}\right] \le 2\mathbb{P}\left[\left\{\phi_{k,k,m}^{(n)} \le (q_{0,H}^\star)^2 - \frac{\epsilon}{2}\right\}\right]$$

Hence

$$\lim_{k\to\infty}\lim_{m\to\infty} \mathbb{P}\left[\left\{\Phi_{k,k,m}^{(n)} \le (q_{0,H}^\star)^2 - \frac{\epsilon}{2}\right\}\right]$$

$$\le \lim_{k\to\infty}\lim_{m\to\infty} 2\mathbb{P}\left[\left\{\phi_{k,k,m}^{(n)} \le (q_{0,H}^\star)^2 - \frac{\epsilon}{2}\right\}\right]$$

$$= 2\mathbb{P}\left[\cup_{k=1}^{\infty}\cap_{m=1}^{\infty}\left\{\phi_{k,k,m}^{(n)} \le (q_{0,H}^\star)^2 - \frac{\epsilon}{2}\right\}\right]$$

Let k_0 be an integer chosen such that $k_0 > q_{0,H}^\star$. Hence,

$$\mathbb{P}\left[\cup_{k=1}^{\infty}\cap_{m=1}^{\infty}\left\{\phi_{k,k,m}^{(n)} \le (q_{0,H}^\star)^2 - \frac{\epsilon}{2}\right\}\right]$$

$$\le \mathbb{P}\left[\cup_{k=k_0}^{\infty}\cap_{m=1}^{\infty}\left\{\phi_{k,k,m}^{(n)} \le (q_{0,H}^\star)^2 - \frac{\epsilon}{2}\right\}\right]$$

$$= \mathbb{P}\left[\cup_{k=k_0}^{\infty}\forall m, \phi_{k,k,m}^{(n)} \le (q_{0,H}^\star)^2 - \frac{\epsilon}{2}\right]$$

$$\le \mathbb{P}\left[\cup_{k=k_0}^{\infty}\lim_{m\to\infty}\phi_{k,k,m}^{(n)} \le (q_{0,H}^\star)^2 - \frac{\epsilon}{2}\right]$$

Letting $B_n^{(AO)} \triangleq \cup_{k=k_0}^{\infty}\{\lim_{m\to\infty}\phi_{k,k,m}^{(n)} \le (q_{0,H}^\star)^2 - \frac{\epsilon}{2}\}$, it follows from (58) that $\mathbb{P}[B_n^{(AO)}, \text{i.o.}] = 0$. Since $(B_n^{(AO)})$ are independent, from the converse of Borel-Cantelli Lemma, we have:

$$\sum_{n=1}^{\infty} \mathbb{P}(B_n^{(AO)}) < \infty.$$

From the inequalities above, and leveraging this fact, we obtain:

$$\sum_{n=1}^{\infty} \mathbb{P}\left[\Phi^{(n)} \le (q_{0,H}^\star)^2 - \frac{\epsilon}{2}\right] < \infty$$

which from the Borel Cantelli Lemma implies (60).

Proof of (61). Using the fact that for all $k \in \mathbb{N}$, $\Phi^{(n)} \le \Phi_{k,k}^{(n)}$, we bound $\mathbb{P}[\Phi^{(n)} \ge (q_{0,H}^\star)^2 + \epsilon]$ as:

$$\mathbb{P}\left[\Phi^{(n)} \ge (q_{0,H}^\star)^2 + \epsilon\right]$$

$$\le \mathbb{P}\left[\cap_{k=1}^{\infty}\left\{\Phi_{k,k}^{(n)} \ge (q_{0,H}^\star)^2 + \epsilon\right\}\right]$$

We already proved the boundedness of the PO cost in (49), which implies that almost surely $\Phi_{k,k}^{(n)} \ne \infty$ for sufficiently large $k \ge k_0$. Hence, for $k \ge k_0$, there exists m such that:

$$\Phi_{k,k}^{(n)} \le \Phi_{k,k,m}^{(n)} - \frac{\epsilon}{2}.$$

Hence,

$$\mathbb{P}\left[\Phi^{(n)} \ge (q_{0,H}^\star)^2 + \epsilon\right] \le \mathbb{P}\left[\cup_{m=1}^{\infty}\left\{\Phi_{k,k}^{(n)} \ge (q_{0,H}^\star)^2 + \epsilon\right\}\right]$$

$$= \lim_{m\to\infty} \mathbb{P}\left[\left\{\Phi_{k,k,m}^{(n)} \ge (q_{0,H}^\star)^2 + \frac{\epsilon}{2}\right\}\right]$$

Using the CGMT theorem, we have:

$$\lim_{m\to\infty} \mathbb{P}\left[\left\{\Phi_{k,k,m}^{(n)} \ge (q_{0,H}^\star)^2 + \frac{\epsilon}{2}\right\}\right]$$

$$\le 2\lim_{m\to\infty} \mathbb{P}\left[\left\{\phi_{k,k,m}^{(n)} \ge (q_{0,H}^\star)^2 + \frac{\epsilon}{2}\right\}\right]$$

$$= 2\mathbb{P}\left[\cup_{m=1}^{\infty}\left\{\phi_{k,k,m}^{(n)} \ge (q_{0,H}^\star)^2 + \frac{\epsilon}{2}\right\}\right]$$

Choose $k > q_{0,H}^\star$. Hence,

$$\mathbb{P}\left[\cup_{m=1}^{\infty}\left\{\phi_{k,k,m}^{(n)} \ge (q_{0,H}^\star)^2 + \epsilon\right\}\right]$$

$$\le \mathbb{P}\left[\cup_{m=1}^{\infty}\left\{\phi_{k,k,m}^{(n)} \ge (q_{0,H}^\star)^2 + \epsilon\right\}\right]$$

$$\le \mathbb{P}\left[\lim_{m\to\infty}\phi_{k,k,m} \ge (q_{0,H}^\star)^2 + \frac{\epsilon}{2}\right]$$

With this at hand, we can in a similar way as before invoke the Converse of Borel-Cantelli Lemma to prove that $\{\Phi^{(n)} \ge (q_{0,H}^\star)^2 + \frac{\epsilon}{2}\}$ does not occur infinitely often. So far, we have thus proven that establishing (58) and (59) leads to proving

that $\Phi^{(n)} \to (q^\star_{0,H})^2$ almost surely. We will now proceed to the proof of (58) and (59).

