Sharp Asymptotics and Optimal Performance for Inference in Binary Models

Hossein Taheri
UCSB

Ramtin Pedarsani
UCSB

Christos Thrampoulidis
UCSB

Abstract

We study convex empirical risk minimization for high-dimensional inference in binary models. Our first result sharply predicts the statistical performance of such estimators in the linear asymptotic regime under isotropic Gaussian features. Importantly, the predictions hold for a wide class of convex loss functions, which we exploit in order to prove a bound on the best achievable performance among them. Notably, we show that the proposed bound is tight for popular binary models (such as Signed, Logistic or Probit), by constructing appropriate loss functions that achieve it. More interestingly, for binary linear classification under the Logistic and Probit models, we prove that the performance of least-squares is no worse than 0.997 and 0.98 times the optimal one. Numerical simulations corroborate our theoretical findings and suggest they are accurate even for relatively small problem dimensions.

1 INTRODUCTION

1.1 Motivation

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

Several recent works focus on the *linear asymptotic regime* and derive *sharp* results on the inference performance of appropriate convex optimization methods, e.g., [Donoho, 2006, Stojnic, 2009, Chandrasekaran et al., 2012, Donoho et al., 2011, Tropp, 2014, Bayati and Montanari, 2012, Oymak and Tropp, 2017, Stojnic, 2013, Oymak et al., 2013, Karoui, 2013, Bean et al., 2013, Thrampoulidis et al., 2015b, Donoho and Montanari, 2016, Thrampoulidis et al., 2018a, Advani and Ganguli, 2016, Weng et al., 2018, Thrampoulidis et al., 2018b, Miolane and Montanari, 2018, Bu et al., 2019, Xu et al., 2019, Celentano and Montanari, 2019]. These works show that, albeit challenging, *sharp* results are advantageous over loose order-wise bounds. Not only do they allow for accurate comparisons between different choices of the optimization parameters, but they also form the basis for establishing optimal such choices as well as fundamental performance limitations.

This paper takes this recent line of work a step further by demonstrating that results of this nature can be achieved in binary observation models. While we depart from the previously studied linear regression model, we remain faithful to the requirement and promise of sharp results. Binary models are popularly applicable in a wide range of signal-processing (e.g., highly quantized measurements) and machine learning (e.g., binary classification) problems. We derive sharp asymptotics for a rich class of convex optimization estimators, which includes least-squares, logistic regression and hinge-loss as special cases. Perhaps more interestingly, we use these results to derive fundamental performance limitations and design optimal loss functions that provably outperform existing choices.

In Section 1.2 we formally introduce the problem setup. The paper's main contributions and organization are presented in Section 1.3. A detailed discussion of prior art follows in Section 1.4.

Notation. The symbols $\mathbb{P}(\cdot)$, $\mathbb{E}[\cdot]$ and $\mathrm{Var}[\cdot]$ denote

Proceedings of the 23rd International Conference on Artificial Intelligence and Statistics (AISTATS) 2020, Palermo, Italy. PMLR: Volume 108.

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

1.2 Problem Statement

Consider the problem of recovering $\mathbf{x}_0 \in \mathbb{R}^n$ from observations $y_i = f(\mathbf{a}_i^T \mathbf{x}_0),\ i \in [m]$, where $f : \mathbb{R} \to \{\pm 1\}$ is a (possibly random) binary function. We study the performance of *empirical-risk minimization (ERM)* estimators $\widehat{\mathbf{x}}_\ell$ that solve the following optimization problem for some *convex* loss function $\ell : \mathbb{R} \to \mathbb{R}$

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

Model. The binary observations $y_i, i \in [m]$ are determined by a label function $f : \mathbb{R} \to \{-1, 1\}$ as follows:

$$y_i = f(\mathbf{a}_i^T \mathbf{x}_0), \quad i \in [m], \tag{2}$$

where $\mathbf{a}_i$'s are known measurement vectors with i.i.d. Gaussian entries; and $\mathbf{x}_0 \in \mathbb{R}^n$ is an unknown vector of coefficients. Some popular examples for the label function f include the following:

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

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

- *Logistic*: $y_i = \begin{cases} +1 & \text{, w.p. } \frac{1}{1+\exp(-\mathbf{a}_i^T \mathbf{x}_0)}, \\ -1 & \text{, w.p. } 1 - \frac{1}{1+\exp(-\mathbf{a}_i^T \mathbf{x}_0)}. \end{cases}$

- *Probit*: $y_i = \begin{cases} +1 & \text{, w.p. } \Phi(\mathbf{a}_i^T \mathbf{x}_0), \\ -1 & \text{, w.p. } 1 - \Phi(\mathbf{a}_i^T \mathbf{x}_0). \end{cases}$

Loss function. We study the recovery performance of estimates $\widehat{\mathbf{x}}_\ell$ of $\mathbf{x}_0$ that are obtained by solving (1) for proper convex loss functions $\ell : \mathbb{R} \to \mathbb{R}$. Different choices for ℓ lead to popular specific estimators including the following:

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

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

$$\mathrm{corr}\left(\, \widehat{\mathbf{x}}_\ell\, ;\, \mathbf{x}_0 \,\right) := \frac{\langle \widehat{\mathbf{x}}_\ell, \mathbf{x}_0 \rangle}{\|\widehat{\mathbf{x}}_\ell\|_2 \|\mathbf{x}_0\|_2} \in [-1, 1]. \tag{3}$$

Obviously, we seek estimates that maximize correlation. While correlation is the measure of primal interest, our results extend rather naturally to other parameter estimation metrics, such as square error, as well as prediction metrics, such as classification error.

Model assumptions. All our results are valid under the assumption that the measurement vectors have i.i.d. Gaussian entries.

Assumption 1 (Gaussian feature vectors)**.** *The vectors $\mathbf{a}_i \in \mathbb{R}^n$, $i \in [m]$ have entries i.i.d. standard normal.*

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

1.3 Contributions and Organization

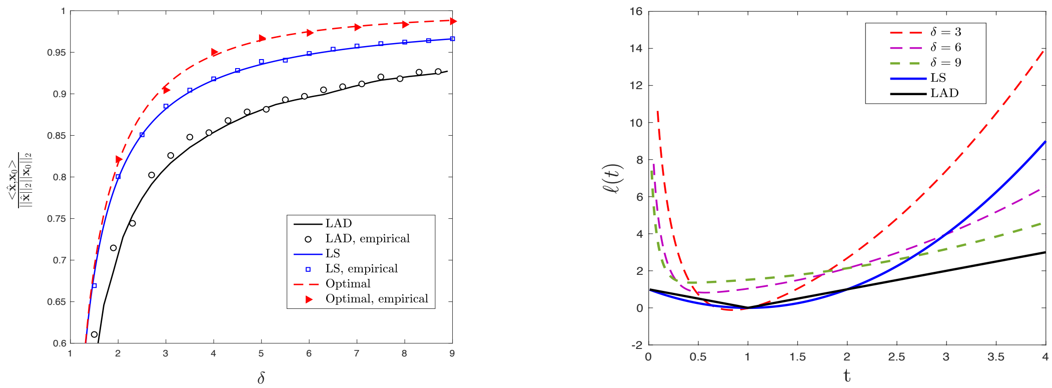
This paper's main contributions are summarized below.

- **Sharp asymptotics**: We show that the absolute value of correlation of $\widehat{\mathbf{x}}_\ell$ to the true vector $\mathbf{x}_0$ is sharply predicted by $\sqrt{1/(1+\sigma_\ell^2)}$ where the "effective noise" parameter σ_ℓ can be explicitly computed by solving a system of three non-linear equations in three unknowns. We find that the system of equations (and thus, the value of σ_ℓ) depends on the loss function ℓ through its Moreau envelope function. Our prediction holds in the linear asymptotic regime in which $m, n \to \infty$ and $m/n \to \delta > 1$. See Section 2.

- **Fundamental limits**: We establish fundamental limits on the performance of convex optimization-based estimators by computing an upper bound on the best possible correlation performance among all convex loss functions. We compute the upper bound by solving a certain nonlinear equation and we show that such a solution exists for all $\delta > 1$. See Section 3.1.

- **Optimal performance**: For certain models including Signed and Logistic, we find the loss functions that achieve the optimal performance, i.e., they attain the previously derived upper bound. See Section 3.2.

