Fundamental Limits of Ridge-Regularized Empirical Risk Minimization in High Dimensions

Hossein Taheri
UC Santa Barbara

Ramtin Pedarsani
UC Santa Barbara

Christos Thrampoulidis
University of British Columbia, UC Santa Barbara

Abstract

Despite the popularity of Empirical Risk Minimization (ERM) algorithms, a theory that explains their statistical properties in modern high-dimensional regimes is only recently emerging. We characterize for the first time the fundamental limits on the statistical accuracy of convex ridge-regularized ERM for inference in high-dimensional generalized linear models. For a stylized setting with Gaussian features and problem dimensions that grow large at a proportional rate, we start with sharp performance characterizations and then derive tight lower bounds on the estimation and prediction error. Our bounds provably hold over a wide class of loss functions, and, for any value of the regularization parameter and of the sampling ratio. Our precise analysis has several attributes. First, it leads to a recipe for optimally tuning the loss function and the regularization parameter. Second, it allows to precisely quantify the suboptimality of popular heuristic choices, such as optimally-tuned least-squares. Third, we use the bounds to precisely assess the merits of ridge-regularization as a function of the sampling ratio. Our bounds are expressed in terms of the Fisher Information of random variables that are simple functions of the data distribution, thus making ties to corresponding bounds in classical statistics.

Proceedings of the 24[th] International Conference on Artificial Intelligence and Statistics (AISTATS) 2021, San Diego, California, USA. PMLR: Volume 130.

1 Introduction

Motivation. Empirical Risk Minimization (ERM) includes statistical inference algorithms that are popular in estimation and learning tasks in a range of applications in signal processing, communications and machine learning. ERM methods are often efficient in implementation, but first one needs to make certain choices: such as, choose an appropriate loss function and regularization function, and tune the regularization parameter. Classical statistics have complemented the practice of ERM with an elegant theory regarding optimal such choices, as well as, fundamental limits, i.e., tight bounds on their performance (e.g., [Huber, 2011]). These classical theories typically assume that the size m of the set of observations (or, training set) is much larger than the dimension n of the unknown parameter-vector to be estimated. In contrast, modern inference problems are typically high-dimensional: m and n are of the same order, and, often $n > m$ [Candès, 2014, Montanari, 2015, Karoui, 2013]. This paper studies the fundamental limits of convex ERM in high-dimensions for generalized linear models. Generalized linear models (GLM) relate the response variable y_i to a linear model $\mathbf{a}_i^T\mathbf{x}_0$ via a link function: $y_i = \varphi(\mathbf{a}_i^T\mathbf{x}_0)$. Here, $\mathbf{x}_0 \in \mathbb{R}^n$ is a vector of true parameters and $\mathbf{a}_i \in \mathbb{R}^n$, $i \in [m]$ are the feature (or, measurement) vectors. Let $\mathbf{x}_0$ be estimated by minimizing the empirical risk $\frac{1}{m}\sum_{i=1}^m \mathcal{L}\left(y_i, \mathbf{a}_i^T\mathbf{x}\right)$ for a particular *convex* loss $\mathcal{L}$. Typically, ERM is combined with a regularization term. Arguably the most popular choice is ridge regularization, which gives rise to ridge-regularized ERM (RERM, in short):

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

This paper aims to answer the following questions on fundamental limits of (1): *What is the minimum achievable estimation/prediction error of* $\widehat{\mathbf{x}}_{\mathcal{L},\lambda}$*? How*

does this depend on the link function φ and how to choose $\mathcal{L}$ and λ to achieve it? What is the sub-optimality gap of popular ad-hoc choices, such as ridge-regularized least-squares (RLS)? How do the answers depend on the sampling ratio m/n?

Challenge. The challenge of answering the questions above involves completing the following three key tasks. The first prerequisite task is to: **[T1]** obtain a *precise* characterization of the estimation/prediction error of $\widehat{\mathbf{x}}_{\mathcal{L},\lambda}$ as a function of the parameters $\mathcal{L}$, λ and the dimensions m and n. Significant research activity over the past decade has led to novel analysis frameworks making this possible, typically, in a stylized setting of Gaussian features and an asymptotic regime, where m and n grow large at a proportional rate $\delta = m/n$ [Montanari, 2015, Karoui, 2013, Sur and Candès, 2019]. The analysis leads to error characterizations in terms of solutions to appropriate systems of (a few) nonlinear equations. Naturally, the equations are parameterized by $\mathcal{L}$, λ and δ. For different choices of these parameters, numerically solving the equations leads to *precise* asymptotic error characterizations. But, questions on fundamental limits such as "what is the optimal loss $\mathcal{L}$ and regularizer λ?", ask us to take a step further. They require determining the choice of $\mathcal{L}, \lambda$ that leads to a solution to the equations that, in turn, implies minimum error, for a given δ. There are two additional tasks involved in accomplishing this. First, to allow optimizing over $\mathcal{L}, \lambda$ we need to: **[T2]** prove that the system of equations is valid for a rich family of losses $\mathcal{L}$ and every value $\lambda > 0$. Second, since the solution to the equations (thus, the asymptotic error) is *not* an explicit function of the parameters of interest $\mathcal{L}, \lambda$, we need a mechanism to: **[T3]** minimize the solution to the system of equations over $\mathcal{L}, \lambda$.

Contributions. This paper, for the first time accomplishes tasks **[T2,T3]**, for two popular GLM-instances, namely linear and binary models, and, for a stylized distributional setting of isotropic Gaussian features. With this we establish the promised fundamental performance limits and answer corresponding optimality questions.

- For linear models, we prove a lower bound on the squared estimation error of RERM (see Thm. 2.1) that holds for all choices of $\mathcal{L}, \lambda > 0$ and $\delta > 0$. Specifically, our contribution involves accomplishing Task **[T3]**; Tasks **[T1,T2]** were investigated in [Karoui, 2013, Thrampoulidis et al., 2018]. Our analysis, leads to explicit expressions for the optimal loss $\mathcal{L}_\star$ and regularizer parameter $\lambda_\star$. Additionally, we present analytic conditions on the noise-distribution and δ, for which $\mathcal{L}_\star$ is convex.

- For binary models, we fulfill the promise of performance lower bounds by completing both Tasks **[T2]** (see Thm. 3.1) and Task **[T3]** (see Thm. 3.2). As in linear models, we present explicit recipes for optimally tuning $\mathcal{L}$ and λ. For specific models, such as binary logistic and signed data, we numerically show that the optimal loss function is convex and we use gradient-descent to optimize it. The numerical simulations perfectly match with the theoretical predictions suggesting that our bounds are tight.

- We derive simple closed-form approximations to the aforementioned bounds (see Cor. 2.1 and 3.1). These simple (yet tight) expressions allow us to precisely quantify the sub-optimality of ridge-regularized least-squares (RLS). For instance, we show that optimally-tuned RLS is approximately optimal for logistic data and small signal strength, but the sub-optimality gap grows drastically as signal strength increases. In the appendix, we also include comparisons to ERM without regularization and to a simple averaging method.

Comparison to state-of-the-art. Our results fit in the rapidly growing recent literature on *precise* asymptotics of convex-regularized estimators, e.g., [Donoho et al., 2011, Stojnic, 2009, Bayati and Montanari, 2012, Chandrasekaran et al., 2012, Amelunxen et al., 2013, Oymak and Hassibi, 2016, Abbasi et al., 2016, Stojnic, 2013, Oymak et al., 2013, Thrampoulidis et al., 2015b, Karoui, 2013, Donoho and Montanari, 2016, El Karoui, 2018, Thrampoulidis et al., 2018, Oymak and Tropp, 2017, Dobriban et al., 2018, Lei et al., 2018, Miolane and Montanari, 2018, Hastie et al., 2019, Wang et al., 2019, Celentano and Montanari, 2019, Hu and Lu, 2019, Bu et al., 2019, Emami et al., 2020, Lolas, 2020, Kini and Thrampoulidis, 2020, Gerbelot et al., 2020]. Most of these works study linear models. Extensions to generalized linear models for the special case of regularized LS were studied in [Thrampoulidis et al., 2015a], while more recently there has been a surge of interest in (R)ERM methods tailored to binary models (such as logistic regression or SVM) [Huang, 2017, Candès and Sur, 2018, Sur and Candès, 2019, Mai et al., 2019, Kammoun and Alouini, 2020, Salehi et al., 2019, Taheri et al., 2020, Deng et al., 2019, Montanari et al., 2019, Mignacco et al., 2020, Emami et al., 2020, Salehi et al., 2020]. The focus of these works has been Task **[T1]**. Out of these

works relatively few have focused on fundamental limits, which requires accomplishing the additional tasks **[T2]** and **[T3]**. For linear models, the papers [Bean et al., 2013, Donoho and Montanari, 2016, Advani and Ganguli, 2016] were the first to derive lower bounds and optimal loss functions for the squared error of *unregularized* ERM. In a related work, [Donoho and Montanari, 2015] studies noise-robustness of these methods. More recently, [Celentano and Montanari, 2019] performed an in-depth analysis of fundamental limits of convex-regularized least-squares for linear models over structured (e.g., sparse, low-rank) signals. For binary models, performance lower bounds for *unregularized* ERM were only recently derived in [Taheri et al., 2020].

To the best of our knowledge, none of these prior works has established fundamental limits for *ridge-regularized* ERM, for either linear or binary models. Accounting for the regularization term brings the following technical challenges. First, to accomplish Task **[T2]**, we prove that a solution to the corresponding system of equations exists and is unique for all values of $\delta > 0$, and, only under mild assumptions on $\mathcal{L}$. For binary models, this is the first proof of both existence and uniqueness compared to prior works [Sur and Candès, 2019, Salehi et al., 2019, Taheri et al., 2020, Mignacco et al., 2020]. Second, the presence of the regularizer complicates Task **[T3]**. Compared to the unregularized case, we need to optimize not only over $\mathcal{L}$, but also over $\lambda > 0$. More elaborate discussions on technical comparisons of our results to prior work are deferred till after the formal statement of our results.

1.1 Dataset model

Linear models: $y_i = \mathbf{a}_i^T \mathbf{x}_0 + z_i$, where $z_i \overset{\text{iid}}{\sim} P_Z$, $i \in [m]$. As is typical, for linear models, we measure performance of $\widehat{\mathbf{x}}_{\mathcal{L},\lambda}$ with the *squared error*: $\|\widehat{\mathbf{x}}_{\mathcal{L},\lambda} - \mathbf{x}_0\|_2^2$.

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

Our precise analysis requires isotropic Gaussian features and a proportional asymptotic regime, as follows

Assumption 1 (High-dimensional asymptotics)**.** *Throughout the paper, we assume the high-dimensional limit where $m, n \to \infty$ at a fixed ratio $\delta = m/n > 0$.*

Assumption 2 (Gaussian features)**.** *The feature vectors $\mathbf{a}_i \in \mathbb{R}^n, i \in [m]$ are iid $\mathcal{N}(\mathbf{0}, \mathbf{I}_n)$.*

This set of assumption is well-adapted in the recent literature on precise high-dimensional statistics. Specifically regarding the Gaussianity assumption, it is an essential first step in the vast majority of existing analyses targeting Task **[T1]** (e.g., [Montanari, 2015, Karoui, 2013, Sur and Candès, 2019] and many references therein). Besides, extensive numerical simulations and partial theoretical evidence [Bayati et al., 2015, Oymak and Tropp, 2017, Panahi and Hassibi, 2017, Abbasi et al., 2019, Goldt et al., 2020] seem to suggest that the systems of equations characterizing the error enjoy a remarkable *universality* property: they hold for a broader class of distributions, e.g., sub-gaussians. All the results of this paper on fundamental performance limits and optimality automatically hold for any feature distribution that leads to the same asymptotic error characterizations as the Gaussian distribution. A formal proof of universality of our results is beyond our scope. However, we present numerical experiments in support of this conjecture; see Figure 1.

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

2 Linear Models

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

Assumption 3 (Noise distribution)**.** *The noise $z_i, i \in [m]$ is distributed $Z \overset{iid}{\sim} P_Z$, for a distribution P_Z with zero mean and finite nonzero second moment.*

For lower semicontinuous, proper, and convex losses we focus on an instance of (1) tailored to linear models:

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

We assume without loss of generality that $\|\mathbf{x}_0\|_2 = 1$. Indeed, suppose that $\|\mathbf{x}_0\|_2 = r > 0$. Then, (2) can be transformed to the case $\widetilde{\mathbf{x}}_0 := \mathbf{x}_0/r$ (hence $\|\widetilde{\mathbf{x}}_0\|_2 = 1$) by setting $\widetilde{\mathcal{L}}(t) := \mathcal{L}(rt)$, $\widetilde{\lambda} := r^2\lambda$ and $\widetilde{Z} = Z/r$. Thus our results can be reformulated by replacing Z with $\widetilde{Z}$.

2.1 Background on asymptotic performance

Prior works have investigated the limit of the squared error $\|\widehat{\mathbf{x}}_{\mathcal{L},\lambda} - \mathbf{x}_0\|^2$ [Karoui, 2013, Thrampoulidis et al., 2018, Donoho and Montanari, 2016]. Let the following system of two equations in two unknowns α and τ:

$$\mathbb{E}\Big[\Big(\mathcal{M}'_{\mathcal{L},1}\left(\alpha\, G + Z; \tau\right)\Big)^2\Big] = \frac{\alpha^2 - \lambda^2\delta^2\tau^2}{\tau^2\,\delta}, \tag{3a}$$

$$\mathbb{E}\Big[G \cdot \mathcal{M}'_{\mathcal{L},1}\left(\alpha\, G + Z; \tau\right)\Big] = \frac{\alpha(1-\lambda\delta\tau)}{\tau\,\delta}, \tag{3b}$$

where $G \sim \mathcal{N}(0,1)$ and $Z \sim P_Z$. It has been shown in [Karoui, 2013, Thrampoulidis et al., 2018] that under appropriate regularity conditions on $\mathcal{L}$ and the noise distribution P_Z, (cf. Tasks **[T1,T2]**) the system of equations above has a unique solution $(\alpha_{\mathcal{L},\lambda} > 0, \tau_{\mathcal{L},\lambda} > 0)$ and $\alpha^2_{\mathcal{L},\lambda}$ is the limit of the squared-error, i.e.,

$$\|\widehat{\mathbf{x}}_{\mathcal{L},\lambda} - \mathbf{x}_0\|_2^2 \xrightarrow{P} \alpha^2_{\mathcal{L},\lambda}. \tag{4}$$

Using this, we derive tight lower bounds on $\alpha^2_{\mathcal{L},\lambda}$ over the choices of $\mathcal{L}$ and λ (cf. Task **[T3]**). Our results hold for all losses and regularizer parameters for which (3) has a unique solution characterizing the limit of the squared-error. To formalize this, define the following collection of losses $\mathcal{L}$ and noise distributions P_Z:

$$\mathcal{C}_{\text{lin}} := \Big\{(\mathcal{L}, P_Z)\,\Big|\,\forall\lambda > 0\text{: (3) has a unique bounded solution } (\alpha_{\mathcal{L},\lambda} > 0, \tau_{\mathcal{L},\lambda} > 0) \text{ and (4) holds}\Big\}.$$

Please refer to [Karoui, 2013, Thm. 1.1] and [Thrampoulidis et al., 2018, Thm. 2] for explicit characterizations of $\mathcal{C}_{\text{lin}}$. We conjecture that some of these regularity conditions (e.g., the differentiability requirement) can in fact be relaxed. While this is beyond the scope of this paper, if this is shown then automatically the results of this paper formally hold for a richer class of loss functions.

2.2 Fundamental Limits and Optimal Tuning

Our first main result, stated as Theorem 2.1 below, establishes a tight bound on the achievable values of $\alpha^2_{\mathcal{L},\lambda}$ for all pairs $(\mathcal{L}, P_Z) \in \mathcal{C}_{\text{lin}}$.

Theorem 2.1 (Lower bound on $\alpha_{\mathcal{L},\lambda}$)**.** *Let Assumptions 1, 2 and 3 hold. For $G \sim \mathcal{N}(0,1)$ and noise random variable $Z \sim P_Z$, consider a new random variable $V_a := a\,G + Z$, parameterized by $a \in \mathbb{R}$. Fix any $\delta > 0$ and define $\alpha_\star = \alpha_\star(\delta, P_Z)$ as follows:*

$$\alpha_\star := \min_{0 \le x < 1/\delta} \left[a > 0 :\ \frac{\delta(a^2 - x^2\,\delta^2)\,\mathcal{I}(V_a)}{(1 - x\,\delta)^2} = 1\right]. \tag{5}$$

For any $\mathcal{L}$ such that $(\mathcal{L}, P_Z) \in \mathcal{C}_{\text{lin}}$, $\lambda > 0$ and $\alpha^2_{\mathcal{L},\lambda}$ denoting the respective high-dimensional limit of the squared-error as in (4), it holds that $\alpha_{\mathcal{L},\lambda} \ge \alpha_\star$.

The proof is given to Section C.2. It includes showing feasibility of the minimization in (5) for any $\delta > 0$.

