entropy

MDPI

Article

Sharp Guarantees and Optimal Performance for Inference in Binary and Gaussian-Mixture Models †

Hossein Taheri *, Ramtin Pedarsani * and Christos Thrampoulidis *

Department of Electrical and Computer Engineering, University of California, Santa Barbara, CA 93106, USA;
* Correspondence: hossein@ucsb.edu (H.T.); ramtin@ucsb.edu (R.P.); cthrampo@ucsb.edu (C.T.)
† This paper is an extended version of our paper published in the 23rd International Conference on Artificial Intelligence and Statistics (AISTATS), 2020.

Abstract: We study convex empirical risk minimization for high-dimensional inference in binary linear classification under both discriminative binary linear models, as well as generative Gaussian-mixture models. Our first result sharply predicts the statistical performance of such estimators in the proportional asymptotic regime under isotropic Gaussian features. Importantly, the predictions hold for a wide class of convex loss functions, which we exploit to prove bounds on the best achievable performance. Notably, we show that the proposed bounds are tight for popular binary models (such as signed and logistic) and for the Gaussian-mixture model by constructing appropriate loss functions that achieve it. Our numerical simulations suggest that the theory is accurate even for relatively small problem dimensions and that it enjoys a certain universality property.

Keywords: signal processing in machine learning; statistics; optimization

Citation: Taheri, H.; Pedarsani, R.; Thrampoulidis, C. Sharp Guarantees and Optimal Performance for Inference in Binary and Gaussian-Mixture Models. *Entropy* **2021**, *23*, 178. https://doi.org/10.3390/e23020178

Academic Editor: Nariman Farsad, Marco Mondelli and Morteza Mardani
Received: 10 December 2020
Accepted: 26 January 2021
Published: 30 January 2021

1. Introduction

1.1. Motivation

Classical estimation theory studies problems in which the number of unknown parameters n is small compared to the number of observations m. In contrast, modern inference problems are typically *high-dimensional*, that is n can be of the same order as m. Examples are abundant in a wide range of signal processing and machine learning applications such as medical imaging, wireless communications, recommendation systems, etc. Classical tools and theories are not applicable in these modern inference problems [1]. As such, over the last two decades or so, the study of high-dimensional estimation problems has received significant attention.

Perhaps the most well-studied setting is that of noisy linear observations (namely, linear regression). The literature on the topic is vast with remarkable contributions from the statistics, signal processing and machine learning communities. Several recent works focus on the *proportional/linear asymptotic regime* and derive *sharp* results on the inference performance of appropriate convex optimization methods (e.g., [2–23]). These works show that, albeit challenging, *sharp* results are advantageous over loose order-wise bounds. Not only do they allow for accurate comparisons between different choices of the optimization parameters, but they also form the basis for establishing optimal such choices as well as fundamental performance limitations (e.g., [12,14–16,24–26]).

This paper takes this recent line of work a step further by demonstrating that results of this nature can be achieved in binary observation models. While we depart from the previously studied linear regression model, we remain faithful to the requirement and promise of sharp results. Binary models are popularly applicable in a wide range of signal-processing (e.g., highly quantized measurements) and machine learning (e.g., binary classification) problems. We derive sharp asymptotics for a rich class of convex optimization estimators, which include least-squares, logistic regression and hinge loss

as special cases. Perhaps more interestingly, we use these results to derive fundamental performance limitations and design optimal loss functions that provably outperform existing choices. Our results hold both for discriminative and generative data models.

In Section 1.2, we formally introduce the problem setup. The paper's main contributions and organization are presented in Section 1.4. A detailed discussion of prior art follows in Section 1.5.

Notation 1. *The symbols* $\mathbb{P}(\cdot)$, $\mathbb{E}[\cdot]$ *and* $Var[\cdot]$ *denote probability, expectation and variance, respectively. We use boldface notation for vectors.* $\|\mathbf{v}\|_2$ *denotes the Euclidean norm of a vector* $\mathbf{v}$. *We write* $i \in [m]$ *for* $i = 1, 2, \ldots, m$. *When writing* $x_* = \arg\min_x f(x)$, *we let the operator* arg min *return any one of the possible minimizers of* f. *For all* $x \in \mathbb{R}$, $\Phi(x)$ *is the cumulative distribution function of standard normal and Gaussian Q-function at* x *is defined as* $Q(x) = 1 - \Phi(x)$.

1.2. Data Models

Consider m data pairs $(y_i, \mathbf{a}_i)_{i=1}^m$ generated i.i.d from one of the following two models such that $y_i \in \{-1, +1\}$ and $\mathbf{a}_i \in \mathbb{R}^n$ for all $i \in [m]$.

Binary models with Gaussian features: Here, the feature/measurement vectors $\mathbf{a}_i, i \in [n]$ have i.i.d Gaussian entries, i.e., $\mathbf{a}_i \sim \mathcal{N}(0, \mathbf{I}_n)$. Given the feature vector $\mathbf{a}_i$, the corresponding label takes the form

$$y_i = f(\mathbf{a}_i^T \mathbf{x}_0), \quad i \in [m], \tag{1}$$

for some unknown true signal $\mathbf{x}_0 \in \mathbb{R}^n$ and a label/link function $f : \mathbb{R} \to \{-1, +1\}$ a (possibly random) binary function. Some popular examples for the label function f include the following:

- *(Noisy) Signed*: $\begin{cases} \text{sign}(\mathbf{a}_i^T \mathbf{x}_0) & \text{, w.p. } 1-\varepsilon, \\ -\text{sign}(\mathbf{a}_i^T \mathbf{x}_0) & \text{, w.p. } \varepsilon, \end{cases}$ where $\varepsilon \in [0, 1/2]$.
- *Logistic*: $y_i = \begin{cases} +1 & \text{, w.p. } \frac{1}{1+\exp(-\mathbf{a}_i^T \mathbf{x}_0)}, \\ -1 & \text{, w.p. } 1 - \frac{1}{1+\exp(-\mathbf{a}_i^T \mathbf{x}_0)}. \end{cases}$
- *Probit*: $y_i = \begin{cases} +1 & \text{, w.p. } \Phi(\mathbf{a}_i^T \mathbf{x}_0), \\ -1 & \text{, w.p. } 1 - \Phi(\mathbf{a}_i^T \mathbf{x}_0). \end{cases}$

We remark that when the signal strength $\|\mathbf{x}_0\|_2 \to +\infty$, logistic and Probit label functions approach the signed model (i.e., noisy-signed function with $\varepsilon = 0$).

Throughout, we assume that $\|\mathbf{x}_0\|_2 = 1$. This assumption is without loss of generality since the norm of $\mathbf{x}_0$ can always be absorbed in the link function. Indeed, letting $\|\mathbf{x}_0\|_2 = r$, we can always write the measurements as $f(\mathbf{a}^T \mathbf{x}_0) = \widetilde{f}(\mathbf{a}^T \widetilde{\mathbf{x}}_0)$, where $\widetilde{\mathbf{x}}_0 = \mathbf{x}_0 / r$ (hence, $\|\widetilde{\mathbf{x}}_0\|_2 = 1$) and $\widetilde{f}(t) = f(rt)$. We make no further assumptions on the distribution of the true vector $\mathbf{x}_0$.

Gaussian-mixture model: In Section 5, we also study the following generative Gaussian-mixture model (GMM):

$$y_i = \begin{cases} +1 & \text{, w.p. } \pi, \\ -1 & \text{, w.p. } 1-\pi, \end{cases} \quad , \quad \mathbf{a}_i | y_i \sim \mathcal{N}(y_i \mathbf{x}_0, \mathbf{I}_n), \quad i \in [m]. \tag{2}$$

Above, $\pi \in [0, 1]$ is the prior of class $+1$ and $\mathbf{x}_0 \in \mathbb{R}^n$ is the true signal, which here represents the mean of the features.

1.3. Empirical Risk Minimization

We study the performance of *empirical-risk minimization (ERM)* estimators $\widehat{\mathbf{x}}_\ell$ of $\mathbf{x}_0$ that solve the following optimization problem for some *convex* loss function $\ell : \mathbb{R} \to \mathbb{R}$

$$\widehat{\mathbf{x}}_\ell := \arg\min_{\mathbf{x}} \frac{1}{m} \sum_{i=1}^{m} \ell(y_i \mathbf{a}_i^T \mathbf{x}). \tag{3}$$

Loss function. Different choices for ℓ lead to popular specific estimators including the following:

- *Least Squares (LS):* $\ell(t) = (t-1)^2$,
- *Least-Absolute Deviations (LAD):* $\ell(t) = |t-1|$,
- *Logistic Loss:* $\ell(t) = \log(1+\exp(-t))$,
- *Exponential Loss:* $\ell(t) = \exp(-t)$,
- *Hinge Loss:* $\ell(t) = \max\{1-t\,,\,0\}$.

Performance Measure. We measure performance of the estimator $\widehat{\mathbf{x}}_\ell$ by the value of its correlation to $\mathbf{x}_0$, i.e.,

$$\text{corr}(\,\widehat{\mathbf{x}}_\ell\,;\,\mathbf{x}_0\,) := \frac{\langle \widehat{\mathbf{x}}_\ell, \mathbf{x}_0 \rangle}{\|\widehat{\mathbf{x}}_\ell\|_2 \|\mathbf{x}_0\|_2} \in [-1,1]. \tag{4}$$

Obviously, we seek estimates that maximize correlation. While correlation is the measure of primal interest, our results extend rather naturally to other prediction metrics, such as classification error given by (e.g., see [27] (Section D.2.)),

$$\mathcal{E}_\ell := \mathbb{E}_{\mathbf{a},y}\Big[\mathbb{1}_{\{y \neq \text{sign}(\langle \widehat{\mathbf{x}}_\ell, \mathbf{a} \rangle)\}}\Big]. \tag{5}$$

Expectation in (5) is derived based on a test sample $(\mathbf{a}, y)$ from the same distribution of the training set.

1.4. Contributions and Organization

As mentioned, our techniques naturally apply to both binary Gaussian and Gaussian-mixture models. For concreteness, we focus our presentation on the former models (see Sections 2–4.1). Then, we extend our results to Gaussian mixtures in Section 5. Numerical simulations corroborating our theoretical findings for both models are presented in Section 6.

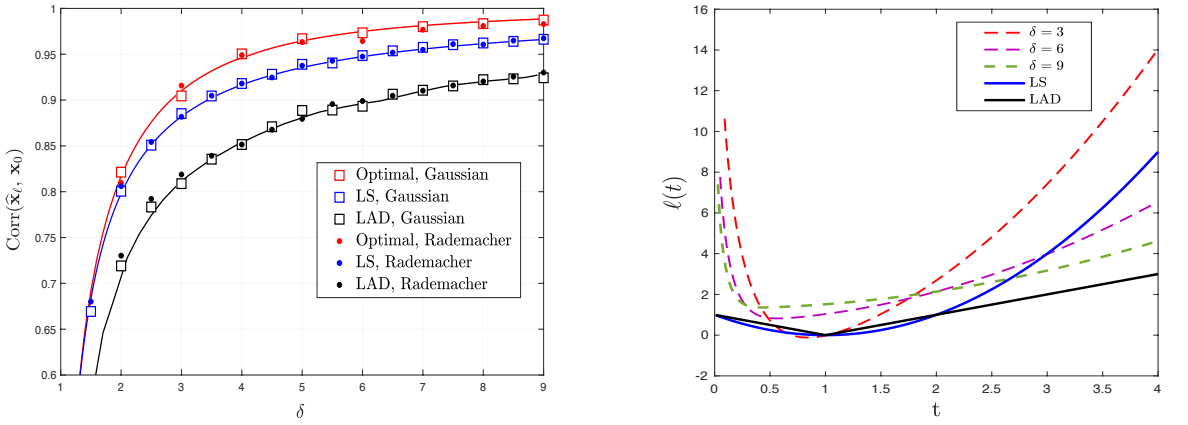
Now, we state the paper's main contributions:

- **Precise Asymptotics**: We show that the absolute value of correlation of $\widehat{\mathbf{x}}_\ell$ to the true vector $\mathbf{x}_0$ is sharply predicted by $\sqrt{1/(1+\sigma_\ell^2)}$ where the "effective noise" parameter σ_ℓ can be explicitly computed by solving a system of three non-linear equations in three unknowns. We find that the system of equations (and, thus, the value of σ_ℓ) depends on the loss function ℓ through its Moreau envelope function. Our prediction holds in the linear asymptotic regime in which $m, n \to \infty$ and $m/n \to \delta > 1$ (see Section 2).
- **Fundamental Limits**: We establish fundamental limits on the performance of convex optimization-based estimators by computing an upper bound on the best possible correlation performance among all convex loss functions. We compute the upper bound by solving a certain nonlinear equation and we show that such a solution exists for all $\delta > 1$ (see Section 3.1).
- **Optimal Performance and (sub)-optimality of LS for binary models**: For certain binary models including signed and logistic, we find the loss functions that achieve the optimal performance, i.e., they attain the previously derived upper bound (see Section 3.2). Interestingly, for logistic and Probit models with $\|\mathbf{x}_0\|_2 = 1$, we prove

that the correlation performance of least-squares (LS) is at least as good 0.9972 and 0.9804 times the optimal performance. However, as $\|\mathbf{x}_0\|_2$ grows large, logistic and Probit models approach the signed model, in which case LS becomes sub-optimal (see Section 4.1).

- **Extension to the Gaussian-Mixture Model**: In Section 5, we extend the fundamental limits and the system of equations to the Gaussian-mixture model. Interestingly, our results indicate that, for this model, LS is optimal among all convex loss functions for all $\delta > 1$.
- **Numerical Simulations**: We do numerous experiments to specialize our results to popular models and loss functions, for which we provide simulation results that demonstrate the accuracy of the theoretical predictions (see Section 6 and Appendix E).

Figure 1 contains a pictorial preview of our results described above for the special case of signed measurements. First, Figure 1a depicts the correlation performance of LS and LAD estimators as a function of the aspect ratio δ. Both theoretical predictions and numerical results are shown; note the close match between theory and empirical results for both i.i.d. Gaussian (shown by circles) and i.i.d. Rademacher (shown by squares) distributions of the feature vectors for even small dimensions. Second, the red line on the same figure shows the upper bound derived in this paper—there is no convex loss function that results in correlation exceeding this line. Third, we show that the upper bound can be achieved by the loss functions depicted in Figure 1b for several values of δ. We solve (3) for this choice of loss functions using gradient descent and numerically evaluate the achieved correlation performance. The recorded values are compared in Table 1 to the corresponding values of the upper bound; again, note the close agreement between the values as predicted by the findings of this paper, which suggests that the fundamental limits derived in this paper hold for sub-Gaussian features. We present corresponding results for the logistic and Probit models in Section 6 and for the noisy-signed model in Appendix E.



(**a**) (**b**)

Figure 1. (**a**) Comparison between theoretical (solid lines) and empirical (markers) performance for least-squares (LS) and least-absolute deviations (LAD), as predicted by Theorem 1, and the optimal performance, as predicted by the upper bound of Theorem 2, for the signed model. The squares and circles denote the empirical performance for Gaussian and Rademacher features, respectively. (**b**) Illustrations of optimal loss functions for the signed model for different values of δ according to Theorem 3.

Table 1. Theoretical predictions and empirical performance of the optimal loss function for signed model. Empirical results are averaged over 20 experiments for $n = 128$.

δ	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**
Predicted Performance	0.8168	0.9101	0.9457	0.9645	0.9748	0.9813	0.9855	0.9885
Empirical (Gaussian)	0.8213	0.9045	0.9504	0.9669	0.9734	0.9801	0.9834	0.9873
Empirical (Rademacher)	0.8096	0.9158	0.9490	0.9633	0.9644	0.9768	0.9808	0.9829

A remark on the Gaussianity assumption. Our results on precise asymptotics (to which our study of fundamental limits rely upon) hold rigorously for the two data models in Section 1.2, in which the feature vectors have entries i.i.d. standard Gaussian. However, we conjecture that the Gaussianity assumption can be relaxed. As partial numerical evidence, note in Figure 1a the perfect match of our theory with the empirical performance over data in which the feature vectors $\mathbf{a}_i, i \in [m]$ have entries i.i.d. Rademacher (i.e., centered Bernoulli with probability 1/2). Figure 2 shows corresponding results for the Gaussian-mixture model. Our conjecture that the so-called *universality* property holds in our setting is also in line with similar numerical observations and partial theoretical evidence previously made for linear regression settings [7,28–31]. A formal proof of universality of our results is beyond the scope of this paper. However, we remark that, as long as the asymptotic predictions of Section 2 enjoy this property, then all our results on fundamental performance limits and optimal functions automatically hold under the same relaxed assumptions.

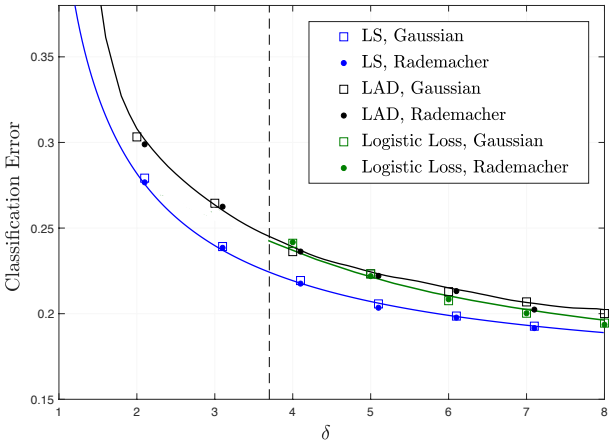


Figure 2. Theoretical (solid lines) and empirical (markers) results of classification risk in GMM as in Theorem 4 and (39) for LS, LAD and logistic loss functions as a function of δ for $r = 1$. The vertical line represents the threshold $\delta^\star \approx 3.7$ as evaluated by (36). Logistic loss gives unbounded solution if and only if $\delta < \delta^\star$.

1.5. Related Works

Over the past two decades, there has been a long list of works that derive statistical guarantees for high-dimensional estimation problems. Many of these are concerned with convex optimization-based inference methods. Our work is most closely related to the following three lines of research.

(a) Sharp asymptotics for linear measurements.

Most of the results in the literature of high-dimensional statistics are order-wise in nature. Sharp asymptotic predictions have only more recently appeared in the literature for the case of noisy linear measurements with Gaussian measurement vectors. There are by now three different approaches that have been used towards asymptotic analysis of convex regularized estimators: (i) the one that is based on the approximate message passing (AMP) algorithm and its state-evolution analysis (e.g., [5,8,14,20,32–34]); (ii) the one that is based on Gaussian process (GP) inequalities, specifically on the convex Gaussian min-max Theorem (CGMT) (e.g., [9,10,13,15,18,19]); and (iii) the "leave-one-out" approach [11,35]. The three approaches are quite different to each other and each comes with its unique distinguishing features and disadvantages. A detailed comparison is beyond our scope.

Our results in Theorems 2 and 3 for achieving the best performance across all loss functions is complementary to [12] (Theorem 1) and the work of Advani and Ganguli [16], who proposed a method for deriving optimal loss function and measuring its performance, albeit for *linear* models. Instead, we study binary models. The optimality of regularization for linear measurements is recently studied in [22].

In terms of analysis, we follow the GP approach and build upon the CGMT. Since the previous works are concerned with linear measurements, they consider estimators that solve minimization problems of the form

$$\widehat{\mathbf{x}} := \arg\min_{\mathbf{x}} \sum_{i=1}^{m} \widetilde{\ell}(y_i - \mathbf{a}_i^T \mathbf{x}) + rR(\mathbf{x}) \tag{6}$$

Specifically, the loss function $\widetilde{\ell}$ penalizes the residual. In this paper, we show that the CGMT is applicable to optimization problems in the form of (3). For our case of binary observations, (3) is more general than (6). To see this, note that, for $y_i \in \pm 1$ and popular symmetric loss functions $\widetilde{\ell}(t) = \widetilde{\ell}(-t)$, e.g., least-squares (LS), (3) results in (6) by choosing $\ell(t) = \widetilde{\ell}(t-1)$ in the former. Moreover, (3) includes several other popular loss functions such as the logistic loss and the hinge loss which cannot be expressed by (6).

(b) One-bit compressed sensing.

Our work naturally relates to the literature on one-bit compressed sensing (CS) [36]. The vast majority of performance guarantees for one-bit CS are order-wise in nature (e.g., [37–42]). To the best of our knowledge, the only existing sharp results are presented in [43] for Gaussian measurement vectors, which studies the asymptotic performance of regularized LS. Our work can be seen as a direct extension of the work in [43] to loss functions beyond least-squares (see Section 4.1 for details).

Similar to the generality of our paper, Genzel [41] also studied the high-dimensional performance of general loss functions. However, in contrast to our results, their performance bounds are loose (order-wise); as such, they are not informative about the question of optimal performance which we also address here.

(c) Classification in high-dimensions.

In [44,45], the authors studied the high-dimensional performance of maximum-likelihood (ML) estimation for the logistic model. The ML estimator is a special case of (3) and we consider general binary models. In addition, their analysis is based on the AMP framework. The asymptotics of logistic loss under different classification models is also recently studied in [46]. In yet another closely related recent work [47], the authors extended the results of Sur and Candès [45] to regularized ML by using the CGMT. Instead, we present results for general convex loss functions and for binary linear models. Importantly, we also study performance bounds and optimal loss functions.

We also remark on the following closely related parallel works. While the conference version of this paper was being reviewed, the CGMT was applied by Montanari et al. [48] and Deng et al. [49] to determine the generalization performance of max-margin linear classifiers in a binary classification setting. In essence, these results are complementary to the results of our paper in the following sense. Consider a binary classification setting under the logistic model and Gaussian regressors. As discussed in Section 4.2, the optimal set of (3) is bounded with probability approaching one if and only if $\delta > \delta_f^\star$, for appropriate threshold $\delta_f^\star$ determined for first time in [44] (see also Figure 3a). Our results hold in this regime. In contrast, the papers by Montanari et al. [48] and Deng et al. [49] study the regime $\delta < \delta_f^\star$.

We close this section by mentioning works that build on our results and appeared after the initial submission of this paper. The paper by Mignacco et al. [50] studies sharp asymptotics of ridge-regularized ERM with an intercept for Gaussian-mixture models. In [27], we extend the results of this paper on fundamental limits and optimality to the case of ridge-regularized ERM (see also the concurrent work by Aubin et al. [51]).

2. Sharp Performance Guarantees

2.1. Definitions

Moreau Envelopes. Before stating the first result, we need a definition. We write

$$\mathcal{M}_\ell(x;\lambda) := \min_v \frac{1}{2\lambda}(x-v)^2 + \ell(v),$$

for the *Moreau envelope function* of the loss $\ell : \mathbb{R} \to \mathbb{R}$ at x with parameter $\lambda > 0$. The minimizer (which is unique by strong convexity) is known as the *proximal operator* of ℓ at x with parameter λ and we denote it as $\text{prox}_\ell(x;\lambda)$. A useful property of the Moreau envelope function is that it is continuously differentiable with respect to both x and λ [52]. We denote these derivatives as follows

$$\mathcal{M}'_{\ell,1}(x;\lambda) := \frac{\partial \mathcal{M}_\ell(x;\lambda)}{\partial x},$$
$$\mathcal{M}'_{\ell,2}(x;\lambda) := \frac{\partial \mathcal{M}_\ell(x;\lambda)}{\partial \lambda}.$$

2.2. A System of Equations

As we show shortly the asymptotic performance of the optimization in (3) is tightly connected to the solution of a certain system of nonlinear equations, which we introduce here. Specifically, define random variables G, S and Y as follows:

$$G, S \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0,1) \quad \text{and} \quad Y = f(S), \tag{7}$$

and consider the following system of non-linear equations in three unknowns $(\mu, \alpha \geq 0, \lambda \geq 0)$:

$$\mathbb{E}\Big[Y\, S \cdot \mathcal{M}'_{\ell,1}(\alpha G + \mu S Y; \lambda) \Big] = 0, \tag{8a}$$
$$\lambda^2\, \delta\, \mathbb{E}\Big[\Big(\mathcal{M}'_{\ell,1}(\alpha G + \mu S Y; \lambda) \Big)^2 \Big] = \alpha^2, \tag{8b}$$
$$\lambda\, \delta\, \mathbb{E}\Big[G \cdot \mathcal{M}'_{\ell,1}(\alpha G + \mu S Y; \lambda) \Big] = \alpha. \tag{8c}$$

The expectations are with respect to the randomness of the random variables G, S and Y. We remark that the equations are well defined even if the loss function ℓ is not differentiable. In Appendix A, we summarize some well-known properties of the Moreau envelope function and use them to simplify (8) for differentiable loss functions.