- **Optimality of LS**: For binary logistic and sigmoid models, we prove that the correlation performance of



(a) (b)

Figure 1: Left: Comparison between analytical (solid lines) and empirical (markers) performance for least-squares (LS) and least-absolute deviations (LAD), along with optimal performance (dashed line) as predicted by the upper bound of Theorem 3.1 for the Signed model($\varepsilon = 0$). The red markers depict the empirical performance of the optimal loss function that attains the upper bound. Right: Illustrations of optimal loss functions for the Signed model for different values of δ according to Theorem 3.2.

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

δ	2	3	4	5	6	7	8	9
Predicted Performance	0.8168	0.9101	0.9457	0.9645	0.9748	0.9813	0.9855	0.9885
Empirical Performance	0.8213	0.9045	0.9504	0.9669	0.9734	0.9801	0.9834	0.9873

least-squares (LS) is at least as good as 0.9972 and 0.9804 times the optimal performance. See Section 4.1.

• **Numerical simulations**: We specialize our results on general models and loss functions to popular instances, for which we provide simulation results that demonstrate the accuracy of the theoretical predictions. See Section 5.

Figure 1 contains a pictorial preview of our results described above for the special case of Signed measurements. First, Figure 1a depicts the correlation performance of LS and LAD estimators as a function of the aspect ratio δ. Both theoretical predictions and numerical results are shown; note the close match for even small dimensions. Second, the dashed line on the same figure shows the upper bound derived in this paper – there is no convex loss function that results in correlation exceeding this line. Third, we show that the upper bound can be achieved by the loss functions depicted in Figure 1b for several values of δ. We solve (1) for this choice of loss functions using gradient descent and numerically evaluate the achieved correlation performance. The recorded values are compared in Table 1 to the corresponding values of the upper bound; again, note the close agreement between the values as predicted by the findings of this paper. We present corresponding results for the Logistic model in Section 5 and for the Noisy-signed model in Appendix E.

1.4 Related Work

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

(a) Sharp asymptotics for linear measurements. Most of the results in the literature of high-dimensional statistics are order-wise in nature. Sharp asymptotic predictions have only more recently appeared in the literature for the case of noisy linear measurements with Gaussian measurement vectors. There are by now three different approaches that have been used towards asymptotic analysis of convex regularized estimators:
i) the one that is based on the approximate message passing (AMP) algorithm and its state-evolution analysis, e.g., [Donoho et al., 2009, 2011, Bayati and Montanari, 2011, 2012, Donoho and Montanari, 2016, Bu et al., 2019, Mousavi et al., 2018].
ii) the one that is based on Gaussian process (GP) inequalities, specifically the convex Gaussian min-max

Theorem (CGMT) e.g., [Stojnic, 2013, Oymak et al., 2013, Thrampoulidis et al., 2015b, 2018a,b, Miolane and Montanari, 2018].

Our results in Theorems 3.1 and 3.2 for achieving the best performance across all loss functions is complementary to [Bean et al., 2013, Theorem 1] and [Advani and Ganguli, 2016] in which the authors also proposed a method for deriving optimal loss function and measuring its performance, albeit for *linear* models. Instead, we study binary models. The optimality of regularization for linear measurements, is recently studied in [Celentano and Montanari, 2019].

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

$$\widehat{\mathbf{x}} := \arg\min_{\mathbf{x}} \sum_{i=1}^{m} \widetilde{\ell}(y_i - \mathbf{a}_i^T \mathbf{x}) + rR(\mathbf{x}) \tag{4}$$

Specifically, the loss function $\widetilde{\ell}$ penalizes the residual. In this paper, we show that the CGMT is applicable to optimization problems in the form of (1). For our case of binary observations, (1) is more general than (4). To see this, note that for $y_i \in \{\pm 1\}$ and popular symmetric loss functions $\widetilde{\ell}(t) = \widetilde{\ell}(-t)$, e.g. least-squares (LS), (1) results in (4) by choosing $\ell(t) = \widetilde{\ell}(t-1)$ in the former. Moreover, (1) includes several other popular loss functions such as the logistic loss and the hinge-loss which cannot be expressed by (4).

Similar to the generality of our paper, [Genzel, 2017] also studies the high-dimensional performance of general loss functions. However, in contrast to our results, their performance bounds are loose (order-wise); as such, they are not informative about the question of optimal performance which we also address here.