In general, the lower bound $\alpha_\star$ can be computed by numerically solving (5). For special cases, such as Gaussian noise, it is possible to analytically solve (5) and obtain a closed-form formula for $\alpha_\star$, which is easier to interpret. Because this is *not* always possible, our next result establishes a simple closed-form lower bound on $\alpha_\star$ that is valid under only mild assumptions on P_Z. For convenience, let us define $h_\delta : \mathbb{R}_{>0} \to \mathbb{R}_{>0}$,

$$h_\delta(x) := \frac{1}{2}\Big(1 - x - \delta + \sqrt{(1+\delta+x)^2 - 4\delta}\Big). \tag{6}$$

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

Corollary 2.1 (Closed-form lower bound on $\alpha_\star^2$)**.** *Let $\alpha_\star$ be as in (5) under the assumptions of Theorem 2.1. Assume that P_Z is differentiable and takes strictly positive values on the real line. Then, it holds that*

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

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

The proof, presented in Section C.5, shows that the gap between the actual value of $\alpha_\star$ and $h_\delta(1/\mathcal{I}(Z))$ depends solely on the distribution of Z. Informally, the more Z resembles a Gaussian, the smaller the gap. The simple approximation of Corollary 2.1 is key for comparing the performance of optimally tuned RERM to optimally-tuned RLS in Section 2.3.Moreover, it can be used to show that the lower bound of Theorem 2.1 cannot be improved in general. This can be argued as follows. Consider additive Gaussian noise $Z \sim \mathcal{N}(0, \zeta^2)$ for which $\mathcal{I}(Z) = 1/\mathbb{E}[Z^2] = 1/\zeta^2$. On the one hand, Corollary 2.1 shows that $\alpha_\star^2 \ge h_\delta(\zeta^2)$. On the other hand, we will soon show in Lemma 2.2 that optimally-tuned RLS achieves this bound, i.e., $\alpha^2_{\ell_2,\lambda_{\text{opt}}} = h_\delta(\zeta^2)$. Thus, the case of Gaussian noise shows that the bound of Theorem 2.1 cannot be improved in general.

Our next result reinforces the claim that the bound of

Theorem 2.1 is indeed tight for a broad class of noise distributions. Specifically, the lemma below delivers an explicit recipe for optimally choosing the loss and the regularizer parameter, as well as, sufficient conditions under which the optimal loss is convex $\mathcal{L}_\star$. Note that both $\mathcal{L}_\star$ and $\lambda_\star$ depend on the sampling ratio δ.

Lemma 2.1 (Optimal tuning of RERM). *For given $\delta > 0$ and P_Z, let $(\alpha_\star > 0, x_\star \in [0, 1/\delta))$ be the optimal solution in the minimization in (5). Denote $\lambda_\star = x_\star$ and define $V_\star := \alpha_\star G + Z$. Consider the loss function $\mathcal{L}_\star : \mathbb{R} \to \mathbb{R}$ defined as*

$$\mathcal{L}_\star(v) := -\mathcal{M}_{\frac{\alpha_\star^2 - \lambda_\star^2\,\delta^2}{1-\lambda_\star\,\delta}\cdot\log(P_{V_\star})}(v; 1)\,.$$

Then for $\mathcal{L}_\star$ and $\lambda_\star$, Equations (3) satisfy $(\alpha, \tau) = (\alpha_\star, 1)$. Moreover, $\mathcal{L}_\star$ is convex provided that P_Z is log-concave and $\alpha_\star^2 < \lambda_\star \delta$.

We numerically validate the theoretical results of this section in Figure 1(Left), and in the Appendix in Figures 2(Top Left) and 3(Left). Specifically, we consider Laplacian noise $Z \sim \text{Laplace}(0, \text{b})$, where $\mathbb{E}[Z^2] = 2b^2$. In Figure 3(Left), we plot the optimal loss $\mathcal{L}_\star$ (computed as per Lemma 2.1) for $\delta = 2$ and $b = 1, 2$. Note that $\mathcal{L}_\star$ differs from the loss function of the maximum-likelihood estimator. Instead, it is a (non-trivial) smoothed version of it; see also [Bean et al., 2013, Advani and Ganguli, 2016]. In Figures 1(Left) and 2(Top Left), we use gradient descent to numerically evaluate the error of the pair $(\mathcal{L}_\star, \lambda_\star)$ as a function of δ, for $b = 1$ and $b = 2$, respectively. We compare the achieved error to the lower bound of Theorem 2.1. Note the perfect match.

2.3 The Sub-optimality Gap of RLS

Here, we use Theorem 2.1 to assess the statistical gap between least-squares and the optimal choice of $\mathcal{L}$. As a first step, the lemma below computes the high-dimensional limit of optimally regularized RLS.

Lemma 2.2 (Asymptotic Error of Optimally Regularized RLS). *Fix $\delta > 0$ and noise distribution P_Z. Let $\widehat{\mathbf{x}}_{\ell_2,\lambda}$ be the solution to λ-regularized least-squares (i.e., $\mathcal{L}(t) = t^2$ in (2)). Further let $\alpha_{\ell_2,\lambda}$ denote the high-dimensional limit of $\|\widehat{\mathbf{x}}_{\ell_2,\lambda} - \mathbf{x}_0\|_2^2$. Then, $\lambda \mapsto \alpha_{\ell_2,\lambda}$ is minimized at $\lambda_{\text{opt}} = 2\,\mathbb{E}[Z^2]$, and, it holds that*

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

Combining this with the closed-form lower bound of Corollary 2.1 delivers an explicit lower bound on the sub-optimality ratio $\alpha_\star^2/\alpha^2_{\ell_2,\lambda_{\text{opt}}}$, as follows,

$$\frac{\alpha_\star^2}{\alpha^2_{\ell_2,\lambda_{\text{opt}}}} \in [\omega_\delta, 1], \text{ with } \omega_\delta := \frac{h_\delta\left(1/\mathcal{I}(Z)\right)}{h_\delta\left(\mathbb{E}\left[Z^2\right]\right)}.$$

Note that the bound depends on the noise distribution only via its Fisher Information and its second moment. The fact that $\omega_\delta \le 1$ follows directly by the increasing nature of the function h_δ and the Cramer-Rao bound $\mathbb{E}[Z^2] \ge 1/\mathcal{I}(Z)$ (see Proposition A.3(c)). Using analytic properties of h_δ we can simplify the bound above even further. We show in Section C.6 that

$$\alpha_\star^2/\alpha^2_{\ell_2,\lambda_{\text{opt}}} \ge \omega_\delta \ge \max\left\{1-\delta\,,\,\left(\mathcal{I}(Z)\,\mathbb{E}[Z^2]\right)^{-1}\right\}. \quad (7)$$

The first term in the RHS of (7) reveals that in the highly over-parameterized regime ($\delta \ll 1$), it holds $\omega_\delta \approx 1$. Thus, optimally-regularized LS becomes optimal. More generally, in the over-parameterized regime $0 < \delta < 1$, the squared-error of optimally-tuned LS is no worse than $(1-\delta)^{-1}$ times the optimal performance among all convex ERM.

The second term in (7) is more useful in the underparameterized regime $\delta \ge 1$ and captures the effect of the noise distribution via the ratio $\left(\mathcal{I}(Z)\,\mathbb{E}[Z^2]\right)^{-1} \le 1$ (which is closely related to the classical Fisher information distance studied e.g. in [Johnson and Barron, 2004]). Using the fact that $\mathcal{I}(Z) = 1/\mathbb{E}[Z^2]$ (thus, ω_δ attains its maximum value 1) iff $Z \sim \mathcal{N}(0, \zeta^2)$. Hence, optimally-tuned LS is optimal when Z is Gaussian. To further illustrate that our results are informative for general noise distributions, consider Laplacian noise $Z \sim \texttt{Laplace}(0, b^2)$. Using $\mathbb{E}[Z^2] = 2b^2$ and $\mathcal{I}(Z) = b^{-2}$, it follows from(7) that $\omega_\delta \ge 1/2$, for all $b > 0$ and $\delta > 0$. Hence, we find that optimally-tuned RLS achieves squared-error that is at most twice as large as the optimal error, i.e. if $Z \sim \texttt{Laplace}(0, b^2)$, $b > 0$ then for all $\delta > 0$ it holds that $\alpha^2_{\ell_2,\lambda_{\text{opt}}} \le 2\,\alpha_\star^2$. See also Figures 1 and 2 for a numerical illustration.

3 Binary Models

Consider data $(y_i, \mathbf{a}_i)$, $i \in [m]$ from a binary model $y_i = f(\mathbf{a}_i^T \mathbf{x}_0)$, where $f : \mathbb{R} \to \{\pm 1\}$ is possibly random. We make the following mild assumption on f; see Section D.1 for a discussion.

Assumption 4 (Link function). *The link function f satisfies $\nu_f := \mathbb{E}\left[S\, f(S)\right] \neq 0$, for $S \sim \mathcal{N}(0, 1)$.*

Under Assumptions 1, 2 and 4 we study the following

ridge-regularized ERM for binary measurements,

$$\widehat{\mathbf{w}}_{\mathcal{L},\lambda} := \arg\min_{\mathbf{w}\in\mathbb{R}^n} \quad \frac{1}{m}\sum_{i=1}^{m}\mathcal{L}\left(y_i\mathbf{a}_i^T\mathbf{w}\right) + \frac{\lambda}{2}\|\mathbf{w}\|^2. \quad (8)$$

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

3.1 Asymptotic Performance

In contrast to linear models where we focused on squared error, for binary models, a more relevant performance measure is the normalized correlation $\mathrm{corr}\left(\widehat{\mathbf{w}}_{\mathcal{L},\lambda}, \mathbf{x}_0\right) := \frac{|\widehat{\mathbf{w}}_{\mathcal{L},\lambda}^T\mathbf{x}_0|}{\|\widehat{\mathbf{w}}_{\mathcal{L},\lambda}\|_2\|\mathbf{x}_0\|_2}$. Our first result determines the limit of $\mathrm{corr}\left(\widehat{\mathbf{w}}_{\mathcal{L},\lambda}, \mathbf{x}_0\right)$. Specifically, we show that for a wide class of loss functions it holds that

$$\rho_{\mathcal{L},\lambda} := \mathrm{corr}\left(\widehat{\mathbf{w}}_{\mathcal{L},\lambda}, \mathbf{x}_0\right) \xrightarrow{P} \sqrt{\frac{1}{1+\sigma_{\mathcal{L},\lambda}^2}}, \quad (9)$$

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

$$\mathbb{E}\Big[S\, f(S)\, \mathcal{M}'_{\mathcal{L},1}\left(\alpha G + \mu S f(S); \tau\right)\Big] = -\lambda\mu, \quad (10a)$$

$$\tau^2\,\delta\,\mathbb{E}\Big[\left(\mathcal{M}'_{\mathcal{L},1}\left(\alpha G + \mu S f(S); \tau\right)\right)^2\Big] = \alpha^2, \quad (10b)$$

$$\tau\,\delta\,\mathbb{E}\Big[G\, \mathcal{M}'_{\mathcal{L},1}\left(\alpha G + \mu S f(S); \tau\right)\Big] = \alpha(1-\lambda\tau\delta). \quad (10c)$$

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

$$\mathcal{C}_{\text{bin}} := \Big\{(\mathcal{L},\, f)\,\Big|\, \forall\lambda > 0\text{: (10) has a unique bounded solution } (\alpha_{\mathcal{L},\lambda} > 0, \mu_{\mathcal{L},\lambda}, \tau_{\mathcal{L},\lambda} > 0) \text{ and (9) holds}\Big\}.$$

Theorem 3.1 (Asymptotics for binary RERM)**.** *Let Assumptions 1 and 2 hold and $\|\mathbf{x}_0\|_2 = 1$. Assume the link function $f : \mathbb{R} \to \{\pm 1\}$ satisfies Assumption 4. Further assume a loss function $\mathcal{L}$ with the following properties: $\mathcal{L}$ is convex, twice differentiable and bounded from below such that $\mathcal{L}'(0) \neq 0$ and for $G \sim \mathcal{N}(0,1)$, we have $\mathbb{E}[\mathcal{L}(G)] < \infty$. Then, it holds that $(\mathcal{L}, f) \in \mathcal{C}_{\text{bin}}$.*

We prove Theorem 3.1 in Section B. Previous works have considered special instances of this: [Sur and Candès, 2019, Salehi et al., 2019] study unregularized and regularized logistic-loss for the logistic binary model, while [Taheri et al., 2020] studies strictly-convex ERM *without* regularization. Here, we follow the same approach as in [Salehi et al., 2019, Taheri et al., 2020], who apply the convex Gaussian min-max theorem (CGMT) to relate the performance of RERM to an auxiliary optimization (AO) problem whose first-order optimality conditions lead to the system of equations in (10). Our key technical contribution in proving Theorem 3.1 is proving existence and uniqueness of solutions to (10) for the broad class of convex losses as in the statement of the theorem (cf. Task **[T2]**). This is a non-trivial task in view of the highly nonlinear nature of (10). Specifically, we remark that none of the previous works has established existence. Also, note that the uniqueness result of [Taheri et al., 2020, Prop. 2.1] is limited to large enough values of the sampling ratio δ such that the data are linearly separable. As a final remark, compared to [Sur and Candès, 2019, Salehi et al., 2019, Taheri et al., 2020], we also show that the solution to (10) (specifically, the parameter $\sigma_{\mathcal{L},\lambda}^2$) further determines the limit of the classification error of $\widehat{\mathbf{w}}_{\mathcal{L},\lambda}$. Specifically, letting $\mathbf{a} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_n)$ be a fresh feature vector and $y = f\left(\mathbf{a}^T\mathbf{x}_0\right)$ its label, we show in Section D.2 that

$$\mathcal{E}_{\mathcal{L},\lambda} := \mathbb{P}_{\mathbf{a},y}\left(y \neq \mathrm{sign}\left(\mathbf{a}^T\widehat{\mathbf{w}}_{\mathcal{L},\lambda}\right)\right) \xrightarrow{P} \quad (11)$$
$$\mathbb{P}_{G,S}\left(\sigma_{\mathcal{L},\lambda}\, G + S f(S) < 0\right), \quad G, S \overset{\text{iid}}{\sim} \mathcal{N}(0,1).$$

3.2 Fundamental limits and optimal tuning

Eqns. (9) and (11) predict the high-dimensional limit of the correlation and classification-error of the RERM solution $\widehat{\mathbf{w}}_{\mathcal{L},\lambda}$. In fact, smaller values for $\sigma_{\mathcal{L},\lambda}$ result in better performance: higher correlation and classification accuracy (see Section D.2). Here, we lower bound $\sigma_{\mathcal{L},\lambda}$ and characterize the statistical limits of (8) (cf. Task **[T3]**).

Theorem 3.2 (Lower Bound on $\sigma_{\mathcal{L},\lambda}$)**.** *Let Assumptions 1, 2 and 4 hold. For $G, S \overset{iid}{\sim} \mathcal{N}(0,1)$ define the random variable $W_s := s\,G + S\,f(S)$ parameterized by $s \in \mathbb{R}$. Fix any $\delta > 0$ and define*

$$\sigma_\star := \min_{0\le x<1/\delta}\Big[s > 0 : \frac{1 - s^2(1 - s^2\mathcal{I}(W_s))}{\delta s^2(s^2\mathcal{I}(W_s) + \mathcal{I}(W_s) - 1)} - 2x + x^2\delta(1 + \frac{1}{s^2}) = 1\Big]. \quad (12)$$

For any $(\mathcal{L}, f) \in \mathcal{C}_{\text{bin}}$, $\lambda > 0$ and $\sigma_{\mathcal{L},\lambda}^2$ the respective high-dimensional limit of the error as in (9), it holds that $\sigma_{\mathcal{L},\lambda} \ge \sigma_\star$.

We prove Theorem 3.2 in Section D.3, where we also show that the minimization in (12) is always feasible. In view of (9) and (11) the theorem's lower bound translates to an upper bound on correlation and test accuracy. Note that $\sigma_\star$ depends on the link function only through the Fisher information of the random variable $s\,G + S\,f(S)$. This parallels the lower bound of Theorem 2.1 on linear models. Here, the role of the noise variable Z is played by the random variable $S\,f(S)$. This "effective noise term" $S\,f(S)$ fully captures the specifics of the link function f. Also, we see again, that the lower bound depends $\mathcal{I}(sG + Sf(S))$, that is the Fisher Information of the "noise distribution" augmented by a Gaussian sG.

Next we present a useful closed-form lower bound for $\sigma_\star$. For convenience let function $H_\delta : \mathbb{R}_{>1} \to \mathbb{R}_{>0}$ parameterized by $\delta > 0$ be defined as follows, $H_\delta(x) :=$

$$2\left(-\delta - x + \delta\, x + \sqrt{(-\delta - x + \delta\, x)^2 + 4\delta(x-1)}\right)^{-1}.$$

Corollary 3.1 (Lower bound on $\sigma_\star$)**.** *Let $\sigma_\star$ be as in (12). Fix any $\delta > 0$ and assume that f is such that the random variable $Sf(S)$ has a differentiable and strictly positive probability density on the real line. Then,*

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

Corollary 3.1 nicely parallels Corollary 2.1 for linear models. The proof of the corollary, presented in Section D.6, reveals that the more the distribution of $Sf(S)$ resembles a Gaussian distribution, the tighter the gap is, with equality achieved iff $Sf(S)$ is Gaussian.