2.3. Asymptotic Prediction

We are now ready to state our first main result.

Theorem 1 (Sharp Asymptotics)**.** *Assume data generated from the binary model with Gaussian features and assume $\delta > 1$ such that the set of minimizers in (3) is bounded and the system of Equation (8) has a unique solution $(\mu, \alpha \geq 0, \lambda \geq 0)$, such that $\mu \neq 0$. Let $\widehat{\mathbf{x}}_\ell$ be as in (3). Then, in the limit of $m, n \to +\infty$, $m/n \to \delta$, it holds with probability one that*

$$\lim_{n\to\infty} \text{corr}(\,\widehat{\mathbf{x}}_\ell\,;\, \mathbf{x}_0\,) = \frac{\mu}{\sqrt{\mu^2 + \alpha^2}}. \tag{9}$$

Moreover,

$$\lim_{n\to\infty} \left\| \widehat{\mathbf{x}}_\ell - \mu \cdot \frac{\mathbf{x}_0}{\|\mathbf{x}_0\|_2} \right\|_2^2 = \alpha^2. \tag{10}$$

Theorem 1 holds for any convex loss function. In Section 4, we specialize the result to specific popular choices and also present numerical simulations that confirm the validity of the predictions (see Figures 1a, 3a, 4a and A4a,b). Before that, we include a few remarks on the conditions, interpretation and implications of the theorem. The proof is deferred to Appendix B and uses the convex Gaussian min-max theorem (CGMT) [13,15].

Remark 1 (The Role of μ and α)**.** *According to* (9)*, the prediction for the limiting behavior of the correlation value is given in terms of an effective noise parameter* $\sigma_\ell := \alpha/\mu$*, where* μ *and* α *are unique solutions of* (8)*.* The smaller is the value of σ_ℓ is, the larger the correlation value becomes. *While the correlation value is fully determined by the ratio of* α *and* μ*, their individual role is clarified in* (10)*. Specifically, according to* (10)*,* $\widehat{\mathbf{x}}_\ell$ *is a biased estimate of the true* $\mathbf{x}_0$ *and* μ *represents exactly the correlation bias term. In other words, solving* (3) *returns an estimator that is close to a* μ*-scaled version of* $\mathbf{x}_0$*. When* $\mathbf{x}_0$ *and* $\widehat{\mathbf{x}}_\ell$ *are scaled appropriately, the* ℓ_2*-norm of their difference converges to* α*.*

Remark 2 (Why $\delta > 1$)**.** *The theorem requires that* $\delta > 1$ *(equivalently,* $m > n$ *asymptotically). Here, we show that this condition is* necessary *for Equations* (8) *to have a bounded solution. To see this, take squares in both sides of* (8c) *and divide by* (8b) *to find that*

$$\delta = \frac{\mathbb{E}\Big[\Big(\mathcal{M}'_{\ell,1}(\alpha G + \mu SY;\lambda)\Big)^2\Big]}{\Big(\mathbb{E}\Big[G\cdot\mathcal{M}'_{\ell,1}(\alpha G + \mu SY;\lambda)\Big]\Big)^2} \geq 1.$$

The inequality follows by applying Cauchy–Schwarz and using the fact that $\mathbb{E}[G^2] = 1$*.*

Remark 3 (On the Existence of a Solution to (8))**.** *While* $\delta > 1$ *is a necessary condition for the equations in* (8) *to have a solution, it is* not *sufficient in general. This depends on the specific choice of the loss function. For example, in Section 4.1, we show that, for the squared loss* $\ell(t) = (t-1)^2$*, the equations have a unique solution iff* $\delta > 1$*. On the other hand, for logistic loss and hinge loss, it is argued in Section 4.2 that there exists a threshold value* $\delta^\star_f > 2$ *such that the set of minimizers in* (3) *is unbounded if* $\delta < \delta^\star_f$*. In this case, the assumptions of Theorem 1 do not hold. We conjecture that, for these choices of loss, Equations* (8) *are solvable iff* $\delta > \delta^\star_f$*. Justifying this conjecture and further studying more general sufficient and necessary conditions under which the Equation* (8) *admit a solution is left to future work. However, in what follows, given such a solution, we prove that it is unique for a wide class of convex loss functions of interest.*

Remark 4 (On the Uniqueness of Solutions to (8))**.** *We show that, if the system of equations in* (8) *has a solution, then it is unique provided that* ℓ *is strictly convex, continuously differentiable and its derivative satisfies* $\ell'(0) \neq 0$*. For instance, this class includes the square, the logistic and the exponential losses. However, it excludes non-differentiable functions such as the LAD and hinge loss. We believe that the differentiability assumption can be relaxed without major modification in our proof, but we leave this for future work. Our result is summarized in Proposition 1 below.*

Proposition 1 (Uniqueness)**.** *Assume that the loss function* $\ell : \mathbb{R} \to \mathbb{R}$ *has the following properties: (i) it is proper strictly convex; and (ii) it is continuously differentiable and its derivative* ℓ' *is such that* $\ell'(0) \neq 0$*. Further, assume that the (possibly random) link function* f *is such that* $SY = Sf(S), S \sim \mathcal{N}(0,1)$ *has strictly positive density on the real line. The following statement is true. For any* $\delta > 1$*, if the system of equations in* (8) *has a bounded solution, then it is unique.*

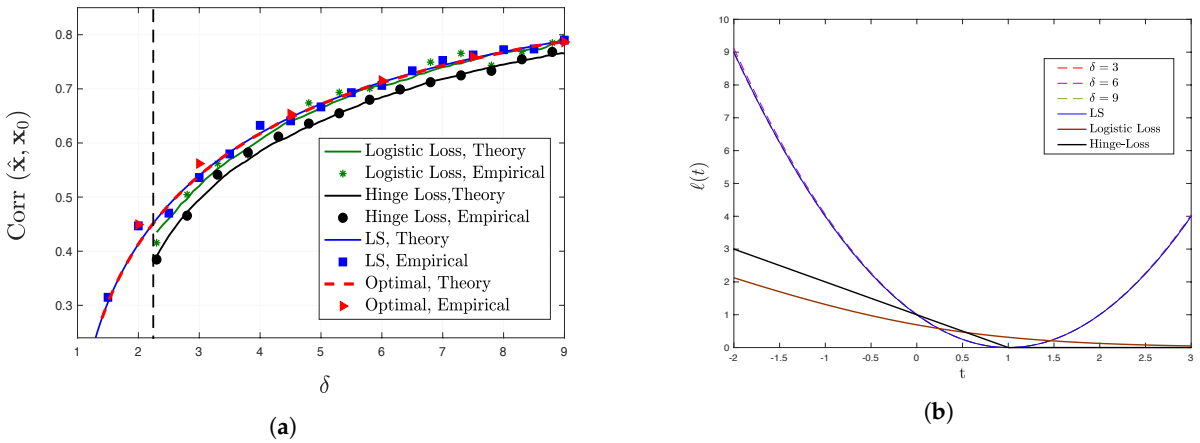
The detailed proof of Proposition 1 is deferred to Appendix B.5. Here, we highlight some key ideas. The CGMT relates—in a rather natural way—the original ERM optimization (3) to the following deterministic min-max optimization on four variables

$$\min_{\alpha>0,\mu,\tau>0} \max_{\gamma>0} F(\alpha,\mu,\tau,\gamma) := \frac{\gamma\tau}{2} - \frac{\alpha\gamma}{\sqrt{\delta}} + \mathbb{E}\left[\mathcal{M}_{\ell}\left(\alpha G + \mu Y S; \frac{\tau}{\gamma}\right)\right]. \tag{11}$$

In Appendix B.4, we show that the optimization above is convex-concave for any lower semi-continuous, proper and convex function $\ell : \mathbb{R} \to \mathbb{R}$. Moreover, it is shown that one arrives at the system of equations in (8) by simplifying the first-order optimality conditions of the min-max optimization in (11). This connection is key to the proof of Proposition 1. Indeed, we prove uniqueness of the solution (if such a solution exists) to (8), by proving instead that the function $F(\alpha,\mu,\tau,\gamma)$ above is (jointly) strictly *convex in (α,μ,τ) and* strictly *concave in γ, provided that ℓ satisfies the conditions of the proposition. Next, let us briefly discuss how strict convex-concavity of (11) can be shown. For concreteness, we only discuss strict convexity here; the ideas are similar for strict concavity. At the heart of the proof of strict convexity of F is understanding the properties of the* expected Moreau envelope function $\Omega : \mathbb{R}_+ \times \mathbb{R} \times \mathbb{R}_+ \times \mathbb{R}_+ \to \mathbb{R}$ *defined as follows:*

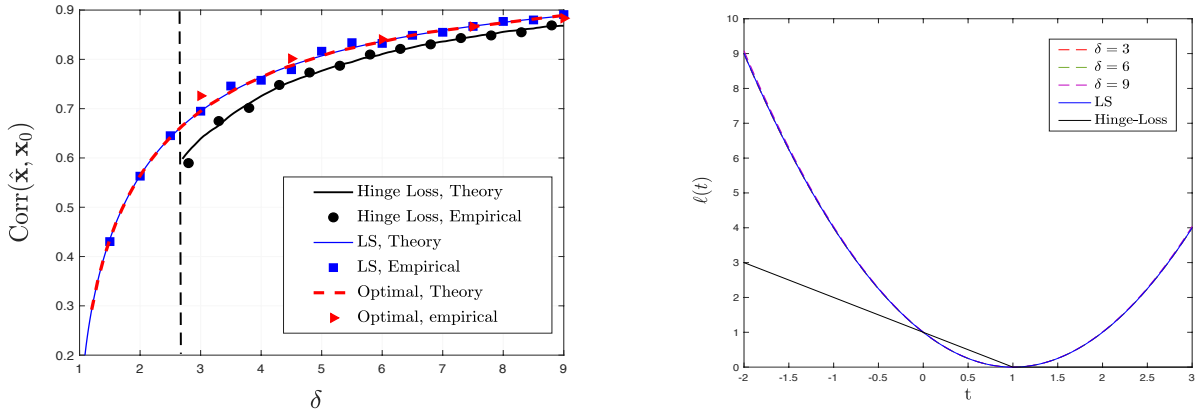
$$\Omega(\alpha,\mu,\tau,\gamma) := \mathbb{E}\left[\mathcal{M}_{\ell}\left(\alpha G + \mu Y S; \frac{\tau}{\gamma}\right)\right].$$

Specifically, we prove in Proposition A7 in Appendix A.6 that if ℓ is strictly convex, differentiable and does not attain its minimum at 0, then Ω is strictly convex in (α,μ,τ) and strictly concave in γ. It is worth noting that the Moreau envelope function $\mathcal{M}_{\ell}(\alpha g + \mu y s; \tau)$ for fixed g, s and $y = f(s)$ is not *necessarily strictly convex. Interestingly, we show that the* expected *Moreau envelope has this desired feature. We refer the reader to Appendices A.6 and B.5 for more details.*



(**a**) (**b**)

Figure 3. (**a**) Comparison between analytical and empirical results for the performance of LS, logistic loss, hinge loss and optimal loss function for logistic model. The vertical dashed line represents $\delta_f^\star \approx 2.275$, as evaluated by (35). (**b**) Illustrations of optimal loss functions for different values of δ, derived according to Theorem 3 for logistic model. To signify the similarity of optimal loss function to the LS loss, the optimal loss functions (hardly visible) are scaled such that $\ell(1) = 0$ and $\ell(2) = 1$.



(**a**) (**b**)

Figure 4. (**a**) Comparison between analytical and empirical results for the performance of LS, hinge loss and optimal loss function for Probit model. The vertical dashed line represents $\delta^\star_f \approx 2.699$, as evaluated by (35). (**b**) Illustrations of optimal loss functions for different values of δ derived according to Theorem 3 for Probit model. To signify the similarity of optimal loss function to the LS loss, the optimal loss functions (hardly visible) are scaled such that $\ell(1) = 0$ and $\ell(2) = 1$.

3. On Optimal Performance

3.1. Fundamental Limits

In this section, we establish fundamental limits on the performance of (3) by deriving an upper bound on the absolute value of correlation $\mathrm{corr}(\,\widehat{\mathbf{x}}_\ell\,;\,\mathbf{x}_0\,)$ that holds for *all* choices of loss functions satisfying Theorem 1. The result builds on the prediction of Theorem 1. In view of (9), upper bounding correlation is equivalent to lower bounding the effective noise parameter $\sigma_\ell = \alpha/\mu$. Theorem 2 derives such a lower bound.

Before stating the theorem, we need a definition. For a random variable H with density $p_H(h)$ that has a derivative $p'_H(h), \forall h \in \mathbb{R}$, we denote its score function $\xi_H(h) := \frac{\partial}{\partial h} \log p_H(h) = \frac{p'_H(h)}{p_H(h)}$. Then, the Fisher information of H, denoted by $\mathcal{I}(H) \in \mathbb{R}_+$, is defined as follows (e.g., [53] (Sec. 2)):

$$\mathcal{I}(H) := \mathbb{E}\Big[\,(\xi_H(H))^2\,\Big].$$

Theorem 2 (Best Achievable Performance)**.** *Let the assumptions and notation of Theorem 1 hold and recall the definition of random variables G, S and Y in (7). For $\sigma > 0$, define a new random variable $W_\sigma := \sigma G + SY$, and the function $\kappa : (0, \infty] \to [0, 1]$ as follows,*

$$\kappa(\sigma) := \frac{\sigma^2\big(\sigma^2\mathcal{I}(W_\sigma) + \mathcal{I}(W_\sigma) - 1\big)}{1 + \sigma^2(\sigma^2\mathcal{I}(W_\sigma) - 1)}.$$

Further, define σ_{opt} as follows,

$$\sigma_{\mathrm{opt}} := \min\left\{\sigma \geq 0 : \kappa(\sigma) = \frac{1}{\delta}\right\}. \tag{12}$$

Then, for $\sigma_\ell := \frac{\alpha}{\mu}$, it holds that $\sigma_\ell \geq \sigma_{\mathrm{opt}}$.

The theorem above establishes an upper bound on the best possible correlation performance among all convex loss functions. In Section 3.2, we show that this bound is often tight, i.e., there exists a loss function that achieves the specified best possible performance.

Remark 5. *Theorem 2 complements the results in [12], [14] (Lem. 3.4) and [15] (Rem. 5.3.3), in which the authors considered only linear regression. In particular, Theorem 2 shows that it is possible to achieve results of this nature for the more challenging setting of binary classification considered here.*

Proof of Theorem 2. Fix a loss function ℓ and let $(\mu \neq 0, \alpha > 0, \lambda \geq 0)$ be a solution to (8), which by assumptions of Theorem 1 is unique. The first important observation is that the error of a loss function is unique up to a multiplicative constant. To see this, consider an arbitrary loss function $\ell(t)$ and let $\widehat{\mathbf{x}}_\ell$ be a minimizer in (3). Now, consider (3) with the following loss function instead, for some arbitrary constants $C_1 > 0, C_2 \neq 0$:

$$\widehat{\ell}(t) := \frac{1}{C_1}\ell(C_2 t). \tag{13}$$

It is not hard to see that $\frac{1}{C_2}\widehat{\mathbf{x}}_\ell$ is the minimizer for $\widehat{\ell}$. Clearly, $\frac{1}{C_2}\widehat{\mathbf{x}}_\ell$ has the same correlation value with $\mathbf{x}_0$ as $\widehat{\mathbf{x}}_\ell$, showing that the two loss functions ℓ and $\widehat{\ell}$ perform the same. With this observation in mind, consider the function $\widehat{\ell} : \mathbb{R} \to \mathbb{R}$ such that $\widehat{\ell}(t) = \frac{\lambda}{\mu^2}\ell(\mu t)$. Then, notice that

$$\mathcal{M}'_{\ell,1}(x;\lambda) = \frac{1}{\lambda}\mathcal{M}'_{\widehat{\ell},1}(x/\mu;1).$$

Using this relation in (8) and setting $\sigma := \sigma_\ell = \alpha/\mu$, the system of equations in (8) can be equivalently rewritten in the following convenient form,

$$\mathbb{E}\Big[Y S \cdot \mathcal{M}'_{\widehat{\ell},1}(W_\sigma;1)\Big] = 0, \tag{14a}$$

$$\mathbb{E}\Big[\Big(\mathcal{M}'_{\widehat{\ell},1}(W_\sigma;1)\Big)^2\Big] = \sigma^2/\delta\,, \tag{14b}$$

$$\mathbb{E}\Big[G \cdot \mathcal{M}'_{\widehat{\ell},1}(W_\sigma;1)\Big] = \sigma/\delta\,. \tag{14c}$$

Next, we show how to use (14) to derive an equivalent system of equations based on W_σ. Starting with (14c), we have

$$\mathbb{E}\Big[G \cdot \mathcal{M}'_{\widehat{\ell},1}(W_\sigma;1)\Big] = \frac{1}{\sigma}\iint u\,\mathcal{M}'_{\widehat{\ell},1}(u+z;1)\phi_\sigma(u)p_{SY}(z)\mathrm{d}u\mathrm{d}z, \tag{15}$$

where $\phi_\sigma(u) := p_{\sigma G}(u) = \frac{1}{\sigma\sqrt{2\pi}}e^{-\frac{u^2}{2\sigma^2}}$. Since it holds that $\phi_\sigma(u) = \frac{-\sigma^2}{u}\phi'_\sigma(u)$, using (A74), it follows that

$$\begin{aligned}\mathbb{E}\Big[G \cdot \mathcal{M}'_{\widehat{\ell},1}(W_\sigma;1)\Big] &= -\sigma\iint \mathcal{M}'_{\widehat{\ell},1}(u+z;1)\phi'_\sigma(u)p_{SY}(z)\mathrm{d}u\mathrm{d}z \\ &= -\sigma\iint \mathcal{M}'_{\widehat{\ell},1}(w;1)\phi'_\sigma(u)p_{SY}(w-u)\mathrm{d}u\mathrm{d}w \\ &= -\sigma\int \mathcal{M}'_{\widehat{\ell},1}(w;1)p'_{W_\sigma}(w)\mathrm{d}w,\end{aligned} \tag{16}$$

where in the last step we use

$$p'_{W_\sigma}(w) = \int \phi'_\sigma(u)p_{SY}(w-u)\,\mathrm{d}u.$$

Therefore, we have by (16) that

$$\mathbb{E}\Big[G \cdot \mathcal{M}'_{\widehat{\ell},1}(W_\sigma;1)\Big] = -\sigma\,\mathbb{E}\Big[\mathcal{M}'_{\widehat{\ell},1}(W_\sigma;1)\xi_{W_\sigma}(W_\sigma)\Big]. \tag{17}$$

This combined with (14c) gives $\mathbb{E}\Big[\mathcal{M}'_{\widehat{\ell},1}(W_\sigma;1)\xi_{W_\sigma}(W_\sigma)\Big] = -1/\delta$. Second, multiplying (14c) with σ^2 and adding it to (14a) yields that,

$$\mathbb{E}\Big[W_\sigma \cdot \mathcal{M}'_{\widehat{\ell},1}(W_\sigma;1)\Big] = \sigma^2/\delta, \tag{18}$$

Putting these together, we conclude with the following system of equations which is equivalent to (14),

$$\mathbb{E}\Big[W_\sigma \cdot \mathcal{M}'_{\widehat{\ell},1}(W_\sigma;1)\Big] = \sigma^2/\delta\,, \tag{19a}$$

$$\mathbb{E}\Big[\Big(\mathcal{M}'_{\widehat{\ell},1}(W_\sigma;1)\Big)^2\Big] = \sigma^2/\delta\,, \tag{19b}$$

$$\mathbb{E}\Big[\mathcal{M}'_{\widehat{\ell},1}(W_\sigma;1)\xi_{W_\sigma}(W_\sigma)\Big] = -1/\delta\,. \tag{19c}$$

Note that, for $\sigma > 0$, $\xi_{W_\sigma} = p'_{W_\sigma}/p_{W_\sigma}$ exists everywhere. This is because for all $w \in \mathbb{R}$: $p_{W_\sigma}(w) > 0$ and $p_{W_\sigma}(\cdot)$ is continuously differentiable. Combining (19a) and (19c), we derive the following equation which holds for $\alpha_1, \alpha_2 \in \mathbb{R}$,

$$\mathbb{E}\Big[(\alpha_1 W_\sigma + \alpha_2 \xi_{W_\sigma}(W_\sigma)) \cdot \mathcal{M}'_{\widehat{\ell},1}(W_\sigma;1)\Big] = \alpha_1\sigma^2/\delta - \alpha_2/\delta.$$

By Cauchy–Schwarz inequality, we have that

$$\Big(\mathbb{E}\Big[(\alpha_1 W_\sigma + \alpha_2 \xi_{W_\sigma}(W_\sigma)) \cdot \mathcal{M}'_{\widehat{\ell},1}(W_\sigma;1)\Big]\Big)^2 \leq \mathbb{E}\Big[(\alpha_1 W_\sigma + \alpha_2 \xi_{W_\sigma}(W_\sigma))^2\Big]\mathbb{E}\Big[\Big(\mathcal{M}'_{\widehat{\ell},1}(W_\sigma;1)\Big)^2\Big]. \tag{20}$$

Using the fact that $\mathbb{E}[W_\sigma \xi_{W_\sigma}(W_\sigma)] = -1$ (by integration by parts), $\mathbb{E}[(\xi_{W_\sigma}(W_\sigma))^2] = \mathcal{I}(W_\sigma)$, $\mathbb{E}[W_\sigma^2] = \sigma^2 + 1$ and (19b), the right hand side of (20) is equal to

$$\Big(\alpha_1^2(\sigma^2+1) + \alpha_2^2\,\mathcal{I}(W_\sigma) - 2\alpha_1\alpha_2\Big)\,\sigma^2/\delta.$$

Therefore, we conclude with the following inequality for σ,

$$\delta\sigma^2\Big(\alpha_1^2(\sigma^2+1) + \alpha_2^2\,\mathcal{I}(W_\sigma) - 2\alpha_1\alpha_2\Big) \geq (\alpha_1\sigma^2 - \alpha_2)^2, \tag{21}$$

which holds for all $\alpha_1, \alpha_2 \in \mathbb{R}$. In particular, (21) holds for the following choice of values for α_1 and α_2:

$$\alpha_1 = \frac{1 - \sigma^2\mathcal{I}(W_\sigma)}{\delta(\sigma^2\mathcal{I}(W_\sigma) + \mathcal{I}(W_\sigma) - 1)}, \qquad \alpha_2 = \frac{1}{\delta(\sigma^2\mathcal{I}(W_\sigma) + \mathcal{I}(W_\sigma) - 1)}.$$

(The choice above is motivated by the result of Section 3.2; see Theorem 3). Rewriting (21) with the chosen values of α_1 and α_2 yields the following inequality,

$$\frac{1}{\delta} \leq \frac{\sigma^2(\sigma^2\mathcal{I}(W_\sigma) + \mathcal{I}(W_\sigma) - 1)}{1 + \sigma^2(\sigma^2\mathcal{I}(W_\sigma) - 1)} = \kappa(\sigma), \tag{22}$$

where on the right-hand side above, we recognize the function κ defined in the theorem.

Next, we use (22) to show that σ_{opt} defined in (12) yields a lower bound on the achievable value of σ. For the sake of contradiction, assume that $\sigma < \sigma_{\text{opt}}$. By the above, $1/\delta \leq \kappa(\sigma)$. Moreover, by the definition of σ_{opt}, we must have that $1/\delta < \kappa(\sigma)$. Since $\kappa(0) = 0$ and $\kappa(\cdot)$ is a continuous function we conclude that for some $\sigma_1 \in (0, \sigma)$, it holds

that $\kappa(\sigma_1) = 1/\delta$. Therefore, for $\sigma_1 < \sigma_{\text{opt}}$, we have $\kappa(\sigma_1) = 1/\delta$, which contradicts the definition of σ_{opt}. This proves that $\sigma \geq \sigma_{\text{opt}}$, as desired.