Proof of (58). From (24) and the discussion following it, we have for sufficiently large k:

$$\lim_{m\to\infty} \phi_{k,k,m} = \min_{\substack{0\le q_0\le k \\ \hat{D}_H(q_0,\rho,\eta)\le 0 \\ -1\le\rho\le 1 \\ \eta\in\mathbb{R}}} q_0^2$$

Hence,

$$\mathbb{P}\left[\lim_{m\to\infty} \phi_{k,k,m} \le (q^\star_{0,H})^2 - \epsilon\right]$$
$$\le \mathbb{P}[\min_{\substack{0\le q_0\le k \\ \hat{D}_H(q_0,\rho,\eta)\le 0 \\ -1\le\rho\le 1 \\ \eta\in\mathbb{R}}} q_0^2 \le (q^\star_{0,H})^2 - \epsilon]$$
$$\le \mathbb{P}\left[\min_{\substack{0\le q_0\le k \\ \min_{-1\le\rho\le 1,\eta\in\mathbb{R}} \hat{D}_H(q_0,\rho,\eta)\le 0}} q_0^2 \le (q^\star_{0,H})^2 - \epsilon\right]$$

Function $\eta \mapsto \hat{D}_H(q_0,\rho,\eta)$ is convex in η and conveges pointwise to $D_H(q_0,\rho,\eta)$. Since $\lim_{\eta\to\infty} D_H(q_0,\rho,\eta) = \infty$ and $\lim_{\eta\to-\infty} D_H(q_0,\rho,\eta) = \infty$, from Lemma 10 in [10], we have:

$$\min_{\eta\in\mathbb{R}} \hat{D}_H(q_0,\rho,\eta) \xrightarrow{\text{a.s.}} \min_{\eta\in\mathbb{R}} D_H(q_0,\rho,\eta) \tag{62}$$

Function $\rho \mapsto \min_{\eta\in\mathbb{R}} \hat{D}_H(q_0,\rho,\eta)$ defined on $[-1,1]$ is convex in ρ, and converges pointwise from (62) to $\rho \mapsto \min_{\eta\in\mathbb{R}} D_H(q_0,\rho,\eta)$. As the pointwise convergence of convex functions implies uniform convergence over compact sets, we have for any $q_0 \in (0,k]$:

$$\min_{\substack{-1\le\rho\le 1 \\ \eta\in\mathbb{R}}} \hat{D}_H(q_0,\rho,\eta) \xrightarrow{\text{a.s.}} \min_{\substack{-1\le\rho\le 1 \\ \eta\in\mathbb{R}}} D_H(q_0,\rho,\eta) = \beta(q_0)$$

Define function $\hat{\beta}$ as $\hat{\beta} : q_0 \mapsto \min_{\substack{-1\le\rho\le 1 \\ \eta\in\mathbb{R}}} \hat{D}_H(q_0,\rho,\eta)$. We shall prove that

$$\hat{\beta}(q_0) \ge 0 \quad \text{for all } q \in (0, \frac{q^\star_{0,H}}{2}]. \tag{63}$$

To see this, we argue that $\beta(\frac{q^\star_{0,H}}{2}) > 0$, hence, $\hat{\beta}(\frac{q^\star_{0,H}}{2}) \ge 0$ almost surely. As $\hat{\beta}$ is decreasing we conclude that $\hat{\beta}(q_0) \ge 0$ for all $q_0 \in (0, \frac{q^\star_{0,H}}{2}]$. In view of this and choosing integer k_0 greater than $2q^\star_{0,H}$, we obtain

$$\mathbb{P}\left[\lim_{m\to\infty} \phi_{k_0,k_0,m} \le (q^\star_{0,H})^2 - \epsilon\right]$$
$$\le \mathbb{P}\left[\min_{\substack{\frac{q^\star_{0,H}}{2}\le q_0\le k_0 \\ \hat{\beta}(q_0)\le 0}} q_0^2 \le (q^\star_{0,H})^2 - \epsilon\right]$$

Function $q_0 \mapsto \hat{\beta}(q_0)$ is convex and converges pointwise to $q_0 \mapsto \beta(q_0)$ for all $q_0 > 0$. Hence, it converges uniformly over the set $[\frac{q^\star_{0,H}}{2}, k_0]$. As a result, for all $\tilde{\delta}$ sufficiently small, we can choose n, and p sufficiently large such that for all $q_0 \in [\frac{q^\star_{0,H}}{2}, k_0]$

$$\beta(q_0) - \tilde{\delta} \le \hat{\beta}(q_0) \le \beta(q_0) + \tilde{\delta}$$

We thus have:

$$\mathbb{P}\left[\lim_{m\to\infty} \phi_{k_0,k_0,m} \le (q^\star_{0,H})^2 - \epsilon\right]$$
$$\le \mathbb{P}[\min_{\substack{\frac{q^\star_{0,H}}{2}\le q_0\le k_0 \\ \beta(q_0)\le\tilde{\delta}}} q_0^2 \le (q^\star_{0,H})^2 - \epsilon]$$

Before going further, it is noteworthy to mention that the right-hand side event is casted in the form of a determinstic statement that does not involve any random variables. It suffices thus to check that for $\tilde{\delta}$ sufficiently small, this statement is false. This can be easily checked using Lemma 9 which enables to show that there exists δ_0 such that for all $\tilde{\delta} \le \delta_0$,

$$\min_{\substack{\frac{q^\star_{0,H}}{2}\le q_0\le k_0 \\ \beta(q_0)\le\tilde{\delta}}} q_0^2 \ge \min_{\substack{\frac{q^\star_{0,H}}{2}\le q_0\le k_0 \\ \beta(q_0)\le 0}} q_0^2 - \frac{\epsilon}{2}$$

Since $q_0 \mapsto \beta(q_0)$ is decreasing, $(q^\star_{0,H})^2 = \min_{\substack{\frac{q^\star_{0,H}}{2}\le q_0\le k_0 \\ \beta(q_0)\le 0}} q_0^2$ which directly implies that:

$$\mathbb{P}\left[\lim_{m\to\infty} \phi_{k_0,k_0,m} \le (q^\star_{0,H})^2 - \epsilon, \text{i.o.}\right] = 0.$$

Finally, to finish the proof of (58), we need to show that:

$$\lim_{m\to\infty} \phi_{k,k,m} = \lim_{m\to\infty} \phi_{k_0,k_0,m} \tag{64}$$

where k_0 is some integer sufficiently large. For that, we shall recall that for k_0 sufficiently large, and for all $k \ge k_0$

$$\lim_{m\to\infty} \phi_{k,k,m} = \min_{\substack{0\le q_0\le k \\ \hat{D}_H(q_0,\rho,\eta)\le 0 \\ -1\le\rho\le 1 \\ \eta\in\mathbb{R}}} q_0^2 \tag{65}$$