(b) Classification in high-dimensions. In [Candès and Sur, 2018, Sur and Candès, 2019] the authors study the high-dimensional performance of maximum-likelihood (ML) estimation for the logistic model. The ML estimator is a special case of (1) and we consider general binary models. Also, their analysis is based on the AMP. The asymptotics of logistic loss under different classification models has also been recently studied in [Mai et al., 2019]. In yet another closely related recent work [Salehi et al., 2019], the authors extend the results of [Sur and Candès, 2019] to regularized ML by using the CGMT. Instead, we present results for general loss functions and for general measurement models. Importantly, we also study performance bounds and optimal loss functions. A preliminary version of the results of this paper was published in [Taheri et al., 2019].

2 SHARP PERFORMANCE GUARANTEES

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

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

for the *Moreau envelope function* of the loss $\ell : \mathbb{R} \to \mathbb{R}$ at x with parameter $\lambda > 0$. The minimizer (which is unique by strong convexity) is known as the *proximal operator* of ℓ at x with parameter λ and we denote it as $\mathrm{prox}_{\ell}\left(x;\lambda\right)$. A useful property of the Moreau envelope function is that it is continuously differentiable with respect to both x and λ [Rockafellar and Wets, 2009]. We denote these derivatives as follows

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

A system of equations. As we show shortly, the asymptotic performance of the optimization in (1) is tightly connected to the solution of a certain system of nonlinear equations, which we introduce here. Specifically, define random variables G, S and Y as follows:

$$G, S \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0,1) \quad \text{and} \quad Y = f(S), \tag{5}$$

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

$$\mathbb{E}\Big[Y\, S \cdot \mathcal{M}'_{\ell,1}\left(\alpha G + \mu S Y;\lambda\right)\Big] = 0, \tag{6a}$$

$$\lambda^2\, \delta\, \mathbb{E}\Big[\, \left(\mathcal{M}'_{\ell,1}\left(\alpha G + \mu S Y;\lambda\right)\right)^2\, \Big] = \alpha^2, \tag{6b}$$

$$\lambda\, \delta\, \mathbb{E}\Big[G \cdot \mathcal{M}'_{\ell,1}\left(\alpha G + \mu S Y;\lambda\right)\Big] = \alpha. \tag{6c}$$

The expectations are with respect to the randomness of the random variables G, S and Y. We remark that the equations are well defined even if the loss function ℓ is not differentiable. In Section A we summarize some well-known properties of the Moreau Envelope function and use them to simplify (6) for differentiable loss functions.

2.1 Asymptotic Prediction

We are now ready to state our first main result.

Theorem 2.1 (Sharp asymptotics). *Let Assumption 1 hold and assume $\delta > 1$ such that the set of minimizers in* (1) *is bounded and the system of equations* (6) *has a unique solution $(\mu, \alpha \geq 0, \lambda \geq 0)$, such that $\mu \neq 0$.*

Let $\widehat{\mathbf{x}}_\ell$ be as in (1). *Then, in the limit of $m, n \to +\infty$, $m/n \to \delta$, it holds with probability one that*

$$\lim_{n\to\infty} \operatorname{corr}\left(\widehat{\mathbf{x}}_\ell\,;\,\mathbf{x}_0\right) = \frac{\mu}{\sqrt{\mu^2+\alpha^2}}. \tag{7}$$

Moreover,

$$\lim_{n\to\infty} \left\| \widehat{\mathbf{x}}_\ell - \mu\cdot\frac{\mathbf{x}_0}{\|\mathbf{x}_0\|_2}\right\|_2^2 = \alpha^2. \tag{8}$$

Theorem 2.1 holds for general loss functions. In Section 4 we specialize the result to specific popular choices and also present numerical simulations that confirm the validity of the predictions (see Figures. 1a–2a and 6a–6b). Before that, we include a few remarks on the conditions, interpretation and implications of the theorem. The proof is deferred to Appendix B and uses the convex Gaussian min-max theorem (CGMT) [Thrampoulidis et al., 2015b, 2018a].