To complete the proof, it remains to show that the equation $\kappa(\sigma) = 1/\delta$ admits a solution for all $\delta > 1$. For this purpose, we use the continuous mapping theorem and the fact that the Fisher information is a continuous function [54]. Recall that, for two independent and non-constant random variables, it holds that $\mathcal{I}(X+Y) < \mathcal{I}(X)$ [53] (Eq. 2.18). Since G and SY are independent random variables, we find that $\mathcal{I}(\sigma G + SY) < \mathcal{I}(SY)$ which implies that $\mathcal{I}(\sigma G + SY)$ is uniformly bounded for all values of σ. Therefore,

$$\lim_{\sigma \to 0} \kappa(\sigma) = \lim_{\sigma \to 0} \frac{\sigma^2\big(\sigma^2 \mathcal{I}(W_\sigma) + \mathcal{I}(W_\sigma) - 1\big)}{1 + \sigma^2(\sigma^2 \mathcal{I}(W_\sigma) - 1)} = 0.$$

Furthermore, $\sigma^2 \mathcal{I}(\sigma G + SY) = \mathcal{I}(G + \frac{1}{\sigma} SY) \to \mathcal{I}(G) = 1$ when $\sigma \to \infty$. Hence,

$$\lim_{\sigma \to \infty} \kappa(\sigma) = \lim_{\sigma \to \infty} \frac{\sigma^2\big(\sigma^2 \mathcal{I}(W_\sigma) + \mathcal{I}(W_\sigma) - 1\big)}{1 + \sigma^2(\sigma^2 \mathcal{I}(W_\sigma) - 1)} = 1.$$

Note that $\sigma^2 \mathcal{I}(\sigma G + SY) < \sigma^2 \mathcal{I}(\sigma G) = 1$, which further yields that $\kappa(\sigma) < 1$ for all $\sigma \geq 0$. Finally, since $\mathcal{I}(\cdot)$ is a continuous function, we deduce that range of $\kappa : \mathbb{R}^+ \cup 0 \to \mathbb{R}$ is $[0, 1)$, implying the *existence* of a solution to (12) for all $\delta > 1$. This completes the proof of Theorem 2. □

A useful closed-form bound on the best achievable performance: In general, determining σ_{opt} requires computing the Fisher information of the random variable $\sigma G + SY$ for $\sigma > 0$. If the probability distribution of SY is continuously differentiable (e.g., logistic model; see Appendix C.2), then we obtain the following simplified bound.

Corollary 1 (Closed-form Lower Bound on σ_{opt})**.** *Let $p_{SY} : \mathbb{R} \to \mathbb{R}$ be the probability distribution of SY. If $p_{SY}(x)$ is differentiable for all $x \in \mathbb{R}$, then,*

$$\sigma^2_{\text{opt}} \geq \frac{1}{(\delta - 1)(\mathcal{I}(SY) - 1)}. \tag{23}$$

Proof. Based on Theorem 2, the following equation holds for $\sigma = \sigma_{\text{opt}}$

$$\frac{1}{\delta} = \kappa(\sigma)$$

or, equivalently, by rewriting the right-hand side,

$$\frac{1}{\delta} = 1 - \frac{1}{\frac{1}{1 - \sigma^2 \mathcal{I}(W_\sigma)} - \sigma^2}. \tag{24}$$

Define the following function

$$h(x) := 1 - \frac{1}{\frac{1}{1 - \sigma^2 x} - \sigma^2}.$$

The function h is increasing in the region $\mathcal{R}_\sigma = \{z : z > \sigma^{-2} - \sigma^{-4}\}$. According to Stam's inequality [55], for two independent random variables X and Y with continuously differentiable p_X and p_Y, it holds that

$$\mathcal{I}(X+Y) \leq \frac{\mathcal{I}(X) \cdot \mathcal{I}(Y)}{\mathcal{I}(X) + \mathcal{I}(Y)},$$

where equality is achieved if and only if X and Y are independent Gaussian random variables. Therefore, since by assumption p_{SY} is differentiable on the real line, Stam's inequality yields

$$\mathcal{I}(W_\sigma) = \mathcal{I}(\sigma G + SY) \leq \frac{\mathcal{I}(\sigma G) \cdot \mathcal{I}(SY)}{\mathcal{I}(\sigma G) + \mathcal{I}(SY)}. \tag{25}$$

Next, we prove that for all $\sigma > 0$, both sides of (25) are in the region $\mathcal{R}_\sigma$. First, we prove that $\mathcal{I}(W_\sigma) \in \mathcal{R}_\sigma$. By Cramer–Rao bound (e.g., see [53] (Eq. 2.15)) for Fisher information of a random variable X, we have that $\mathcal{I}(X) \geq 1/(\mathrm{Var}[X])$. In addition, for the random variable W_σ, we know that $\mathrm{Var}[W_\sigma] = 1 + \sigma^2 - (\mathbb{E}[SY])^2$, thus

$$\mathcal{I}(W_\sigma) \geq \frac{1}{1 + \sigma^2 - (\mathbb{E}[SY])^2}. \tag{26}$$

Using the relation $(\mathbb{E}[SY])^2 \leq \mathbb{E}[S^2]\mathbb{E}[Y^2] = 1$, one can check that the following inequality holds:

$$\frac{1}{1 + \sigma^2 - (\mathbb{E}[SY])^2} \geq \sigma^{-2} - \sigma^{-4}. \tag{27}$$

Therefore, from (26) and (27), we derive that $\mathcal{I}(W_\sigma) \in \mathcal{R}_\sigma$ for all $\sigma > 0$. Furthermore, by the inequality in (25) and the definition of $\mathcal{R}_\sigma$ it directly follows that for all $\sigma > 0$

$$\frac{\mathcal{I}(\sigma G)\,\mathcal{I}(SY)}{\mathcal{I}(\sigma G) + \mathcal{I}(SY)} \in \mathcal{R}_\sigma.$$

Finally, noting that $h(\cdot)$ is increasing in $\mathcal{R}_\sigma$, combined with (25), we have

$$\frac{1}{\delta} = h(\mathcal{I}(W_\sigma)) \leq h\left(\frac{\mathcal{I}(\sigma G) \cdot \mathcal{I}(SY)}{\mathcal{I}(\sigma G) + \mathcal{I}(SY)}\right),$$

which after using the relation $\mathcal{I}(\sigma G) = \sigma^{-2}$ and further simplification yields the inequality in the statement of the corollary. □

The proof of the corollary reveals that (23) holds with equality when SY is Gaussian. In Appendix C.2, we compute p_{SY} for the logistic and the Probit models with $\|\mathbf{x}_0\|_2 = 1$ and numerically show that it is close to the density of a Gaussian random variable. Consequently, the lower bound of Corollary 1 is almost exact when measurements are obtained according to the logistic and Probit models (see Figure A2 in the Appendix C).

3.2. On the Optimal Loss Function

It is natural to ask whether there exists a loss function that attains the bound of Theorem 2. If such a loss function exists, then we say it is *optimal* in the sense that it maximizes the correlation performance among all convex loss functions in (3).

Our next theorem derives a candidate for the optimal loss function, which we denote ℓ_{opt}. Before stating the result, we provide some intuition about the proof which builds on Theorem 2. The critical observation in the proof of Theorem 2 is that the effective noise $\sigma_{\widehat{\ell}}$ of $\widehat{\ell}$ is minimized (i.e., it attains the value σ_{opt}) if the Cauchy–Schwartz inequality in (20) holds with equality. Hence, we seek $\widehat{\ell} = \ell_{\text{opt}}$ so that for some $c \in \mathbb{R}$,

$$\mathcal{M}'_{\ell_{\text{opt}},1}(w;1) = c(\alpha_1 w + \alpha_2 \cdot \xi_{W_{\text{opt}}}(w)). \tag{28}$$

By choosing $c = -1$, integrating and ignoring constants irrelevant to the minimization of the loss function, the previous condition is equivalent to the following $\mathcal{M}_{\ell_{\text{opt}}}(w;1) = -\alpha_1 w^2/2 - \alpha_2 \log(p_{W_{\text{opt}}}(w))$. It turns out that this condition can be "inverted" to yield

the explicit formula for ℓ_{opt} as, $\ell_{\text{opt}}(w) = -\mathcal{M}_{\alpha_1 q + \alpha_2 \log(p_{W_{\text{opt}}})}(w;1)$. Of course, one has to properly choose α_1 and α_2 to make sure that this function satisfies the system of equations in (19) with $\sigma = \sigma_{\text{opt}}$. The correct choice is specified in the theorem below. The proof is deferred to Appendix D.1.

Theorem 3 (Optimal Loss Function)**.** *Recall the definition of σ_{opt} in (12). Define the random variable $W_{\text{opt}} := \sigma_{\text{opt}} G + SY$ and let $p_{W_{\text{opt}}}$ denote its density. Consider the following loss function $\ell_{\text{opt}} : \mathbb{R} \to \mathbb{R}$*

$$\ell_{\text{opt}}(w) = -\mathcal{M}_{\alpha_1 q + \alpha_2 \log(p_{W_{\text{opt}}})}(w;1), \tag{29}$$

where $q(x) = x^2/2$ and

$$\begin{aligned} \alpha_1 &= \frac{1 - \sigma_{\text{opt}}^2 \mathcal{I}(W_{\text{opt}})}{\delta(\sigma_{\text{opt}}^2 \mathcal{I}(W_{\text{opt}}) + \mathcal{I}(W_{\text{opt}}) - 1)}, \\ \alpha_2 &= \frac{1}{\delta(\sigma_{\text{opt}}^2 \mathcal{I}(W_{\text{opt}}) + \mathcal{I}(W_{\text{opt}}) - 1)}. \end{aligned} \tag{30}$$

If ℓ_{opt} defined as in (29) is convex and the equation $\kappa(\sigma) = 1/\delta$ has a unique solution, then $\sigma_{\ell_{\text{opt}}} = \sigma_{\text{opt}}$.

In general, there is no guarantee that the function $\ell_{\text{opt}}(\cdot)$ as defined in (29) is convex. However, if this is the case, the theorem above guarantees that it is optimal (Strictly speaking, the performance is optimal among all convex loss functions ℓ for which (8) has a unique solution as required by Theorem 2.). A *sufficient* condition for $\ell_{\text{opt}}(w)$ to be convex is provided in Appendix D.2. Importantly, in Appendix D.2.1, we show that this condition holds for observations following the signed model. Thus, for this case, the resulting function is convex. Although we do *not* prove the convexity of optimal loss function for the logistic and Probit models, our numerical results (e.g., see Figure 3b) suggest that this is the case. Concretely, we conjecture that the loss function ℓ_{opt} is convex for logistic and Probit models, and therefore by Theorem 3 its performance is optimal.

4. Special Cases

4.1. Least-Squares

By choosing $\ell(t) = (t-1)^2$ in (3), we obtain the standard least-squares estimate. To see this, note that since $y_i = \pm 1$, it holds for all i that $(y_i \mathbf{a}_i^T \mathbf{x} - 1)^2 = (y_i - \mathbf{a}_i^T \mathbf{x})^2$. Thus, $\widehat{\mathbf{x}}$ is minimizing the sum of squares of the residuals:

$$\widehat{\mathbf{x}} = \arg\min_{\mathbf{x}} \sum (y_i - \mathbf{a}_i^T \mathbf{x})^2. \tag{31}$$

For this choice of a loss function, we can solve the equations in (8) in closed form. Furthermore, the equations have a (unique, bounded) solution for any $\delta > 1$ provided that $\mathbb{E}[SY] > 0$. The final result is summarized in the corollary below (see Appendix F.1 for the proof).

Corollary 2 (Least-squares)**.** *Assume data generated from the binary model and $\delta > 1$. For the label function assume that $\mathbb{E}[SY] > 0$ in the notation of (7). Let $\widehat{\mathbf{x}}$ be as in (41). Then, in the limit of $m, n \to +\infty$, $m/n \to \delta$, Equations (9) and (10) hold with probability one with α and μ given as follows:*

$$\mu = \mathbb{E}[SY], \tag{32}$$

$$\alpha = \sqrt{1 - (\mathbb{E}[SY])^2} \cdot \sqrt{\frac{1}{\delta - 1}}. \tag{33}$$

Corollary 2 appears in [43] (see also [40,41,56] and Appendix F for an interpretation of the result). However, these previous works obtain results that are limited to least-squares loss. In contrast, our results are general and LS prediction is obtained as a simple corollary of our general Theorem 1. Moreover, our study of fundamental limits allows us to quantify the sub-optimality gap of least-square (LS) as follows.

On the Optimality of LS. On the one hand, Corollary 2 derives an explicit formula for the effective noise variance $\sigma_{\text{LS}} = \alpha/\mu$ of LS in terms of $E[YS]$ and δ. On the other hand, Corollary 1 provides an explicit lower bound on the optimal value σ_{opt} in terms of $\mathcal{I}(SY)$ and δ. Combining the two, we conclude that

$$\frac{\sigma_{\text{LS}}^2}{\sigma_{\text{opt}}^2} \leq \xi := (\mathcal{I}(SY) - 1)\frac{1 - (\mathbb{E}[SY])^2}{(\mathbb{E}[SY])^2}.$$

In terms of correlation,

$$\frac{\text{corr}_{\text{opt}}}{\text{corr}_{\text{LS}}} = \sqrt{\frac{1+\sigma_{\text{LS}}^2}{1+\sigma_{\text{opt}}^2}} \leq \frac{\sigma_{\text{LS}}}{\sigma_{\text{opt}}} \leq \sqrt{\xi},$$

where the first inequality follows from the fact that $\sigma_{\text{LS}} \geq \sigma_{\text{opt}}$. Therefore, the performance of LS is at least as good as $\frac{1}{\sqrt{\xi}}$ times the optimal one. In particular, assuming $\|\mathbf{x}_0\| = 1$ and for logistic and Probit models (for which Corollary 1 holds), we can explicitly compute $\frac{1}{\sqrt{\xi}} = 0.9972$ and 0.9804, respectively. However, we recall that for large $\|\mathbf{x}_0\|$ logistic and Probit models approach the signed model, and, as Figure 1a demonstrates, LS becomes suboptimal.

Another interesting consequence of combining Corollaries 1 and 2 is that LS would be optimal if SY were a Gaussian random variable. To see this, recall from Corollary 1 that, if SY is Gaussian, then:

$$\sigma_{\text{opt}}^2 = \frac{1}{(\delta - 1)(\mathcal{I}(SY) - 1)}.$$

However, for SY Gaussian, we can explicitly compute $\mathcal{I}(SY) = 1/\text{Var}[SY]$, which leads to

$$\sigma_{\text{opt}}^2 = \frac{1 - (\mathbb{E}[SY])^2}{(\mathbb{E}[SY])^2(\delta - 1)}.$$

The right hand side is exactly σ_{LS}^2. Therefore, the optimal performance is achieved by the square loss function if SY is a Gaussian random variable. Remarkably, for logistic and Probit models with small SNR (i.e., small $\|\mathbf{x}_0\|$), density of SY is close to the density of a normal random variable (see Figure A2 in the Appendix C), implying the optimality of LS for these models.

4.2. Logistic and Hinge Loss

Theorem 1 only holds in regimes for which the set of minimizers of (3) is bounded. As we show here, this is *not* always the case. Specifically, consider non-negative loss functions $\ell(t) \geq 0$ with the property $\lim_{t\to+\infty} \ell(t) = 0$. For example, the hinge, exponential and logistic loss functions all satisfy this property. Now, we show that for such loss functions the set of minimizers is unbounded if $\delta < \delta_f^\star$ for some appropriate $\delta_f^\star > 2$. First, note that the set of minimizers is unbounded if the following condition holds:

$$\exists \mathbf{x}_s \neq \mathbf{0} \quad \text{such that} \quad y_i \mathbf{a}_i^T \mathbf{x}_s \geq 0, \quad \forall i \in [m]. \tag{34}$$

Indeed, if (34) holds then $\mathbf{x} = c \cdot \mathbf{x}_s$ with $c \to +\infty$, attains zero cost in (3); thus, it is optimal and the set of minimizers is unbounded. To proceed, we rely on a recent result by Candes and Sur [44] who proved that (34) holds iff (To be precise, Candès and Sur [44] proved the statement for measurements $y_i, i \in [m]$ that follow a logistic model. Close inspection

of their proof shows that this requirement can be relaxed by appropriately defining the random variable Y in (7) (see also [48,49]).)

$$\delta \leq \delta_f^\star := \left(\min_{c \in \mathbb{R}} \mathbb{E}\left[(G + c\, S\, Y)_-^2 \right] \right)^{-1}, \tag{35}$$

where G, S and Y are random variables as in (7) and $(t)_- := \min\{0, t\}$. We highlight that logistic and hinge losses give unbounded solutions in the noisy-signed model with $\varepsilon = 0$, since the condition (34) holds for $\mathbf{x}_s = \mathbf{x}_0$. However, their performances are comparable to the optimal performance in both logistic and Probit models (see Figures 3a and 4a).

5. Extensions to Gaussian-Mixture Models

In this section, we show that our results on sharp asymptotics and lower bounds on error can be extended to include the Gaussian-Mixture model (GMM) presented in Section 1.2. The discussions on the phase transition for the existence of a bounded solution in Section 4.2 applies here as well. We rely on a phase-transition result [49] (Prop. 3.1), which proves that (34) holds if and only if

$$\delta \leq \delta^\star := \left(\min_{t \in \mathbb{R}} \mathbb{E}\left[(W_1 + t\, W_2)_-^2 \right] \right)^{-1}, \tag{36}$$

where W_1 and W_2 are random variables defined in (7) and $(x)_-^2 := (\min\{x, 0\})^2$. Therefore, for loss functions satisfying this property, e.g., hinge loss and logistic loss, the solution to (3) is unbounded if and only if $\delta \leq \delta^\star$.

5.1. System of Equations for GMM

It turns out that, similar to the generative models, the asymptotic performance of (3) for GMM depends on the loss function ℓ via its Moreau envelope. Specifically, let W_1 and W_2 be independent Gaussian random variables such that

$$W_1 \sim \mathcal{N}(0, 1),\ W_2 \sim \mathcal{N}(r, 1), \tag{37}$$

where $r := \|\mathbf{x_0}\|_2 > 0$.
Consider the following system of non-linear equations in three unknowns $(\mu, \alpha \geq 0, \lambda \geq 0)$:

$$0 = \mathbb{E}\left[W_2 \cdot \mathcal{M}'_{\ell,1}(\alpha W_1 + \mu W_2; \lambda) \right], \tag{38a}$$

$$\alpha^2 = \lambda^2\, \delta\, \mathbb{E}\left[\left(\mathcal{M}'_{\ell,1}(\alpha W_1 + \mu W_2; \lambda) \right)^2 \right], \tag{38b}$$

$$\alpha = \lambda\, \delta\, \mathbb{E}\left[W_1 \cdot \mathcal{M}'_{\ell,1}(\alpha W_1 + \mu W_2; \lambda) \right]. \tag{38c}$$

The expectations above are with respect to the randomness of the random variables W_1 and W_2.

As we show shortly, the solution to these equations is tightly connected to the asymptotic behavior of the optimization in (3).

5.2. Theoretical Prediction of Error for Convex Loss Functions

Theorem 4 (Asymptotic Prediction)**.** *Assume data generated from the Gaussian-mixture model and assume $\delta > 1$ such that the set of minimizers in (3) is bounded and the system of Equation (38)*

has a unique solution (μ, α, λ)*, such that* $\mu \neq 0$*. Let* $\widehat{\mathbf{x}}_\ell$ *be as in* (3) *and* $\sigma_\ell = \alpha/\mu$*. Then, in the limit of* $m, n \to +\infty$*,* $m/n \to \delta$*, it holds with probability one that*

$$\lim_{n\to\infty} \text{corr}(\,\widehat{\mathbf{x}}_\ell\,;\,\mathbf{x}_0\,) = \frac{\mu}{\sqrt{\mu^2+\alpha^2}}, \quad \lim_{n\to\infty} \mathcal{E}_\ell = Q\bigg(\frac{r}{\sqrt{1+\sigma_\ell^2}}\bigg), \tag{39}$$

where $\mathcal{E}_\ell$ *denotes the classification test error defined in* (5)*.*

Remark 6 (Proof of Theorem 4)**.** *The high-level steps of the proof of Theorem 4 follow closely the proof of Theorem 1. Particularly, for GMM one can show the correlation of the ERM estimate with the true vector* $\mathbf{x}_0$ *is predicted by a system of Equations as in* (38)*, only with* W_2 *replaced by a non-gaussian random variable (denoted as SY in Theorem 1). Specifically, by rotational invariance of the Gaussian feature vectors* $\mathbf{a}_i$*, we can assume, without loss of generality, that* $\mathbf{x}_0 = [r, 0, 0, ..., 0]^T$*. Then, we can can guarantee that with probability one it holds that*

$$\lim_{n\to\infty} \widehat{\mathbf{x}}_\ell(1) = \mu, \quad \lim_{n\to\infty} \sum_{j=2}^{n} \widehat{\mathbf{x}}_\ell^2(j) = \alpha^2, \tag{40}$$

where μ *and* α *are specified by* (38)*. To see how this implies* (39)*, we argue as follows. Recalling that* $\mathbf{x}|y \sim \mathcal{N}(y\mathbf{x}_0, \boldsymbol{I})$*, we have*

$$y\langle \widehat{\mathbf{x}}_\ell, \mathbf{a}\rangle \sim \mathcal{N}\Big(r\widehat{\mathbf{x}}_\ell(1), \|\widehat{\mathbf{x}}_\ell\|_2^2\Big).$$

Using this and (40) *leads to the asymptotic value of correlation and classification error as presented in* (39)*.*

Remark 7. *(On the Uniqueness of Solutions to Equations* (38)*) Our results in proving the uniqueness of solutions to the equations for generative models* (8) *in Proposition 1, extend to GMM. Noting that* $W_2 \sim \mathcal{N}(r, 1)$ *in* (38) *plays the role of SY in* (8)*, we straightforwardly deduce the following result for uniqueness of solutions to* (38)*.*

Proposition 2. *Assume that the loss function* $\ell : \mathbb{R} \to \mathbb{R}$ *has the following properties: (i) it is proper strictly convex; and (ii) it is continuously differentiable and its derivative* ℓ' *is such that* $\ell'(0) \neq 0$*. The following statement is true. For any* $\delta > 1$*, if the system of equations in* (38) *has a bounded solution, then it is unique.*

5.3. Special Case: Least-Squares

By choosing $\ell(t) = (t-1)^2$ in (3), we obtain the standard least-squares estimate. To see this, note that since $y_i = \pm 1$, it holds for all i that $(y_i \mathbf{a}_i^T \mathbf{x} - 1)^2 = (y_i - \mathbf{a}_i^T \mathbf{x})^2$.

Thus, the estimator $\widehat{\mathbf{x}}_{LS}$ is minimizing the sum of squares of the residuals:

$$\widehat{\mathbf{x}}_{LS} = \arg\min_{\mathbf{x}} \sum (y_i - \mathbf{a}_i^T \mathbf{x})^2. \tag{41}$$

For the choice $\ell(t) = (t-1)^2$, it turns out that we can solve the equations in (38) in closed form. The final result is summarized in the corollary below and proved in Appendix G.1.

Corollary 3 (Least-Squares)**.** *Let* $\widehat{\mathbf{x}}_{LS}$ *be as in* (41) *and* $\delta > 1$*. Then, in the limit of* $m, n \to +\infty$*,* $m/n \to \delta$*, Equation* (39) *holds with probability one with* σ_{LS}^2 *given as follows:*

$$\sigma_{LS}^2 = \frac{1+r^2}{r^2} \cdot \frac{1}{(\delta-1)}. \tag{42}$$

5.4. Optimal Risk for GMM

Next, we characterize the best achievable classification error by different choices of loss function. Considering (39), we see that an optimal choice of ℓ is the one that minimizes

σ_ℓ^2. The next theorem characterizes the best achievable σ_ℓ among convex loss functions by deriving an equivalent set of equations to (38) and combining them with proper coefficients. Similar to the proof of Theorem 2, a key step in the proof is properly setting up a Cauchy–Schwarz inequality that exploits the structure of the new set of equations. The proof is deferred to Appendix G.2.