It is clear that $\lim_{m\to\infty} \phi_{k,k,m} \le \lim_{m\to\infty} \phi_{k_0,k_0,m}$. To prove (64), assume that $\lim_{m\to\infty} \phi_{k,k,m} < \lim_{m\to\infty} \phi_{k_0,k_0,m}$. Let $q_{0,1}$ and $q_{0,2}$ be positive scalars such that $\lim_{m\to\infty} \phi_{k,k,m} = q_{0,1}^2$ and $\lim_{m\to\infty} \phi_{k_0,k_0,m} = q_{0,2}^2$. Then,

$$q_{0,1} < q_{0,2}. \tag{66}$$

As $q_{0,2} \le k_0$, $q_{0,1} \le k_0$. From (65), we obtain $q_{0,1}^2 \ge \lim_{m\to\infty} \phi_{k_0,k_0,m} = q_{0,2}^2$, and hence $q_{0,1} \ge q_{0,2}$ which contradicts (66).

Proof of (59). We will now proceed to the proof of (59) for sufficiently large k. To begin with, we recall that:

$$\phi_{k,k,m} = \min_{\substack{0\le q_0\le k \\ -k\le b\le k \\ -1\le\rho\le 1}} q_0^2 + q_0 m \hat{D}_H\left(q_0, \rho, \frac{b}{q_0\sigma}\right)$$

Let k be greater than $2\max(q^\star_{0,H}, \eta^\star_H q^\star_{0,H}\sigma)$. Then, we can bound $\phi_{k,k,m}$ as:

$$\phi_{k,k,m} \leq \min_{\frac{q^\star_{0,H}}{2}\leq q_0\leq k} \quad q_0^2 + mq_0\hat{D}_H(q_0, \rho^\star_H, \eta^\star_H) \tag{67}$$

$$\leq \min_{\substack{\frac{q^\star_{0,H}}{2}\leq q_0\leq k \\ \hat{D}_H(q_0,\eta^\star_H)\leq 0}} \quad q_0^2 + mq_0\hat{D}_H(q_0, \rho^\star_H, \eta^\star_H) \tag{68}$$

$$\leq \min_{\substack{\frac{q^\star_{0,H}}{2}\leq q_0\leq k \\ \hat{D}_H(q_0,\eta^\star_H)\leq 0}} \quad q_0^2 \tag{69}$$

As a result,

$$\lim_{m\to\infty}\phi_{k,k,m} \leq \min_{\substack{\frac{q^\star_{0,H}}{2}\leq q_0\leq k \\ \hat{D}_H(q_0,\rho^\star_H,\eta^\star_H)\leq 0}} \quad q_0^2$$

Function $q_0 \mapsto \hat{D}_H(q_0, \rho^\star_H, \eta^\star_H)$ is convex in q_0 and converges pointwise to $q_0 \mapsto D_H(q_0, \rho^\star_H, \eta^\star_H)$. As the convergence of convex functions is uniform over compact sets, it thus converges uniformly to $q_0 \mapsto D_H(q_0, \rho^\star_H, \eta^\star_H)$ over the set $\{q_0 \mid \frac{q^\star_{0,H}}{2} \leq q_0 \leq k\}$. Hence, for all $\tilde{\delta}$ sufficiently small, and for all n and p sufficiently large,

$$\hat{D}_H(q_0, \rho^\star_H, \eta^\star_H) \leq D_H(q_0, \rho^\star_H, \eta^\star_H) + \tilde{\delta}$$

Hence,

$$\left\{q_0 \mid D_H(q_0, \rho^\star_H, \eta^\star_H) \leq -\tilde{\delta}\right\} \subset \left\{q_0 \mid \hat{D}_H(q_0, \rho^\star_H, \eta^\star_H) \leq 0\right\}$$

and thus:

$$\lim_{m\to\infty}\phi_{k,k,m} \leq \min_{\substack{\frac{q^\star_{0,H}}{2}\leq q_0\leq k \\ D_H(q_0,\frac{\eta^\star_H}{q^\star_{0,H}\sigma})\leq -\tilde{\delta}}} \quad q_0^2 \tag{70}$$

Applying Lemma 9, we can see that:

$$\min_{\substack{\frac{q^\star_{0,H}}{2}\leq q_0\leq k \\ D_H(q_0,\frac{\eta^\star_H}{q^\star_{0,H}\sigma})\leq 0}} \quad q_0^2 = \inf_{\tilde{\delta}\geq 0} \min_{\substack{\frac{q^\star_{0,H}}{2}\leq q_0\leq k \\ D_H(q_0,\rho^\star,\frac{\eta^\star_H}{q^\star_{0,H}\sigma})\leq -\tilde{\delta}}} \quad q_0^2$$

Invoking the ϵ-definition of the infimum, we thus have for $\tilde{\delta} > 0$,

$$\min_{\substack{\frac{q^\star_{0,H}}{2}\leq q_0\leq k \\ D_H(q_0,\frac{\eta^\star_H}{q^\star_{0,H}\sigma})\leq -\tilde{\delta}}} \quad q_0^2 \leq \min_{\substack{\frac{q^\star_{0,H}}{2}\leq q_0\leq k \\ D_H(q_0,\rho^\star,\frac{\eta^\star_H}{q^\star_{0,H}\sigma})\leq 0}} \quad q_0^2 + \frac{\epsilon}{2} \leq (q^\star_{0,H})^2 + \frac{\epsilon}{2}$$

and thus from (70)

$$\lim_{m\to\infty}\phi_{k,k,m} \leq (q^\star_{0,H})^2 + \frac{\epsilon}{2},$$

which proves (59).