Our next result strengthens the lower bound of Theorem 3.2 by showing existence of a loss function and regularizer parameter for which the system of equations (10) has a solution leading to $\sigma_\star$.

Lemma 3.1 (Optimal tuning for binary RERM)**.** *For given $\delta > 0$ and binary link function f, let $(\sigma_\star > 0, x_\star \in [0, 1/\delta))$ be the optimal solution in the minimization in (12). Denote $\lambda_\star = x_\star$ and define $W_\star := \sigma_\star G + Sf(S)$. Consider the loss function $\mathcal{L}_\star : \mathbb{R} \to \mathbb{R}$*

$$\mathcal{L}_\star(x) := -\mathcal{M}_{\frac{\lambda_\star\delta - 1}{\delta(\eta - \mathcal{I}(W_\star))}(\eta Q + \log P_{W_\star})}(x; 1)\,, \qquad (13)$$

where $\eta := 1 - \mathcal{I}(W_\star)\cdot(\sigma_\star^2 - \sigma_\star^2\lambda_\star\delta - \lambda_\star\delta) - \lambda_\star\delta$ and $Q(w) := w^2/2$. Then for $\mathcal{L}_\star$ and $\lambda_\star$, the equations (10) satisfy $(\alpha, \mu, \tau) = (\sigma_\star, 1, 1)$.

Lemma 3.1 suggests that if $\mathcal{L}_\star$ satisfies the assumptions of Theorem 3.1, then $\sigma_{\mathcal{L}_\star,\lambda_\star} = \sigma_\star$. In Figures 1 and 2 we verify this numerically for the Signed and Logistic models. Specifically, we numerically evaluate the performance of gradient descent on $\mathcal{L}_\star$ showing that the pair $(\mathcal{L}_\star, \lambda_\star)$ achieves the optimal error predicted by Theorem 3.1 (with remarkable accuracy despite the finite dimensions). See also Figure 3(Right) for an illustration of $\mathcal{L}_\star$.

3.3 The sub-optimality gap of RLS

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

Lemma 3.2 (Asymptotic error of RLS)**.** *Let Assumptions 1, 2 and 4 hold. Recall that $\nu_f = \mathbb{E}[Sf(S)] \neq 0$. Fix any $\delta > 0$ and consider solving (8) with the square-loss $\mathcal{L}(t) = (t-1)^2$ and $\lambda \geq 0$. Then, the system of equations in (10) has a unique solution $(\alpha_{\ell_2,\lambda}, \mu_{\ell_2,\lambda}, \tau_{\ell_2,\lambda})$ and $\sigma^2_{\ell_2,\lambda} = \frac{\alpha^2_{\ell_2,\lambda}}{\mu^2_{\ell_2,\lambda}} =$*

$$\frac{1}{2\delta\nu_f^2}\left(1 - \delta\nu_f^2 + \frac{2 + 2\delta + \lambda\delta + \delta\nu_f^2\,((2+\lambda)\delta - 6)}{\sqrt{4 + 4\delta(\lambda - 2) + \delta^2(\lambda+2)^2}}\right). \qquad (14)$$

Moreover, it holds that

$$\sigma^2_{\ell_2,\lambda} \geq \sigma^2_{\ell_2,\lambda_{\rm opt}} := H_\delta((1-\nu_f^2)^{-1}),$$

with equality attained for $\lambda_{\rm opt} = 2(1-\nu_f^2)/(\delta\,\nu_f^2)$.

In resemblance to Lemma 2.2 in which RLS performance for linear measurements only depends on the second moment $\mathbb{E}[Z^2]$ of the additive noise distribution, Lemma 3.2 reveals that the corresponding key parameter for binary models is $1 - \nu_f^2$. Interestingly, the expression for $\sigma^2_{\ell_2,\lambda_{\rm opt}}$ conveniently matches with the simple bound on $\sigma_\star^2$ in Corollary 3.1. Specifically, it holds for any $\delta > 0$ that

$$1 \;\geq\; \frac{\sigma_\star^2}{\sigma^2_{\ell_2,\lambda_{\rm opt}}} \geq \Omega_\delta := \frac{H_\delta\left(\,\mathcal{I}(S\,f(S))\,\right)}{H_\delta\left((1-\nu_f^2)^{-1}\right)}. \qquad (15)$$

It can be checked that $H_\delta(\cdot)$ is strictly-decreasing in its domain for a fixed $\delta > 0$. Furthermore, the Cramer-Rao bound (see Prop. A.3 (d)) requires that $\mathcal{I}(Sf(S)) \geq (\mathrm{Var}[Sf(S)])^{-1} = (1-\nu_f^2)^{-1}$. Combining these, confirms that $\Omega_\delta \leq 1$. Furthermore $\Omega_\delta = 1$ (thus, $\sigma_\star^2 = \sigma^2_{\ell_2,\lambda_{\rm opt}}$) iff the random variable $Sf(S)$ is Gaussian. This conclusion is similar to what we found for linear models. However, for binary models satisfying Assumption 4, it can be easily checked (see Section D.1) that $Sf(S)$ is never Gaussian. Thus (15) suggests that square-loss cannot be optimal. Nevertheless, one can use (15) to argue that square-loss is (perhaps sur-

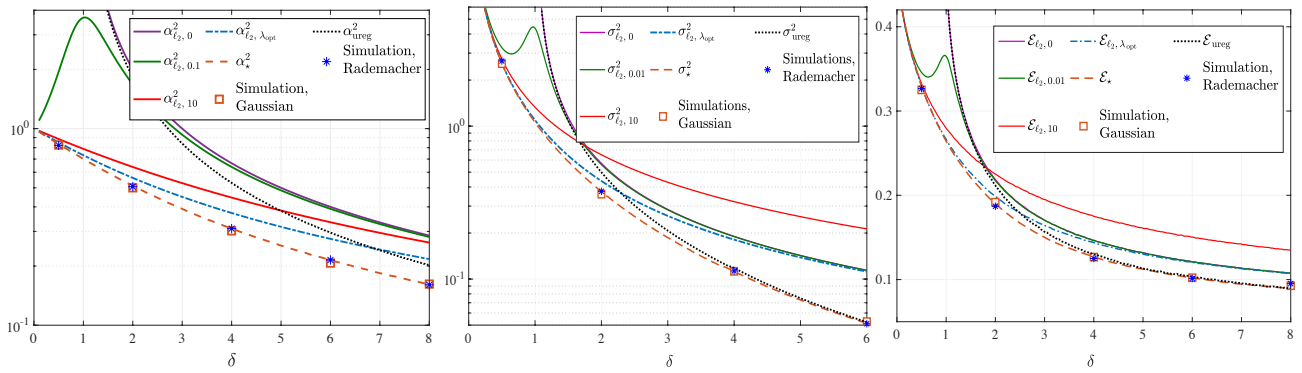

Figure 1: The lower bounds on error derived in this paper, compared to RLS for the linear model with $Z \sim \texttt{Laplace}(0,1)$ (Left), and for the binary Signed model(Middle) and binary Logistic model with $\|\mathbf{x}_0\|= 10$ (Right). The markers denote the emipirical performance of the optimally tuned RERM as derived in Lemmas 2.1 and 3.1 for Gaussian and Rademacher data. See Section G for additional numerical results.

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

Table 1: Theoretical and numerical values of $\alpha_\star^2/\alpha_{\mathcal{L},\lambda_{\text{opt}}}^2$ (linear models) and $\sigma_\star^2/\sigma_{\mathcal{L},\lambda_{\text{opt}}}^2$(binary models) for different values of δ and various link functions. The curves for $\alpha_\star$ and $\sigma_\star$ correspond to Theorems 2.1 and 3.2. The empirical values of $\alpha_\star$ and $\sigma_\star$ are derived by numerically solving the optimally-tuned RERM of Lemmas 2.1 and 3.1 by GD for isotropic Gaussian features, $n = 100$ and averaging over 50 independent experiments.

prisingly) approximately optimal for certain popular models. For instance, consider the logistic link function $\widetilde{f}_r$ satisfying $\mathbb{P}(\widetilde{f}_r(x) = 1) = (1 + \exp(-rx))^{-1}$, where $r := \|\mathbf{x}_0\|_2$. Using (15) and maximizing the sub-optimality gap $1/\Omega_\delta$ over $\delta > 0$, we find that if $f = \widetilde{f}_{r=1}$ then for all $\delta > 0$ it holds that

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

Thus, for a logistic link function and $\|\mathbf{x}_0\|_2= 1$ optimally-tuned RLS is approximately optimal. This is in agreement with the key message of Corollary 3.1 on the critical role played by $Sf(S)$, since for the logistic model and small values of r, its density is "close" to a Gaussian. We remark that [Taheri et al., 2020] further shows that LS remains approximately optimal among convex loss functions without regularization for the logistic and probit models with $r = 1$. However, [Taheri et al., 2020] did not investigate the effect of the SNR term $r := \|\mathbf{x}_0\|_2$. Specifically, as the signal strength increases, $\widetilde{f}_r$ converges to the sign function ($\widetilde{f}_r(\cdot) \to \text{sign}(\cdot)$). This suggests that there might be room for improvement between RLS and what Theorem 3.2 suggests to be possible. This can be precisely quantified using (15). For example, for $r = 10$ it can be checked that $\sigma_{\ell_2,\lambda_{\text{opt}}}^2 \leq 2.442\ \sigma_\star^2,\ \forall \delta > 0$. Lemma 3.1 provides the recipe to bridge the gap in this case. Indeed, Figures 1 and 2 show that the optimal loss function $\mathcal{L}$ predicted by the lemma outperforms RLS for all values δ and its performance matches the best possible one specified by Theorem 3.2.

Due to space limitations, we defer Figures 2 and 3 to the appendix; see Section G.

4 Numerical Experiments

In Figure 1(Left), we compare the lower bound of Theorem 2.1 with the error of RLS (see Lemma 2.2) for $Z \sim \texttt{Laplace}(0,1)$ and $\|\mathbf{x}_0\|_2= 1$. To numeri-

cally validate that $\alpha_\star$ is achievable by the proposed choices of loss function and regularization parameter in Lemma 2.1, we proceed as follows. We generate noisy linear measurements with iid Gaussian feature vectors $\mathbf{a}_i \in \mathbb{R}^{100}$. The estimator $\widehat{\mathbf{x}}_{\mathcal{L}_\star,\lambda_\star}$ is computed by running gradient descent (GD) on the corresponding optimization in (2) when the proposed optimal loss and regularizer of Lemma 2.1 are used. See Figure 3(Left) for an illustration of the optimal loss for this model. The resulting vector $\widehat{\mathbf{x}}_{\mathcal{L}_\star,\lambda_\star}$ is used to compute $\|\widehat{\mathbf{x}}_{\mathcal{L}_\star,\lambda_\star} - \mathbf{x}_0\|^2$. The average of these values over 50 independent Monte-carlo trials is shown in red squares. Note the remarkable agreement between theoretical and empirical values despite the finite dimensions (see also the first and second rows of Table 1). In the next two plots, we present results for binary models. Figure 1(Middle) plots the effective error parameter σ for the Signed model and Figure 1(Right) plots the classification error '$\mathcal{E}$' for the Logistic model with $\|\mathbf{x}_0\|_2 = 10$. The red squares correspond to the numerical evaluations of ERM with $\mathcal{L} = \mathcal{L}_\star$ and $\lambda = \lambda_\star$ (as in Lemma 3.1) derived by running GD on the proposed optimal loss and regularization parameter. See Figure 3(Right) for an illustration of the optimal loss in this case. The solution $\widehat{\mathbf{w}}_{\mathcal{L}_\star,\lambda_\star}$ of GD is used to calculate $\sigma_{\mathcal{L}_\star,\lambda_\star}$ and $\mathcal{E}_{\mathcal{L}_\star,\lambda_\star}$ in accordance with (9) and (11), respectively. Again, note the close match between theory and experiments (see the third and fourth rows of Table 1).

The goal of the next experiment is to numerically support the universality property of our results discussed in Section 1.1. For this purpose, we repeat the experiments above with choosing the entries $\mathbf{a}_i$ as independent Rademacher random variables. We plot the numerical averages in blue stars. Again, for all three plots, note the remarkable agreement of these values to both the corresponding numerical values for Gaussian features, and, our theoretical performance bounds.

Finally, for all three models studied in Figure 1, we include the plots the theoretical predictions for the error of the following: (i) RLS with small and large regularization (see Eqns. (56) and (14)); (ii) optimally tuned RLS (see Lemmas 2.2 and 3.2); (iii) optimally-tuned unregularized ERM (marked as $\alpha_{\text{ureg}}, \sigma_{\text{ureg}}, \mathcal{E}_{\text{ureg}}$). The curves for the latter are obtained from [Bean et al., 2013] and [Taheri et al., 2020] for linear and binary models, respectively. We refer the reader to Sections F.1 and F.2 for a precise study of the benefits of regularization in view of Theorems 2.1 and 3.2, for both linear and binary models.

5 Future work

There is a host of exciting directions for future work. Proving universality of our results is an important, yet possibly challenging, task. Extensions to correlated features is yet another important direction. While we suspect that our techniques are still useful, such an extension requires revisiting Task **[T1]** to obtain the appropriate system of equations (one that properly accounts for the covariance matrix) for that case; see [Montanari et al., 2019, Liang and Sur, 2020] for some very recent progress in this direction, but only for special ERM instances.

Acknowledgements

This work was supported by NSF grants CNS-2003035, CCF-2009030 and CCF-1909320.

References

[Abbasi et al., 2019] Abbasi, E., Salehi, F., and Hassibi, B. (2019). Universality in learning from linear measurements. In *Advances in Neural Information Processing Systems*, pages 12372–12382.

[Abbasi et al., 2016] Abbasi, E., Thrampoulidis, C., and Hassibi, B. (2016). General performance metrics for the lasso. In *2016 IEEE Information Theory Workshop (ITW)*, pages 181–185. IEEE.

[Advani and Ganguli, 2016] Advani, M. and Ganguli, S. (2016). Statistical mechanics of optimal convex inference in high dimensions. *Physical Review X*, 6(3):031034.

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

[Bayati et al., 2015] Bayati, M., Lelarge, M., Montanari, A., et al. (2015). Universality in polytope phase transitions and message passing algorithms. *The Annals of Applied Probability*, 25(2):753–822.

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

[Bean et al., 2013] Bean, D., Bickel, P. J., El Karoui, N., and Yu, B. (2013). Optimal m-estimation in high-dimensional regression. *Proceedings of the National Academy of Sciences*, 110(36):14563–14568.

[Blachman, 1965] Blachman, N. (1965). The convolution inequality for entropy powers. *IEEE Transactions on Information Theory*, 11(2):267–271.

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

[Candès, 2014] Candès, E. J. (2014). Mathematics of sparsity (and a few other things). In *Proceedings of the International Congress of Mathematicians, Seoul, South Korea*, volume 123. Citeseer.

[Candès and Sur, 2018] Candès, E. J. and Sur, P. (2018). The phase transition for the existence of the maximum likelihood estimate in high-dimensional logistic regression. *arXiv preprint arXiv:1804.09753*.

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

[Chandrasekaran et al., 2012] Chandrasekaran, V., Recht, B., Parrilo, P. A., and Willsky, A. S. (2012). The convex geometry of linear inverse problems. *Foundations of Computational Mathematics*, 12(6):805–849.

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

[Dobriban et al., 2018] Dobriban, E., Wager, S., et al. (2018). High-dimensional asymptotics of prediction: Ridge regression and classification. *The Annals of Statistics*, 46(1):247–279.

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

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

[Donoho and Montanari, 2015] Donoho, D. L. and Montanari, A. (2015). Variance breakdown of huber (m)-estimators: n/p \rightarrow m\in (1,\infty). *arXiv preprint arXiv:1503.02106*.

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

[Emami et al., 2020] Emami, M., Sahraee-Ardakan, M., Pandit, P., Rangan, S., and Fletcher, A. (2020). Generalization error of generalized linear models in high dimensions. In *International Conference on Machine Learning*, pages 2892–2901. PMLR.

[Genzel, 2016] Genzel, M. (2016). High-dimensional estimation of structured signals from non-linear observations with general convex loss functions. *IEEE Transactions on Information Theory*, 63(3):1601–1619.

[Gerbelot et al., 2020] Gerbelot, C., Abbara, A., and Krzakala, F. (2020). Asymptotic errors for teacher-student convex generalized linear models (or: How to prove kabashima's replica formula). *arXiv preprint arXiv:2006.06581*.

[Goldt et al., 2020] Goldt, S., Reeves, G., Mézard, M., Krzakala, F., and Zdeborová, L. (2020). The gaussian equivalence of generative models for learning with two-layer neural networks. *arXiv preprint arXiv:2006.14709*.

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

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

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

[Huber, 2011] Huber, P. J. (2011). *Robust statistics*. Springer.

[Johnson and Barron, 2004] Johnson, O. and Barron, A. (2004). Fisher information inequalities and the central limit theorem. *Probability Theory and Related Fields*, 129(3):391–409.

[Kammoun and Alouini, 2020] Kammoun, A. and Alouini, M.-S. (2020). On the precise error analysis of support vector machines. *arXiv preprint arXiv:2003.12972*.