Theorem 5 (Lower Bound on Risk)**.** *Under the assumptions of Theorem 4, the following inequality holds for the effective risk parameter (σ_ℓ) of a loss function ℓ:*

$$\lim_{n\to\infty} \sigma_\ell^2 \geq \sigma_\star^2 := \frac{1+r^2}{r^2} \cdot \frac{1}{\delta - 1} \tag{43}$$

Remark 8 (Optimality of Least-squares for GMM)**.** *Theorem 5 provides a lower bound for the asymptotic value of σ_ℓ which holds for all $\delta > 1$ and $r > 0$. This result together with Corollary 3 implies that least-squares achieves the least value of risk (i.e., σ_ℓ and $\mathcal{E}_\ell$) for all $\delta > 1$ and $r > 0$ among all convex loss functions ℓ for which the set of minimizers in (3) is bounded.*

6. Numerical Experiments

In this section, we present numerical simulations that validate the predictions of Theorems 1–5. To begin, we use the following three popular models as our case study: signed, logistic and Probit. We generate random measurements according to (1). Without loss of generality (due to rotational invariance of the Gaussian measure), we set $\mathbf{x}_0 = [1, 0, ..., 0]^T$. We then obtain estimates $\widehat{\mathbf{x}}_\ell$ of $\mathbf{x}_0$ by numerically solving (3) and measure performance by the correlation value $\mathrm{corr}(\widehat{\mathbf{x}}_\ell\,;\mathbf{x}_0)$. Throughout the experiments, we set $n = 128$ and the recorded values of correlation are averages over 25 independent realizations. For each label function, we first provide plots that compare results of Monte Carlo simulations to the asymptotic predictions for loss functions discussed in Section 4, as well as to the optimal performance of Theorem 2. We next present numerical results on optimal loss functions. To empirically derive the correlation of optimal loss function, we run gradient descent-based optimization with 1000 iterations. As a general comment, we note that, despite being asymptotic, our predictions appear accurate even for relatively small problem dimensions. For the analytical predictions, we apply Theorem 1. In particular, for solving the system of non-linear equations in (3), we empirically observe (see also [15,47] for similar observation) that, if a solution exists, then it can be efficiently found by the following fixed-point iteration method. Let $\mathbf{v} := [\mu, \alpha, \lambda]^T$ and $\mathcal{F} : \mathbb{R}^3 \to \mathbb{R}^3$ be such that (3) is equivalent to $\mathbf{v} = \mathcal{F}(\mathbf{v})$. With this notation, we initialize $\mathbf{v} = \mathbf{v}_0$ and for $k \geq 1$ repeat the iterations $\mathbf{v}_{k+1} = \mathcal{F}(\mathbf{v}_k)$ until convergence.

Logistic model. For the logistic model, comparison between the predicted values and the numerical results is illustrated in Figure 3a. Results are shown for LS, logistic and hinge loss functions. Note that minimizing the logistic loss corresponds to the maximum-likelihood estimator (MLE) for logistic model. An interesting observation in Figure 3a is that in the high-dimensional setting (finite δ) LS has comparable (if not slightly better) performance to MLE. Additionally, we observe that in this model, performance of LS is almost the same as the best possible performance derived according to Theorem 2. This confirms the analytical conclusion of Section 4.1. The comparison between the optimal loss function as in Theorem 3 and other loss functions is illustrated in Figure 3b. We note the obvious similarity between the shapes of optimal loss functions and LS which further explains the similarity between their performance.

Probit model. Theoretical predictions for the performance of hinge and LS loss functions are compared with the empirical results and optimal performance of Theorem 2 in Figure 4a. Similar to the logistic model, in this model, LS also outperforms hinge loss and its performance resembles the performance of optimal loss function derived according to Theorem 3. Figure 4b illustrates the shapes of LS, hinge loss and the optimal loss functions

for the Probit model. The obvious similarity between the shape of LS and optimal loss functions for all values of δ explains the close similarity of their performance.

Additionally, by comparing the LS performance for the three models in Figures 1a, 3a and 4a, it is clear that higher (respectively, lower) correlation values are achieved for signed (respectively, logistic) measurements. This behavior is indeed predicted by Corollary 2: correlation performance is higher for higher values of $\mu = \mathbb{E}[SY]$. It can be shown that, for the signed, probit and logistic models (with $\|\mathbf{x}_0\|_2 = 1$), we have $\mu = \sqrt{2/\pi}, \sqrt{1/\pi}$ and 0.4132, respectively.

Optimal loss function. By putting together Theorems 2 and 3, we obtain a method on deriving the optimal loss function for generative binary models. This requires the following steps.

1. Find σ_{opt} by solving (12).
2. Compute the density of $W_{\text{opt}} = \sigma_{\text{opt}} G + SY$.
3. Compute ℓ_{opt} according to (29).

Note that computing σ_{opt} needs the density function p_W of the random variable $W = \sigma G + SY$. In principle p_W can be calculated as the convolution of the Gaussian density with the pdf p_{SY} of SY. Moreover, it follows from the recipe above that the optimal loss function depends on δ in general. This is because σ_{opt} itself depends on δ via (12).

Numerical Experiments for GMM

Theorem 5 implies the optimality of least-squares among convex loss functions in the under-parameterized regime $\delta > 1$. In Figure 2, we demonstrate the classification risk of least-squares alongside other well-known loss functions LAD and logistic, for $r = 1$. Solid lines correspond to the theoretical predictions of Theorem 4. For least-squares we rely on the result of Corollary 3 and for LAD and logistic loss, the system of equations are solved by iterating over the equations, where we observe that after relatively small number of iterations the triple (μ, α, λ) converges to $(\mu^\star, \alpha^\star, \lambda^\star)$. We use 10^5 and 10^3 samples to compute the expectations in (38) for LAD and logistic loss, respectively. After deriving $\sigma_\ell = \alpha/\mu$, the classification risk $\mathcal{E}_\ell$ is obtained according to the formula in (39). Dots correspond to the empirical evaluations of the classification risk of loss functions for $n = 60$ and for different values of $\delta = m/n > 1$. The resulting numbers are averaged over 30 independent experiments. As is observed, the empirical results closely follow the theoretical predictions of Theorem 4. Furthermore, as predicted by Theorem 5, least-squares has the minimum expected classification risk among other convex loss functions and for all $\delta > 1$.

7. Conclusions

We derive theoretical predictions for the generalization error of estimators obtained by ERM for generative binary models and a Gaussian Mixture model. Furthermore, we use this theoretical characterizations to find the optimal performance and optimal loss function among all convex losses. Although our analysis is true for Gaussian matrices, we empirically show they hold for sub-Gaussian matrices as well. As an exciting future direction, we plan to extend our analysis on sharp asymptotics and optimal loss function to non-isotropic (Gaussian) features with arbitrary covariance. A more challenging, albeit interesting, direction is going beyond (binary) linear models studied in this paper, by considering asymptotics and optimal error for kernel models and neural networks (see [48,57] for partial progress in this direction).

Author Contributions: Formal analysis, H.T., R.P. and C.T.; Funding acquisition, R.P. and C.T.; Investigation, H.T., R.P. and C.T.; Methodology, H.T., R.P. and C.T.; Project administration, H.T., R.P. and C.T.; Software, H.T.; Supervision, H.T., R.P. and C.T.; Writing—original draft, H.T., R.P. and C.T.; Writing—review & editing, H.T., R.P. and C.T. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by NSF CNS-2003035, NSF CCF-2009030 and NSF CCF-1909320.

Conflicts of Interest: The authors declare no conflict of interest.

Appendix A. Properties of Moreau Envelopes

Appendix A.1. Derivatives

Recall the definition of the Moreau envelope $\mathcal{M}_\ell(x;\lambda)$ and proximal operator $\text{prox}_\ell(x;\lambda)$ of a function ℓ:

$$\mathcal{M}_\ell(x;\lambda) = \min_y \frac{1}{2\lambda}(x-y)^2 + \ell(y), \tag{A1}$$

and $\text{prox}_\ell(x;\lambda) = \arg\min_y \frac{1}{2\lambda}(x-y)^2 + \ell(y)$.

Proposition A1 (Basic properties of $\mathcal{M}_\ell$ and prox_ℓ [52])**.** *Let $\ell : \mathbb{R} \to \mathbb{R}$ be lower semi-continuous (lsc), proper and convex. The following statements hold for any $\lambda > 0$.*

(a) The proximal operator $\text{prox}_\ell(x;\lambda)$ is unique and continuous. In fact, $\text{prox}_\ell(x;\lambda) \to \text{prox}_\ell(x';\lambda')$ whenever $(x,\lambda) \to (x',\lambda')$ with $\lambda' > 0$.

(b) The value $\mathcal{M}_\ell(x;\lambda)$ is finite and depends continuously on (λ, x), with $\mathcal{M}_\ell(x;\lambda) \to f(x)$ for all x as $\lambda \to 0_+$.

(c) The Moreau envelope function is differentiable with respect to both arguments. Specifically, for all $x \in \mathbb{R}$, the following properties are true:

$$\mathcal{M}'_{\ell,1}(x;\lambda) = \frac{1}{\lambda}(x - \text{prox}_\ell(x;\lambda)), \tag{A2}$$

$$\mathcal{M}'_{\ell,2}(x;\lambda) = -\frac{1}{2\lambda^2}(x - \text{prox}_\ell(x;\lambda))^2. \tag{A3}$$

If in addition ℓ is differentiable and ℓ' denotes its derivative, then

$$\mathcal{M}'_{\ell,1}(x;\lambda) = \ell'(\text{prox}_\ell(x;\lambda)), \tag{A4}$$

$$\mathcal{M}'_{\ell,2}(x;\lambda) = -\frac{1}{2}(\ell'(\text{prox}_\ell(x;\lambda))^2. \tag{A5}$$

Appendix A.2. Alternative Representations of (8)

Replacing the above relations for derivative of $\mathcal{M}_\ell$ in (8), we can write the equations in terms of the proximal operator. If ℓ is differentiable, then Equations (8) can be equivalently written as follows:

$$\mathbb{E}\Big[Y\,S \cdot \ell'(\text{prox}_\ell(\alpha G + \mu S Y;\lambda))\Big] = 0, \tag{A6a}$$

$$\lambda^2\,\delta\,\mathbb{E}\Big[\big(\ell'(\text{prox}_\ell(\alpha G + \mu S Y;\lambda))\big)^2\Big] = \alpha^2, \tag{A6b}$$

$$\lambda\,\delta\,\mathbb{E}\Big[G \cdot \ell'(\text{prox}_\ell(\alpha G + \mu S Y;\lambda))\Big] = \alpha. \tag{A6c}$$

Finally, if ℓ is two times differentiable, then applying integration by parts in Equation (14c) results in the following reformulation of (8c):

$$1 = \lambda\,\delta\,\mathbb{E}\left[\frac{\ell''(\text{prox}_\ell(\alpha G + \mu S Y;\lambda))}{1 + \lambda\,\ell''(\text{prox}_\ell(\alpha G + \mu S Y;\lambda))}\right]. \tag{A7}$$

Appendix A.3. Examples of Proximal Operators

LAD.

For $\ell(t) = |t-1|$, the proximal operator admits a simple expression, as follows:

$$\text{prox}_{\ell}(x;\lambda) = 1 + \mathcal{H}(x-1;\lambda), \tag{A8}$$

where

$$\mathcal{H}(x;\lambda) = \begin{cases} x-\lambda, & \text{if } x > \lambda, \\ x+\lambda, & \text{if } x < -\lambda, \\ 0, & \text{otherwise.} \end{cases}$$

is the standard soft-thresholding function.

Hinge Loss.

When $\ell(t) = \max\{0, 1-t\}$, the proximal operator can be expressed in terms of the soft-thresholding function as follows:

$$\text{prox}_{\ell}(x;\lambda) = 1 + \mathcal{H}\left(x + \frac{\lambda}{2} - 1; \frac{\lambda}{2}\right).$$

Appendix A.4. Fenchel–Legendre Conjugate Representation

For a function $h : \mathbb{R} \to \mathbb{R}$, its Fenchel–Legendre conjugate, $h^\star : \mathbb{R} \to \mathbb{R}$ is defined as:

$$h^\star(x) = \max_y [xy - h(y)].$$

The following proposition relates Moreau Envelope of a function to its Fenchel–Legendre conjugate.

Proposition A2. *For $\lambda > 0$ and a function h, we have:*

$$\mathcal{M}_h(x;\lambda) = \frac{q(x)}{\lambda} - \frac{1}{\lambda}(q + \lambda h)^\star(x), \tag{A9}$$

where $q(x) = x^2/2$.

Proof.

$$\begin{aligned} \mathcal{M}_h(x;\lambda) &= \tfrac{1}{2\lambda} \min_y \left[(x-y)^2 + 2\lambda h(y)\right] \\ &= \tfrac{x^2}{2\lambda} + \tfrac{1}{2\lambda} \min_y \left[y^2 - 2xy + 2\lambda h(y)\right] \\ &= \tfrac{x^2}{2\lambda} - \tfrac{1}{\lambda} \max_y \left[xy - \left(y^2/2 + \lambda h(y)\right)\right] \\ &= \tfrac{q(x)}{\lambda} - \tfrac{1}{\lambda}(q + \lambda h)^\star(x). \end{aligned}$$

□

Appendix A.5. Convexity of the Moreau Envelope

Lemma A1. *The function $H : \mathbb{R}^3 \to \mathbb{R}$ defined as follows*

$$H(x, v, \lambda) = \frac{1}{2\lambda}(x-v)^2, \tag{A10}$$

is jointly convex in its arguments.

Proof. Note that the function $h(x, v) = (x-v)^2$ is jointly convex in (x, v). Thus, its perspective function

$$\lambda\, h(x/\lambda, v/\lambda) = (x-v)^2/\lambda = 2H(x, v, \lambda)$$

is jointly convex in (x, v, λ) [58] (Sec. 2.3.3), which completes the proof. □

Proposition A3. *(a) Ref. [52] (Prop. 2.22) Let $f(x, y)$ be jointly convex in its arguments. Then, the function $g(x) = \min_y f(x, y)$ is convex.*

(b) Ref. [58] (Sec. 3.2.3) Suppose $f_i : \mathbb{R} \to \mathbb{R}$ is a set of concave functions, with $i \in A$ an index set. Then, the function $f : \mathbb{R} \to \mathbb{R}$ defined as $f(x) := \inf_{i \in A} f_i(x)$ is concave.

Lemma A2. *Let $\ell : \mathbb{R} \to \mathbb{R}$ be a lsc, proper, convex function. Then, $\mathcal{M}_\ell(x; \lambda)$ is* jointly *convex in (x, λ).*

Proof. Recall that

$$\mathcal{M}_\ell(x; \lambda) = \min_v \; G(\mathbf{a}) := \frac{1}{2\lambda}(x - v)^2 + \ell(v), \tag{A11}$$

where, for compactness, we let $\mathbf{a} \in \mathbb{R}^3$ denote the triplet (x, v, λ). Now, let $\mathbf{a}_i = (x_i, v_i, \lambda_i), i = 1, 2$, $\theta \in (0, 1)$ and $\overline{\theta} := 1 - \theta$. With this notation, we may write

$$\begin{aligned} G(\theta \mathbf{a}_1 + \overline{\theta} \mathbf{a}_2) &= H\big(\theta x_1 + \overline{\theta} x_2, \theta \lambda_1 + \overline{\theta} \lambda_2, \theta v_1 + \overline{\theta} v_2\big) + \ell(\theta v_1 + \overline{\theta} v_2) \\ &\leq \theta H(x_1, v_1, \lambda_1) + \overline{\theta} H(x_2, v_2, \lambda_2) + \theta \ell(v_1) + \overline{\theta} \ell(v_2) \\ &= \theta G(\mathbf{a}_1) + \overline{\theta} G(\mathbf{a}_2). \end{aligned}$$

For the first equality above, we recall the definition of $H : \mathbb{R}^3 \to \mathbb{R}$ in (A10) and the inequality right after follows from Lemma A1 and convexity of ℓ. Thus, the function G is jointly convex in its arguments. Using this fact, as well as (A11), and applying Proposition A3(a) completes the proof. □

Appendix A.6. The Expected Moreau-Envelope (EME) Function and its Properties

The performance of the ERM estimator (3) is governed by the system of equations (8) in which the Moreau envelope function $\mathcal{M}_\ell(x; \lambda)$ of the loss function ℓ plays a central role. More precisely, as already hinted by (8) and becomes clear in Appendix B, what governs the behavior is the function

$$(\alpha > 0, \mu, \tau > 0, \gamma > 0) \mapsto \mathbb{E}\big[\mathcal{M}_\ell(\alpha G + \mu S Y; \tau/\gamma)\big], \tag{A12}$$

which we call the expected Moreau envelope (EME). Recall here that $Y = f(S)$. Hence, the EME is the key summary parameter that captures the role of both the loss function $\ell : \mathbb{R} \to \mathbb{R}$ and of the link function $f : \mathbb{R} \to \{\pm 1\}$ on the statistical performance of (3).

In this section, we study several favorable properties of the EME. In (A12), the expectation is over $G, S \overset{\text{iid}}{\sim} \mathcal{N}(0, 1)$. We first study the EME under more general distribution assumptions in Appendices A.6.1–A.6.3 and we then specialize our results to Gaussian random variables G and S in Appendix A.6.4.

Appendix A.6.1. Derivatives

Proposition A4. *Let $\ell : \mathbb{R} \to \mathbb{R}$ be a lsc, proper and convex function. Further, let X, Z be independent random variables with bounded second moments $\mathbb{E}[X^2] < \infty$, $\mathbb{E}[Z^2] < \infty$. Then, the expected Moreau envelope function $\mathbb{E}[\mathcal{M}_\ell(cX + Z; \lambda)]$ is differentiable with respect to both c and λ and the derivatives are given as follows:*

$$\frac{\partial}{\partial c} \mathbb{E}\Big[\mathcal{M}_\ell(cX + Z; \lambda)\Big] = \mathbb{E}\Big[X \mathcal{M}'_{\ell,1}(cX + Z; \lambda)\Big], \tag{A13}$$

$$\frac{\partial}{\partial \lambda} \mathbb{E}\Big[\mathcal{M}_\ell(cX + Z; \lambda)\Big] = \mathbb{E}\Big[\mathcal{M}'_{\ell,2}(cX + Z; \lambda)\Big]. \tag{A14}$$

Proof. The proof is an application of the Dominated Convergence Theorem (DCT). First, by Proposition A1(b), for every $c \in \mathbb{R}$ and any $\lambda > 0$, the function $\mathbb{E}[\mathcal{M}_\ell(cX+Z;\lambda)]$ takes a finite value. Second, by Proposition A1(c), $\mathcal{M}_\ell(cx+z;\lambda)$ is continuously differentiable with respect to both c and λ:

$$\frac{\partial}{\partial c}\mathcal{M}_\ell(cX+Z;\lambda) = X\mathcal{M}'_{\ell,1}(cX+Z;\lambda) = X\frac{1}{\lambda}\big(cX+Z-\mathrm{prox}_\ell(cX+Z;\lambda)\big),$$
$$\frac{\partial}{\partial \lambda}\mathcal{M}_\ell(cX+Z;\lambda) = \mathcal{M}'_{\ell,2}(cX+Z;\lambda) = -\frac{1}{2\lambda^2}\big(cX+Z-\mathrm{prox}_\ell(cX+Z;\lambda)\big)^2.$$

From this, note that the Cauchy–Schwarz inequality gives

$$\mathbb{E}\Big[\frac{\partial}{\partial c}\mathcal{M}_\ell(cX+Z;\lambda)\Big] \le \Big(\mathbb{E}[X^2])^{1/2}\Big)\Big(\mathbb{E}\Big[\frac{1}{\lambda^2}\underbrace{\big(cX+Z-\mathrm{prox}_\ell(cX+Z;\lambda)\big)^2}_{:=A}\Big]\Big)^{1/2},$$

Therefore, the remaining condition to check so that DCT can be applied is that the term A/λ^2 above is integrable. To begin with, we can easily bound A as: $A \le 2(cX+Z)^2 + 2(\mathrm{prox}_\ell(cX+Z;\lambda))^2$. Next, by non-expansiveness (Lipschitz property) of the proximal operator [52] (Prop. 12.19), we have that $|\mathrm{prox}_\ell(cX+Z;\lambda)| \le |cX+Z| + |\mathrm{prox}_\ell(0;\lambda)|$. Putting together, we find that

$$A \le 6(cX+Z)^2 + 2|\mathrm{prox}_\ell(0;\lambda)|^2 \le 12c^2X^2 + 12Z^2 + 2|\mathrm{prox}_\ell(0;\lambda)|^2.$$

We consider two cases. First, for fixed $\lambda > 0$ and any compact interval $\mathcal{I}$, we have that

$$\mathbb{E}\sup_{c\in\mathcal{I}}[A] \le 12(\sup_{c\in\mathcal{I}} c^2)\mathbb{E}[X^2] + 12\mathbb{E}[Z]^2 + 2|\mathrm{prox}_\ell(0;\lambda)|^2 < \infty.$$

Similarly, for fixed c and any compact interval $\mathcal{J}$ on the positive real line, we have that

$$\mathbb{E}\sup_{\lambda\in\mathcal{J}}[A/\lambda^2] \le 12\sup_{\lambda\in\mathcal{J}}\frac{c^2\mathbb{E}[X^2]+\mathbb{E}[Z]^2}{\lambda^2} + 2\sup_{\lambda\in\mathcal{J}}\frac{|\mathrm{prox}_\ell(0;\lambda)|^2}{\lambda^2} < \infty,$$

where we also used boundedness of the proximal operator (cf. Proposition A1(a)). This completes the proof. □

Appendix A.6.2. Strict Convexity

We study convexity properties of the *expected Moreau envelope function* $\Psi : \mathbb{R}^3 \to \mathbb{R}$:

$$\Psi(\mathbf{v}) := \Psi(\alpha,\mu,\lambda) := \mathbb{E}\Big[\mathcal{M}_\ell(\alpha X+\mu Z;\lambda)\Big], \tag{A15}$$

for a lsc, proper, convex function ℓ and independent random variables X and Z with positive densities. Here, and onwards, we let $\mathbf{v} \in \mathbb{R}^3$ denote a triplet (α,μ,λ) and the expectation is over the randomness of X and Z. From Lemma A2, it is easy to see that $\Psi(\mathbf{v})$ is convex. In this section, we prove a stronger claim:

> " If ℓ is strictly convex and does not attain its minimum at 0, then $\Psi(\mathbf{v})$ is also strictly convex. "

This is summarized in Proposition A5 below.

Proposition A5 (Strict Convexity). *Let $\ell : \mathbb{R} \to \mathbb{R}$ be a function with the following properties: (i) it is proper strictly convex; and (ii) it is continuously differentiable and its derivative ℓ' is such that $\ell'(0) \neq 0$. Further, let X, Z be independent random variables with strictly positive densities. Then, the function $\Psi : \mathbb{R}^3 \to \mathbb{R}$ in (A15) is jointly strictly convex in its arguments.*

Proof. Let $\mathbf{v}_i = (\alpha_i, \mu_i, \lambda_i), i = 1,2, \theta \in (0,1)$ and $\overline{\theta} = 1 - \theta$. Further, assume that $\mathbf{v}_1 \neq \mathbf{v}_2$ and define the proximal operators

$$p_i(X,Z) := \text{prox}_{\ell}(\alpha_i X + \mu_i Z; \lambda_i) = \arg\min_{v} \frac{1}{2\lambda_i}(\alpha_i X + \mu_i Z - v)^2 + \ell(v),$$

for $i = 1,2$. Finally, denote $\lambda_\theta := \theta\lambda_1 + \overline{\theta}\lambda_2, \alpha_\theta := \theta\alpha_1 + \overline{\theta}\alpha_2$ and $\mu_\theta := \theta\mu_1 + \overline{\theta}\mu_2$. With this notation,

$$\begin{aligned}
&\Psi(\theta\mathbf{v}_1 + \overline{\theta}\mathbf{v}_2) \\
&\quad \leq \mathbb{E}\Big[\frac{1}{2\lambda_\theta}\Big(\alpha_\theta X + \mu_\theta Z - (\theta p_1(X,Z) + \overline{\theta} p_2(X,Z))\Big)^2 + \ell\Big(\theta p_1(X,Z) + \theta p_2(X,Z)\Big)\Big] \\
&\quad = \mathbb{E}\Big[H\Big(\alpha_\theta X + \mu_\theta Z, \theta p_1(X,Z) + \theta p_2(X,Z), \lambda_\theta\Big) + \ell\Big(\theta p_1(X,Z) + \overline{\theta} p_2(X,Z)\Big)\Big] \\
&\quad \leq \mathbb{E}\Big[\theta H\Big(\alpha_1 X + \mu_1 Z, p_1(X,Z), \lambda_1\Big) + \overline{\theta} H\Big(\alpha_2 X + \mu_2 Z, p_2(X,Z), \lambda_2\Big) \\
&\qquad\qquad + \ell\Big(\theta p_1(X,Z) + \overline{\theta} p_2(X,Z)\Big)\Big].
\end{aligned} \tag{A16}$$

The first inequality above follows by the definition of the Moreau envelope in (A1). The equality in the second line uses the definition of the function $H : \mathbb{R}^3 \to \mathbb{R}$ in (A10). Finally, the last inequality follows from convexity of H as proved in Lemma A1.
Continuing from (A16), we may use convexity of ℓ to find that

$$\begin{aligned}
&\Psi(\theta\mathbf{v}_1 + \overline{\theta}\mathbf{v}_2) \\
&\leq \mathbb{E}\Big[\theta H(\alpha_1 X + \mu_1 Z, \lambda_1, p_1(X,Z)) + \overline{\theta} H(\alpha_2 X + \mu_2 Z, \lambda_2, p_2(X,Z)) \\
&\qquad\qquad + \theta\ell(p_1(X,Z)) + \overline{\theta}\ell(p_2(X,Z))\Big] \\
&= \theta\Psi(\mathbf{v}_1) + \overline{\theta}\Psi(\mathbf{v}_2).
\end{aligned} \tag{A17}$$

This already proves convexity of (A15). In what follows, we argue that the inequality in (A17) is in fact strict under the assumption of the lemma.