Concluding. It follows from (60) and (61) that when $\delta \leq \delta_\star$,

$$\Phi^{(n)} \xrightarrow{\text{a.s.}} (q^\star_{0,H})^2.$$

Recalling that $\Phi^{(n)} = \|\hat{\mathbf{w}}_H\|^2$, this shows that $\|\hat{\mathbf{w}}_H\|^2 \xrightarrow{\text{a.s.}} (q^\star_{0,H})^2$. To prove that $\frac{\hat{\mathbf{w}}_H^T\boldsymbol{\mu}}{\|\hat{\mathbf{w}}_H\|_2\|\boldsymbol{\mu}\|_2}$ converges to $\rho^\star_H$, we consider the following set:

$$\mathcal{S}_\zeta = \left\{\mathbf{w}\in\mathbb{R}^p \mid \left|\frac{\mathbf{w}^T\boldsymbol{\mu}}{\|\hat{\mathbf{w}}\|_2\|\boldsymbol{\mu}\|_2} - \rho^\star_H\right| \leq \zeta\right\}$$

and define the perturbed version of the PO obtained from (17) by constraining $\mathbf{w}$ to be outside the set $\mathcal{S}_\zeta$:

$$\tilde{\Phi}^{(n)} = \min_{\substack{w,b \\ \mathbf{w}\notin\mathcal{S}_\zeta}} \max_{\mathbf{u}\geq 0} \frac{1}{\sqrt{p}}\mathbf{u}^T\mathbf{Z}\mathbf{w} + \psi(\mathbf{w},\mathbf{u})$$

We consider proving that there exists $\tilde{\nu}$ such that

$$\mathbb{P}\left[\tilde{\Phi}^{(n)} \leq (q^\star_{0,H})^2 + \tilde{\nu}, \text{i.o}\right] = 0 \tag{71}$$

which implies that $\mathbf{w}\in\mathcal{S}_\zeta$ almost surely. For that, we invoke the max-min inequality and lower-bound $\tilde{\Phi}^{(n)}$ by [3]:

$$\tilde{\Phi}^{(n)} \geq \inf_{r,B\geq 0}\sup_{\theta\geq 0}\tilde{\Phi}^{(n)}_{r,B,\theta}$$

with

$$\tilde{\Phi}^{(n)}_{r,B,\theta} := \min_{\substack{\mathbf{w}\in\mathbb{R}^p \\ \mathbf{w}\notin\mathcal{S}_\zeta \\ \|\mathbf{w}\|_2\leq r \\ |b|\leq B}} \max_{\substack{\mathbf{u}\geq 0 \\ \|\mathbf{u}\|_2\leq\theta}} \frac{1}{\sqrt{p}}\mathbf{u}^T\mathbf{Z}\mathbf{w} + \psi(\mathbf{w},\mathbf{u})$$

In the event $\{\tilde{\Phi} \leq (q^\star_{0,H})^2 + \tilde{\nu}\}$, $\tilde{\Phi} \neq \infty$. Hence, there exists $\overline{k}$ chosen sufficiently large as needed such that:

$$\tilde{\Phi} \geq \tilde{\Phi}_{k,k,m} - \epsilon, \ \forall\, k \geq \overline{k} \ \text{ and } \ m\in\mathbb{N}$$

Hence,

$$\mathbb{P}\left[\tilde{\Phi}^{(n)} < (q^\star_{0,H})^2 + \tilde{\nu}\right]$$

$$\leq \mathbb{P}\left[\cup_{k=1}^\infty \cap_{m=1}^\infty \left\{\tilde{\Phi}^{(n)}_{k,k,m} < (q^\star_{0,H})^2 + \tilde{\nu}\right\}\right] \tag{72}$$

$$= \lim_{k\to\infty}\lim_{m\to\infty}\mathbb{P}\left[\tilde{\Phi}^{(n)}_{k,k,m} < (q^\star_{0,H})^2 + \tilde{\nu}\right] \tag{73}$$

We may proceed as previously by associating with each $\tilde{\Phi}^{(n)}_{k,k,m}$ the following perturbed AO problem given by:

$$\tilde{\phi}^{(n)}_{k,k,m} = \min_{\substack{\mathbf{w}\in\mathbb{R}^p \\ \mathbf{w}\notin\mathcal{S}_\zeta \\ \|\mathbf{w}\|_2\leq k \\ |b|\leq k}} \max_{\substack{\mathbf{u}\geq 0 \\ \|\mathbf{u}\|_2\leq m}} \frac{1}{\sqrt{p}}\|\mathbf{w}\|_2\mathbf{g}^T\mathbf{u}$$

$$- \frac{1}{\sqrt{p}}\|\mathbf{u}\|_2\mathbf{h}^T\mathbf{w} + \psi(\mathbf{w},\mathbf{u})$$

Using Gordon's inequality, we obtain:

$$\mathbb{P}\left[\tilde{\Phi}^{(n)}_{k,k,m} < (q^\star_{0,H})^2 + \tilde{\nu}\right] \leq 2\mathbb{P}\left[\tilde{\phi}^{(n)}_{k,k,m} < (q^\star_{0,H})^2 + \tilde{\nu}\right]$$

[3] Note here that equality is not guaranteed since the set $\{\mathbf{w} \mid \mathbf{w}\notin\mathcal{S}_\zeta\}$ is not convex

Hence,

$$\begin{aligned}
&\mathbb{P}\left[\tilde{\Phi}^{(n)} < (q_{0,H}^\star)^2 + \tilde{\nu}\right] \\
&\quad \le 2 \lim_{k\to\infty} \lim_{m\to\infty} \mathbb{P}\left[\tilde{\phi}_{k,k,m}^{(n)} < (q_{0,H}^\star)^2 + \tilde{\nu}\right] && (74)\\
&\quad = 2\mathbb{P}\left[\cup_{k=\bar{k}}^{\infty} \cap_{m=1}^{\infty} \left\{\tilde{\phi}_{k,k,m}^{(n)} < (q_{0,H}^\star)^2 + \tilde{\nu}\right\}\right] && (75)\\
&\quad \le 2\mathbb{P}\left[\cup_{k=\bar{k}}^{\infty} \left\{\lim_{m\to\infty} \tilde{\phi}_{k,k,m}^{(n)} < (q_{0,H}^\star)^2 + \tilde{\nu}\right\}\right] && (76)
\end{aligned}$$

Following the same calculations as before, we can further simplify $\tilde{\phi}_{\bar{k},\bar{k},m}^{(n)}$ for $m \in \mathbb{N}$ as:

$$\tilde{\phi}_{\bar{k},\bar{k},m}^{(n)} = \inf_{\substack{0\le q_0\le \bar{k}\\ -1\le\rho\le 1\\ |\rho-\rho_H^\star|\ge\zeta\\ |b|\le\bar{k}}} q_0^2 + m q_0 \hat{D}_H\left(q_0, \rho, \frac{b}{\sigma q_0}\right)$$