Remark 1 (The role of μ and α). According to (7), the prediction for the limiting behavior of the correlation value is given in terms of an effective noise parameter $\sigma_\ell := \alpha/\mu$, where μ and α are unique solutions of (6). *The smaller the value of σ_ℓ is, the larger becomes the correlation value.* While the correlation value is fully determined by the ratio of α and μ, their individual role is clarified in (8). Specifically, according to (8), $\widehat{\mathbf{x}}_\ell$ is a biased estimate of the true $\mathbf{x}_0$ and μ represents exactly that bias term. In other words, solving (1) returns an estimator that is close to a μ–scaled version of $\mathbf{x}_0$. When $\mathbf{x}_0$ and $\widehat{\mathbf{x}}_\ell$ are scaled appropriately, the ℓ_2-norm of their difference converges to α.

Remark 2 (Why $\delta > 1$). The theorem requires that $\delta > 1$ (equivalently, $m > n$). Here, we show that this condition is *necessary* for the equations (6) to have a bounded solution. To see this, take squares in both sides of (6c) and divide by (6b), to find that

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

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

Remark 3 (On the existence of a solution to (6)). While $\delta > 1$ is a necessary condition for the equations in (6) to have a solution, it is *not* sufficient in general. This depends on the specific choice of the loss function. For example, in Section 4.1, we show that for the squared loss $\ell(t) = (t-1)^2$, the equations have a unique solution iff $\delta > 1$. On the other hand, for logistic-loss and hinge-loss, it is argued in Section 4.2 that there exists a threshold value $\delta_f^\star > 2$ such that the set of minimizers in (1) is unbounded if $\delta < \delta_f^\star$. In this case, Theorem 2.1 does not hold. We conjecture that for these choices of loss, the equations (6) are solvable iff $\delta > \delta_f^\star$. Justifying this conjecture and further studying more general sufficient and necessary conditions under which the equations (6) admit a solution is left to future work. However, in what follows, given such a solution, we prove that it is unique for a wide class of convex-loss functions of interest.

Remark 4 (On the uniqueness of solution to (6)). We show that if the system of equations in (6) has a solution, then it is unique provided that ℓ is strictly convex, continuously differentiable and its derivative satisfies $\ell'(0) \neq 0$. For instance, this class includes the square, the logistic and the exponential losses. However, it excludes non-differentiable functions such as the LAD and hinge-loss. We believe that the differentiability assumption can be relaxed without major modification in our proof, but we leave this for future work. Our result is summarized in Proposition 2.1 below.

Proposition 2.1. *Assume that the loss function $\ell : \mathbb{R} \to \mathbb{R}$ has the following properties: (i) it is proper strictly convex; (ii) it is continuously differentiable and its derivative ℓ' is such that $\ell'(0) \neq 0$. Further assume that the (possibly random) link function f is such that $SY = Sf(S)$, $S \sim \mathcal{N}(0,1)$ has strictly positive density on the real line. The following statement is true. For any $\delta > 1$, if the system of equations in* (6) *has a bounded solution, then it is unique.*

The detailed proof of Proposition 2.1 is deferred to the extended version of the paper [Taheri et al., 2020, Appendix B.5]. Here, we highlight some key ideas. The CGMT relates –in a rather natural way– the original ERM optimization (1) to the following deterministic min-max optimization on four variables

$$\min_{\alpha>0,\mu,\tau>0}\ \max_{\gamma>0}\ F(\alpha,\mu,\tau,\gamma) := \frac{\gamma\tau}{2} - \frac{\alpha\gamma}{\sqrt{\delta}} + \mathbb{E}\left[\mathcal{M}_\ell\left(\alpha G+\mu YS;\frac{\tau}{\gamma}\right)\right]. \tag{9}$$

In [Taheri et al., 2020, Appendix B.4], we show that the optimization above is convex-concave for any lower semi-continuous, proper, convex function $\ell : \mathbb{R} \to \mathbb{R}$. Moreover, it is shown that one arrives at the system of equations in (6) by simplifying the first-order optimality conditions of the min-max optimization in (9). This connection is key to the proof of Proposition 2.1. Indeed, we prove uniqueness of solution (if such a solution exists) to (6), by proving instead that the function $F(\alpha,\mu,\tau,\gamma)$ above is (jointly) *strictly* convex in (α,μ,τ) and *strictly* concave in γ, provided that ℓ satisfies the conditions of the proposition. Next, let us briefly discuss how strict convex-concavity of (9) can be shown. For concreteness, we only discuss strict convexity here; the ideas are similar for strict concavity. At the heart of the proof of strict convexity of F is understanding the properties of the *expected Moreau envelope function* $\Omega : \mathbb{R}_+ \times \mathbb{R} \times \mathbb{R}_+ \times \mathbb{R}_+ \to \mathbb{R}$ defined as follows:

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

Specifically, we prove in [Taheri et al., 2020, Proposition A.7] that if ℓ is strictly convex, differentiable and does not attain its minimum at 0, then Ω is strictly convex in (α, μ, τ) and strictly concave in γ. It is worth noting that the Moreau envelope function $\mathcal{M}_\ell\left(\alpha g + \mu y s; \tau\right)$ for fixed g, s and $y = f(s)$ is *not* necessarily strictly convex. Interestingly, we show that the *expected* Moreau envelope has this desired feature. We refer the reader to [Taheri et al., 2020, Appendices A.6 and B.5] for more details.

3 ON OPTIMAL PERFORMANCE

3.1 Fundamental Limitations

In this section, we establish fundamental limits on the performance of (1) by deriving an upper bound on the absolute value of correlation $\mathrm{corr}\left(\widehat{\mathbf{x}}_\ell\,;\,\mathbf{x}_0\right)$ that holds for *all* choices of loss functions satisfying Theorem 2.1. The result builds on the prediction of Theorem 2.1. In view of (7) upper bounding correlation is equivalent to lower bounding the effective noise parameter $\sigma_\ell = \alpha/\mu$. Theorem 3.1 below derives such a lower bound. The proof is deferred to Appendix C.

For a random variable H with density $p_H(h)$ that has a derivative $p_H'(h), \forall h \in \mathbb{R}$, we denote its score function $\xi_H(h) := \frac{\partial}{\partial h}\log p_H(h) = \frac{p_H'(h)}{p_H(h)}$. Then, the Fisher information of H is defined as follows (e.g. [Barron, 1984, Sec. 2]):

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

Theorem 3.1 (Best achievable performance)**.** *Let the assumptions and notation of Theorem 2.1 hold and recall the definition of random variables G, S and Y in* (5)*. For $\sigma > 0$, define a new random variable $W_\sigma := \sigma G + SY$, and the function $\kappa : (0, \infty] \to [0, 1]$ as follows,*

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

Further define $\sigma_{\rm opt}$ as follows,

$$\sigma_{\rm opt} := \min\left\{\sigma \geq 0 : \kappa(\sigma) = \frac{1}{\delta}\right\}. \tag{10}$$

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

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

Remark 5. Theorem 3.1 complements the results of [Bean et al., 2013], [Donoho and Montanari, 2016, Lem. 3.4] and [Thrampoulidis et al., 2018a, Rem. 5.3.3] in which they consider only linear measurements. In particular, Theorem 3.1 shows that it is possible to achieve results of this nature for the more challenging setting of binary observations considered here.

A useful closed-form bound on the best achievable performance: In general, determining $\sigma_{\rm opt}$ requires computing the Fisher information of the random variable $\sigma G + SY$ for $\sigma > 0$. If the probability distribution of SY is continuously differentiable (e.g., logistic model; see Section C.3), then we obtain the following simplified bound. The proof is deferred to Appendix C.4.

Corollary 3.1 (Closed-form lower bound on $\sigma_{\rm opt}$)**.** *Let $p_{SY} : \mathbb{R} \to \mathbb{R}$ be the probability distribution of SY. If $p_{SY}(x)$ is differentiable for all $x \in \mathbb{R}$, then,*

$$\sigma_{\rm opt}^2 \geq \frac{1}{(\delta - 1)(\mathcal{I}(SY) - 1)}\,. \tag{11}$$

The proof of the corollary reveals that (11) holds with equality when SY is Gaussian. In Section C.3, we compute p_{SY} for the Logistic and the Probit models and numerically show that it is close to the density of a Gaussian random variable. Consequently, the lower bound of Corollary 3.1 is almost exact when measurements are obtained according to the Logistic and Probit models; see Figure 4 in the appendix.

3.2 On the Optimal Loss Function

It is natural to ask whether there exists a loss function that attains the bound of Theorem 3.1. If such a loss function exists, then we say it is *optimal* in the sense that it maximizes the correlation performance among all convex loss functions in (1).