Specifically, in Lemma A3, we prove that, under the assumptions of the proposition, for $\mathbf{v}_1 \neq \mathbf{v}_2$, it holds that

$$\mathbb{E}\Big[\ell\big(\theta p_1(X,Z) + \overline{\theta} p_2(X,Z)\big)\Big] < \theta\,\mathbb{E}\Big[\ell(p_1(X,Z))\Big] + \overline{\theta}\,\mathbb{E}\Big[\ell(p_2(X,Z))\Big].$$

Using this in (A16) completes the proof of the proposition. The idea behind the proof of Lemma A3 is as follows. First, we use the fact that $\mathbf{v}_1 \neq \mathbf{v}_2$ and $\ell'(0) \neq 0$ to argue that there exists a non-zero measure set of $(x,z) \in \mathbb{R}^2$ such that $p_1(x,z) \neq p_2(x,z)$. Then, the desired claim follows by *strict* convexity of ℓ. □

Lemma A3. *Let $\ell : \mathbb{R} \to \mathbb{R}$ be a proper strictly convex function that is continuously differentiable with $\ell'(0) \neq 0$. Further, assume independent continuous random variables X, Z with strictly positive densities. Fix arbitrary triplets $\mathbf{v}_i = (\alpha_i, \mu_i, \lambda_i), i = 1,2$ such that $\mathbf{v}_1 \neq \mathbf{v}_2$. Further, denote*

$$p_i(X,Z) := \text{prox}_{\ell}(\alpha_i X + \mu_i Z; \lambda_i), i = 1,2. \tag{A18}$$

Then, there exists a ball $\mathcal{S} \subset \mathbb{R}^2$ of nonzero measure, i.e., $\mathbb{P}((X,Z) \in \mathcal{S}) > 0$, such that $p_1(x,z) \neq p_2(x,z)$, for all $(x,z) \in \mathcal{S}$. Consequently, for any $\theta \in (0,1)$ and $\bar{\theta} = 1-\theta$, the following strict inequality holds,

$$\mathbb{E}\Big[\ell\big(\theta p_1(X,Z) + \bar{\theta} p_2(X,Z)\big)\Big] < \theta\,\mathbb{E}\Big[\ell\big(p_1(X,Z)\big)\Big] + \bar{\theta}\,\mathbb{E}\Big[\ell\big(p_2(X,Z)\big)\Big]. \tag{A19}$$

Proof. Note that (A19) holds trivially with "$<$" replaced by "$\leq$" due to the convexity of ℓ. To prove that the inequality is strict, it suffices, by strict convexity of ℓ, that there exists subset $\mathcal{S} \subset \mathbb{R}^2$ that satisfies the following two properties:

1. $p_1(x,z) \neq p_2(x,z)$, for all $(x,z) \in \mathcal{S}$.
2. $\mathbb{P}((X,Z) \in \mathcal{S}) > 0$.

Consider the following function $f : \mathbb{R}^2 \to \mathbb{R}$:

$$f(x,z) := p_1(x,z) - p_2(x,z). \tag{A20}$$

By Lemma A4, there exists (x_0, z_0) such that

$$f(x_0, z_0) \neq 0. \tag{A21}$$

Moreover, by continuity of the proximal operator (cf. Proposition A1(a)), it follows that f is continuous. From this and (A21), we conclude that for sufficiently small $\zeta > 0$ there exists a ζ-ball $\mathcal{S}$ centered at (x_0, z_0), such that property 1 holds. Property 2 is also guaranteed to hold for $\mathcal{S}$, since both X, Z have strictly positive densities and are independent. □

Lemma A4. *Let $\ell : \mathbb{R} \to \mathbb{R}$ be a proper, convex function. Further, assume that $\ell : \mathbb{R} \to \mathbb{R}$ is continuously differentiable and $\ell'(0) \neq 0$. Let $\alpha_1, \alpha_2 > 0$, $\lambda_1, \lambda_2 > 0$. Then, the following statement is true*

$$(\alpha_1, \mu_1, \lambda_1) \neq (\alpha_2, \mu_2, \lambda_2) \to \exists (x,z) \in \mathbb{R}^2 : \operatorname{prox}_\ell(\alpha_1 x + \mu_1 z; \lambda_1) \neq \operatorname{prox}_\ell(\alpha_2 x + \mu_2 z; \lambda_2). \tag{A22}$$

Proof. We prove the claim by contradiction, but first, let us set up some useful notation. Let $\mathbf{v} \in \mathbb{R}^3$ denote triplets (α, μ, λ) and further define

$$p_{\alpha,\mu,\lambda}(x,z) := \operatorname{prox}_\ell(\alpha x + \mu z; \lambda),$$

and

$$L_{\alpha,\mu,\lambda}(x,z) := \ell'(\operatorname{prox}_\ell(\alpha x + \mu z; \lambda)).$$

By Proposition A1, the following is true:

$$L_{\alpha,\mu,\lambda}(x,z) = \frac{1}{\lambda}\big(\alpha x + \mu z - p_{\alpha,\mu,\lambda}(x,z)\big). \tag{A23}$$

For the sake of contradiction, assume that the claim of the lemma is false. Then,

$$p_{\alpha_1,\mu_1,\lambda_1}(x,z) = p_{\alpha_2,\mu_2,\lambda_2}(x,z), \quad \forall (x,z) \in \mathbb{R}^2. \tag{A24}$$

From this, it also holds that

$$L_{\alpha_1,\mu_1,\lambda_1}(x,z) = L_{\alpha_2,\mu_2,\lambda_2}(x,z), \quad \forall (x,z) \in \mathbb{R}^2. \tag{A25}$$

Recalling (A23) and applying (A24), we derive the following from (A25):

$$(\lambda_2 - \lambda_1) p_{\alpha_1,\mu_1,\lambda_1}(x,z) = (\lambda_2\alpha_1 - \lambda_1\alpha_2)x + (\lambda_2\mu_1 - \lambda_1\mu_2)z, \quad \forall (x,z) \in \mathbb{R}^2. \tag{A26}$$

We consider the following two cases separately.

Case 1: $\lambda_1 = \lambda_2$: Since $\mathbf{v}_1 \neq \mathbf{v}_2$, it holds that

$$\exists (x,z) \in \mathbb{R}^2 : \quad \alpha_1 x + \mu_1 z \neq \alpha_2 x + \mu_2 z. \tag{A27}$$

However, from (A26) we have that $(\alpha_1 - \alpha_2)x + (\mu_1 - \mu_2)z = 0$ for all $(x,z) \in \mathbb{R}^2$. This contradicts (A27) and completes the proof for this case.

Case 2: $\lambda_1 \neq \lambda_2$: Continuing from (A26), we can compute that for all $(x,z) \in \mathbb{R}^2$

$$\begin{aligned} \ell'(p_{\alpha_1,\mu_1,\lambda_1}(x,z)) &= \frac{1}{\lambda_1}(\alpha_1 x + \mu_1 z - p_{\alpha_1,\mu_1,\lambda_1}(x,z)) \\ &= \frac{\alpha_2 - \alpha_1}{\lambda_2 - \lambda_1} x + \frac{\mu_2 - \mu_1}{\lambda_2 - \lambda_1} z. \end{aligned} \tag{A28}$$

By replacing $p_{\alpha_1,\mu_1,\lambda_1}(x,z)$ from (A26), we derive that:

$$\ell'(\varepsilon_1 x + \varepsilon_2 z) = \varepsilon_3 x + \varepsilon_4 z, \quad \forall (x,z) \in \mathbb{R}^2, \tag{A29}$$

where

$$\begin{aligned} \varepsilon_1 &= \frac{\lambda_2\alpha_1 - \lambda_1\alpha_2}{\lambda_2 - \lambda_1}, & \varepsilon_2 &= \frac{\lambda_2\mu_1 - \lambda_1\mu_2}{\lambda_2 - \lambda_1}, \\ \varepsilon_3 &= \frac{\alpha_2 - \alpha_1}{\lambda_2 - \lambda_1}, & \varepsilon_4 &= \frac{\mu_2 - \mu_1}{\lambda_2 - \lambda_1}. \end{aligned}$$

By replacing $x = z = 0$ in (A29), we find that $\ell'(0) = 0$. This contradicts the assumption of the lemma and completes the proof. □

Appendix A.6.3. Strict Concavity

In this section, we study the following variant $\Gamma : \mathbb{R}_+ \to \mathbb{R}$ of the expected Moreau envelope:

$$\Gamma(\gamma) := \mathbb{E}[\mathcal{M}_\ell(X; 1/\gamma)], \tag{A30}$$

for a lower semi-continuous, proper, convex function ℓ and continuous random variable X. The expectation above is over the randomness of X. In Appendix B.4, we show that the function Γ is concave in γ. Here, we prove the following statement regarding *strict*-concavity of Γ:

> "If ℓ is convex, continuously differentiable and $\ell'(0) \neq 0$, then Γ is *strictly* concave."

This is summarized in Proposition A6 below.

Proposition A6 (Strict concavity)**.** *Let $\ell : \mathbb{R} \to \mathbb{R}$ be a convex, continuously differentiable function for which $\ell'(0) \neq 0$. Further, let X be a continuous random variable in $\mathbb{R}$ with strictly positive density in the real line. Then, the function Γ in* (A30) *is* strictly *concave in $\mathbb{R}_+$.*

Proof. Before everything, we introduce the following convenient notation:

$$\widetilde{\Gamma}_x(\gamma) := \mathcal{M}_\ell(x; 1/\gamma) \quad \text{and} \quad p_\gamma^x := \text{prox}_\ell(x; 1/\gamma).$$

Note from Proposition A1 that $\widetilde{\Gamma}_x$ is differentiable with derivative

$$\widetilde{\Gamma}_x'(\gamma) = \frac{1}{2}\Big(x - \text{prox}_\ell(x; 1/\gamma)\Big)^2. \tag{A31}$$

We proceed in two steps as follows. First, for fixed $x \in \mathbb{R}$ and $\gamma_2 > \gamma_1$, we prove in Lemma A5 that

$$(x - p^x_{\gamma_2})^2 - (x - p^x_{\gamma_1})^2 \leq -\frac{\gamma_1}{\gamma_2 - \gamma_1}(p^x_{\gamma_1} - p^x_{\gamma_2})^2, \tag{A32}$$

This shows that for all $x \in \mathbb{R}$

$$\widetilde{\Gamma}'_x(\gamma_2) - \widetilde{\Gamma}'_x(\gamma_1) \leq 0. \tag{A33}$$

Second, we use Lemma A3 to argue that the inequality is in fact strict for all $x \in \mathcal{S}$ where $\mathcal{S} \subset \mathbb{R}$ and $\mathbb{P}(X \in \mathcal{S}) > 0$. To be concrete, apply Lemma A3 for $\mathbf{v}_i = (1, 0, 1/\gamma_i)$, $i = 1, 2$. Notice that all the assumptions of the lemma are satisfied, hence there exists interval $\mathcal{S} \subset \mathbb{R}$ for which $\mathbb{P}(X \in \mathcal{S}) > 0$ and

$$p^x_{\gamma_1} \neq p^x_{\gamma_2} \Rightarrow (p^x_{\gamma_1} - p^x_{\gamma_2})^2 > 0, \quad \forall x \in \mathcal{S}.$$

Hence, from (A32), it follows that

$$(x - p^x_{\gamma_2})^2 - (x - p^x_{\gamma_1})^2 < 0, \quad \forall x \in \mathcal{S}.$$

From this, and (A31) we conclude that

$$\widetilde{\Gamma}'_x(\gamma_2) - \widetilde{\Gamma}'_x(\gamma_1) < 0, \quad \forall x \in \mathcal{S}. \tag{A34}$$

Thus, from (A33) and (A34), as well as the facts that $\Gamma(\gamma) = \mathbb{E}\left[\widetilde{\Gamma}_X(\gamma)\right]$ and $\mathbb{P}(X \in \mathcal{S}) > 0$, we conclude that Γ is strictly concave in $\mathbb{R}_+$. □

Lemma A5. *Let $\ell : \mathbb{R} \to \mathbb{R}$ be a convex, continuously differentiable function. Fix $x \in \mathbb{R}$ and denote $p_\gamma := \mathrm{prox}_\ell(x; 1/\gamma)$. Then, for any $\gamma, \widetilde{\gamma} > 0$, it holds that*

$$(\widetilde{\gamma} - \gamma)(p_{\widetilde{\gamma}} - p_\gamma)(p_\gamma - x) + \widetilde{\gamma}(p_{\widetilde{\gamma}} - p_\gamma)^2 \leq 0. \tag{A35}$$

Moreover, for $\gamma_2 > \gamma_1$, the following statement is true:

$$(x - p_{\gamma_2})^2 - (x - p_{\gamma_1})^2 \leq -\frac{\gamma_1}{\gamma_2 - \gamma_1}(p_{\gamma_1} - p_{\gamma_2})^2. \tag{A36}$$

Proof. First, we prove (A35). Then, we use it to prove (A36).

Proof of (A35): Consider function $g : \mathbb{R} \to \mathbb{R}$ defined as follows $g(p) = \frac{\widetilde{\gamma}}{2}(x - p)^2 + \ell(p)$. By assumption, g is differentiable with derivative $g'(p) = \widetilde{\gamma}(p - x) + \ell'(p)$. Moreover, g is γ_2-strongly convex. Finally, by optimality of the proximal operator (cf. Proposition A1), it holds that $\gamma(x - p_\gamma) = \ell'(p_\gamma)$ and $\widetilde{\gamma}(x - p_{\widetilde{\gamma}}) = \ell'(p_{\widetilde{\gamma}})$. Using these, it can be computed that $g'(p_{\widetilde{\gamma}}) = 0$ and $g'(p_\gamma) = (\widetilde{\gamma} - \gamma)(p_\gamma - x)$.

In the following inequalities, we combine all the aforementioned properties of the function g to find that

$$g(p_\gamma) \geq g(p_{\widetilde{\gamma}}) + \frac{\widetilde{\gamma}}{2}(p_\gamma - p_{\widetilde{\gamma}})^2 \geq g(p_\gamma) + (\widetilde{\gamma} - \gamma)(p_\gamma - x)(p_{\widetilde{\gamma}} - p_\gamma) + \widetilde{\gamma}(p_\gamma - p_{\widetilde{\gamma}})^2.$$

This leads to the desired statement and completes the proof of (A35).

Proof of (A36): We fix $\gamma_2 > \gamma_1$ and apply (A35) two times as follows. First, applying (A35) for $(\widetilde{\gamma}, \gamma) = (\gamma_2, \gamma_1)$ and using the fact that $\gamma_2 > \gamma_1$, we find that

$$(p_{\gamma_2} - p_{\gamma_1})(p_{\gamma_1} - x) \leq -\frac{\gamma_2}{\gamma_2 - \gamma_1}(p_{\gamma_2} - p_{\gamma_1})^2. \tag{A37}$$

Second, applying (A35) for $(\widetilde{\gamma}, \gamma) = (\gamma_1, \gamma_2)$ and using again the fact that $\gamma_2 > \gamma_1$, we find that

$$\begin{aligned} & (\gamma_1 - \gamma_2)(p_{\gamma_1} - p_{\gamma_2})(p_{\gamma_2} - x) + \gamma_1 (p_{\gamma_1} - p_{\gamma_2})^2 \leq 0 \\ \Rightarrow \quad & (p_{\gamma_2} - p_{\gamma_1})(p_{\gamma_2} - x) \leq -\frac{\gamma_1}{\gamma_2 - \gamma_1}(p_{\gamma_1} - p_{\gamma_2})^2. \end{aligned} \tag{A38}$$

Adding (A37) and (A38), we show the desired property as follows:

$$(p_{\gamma_2} - p_{\gamma_1})(p_{\gamma_2} - x) + (p_{\gamma_2} - p_{\gamma_1})(p_{\gamma_1} - x) \leq -\frac{\gamma_2 + \gamma_1}{\gamma_2 - \gamma_1}(p_{\gamma_1} - p_{\gamma_2})^2.$$

□

Appendix A.6.4. Summary of Properties of (A12)

Proposition A7. *Let $\ell : \mathbb{R} \to \mathbb{R}$ be a lsc, proper, convex function. Let $G, S \overset{iid}{\sim} \mathcal{N}(0,1)$ and function $f : \mathbb{R} \to \{\pm 1\}$ such that the random variable $YS = f(S)S$ has a continuous strictly positive density on the real line. Then, the following properties are true for the expected Moreau envelope function*

$$\Omega : (\alpha > 0, \mu, \tau > 0, \gamma > 0) \mapsto \mathbb{E}\big[\mathcal{M}_\ell(\alpha G + \mu SY; \tau/\gamma)\big] : \tag{A39}$$

(a) The function Ω is differentiable and its derivatives are given as follows:

$$\begin{aligned} \frac{\partial}{\partial \alpha}\Omega(\alpha, \mu, \tau, \gamma) &= \mathbb{E}\Big[G\, \mathcal{M}'_{\ell,1}(\alpha G + \mu SY; \tau/\gamma)\Big], \\ \frac{\partial}{\partial \mu}\Omega(\alpha, \mu, \tau, \gamma) &= \mathbb{E}\Big[SY\, \mathcal{M}'_{\ell,1}(\alpha G + \mu SY; \tau/\gamma)\Big], \\ \frac{\partial}{\partial \tau}\Omega(\alpha, \mu, \tau, \gamma) &= \frac{1}{\gamma}\mathbb{E}\Big[\mathcal{M}'_{\ell,2}(\alpha G + \mu SY; \tau/\gamma)\Big], \\ \frac{\partial}{\partial \gamma}\Omega(\alpha, \mu, \tau, \gamma) &= -\frac{\tau}{\gamma^2}\mathbb{E}\Big[\mathcal{M}'_{\ell,2}(\alpha G + \mu SY; \tau/\gamma)\Big]. \end{aligned}$$

(b) The function Ω is jointly convex (α, μ, τ) and concave on γ.
(c) The function Ω is increasing in α.
For the statements below, further assume that ℓ is strictly convex and continuously differentiable with $\ell'(0) \neq 0$.
(d) The function Ω is strictly convex in (α, μ, τ) and strictly concave in λ.
(e) The function Ω is strictly increasing in α.

Proof. Statements (a), (b) and (d) follow directly by Propositions A4–A6. It remains to prove Statements (c) and (e). Let $\alpha_2 > \alpha_1$. Then, there exist independent copies G', G'' of G and $\tilde{\alpha} > 0$ such that $\alpha_2 G = \alpha_1 G' + \tilde{\alpha} G''$. Hence, we have the following chain of inequalities:

$$\begin{aligned} \Omega(\alpha_2, \mu, \tau, \gamma) &= \mathbb{E}\big[\mathcal{M}_\ell(\alpha_1 G' + \tilde{\alpha} G'' + \mu SY; \tau/\gamma)\big] \geq \mathbb{E}\big[\mathcal{M}_\ell(\alpha_1 G' + \tilde{\alpha}\mathbb{E}[G''] + \mu SY; \tau/\gamma)\big] \\ &= \mathbb{E}\big[\mathcal{M}_\ell(\alpha_1 G' + \mu SY; \tau/\gamma)\big] = \Omega(\alpha_1, \mu, \tau, \gamma), \end{aligned}$$

where the inequality follows from Jensen and convexity of Ω with respect to α (see Statement (b) of the Proposition). This proves Statement (c). For Statement (e), note that the inequality is strict provided that Ω is strictly convex (see Statement (d) of the Proposition). □

Appendix B. Proof of Theorem 1

In this section, we provide a proof sketch of Theorem 1. The main technical tool that facilitates our analysis is the convex Gaussian min-max theorem (CGMT), which is an

extension of Gordon's Gaussian min-max inequality (GMT). We introduce the necessary background on the CGMT in Appendix B.1.

The CGMT has been mostly applied to linear measurements [9,10,13,15,19]. The simple, yet central idea, which allows for this extension, is a certain projection trick inspired by Plan and Vershynin [40]. Here, we apply a similar trick, but, in our setting, we recognize that it suffices to simply rotate $\mathbf{x}_0$ to align with the first basis vector. The simple rotation decouples the measurements y_i from the last $n-1$ coordinates of the measurement vectors $\mathbf{a}_i$ (see Appendix B.2). While this is sufficient for LS in [43], to study more general loss functions, we further need to combine this with a duality argument similar to that in [13]. Second, while the steps that bring the ERM minimization to the form of a PO (see (A48)) bear the aforementioned similarities to those in [13,43], the resulting AO is different from the one studied in previous works. Hence, the mathematical derivations in Appendices B.3 and B.4 are different. This also leads to a different system of equations characterizing the statistical behavior of ERM. Finally, in Appendix B.5, we prove uniqueness of the solution of this system of equations using the properties of the expected Moreau envelope function studied in Appendix A.6.

Appendix B.1. Technical Tool: CGMT

Appendix B.1.1. Gordon's Min-Max Theorem (GMT)

The Gordon's Gaussian comparison inequality [59] compares the min-max value of two doubly indexed Gaussian processes based on how their autocorrelation functions compare. The inequality is quite general (see [59]), but for our purposes we only need its application to the following two Gaussian processes:

$$X_{\mathbf{w},\mathbf{u}} := \mathbf{u}^T\mathbf{G}\mathbf{w} + \psi(\mathbf{w},\mathbf{u}), \tag{A40a}$$

$$Y_{\mathbf{w},\mathbf{u}} := \|\mathbf{w}\|_2\mathbf{g}^T\mathbf{u} + \|\mathbf{u}\|_2\mathbf{h}^T\mathbf{w} + \psi(\mathbf{w},\mathbf{u}), \tag{A40b}$$

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

$$\widetilde{\Phi}(\mathbf{G}) = \min_{\mathbf{w}\in\mathcal{S}_\mathbf{w}} \max_{\mathbf{u}\in\mathcal{S}_\mathbf{u}} X_{\mathbf{w},\mathbf{u}}, \tag{A41a}$$

$$\phi(\mathbf{g},\mathbf{h}) = \min_{\mathbf{w}\in\mathcal{S}_\mathbf{w}} \max_{\mathbf{u}\in\mathcal{S}_\mathbf{u}} Y_{\mathbf{w},\mathbf{u}}. \tag{A41b}$$

According to Gordon's comparison inequality (To be precise, the formulation in (A42), which is due to [13], is slightly different from the original statement in Gordon's paper (see [13] for details).), for any $c \in \mathbb{R}$, it holds:

$$\mathbb{P}\Big(\widetilde{\Phi}(\mathbf{G}) < c\Big) \le 2\mathbb{P}(\phi(\mathbf{g},\mathbf{h}) < c). \tag{A42}$$

In other words, a high-probability lower bound on the AO is a high-probability lower bound on the PO. The premise is that it is often much simpler to lower bound the AO rather than the PO. To be precise, (A42) is a slight reformulation of Gordon's original result proved in [13].