Hence,

$$\begin{aligned}
\lim_{m\to\infty} \tilde{\phi}_{\bar{k},\bar{k},m}^{(n)} &\ge \lim_{m\to\infty} \inf_{\substack{0\le q_0\le \bar{k}\\ -1\le\rho\le 1\\ |\rho-\rho_H^\star\|\ge\zeta\\ \eta\in\mathbb{R}}} q_0^2 + m q_0 \hat{D}_H(q_0,\rho,\eta) && (77)\\
&= \inf_{\substack{0\le q_0\le \bar{k}\\ -1\le\rho\le 1\\ |\rho-\rho_H^\star|\ge\zeta\\ \eta\in\mathbb{R}\\ \hat{D}_H(q_0,\rho,\eta)\le 0}} q_0^2 && (78)\\
&\ge \inf_{\substack{0\le q_0\le \bar{k}\\ \min_{\substack{-1\le\rho\le 1\\ |\rho-\rho_H^\star|\ge\zeta\\ \eta\in\mathbb{R}}} \hat{D}_H(q_0,\rho,\eta)\le 0}} q_0^2 && (79)\\
&\overset{(a)}{=} \inf_{\substack{\frac{q_{0,H}^\star}{2}\le q_0\le \bar{k}\\ \min_{\substack{-1\le\rho\le 1\\ |\rho-\rho_H^\star|\ge\zeta\\ \eta\in\mathbb{R}}} \hat{D}_H(q_0,\rho,\eta)\le 0}} q_0^2 && (80)
\end{aligned}$$

where (a) follows from (63). Next, using the uniform convergence of $(q_0,\rho)\mapsto \min_{\eta\in\mathbb{R}} \hat{D}_H(q_0,\rho,\eta)$ to $(q_0,\rho)\mapsto \min_{\eta\in\mathbb{R}} D_H(q_0,\rho,\eta)$ we show that for all $\tilde{\delta}>0$, we may select n and p sufficiently large such that for all $q_0 \in [\frac{q_{0,H}^\star}{2}, \bar{k}]$,

$$\min_{\substack{-1\le\rho\le 1\\ |\rho-\rho_H^\star|\ge\zeta\\ \eta\in\mathbb{R}}} \hat{D}_H(q_0,\rho,\eta) \ge \min_{\substack{-1\le\rho\le 1\\ |\rho-\rho_H^\star|\ge\zeta\\ \eta\in\mathbb{R}}} D_H(q_0,\rho,\eta) - \tilde{\delta}$$

Hence, almost surely,

$$\lim_{m\to\infty} \tilde{\phi}_{\bar{k},\bar{k},m}^{(n)} \ge \inf_{\substack{\frac{q_{0,H}^\star}{2}\le q_0\le \bar{k}\\ \min_{\substack{-1\le\rho\le 1\\ |\rho-\rho_H^\star|\ge\zeta\\ \eta\in\mathbb{R}}} D_H(q_0,\rho,\eta)\le\tilde{\delta}}} q_0^2$$

Fix $\epsilon>0$. From Lemma 9, there exists δ_0 such that for all $\tilde{\delta}\le\delta_0$,

$$\lim_{m\to\infty} \tilde{\phi}_{\bar{k},\bar{k},m}^{(n)} \ge \inf_{\substack{\frac{q_{0,H}^\star}{2}\le q_0\le \bar{k}\\ \tilde{\beta}(q_0)\le 0}} q_0^2 - \epsilon.$$

with $\tilde{\beta}: q_0 \mapsto \min_{\substack{-1\le\rho\le 1\\ |\rho-\rho_H^\star\|\ge\zeta\\ \eta\in\mathbb{R}}} D_H(q_0,\rho,\eta)$. It is easy to see that $\tilde{\beta}(q_{0,H}^\star) \ge \beta(q_{0,H}^\star) = 0$. Since $\rho_H^\star$ is the unique minimizer of $\rho\mapsto \min_{\eta\in\mathbb{R}} D_H(q_{0,H}^\star,\rho,\eta)$, $\tilde{\beta}(q_{0,H}^\star)>0$. As $\tilde{\beta}$ is decreasing,

$$\inf_{\substack{\frac{q_{0,H}^\star}{2}\le q_0\le \bar{k}\\ \tilde{\beta}(q_0)\le 0}} q_0^2 > (q_{0,H}^\star)^2$$

and as such

$$\lim_{m\to\infty} \tilde{\phi}_{\bar{k},\bar{k},m}^{(n)} > (q_{0,H}^\star)^2$$

The event $\{\tilde{\phi}_{\bar{k},\bar{k},m}^{(n)} > (q_{0,H}^\star)^2 < (q_{0,H}^\star)^2 + \tilde{\nu}\}$ does not occur infinitely often. Since $\tilde{\phi}_{k,k,m}^{(n)} > \tilde{\phi}_{\bar{k},\bar{k},m}^{(n)}$ the event $\{\cup_{k=\bar{k}}^{\infty}\{\tilde{\phi}_{\bar{k},\bar{k},m}^{(n)} > (q_{0,H}^\star)^2 < (q_{0,H}^\star)^2 + \tilde{\nu}\}\}$ does not occur infintely often either. Keeping track of the inequalities (77)-(80) and using as before the converse of the Borel Cantelli Lemma, we prove (71). We can similarly follow the same methodology to prove the convergence of $\hat{b}_H$ to $\eta_H^\star q_{0,H}^\star \sigma$. Details are omitted due to lack of space.

VI. CONCLUSION

This paper presents an asymptotically sharp characterization of the performance of the hard-margin and soft-margin SVM. Our analysis builds upon the recently developed CGMT framework, which was mainly used before in the study of high-dimensional regression problems. Considering its use for the analysis of SVM poses technical challenges, which have been handled through a new promising technical approach. This approach not only allowed for an easier use of the CGMT but also enabled to obtain stronger almost sure convergence results. We believe that the developed tools lay the groundwork to facilitate and pave the way towards the use of the CGMT to general optimization based-classifiers such as logistic regression, Adaboost, for which an explicit formulation is not available.

APPENDIX A TECHNICAL LEMMAS

This appendix gathers some important lemmas that are extensively used when optimizing the auxiliary problem. The following Lemma, whose proof is not complicated, is fundamental to simplify the optimization of the auxiliary problem. As shown above, it allowed in some cases to avoid the necessity of flipping the order of the min-max when solving min-max optimization problems.