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

[Kini and Thrampoulidis, 2020] Kini, G. R. and Thrampoulidis, C. (2020). Analytic study of double descent in binary classification: The impact of loss. In *2020 IEEE International Symposium on Information Theory (ISIT)*, pages 2527–2532. IEEE.

[Lei et al., 2018] Lei, L., Bickel, P. J., and El Karoui, N. (2018). Asymptotics for high dimensional regression m-estimates: fixed design results. *Probability Theory and Related Fields*, 172(3):983–1079.

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

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

[Lu and Li, 2017] Lu, Y. M. and Li, G. (2017). Phase transitions of spectral initialization for high-dimensional nonconvex estimation. *arXiv preprint arXiv:1702.06435*.

[Mai et al., 2019] Mai, X., Liao, Z., and Couillet, R. (2019). A large scale analysis of logistic regression: asymptotic performance and new insights. In *ICASSP*.

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

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

[Mondelli and Montanari, 2017] Mondelli, M. and Montanari, A. (2017). Fundamental limits of weak recovery with applications to phase retrieval. *arXiv preprint arXiv:1708.05932*.

[Montanari, 2015] Montanari, A. (2015). Statistical estimation: from denoising to sparse regression and hidden cliques. *Statistical Physics, Optimization, Inference and Message-passing Algorithms: Lecture Notes of the Les Houches School of Physics-Special Issue, October 2013*, page 127.

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

[Oymak and Hassibi, 2016] Oymak, S. and Hassibi, B. (2016). Sharp mse bounds for proximal denoising. *Foundations of Computational Mathematics*, 16(4):965–1029.

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

[Oymak and Tropp, 2017] Oymak, S. and Tropp, J. A. (2017). Universality laws for randomized dimension reduction, with applications. *Information and Inference: A Journal of the IMA*, 7(3):337–446.

[Panahi and Hassibi, 2017] Panahi, A. and Hassibi, B. (2017). A universal analysis of large-scale regularized least squares solutions. In *Advances in Neural Information Processing Systems*, pages 3381–3390.

[Plan and Vershynin, 2015] Plan, Y. and Vershynin, R. (2015). The generalized lasso with non-linear observations. *arXiv preprint arXiv:1502.04071*.

[Rockafellar, 1997] Rockafellar, R. T. (1997). *Convex analysis*, volume 28. Princeton university press.

[Rockafellar and Wets, 2009] Rockafellar, R. T. and Wets, R. J.-B. (2009). *Variational analysis*, volume 317. Springer Science & Business Media.

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

[Salehi et al., 2020] Salehi, F., Abbasi, E., and Hassibi, B. (2020). The performance analysis of generalized margin maximizers on separable data. In *International Conference on Machine Learning*, pages 8417–8426. PMLR.

[Sion, 1958] Sion, M. (1958). On general minimax theorems. *Pacific J. Math.*, 8(1):171–176.

[Stojnic, 2009] Stojnic, M. (2009). Various thresholds for ℓ_1-optimization in compressed sensing. *arXiv preprint arXiv:0907.3666*.

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

[Sur and Candès, 2019] Sur, P. and Candès, E. J. (2019). A modern maximum-likelihood theory for high-dimensional logistic regression. *Proceedings of the National Academy of Sciences*, page 201810420.

[Taheri et al., 2019] Taheri, H., Pedarsani, R., and Thrampoulidis, C. (2019). Sharp guarantees for solving random equations with one-bit information. In *2019 57th Annual Allerton Conference on Communication, Control, and Computing (Allerton)*, pages 765–772.

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

[Thrampoulidis et al., 2015a] Thrampoulidis, C., Abbasi, E., and Hassibi, B. (2015a). Lasso with non-linear measurements is equivalent to one with linear measurements. In *Advances in Neural Information Processing Systems*, pages 3420–3428.

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

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

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

SUPPLEMENTARY MATERIAL

A Useful facts

A.1 On Moreau Envelopes

In Proposition A.1, some of the differential properties of Moreau-envelope functions, used throughout the paper are summarized (cf. [Rockafellar and Wets, 2009]):

Proposition A.1 (Properties of Moreau-envelopes)**.** *Let $\mathcal{L}$ be a lower semi-continuous and proper function. Then*

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

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

$$\mathcal{M}'_{\mathcal{L},1}(x;\tau) := \frac{\partial \mathcal{M}_{\mathcal{L}}(x;\tau)}{\partial x} = \frac{1}{\tau}(x - \text{prox}_{\mathcal{L}}(x;\tau)), \tag{16}$$

$$\mathcal{M}'_{\mathcal{L},2}(x;\tau) := \frac{\partial \mathcal{M}_{\mathcal{L}}(x;\tau)}{\partial \tau} = -\frac{1}{2\tau^2}(x - \text{prox}_{\mathcal{L}}(x;\tau))^2. \tag{17}$$

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

$$\mathcal{M}'_{\mathcal{L},1}(x;\tau) = \mathcal{L}'(\text{prox}_{\mathcal{L}}(x;\tau)), \tag{18}$$

$$\mathcal{M}'_{\mathcal{L},2}(x;\tau) = -\frac{1}{2}(\mathcal{L}'(\text{prox}_{\mathcal{L}}(x;\tau))^2. \tag{19}$$

(c) Additionally, based on the relations above, if $\mathcal{L}$ is twice differentiable then the following is derived for its second order derivatives :

$$\mathcal{M}''_{\mathcal{L},1}(x;\tau) = \frac{\mathcal{L}''(\text{prox}_{\mathcal{L}}(x;\tau))}{1+\tau\mathcal{L}''(\text{prox}_{\mathcal{L}}(x;\tau))}, \tag{20}$$

$$\mathcal{M}''_{\mathcal{L},2}(x;\tau) = \frac{\left(\mathcal{L}'(\text{prox}_{\mathcal{L}}(x;\tau))\right)^2 \mathcal{L}''(\text{prox}_{\mathcal{L}}(x;\tau))}{1+\tau\,\mathcal{L}''(\text{prox}_{\mathcal{L}}(x;\tau))}. \tag{21}$$

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

Proposition A.2 (Inverse of the Moreau envelope)**.** *[Advani and Ganguli, 2016, Result. 23] For $\tau > 0$ and f a convex, lower semi-continuous function such that $g(\cdot) = \mathcal{M}_f(\cdot;\tau)$, the Moreau envelope can be inverted so that $f(\cdot) = -\mathcal{M}_{-g}(\cdot;\tau)$.*

Lemma A.1 (e.g., [Taheri et al., 2020], Lemma A.1.)**.** *The function $H : \mathbb{R}^3 \to \mathbb{R}$ defined as follows*

$$H(x,p,\tau) = \frac{1}{2\tau}(x-p)^2, \tag{22}$$

is jointly convex in its arguments.

A.2 On Fisher Information

In Proposition A.3 we collect some useful properties of the Fisher Information for location. For the proofs and more details, we refer the interested reader to [Blachman, 1965].

Proposition A.3 (Properties of Fisher Information, [Blachman, 1965])**.** *Let X be a zero-men random variable with probability density p_X satisfying the following conditions: (i) $p_X(x) > 0, -\infty < x < \infty$; (ii) $p'_X(x)$ exists;*

and (iii) The following integral exists:

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

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

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

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

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

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

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

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

$$\mathcal{I}(X_1+X_2) \leq \frac{\mathcal{I}(X_1)\cdot\mathcal{I}(X_2)}{\mathcal{I}(X_1)+\mathcal{I}(X_2)}. \tag{23}$$

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

Lemma A.2. *Let $G \sim \mathcal{N}(0,1)$ and Z be a random variable satisfying the assumptions of Proposition A.3. For any $a \in \mathbb{R}$, use the shorhand $V_a := a\,G + Z$. The following are true:*

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

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

Proof. To show part (a), we use Proposition A.3(e) with $\theta = 0$ to derive that

$$\lim_{a\to 0} a^2\,\mathcal{I}(V_a) \leq \lim_{a\to 0} a^2\,\mathcal{I}(Z) = 0, \tag{24}$$

where the second step follows by the fact that $\mathcal{I}(Z)$ is finite for any Z satisfying the assumption of the lemma. In order to prove part (b), we apply Proposition A.3(c) to deduce that :

$$\lim_{a\to+\infty} a^2\,\mathcal{I}(V_a) = \lim_{a\to+\infty} a^2\,\mathcal{I}(a\,G+Z) = \lim_{a\to+\infty} \mathcal{I}(G+\frac{1}{a}Z) = 1, \tag{25}$$

□

B Asymptotics for Binary RERM: Proof of Theorem 3.1

In this section, we prove that under the assumptions of Theorem 3.1, the system of equations in (10) has a unique and bounded solution.

B.1 Asymptotic Error of RERM via an Auxiliary Min-Max Optimization

As mentioned in Section 3, the proof of Theorem 3.1 has essentially two parts. The first part of the proof uses the CGMT [Thrampoulidis et al., 2015b] and the machinery developed in [Thrampoulidis et al., 2018, Thrampoulidis et al., 2015a, Salehi et al., 2019, Taheri et al., 2019] to relate the properties of the RERM solution to an Auxiliary Optimization (AO). The detailed steps follow mutatis-mutandis analogous derivations in recent works [Thrampoulidis et al., 2018, Thrampoulidis et al., 2015a, Salehi et al., 2019, Taheri et al., 2019, Kammoun and Alouini, 2020] and are omitted here for brevity. Instead, we summarize the finding of this analysis in the following proposition.

Proposition B.1. *Consider the optimization problem in (8). If the min-max optimization in (26) has a unique and bounded solution $(\alpha^\star > 0, \mu^\star, \upsilon^\star > 0, \gamma^\star > 0)$, then the values of $\alpha_{\mathcal{L},\lambda}$ and $\mu_{\mathcal{L},\lambda}$ corresponding to $\mathcal{L}$ and λ defined in (62)-(63) are derived by setting $\alpha_{\mathcal{L},\lambda} = \alpha_\star$ and $\mu_{\mathcal{L},\lambda} = \mu_\star$, where*

$$(\alpha^\star, \mu^\star, \upsilon^\star, \gamma^\star) = \arg \min_{\substack{(\alpha,\mu,\upsilon)\in \\ \mathbb{R}_{\geq 0}\times\mathbb{R}\times\mathbb{R}_{>0}}} \max_{\gamma\in\mathbb{R}_{>0}} \Big[\Theta(\alpha,\mu,\upsilon,\gamma) := \frac{\gamma\upsilon}{2} - \frac{\alpha\gamma}{\sqrt{\delta}} + \frac{\lambda\mu^2}{2} + \frac{\lambda\alpha^2}{2} + \mathbb{E}\left[\mathcal{M}_{\mathcal{L}}\left(\alpha G + \mu S f(S); \frac{\upsilon}{\gamma}\right)\right]\Big], \tag{26}$$

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

The system of equations in (10) is derived by the first-order optimality conditions of the function Θ based on its arguments $(\alpha, \mu, \upsilon, \gamma)$, i.e., by imposing $\nabla\Theta = \mathbf{0}$. In fact, similar to [Taheri et al., 2020], it only takes a few algebraic steps to simplify the four equations in $\nabla\Theta = \mathbf{0}$ to the three equations in (10).

For the rest of this section, we focus on the second part of the proof of Theorem 3.1 regarding existence/uniqueness of solutions to (10), which has not been previously studied in our setting.

B.2 Properties of Θ : Strict Convexity-Strict Concavity and Boundedness of Saddle Points

We will show in Lemma B.2 that for proving uniqueness and boundedness of the solutions to (10), it suffices to prove uniqueness and boundedness of the saddle point $(\alpha^\star, \mu^\star, \upsilon^\star, \gamma^\star)$ of Θ. In fact, a sufficient condition for uniqueness of solutions in (26) is that Θ is (jointly) strictly convex in (α, μ, υ) and strictly-concave in γ (e.g., see [Taheri et al., 2020, Lemma B.2.]). Lemma B.1, which is key to the proof of Theorem 3.1, derives sufficient conditions on $\mathcal{L}$ guaranteeing strict convexity-strict concavity of Θ as well as conditions on $\mathcal{L}$ ensuring boundedness of $(\alpha^\star, \mu^\star, \upsilon^\star, \gamma^\star)$.

Lemma B.1 (Properties of Θ)**.** *Let $\mathcal{L}(\cdot)$ be a lower semi-continuous (lsc), proper and convex function and $\lambda > 0$. Then the following statements hold for the function $\Theta : \mathbb{R}_{\geq 0} \times \mathbb{R} \times \mathbb{R}_{>0} \times \mathbb{R}_{>0} \to \mathbb{R}$ in (26),*

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

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

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

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

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

B.2.1 Proof of Lemma B.1

Proof of Lemma B.1(a). Let $\widetilde{\Theta}(\alpha, \mu, \upsilon) := \sup_{\gamma\in\mathbb{R}_{>0}} \Theta(\alpha, \mu, \upsilon, \gamma)$. For all feasible (α, μ, υ) it holds

$$\begin{aligned} \widetilde{\Theta}(\alpha,\mu,\upsilon) &\geq \Theta(\alpha,\mu,\upsilon,1) \\ &= \frac{\upsilon}{2} - \frac{\alpha}{\sqrt{\delta}} + \frac{\lambda(\alpha^2+\mu^2)}{2} + \mathbb{E}\Big[\mathcal{M}_{\mathcal{L}}(\alpha G + \mu S f(S); \upsilon)\Big]. \end{aligned} \tag{27}$$

Recall that $\mathcal{L}$ is bounded from below, i.e., for all $\mathcal{L}(x) \geq B, \forall x \in \mathbb{R}$ for some real B. By definition of Moreau-envelope function the same bound holds for $\mathcal{M}_{\mathcal{L}}$, i.e. for all $x \in \mathbb{R}$ and $y \in \mathbb{R}_{>0}$, we have that $\mathcal{M}_{\mathcal{L}}(x; y) \geq B$. Using this, we proceed from (27) to derive that:

$$\widetilde{\Theta}(\alpha,\mu,\upsilon) \geq B + \frac{\upsilon}{2} - \frac{\alpha}{\sqrt{\delta}} + \frac{\lambda(\alpha^2+\mu^2)}{2}. \tag{28}$$

Based on (28) that holds for all feasible (α, μ, υ) and using the fact that $\lambda > 0$ it can be readily shown that

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

Thus, the function $\widetilde{\Theta}\left(\alpha, \mu, \upsilon\right)$ is level-bounded in $\mathbb{R}_{\geq 0} \times \mathbb{R} \times \mathbb{R}_{>0}$. This implies the boundedness of solutions $(\alpha^\star, \mu^\star, \upsilon^\star)$ to (26) [Rockafellar and Wets, 2009, Thm. 1.9], as desired.

Proof of Lemma B.1(b). Under the assumptions of the lemma, we know from part (a) that the set of solutions to $(\alpha^\star, \mu^\star, \upsilon^\star)$ in (26) is bounded. Thus we can apply Sion's Min-Max Theorem [Sion, 1958] and flip the order of minimum and maximum to write:

$$\min_{\substack{(\alpha,\mu,\upsilon)\\ \in\,[0,C]\times[-C,C]\times(0,C]}} \max_{\gamma\,\in\,\mathbb{R}_{\geq 0}} \Theta(\alpha, \mu, \upsilon, \gamma) = \max_{\gamma\,\in\,\mathbb{R}_{\geq 0}} \Big[\widehat{\Theta}(\gamma) := \min_{\substack{(\alpha,\mu,\upsilon)\\ \in\,[0,C]\times[-C,C]\times(0,C]}} \Theta\left(\alpha, \mu, \upsilon, \gamma\right)\Big]. \tag{29}$$

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

$$\widehat{\Theta}(\gamma) \;\leq\; \Theta\left(1, 0, 1/\sqrt{\delta}, \gamma\right) = -\frac{\gamma}{2\sqrt{\delta}} + \frac{\lambda}{2} + \mathbb{E}\left[\mathcal{M}_{\mathcal{L}}\left(G; \frac{1}{\gamma\sqrt{\delta}}\right)\right]. \tag{30}$$

Note that for any $y \in \mathbb{R}$: $\mathcal{M}_{\mathcal{L}}\left(y; \frac{1}{\gamma\sqrt{\delta}}\right) = \min_{x\in\mathbb{R}} \frac{\gamma\sqrt{\delta}}{2}(x-y)^2 + \mathcal{L}(x) \leq \mathcal{L}(y)$. Thus we derive from (30):

$$\widehat{\Theta}(\gamma) \leq -\frac{\gamma}{2\sqrt{\delta}} + \frac{\lambda}{2} + \mathbb{E}\left[\mathcal{L}(G)\right]. \tag{31}$$

But $\mathbb{E}\left[\mathcal{L}(G)\right]$ is assumed to be bounded, thus it can be concluded from (31) that the function $\widehat{\Theta}(\gamma)$ is level-bounded, i.e.,

$$\lim_{\gamma \to +\infty} \widehat{\Theta}(\gamma) = -\infty. \tag{32}$$

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

Proof of Lemma B.1(c). First, we show that $\gamma^\star > 0$. On the contrary, assume that $\gamma^\star = 0$. Then based on (26) and Proposition A.1(a),

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

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

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

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

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

thus based on (26),

$$\mathbb{E}\left[G\cdot\mathcal{M}'_{\ell,1}\left(\mu^\star Sf(S);\frac{\upsilon^\star}{\gamma^\star}\right)\right]-\frac{\gamma^\star}{\sqrt{\delta}}\geq 0. \tag{33}$$