Our next theorem derives a candidate for the optimal loss function, which we denote $\ell_{\rm opt}$.The proof is deferred to Appendix D

Theorem 3.2 (Optimal loss function)**.** *Recall the definition of $\sigma_{\rm opt}$ in* (10)*. Define the random variable $W_{\rm opt} := \sigma_{\rm opt}\, G + SY$ and let $p_{W_{\rm opt}}$ denote its density. Consider the following loss function $\ell_{\rm opt} : \mathbb{R} \to \mathbb{R}$*

$$\ell_{\rm opt}(w) = -\mathcal{M}_{\alpha_1 q + \alpha_2 \log(p_{W_{\rm opt}})}\left(w; 1\right), \tag{12}$$

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

$$\alpha_1 = \frac{1 - \sigma_{\text{opt}}^2 \mathcal{I}(W_{\text{opt}})}{\delta(\sigma_{\text{opt}}^2 \mathcal{I}(W_{\text{opt}}) + \mathcal{I}(W_{\text{opt}}) - 1)}, \qquad \alpha_2 = \frac{1}{\delta(\sigma_{\text{opt}}^2 \mathcal{I}(W_{\text{opt}}) + \mathcal{I}(W_{\text{opt}}) - 1)}. \tag{13}$$

If ℓ_{opt} *defined as in* (12) *is convex and the equation* $\kappa(\sigma) = 1/\delta$ *has a unique solution, then* $\sigma_{\ell_{\text{opt}}} = \sigma_{\text{opt}}$.

In general, there is no guarantee that the function $\ell_{\text{opt}}(\cdot)$ as defined in (12) is convex. However, if this is the case, the theorem above guarantees that it is optimal [1]. A *sufficient* condition for $\ell_{\text{opt}}(w)$ to be convex, is provided in Appendix D.2. Importantly, in Appendix D.2.1 we show that this condition holds for observations following the Signed model. Thus, for this case the resulting function is convex. Although we do *not* prove the convexity of optimal loss function for the Logistic and Probit models, our numerical results (e.g., see Figure 2b) suggest that this is the case. Concretely, we conjecture that the loss function ℓ_{opt} is convex for Logistic and Probit models, and therefore by Theorem 3.2 its performance is optimal.

4 SPECIAL CASES

4.1 Least-Squares

For this choice of loss function, we can solve the equations in (6) in closed form. Furthermore, the equations have a unique and bounded solution for any $\delta > 1$ provided that $\mathbb{E}[SY] > 0$. The final result is summarized in the corollary below. See Section F.1 for the proof.

Corollary 4.1 (Least-squares)**.** *Let Assumption 1 hold and* $\delta > 1$*. For the label function assume that* $\mathbb{E}[SY] > 0$ *in the notation of* (5)*. Let* $\widehat{\mathbf{x}}_\ell$ *be as in* (1) *for* $\ell(t) = (t-1)^2$*. Then, in the limit of* $m, n \to +\infty$, $m/n \to \delta > 1$*, Equations* (7) *and* (8) *hold with probability one with* α *and* μ *given as follows:*

$$\mu = \mathbb{E}[SY], \tag{14}$$

$$\alpha = \sqrt{1 - (\mathbb{E}[SY])^2} \cdot \sqrt{\frac{1}{\delta - 1}}\,. \tag{15}$$

On the Optimality of LS. On the one hand, Corollary 4.1 derives an explicit formula for the effective noise variance $\sigma_{\text{LS}} = \alpha/\mu$ of LS in terms of $E[YS]$ and δ. On the other hand, Corollary 3.1 provides an explicit lower bound on the optimal value σ_{opt} in terms of $\mathcal{I}(SY)$ and δ. Combining the two, we conclude that

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

[1]Strictly speaking, the performance is optimal among all convex loss functions ℓ for which (6) has a unique solution as required by Theorem 3.1.

In terms of correlation,

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

where the first inequality follows from the fact that $\sigma_{\text{LS}} \ge \sigma_{\text{opt}}$. Therefore, the performance of LS is at least as good as $\frac{1}{\sqrt{\xi}}$ times the optimal one. In particular, for Logistic and Probit models (for which Corollary 3.1 holds), we can explicitly compute $\frac{1}{\sqrt{\xi}} =$ 0.9972 and 0.9804, respectively.