Lemma 8: Let d_1 and d_2 be two strictly positive integers. Let $X\times Y$ be two non-empty sets in $\mathbb{R}_{d_1}\times\mathbb{R}_{d_2}$. Let $F: X\times Y\to\mathbb{R}$ be a given real-valued function. Assume there exists

$\tilde{X} \subset X$ such that for all $\mathbf{x} \in X$ there exists $\tilde{\mathbf{x}} \in \tilde{X}$ such that:

$$\forall \mathbf{y} \in Y, \quad F(\mathbf{x}, \mathbf{y}) \geq F(\tilde{\mathbf{x}}, \mathbf{y}). \tag{81}$$

Then

$$\min_{\mathbf{x}\in X} \max_{\mathbf{y}\in Y} F(\mathbf{x}, \mathbf{y}) = \min_{\tilde{\mathbf{x}}\in \tilde{X}} \max_{\mathbf{y}\in Y} F(\tilde{\mathbf{x}}, \mathbf{y})$$

In particular, if $\tilde{\mathbf{X}} = \{\tilde{\mathbf{x}}\}$, then:

$$\min_{\mathbf{x}\in X} \max_{\mathbf{y}\in Y} F(\mathbf{x}, \mathbf{y}) = \max_{\mathbf{y}\in Y} F(\tilde{\mathbf{x}}, \mathbf{y})$$

Proof: It is easy to see that $\min_{\mathbf{x}\in X} \max_{\mathbf{y}\in Y} F(\mathbf{x}, \mathbf{y})$ is upper-bounded by:

$$\min_{\mathbf{x}\in X} \max_{\mathbf{y}\in Y} F(\mathbf{x}, \mathbf{y}) \leq \min_{\tilde{\mathbf{x}}\in \tilde{X}} \max_{\mathbf{y}\in Y} F(\tilde{\mathbf{x}}, \mathbf{y})$$

To prove the lower-bound, we will exploit the property described in (81). Let $\mathbf{x} \in X$ and let $\tilde{\mathbf{x}}(\mathbf{x})$ is such that for all $y \in Y$,

$$F(\mathbf{x}, \mathbf{y}) \geq F(\tilde{\mathbf{x}}(\mathbf{x}), \mathbf{y})$$

Hence

$$\max_{\mathbf{y}\in Y} F(\mathbf{x}, \mathbf{y}) \geq \max_{\mathbf{y}\in Y} F(\tilde{\mathbf{x}}(\mathbf{x}), \mathbf{y}) \geq \min_{\tilde{\mathbf{x}}\in \tilde{X}} \max_{\mathbf{y}\in Y} F(\tilde{\mathbf{x}}, \mathbf{y})$$

which proves that:

$$\min_{\mathbf{x}\in X} \max_{\mathbf{y}\in Y} F(\mathbf{x}, \mathbf{y}) \geq \min_{\tilde{\mathbf{x}}\in \tilde{X}} \max_{\mathbf{y}\in Y} F(\tilde{\mathbf{x}}, \mathbf{y})$$

■

Lemma 9: Let $d \in \mathbb{N}^\star$. Let S_x be a compact non-empty set in $\mathbb{R}^d$. Let f and c be two continuous functions over S_x such that the set $\{c(x) \leq 0\}$ is non-empty. Then:

$$\min_{\substack{\mathbf{x}\in S_x \\ c(\mathbf{x})\leq 0}} f(\mathbf{x}) = \sup_{\delta\geq 0} \min_{\substack{\mathbf{x}\in S_x \\ c(\mathbf{x})\leq \delta}} f(\mathbf{x}) = \inf_{\delta>0} \min_{\substack{\mathbf{x}\in S_x \\ c(\mathbf{x})\leq -\delta}} f(\mathbf{x})$$

Proof: We will prove only the first equality, the second one following along the same lines. Obviously, the following inequality holds true,

$$\min_{\substack{\mathbf{x}\in S_x \\ c(\mathbf{x})\leq 0}} f(\mathbf{x}) \geq \sup_{\delta\geq 0} \min_{\substack{\mathbf{x}\in S_x \\ c(\mathbf{x})\leq \delta}} f(\mathbf{x})$$

To see this, it suffices to note that for all $\delta \geq 0$,

$$\min_{\substack{\mathbf{x}\in S_x \\ c(\mathbf{x})\leq 0}} f(\mathbf{x}) \geq \min_{\substack{\mathbf{x}\in S_x \\ c(\mathbf{x})\leq \delta}} f(\mathbf{x})$$

Hence,

$$\min_{\substack{\mathbf{x}\in S_x \\ c(\mathbf{x})\leq 0}} f(\mathbf{x}) \geq \overline{f} \triangleq \sup_{\delta\geq 0} \min_{\substack{\mathbf{x}\in S_x \\ c(\mathbf{x})\leq \delta}} f(\mathbf{x})$$

From the ϵ-definition of the supremum, for any $\epsilon > 0$, there exists δ_ϵ and $\mathbf{x}_\epsilon^\star$ such that $c(\mathbf{x}_\epsilon) \leq \delta_\epsilon$, $f(\mathbf{x}_\epsilon^\star) = \min_{\substack{\mathbf{x}\in S_x \\ c(\mathbf{x})\leq \delta^\epsilon}} f(\mathbf{x})$, and

$$\overline{f} \leq f(\mathbf{x}_\epsilon^\star) + \epsilon$$

Now for all $m \in \mathbb{N}^\star$ such that $m \geq m_0 \triangleq \lceil \rceil [] \frac{1}{\delta^\epsilon}$, define $\mathbf{x}_m^\star$ such that $f(\mathbf{x}_m^\star) = \min_{\substack{\mathbf{x}\in S_x \\ c(\mathbf{x})\leq \frac{1}{m}}} f(\mathbf{x})$. Clearly, $f(\mathbf{x}_m^\star) \geq f(\mathbf{x}_\epsilon^\star)$. Hence,

$$\overline{f} \leq f(\mathbf{x}_m^\star) + \epsilon \leq \min_{\substack{\mathbf{x}\in S_x \\ c(\mathbf{x})\leq 0}} f(\mathbf{x}) + \epsilon \tag{82}$$

Assume that

$$\overline{f} > \min_{\substack{\mathbf{x}\in S_x \\ c(\mathbf{x})\leq 0}} f(\mathbf{x}). \tag{83}$$

Then, plugging $\epsilon = \frac{1}{2}(\overline{f} - \min_{\substack{\mathbf{x}\in S_x \\ c(\mathbf{x})\leq 0}} f(\mathbf{x}))$ into (82) leads to

$$\overline{f} \leq \min_{\substack{\mathbf{x}\in S_x \\ c(\mathbf{x})\leq 0}} f(\mathbf{x})$$

which contradicts (83). ■

Lemma 10: Let X and Y be two convex sets. Let $f : X \times Y \to \mathbb{R}$ be a jointly convex function in $X \times Y$. Assume that $\forall y \in Y$, $\inf_{x\in X} f(x, y) > -\infty$. Then: $g : y \mapsto \inf_{x\in X} f(x, y)$ is convex in Y.