Since by assumption G and $Sf(S)$ are independent and $\mathbb{E}[G]=0$, we deduce from (33) that $\gamma^\star=0$, which is in contradiction to the previously proved fact that $\gamma^\star>0$. This shows that $\alpha^\star>0$, as desired. Finally, we note that if $\upsilon^\star=0$, then based on (26) and in light of Proposition A.1(a), we find that,

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

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

Proof of Lemma B.1(d). Let $\mathbf{w}_1:=(\alpha_1,\mu_1,\tau_1)$ and $\mathbf{w}_2:=(\alpha_2,\mu_2,\tau_2)$ be two distinct points in the space $\mathbb{R}_{\geq 0}\times\mathbb{R}\times\mathbb{R}_{>0}$. We consider two cases :

Case I : $(\alpha_1,\mu_1)=(\alpha_2,\mu_2)$
In this case, it suffices to show that for fixed $\alpha>0$ and μ and under the assumptions of the lemma, the function $\mathbb{E}\left[\mathcal{M}_{\mathcal{L}}\left(\alpha G+\mu Sf(S);\tau\right)\right]$ is strictly-convex in τ. Denote by $p(\alpha,\mu,\tau):=\text{prox}_{\mathcal{L}}\left(\alpha G+\mu Sf(S);\tau\right)$. First, we derive second derivate of the Moreau-envelope function with respect to τ by applying (21), and further use convexity of $\mathcal{L}$ to derive that :

$$\begin{aligned}&\frac{\partial^2}{\partial\tau^2}\mathbb{E}\Big[\mathcal{M}_{\mathcal{L}}\left(\alpha G+\mu Sf(S);\tau\right)\Big]\\&\qquad=\mathbb{E}\left[\frac{\Big(\mathcal{L}'\left(p\left(\alpha,\mu,\tau\right)\right)\Big)^2\mathcal{L}''\left(p\left(\alpha,\mu,\tau\right)\right)}{1+\tau\,\mathcal{L}''\left(p\left(\alpha,\mu,\tau\right)\right)}\right]\geq 0.\end{aligned} \tag{34}$$

Next we show that the inequality above is strict if $\mathcal{L}(\cdot)$ is a non-linear function. First we note that combining (16) and (18) yields that for all $x\in\mathbb{R}$:

$$\begin{aligned}\mathcal{L}'(\text{prox}_{\mathcal{L}}\left(x;\tau\right))&=\frac{1}{\tau}(x-\text{prox}_{\mathcal{L}}\left(x;\tau\right)),\\ \mathcal{L}''(\text{prox}_{\mathcal{L}}\left(x;\tau\right))&=\frac{1-\text{prox}'_{\mathcal{L},1}\left(x;\tau\right)}{\tau\cdot\text{prox}'_{\mathcal{L},1}\left(x;\tau\right)}.\end{aligned}$$

Using these relations and denoting by $p'(\alpha,\mu,\tau):=\text{prox}'_{\mathcal{L},1}\left(\alpha G+\mu Sf(S);\tau\right)$, we can rewrite (34) as following :

$$\begin{aligned}&\frac{\partial^2}{\partial\tau^2}\mathbb{E}\Big[\mathcal{M}_{\mathcal{L}}\left(\alpha G+\mu Sf(S);\tau\right)\Big]\\&\qquad=\frac{1}{\tau^3}\mathbb{E}\left[\frac{\Big(\alpha G+\mu Sf(S)-p(\alpha,\mu,\tau)\Big)^2\Big(1-p'(\alpha,\mu,\tau)\Big)}{p'(\alpha,\mu,\tau)\Big(1+\tau\,\mathcal{L}''(p(\alpha,\mu,\tau))\Big)}\right].\end{aligned} \tag{35}$$

It is straightforward to see that if $\alpha>0$, then $\alpha G+\mu Sf(S)$ has positive density in the real line. Thus from (35) we find that :

$$\frac{\partial^2}{\partial\tau^2}\mathbb{E}\Big[\mathcal{M}_{\mathcal{L}}\left(\alpha G+\mu Sf(S);\tau\right)\Big]=0\iff\exists c\in\mathbb{R}\ \text{ s.t. }\forall x\in\mathbb{R}:\quad\text{prox}_{\mathcal{L}}\left(x;\tau\right)=x+c. \tag{36}$$

Recalling (16), we see that the condition in (36) is satisfied if and only if :

$$\exists c_1, c_2 \in \mathbb{R} : \text{s.t. } \forall x \in \mathbb{R} : \mathcal{M}_{\mathcal{L}}(x;\tau) = c_1 x + c_2. \tag{37}$$

Using inverse properties of Moreau-envelope in Proposition A.2, we derive that the loss function $\mathcal{L}(\cdot)$ satisfying (37) takes the following shape,

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

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

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

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

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

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

$$\begin{aligned}
\Omega(\theta\mathbf{w}_1 + \overline{\theta}\mathbf{w}_2) &\leq \frac{\lambda\mu_\theta^2}{2} + \frac{\lambda\alpha_\theta^2}{2} + \mathbb{E}\left[\frac{1}{2\tau_\theta}\Big(\alpha_\theta G + \mu_\theta Sf(S) - (\theta p_1 + \overline{\theta}p_2)\Big)^2 + \mathcal{L}\Big(\theta p_1 + \overline{\theta}p_2\Big)\right] \\
&= \frac{\lambda\mu_\theta^2}{2} + \frac{\lambda\alpha_\theta^2}{2} + \mathbb{E}\Big[H\Big(\alpha_\theta G + \mu_\theta Sf(S), \theta p_1 + \overline{\theta}p_2, \tau_\theta\Big) + \mathcal{L}\Big(\theta p_1 + \overline{\theta}p_2\Big)\Big] \\
&\leq \frac{\lambda\mu_\theta^2}{2} + \frac{\lambda\alpha_\theta^2}{2} + \mathbb{E}\Big[\theta H\Big(\alpha_1 G + \mu_1 Sf(S), p_1, \tau_1\Big) + \overline{\theta}H\Big(\alpha_2 G + \mu_2 Sf(S), p_2, \tau_2\Big) + \mathcal{L}\Big(\theta p_1 + \overline{\theta}p_2\Big)\Big].
\end{aligned} \tag{38}$$

The first inequality above follows by the definition of the Moreau envelope. The equality in the second line uses the definition of the function $H : \mathbb{R}^3 \to \mathbb{R}$ in (22). Finally, the last inequality follows from convexity of H as proved in Lemma A.1.

Continuing from (38), we use convexity of $\mathcal{L}$ to find that

$$\begin{aligned}
&\Omega(\theta\mathbf{w}_1 + \overline{\theta}\mathbf{w}_2) \leq \frac{\lambda\mu_\theta^2}{2} + \frac{\lambda\alpha_\theta^2}{2} + \\
&\mathbb{E}\Big[\theta H(\alpha_1 G + \mu_1 Sf(S), p_1, \tau_1) + \overline{\theta}H(\alpha_2 G + \mu_2 Sf(S), p_2, \tau_2) + \theta\,\mathcal{L}(p_1) + \overline{\theta}\mathcal{L}(p_2)\Big]
\end{aligned} \tag{39}$$

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

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

Thus proceeding from (39) we conclude strict-convexity of the function Ω :

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

This completes the proof of part (d).

Proof of Lemma B.1(e). Based on the proof of part (c) and under the assumptions of the lemma we have $\alpha^\star \neq 0$. Thus we see that the random variable $\alpha G + \mu S f(S)$ has a positive probability density everywhere in the desired domain of the optimization problem in (26). Next, we use the result in [Taheri et al., 2020, Proposition A.6], which states that if the random variable X has a positive density everywhere and $\mathcal{L}$ is continuously differentiable with $\mathcal{L}'(0) \neq 0$ then

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

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

B.3 From (26) to (10)

The following lemma connects the min-max optimization (26) to the system of equations in (10)

Lemma B.2 (Uniqueness of solutions to (10))**.** *Assume that the optimization problem in (26) yields a unique and bounded solution $(\alpha > 0, \mu, \upsilon > 0, \gamma > 0)$. Then the equations (10) have a unique and bounded solution $(\alpha > 0, \mu, \tau > 0)$ where $\tau = \upsilon/\gamma$.*

Proof. By direct differentiation with respect to the variables $(\mu, \alpha, \upsilon, \gamma)$, the first order optimality conditions of the min-max optimization in (26) are as follows:

$$\begin{aligned} \mathbb{E}\left[Sf(S)\mathcal{M}'_{\ell,1}\left(\alpha G + \mu S f(S); \frac{\upsilon}{\gamma}\right)\right] = -\lambda\mu,\ \lambda\alpha + \mathbb{E}\left[G\mathcal{M}'_{\ell,1}\left(\alpha G + \mu S f(S); \frac{\upsilon}{\gamma}\right)\right] = \frac{\gamma}{\sqrt{\delta}}, \\ \frac{1}{\gamma}\mathbb{E}\left[\mathcal{M}'_{\ell,2}\left(\alpha G + \mu S f(S); \frac{\upsilon}{\gamma}\right)\right] = -\frac{\gamma}{2},\ -\frac{\upsilon}{\gamma^2}\mathbb{E}\left[\mathcal{M}'_{\ell,2}\left(\alpha G + \mu S f(S); \frac{\upsilon}{\gamma}\right)\right] + \frac{\upsilon}{2} = \frac{\alpha}{\sqrt{\delta}}. \end{aligned} \tag{40}$$

Assumptions of the lemma imply that the saddle point of the optimization problem in (26) is unique and bounded, therefore (40) yields a unique bounded solution $(\alpha > 0, \mu, \upsilon > 0, \gamma > 0)$. By denoting $\tau = \upsilon/\gamma$ and using the fact that $\mathcal{M}'_{\mathcal{L},2}(x;\tau) = -\frac{1}{2}(\mathcal{M}'_{\mathcal{L},1}(x;\tau))^2$ (as implied by (16)-(17)) we reach the Equations (10) i.e.,

$$\mathbb{E}\Big[S\, f(S) \cdot \mathcal{M}'_{\mathcal{L},1}\left(\alpha G + \mu S f(S); \tau\right)\Big] = -\lambda\mu, \tag{41a}$$

$$\tau^2\, \delta \cdot \mathbb{E}\Big[\ \left(\mathcal{M}'_{\mathcal{L},1}\left(\alpha G + \mu S f(S); \tau\right)\right)^2\ \Big] = \alpha^2, \tag{41b}$$

$$\tau\, \delta \cdot \mathbb{E}\Big[G \cdot \mathcal{M}'_{\mathcal{L},1}\left(\alpha G + \mu S f(S); \tau\right)\Big] = \alpha(1 - \lambda\tau\delta). \tag{41c}$$

The uniqueness of $(\alpha > 0, \mu, \tau > 0)$ as the solution to (41) follows from the uniqueness of the solution $(\alpha > 0, \mu, \upsilon > 0, \gamma > 0)$ to (40). In particular if there are two distinct solutions $(\alpha_1, \mu_1, \tau_1)$ and $(\alpha_2, \mu_2, \tau_2)$ to the Equations (41), then we reach contradiction by noting that $(\alpha_1, \mu_1, \upsilon_1 := \alpha_1/\sqrt{\delta}, \gamma_1 := \alpha_1/(\tau_1\sqrt{\delta}))$ and $(\alpha_2, \mu_2, \upsilon_2 := \alpha_2/\sqrt{\delta}, \gamma_2 := \alpha_2/(\tau_2\sqrt{\delta}))$ are two distinct points satisfying the Equations (40). This completes the proof of the lemma. □

B.4 Completing the proof of Theorem 3.1

We are now ready to complete the proof of Theorem 3.1. Based on Lemma B.2, for the system of equations in (10) to have a unique and bounded solution, it suffices that $(\alpha^\star > 0, \mu^\star, \upsilon^\star > 0, \gamma^\star > 0)$ as the solution of (26) is unique and bounded. Since Θ is convex-concave and the optimality sets are bounded from Lemma B.1(a)-(e), a saddle point of Θ exists [Rockafellar, 1997, Cor. 37.3.2]. Additionally, based on the assumptions of the theorem and in view of Lemma B.1(d),(e), Θ is jointly strictly-convex in (α, μ, υ) and strictly-concave in γ which implies the uniqueness of $(\alpha^\star > 0, \mu^\star, \upsilon^\star > 0, \gamma^\star > 0)$ as a solution to (26). This completes the proof of the theorem.

As mentioned in the main body of the paper, we conjecture that some of the technical conditions of Theorem 3.1, albeit mild in their current form, can be relaxed even further. Refining these conditions can be an interesting topic of future work, but is out of the scope of this paper. We mention in passing that the conclusions of Theorem

3.1 also hold true if we replace the two-times differentiability condition by an assumption that the loss is one-time differentiable and strictly convex.

C Fundamental Limits for Linear Models: Proofs for Section 2

C.1 Auxiliary Results

Lemma C.1 (Boundedness of τ in (45))**.** *Let $\mathcal{L}(\cdot)$ be a non-linear, convex and twice differentiable function, $\lambda > 0$ and $\delta > 0$ and the pair (α, τ) be a solution to (3) where $\alpha > 0$. Then, $0 < \tau < \frac{1}{\lambda\delta}$.*

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

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

Therefore the equation in the LHS in (3) is equivalent to

$$\tau\delta\, \mathbb{E}\Big[\mathcal{M}''_{\mathcal{L},1}\left(\alpha\, G + Z; \tau\right)\Big] = 1 - \lambda\tau\delta. \tag{42}$$

Next we prove that under the assumptions of the lemma, $\mathbb{E}\Big[\mathcal{M}''_{\mathcal{L},1}\left(\alpha G + \mu S f(S); \tau\right)\Big]$ is positive. First using properties of Morea-envelopes in (20), we have

$$\mathbb{E}\Big[\mathcal{M}''_{\mathcal{L},1}\left(\alpha\, G + Z; \tau\right)\Big] = \mathbb{E}\left[\frac{\mathcal{L}''(\mathrm{prox}_{\mathcal{L}}\left(\alpha\, G + Z; \tau\right))}{1 + \tau\mathcal{L}''(\mathrm{prox}_{\mathcal{L}}\left(\alpha G + Z; \tau\right))}\right] \geq 0. \tag{43}$$

In particular, we see that equality is achieved in (43) is achieved if and only if

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

Or equivalently,

$$\exists\, c_1, c_2 \in \mathbb{R} : \text{s.t. } \forall x \in \mathbb{R} : \mathcal{M}_{\mathcal{L}}\left(x; \tau\right) = c_1 x + c_2. \tag{44}$$

Finally, using Proposition A.2 to "invert" the Moreau envelope function, we find that the loss function $\mathcal{L}(\cdot)$ satisfying (44) is such that

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

where $I(\cdot)$ is the identity function i.e. $I(t) = t$, $\forall t \in \mathbb{R}$. But according to the assumptions of the lemma, $\mathcal{L}$ is a *non-linear* convex function. Thus, it must hold that $\mathbb{E}\left[\mathcal{M}''_{\mathcal{L},1}\left(\alpha\, G + Z; \tau\right)\right] > 0$. Using this and the assumptions on λ and δ, the advertised claim follows directly from (42). □

C.2 Proof of Theorem 2.1

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

$$\delta\tau^2 \cdot \mathbb{E}\Big[\Big(\mathcal{M}'_{\mathcal{L},1}\left(\alpha\, G + Z; \tau\right)\Big)^2\Big] = \alpha^2 - \lambda^2\delta^2\tau^2, \tag{45a}$$

$$\delta\tau \cdot \mathbb{E}\Big[G \cdot \mathcal{M}'_{\mathcal{L},1}\left(\alpha\, G + Z; \tau\right)\Big] = \alpha\,(1 - \lambda\delta\tau). \tag{45b}$$

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

$$\Psi(a, x) := \frac{(a^2 - x^2\,\delta^2)\,\mathcal{I}(V_a)}{(1 - x\,\delta)^2}. \tag{46}$$

Then, $\alpha_\star > 0$ as in (5) is equivalently expressed as

$$\alpha_\star := \min_{0\le x<1/\delta} \left\{ a \ge 0 : \ \Psi(a,x) = \frac{1}{\delta} \right\}. \tag{47}$$

Before everything, let us show that $\alpha_\star$ is well-defined, i.e., that the feasible set of the minimization in (47) is non-empty for all $\delta > 0$ and random variables Z satisfying Assumption 3. Specifically, we will show that there exists $a \ge 0$ such that $\Psi\left(a, \frac{a}{(1+a)\delta}\right) = 1/\delta$. It suffices to prove that the range of the function $\widetilde{\Psi}(a) := \Psi\left(a, \frac{a}{(1+a)\delta}\right)$ is $(0,\infty)$. Clearly, the function $\widetilde{\Psi}$ is continuous in $\mathbb{R}_{\ge 0}$. Moreover, it can be checked that $\widetilde{\Psi}(a) = (a^2+2a)\Psi_0(a)$ where $\Psi_0(a) := a^2\mathcal{I}(V_a)$. By Lemma A.2, $\lim_{a\to 0}\Psi_0(a) = 0$ and $\lim_{a\to+\infty}\Psi_0(a) = 1$. Hence, we find that $\lim_{a\to 0}\widetilde{\Psi}(a) = 0$ and $\lim_{a\to+\infty}\widetilde{\Psi}(a) = +\infty$, as desired.