Appendix B.1.2. Convex Gaussian Min-Max Theorem (CGMT)

The proof of Theorem 1 builds on the CGMT [13]. For ease of reference, we summarize here the essential ideas of the framework following the presentation in [15] (please see [15] (Section 6) for the formal statement of the theorem and further details). The CGMT is an extension of the GMT and it asserts that the AO in (A41b) can be used to tightly infer properties of the original (PO) in (A41a), including the optimal cost and the optimal

solution. According to the CGMT [15] (Theorem 6.1), if the sets $\mathcal{S}_{\mathbf{w}}$ and $\mathcal{S}_{\mathbf{u}}$ are convex and ψ is continuous *convex-concave* on $\mathcal{S}_{\mathbf{w}} \times \mathcal{S}_{\mathbf{u}}$, then, for any $\nu \in \mathbb{R}$ and $t > 0$, it holds that

$$\mathbb{P}\Big(|\widetilde{\Phi}(\mathbf{G}) - \nu| > t \Big) \leq 2\mathbb{P}\Big(|\phi(\mathbf{g}, \mathbf{h}) - \nu| > t \Big). \tag{A43}$$

In words, concentration of the optimal cost of the AO problem around μ implies concentration of the optimal cost of the corresponding PO problem around the same value μ. Moreover, starting from (A43) and under strict convexity conditions, the CGMT shows that concentration of the optimal solution of the AO problem implies concentration of the optimal solution of the PO to the same value. For example, if minimizers of (A41b) satisfy $\|\mathbf{w}^*(\mathbf{g}, \mathbf{h})\|_2 \to \zeta^*$ for some $\zeta^* > 0$, then the same holds true for the minimizers of (A41a): $\|\mathbf{w}^*(\mathbf{G})\|_2 \to \zeta^*$ [15] ([Theorem 6.1(iii)). Thus, one can analyze the AO to infer corresponding properties of the PO, the premise being of course that the former is simpler to handle than the latter.

Appendix B.2. Applying the CGMT to ERM for Binary Classification

In this section, we show how to apply the CGMT to (3). For convenience, we drop the subscript ℓ from $\widehat{\mathbf{x}}_\ell$ and simply write

$$\widehat{\mathbf{x}} = \arg\min_{\mathbf{x}} \frac{1}{m} \sum_{i=1}^{m} \ell(y_i \mathbf{a}_i^T \mathbf{x}), \tag{A44}$$

where the measurements $y_i, i \in [m]$ follow (1). By rotational invariance of the Gaussian distribution of the measurement vectors $\mathbf{a}_i, i \in [m]$, we assume without loss of generality that $\mathbf{x}_0 = [1, 0, ..., 0]^T$. We can rewrite (A44) as a constrained optimization problem by introducing n variables u_i as follows:

$$\widehat{\mathbf{x}} = \arg\min_{\mathbf{x},\mathbf{u}} \frac{1}{m} \sum_{i=1}^{m} \ell(u_i) \text{subject to } u_i = y_i \mathbf{a}_i^T \mathbf{x}, i \in [n].$$

This problem is now equivalent to the following min-max formulation:

$$\min_{\mathbf{u},\mathbf{x}} \max_{\boldsymbol{\beta}} \frac{1}{m} \sum_{i=1}^{m} \ell(u_i) + \frac{1}{m} \sum_{i=1}^{m} \beta_i u_i - \frac{1}{m} \sum_{i=1}^{m} \beta_i y_i \mathbf{a}_i^T \mathbf{x}. \tag{A45}$$

Now, let us define

$$\mathbf{a}_i = [s_i; \widetilde{\mathbf{a}}_i], i \in [m] \quad \text{and} \quad \mathbf{x} = [x_1; \widetilde{\mathbf{x}}],$$

such that s_i and x_1 are the first entries of $\mathbf{a}_i$ and $\mathbf{x}$, respectively. Note that in this new notation (1) becomes:

$$y_i = f(s_i), \tag{A46}$$

and

$$\text{corr}(\widehat{\mathbf{x}}; \mathbf{x}_0) = \frac{\widehat{x}_1}{\sqrt{\widehat{x}_1^2 + \|\widetilde{\widehat{\mathbf{x}}}\|_2^2}}, \tag{A47}$$

where we decompose $\widehat{\mathbf{x}} = [\widehat{x}_1; \widetilde{\widehat{\mathbf{x}}}]$. In addition, (A45) is written as

$$\min_{\mathbf{u},\mathbf{x}} \max_{\boldsymbol{\beta}} \frac{1}{m} \sum_{i=1}^{m} \ell(u_i) + \frac{1}{m} \sum_{i=1}^{m} \beta_i u_i + \frac{1}{m} \sum_{i=1}^{m} \beta_i y_i \widetilde{\mathbf{a}}_i^T \widetilde{\mathbf{x}} - \frac{1}{m} \sum_{i=1}^{m} \beta_i y_i s_i x_1$$

or, in matrix form, as

$$\min_{\mathbf{u},\boldsymbol{x}} \max_{\boldsymbol{\beta}} \frac{1}{m}\boldsymbol{\beta}^T \mathbf{D}_y \widetilde{\mathbf{A}}\widetilde{\mathbf{x}} + \frac{1}{m} x_1 \boldsymbol{\beta}^T \mathbf{D}_y \mathbf{s} + \frac{1}{m}\boldsymbol{\beta}^T \mathbf{u} + \frac{1}{m}\sum_{i=1}^{m} \ell(u_i). \tag{A48}$$

where $\mathbf{D_y} := \mathrm{diag}(y_1, y_2, ..., y_m)$ is a diagonal matrix with $y_1, y_2, ...y_m$ on the diagonal, $\mathbf{s} = [s_1, \ldots, s_m]^T$ and $\widetilde{\mathbf{A}}$ is an $m \times (n-1)$ matrix with rows $\widetilde{\mathbf{a}}_i^T, i \in [m]$.

In (A48), we recognize that the first term has the bilinear form required by the GMT in (A41a). The rest of the terms form the function ψ in (A41a): they are independent of $\widetilde{\mathbf{A}}$ and convex-concave as desired by the CGMT. Therefore, we express (A44) in the desired form of a PO and for the rest of the proof we analyze the probabilistically equivalent AO problem. In view of (A41b), this is given as follows,

$$\min_{\mathbf{u},\boldsymbol{x}} \max_{\boldsymbol{\beta}} \frac{1}{m}\|\widetilde{\mathbf{x}}\|_2 \mathbf{g}^T \mathbf{D}_y \boldsymbol{\beta} + \frac{1}{m}\left\|\mathbf{D}_y \boldsymbol{\beta}\right\|_2 \mathbf{h}^T \widetilde{\mathbf{x}} - \frac{1}{m} x_1 \boldsymbol{\beta}^T \mathbf{D}_y \mathbf{s} + \frac{1}{m}\boldsymbol{\beta}^T \mathbf{u} + \frac{1}{m}\sum_{i=1}^{m} \ell(u_i)\,, \tag{A49}$$

where as in (A41b) $\mathbf{g} \sim \mathcal{N}(0, I_m)$ and $\mathbf{h} \sim \mathcal{N}(0, I_{n-1})$.

Appendix B.3. Analysis of the Auxiliary Optimization

Here, we show how to analyze the AO in (A49). To begin with, note that $y_i \in \{\pm 1\}$, therefore $\mathbf{D_y g} \sim \mathcal{N}(0, I_m)$ and $\left\|\mathbf{D_y}\boldsymbol{\beta}\right\|_2 = \|\boldsymbol{\beta}\|_2$. In addition, let us denote the first entry x_1 of $\mathbf{x}$ as

$$\mu := x_1.$$

The first step is to optimize over the direction of $\widetilde{\mathbf{x}}$. For this, we express the AO as:

$$\min_{\mathbf{u},\mu,\alpha\geq 0} \min_{\|\widetilde{\mathbf{x}}\|_2 = \alpha} \max_{\boldsymbol{\beta}} \frac{1}{m}\|\widetilde{\mathbf{x}}\|_2 \mathbf{g}^T \mathbf{D}_y \boldsymbol{\beta} + \frac{1}{m}\left\|\mathbf{D}_y \boldsymbol{\beta}\right\|_2 \mathbf{h}^T \widetilde{\mathbf{x}} - \frac{1}{m}\mu \boldsymbol{\beta}^T \mathbf{D}_y \mathbf{s} + \frac{1}{m}\boldsymbol{\beta}^T \mathbf{u} + \frac{1}{m}\sum_{i=1}^{m} \ell(u_i)\,, \tag{A50}$$

Now, denote $\widetilde{\mathbf{x}}_\star = -\alpha\|\mathbf{h}\|_2$ and observe that for every $\boldsymbol{\beta}$ the objective above is minimized (with respect to to $\widetilde{\mathbf{x}}$) at $\widetilde{\mathbf{x}}_\star$. Thus, it follows by [23] (Lem. 8) that (A50) simplifies to

$$\min_{\alpha\geq 0,\mu,\mathbf{u}} \max_{\boldsymbol{\beta}} \frac{1}{m}\alpha \mathbf{g}^T \boldsymbol{\beta} - \frac{\alpha}{m}\|\boldsymbol{\beta}\|_2 \|\mathbf{h}\|_2 - \frac{1}{m}\mu \mathbf{s}^T \mathbf{D_y}\boldsymbol{\beta} + \frac{1}{m}\boldsymbol{\beta}^T \mathbf{u} + \frac{1}{m}\sum_{i=1}^{m} \ell(u_i). \tag{A51}$$

Next, let $\gamma := \frac{\|\boldsymbol{\beta}\|_2}{\sqrt{m}}$ and optimize over the direction of $\boldsymbol{\beta}$ to yield

$$\min_{\alpha\geq 0,\mathbf{u},\mu} \max_{\gamma\geq 0} \frac{\gamma}{\sqrt{m}}\left\|\alpha \mathbf{g} - \mu \mathbf{D_y}\mathbf{s} + \mathbf{u}\right\|_2 - \frac{\alpha}{\sqrt{m}}\gamma\|\mathbf{h}\|_2 + \frac{1}{m}\sum_{i=1}^{m} \ell(u_i). \tag{A52}$$

To continue, we utilize the fact that for all $x \in \mathbb{R}$, $\min_{\tau>0} \frac{\tau}{2} + \frac{x^2}{2\tau m} = \frac{x}{\sqrt{m}}$. Hence,

$$\frac{\gamma}{\sqrt{m}}\left\|\alpha \mathbf{g} - \mu \mathbf{D_y}\mathbf{s} + \mathbf{u}\right\|_2 = \min_{\tau>0} \frac{\gamma\tau}{2} + \frac{\gamma}{2\tau m}\left\|-\alpha \mathbf{g} + \mu \mathbf{D_y}\mathbf{s} - \mathbf{u}\right\|_2^2.$$

With this trick, the optimization over $\mathbf{u}$ becomes separable over its coordinates $u_i, i \in [m]$. By inserting this in (A52), we have

$$\min_{\alpha\geq 0,\mathbf{u},\mu} \max_{\gamma\geq 0} \min_{\tau>0} \frac{\gamma\tau}{2} - \frac{\alpha}{\sqrt{m}}\gamma\|\mathbf{h}\|_2 + \frac{\gamma}{2\tau m}\sum_{i=1}^{m}(-\alpha g_i + \mu y_i s_i - u_i)^2 + \frac{1}{m}\sum_{i=1}^{m} \ell(u_i),$$

Now, we show that the objective function above is convex-concave. Clearly, the function is linear (thus, concave in γ). Moreover, from Lemma A1, the function $\frac{1}{2\tau}(\alpha g_i + \mu y_i s_i - u_i)^2$ is jointly convex in (α, μ, u_i, τ). The rest of the terms are clearly convex and this completes

the argument. Hence, with a permissible change in the order of min-max, we arrive at the following convenient form (Here, we skip certain technical details in this argument regarding boundedness of the constraint sets in (A49). While they are not trivial, they can be handled with the same techniques used in [15,60].):

$$\min_{\mu,\alpha\geq 0,\tau>0}\max_{\gamma\geq 0}\frac{\gamma\tau}{2}-\frac{\alpha}{\sqrt{m}}\gamma\|\mathbf{h}\|_2+\frac{1}{m}\sum_{i=1}^{m}\mathcal{M}_{\ell}\left(-\alpha g_i+\mu s_i y_i;\frac{\tau}{\gamma}\right), \tag{A53}$$

where recall the definition of the Moreau envelope in (A1). As to now, we have reduced the AO into a random min-max optimization over only four scalar variables in (A53). For fixed μ,α,τ,γ, direct application of the weak law of large numbers shows that the objective function of (A53) converges in probability to the following as $m,n\to\infty$ and $\frac{m}{n}=\delta$:

$$\gamma\frac{\tau}{2}-\frac{\alpha\gamma}{\sqrt{\delta}}+\mathbb{E}\left[\mathcal{M}_{\ell}\left(\alpha G+\mu YS;\frac{\tau}{\gamma}\right)\right],$$

where $G,S\sim\mathcal{N}(0,1)$ and $Y\sim f(S)$ (in view of (A46)). Based on that, it can be shown (similar arguments are developed in [15,60]) that the random optimizers α_n and μ_n of (A53) converge to the deterministic optimizers α and μ of the following (deterministic) optimization problem (whenever these are bounded as the statement of the theorem requires):

$$\min_{\alpha\geq 0,\mu,\tau>0}\max_{\gamma\geq 0}\gamma\frac{\tau}{2}-\frac{\alpha\gamma}{\sqrt{\delta}}+\mathbb{E}\left[\mathcal{M}_{\ell}\left(\alpha G+\mu YS;\frac{\tau}{\gamma}\right)\right]. \tag{A54}$$

At this point, recall that α represents the norm of $\tilde{\mathbf{x}}$ and μ the value of x_1. Thus, in view of (i) (A47); (ii) the equivalence between the PO and the AO; and (iii) our derivations thus far, we have that with probability approaching 1,

$$\lim_{n\to+\infty}\operatorname{corr}(\widehat{\mathbf{x}};\mathbf{x}_0)=\frac{\mu}{\sqrt{\mu^2+\alpha^2}},$$

where μ and α are the minimizers in (A54). The three equations in (8) are derived by the first-order optimality conditions of the optimization in (A54). We show this next.

Appendix B.4. Convex-Concavity and First-Order Optimality Conditions

First, we prove that the objective function in (A54) is convex–concave. For convenience define the function $F:\mathbb{R}^4\to\mathbb{R}$ as follows

$$F(\alpha,\mu,\tau,\gamma):=\frac{\gamma\tau}{2}-\frac{\alpha\gamma}{\sqrt{\delta}}+\mathbb{E}\left[\mathcal{M}_{\ell}\left(\alpha G+\mu YS;\frac{\tau}{\gamma}\right)\right]. \tag{A55}$$

Based on Lemma A2, it immediately follows that, if ℓ is convex, F is jointly convex in (α,μ,τ). To prove concavity of F based on γ, it suffices to show that $\mathcal{M}_{\ell}(x;1/\gamma)$ is concave in γ for all $x\in\mathbb{R}$. To show this, we note that

$$\mathcal{M}_{\ell}(x;1/\gamma)=\min_{u}\frac{\gamma}{2}(x-u)^2+\ell(u),$$

which is the point-wise minimum of linear functions of γ. Thus, using Proposition A3(b), we conclude that $\mathcal{M}_{\ell}(x;1/\gamma)$ is concave in γ. This completes the proof of convex-concavity of the function F in (A55) when ℓ is convex. By direct differentiation and applying

Proposition A7(a), the first-order optimality conditions of the min–max optimization in (A54) are as follows:

$$\mathbb{E}\left[SY \cdot \mathcal{M}'_{\ell,1}\left(\alpha G + \mu SY; \frac{\tau}{\gamma}\right)\right] = 0, \tag{A56a}$$

$$\mathbb{E}\left[G \cdot \mathcal{M}'_{\ell,1}\left(\alpha G + \mu SY; \frac{\tau}{\gamma}\right)\right] = \frac{\gamma}{\sqrt{\delta}}, \tag{A56b}$$

$$\frac{\gamma}{2} + \frac{1}{\gamma}\mathbb{E}\left[\mathcal{M}'_{\ell,2}\left(\alpha G + \mu SY; \frac{\tau}{\gamma}\right)\right] = 0, \tag{A56c}$$

$$-\frac{\alpha}{\sqrt{\delta}} - \frac{\tau}{\gamma^2}\mathbb{E}\left[\mathcal{M}'_{\ell,2}\left(\alpha G + \mu SY; \frac{\tau}{\gamma}\right)\right] + \frac{\tau}{2} = 0. \tag{A56d}$$

Next, we show how these equations simplify to the following system of equations (same as (8):

$$\mathbb{E}\left[Y\,S \cdot \mathcal{M}'_{\ell,1}(\alpha G + \mu SY; \lambda)\right] = 0, \tag{A57a}$$

$$\lambda^2\,\delta\,\mathbb{E}\left[\left(\mathcal{M}'_{\ell,1}(\alpha G + \mu SY; \lambda)\right)^2\right] = \alpha^2, \tag{A57b}$$

$$\lambda\,\delta\,\mathbb{E}\left[G \cdot \mathcal{M}'_{\ell,1}(\alpha G + \mu SY; \lambda)\right] = \alpha. \tag{A57c}$$

Let $\lambda := \frac{\tau}{\gamma}$. First, (A57a) is immediate from equation (A56a). Second, substituting γ from (A56c) in (A56d) yields $\tau = \frac{\alpha}{\sqrt{\delta}}$ or $\gamma = \frac{\alpha}{\lambda\sqrt{\delta}}$, which together with (A56b) leads to (A57c). Finally, (A57b) can be obtained by substituting $\gamma = \frac{\alpha}{\lambda\sqrt{\delta}}$ in (A56c) and using the fact that (see Proposition A1):

$$\mathcal{M}'_{\ell,2}(\alpha G + \mu SY; \lambda) = -\frac{1}{2}(\mathcal{M}'_{\ell,1}(\alpha G + \mu SY; \lambda))^2.$$

Appendix B.5. On the Uniqueness of Solutions to (A57)*: Proof of Proposition 1*

Here, we prove the claim of Proposition 1 through the following lemmas. As discussed in Remark 4, the main part of the proof is showing strict convex-concavity of F in (11). Lemma A6 proves that this is the case, and Lemmas A7 and A8 show that this is sufficient for the uniqueness of solutions to (A57). When put together, these complete the proof of Proposition 1.

Lemma A6 (Strict Convex-Concavity of (A55))**.** *Let $\ell : \mathbb{R} \to \mathbb{R}$ be proper and strictly convex function. Further, assume that ℓ is continuously differentiable with $\ell'(0) \neq 0$. In addition, assume that SY has positive density in the real line. Then, the function $F : \mathbb{R}^4 \to \mathbb{R}$ defined in* (A55) *is strictly convex in (α, μ, τ) and strictly concave in γ.*

Proof. The claim follows directly from the strict convexity-concavity properties of the expected Moreau-envelope proved in Propositions A5 and A6. Specifically, we apply Proposition A7. □

Lemma A7. *If the objective function in* (A55) *is strictly convex in (α, μ, τ) and strictly concave in γ, then* (A56) *has a unique solution $(\alpha, \mu, \tau, \gamma)$.*

Proof. Let $(\alpha_i, \mu_i, \tau_i, \gamma_i)$, $i = 1, 2$, be two different saddle points of (A55). For convenience, let $\mathbf{x}_i := (\alpha_i, \mu_i, \tau_i)$ for $i = 1, 2$. By strict-concavity in γ, for fixed values of $\mathbf{x} := (\alpha, \mu, \tau)$, the value of γ maximizing $F(\mathbf{x}, \gamma)$ is unique. Thus, if $\mathbf{x}_1 = \mathbf{x}_2$, then it must hold that $\gamma_1 = \gamma_2$, which is a contraction to our assumption of $(\mathbf{x}_1, \gamma_1) \neq (\mathbf{x}_2, \gamma_2)$. Similarly, we can use

strict-convexity to derive that $\gamma_1 \neq \gamma_2$. Then, based on the definition of the saddle point and strict convexity-concavity, the following two relations hold for $i = 1, 2$:

$$F(\mathbf{x}_i, \gamma) < F(\mathbf{x}_i, \gamma_i) < F(\mathbf{x}, \gamma_i), \quad \text{for all } \mathbf{x} \neq \mathbf{x}_i, \gamma \neq \gamma_i.$$

We choose $\mathbf{x} = \mathbf{x}_2, \gamma = \gamma_2$ for $i = 1$ and $\mathbf{x} = \mathbf{x}_1, \gamma = \gamma_1$ for $i = 2$ to find

$$F(\mathbf{x}_1, \gamma_2) < F(\mathbf{x}_1, \gamma_1) < F(\mathbf{x}_2, \gamma_1),$$
$$F(\mathbf{x}_2, \gamma_1) < F(\mathbf{x}_2, \gamma_2) < F(\mathbf{x}_1, \gamma_2).$$

From the above, it follows that $F(\mathbf{x}_1, \gamma_1) < F(\mathbf{x}_2, \gamma_2)$ and $F(\mathbf{x}_1, \gamma_1) > F(\mathbf{x}_2, \gamma_2)$, which is a contradiction. This completes the proof. □

Lemma A8. *If* (A56) *has a unique solution* $(\alpha^\star, \mu^\star, \tau^\star, \gamma^\star)$*, then* (A57) *has a unique solution* $(\alpha^\star, \mu^\star, \lambda^\star)$.

Proof. First, following the same approach of deriving Equations (A57) from (A56) in Appendix B.4, it is easy to see that existence of solution $(\alpha_1, \mu_1, \tau_1, \gamma_1)$ to (A56) implies existence of solution $(\alpha_1, \mu_1, \lambda_1 := \frac{\tau_1}{\gamma_1})$ to (A57). Now, for the sake of contradiction to the statement of the lemma, assume that there are two different triplets $\mathbf{v}_1 := (\alpha_1, \mu_1, \lambda_1)$ and $\mathbf{v}_2 := (\alpha_2, \mu_2, \lambda_2)$ with $\alpha_1, \alpha_2, \lambda_1, \lambda_2 > 0$ and satisfying (A57). Then, we can show that both $\mathbf{w}_i := (\alpha_i, \mu_i, \tau_i, \gamma_i)$ $i = 1, 2$, such that:

$$\tau_i := \frac{\alpha_i}{\sqrt{\delta}}, \quad \gamma_i = \frac{\alpha_i}{\lambda_i \sqrt{\delta}}, \qquad i = 1, 2,$$

satisfy the system of equations in (A56). However, since $\mathbf{v}_1 \neq \mathbf{v}_2$, it must be that $\mathbf{w}_1 \neq \mathbf{w}_2$. This contradicts the assumption of uniqueness of solutions to (A56) and completes the proof. □

Appendix C. Discussions on the Fundamental Limits for Binary Models

Appendix C.1. On the Uniqueness of Solutions to Equation $\kappa(\sigma) = \frac{1}{\delta}$

The existence of a solution to the equation $\kappa(\sigma) = \frac{1}{\delta}$ is proved in the previous section. However, it is not clear if the solution to this equation is unique, i.e., for any $\delta > 1$ there exists only one $\sigma_{\text{opt}} > 0$ such that $\kappa(\sigma_{\text{opt}}) = \frac{1}{\delta}$. If this is the case, then Equation (12) in Theorem 2 can be equivalently written as

$$\sigma_{\text{opt}} = \sigma, \text{ s.t. } \kappa(\sigma) = \frac{1}{\delta}.$$

Although we do not prove this claim, our numerical experiments in Figure A1 show that $\kappa(\cdot)$ is a monotonic function for noisy-signed, logistic and Probit measurements, implying the uniqueness of solution to the equation $\kappa(\sigma) = \frac{1}{\delta}$ for all $\delta > 1$.