Proof: See [26] ■

Lemma 11: Let $\mathbf{a} = [a_1, \ldots, a_n]^T$ be a vector in $\mathbb{R}^{n\times 1}$ and θ be a positive scalar. Then:

$$\max_{\substack{\mathbf{u}\geq 0 \\ \|\mathbf{u}\|_2=\theta}} \mathbf{a}^T\mathbf{u} = \theta\sqrt{\sum\nolimits_{i=1}^n (a_i)_+^2}$$

Lemma 12: Let $\mathbf{a} \in \mathbb{R}^{n\times 1}$. Let $\tilde{\beta}$ and τ be positive scalars. Then, if $\tilde{\beta} = 0$,

$$\max_{0\leq \mathbf{u}\leq \tau} \mathbf{u}^T\mathbf{a} - \tilde{\beta}\|\mathbf{u}\|_2 = \max_{0\leq \mathbf{u}\leq \tau} \mathbf{u}^T\mathbf{a} = \sum_{i=1}^n \tau(a_i)_+$$

If $\tilde{\beta} \neq 0$, then:

$$\max_{0\leq \mathbf{u}\leq \tau} \mathbf{u}^T\mathbf{a} - \tilde{\beta}\|\mathbf{u}\|_2 = \sup_{\xi\geq 0} \sum_{i=1}^n \left(a_i\tau - \tau^2\frac{1}{2\xi}\right)\mathbf{1}_{\{a_i\xi\geq\tau\}}$$

$$+ \sum_{i=1}^n \frac{a_i^2\xi}{2}\mathbf{1}_{\{0\leq a_i\xi<\tau\}} - \frac{\tilde{\beta}^2\xi}{2} \tag{84}$$

Moreover, function $\xi \mapsto \sum_{i=1}^n (a_i\tau - \tau^2\frac{1}{2\xi})\mathbf{1}_{\{a_i\xi\geq\tau\}} + \sum_{i=1}^n \frac{a_i^2\xi}{2}\mathbf{1}_{\{0\leq a_i\xi\leq\tau\}} - \frac{\tilde{\beta}^2\xi}{2}$ is concave in ξ when $\xi \in (0, \infty)$.

Proof: Using the fact that:

$$\|\mathbf{u}\| = \inf_{\chi>0} \frac{\chi}{2} + \frac{\|\mathbf{u}\|_2^2}{2\chi}$$

we obtain:

$$\max_{0\leq \mathbf{u}\leq \tau} \mathbf{u}^T\mathbf{a} - \tilde{\beta}\|\mathbf{u}\|_2 = \max_{0\leq \mathbf{u}\leq \tau} \sup_{\chi\geq 0} \mathbf{u}^T\mathbf{a} - \tilde{\beta}\left[\frac{\chi}{2} + \frac{\|\mathbf{u}\|_2^2}{2\chi}\right] \tag{85}$$

$$= \sup_{\chi>0} \max_{0\leq \mathbf{u}\leq \tau} \sum_{i=1}^n a_iu_i - \frac{\tilde{\beta}}{2\chi}u_i^2 - \tilde{\beta}\frac{\chi}{2} \tag{86}$$

Function $x \mapsto a_i x - \frac{\tilde{\beta}}{2\chi}x^2$ is increasing on $(-\infty, \frac{a_i\chi}{\tilde{\beta}})$ and decreasing on $(\frac{a_i\chi}{\tilde{\beta}}, \infty)$ taking its maximum at $x^\star = \frac{a_i\chi}{\tilde{\beta}}$. Hence,

$$\max_{0\le x\le\tau} a_i x - \frac{\tilde{\beta}}{2\chi}x^2 = \begin{cases} 0 & \text{if } \frac{a_i\chi}{\tilde{\beta}} < 0 \\ a_i\tau - \tau^2\frac{\tilde{\beta}}{2\chi} & \text{if } 0 \le \tau < \frac{a_i\chi}{\tilde{\beta}} \\ \frac{a_i^2\chi}{2\tilde{\beta}} & \text{if } 0 \le \frac{a_i\chi}{\tilde{\beta}} \le \tau \end{cases}$$

Hence,

$$\max_{0\le\mathbf{u}\le\tau} \mathbf{u}^T\mathbf{a} - \tilde{\beta}\|\mathbf{u}\|_2 = \sup_{\chi>0} \sum_{i=1}^n \left(a_i\tau - \tau^2\frac{\tilde{\beta}}{2\chi}\right)\mathbf{1}_{\left\{\frac{a_i\chi}{\tilde{\beta}}\ge\tau\right\}} + \sum_{i=1}^n \frac{a_i^2\chi}{2\tilde{\beta}}\mathbf{1}_{\left\{0\le\frac{a_i\chi}{\tilde{\beta}}<\tau\right\}} - \frac{\tilde{\beta}\chi}{2}$$

Performing the change of variable $\xi := \frac{\chi}{\tilde{\beta}}$ yields (84).

We will now proceed to proving the concavity of function $\varphi_i : \xi \mapsto (a_i\tau - \tau^2\frac{1}{2\xi})\mathbf{1}_{\{a_i\xi\ge\tau\}} + \frac{a_i^2\xi}{2}\mathbf{1}_{\{0\le a_i\xi<\tau\}} - \frac{\tilde{\beta}^2\xi}{2}$. To this end, note that:

$$\varphi_i(\xi) = \max_{0\le u_i\le\tau} a_i u_i - \frac{u_i^2}{2\xi} - \tilde{\beta}^2\frac{\xi}{2}.$$

Function $(\xi, u_i) \mapsto \frac{u_i^2}{2\xi}$ is jointly convex in $\mathbb{R}_{>0}\times[0,\tau]$ since it is the perspective function of $x \mapsto x^2$. Hence, (ξ, u_i) is jointly concave in $\mathbb{R}_{>0}\times[0,\tau]$. Using Lemma 10, we thus get that $\xi \mapsto \varphi_i(\xi)$ is concave in $\mathbb{R}_{>0}$ ■

REFERENCES

[1] V. Vapnik, *Statistical Learning Theory*. New York, NY, USA: Wiley, 1998.

[2] P. McCullagh and J. A. Nelder, *Generalized Linear Models*, 2nd ed., London, U.K.: Chapman & Hall, 1989.

[3] Y. Freund and R. E. Schapire, "A decision-theoretic generalization of on-line learning and an application to boosting," *J. Comput. Syst. Sci.*, vol. 55, no. 1, pp. 119–139, Aug. 1997.