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

$$\alpha \ge \alpha_\star. \tag{48}$$

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

$$\begin{aligned}
\mathbb{E}\Big[G\cdot \mathcal{M}'_{\mathcal{L},1}\left(V_\alpha;\tau\right)\Big] &= -\alpha \iint \mathcal{M}'_{\mathcal{L},1}\left(u+z;\tau\right)\phi'_\alpha(u)p_Z(z)\mathrm{d}u\mathrm{d}z \\
&= -\alpha \iint \mathcal{M}'_{\mathcal{L},1}\left(v;\tau\right)\phi'_\alpha(u)p_Z(v-u)\mathrm{d}u\mathrm{d}v \\
&= -\alpha \int \mathcal{M}'_{\mathcal{L},1}\left(v;\tau\right)p'_V(v)\mathrm{d}v = -\alpha\,\mathbb{E}\Big[\mathcal{M}'_{\mathcal{L},1}\left(V_\alpha;\tau\right)\cdot\xi_{V_\alpha}(V_\alpha)\Big],
\end{aligned} \tag{49}$$

where for a random variable V, we denote its score function with $\xi_V(v) := p'_V(v)/p_V(v)$ for $v\in\mathbb{R}$. Using (49) and $\alpha > 0$, (45b) can be equivalently written as following,

$$1-\lambda\,\delta\,\tau = -\delta\tau\cdot\mathbb{E}\Big[\mathcal{M}'_{\mathcal{L},1}\left(V_\alpha;\tau\right)\cdot\xi_{V_\alpha}(V_\alpha)\Big]. \tag{50}$$

Next, by applying Cauchy-Shwarz inequality, recalling $\mathbb{E}[(\xi_{V_\alpha}(V_\alpha))^2] = \mathcal{I}\left(V_\alpha\right)$ and using (45a), we have that

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

where we have also used the fact that $\tau > 0$. To continue, we use (50) to rewrite the LHS above and deduce that:

$$\left(\frac{1-\lambda\,\delta\,\tau}{\delta\tau}\right)^2 \le \frac{(\alpha^2-\lambda^2\,\delta^2\,\tau^2)\,\mathcal{I}(V_\alpha)}{\delta\tau^2}. \tag{51}$$

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

$$\frac{(\alpha^2-\lambda^2\,\delta^2\,\tau^2)\,\mathcal{I}(V_\alpha)}{(1-\lambda\,\delta\,\tau)^2} \ge \frac{1}{\delta}. \tag{52}$$

In the remaining, we use (52) to prove (48). For the sake of contradiction to (48), assume that there exists a valid triplet (α,λ,τ) such that $\alpha < \alpha_\star$. Recall by inequality (52) that α satisfies:

$$\Psi\Big(\alpha\,,\,\lambda\,\tau\Big) \ge \frac{1}{\delta}. \tag{53}$$

We show first that (53) holds with strict inequality. To see this, suppose that $\Psi(\alpha,\lambda\,\tau) = 1/\delta$. From Lemma C.1, it also holds that $\lambda\,\tau \in (0,1/\delta)$. Hence, the pair $(\alpha,\lambda\,\tau)$ is a feasible point in the minimization in (47). Combining this with optimality of $\alpha_\star$ lead to the conclusion that $\alpha_\star \ge \alpha$, which contradicts our assumption

$\alpha < \alpha_\star$. Therefore we consider only the case where (53) holds with strict inequality i.e., $\Psi(\alpha, \lambda\tau) > 1/\delta$.

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

$$\exists\, \widetilde{\alpha}: \text{ s.t. } 0 \leq \widetilde{\alpha} < \alpha, \text{ and } \Psi\left(\widetilde{\alpha}\,,\, \lambda\,\tau\right) = \frac{1}{\delta}. \tag{54}$$

By recalling our assumption that $\alpha < \alpha_\star$, we can deduce that (54) in fact holds for $\widetilde{\alpha} < \alpha_\star$. However, this is in contradiction with the optimality of $\alpha_\star$ defined in (47). This shows that for all achievable α it must hold that $\alpha \geq \alpha_\star$. This proves the claim in (48) and completes the proof of the theorem.

C.3 Proof of Lemma 2.1

To prove the claim of the lemma, it suffices to show that the proposed loss function and regularization parameter, satisfy the system of equations in (45) with $\alpha = \alpha_\star$. For this purpose we show that $(\mathcal{L}, \lambda, \alpha, \tau) = (\mathcal{L}_\star, \lambda_\star, \alpha_\star, 1)$ satisfy (45).

First, we recognize that for the candidate optimal loss function in Lemma 3.1 we have $\forall v \in \mathbb{R}$ that

$$\mathcal{M}'_{\mathcal{L}_\star,1}(v; 1) = -\frac{\alpha_\star^2 - \lambda_\star^2\,\delta^2}{1 - \lambda_\star \delta} \cdot \xi_{V_\star}(v). \tag{55}$$

Thus by replacing the proposed parameters in (45a) we have :

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

where for the last line we used the definitions of $\alpha_\star$ and $\lambda_\star$ in the statement of the lemma. This proves the claim for (45a). To show that Equation (45b) is satisfied we use its equivalent expression in (50) and also replace (55) in (50). Specifically, this shows that

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

from which we conclude that Equation (45b) is satisfied. This completes the first part of the proof.

Next, we discuss sufficient conditions on convexity on $\mathcal{L}_\star$. Numerical illustrations (e.g., Figure 3), show that when Z has a Laplace distribution, $\mathcal{L}_\star$ is convex. Generally, we conjecture that $\mathcal{L}_\star$ is convex when Z has a log-concave probability density. In the next step, we derive sufficient conditions guaranteeing convexity of $\mathcal{L}_\star$. Our method to providing the sufficient conditions is analogous to the case of unregularized optimal loss studied in [Bean et al., 2013], in which it is shown that the log-concavity of P_Z is a sufficient condition for the convexity of unregularized optimal loss function. However as the next lemma shows, for our studied case of regularized optimal loss, we need a condition on $\lambda_\star, \alpha_\star, \delta$ to guarantee convexity of $\mathcal{L}_\star$. We denote

$$c := \left(\alpha_\star^2 - \lambda_\star^2\delta^2\right)/(1 - \lambda_\star\delta)\,, \quad h(x) := c \cdot \log P_{V_\star}(x).$$

Since $(\alpha_\star, \lambda_\star)$ satisfy (5), we straightforwardly see that $c > 0$. Furthermore, using Proposition A.1(c) and the formula of $\mathcal{L}_\star$ (as in Lemma 2.1) we find that second derivative of $\mathcal{L}_\star$ takes the following shape,

$$\mathcal{L}''_\star(x) = -\frac{h''\left(\mathrm{prox}_h(x; 1)\right)}{\left(1 + h''(\mathrm{prox}_h(x; 1))\right)}.$$

Note that if P_Z is log-concave then $V_\star = \alpha_\star G + Z$ also has a log-concave density, therefore to prove that $\mathcal{L}''_\star > 0$

it is sufficient to show that $h''(\cdot) > -1$. Some algebra shows that

$$h''(x) = -c/\alpha_\star^2 + c \cdot \left(\log \int \exp((2xy - y^2)/(2\alpha_\star^2)) P_Z(y) dy \right)'',$$

where by computing the derivative and using the Cauchy-Schwarz inequality the second term can be straightforwardly proved to be positive. Thus, $h''(\cdot) > -1$ provided that $c < \alpha_\star^2$ or equivalently $\alpha_\star^2 < \lambda_\star \delta$, as claimed. This completes the proof of the lemma.

C.4 Proof of Lemma 2.2

By letting $\mathcal{L}(t) = t^2$ we find that $\mathcal{M}_{\mathcal{L}}(x;\tau) = \frac{x^2}{2\tau+1}$ for all $x \in \mathbb{R}$ and $\tau \in \mathbb{R}_{>0}$. Using this in Equations (45) and a after a few algebraic simplifications we arrive at the following closed-form expression for $\alpha^2_{\ell_2,\lambda}$ for all $\lambda \geq 0$ and random variables Z with finite second moment,

$$\alpha^2_{\ell_2,\lambda} = \frac{1}{2}\left(1 - \mathbb{E}[Z^2] - \delta\right) + \frac{\mathbb{E}[Z^2](\lambda + 2\delta + 2) + 2(\delta - 1)^2 + \lambda(\delta + 1)}{2\sqrt{(\lambda + 2\delta - 2)^2 + 8\lambda}}. \tag{56}$$

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

C.5 Proof of Corollary 2.1

As mentioned in the main body of the paper, the difficulty in deriving a closed-form expression for $\alpha_\star$ in (5) is due to the fact that in general $\mathcal{I}(V_a) = \mathcal{I}(aG + Z)$ may not be expressible in closed-form with respect to a. The core idea behind this corollary is using Stam's inequality (see Proposition A.3) to bound $\mathcal{I}(V_a)$ in terms of $\mathcal{I}(aG) = a^{-2}$ and $\mathcal{I}(Z)$. Specifically, applying (23) to the random variables $a\,G$ and Z we find that:

$$\mathcal{I}(V_a) = \mathcal{I}(a\,G + Z) \leq \frac{\mathcal{I}(Z)}{1 + a^2\mathcal{I}(Z)}. \tag{57}$$

Substituting the RHS above in place of $\mathcal{I}(V_a)$ in the definition of $\alpha_\star$ in (5), let us define $\widehat{\alpha}$ as follows:

$$\widehat{\alpha} := \min_{0 \leq x < 1/\delta} \left\{ a \geq 0 : \frac{(a^2 - x^2\,\delta^2)\,\mathcal{I}(Z)}{(1 - x\,\delta)^2(1 + a^2\mathcal{I}(Z))} \geq \frac{1}{\delta} \right\}. \tag{58}$$

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

$$\alpha_\star^2 \geq \widehat{\alpha}^2. \tag{59}$$

Second, we solve the minimization in (58) to yield a closed-form expression for $\widehat{\alpha}$.

Towards proving (59), note from the definition of $\alpha_\star$ and inequality (57) that there exists $x_\star \in [0, 1/\delta)$ such that

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

Thus, the pair $(\alpha_\star, x_\star)$ is feasible in (58). This and optimality of $\widehat{\alpha}$ in (58) lead to (59), as desired.

The next step is finding a closed-form expression for $\widehat{\alpha}$. Based on (58) and few algebraic simplifications we have :

$$\begin{aligned}\widehat{\alpha}^2 &= \min_{0\le x<1/\delta}\left\{a^2 : a^2\,\mathcal{I}(Z)\cdot(\delta-(1-x\,\delta)^2)\ge(1-x\,\delta)^2+\delta^3x^2\mathcal{I}(Z)\right\}\\ &= \min_{\max\{0,\frac{1-\sqrt{\delta}}{\delta}\}\le x<1/\delta}\left\{a^2 : a^2\ge\frac{(1-x\delta)^2+\delta^3\,x^2\,\mathcal{I}(Z)}{\mathcal{I}(Z)\cdot(\delta-(1-x\delta)^2)}\right\}\\ &= \min_{\max\{0,\frac{1-\sqrt{\delta}}{\delta}\}\le x<1/\delta}\left\{\frac{(1-x\delta)^2+\delta^3\,x^2\,\mathcal{I}(Z)}{\mathcal{I}(Z)\cdot(\delta-(1-x\delta)^2)}\right\}. \end{aligned}\tag{60}$$

The last equality above is true because the fraction in the constraint in the second line is independent of a. Next, by minimizing with respect to the variable x in (60), we reach $\widehat{\alpha}^2 = h_\delta(1/\mathcal{I}(Z))$.

Finally, we know from Proposition A.3(f) that equality in (57) is achieved if and only if the noise is Gaussian i.e. $Z\sim\mathcal{N}(0,\zeta^2)$ for some $\zeta>0$. Thus, if this is indeed the case, then $\alpha_\star=\widehat{\alpha}$ and the lower bound is achieved with replacing the Fisher information of Z i.e, $\mathcal{I}(Z)=\zeta^{-2}$. This completes the proof of the corollary.

C.6 Proof of Equation (7)

First, we prove the bound $\omega_{_\delta}\ge\left(\mathcal{I}(Z)\,\mathbb{E}[Z^2]\right)^{-1}$. Fix $\delta>0$ and consider the function $\widetilde{h}_\delta(x):=h_\delta(x)/x$ for $x\ge0$. Direct differentiation and some algebra steps suffice to show that $\widetilde{h}_\delta(x)$ is decreasing. Using this and the fact that $1/\mathcal{I}(Z)\le\mathbb{E}[Z^2]$ (cf. Proposition A.3 (c)), we conclude with the desired.

Next, we prove the lower bound $\omega_\delta\ge1-\delta$. Fix any $\delta>0$. First, it is straightforward to compute that $h_\delta(0)=\max\{1-\delta,0\}\ge1-\delta$. Also, simple algebra shows that $h_\delta(x)\le1, x\ge0$. From these two and the increasing nature of $h_\delta(x)$ we conclude that $1-\delta\le h_\delta(x)\le1$, for all $x\ge0$. The desired lower bound follows immediately by applying these bounds to the definition of ω_δ.

D Fundametal Limits for Binary Models: Proofs for Section 3

D.1 Discussion on Assumption 4

As per Assumption 4, the link function must satisfy $\mathbb{E}[Sf(S)]\neq0$. This is a rather mild assumption in our setting. For example, it is straightforward to show that it is satisfied for the Signed, Logistic and Probit models. More generally, for a link function $f:\mathbb{R}\to\{\pm1\}$ and $S\sim\mathcal{N}(0,1)$, the probability density of $Sf(S)$ can be computed as follows for any $x\in\mathbb{R}$:

$$P_{Sf(S)}(x)=\Big(1+\widehat{f}(x)-\widehat{f}(-x)\Big)\frac{\exp(-x^2/2)}{\sqrt{2\pi}},\qquad \widehat{f}(x):=\mathbb{P}\left(f(x)=1\right).\tag{61}$$

From this and the fact that $\exp(-x^2/2)$ is an even function of x, we can conclude that Assumption 4 is valid if $\widehat{f}(x)$ is monotonic and non-constant based on x (e.g., as in the Signed, Logistic and Probit models) . In contrast, Assumption 4 fails if the function $\widehat{f}$ is even. Finally, we remark that using (61), it can be checked that $S\,f(S)\sim\mathcal{N}(\mu,\zeta^2)$ if and only if $(\mu,\zeta)=(0,1)$, and consequently only if $\widehat{f}$ is an even function. Based on these, we conclude that for all link functions f satisfying Assumption 4, the resulting distribution of $Sf(S)$ is non-Gaussian. Finally, we remark that $\nu_f=\mathbb{E}[Sf(S)]$ is the first Hermite coefficient of the function f and the requirement $\nu_f\neq0$ arises in a series of recent works on high-dimensional single-index models, e.g., [Plan and Vershynin, 2015, Genzel, 2016]; see also [Mondelli and Montanari, 2017, Lu and Li, 2017] for algorithms specializing to scenarios in which $\nu_f=0$.

D.2 Discussion on the Classification Error (11)

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

For the estimator $\widehat{\mathbf{w}}_{\mathcal{L},\lambda}$ obtained from (8), and $\mathbf{x}_0$ denoting the true vector with unit norm, the parameters $\mu_{\mathcal{L},\lambda}$ and $\alpha_{\mathcal{L},\lambda}$ denote the high-dimensional terms of bias and variance,

$$\mathbf{x}_0^T \widehat{\mathbf{w}}_{\mathcal{L},\lambda} \xrightarrow{P} \mu_{\mathcal{L},\lambda}, \tag{62}$$

$$\|\widehat{\mathbf{w}}_{\mathcal{L},\lambda} - \mu_{\mathcal{L},\lambda}\, \mathbf{x}_0\|_2^2 \xrightarrow{P} \alpha_{\mathcal{L},\lambda}^2. \tag{63}$$

We note that by rotational invariance of Gaussian distribution we may assume without loss of generality that $\mathbf{x}_0 = [1,\, 0,\, 0,\, \cdots,\, 0]^T \in \mathbb{R}^n$. Therefore we deduce from (62) and (63) that

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

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

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

Recalling Assumption 2 we have $\mathbf{a} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$. Thus by denoting $S, G \overset{\text{iid}}{\sim} \mathcal{N}(0,1)$ and assuming without loss of generality that $\mu_{\mathcal{L},\lambda} > 0$, we derive (11).

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

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

This shows the desired. Importantly, we remark that in view of (61), this condition on the density of $Sf(S)$ is satisfied for many well-known binary models including Logistic, Probit and Signed.