Appendix C.2. Distribution of SY in Special Cases

We derive the following densities for SY for the special cases ($\|\mathbf{x}_0\|_2 = 1$):

- *Signed*: $p_{SY}(w) = \sqrt{\frac{2}{\pi}} \exp(-w^2/2)\, \mathbb{1}_{\{w \geq 0\}}$.
- *Logistic*: $p_{SY}(w) = \sqrt{\frac{2}{\pi}} \frac{\exp(-w^2/2)}{1+\exp(-w)}$.
- *Probit*: $p_{SY}(w) = \sqrt{\frac{2}{\pi}}\, \Phi(w) \exp(-w^2/2)$.

In particular, we numerically observe that for logistic and Probit models; the resulting densities are similar to the density of a gaussian distribution derived according to $\mathcal{N}(\mathbb{E}[SY], \text{Var}[SY])$. Figure A2 illustrates this similarity for these two models. As discussed

in Corollary 1, this similarity results in the tightness of the lower bound achieved for σ_{opt} in Equation (23).

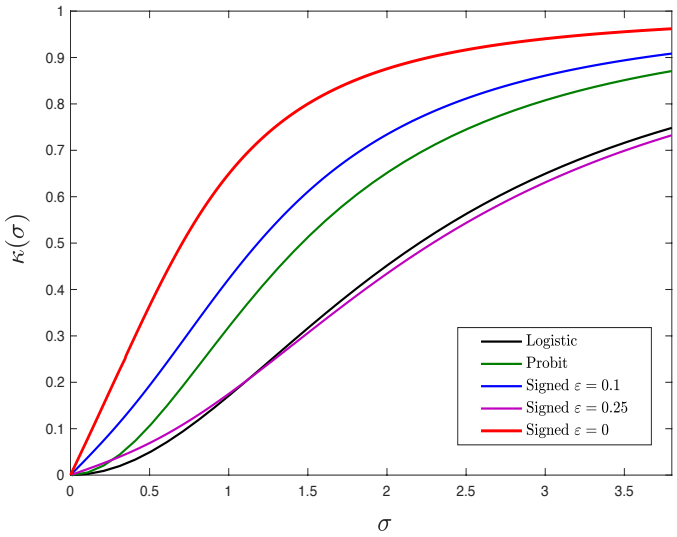


Figure A1. The value of $\kappa(\sigma)$ as in Theorem 2 for various measurement models. Since $\kappa(\sigma)$ is a monotonic function of σ, the solution to $\kappa(\sigma) = 1/\delta$ determines the minimum possible value of σ.

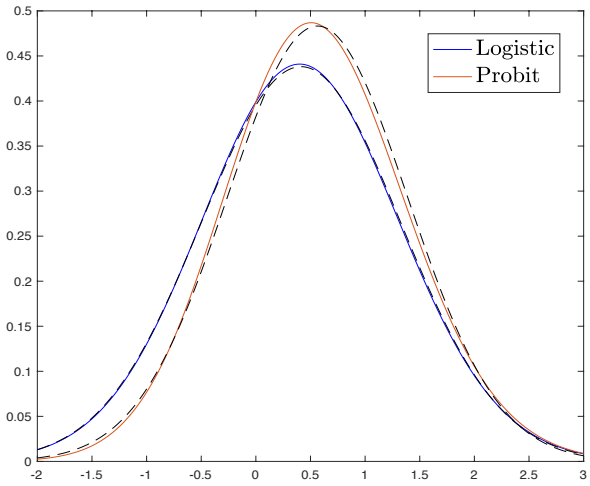


Figure A2. Probability distribution function of SY for the logistic and Probit models ($\|\mathbf{x}_0\|_2 = 1$) compared with the probability distribution function of the Gaussian random variable (dashed lines) with the same mean and variance i.e., $\mathcal{N}(\mathbb{E}[SY], \text{Var}[SY])$.

Appendix D. Proofs and Discussions on the Optimal Loss Function

Appendix D.1. Proof of Theorem 3

We show that the triplet $(\mu = 1, \alpha = \sigma_{\text{opt}}, \lambda = 1)$ is a solution to Equations (8) for ℓ chosen as in (29). Using Proposition A2 in the Appendix, we rewrite ℓ_{opt} using the Fenchel–Legendre conjugate as follows:

$$\ell_{\text{opt}}(w) = \left(q + \alpha_1 q + \alpha_2 \log p_{W_{\text{opt}}}\right)^{\star}(w) - q(w), \tag{A58}$$

where $q(w) = w^2/2$. For a function f, its Fenchel–Legendre conjugate is defined as:

$$f^{\star}(x) = \max_{y} \; xy - f(y).$$

Next, we use the fact that, for any proper, closed and convex function f, it holds that $(f^\star)^\star = f$ [61] (theorem 12.2). Therefore, noting that $q + \alpha_1 q + \alpha_2 \log p_{W_{\text{opt}}}$ is a convex function (see the proof of Lemma A9 in the Appendix), combined with (A58), it yields that

$$(\ell_{\text{opt}} + q)^\star = q + \alpha_1 q + \alpha_2 \log p_{W_{\text{opt}}}. \tag{A59}$$

Additionally, using Proposition A2, we find that $\mathcal{M}_{\ell_{\text{opt}}}(w;1) = q(w) - (q + \ell_{\text{opt}})^\star(w)$, which by (A59) reduces to:

$$\mathcal{M}_{\ell_{\text{opt}}}(w;1) = -\alpha_1 q(w) - \alpha_2 \log p_{W_{\text{opt}}}(w).$$

Thus, by differentiation, we find that ℓ_{opt} satisfies (28) with $c = -1$, i.e.,

$$\mathcal{M}'_{\ell_{\text{opt}},1}(w;1) = -\alpha_1 w - \alpha_2 \cdot \xi_{W_{\text{opt}}}(w). \tag{A60}$$

Next, we establish the desired by directly substituting (A60) into the system of equations in (19). First, using the values of α_1 and α_2 in (30), as well as the fact that $\kappa(\sigma_{\text{opt}}) = 1/\delta$, we have the following chain of equations:

$$\begin{aligned}
\mathbb{E}\left[\left(\mathcal{M}'_{\ell_{\text{opt}},1}(W_{\text{opt}};1)\right)^2\right] &= \mathbb{E}\left[(\alpha_1 W_{\text{opt}} + \alpha_2\, \xi_{W_{\text{opt}}}(W_{\text{opt}}))^2\right] \\
&= \alpha_1^2\,(\sigma_{\text{opt}}^2 + 1) + \alpha_2^2\, \mathcal{I}(W_{\text{opt}}) + 2\,\alpha_1\alpha_2\, \mathbb{E}\left[W_{\text{opt}} \cdot \xi_{W_{\text{opt}}}(W_{\text{opt}})\right] \\
&= \frac{1 + \sigma_{\text{opt}}^2\left(\sigma_{\text{opt}}^2\, \mathcal{I}(W_{\text{opt}}) - 1\right)}{\delta^2\left(\sigma_{\text{opt}}^2\, \mathcal{I}(W_{\text{opt}}) + \mathcal{I}(W_{\text{opt}}) - 1\right)} = \frac{\sigma_{\text{opt}}^2}{\delta^2\, \kappa(\sigma_{\text{opt}})} \\
&= \sigma_{\text{opt}}^2/\delta.
\end{aligned} \tag{A61}$$

This shows (8b). Second, using again the specified values of α_1 and α_2, a similar calculation yields

$$\begin{aligned}
\mathbb{E}\left[\mathcal{M}'_{\ell_{\text{opt}},1}(W_{\text{opt}};1)\,\xi_{W_{\text{opt}}}(W_{\text{opt}})\right] &= -\mathbb{E}\left[\left(\alpha_1 W_{\text{opt}} + \alpha_2\, \xi_{W_{\text{opt}}}(W_{\text{opt}})\right)\, \xi_{W_{\text{opt}}}(W_{\text{opt}})\right] \\
&= \alpha_1 - \alpha_2\, \mathcal{I}(W_{\text{opt}}) \\
&= -1/\delta.
\end{aligned} \tag{A62}$$

Recall from (17) that $\mathbb{E}\left[G \cdot \mathcal{M}'_{\ell_{\text{opt}},1}(W_{\text{opt}};1)\right] = -\sigma_{\text{opt}}\, \mathbb{E}\left[\mathcal{M}'_{\ell_{\text{opt}},1}(W_{\text{opt}};1)\,\xi_{W_{\text{opt}}}(W_{\text{opt}})\right]$. This combined with (A62) yields (8c). Finally, we use again (A60) and the specified values of α_1 and α_2 to find that

$$\begin{aligned}
\mathbb{E}\left[W_{\text{opt}} \cdot \mathcal{M}'_{\ell_{\text{opt}},1}(W_{\text{opt}};1)\right] &= \mathbb{E}\left[W_{\text{opt}} \cdot (-\alpha_1 W_{\text{opt}} - \alpha_2\, \xi_{W_{\text{opt}}}(W_{\text{opt}}))\right] \\
&= -\alpha_1\, \mathbb{E}\left[W_{\text{opt}}^2\right] - \alpha_2\, \mathbb{E}\left[W_{\text{opt}}\, \xi_{W_{\text{opt}}}(W_{\text{opt}})\right] \\
&= -\alpha_1(\sigma_{\text{opt}}^2 + 1) - \alpha_2 \int_{-\infty}^{\infty} w\, p'_{W_{\text{opt}}}(w)\, \mathrm{d}w \\
&= -\alpha_1(\sigma_{\text{opt}}^2 + 1) + \alpha_2 \\
&= \sigma_{\text{opt}}^2/\delta.
\end{aligned} \tag{A63}$$

However, using (17), it holds that

$$\mathbb{E}\Big[W_{\text{opt}} \cdot \mathcal{M}'_{\ell_{\text{opt}},1}(W_{\text{opt}};1)\Big] =$$
$$-\sigma^2_{\text{opt}}\,\mathbb{E}\Big[\mathcal{M}'_{\ell_{\text{opt}},1}(W_{\text{opt}};1)\xi_{W_{\text{opt}}}(W_{\text{opt}})\Big] + \mathbb{E}\Big[Y\,S \cdot \mathcal{M}'_{\ell_{\text{opt}},1}(W_{\text{opt}};\lambda)\Big].$$

This combined with (A63) and (A62) shows that $\mathbb{E}\Big[Y\,S \cdot \mathcal{M}'_{\ell_{\text{opt}},1}(W_{\text{opt}};\lambda)\Big] = 0$, as desired to satisfy (8a). This completes the proof of the theorem.

Appendix D.2. On the Convexity of Optimal Loss Function

Here, we provide a sufficient condition for $\ell_{\text{opt}}(w)$ to be convex.

Lemma A9. *The optimal loss function as defined in Theorem 3 is convex if*

$$(\log(p_{W_\sigma}))''(w) \leq -\frac{1}{\sigma^2+1}, \quad \textit{for all } w \in \mathbb{R} \textit{ and } \sigma \geq 0.$$

Proof. Using (A9) optimal loss function is written in the following form

$$\ell_{\text{opt}}(w) = \Big(q + \alpha_1 q + \alpha_2 \log(p_{W_{\text{opt}}})\Big)^\star(w) - q(w). \tag{A64}$$

Next, we prove that $q + \alpha_1 q + \alpha_2 \log(p_{W_{\text{opt}}})$ is a convex function. We first show that both α_1 and α_2 are positive numbers for all values of σ_{opt}. We first note that, since G and SY are independent random variables, $\sigma^2_{\text{opt}}\mathcal{I}(W_{\text{opt}}) < \sigma^2_{\text{opt}}\mathcal{I}(\sigma_{\text{opt}}\,G) = 1$. Therefore,

$$1 - \sigma^2_{\text{opt}}\mathcal{I}(W_{\text{opt}}) > 0. \tag{A65}$$

Additionally, following the Cramer–Rao bound [53] for Fisher information yields that:

$$\mathcal{I}(W_{\text{opt}}) > \frac{1}{\mathbb{E}\big[(W_{\text{opt}} - \mathbb{E}[W_{\text{opt}}])^2\big]} = \frac{1}{1+\sigma^2_{\text{opt}} - (\mathbb{E}[SY])^2}.$$

Using this inequality for $\mathcal{I}(W_{\text{opt}})$, we derive that

$$\sigma^2_{\text{opt}}\mathcal{I}(W_{\text{opt}}) + \mathcal{I}(W_{\text{opt}}) - 1 > 0. \tag{A66}$$

From (A65) and (A66), it follows that $\alpha_1, \alpha_2 > 0$.
Based on the definition of the random variable W_{opt}:

$$\log p_{W_{\text{opt}}}(w) = -w^2/(2\sigma^2_{\text{opt}}) + \log \int_{-\infty}^{\infty} \exp\Big((2wz - z^2)/2\sigma^2_{\text{opt}}\Big)\, p_{SY}(z)\,\mathrm{d}z + c,$$

where c is a constant independent of w. By differentiating twice, we see that

$$\log \int_{-\infty}^{\infty} \exp\Big((2wz - z^2)/2\sigma^2_{\text{opt}}\Big)\, p_{SY}(z)\,\mathrm{d}z$$

is a convex function of w. Therefore, to prove that $q + \alpha_1 q + \alpha_2 \log(p_{W_{\text{opt}}})$ is a convex function, it is sufficient to prove that $(1 + \alpha_1 - \alpha_2/\sigma^2_{\text{opt}})q$ is a convex function or equivalently $1 + \alpha_1 - \alpha_2/\sigma^2_{\text{opt}} \geq 0$. Replacing values of α_1, α_2 and recalling the equation for σ_{opt} yields that

$$1 + \alpha_1 - \alpha_2/\sigma^2_{\text{opt}} = 0,$$

which implies the convexity of $q + \alpha_1 q + \alpha_2 \log(p_{W_{\text{opt}}})$. To obtain the derivative of ℓ_{opt}, we use the result in [61] (Cor. 23.5.1), which states that, for a convex function f,

$$(f^\star)' = (f')^{-1}.$$

Therefore, following (A64),

$$\ell'_{\text{opt}}(w) = (q' + \alpha_1 q' + \alpha_2 (\log(p_{W_{\text{opt}}}))')^{-1}(w) - w. \tag{A67}$$

Differentiating again and using the properties of inverse function yields that

$$\ell''_{\text{opt}}(w) = \frac{1}{1 + \alpha_1 + \alpha_2 (\log(p_{W_{\text{opt}}}))''(g(w))} - 1, \tag{A68}$$

where

$$g(w) := (q' + \alpha_1 q' + \alpha_2 (\log(p_{W_{\text{opt}}}))')^{-1}(w).$$

Note that the denominator of (A68) is nonnegative since it is second derivative of a convex function. Therefore, it is evident from (A68) that a sufficient condition for the convexity of ℓ_{opt} is that

$$\alpha_1 + \alpha_2 (\log(p_{W_{\text{opt}}}))''(w) \leq 0, \quad \text{for all } w \in \mathbb{R},$$

or

$$1 - \sigma^2_{\text{opt}} \mathcal{I}(W_{\text{opt}}) + (\log(p_{W_{\text{opt}}}))''(w) \leq 0.$$

This condition is satisfied if the statement of the lemma holds for $\sigma = \sigma_{\text{opt}}$:

$$1 - \sigma^2_{\text{opt}} \mathcal{I}(W_{\text{opt}}) + (\log(p_{W_{\text{opt}}}))''(w) \leq 1 - \sigma^2_{\text{opt}} \mathcal{I}(W_{\text{opt}}) - \frac{1}{1 + \sigma^2_{\text{opt}}} < 0,$$

where we use (A66) in the last inequality. This concludes the proof. □

Appendix D.2.1. Provable Convexity of the Optimal Loss Function for Signed Model

In the case of signed model, it can be proved that the conditions of Lemma A9 is satisfied. Since $W_\sigma = \sigma G + SY$, we derive the probability density of W_σ as follows:

$$p_{W_\sigma}(w) = p_{\sigma G}(w) * p_{SY}(w) = \frac{\exp(-w^2/(2 + 2\sigma^2))}{\sqrt{2\pi(1 + \sigma^2)}} \cdot f(w),$$

where

$$f(w) = 2 - 2\mathrm{Q}(w/(\sigma\sqrt{2 + 2\sigma^2})).$$

Direct calculation shows that f is a log-concave function for all $w \in \mathbb{R}$. Therefore,

$$\begin{aligned}(\log(p_{W_\sigma}))''(w) &= -\frac{1}{\sigma^2 + 1} + (\log(f))''(w) \\ &\leq -\frac{1}{\sigma^2 + 1}.\end{aligned}$$

This proves the convexity of optimal loss function derived according to Theorem 3 when measurements follow the signed model.

Appendix E. Noisy-Signed Measurement Model

Consider a noisy-signed label function as follows:

$$y_i = f_\varepsilon(\mathbf{a}_i^T \mathbf{x}_0) = \begin{cases} \text{sign}(\mathbf{a}_i^T \mathbf{x}_0) & \text{, w.p.} 1 - \varepsilon, \\ -\text{sign}(\mathbf{a}_i^T \mathbf{x}_0) & \text{, w.p.} \varepsilon, \end{cases}$$

where $\varepsilon \in [0, 1/2]$.

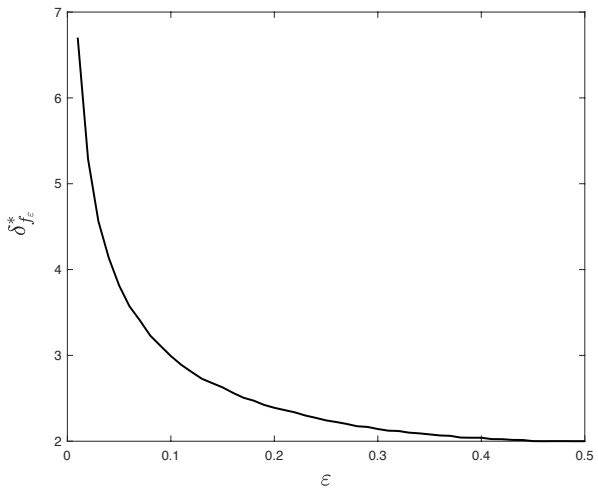


Figure A3. The value of the threshold $\delta^\star_{f_\varepsilon}$ in (A69) as a function of probability of error $\varepsilon \in [0, 1/2]$. For logistic and hinge losses, the set of minimizers in (3) is bounded (as required by Theorem 1) iff $\delta > \delta^\star_{f_\varepsilon}$.

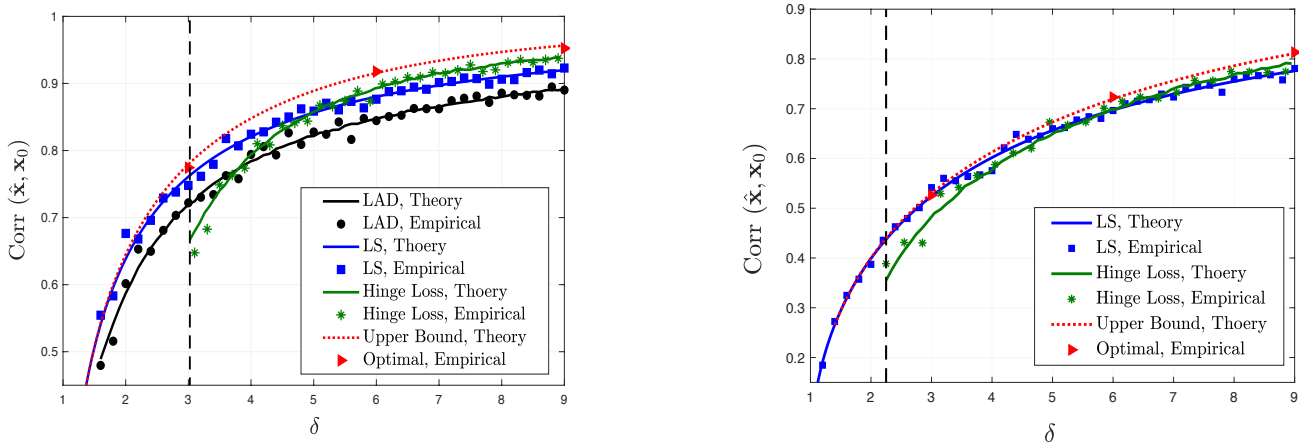
In the case of signed measurements, i.e., $y_i = \text{sign}(\mathbf{a}_i^T \mathbf{x}_0)$, it can be observed that for all possible values of δ, the condition (34) in Section 4.2 holds for $\mathbf{x}_s = \mathbf{x}_0$. This implies the separability of data and therefore the solution to the optimization problem (3) is unbounded for all δ. However, in the case of noisy signed label function, boundedness or unboundedness of solutions to (3) depends on δ. As discussed in Section 4.2, the minimum value of δ for bounded solutions is derived from the following:

$$\delta^\star_{f_\varepsilon}(\varepsilon) := \left(\min_{c \in \mathbb{R}} \mathbb{E}\Big[(G + c\, S\, Y)_-^2 \Big] \right)^{-1}, \tag{A69}$$

where $Y = f_\varepsilon(S)$. It can be checked analytically that $\delta^\star_{f_\varepsilon}$ is a decreasing function of ε with $\delta^\star_{f_\varepsilon}(0^+) = +\infty$ and $\delta^\star_{f_\varepsilon}(1/2) = 2$.

In Figure A3, we numerically evaluate the threshold value $\delta^\star_{f_\varepsilon}$ as a function of the probability of error ε. For $\delta < \delta^\star_{f_\varepsilon}$, the set of minimizers of the (3) with logistic or hinge loss is unbounded.

The performances of LS, LAD and hinge loss functions for noisy-signed measurement model with $\varepsilon = 0.1$ and $\varepsilon = 0.25$ are demonstrated in Figure A4a,b, respectively. Comparing performances of least-squares and hinge loss functions suggest that hinge loss is robust to measurement corruptions, as for moderate to large values of δ it outperforms the LS estimator. Theorem 1 opens the way to analytically confirm such conclusions, which is an interesting future direction.



(**a**) $\varepsilon = 0.1$ (**b**) $\varepsilon = 0.25$

Figure A4. Comparisons between analytical and empirical results for the least-squares (LS), least-absolute deviations and hinge loss functions along with the upper bound on performance and the empirical performance of optimal loss function as in Theorem 3, for noisy-signed measurement model with $\varepsilon = 0.1$ (**a**) and $\varepsilon = 0.25$ (**b**). The vertical dashed lines are evaluated by (A69) and represent $\delta^{\star}_{f_\varepsilon} \approx 3$ and 2.25 for $\varepsilon = 0.1$ and 0.25, respectively.

Appendix F. On LS Performance for Binary Models

Appendix F.1. Proof of Corollary 2

To get the values of α and μ as in the statement of the corollary, we show how to simplify Equations (8) for $\ell(t) = (t-1)^2$. In this case, the proximal operator admits a simple expression:

$$\text{prox}_{\ell}(x;\lambda) = (x + 2\lambda) \Big/ (1 + 2\lambda).$$

In addition, $\ell'(t) = 2(t-1)$. Substituting these in (14a) gives the formula for μ as follows:

$$\begin{aligned} 0 = \mathbb{E}[YS(\alpha G + \mu SY - 1)] &= \mu\, \mathbb{E}[S^2] - \mathbb{E}[YS] \\ &\Longrightarrow \mu = \mathbb{E}[YS], \end{aligned}$$

where we have also used from (7) that $\mathbb{E}[S^2] = 1$ and G is independent of S. In addition, since $\ell''(t) = 2$, direct application of (A7) gives

$$1 = \lambda\delta \frac{2}{1+2\lambda} \Longrightarrow \lambda = \frac{1}{2(\delta - 1)}.$$

Finally, substituting the value of λ into (14b), we obtain the desired value for α as follows:

$$\begin{aligned} \alpha^2 &= 4\lambda^2\delta\, \mathbb{E}\Big[(\text{prox}_{\ell}(\alpha G + \mu SY;\lambda) - 1)^2\Big] \\ &= \frac{4\lambda^2}{(1+2\lambda)^2}\delta\, \mathbb{E}\Big[(\alpha G + \mu SY - 1)^2\Big] \\ &= \frac{4\lambda^2\delta}{(1+2\lambda)^2}(\alpha^2 + \mu^2 + 1 - 2\mu\mathbb{E}[SY]) \\ &= \frac{1}{\delta}(\alpha^2 + 1 - (\mathbb{E}[SY])^2) \\ &\Longrightarrow \alpha = \sqrt{1 - (\mathbb{E}[SY])^2} \cdot \sqrt{\frac{1}{\delta - 1}}. \end{aligned}$$

Appendix F.2. Discussion

Linear vs. Binary

On the one hand, Corollary 2 shows that least-squares performance for binary measurements satisfies

$$\lim_{n\to\infty} \left\| \widehat{\mathbf{x}} - \frac{\mu}{\|\mathbf{x}_0\|_2} \cdot \mathbf{x}_0 \right\|_2^2 = \tau^2 \cdot \frac{1}{\delta - 1}, \tag{A70}$$

where μ is as in (32) and $\tau^2 := 1 - (\mathbb{E}[SY])^2$. On the other hand, it is well-known (e.g., see references in [15] (Sec. 5.1)) that least-squares for (scaled) linear measurements with additive Gaussian noise (i.e., $y_i = \rho \mathbf{a}_i^T \mathbf{x}_0 + \sigma z_i$, $z_i \sim \mathcal{N}(0,1)$) leads to an estimator that satisfies

$$\lim_{n\to\infty} \|\widehat{\mathbf{x}} - \rho \cdot \mathbf{x}_0\|_2^2 = \sigma^2 \cdot \frac{1}{\delta - 1}. \tag{A71}$$

Direct comparison of (A70) to (A71) suggests that least-squares with binary measurements performs the same as if measurements were linear with scaling factor $\rho = \mu / \|\mathbf{x}_0\|_2$ and noise variance $\sigma^2 = \tau^2 = \alpha^2(\delta - 1)$. This worth-mentioning conclusion is not new, as it is proved in [40,43,56,62]. We include a short discussion on the relation to this prior work in the following paragraph. We highlight that all these existing results are limited to a least-squares loss unlike our general analysis.