[4] R. Couillet and F. Benyach-Georges, "Kernel spectral clustering of large dimensional data," *Electron. J. Statist.*, vol. 10, no. 1, pp. 1393–1454, 2016, doi: 10.1214/16-EJS1144, [Online]. Available: https://projecteuclid.org/euclid.ejs/1464710237.

[5] K. Elkhalil, A. Kammoun, R. Couillet, T. Y. Al-Naffouri, and M.-S. Alouini, "A large dimensional study of regularized discriminant analysis classifiers," *IEEE Trans. Signal Process.*, vol. 68, pp. 2464–2479, Apr. 2020. [Online]. Available: https://doi.org/10.1109/TSP.2020.2984160

[6] K. Elkhalil, A. Kammoun, R. Couillet, T. Y. Al-Naffouri, and M.-S. Alouini, "Asymptotic performance of regularized quadratic discriminant analysis based classifiers," in *Proc. 27th IEEE Int. Workshop Mach. Learn. Signal Process. (MLSP)*, Tokyo, Japan, Sep. 25–28, 2017, pp. 1–6. doi: 10.1109/MLSP.2017.8168172.

[7] N. E. Karoui, "Asymptotic behavior of unregularized and ridge-regularized high-dimensional robust regression estimators: Rigorous results," *arXiv:1311.2445*.

[8] D. Donoho and A. Montanari, "High dimensional robust m-estimation: Asymptotic variance via approximate message passing," *Probability Theory Related Fields*, vol. 166, no. 3, pp. 935–969, Dec. 2016.

[9] M. Stojnic, "A framework to characterize performance of LASSO algorithms," 2013. [Online]. Available: https://arxiv.org/pdf/1303.7291.pdf

[10] C. Thrampoulidis, E. Abbasi, and B. Hassibi, "Precise error analysis of regularized M-estimators in high dimensions," *IEEE Trans. Inf. Theory*, vol. 64, no. 8, pp. 5592–5628, Aug. 2018.

[11] C. M. Bishop, *Pattern Recognition and Machine Learning (Information Science and Statistics)*. New York, NY, USA: Springer-Verlag, Inc., 2006.

[12] H. Sifaou, A. Kammoun, and M. Alouini, "Phase transition in the hard-margin support vector machines," in *Proc. 8th IEEE Int. Workshop Comput. Adv. Multi-Sensor Adaptive Process.*, Le Gosier, Guadeloupe, Dec. 15–18, 2019, pp. 415–419. [Online]. Available: https://doi.org/10.1109/CAMSAP45676.2019.9022461

[13] H. Huang, "Asymptotic behavior of support vector machine for spiked population model," *J. Mach. Learn. Res.*, vol. 18, pp. 45:1–45:21, 2017.

[14] X. Mai and Z. Liao, "High Dimensional Classification via Regularized and Unregularized Empirical Risk Minimization: Precise Error and Optimal Loss," *Mach. Learn.*, 2020. [Online]. Available: https://arxiv.org/abs/1905.13742

[15] H. Sifaou, A. Kammoun, and M.-S. Alouini, "High-dimensional linear discriminant analysis classifier for spiked covariance model," *J. Mach. Learn. Res.*, vol. 21, pp. 112:1–112:24, 2020. [Online]. Available: http://jmlr.org/papers/v21/19-428.html

[16] Z. Liao and R. Couillet, "A large dimensional analysis of least squares support vector machines," *IEEE Trans. Signal Process.*, vol. 67, no. 4, pp. 1065–1074, Feb. 2019.

[17] J. E. Candès and P. Sur, "The phase transition for the existence of the maximum likelihood estimate in high-dimensional logistic regression," *Ann. Statist.*, vol. 48, no. 1, pp. 27–42, Feb. 2020. [Online]. Available: https://doi.org/10.1214/18-AOS1789

[18] A. Kammoun and M.-S. Alouini, "On the precise error analysis of support vector machines," Tech. Rep., 2020. [Online]. Available: https://arxiv.org/abs/2003.12972

[19] C. Thrampoulidis, S. Oymak, and B. Hassibi, "Regularized linear regression: A precise analysis of the estimation error," in *Proc. 28th Conf. Learn. Theory*, Paris, France, 2015, pp. 1683–1709.

[20] Y. Gordon, *On Milman's Inequality and Random Subspaces Which Escape Through a Mesh in* $\mathbb{R}^n$. Berlin, Germany: Springer, 1988.

[21] Y. Gordon, "Some inequalities for Gaussian processes and applications," *Isr. J. Math.*, vol. 50, no. 4, pp. 265–289, Dec. 1985.

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

[23] C. Thrampoulidis, W. Xu, and B. Hassibi, "Symbol error rate performance of box-relaxation decoders in massive MIMO," *IEEE Trans. Signal Process.*, vol. 66, no. 13, pp. 3377–3392, Jul. 2018.

[24] W. K. Newey and D. L. McFadden, "Large sample estimation and hypothesis testing," in *Handbook of Econometrics*, D. L. McFadden and R. F. Engle, Eds., New York, NY, USA: Elsevier, 1994, pp. 2111–2245.

[25] R. K. Sundaram, *A First Course in Optimization Theory*. Cambridge, U.K.: Cambridge Univ. Press, 1996.

[26] S. Boyd and L. Vandenberghe, *Convex Optimization*. Cambridge, U.K.: Cambridge Univ. Press, 2003.

ABLA KAMMOUN (Member, IEEE) was born in Sfax, Tunisia. She received the Engineering degree in signal and systems from Tunisia Polytechnic School, La Marsa, and the master's and Ph.D. degrees in digital communications from Télécom Paris Tech (formerly, École Nationale Supérieure des Télécommunications). From 2010 to 2012, she was a Postdoctoral Researcher with TSI Department, Telecom Paris Tech. She was then with Supélec, Alcatel-Lucent Chair on Flexible Radio until 2013. She is currently a Research Scientist with KAUST. Her research interests include performance analysis of wireless communication systems, random matrix theory, and statistical signal processing.

MOHAMED-SLIM ALOUINI (Fellow, IEEE) was born in Tunis, Tunisia. He received the Ph.D. degree in electrical engineering from the California Institute of Technology (Caltech), Pasadena, CA, USA, in 1998. He was a Faculty Member with the University of Minnesota, Minneapolis, MN, USA, then in the Texas A&M University, Qatar, Education City, Doha, Qatar before joining King Abdullah University of Science and Technology (KAUST), Thuwal, Makkah Province, Saudi Arabia as a Professor of Electrical Engineering in 2009. His current research interests include the modeling, design, and performance analysis of wireless communication systems.