D.3 Proof of Theorem 3.2

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

Lemma D.1 (Boundedness of τ in (41))**.** *Fix $\delta > 0$ and $\lambda > 0$ and let $\mathcal{L}$ be a convex, twice differentiable and non-linear function. Then all solutions τ of the system of equations in (41) satisfy $0 < \tau < \frac{1}{\lambda\delta}$.*

Proof. The proof follows directly from the proof of Lemma C.1 by replacing Z with $\mu\, Sf(S)$. Note that the Equation (41c) can be obtained by replacing Z with $\mu Sf(S)$ in Equation (45b). □

Next, we proceed to the proof main of Theorem 3.2. For convenience, let us define the function $\Phi : \mathbb{R}_{\geq 0} \times [0, 1/\delta) \to \mathbb{R}$ as following :

$$\Phi(s,x) := \frac{1 - s^2(1 - s^2\mathcal{I}(W_s))}{\delta s^2(s^2\mathcal{I}(W_s) + \mathcal{I}(W_s) - 1)} - 2x + x^2\delta(1 + s^{-2}). \tag{64}$$

Then, $\sigma_\star$ as in (12) is equivalently expressed as:

$$\sigma_\star := \min_{0 \leq x < 1/\delta} \left\{ s \geq 0 :\ \Phi(s,x) = 1 \right\}, \tag{65}$$

Before everything, we show that $\sigma_\star$ is well defined, i.e., the feasible set of the minimization in (65) is non-empty for all $\delta > 0$ and link functions $f(\cdot)$ satisfying Assumption 4. Specifically, we will show that for any $\delta > 0$ there exists $s \geq 0$ such that $\widetilde{\Phi} := \Phi(s, \frac{s}{\delta(1+s)}) = 1$. It suffices to prove that the range of the function $\widetilde{\Phi}$ is $(0, \infty)$. Clearly, the function is continuous in $\mathbb{R}_{\geq 0}$. Moreover, it can be checked that

$$\widetilde{\Phi}(s) = \Phi_0(s) + \frac{2}{\delta(1+s^2)}, \quad \text{where} \quad \Phi_0(s) := \frac{1 - s^2\mathcal{I}(W_s)}{\delta\, s^2(s^2\mathcal{I}(W_s) + \mathcal{I}(W_s) - 1)}. \tag{66}$$

But, by Lemma A.2, $\lim_{s\to 0} \; s^2\,\mathcal{I}(W_s) = 0$ and $\lim_{s\to+\infty} s^2\,\mathcal{I}(W_s) = 1$. Using these, we can show that $\lim_{s\to 0} \; \Phi_0(s) = +\infty$ and $\lim_{s\to+\infty} \; \Phi_0(s) = 0$. Combined with (66), we find that $\lim_{s\to 0} \; \Phi_0(s) = +\infty$ and $\lim_{s\to+\infty} \; s^2\,\mathcal{I}(W_s) = 0$. This concludes the proof of feasibility of the minimization in (64).

We are now ready to prove the main claim of the theorem. Fix convex loss function $\mathcal{L}$ and regularization parameter $\lambda \geq$. Let $(\alpha > 0, \mu, \tau > 0)$ be the unique solution to (41) and denote $\sigma = \alpha/\mu$. We will prove that

$$\sigma \geq \sigma_\star. \tag{67}$$

The first step in the proof will be to transform the equations (41) in a more appropriate form. In order to motivate the transformation, note that the performance of the optimization problem in (8) is unique up to rescaling. In particular consider the following variant of the optimization problem in (8) :

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

It is straightforward to see that, regardless of the values of c_1 and c_2, $\mathrm{corr}\left(\widehat{\mathbf{w}}_{\mathcal{L},\lambda}\,,\, \mathbf{x}_0\right) = \mathrm{corr}\left(\widehat{\mathbf{v}}_{\mathcal{L},\lambda}\,,\, \mathbf{x}_0\right)$, where recall that $\widehat{\mathbf{w}}_{\mathcal{L},\lambda}$ solves (8). Thus in view of (9), we see that the error σ resulting from $\widehat{\mathbf{w}}_{\mathcal{L},\lambda}$ and $\widehat{\mathbf{v}}_{\mathcal{L},\lambda}$ are the same. Motivated by this observation, we consider the following rescaling for the loss function and regularization parameter:

$$\widetilde{\mathcal{L}}(\cdot) := \frac{\tau}{\mu^2}\, \mathcal{L}(\mu\,\cdot), \qquad \widetilde{\lambda} := \tau\lambda, \tag{68}$$

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

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

Using these transformations, we can rewrite the system of equations (41) in terms of σ, $\widetilde{\mathcal{L}}$ and $\widetilde{\lambda}$ as follows:

$$\mathbb{E}\Big[S f(S) \cdot \mathcal{M}'_{\widetilde{\mathcal{L}},1}\left(W_\sigma; 1\right) \Big] = -\widetilde{\lambda}, \tag{69a}$$

$$\mathbb{E}\Big[\Big(\mathcal{M}'_{\widetilde{\mathcal{L}},1}\left(W_\sigma; 1\right) \Big)^2 \Big] = \sigma^2/\delta, \tag{69b}$$

$$\mathbb{E}\Big[G \cdot \mathcal{M}'_{\widetilde{\mathcal{L}},1}\left(W_\sigma; 1\right) \Big] = \sigma(1 - \widetilde{\lambda}\delta)/\delta. \tag{69c}$$

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

Next, we further simplify (69) as follows. Similar to the procedure leading to (49), here also we may deduce that,

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

Thus (69c) can be rewritten as

$$\mathbb{E}\Big[\xi_{W_\sigma}(W_\sigma) \cdot \mathcal{M}'_{\widetilde{\mathcal{L}},1}\left(W_\sigma; 1\right) \Big] = (\widetilde{\lambda}\delta - 1)/\delta. \tag{70}$$

Additionally, we linearly combine (69a) and (69c) (with coefficient σ) to yield :

$$\mathbb{E}\Big[W_\sigma \cdot \mathcal{M}'_{\widetilde{\mathcal{L}},1}\left(W_\sigma;1\right)\Big] = \sigma^2/\delta - \sigma^2\widetilde{\lambda} - \widetilde{\lambda}, \tag{71}$$

Putting together (69b), (70) and (71), we have shown that σ satisfies the following system of equations:

$$\mathbb{E}\Big[W_\sigma \cdot \mathcal{M}'_{\widetilde{\mathcal{L}},1}\left(W_\sigma;1\right)\Big] = \frac{\sigma^2}{\delta} - \sigma^2\widetilde{\lambda} - \widetilde{\lambda}, \tag{72a}$$

$$\mathbb{E}\Big[\Big(\mathcal{M}'_{\widetilde{\mathcal{L}},1}\left(W_\sigma;1\right)\Big)^2\Big] = \frac{\sigma^2}{\delta}, \tag{72b}$$

$$\mathbb{E}\Big[\xi_{W_\sigma}(W_\sigma) \cdot \mathcal{M}'_{\widetilde{\mathcal{L}},1}\left(W_\sigma;1\right)\Big] = \widetilde{\lambda} - \frac{1}{\delta}. \tag{72c}$$

Next, we will use this fact to derive a lower bound on σ. To this end, let $\beta_1, \beta_2 \in \mathbb{R}$ be two real constants. By combining (72a) and (72c) we find that

$$\mathbb{E}\Big[(\beta_1 W_\sigma + \beta_2 \xi_{W_\sigma}(W_\sigma)) \cdot \mathcal{M}'_{\widetilde{\mathcal{L}},1}\left(W_\sigma;1\right)\Big] = \beta_1(\frac{\sigma^2}{\delta} - \sigma^2\widetilde{\lambda} - \widetilde{\lambda}) + \beta_2(\widetilde{\lambda} - \frac{1}{\delta}). \tag{73}$$

Applying Cauchy-Schwarz inequality to the LHS of (73) gives :

$$\begin{aligned}\left(\beta_1\left(\frac{\sigma^2}{\delta} - \sigma^2\widetilde{\lambda} - \widetilde{\lambda}\right) + \beta_2(\widetilde{\lambda} - \frac{1}{\delta})\right)^2 &\le \mathbb{E}\Big[\left(\beta_1 W_\sigma + \beta_2\xi_{W_\sigma}(W_\sigma)\right)\big)^2\Big] \cdot \mathbb{E}\Big[\Big(\mathcal{M}'_{\widetilde{\mathcal{L}},1}\left(W_\sigma;1\right)\Big)^2\Big] \\ &= \mathbb{E}\Big[\left(\beta_1 W_\sigma + \beta_2\xi_{W_\sigma}(W_\sigma)\right)\big)^2\Big]\frac{\sigma^2}{\delta},\end{aligned} \tag{74}$$

where we used (72b) in the last line. To simplify the expectation in the RHS of (74), we use the facts that $\mathbb{E}[W_\sigma^2] = \sigma^2 + 1$ and $\mathbb{E}[(\xi_{W_\sigma}(W_\sigma))^2] = \mathcal{I}(W_\sigma)$. Also by integration by parts one can derive that $\mathbb{E}[W_\sigma \cdot \xi_{W_\sigma}(W_\sigma)] = -1$. Thus we arrive at the following inequality from (74):

$$\left(\beta_1\left(\sigma^2/\delta - \sigma^2\widetilde{\lambda} - \widetilde{\lambda}\right) + \beta_2(\widetilde{\lambda} - 1/\delta)\right)^2 \le \beta_1^2\left(\sigma^2+1\right) + \beta_2^2\,\mathcal{I}(W_\sigma) - 2\beta_1\beta_2. \tag{75}$$

Now, we choose the coefficients β_1 and β_2 as follows: $\beta_1 = 1 - \widetilde{\lambda}\delta - (\sigma^2 - \sigma^2\widetilde{\lambda}\delta - \widetilde{\lambda}\delta)\,\mathcal{I}(W_\sigma)$ and $\beta_2 = 1$. (We show later in Theorem 3.1, that this choice lead to an achievable lower bound). Substituting these values in (75) and simplifying the resulting expressions yield the following inequality for σ:

$$\frac{1 - \sigma^2(1 - \sigma^2\mathcal{I}(W_\sigma))}{\delta\sigma^2(\sigma^2\mathcal{I}(W_\sigma) + \mathcal{I}(W_\sigma) - 1)} - 2\widetilde{\lambda} + \widetilde{\lambda}^2\delta(1 + \sigma^{-2}) \le 1. \tag{76}$$

We will now finish the proof of the theorem by using (76) to prove (67). For the sake of contradiction to (67), assume that $\sigma < \sigma_\star$. From (76) and the notation introduced in (64), we have shown that $\Phi\left(\sigma, \widetilde{\lambda}\right) \le 1$. Recall from (68) that $\widetilde{\lambda} = \lambda\tau$. But, from Lemma D.1 it holds that $\widetilde{\lambda} = \lambda\tau < \frac{1}{\delta}$. Therefore, the pair $(\sigma, \widetilde{\lambda})$ is feasible in the minimization problem in (65). By this, optimality of $\sigma_\star$ and our assumption that $\sigma < \sigma_\star$ in (65) it must hold that $\Phi\left(\sigma, \widetilde{\lambda}\right) < 1$. But then, since $\lim_{s\to 0}\Phi\left(s, \widetilde{\lambda}\right) = +\infty$ and by continuity of the function $\Phi(\cdot, x)$ for all fixed $x \in [0, 1/\delta)$, we have:

$$\exists\, \sigma_1 : \text{ s.t. } 0 < \sigma_1 < \sigma, \text{ and } \Phi(\sigma_1, \widetilde{\lambda}) = 1. \tag{77}$$

Therefore $\Phi(\sigma_1, \widetilde{\lambda}) = 1$ for $\sigma_1 < \sigma_\star$, which contradicts the optimality of $\sigma_\star$ in (65) and completes the proof.

D.4 Proof of Lemma 3.1

To prove the claim of the lemma we show that the proposed candidate-optimal loss and regularization parameter pair $(\mathcal{L}_\star, \lambda_\star)$ satisfies the system of equations in (41) with $(\alpha, \mu, \tau) = (\sigma_\star, 1, 1)$. In line with the proof of Theorem 3.2 and the equivalent representation of (72) for the equations in (41), we show that $(\mathcal{L}_\star, \lambda_\star)$ satisfy all three equations in (72) with $(\sigma, \mu, \tau) = (\sigma_\star, 1, 1)$. We emphasize that since $\mu = \tau = 1$, based on (68) the $\mathcal{L}_\star$ and $\lambda_\star$ remain the same under these changes of parameters thus $(\widetilde{\mathcal{L}}_\star(\cdot), \widetilde{\lambda}_\star) = (\mathcal{L}_\star, \lambda_\star)$.

Note that we need $\mathcal{M}_{\mathcal{L}}(\cdot)$ to be able to assess the equations in (72). For this purpose we use inverse properties of Moreau-envelope functions in Proposition A.2 to derive the following from the definition of $\mathcal{L}_\star$ in (13) :

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

Thus,

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

Using this and the fact that $\mathbb{E}[W_\star \cdot \xi_{W_\star}(W_\star)] = -1$ (derived by integration by parts), the LHS of the equation (72a) changes to

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

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

Similarly, for the second equation (72b), we begin with replacing the expression for $\mathcal{M}'_{\mathcal{L}_\star,1}(W_\star;1)$ to see that

$$\begin{aligned}\mathbb{E}\left[\left(\mathcal{M}'_{\mathcal{L}_\star,1}(W_\star;1)\right)^2\right] &= \frac{(\lambda_\star\delta - 1)^2}{\delta^2(\eta - \mathcal{I}(W_\star))^2}\left(\eta^2\,\mathbb{E}\left[W_\star^2\right] + \mathcal{I}(W_\star) + 2\eta\,\mathbb{E}\left[W_\star \cdot \xi_{W_\star}(W_\star)\right]\right) \\ &= \frac{(\lambda_\star\delta - 1)^2}{\delta^2(\eta - \mathcal{I}(W_\star))^2}\left(\eta^2(\sigma_\star^2 + 1) + \mathcal{I}(W_\star) - 2\eta\right). \qquad (78)\end{aligned}$$

After replacing η, we can simplify (78) to reach the following

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

where the last two steps follow from the definition of $\sigma_\star$ in (12) and $\Phi(\cdot,\cdot)$ in (64).

For the third Equation (72c) we deduce in a similar way that

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

confirming the RHS of Equation (72c). This completes the proof.

D.5 Proof of Lemma 3.2

Let $\ell_2(t) = (1-t)^2$ for $t \in \mathbb{R}$. Using the Equations in (41) and replacing $\mathcal{M}_{\ell_2}(x;\tau) = \frac{(x-1)^2}{2\tau+1}$ we can solve the equations to find the closed-form formulas for (μ, α, τ) for a fixed $\lambda \geq 0$. For compactness, define $F(\cdot,\cdot) : \mathbb{R}_{>0} \times \mathbb{R}_{>0} \to \mathbb{R}_{>0}$ where $F(\delta,\lambda) := \lambda\delta + \sqrt{8\lambda\delta + (\delta(\lambda+2)-2)^2}$. We derive the following for $\mu_{\ell_2,\lambda}$ and $\alpha_{\ell_2,\lambda}$ and for all $\delta > 0$,

$$\mu_{\ell_2,\lambda} = \frac{4\delta\,\mathbb{E}[Z^2]}{2+2\delta+F(\delta,\lambda)},$$

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

Using these, we reach $\sigma^2_{\ell_2,\lambda} = \alpha^2_{\ell_2,\lambda}/\mu^2_{\ell_2,\lambda}$ as stated in (14). By minimizing $\sigma^2_{\ell_2,\lambda}$ with respect to $\lambda \geq 0$ we derive λ_{opt} and the resulting $\sigma^2_{\ell_2,\lambda_{\text{opt}}}$ in the statement of the lemma.

D.6 Proof of Corollary 3.1

The proof is analogous to the proof of Corollary 2.1. Here again we use Stam's inequality in Proposition A.3 to provide a bound for $\mathcal{I}(W_\sigma) = \mathcal{I}(\sigma G + Sf(S))$ based on $\mathcal{I}(\sigma G) = \sigma^{-2}$ and $\mathcal{I}(Sf(S))$. First we define

$$\widehat{\sigma} := \min_{x\geq 0}\left\{ s \geq 0 : \frac{1}{\delta} + \frac{1}{\delta s^2(\mathcal{I}(Sf(S))-1)} - 2x + \delta x^2(1+s^{-2}) \leq 1 \right\}. \tag{79}$$

Next we use Stam's inequality to deduce that :

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

We can use this inequality in the constraint condition of $\sigma_\star$ in (12) to deduce that:

$$\frac{1}{\delta} + \frac{1}{\delta\sigma_\star^2(\mathcal{I}(Sf(S))-1)} - 2\lambda_\star + \delta\lambda_\star^2(1+\sigma_\star^{-2}) \leq 1, \tag{80}$$

Thus we find that $(\sigma, x) = (\sigma_\star, \lambda_\star)$ is a feasible solution of the constraint in (79), resulting in :

$$\sigma_\star \geq \widehat{\sigma}. \tag{81}$$

To complete the proof of the theorem, we need to find the closed-form $\widehat{\sigma}$. Proceeding from (79) we derive the following

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

The first line follows by algebraic simplifications in (79). The second line is true since by Cramer-Rao bound (see Proposition A.3 (d)) $\mathcal{I}(Sf(S)) \geq (Var[Sf(S)])^{-1}$; thus $\mathcal{I}(Sf(S)) \geq 1$. Noting that the right hand-side of the inequality is independent of σ and can take positive values for some $x \geq 0$ we conclude the last line. Optimizing with respect to the non-negative variable x in the last line completes the proof and yields the desired result in the statement of the corollary.