Prior work. There is a lot of recent work on the use of least-squares-type estimators for recovering signals from nonlinear measurements of the form $y_i = h(\mathbf{a}_i^T \mathbf{x}_0)$ with Gaussian vectors $\mathbf{a}_i$. The original work that suggests least-squares as a reasonable estimator in this setting is due to Brillinger [56]. In his 1982 paper, Brillinger studied the problem in the classical statistics regime (namely, n is fixed not scaling with $m \to +\infty$) and he proved for the least-squares solution satisfies

$$\lim_{m\to+\infty} \frac{1}{m} \left\| \widehat{\mathbf{x}} - \frac{\mu}{\|\mathbf{x}_0\|_2} \cdot \mathbf{x}_0 \right\|_2^2 = \tau^2,$$

where

$$\begin{aligned} \mu &= \mathbb{E}[SY], \qquad S \sim \mathcal{N}(0,1), \\ \tau^2 &= \mathbb{E}[(Y - \mu S)^2]. \end{aligned} \tag{A72}$$

and the expectations are with respect to S and possible randomness of f. Evaluating (A72) for $Y = f_\varepsilon(S)$ leads to the same values for μ and τ^2 in (A70). In other works, (A70) for $\delta \to +\infty$ indeed recovers Brillinger's result. The extension of Brillinger's original work to the high-dimensional setting (both m, n large) was first studied by Plan and Vershynin [40], who derived (non-sharp) non-asymptotic upper bounds on the performance of constrained least-squares (such as the Lasso). Shortly after, Thrampoulidis et al. [43] extended this result to *sharp* asymtpotic predictions and to regularized least-squares. In particular, Corollary 2 is a special case of the main theorem in [43]. Several other interesting extensions of the result by Plan and Vershynin have recently appeared in the literature (e.g., [41,62–64]). However, the one in [43] is the only one to give results that are sharp in the flavor of this paper. Our work, extends the result of Thrampoulidis et al. [43] to general loss functions beyond least-squares. The techniques of Thrampoulidis et al. [43] that have guided the use of the CGMT in our context were also recently applied by Dhifallah et al. [60] in the context of phase-retrieval.

Appendix G. Fundamental Limits for Gaussian-Mixture Models: Proofs for Section 5

Appendix G.1. Proof of Corollary 3

The proof follows directly by noting that, when $\ell(t) = (t-1)^2$, it holds that $\mathcal{M}_\ell(x;\lambda) = \frac{(x-1)^2}{2\lambda+1}$. By inserting this into (38a) and simplifying the equations, we find the value of μ:

$$\mu = \frac{r}{1+r^2}.$$

Similarly, we derive λ using Equation (38c):

$$\lambda = \frac{1}{2(\delta-1)}.$$

Substituting these values of μ and λ into (38b) yields that

$$\alpha^2 = \frac{1}{\delta-1}\cdot\frac{1}{r^2+1}\,.$$

Recalling that $\sigma_{LS} = \alpha/\mu$ concludes the proof.

Appendix G.2. Proof of Theorem 5

The high-level steps of the proof follow the proof of Theorem 2. First, we note that by scaling the loss function ℓ the value of σ_ℓ does not change. In particular, if $\widetilde{\ell}(t) := C_1\ell(C_2 t)$ for arbitrary constants $C_1 > 0, C_2 \neq 0$, it is not hard to see that $\widehat{\mathbf{x}}_{\widetilde{\ell}} = 1/C_2\,\widehat{\mathbf{x}}_\ell$ is the minimizer of (3). Thus, we conclude from (40) that $\sigma_{\widetilde{\ell}} = \sigma_\ell$. With this observation, consider the function $\widetilde{\ell}:\mathbb{R}\to\mathbb{R}$ such that $\widetilde{\ell}(t) = \frac{\lambda}{\mu^2}\ell(\mu\,t)$. Then, notice that

$$\mathcal{M}'_{\ell,1}(x;\lambda) = \frac{1}{\lambda}\mathcal{M}'_{\widetilde{\ell},1}(x/\mu;1).$$

Using this relation in (38) and setting $\sigma := \sigma_\ell = \alpha/\mu$, the system of equations in (38) can be equivalently rewritten in the following convenient form, where $Z_\sigma = \sigma W_1 + W_2$:

$$\mathbb{E}\left[W_2\cdot\mathcal{M}'_{\widetilde{\ell},1}(Z_\sigma;1)\right] = 0, \tag{A73a}$$

$$\mathbb{E}\left[\left(\mathcal{M}'_{\widetilde{\ell},1}(Z_\sigma;1)\right)^2\right] = \sigma^2/\delta\,, \tag{A73b}$$

$$\mathbb{E}\left[W_1\cdot\mathcal{M}'_{\widetilde{\ell},1}(Z_\sigma;1)\right] = \sigma/\delta\,. \tag{A73c}$$

Next, we show how to use (A73) to derive an equivalent system of equations in terms of only Z_σ. Starting with (A73c), we have

$$\mathbb{E}\left[W_1\cdot\mathcal{M}'_{\widetilde{\ell},1}(Z_\sigma;1)\right] = \frac{1}{\sigma}\iint x\,\mathcal{M}'_{\widetilde{\ell},1}(x+y;1)\,p_{\sigma W_1}(x)\,p_{W_2}(y)\mathrm{d}x\mathrm{d}y, \tag{A74}$$

where recall that $p_{\sigma W_1}(x) = \frac{1}{\sigma\sqrt{2\pi}}e^{-\frac{x^2}{2\sigma^2}}$. Since it holds that $p_{\sigma W_1}(x) = \frac{-\sigma^2}{x}p'_{\sigma W_1}(x)$, using (A74) yields that

$$\begin{aligned}\mathbb{E}\left[W_1\cdot\mathcal{M}'_{\widetilde{\ell},1}(Z_\sigma;1)\right] &= -\sigma\iint\mathcal{M}'_{\widetilde{\ell},1}(x+y;1)\,p'_{\sigma W_1}(x)\,p_{W_2}(y)\mathrm{d}x\mathrm{d}y\\ &= -\sigma\iint\mathcal{M}'_{\widetilde{\ell},1}(z;1)\,p'_{\sigma W_1}(x)\,p_{W_2}(z-x)\mathrm{d}x\mathrm{d}z = -\sigma\int\mathcal{M}'_{\widetilde{\ell},1}(z;1)\,p'_{Z_\sigma}(z)\mathrm{d}z,\end{aligned}$$

where in the last step we use

$$p'_{Z_\sigma}(w) = \int p'_{\sigma W_1}(x) p_{W_2}(z-x)\,\mathrm{d}x.$$

Therefore,

$$\mathbb{E}\Big[W_1 \cdot \mathcal{M}'_{\widetilde{\ell},1}(Z_\sigma;1)\Big] = -\sigma\,\mathbb{E}\Big[\mathcal{M}'_{\widetilde{\ell},1}(Z_\sigma;1)\xi_{Z_\sigma}(Z_\sigma)\Big].$$

This combined with (A73c) gives

$$\mathbb{E}\Big[\mathcal{M}'_{\widetilde{\ell},1}(Z_\sigma;1)\xi_{Z_\sigma}(Z_\sigma)\Big] = -1/\delta.$$

Second, multiplying (A73c) with σ^2 and adding it to (A73a) yields

$$\mathbb{E}\Big[Z_\sigma \cdot \mathcal{M}'_{\widetilde{\ell},1}(Z_\sigma;1)\Big] = \sigma^2/\delta. \tag{A75}$$

Putting these together, we conclude with the following system of equations which is equivalent to (A73),

$$\mathbb{E}\Big[Z_\sigma \cdot \mathcal{M}'_{\widetilde{\ell},1}(Z_\sigma;1)\Big] = \sigma^2/\delta, \tag{A76a}$$

$$\mathbb{E}\Big[\Big(\mathcal{M}'_{\widetilde{\ell},1}(Z_\sigma;1)\Big)^2\Big] = \sigma^2/\delta, \tag{A76b}$$

$$\mathbb{E}\Big[\mathcal{M}'_{\widetilde{\ell},1}(Z_\sigma;1)\xi_{Z_\sigma}(Z_\sigma)\Big] = -1/\delta. \tag{A76c}$$

Next, considering (A76a) and (A76c), the following holds for any $c_1, c_2 \in \mathbb{R}$,

$$\mathbb{E}\Big[(c_1 Z_\sigma + c_2\,\xi_{Z_\sigma}(Z_\sigma)) \cdot \mathcal{M}'_{\widetilde{\ell},1}(Z_\sigma;1)\Big] = c_1\sigma^2/\delta - c_2/\delta. \tag{A77}$$

Applying Cauchy–Schwarz inequality to the LHS of (A77) gives

$$\begin{aligned}\Big(c_1\sigma^2/\delta - c_2/\delta\Big)^2 &= \Big(\mathbb{E}\Big[\Big(c_1 Z_\sigma + c_2\,\xi_{Z_\sigma}(Z_\sigma)\Big) \cdot \mathcal{M}'_{\widetilde{\ell},1}(Z_\sigma;1)\Big]\Big)^2 \\ &\le \mathbb{E}\Big[\Big(c_1 Z_\sigma + c_2\,\xi_{Z_\sigma}(Z_\sigma)\Big)^2\Big]\mathbb{E}\Big[\Big(\mathcal{M}'_{\widetilde{\ell},1}(Z_\sigma;1)\Big)^2\Big].\end{aligned} \tag{A78}$$

By considering (A76b), $\mathbb{E}[Z_\sigma \xi_{Z_\sigma}(Z_\sigma)] = -1$ (follows from integration by parts) and $\mathbb{E}\big[(\xi_{Z_\sigma}(Z_\sigma))^2\big] = \mathcal{I}(Z_\sigma) = (\sigma^2+1)^{-1}$, we simplify (A78) to the following:

$$(c_1\sigma^2/\delta - c_2/\delta)^2 \le \Big(c_1^2(\sigma^2+1+r^2) + c_2^2/(\sigma^2+1) - 2c_1c_2\Big)\sigma^2/\delta.$$

Choosing $c_1 = 1$ and $c_2 = (1+r^2)(1+\sigma^2)$ and simplifying both sides, we derive the lower bound for σ^2:

$$\sigma^2 \ge \frac{1+r^2}{r^2} \cdot \frac{1}{(\delta-1)}.$$

This completes the proof of theorem.

References

1. Donoho, D.L. High-dimensional data analysis: The curses and blessings of dimensionality. *AMS Math. Chall. Lect.* **2000**, *1*, 32.
2. Donoho, D.L. Compressed sensing. *Inf. Theory IEEE Trans.* **2006**, *52*, 1289–1306. [CrossRef]
3. Stojnic, M. Various thresholds for ℓ_1-optimization in compressed sensing. *arXiv* **2009**, arXiv:0907.3666.

4. Chandrasekaran, V.; Recht, B.; Parrilo, P.A.; Willsky, A.S. The convex geometry of linear inverse problems. *Found. Comput. Math.* **2012**, *12*, 805–849. [CrossRef]
5. Donoho, D.L.; Maleki, A.; Montanari, A. The noise-sensitivity phase transition in compressed sensing. *Inf. Theory IEEE Trans.* **2011**, *57*, 6920–6941. [CrossRef]
6. Tropp, J.A. Convex recovery of a structured signal from independent random linear measurements. *arXiv* **2014**, arXiv:1405.1102.
7. Oymak, S.; Tropp, J.A. Universality laws for randomized dimension reduction, with applications. *Inf. Inference J. IMA* **2017**, *7*, 337–446. [CrossRef]
8. Bayati, M.; Montanari, A. The LASSO risk for gaussian matrices. *Inf. Theory IEEE Trans.* **2012**, *58*, 1997–2017. [CrossRef]
9. Stojnic, M. A framework to characterize performance of LASSO algorithms. *arXiv* **2013**, arXiv:1303.7291.
10. Oymak, S.; Thrampoulidis, C.; Hassibi, B. The Squared-Error of Generalized LASSO: A Precise Analysis. *arXiv* **2013**, arXiv:1311.0830.
11. Karoui, N.E. Asymptotic behavior of unregularized and ridge-regularized high-dimensional robust regression estimators: Rigorous results. *arXiv* **2013**, arXiv:1311.2445.
12. Bean, D.; Bickel, P.J.; El Karoui, N.; Yu, B. Optimal M-estimation in high-dimensional regression. *Proc. Natl. Acad. Sci. USA* **2013**, *110*, 14563–14568. [CrossRef]
13. Thrampoulidis, C.; Oymak, S.; Hassibi, B. Regularized Linear Regression: A Precise Analysis of the Estimation Error. In Proceedings of the 28th Conference on Learning Theory, Paris, France, 3–6 July 2015; pp. 1683–1709.
14. Donoho, D.; Montanari, A. High dimensional robust m-estimation: Asymptotic variance via approximate message passing. *Probab. Theory Relat. Fields* **2016**, *166*, 935–969. [CrossRef]
15. Thrampoulidis, C.; Abbasi, E.; Hassibi, B. Precise Error Analysis of Regularized *M*-Estimators in High Dimensions. *IEEE Trans. Inf. Theory* **2018**, *64*, 5592–5628. [CrossRef]
16. Advani, M.; Ganguli, S. Statistical mechanics of optimal convex inference in high dimensions. *Phys. Rev. X* **2016**, *6*, 031034. [CrossRef]
17. Weng, H.; Maleki, A.; Zheng, L. Overcoming the limitations of phase transition by higher order analysis of regularization techniques. *Ann. Stat.* **2018**, *46*, 3099–3129. [CrossRef]
18. Thrampoulidis, C.; Xu, W.; Hassibi, B. Symbol Error Rate Performance of Box-relaxation Decoders in Massive MIMO. *IEEE Trans. Signal Process.* **2018**, *66*, 3377–3392. [CrossRef]
19. Miolane, L.; Montanari, A. The distribution of the Lasso: Uniform control over sparse balls and adaptive parameter tuning. *arXiv* **2018**, arXiv:1811.01212.
20. Bu, Z.; Klusowski, J.; Rush, C.; Su, W. Algorithmic analysis and statistical estimation of slope via approximate message passing. In Proceedings of the Advances in Neural Information Processing Systems, Vancouver, BD, Canada, 8–14 December 2019; pp. 9361–9371.
21. Xu, J.; Maleki, A.; Rad, K.R.; Hsu, D. Consistent risk estimation in high-dimensional linear regression. *arXiv* **2019**, arXiv:1902.01753.
22. Celentano, M.; Montanari, A. Fundamental Barriers to High-Dimensional Regression with Convex Penalties. *arXiv* **2019**, arXiv:1903.10603.
23. Kammoun, A.; Alouini, M.S. On the precise error analysis of support vector machines. *arXiv* **2020**, arXiv:2003.12972.
24. Amelunxen, D.; Lotz, M.; McCoy, M.B.; Tropp, J.A. Living on the edge: A geometric theory of phase transitions in convex optimization. *arXiv* **2013**, arXiv:1303.6672.
25. Donoho, D.L.; Johnstone, L.; Montanari, A. Accurate Prediction of Phase Transitions in Compressed Sensing via a Connection to Minimax Denoising. *IEEE Trans. Inf. Theory* **2013**, *59*, 3396–3433. [CrossRef]
26. Mondelli, M.; Montanari, A. Fundamental limits of weak recovery with applications to phase retrieval. *arXiv* **2017**, arXiv:1708.05932.
27. Taheri, H.; Pedarsani, R.; Thrampoulidis, C. Fundamental limits of ridge-regularized empirical risk minimization in high dimensions. *arXiv* **2020**, arXiv:2006.08917.
28. Bayati, M.; Lelarge, M.; Montanari, A. Universality in polytope phase transitions and message passing algorithms. *Ann. Appl. Probab.* **2015**, *25*, 753–822. [CrossRef]
29. Panahi, A.; Hassibi, B. A universal analysis of large-scale regularized least squares solutions. In Proceedings of the Advances in Neural Information Processing Systems, Long Beach, CA, USA, 4–9 December 2017; pp. 3381–3390.
30. Abbasi, E.; Salehi, F.; Hassibi, B. Universality in learning from linear measurements. In Proceedings of the Advances in Neural Information Processing Systems, Vancouver, BD, Canada, 8–14 December 2019; pp. 12372–12382.
31. Goldt, S.; Reeves, G.; Mézard, M.; Krzakala, F.; Zdeborová, L. The Gaussian equivalence of generative models for learning with two-layer neural networks. *arXiv* **2020**, arXiv:2006.14709.
32. Donoho, D.L.; Maleki, A.; Montanari, A. Message-passing algorithms for compressed sensing. *Proc. Natl. Acad. Sci. USA* **2009**, *106*, 18914–18919. [CrossRef]
33. Bayati, M.; Montanari, A. The dynamics of message passing on dense graphs, with applications to compressed sensing. *Inf. Theory IEEE Trans.* **2011**, *57*, 764–785. [CrossRef]
34. Mousavi, A.; Maleki, A.; Baraniuk, R.G. Consistent parameter estimation for LASSO and approximate message passing. *Ann. Stat.* **2018**, *46*, 119–148. [CrossRef]

35. El Karoui, N. On the impact of predictor geometry on the performance on high-dimensional ridge-regularized generalized robust regression estimators. *Probab. Theory Relat. Fields* **2018**, *170*, 95–175. [CrossRef]
36. Boufounos, P.T.; Baraniuk, R.G. 1-bit compressive sensing. In Proceedings of the 2008 IEEE 42nd Annual Conference on Information Sciences and Systems (CISS), Princeton, NJ, USA, 19–21 March 2008, pp. 16–21.
37. Jacques, L.; Laska, J.N.; Boufounos, P.T.; Baraniuk, R.G. Robust 1-bit compressive sensing via binary stable embeddings of sparse vectors. *IEEE Trans. Inf. Theory* **2013**, *59*, 2082–2102. [CrossRef]
38. Plan, Y.; Vershynin, R. One-Bit Compressed Sensing by Linear Programming. *Commun. Pure Appl. Math.* **2013**, *66*, 1275–1297. [CrossRef]
39. Plan, Y.; Vershynin, R. Robust 1-bit compressed sensing and sparse logistic regression: A convex programming approach. *IEEE Trans. Inf. Theory* **2012**, *59*, 482–494. [CrossRef]
40. Plan, Y.; Vershynin, R. The generalized lasso with non-linear observations. *IEEE Trans. Inf. Theory* **2016**, *62*, 1528–1537. [CrossRef]
41. Genzel, M. High-dimensional estimation of structured signals from non-linear observations with general convex loss functions. *IEEE Trans. Inf. Theory* **2017**, *63*, 1601–1619. [CrossRef]
42. Xu, C.; Jacques, L. Quantized compressive sensing with rip matrices: The benefit of dithering. *arXiv* **2018**, arXiv:1801.05870.
43. Thrampoulidis, C.; Abbasi, E.; Hassibi, B. Lasso with non-linear measurements is equivalent to one with linear measurements. In Proceedings of the Advances in Neural Information Processing Systems,Montreal, QC, Canada, 7–12 December 2015; pp. 3420–3428.
44. Candès, E.J.; Sur, P. The phase transition for the existence of the maximum likelihood estimate in high-dimensional logistic regression. *arXiv* **2018**, arXiv:1804.09753.
45. Sur, P.; Candès, E.J. A modern maximum-likelihood theory for high-dimensional logistic regression. *Proc. Natl. Acad. Sci. USA* **2019**, 201810420. [CrossRef]
46. Mai, X.; Liao, Z.; Couillet, R. A Large Scale Analysis of Logistic Regression: Asymptotic Performance and New Insights. In Proceedings of the ICASSP 2019–2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Brighton, UK, 12–17 May 2019; pp. 3357–3361.
47. Salehi, F.; Abbasi, E.; Hassibi, B. The Impact of Regularization on High-dimensional Logistic Regression. *arXiv* **2019**, arXiv:1906.03761.
48. Montanari, A.; Ruan, F.; Sohn, Y.; Yan, J. The generalization error of max-margin linear classifiers: High-dimensional asymptotics in the overparametrized regime. *arXiv* **2019**, arXiv:1911.01544.
49. Deng, Z.; Kammoun, A.; Thrampoulidis, C. A Model of Double Descent for High-dimensional Binary Linear Classification. *arXiv* **2019**, arXiv:1911.05822.
50. Mignacco, F.; Krzakala, F.; Lu, Y.M.; Zdeborová, L. The role of regularization in classification of high-dimensional noisy Gaussian mixture. *arXiv* **2020**, arXiv:2002.11544.
51. Aubin, B.; Krzakala, F.; Lu, Y.M.; Zdeborová, L. Generalization error in high-dimensional perceptrons: Approaching Bayes error with convex optimization. *arXiv* **2020**, arXiv:2006.06560.
52. Rockafellar, R.T.; Wets, R.J.B. *Variational Analysis*; Springer Science & Business Media: Berlin/Heidelberg, Germany, 2009; Volume 317.
53. Barron, A.R. *Monotonic Central Limit Theorem for Densities*; Technical Report; Stanford University: Stanford, CA , USA, 1984.
54. Costa, M.H.M. A new entropy power inequality. *IEEE Trans. Inf. Theory* **1985**, *31*, 751–760. [CrossRef]
55. Blachman, N. The convolution inequality for entropy powers. *IEEE Trans. Inf. Theory* **1965**, *11*, 267–271. [CrossRef]
56. Brillinger, D.R. A Generalized Linear Model with “Gaussian" Regressor Variables. In *A Festschrift For Erich L. Lehmann*; Springer: New York, NY, USA, 1982; p. 97.
57. Dhifallah, O.; Lu, Y.M. A precise performance analysis of learning with random features. *arXiv* **2020**, arXiv:2008.11904.
58. Boyd, S.; Vandenberghe, L. *Convex Optimization*; Cambridge University Press: Cambridge, UK, 2009.
59. Gordon, Y. *On Milman's Inequality and Random Subspaces which Escape through a Mesh in* $\mathbb{R}^n$; Springer: Berlin/Heidelberg, Germany, 1988.
60. Dhifallah, O.; Thrampoulidis, C.; Lu, Y.M. Phase retrieval via polytope optimization: Geometry, phase transitions, and new algorithms. *arXiv* **2018**, arXiv:1805.09555.
61. Rockafellar, R.T. *Convex Analysis*; Princeton University Press: Princeton, NJ, USA, 1970.
62. Genzel, M.; Jung, P. Recovering structured data from superimposed non-linear measurements. *arXiv* **2017**, arXiv:1708.07451.
63. Goldstein, L.; Minsker, S.; Wei, X. Structured signal recovery from non-linear and heavy-tailed measurements. *IEEE Trans. Inf. Theory* **2018**, *64*, 5513–5530. [CrossRef]
64. Thrampoulidis, C.; Rawat, A.S. The generalized lasso for sub-gaussian measurements with dithered quantization. *arXiv* **2018**, arXiv:1807.06976.