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 sub-optimality 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.

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):

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

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*

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

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 $\mathcal{M}_{\mathcal{L}}\left(x;\tau\right) := \min_v \frac{1}{2\tau}(x-v)^2 + \mathcal{L}(v)$, for the *Moreau envelope function* and $\text{prox}_{\mathcal{L}}\left(x;\tau\right) := \arg\min_v \frac{1}{2\tau}(x-v)^2 + \mathcal{L}(v)$ 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: $\mathcal{M}_{\mathcal{L},1}'\left(x;\tau\right) := \frac{\partial \mathcal{M}_{\mathcal{L}}(x;\tau)}{\partial x}$. 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:

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

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

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.,

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

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:*

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

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 $h_\delta : \mathbb{R}_{>0} \to \mathbb{R}_{>0}$,

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

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 $\alpha^2_\star$)**.** *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*

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

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 $\alpha^2_\star \ge h_\delta(\zeta^2)$. 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 $(\mathcal{I}(Z)\,\mathbb{E}[Z^2])^{-1} \le 1$ (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,

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

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 $f:\mathbb{R}\to\{\pm1\}$ 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

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

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*

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

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 $\sigma_{\mathcal{L},\lambda}\geq\sigma_\star$.

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,*

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

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 $\mathcal{L}_\star : \mathbb{R} \to \mathbb{R}$*

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

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 $\lambda \ge 0$. Then, the system of equations in (10) has a unique solution $(\alpha_{\ell_2,\lambda}, \mu_{\ell_2,\lambda}, \tau_{\ell_2,\lambda})$ and $\sigma_{\ell_2,\lambda}^2 = \frac{\alpha_{\ell_2,\lambda}^2}{\mu_{\ell_2,\lambda}^2} =$*

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

Moreover, it holds that

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

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 $\sigma_{\ell_2,\lambda_{\rm opt}}^2$ conveniently matches with the simple bound on $\sigma_\star^2$ in Corollary 3.1. Specifically, it holds for any $\delta > 0$ that

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

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 $\mathcal{I}(Sf(S)) \ge (\mathrm{Var}[Sf(S)])^{-1} = (1-\nu_f^2)^{-1}$. Combining these, confirms that $\Omega_\delta \le 1$. Furthermore $\Omega_\delta = 1$ (thus, $\sigma_\star^2 = \sigma_{\ell_2,\lambda_{\rm opt}}^2$) 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-

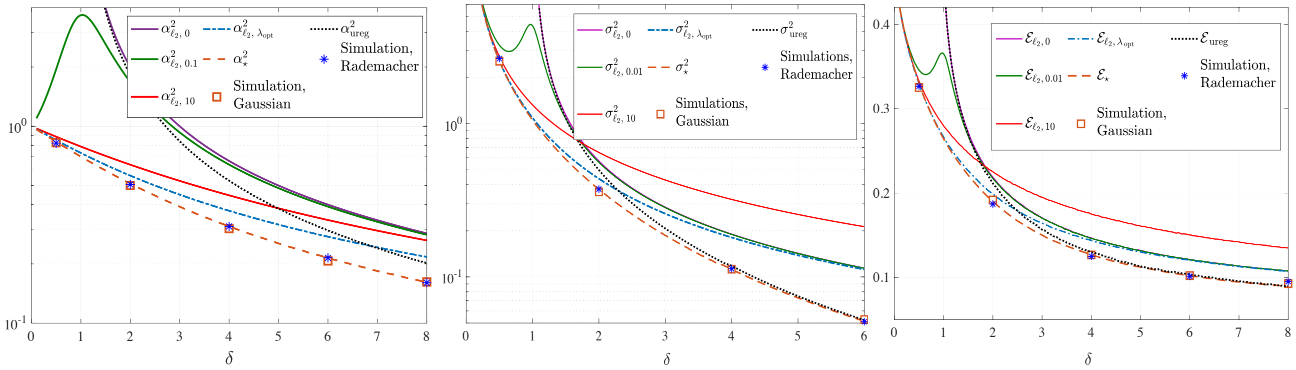

Figure 1: The lower bounds on error derived in this paper, compared to RLS for the linear model with $Z \sim \texttt{Laplace}(0,1)$ (Left), and for the binary Signed model(Middle) and binary Logistic model with $\|\mathbf{x}_0\|= 10$ (Right). The markers denote the emipirical performance of the optimally 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
$Z \sim$ 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
$Z \sim$ 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
$f =$ Sign	Theory	0.9934	0.8531	0.6199	0.4602	0.3618
	Experiment	0.9918	0.8204	0.6210	0.4710	0.3829
$f =$ 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 $\alpha^2_\star/\alpha^2_{\mathcal{L},\lambda_{\text{opt}}}$ (linear models) and $\sigma^2_\star/\sigma^2_{\mathcal{L},\lambda_{\text{opt}}}$(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

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

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 $\sigma^2_{\ell_2,\lambda_{\text{opt}}} \leq 2.442\ \sigma^2_\star,\ \forall\delta > 0$. 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 numerically 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 nonlinear 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*.