E Comparison to a Simple Averaging Estimator

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

$$\widehat{\mathbf{w}}_{\text{ave}} = \frac{1}{m}\sum_{i=1}^{m} y_i \mathbf{a}_i. \tag{82}$$

This estimator is closely related to the family of RERM estimators studied in this paper. To see this, note that $\widehat{\mathbf{w}}_{\text{ave}}$ can be expressed as the solution to ridge-regularized ERM with $\lambda = 1$ and linear loss function $\mathcal{L}(x) = -x$ for all $x \in \mathbb{R}$:

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

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

It is in fact possible to exploit these relations of the estimator to the RERM family in order to evaluate its asymptotic performance using the machinery of this paper (i.e., by using the Equations (10)). However, a more direct evaluation that uses the closed form expression in (82) is preferable here. In fact, it can be easily checked that the following limit is true in the high-dimensional asymptotic regime:

$$\forall \delta > 0 \ : \qquad \text{corr}\left(\widehat{\mathbf{w}}_{\text{ave}}, \mathbf{x}_0\right) \xrightarrow{P} \frac{1}{1+\frac{1}{\delta}\nu_f^{-2}}, \tag{83}$$

where recall our notation $\nu_f = \mathbb{E}[Sf(S)]$, $S \sim \mathcal{N}(0,1)$. The use of the simple averaging estimator for signal recovery in generalized linear models (also, in single-index models) has been previously investigated for example in [Lu and Li, 2017].

A favorable feature of $\widehat{\mathbf{w}}_{\text{ave}}$ is its computational efficiency. In what follows, we use our lower bounds on the performance of general RERM estimators, to evaluate its suboptimality gap compared to more complicated alternatives. To begin, in view of (83) and (9) let us define the corresponding "effective error parameter"

$$\sigma_{\text{ave}}^2 = \frac{1}{\delta}\nu_f^{-2}. \tag{84}$$

First, we compare this value with the error of regularized LS. Let $\widehat{\mathbf{w}}_{\text{LS}}$ be the solution to *unregularized* LS for $n > m$. It can be checked (e.g., [Thrampoulidis et al., 2015a]) that

$$\text{corr}\left(\widehat{\mathbf{w}}_{\text{LS}}, \mathbf{x}_0\right) \xrightarrow{P} \frac{1}{1+\sigma_{\text{LS}}^2}, \quad \text{where} \quad \sigma_{\text{LS}}^2 := \frac{1}{\delta-1}(\nu_f^{-2}-1). \tag{85}$$

Directly comparing this to (84), we find that $\frac{\sigma_{\text{LS}}^2}{\sigma_{\text{ave}}^2} = \left(\frac{1}{1-1/\delta}\right)(1-\nu_f^2)$, for all $\delta > 1$. In other words,

$$\sigma_{\text{ave}}^2 \gtrless \sigma_{\text{LS}}^2 \quad \Longleftrightarrow \quad \delta \gtrless \nu_f^{-2}. \tag{86}$$

Next, we study the performance gap of the averaging estimator from the optimal RERM. For this, we use Corollary 3.1 to compare σ_{ave}^2 to the lower bound $\sigma_\star$. We find that for any $\delta > 0$ and any link function f satisfying the assumptions of Corollary 3.1:

$$1 \ \geq \ \frac{\sigma_\star^2}{\sigma_{\text{ave}}^2} \ \geq \ \delta\,\nu_f^2 \cdot H_\delta\Big(\mathcal{I}(Sf(S))\Big). \tag{87}$$

We complement these bounds with numerical simulations in Section G.

F Gains of Regularization

F.1 Linear models

In this section, we study the impact of the regularization parameter on the best achievable performance. For this purpose, we compare $\alpha_\star$, the best achievable performance of ridge-regularized case, to the best achievable performance among non-regularized empirical risk minimization with convex losses denoted by α_{ureg}. By definition of α_{ureg}, for all convex losses $\mathcal{L}$, in the regime of $\delta > 1$ it holds that, $\alpha_{\text{ureg}} \leq \alpha_{\mathcal{L},0}$. In [Bean et al., 2013], the authors compute a tight lower bound on α_{ureg} and show that it is attained provided that p_Z is log-concave. Our next result bounds the ratio $\alpha_\star^2 / \alpha_{\text{ureg}}^2$, illustrating the impact of regularization for a wide range of choices of $Z \sim \mathcal{D}$ and any $\delta > 1$.

Corollary F.1. *Let the assumptions of Corollary 2.1 hold and $\delta > 1$. Then it holds that:*

$$\frac{(\delta-1)}{\mathbb{E}[Z^2]}\, h_\delta\left(\frac{1}{\mathcal{I}(Z)}\right) \leq \frac{\alpha_\star^2}{\alpha_{\text{ureg}}^2} \leq \min\Big\{(\delta-1)\,\mathcal{I}(Z),\, 1\Big\}. \tag{88}$$

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

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

thus we may apply the Stam's inequality (as stated in Proposition A.3(f)) for $\mathcal{I}(V_{\alpha_{\text{ureg}}})$ to derive the following lower bound :

$$\alpha_{\text{ureg}}^2 \geq \frac{1}{(\delta-1)\mathcal{I}(Z)}. \tag{89}$$

Also note that it holds that $\alpha_\star^2 \leq \alpha_{\ell_2,\lambda_{\text{opt}}}^2$. Thus by recalling Lemma 2.2 and the fact that the function $h_\delta(\cdot) \leq 1$ for all $\delta \geq 0$ we deduce that $\alpha_\star^2 \leq 1$. Additionally since $\alpha_\star^2 \leq \alpha_{\text{ureg}}^2$, we conclude the upper bound in the statement of the Corollary. To proceed, we use the Cramer-Rao bound (see Proposition A.3(d)) for $\mathcal{I}(V_{\alpha_{\text{ureg}}})$ to derive the following upper bound for α_{ureg}^2 which holds for all $\delta > 1$:

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

This combined with the result of Corollary 2.1 derives the lower bound in the statement of the corollary and completes the proof. □

Importantly, based on (88) we find that as $\delta \to 1$ the ratio $\alpha_\star^2 / \alpha_{\text{ureg}}^2$ reaches zero, implying the large gap between $\alpha^\star$ and α_{ureg} in this regime. In the highly under-parameterized regime where $\delta \to \infty$, by computing the limit in the lower bound our bound gives

$$\frac{1}{\mathbb{E}[Z^2]\,\mathcal{I}(Z)} \leq \lim_{\delta\to\infty} \frac{\alpha_\star^2}{\alpha_{\text{ureg}}^2} \leq 1\,. \tag{90}$$

For example, we see that in this regime when Z is close to a Gaussian distribution such that $\mathcal{I}(Z) \approx 1/\mathbb{E}[Z^2]$, then provably $\alpha_\star \approx \alpha_{\text{ureg}}$, implying that impact of regularization is infinitesimal in the resulting error. We remark that for other distributions that are far from Gaussian in the sense $\mathcal{I}(Z) \gg 1/\mathbb{E}[Z^2]$ the simple lower bound in (90) is not tight; this is because the bound of Corollary 2.1 is not tight in this case.

F.2 Binary models

In order to demonstrate the impact of regularization on the performance of ERM based inference, we compare $\sigma_\star$ with the optimal error of the non-regularized ERM for $\delta > 1$ which we denote by σ_{ureg}. Thus σ_{ureg} satisfies for all convex losses that $\sigma_{\text{ureg}} \leq \sigma_{\mathcal{L},0}$. The general approach for determining σ_{ureg} is discussed in [Taheri et al., 2020]

in which the authors also show the achievability of σ_{ureg} for well-known models such as the Signed and Logistic models.

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

Corollary F.2. *Let the assumptions of Theorem 3.2 hold and $\delta > 1$. Further assume the label function f is such that $P_{S\cdot f(S)}(x)$ is differentiable and positive for all $x \in \mathbb{R}$. Then it holds that:*

$$\frac{(\delta-1)\nu_f^2}{1-\nu_f^2} H_\delta\Big(\mathcal{I}(\,Sf(S)\,)\Big) \;\le\; \frac{\sigma_\star^2}{\sigma_{\text{ureg}}^2} \;\le\; \min\left\{\frac{\delta-1}{\delta}\cdot\frac{\mathcal{I}(Sf(S))-1}{\nu_f^2},\, 1\right\}. \tag{91}$$

Proof. To provide the bounds of the ratio $\sigma_\star^2/\sigma_{\text{ureg}}^2$, we follow a similar argument stated in the proof of Corollary F.1. First, we use the result in [Taheri et al., 2020] which states that for σ_{ureg}^2 and all $\delta > 1$ it holds that

$$\sigma_{\text{ureg}}^2 \ge \frac{1}{(\delta-1)(\mathcal{I}(Sf(S))-1)}. \tag{92}$$

Since it trivially holds that $\sigma_\star^2 \le \sigma_{\ell_2,\lambda_{\text{opt}}}^2$ and also by noting that $\sigma_{\ell_2,\lambda_{\text{opt}}}^2$ as derived by Lemma 3.2 satisfies $\sigma_{\ell_2,\lambda_{\text{opt}}}^2 \le \frac{1}{\delta\nu_f^2}$ for all $\delta > 0$ (which is followed by the fact that $H_\delta(x) \le \frac{x}{(x-1)\delta}$), we conclude that

$$\sigma_\star^2 \le \frac{1}{\delta\nu_f^2}. \tag{93}$$

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

$$\frac{1-\sigma_{\text{ureg}}^2(1-\sigma_{\text{ureg}}^2\mathcal{I}(W_{\text{ureg}}))}{\delta\sigma_{\text{ureg}}^2(\sigma_{\text{ureg}}^2\mathcal{I}(W_{\text{ureg}})+\mathcal{I}(W_{\text{ureg}})-1)} = 1 \tag{94}$$

as well as the Cramer-Rao lower bound (Proposition A.3(d)) for $\mathcal{I}(W_{\text{ureg}})$ we may deduce that :

$$\sigma_{\text{ureg}}^2 \le \frac{(\delta-1)\nu_f^2}{1-\nu_f^2}. \tag{95}$$

This combined with the lower bound on $\sigma_\star^2$ as stated in Corollary 3.1 proves the lower bound in the statement of the corollary and completes the proof. □

Importantly, as shown by (91), in the case of δ being close to 1, one can see that both of the bounds in (91) vanish. This shows the large gap between σ_{ureg} and $\sigma_\star$ and further implies the benefit of regularization in this regime. When $\delta \to \infty$ i.e. in the highly under-parameterized regime, by deriving the limits as well as using Proposition A.3 (d), we see that (91) yields:

$$\frac{\nu_f^2}{1-\nu_f^2}\cdot\frac{1}{\mathcal{I}(Sf(S))-1} \le \lim_{\delta\to\infty}\frac{\sigma_\star^2}{\sigma_{\text{ureg}}^2} \le 1. \tag{96}$$

Thus in this case both the values of $\sigma_\star$ and σ_{ureg} are approaching zero with the ratio depending on the properties of $Sf(S)$. For models such as Logistic with small signal strength (i.e. small $\|\mathbf{x}_0\|$) where $\mathcal{I}(Sf(S)) \approx 1/(1-\nu_f^2)$, one can derive that based on (96) the ratio reaches 1, which confirms the intuition that for large values of δ the impact of regularization is almost negligible.

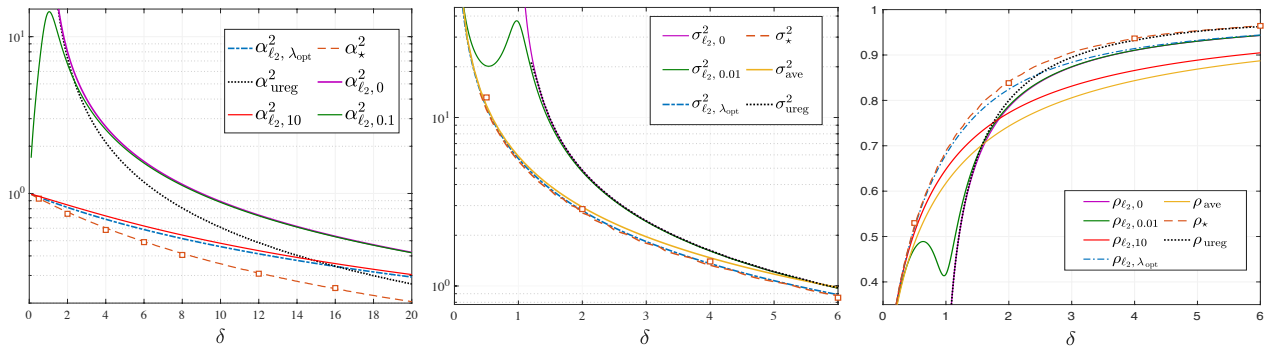

Figure 2: Fundamental error bounds derived in this paper compared to RLS, averaging estimator and optimal unregularized ERM for: (Top Left) a linear model with $Z \sim \texttt{Laplace}(0,2)$, (Top Right) a binary Logistic model with $\|\mathbf{x}_0\|_2 = 1$, (Bottom) a binary Logistic model with $\|\mathbf{x}_0\|_2 = 10$ (here shown is correlation measure (9)). The red squares correspond to numerical evaluation of the performance of the optimally tuned RERM as derived in Lemmas 2.1 and 3.1.

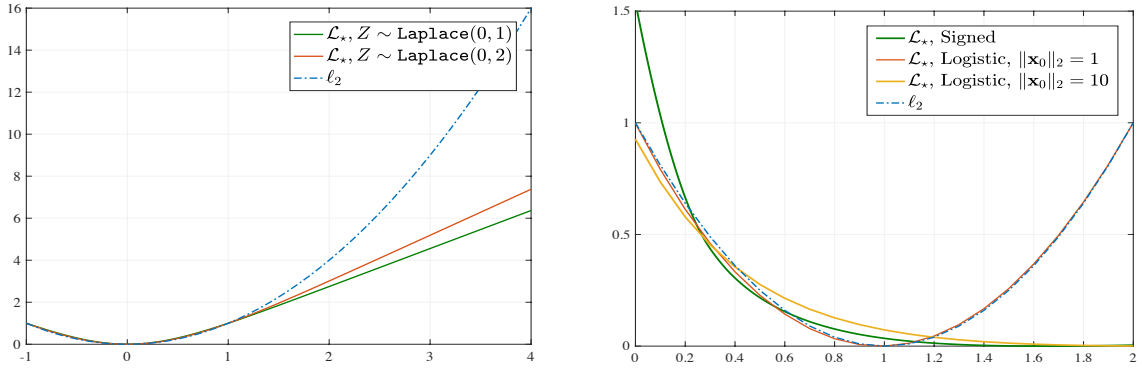

Figure 3: Illustrations of the proposed loss functions achieving optimal performance (as in Lemmas 2.1 and 3.1), for three special cases: a linear model with additive Laplace noise, the binary logistic model and the binary signed model. Here, in both plots, we fix $\delta = 2$. The curves are appropriately shifted and rescaled to allow direct comparison to the least-squares loss function.

G Additional Experiments

In this section, we present additional numerical results comparing the bounds of Theorems 2.1 and 3.2 to the performance of the following: (i) Ridge-regularized Least-Squares (RLS); (ii) optimal unregularized ERM (Section F); (iii) a simple averaging estimator (see Section E). Figure 2(Top Left) plots the asymptotic squared error α^2 of these estimators for linear measurements with $Z \sim \texttt{Laplace}(0,2)$. Similarly, Figure 2(Top Right) and Figure 2(Bottom) plot the effective error term σ for Logistic data with $\|\mathbf{x}_0\|_2 = 1$, and the limiting value ρ of the correlation measure for Logistic data with $\|\mathbf{x}_0\|_2 = 10$, respectively. The red squares represent the performance of optimally tuned ERM (as per Lemmas 2.1 and 3.1) derived numerically by running GD, as previously described in the context of Figure 1.

The numerical findings in Figures 1 and 2 validate the theoretical findings of Sections 2.3 and 3.3, regarding sub-optimality of RLS for Laplace noise and Logistic binary model (with large $\|\mathbf{x}_0\|$) and optimality of λ-tuned RLS for Logistic model with small $\|\mathbf{x}_0\|$. Furthermore, by comparing the optimal performance of unregularized ERM to the optimal errors of RERM in both Figures 1 and 2, we confirm the the theoretical guarantees of Section F regarding the impact of regularization in the regime of small δ for both linear and binary models.

G.1 Optimal Tuning in Special Cases

Figure 3 depicts the candidate for optimal loss function derived in Lemmas 2.1 and 3.1, for specific linear and binary models discussed in this paper. To allow for a direct comparison with the least-squares loss function, the optimal losses for the linear models are shifted such that $\mathcal{L}_\star \geq 0$ and rescaled such that $\mathcal{L}_\star(1) = 1$. Similarly, for the Logistic model with $\|\mathbf{x}_0\| = 1$, the optimal loss is rescaled such that $\mathcal{L}_\star(1) = 0$ and $\mathcal{L}_\star(2) = 1$. Interestingly, for this model, $\mathcal{L}_\star$, when rescaled (which results in no change in performance by appropriately rescaling $\lambda_\star$) is similar to the least-squares loss. This confirms the (approximate) optimality of optimally-tuned RLS for this model and further verifies the numerical observations in Figure 2 (Top Right) and the theoretical guarantees of Section 3.3 for this model.