4.2 Logistic & Hinge Loss Functions

Theorem 2.1 only holds in regimes for which the set of minimizers of (1) is bounded. As we show here, this is *not* always the case. Specifically, consider non-negative loss functions $\ell(t) \ge 0$ with the property $\lim_{t\to+\infty} \ell(t) = 0$. For example, the hinge, exponential and logistic loss functions all satisfy this property. Now, we show that for such loss functions the set of minimizers is unbounded if $\delta < \delta_f^\star$ for some appropriate $\delta_f^\star > 2$. First, note that the set of minimizers is unbounded if the following condition holds:

$$\exists\ \mathbf{x}_s \neq \mathbf{0} \quad \text{such that} \quad y_i \mathbf{a}_i^T \mathbf{x}_s \ge 0, \quad \forall\ i \in [m]. \tag{16}$$

Indeed, if (16) holds then $\mathbf{x} = c \cdot \mathbf{x}_s$ with $c \to +\infty$, attains zero cost in (1); thus, it is optimal and the set of minimizers is unbounded. To proceed, we rely on a recent result by Candes and Sur [Candès and Sur, 2018] who prove that (16) holds iff

$$\delta \le \delta_f^\star := \left(\min_{c\in\mathbb{R}} \mathbb{E}\left[(G + c\,S\,Y)_-^2\right]\right)^{-1}, \tag{17}$$

where G, S and Y are random variables as in (5) and $(t)_- := \min\{0, t\}$. We highlight that Logistic and Hinge losses give unbounded solutions in the Noisy-Signed model with $\varepsilon = 0$, since the condition (16) holds for $\mathbf{x}_s = \mathbf{x}_0$. However their performances are comparable to the optimal performance in the Logistic model(see Figure 2a).

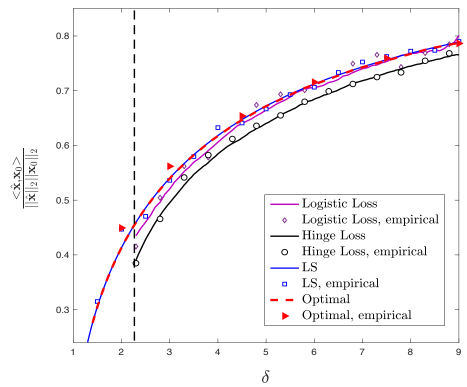
5 NUMERICAL EXPERIMENTS

In this section, we present numerical simulations that validate the predictions of Theorems 2.1, 3.1 and 3.2. We use Signed and Logistic models as our case study. The experiments on Probit model are presented in the extended version of this paper [Taheri et al., 2020]. We generate random measurements according to (2) and Assumption 1. Without loss of generality (due to rotational invariance of the Gaussian measure) we set $\mathbf{x}_0 = [1, 0, ..., 0]^T$. We then obtain estimates $\widehat{\mathbf{x}}_\ell$ of $\mathbf{x}_0$ by numerically solving (1) and measure performance by the correlation value $\mathrm{corr}(\widehat{\mathbf{x}}_\ell\,;\mathbf{x}_0)$. Throughout the experiments, we set $n = 128$ and the recorded values of correlation are averages over 25 independent realizations. For each label function we first provide plots that compare results of Monte Carlo simulations to the asymptotic predictions for loss functions discussed in Section 4, as well as, to the optimal performance of Theorem 3.1. We next present numerical results on optimal loss functions. In order to empirically derive the correlation of optimal loss function, we run gradient descent-based optimization with 1000 iterations. As a general comment, we note that despite being asymptotic, our predictions appear accurate even for relatively small problem dimensions. For the analytical predictions we apply Theorem 2.1. In particular for solving the system of non-linear equations in (1), we empirically observe that if a solution exists, then it can be efficiently found by the following fixed-point iteration method. Let $\mathbf{v} := [\mu, \alpha, \lambda]^T$ and $\mathcal{F}:\mathbb{R}^3 \to \mathbb{R}^3$ be such that (1) is equivalent to $\mathbf{v} = \mathcal{F}(\mathbf{v})$. With this notation, we initialize $\mathbf{v} = \mathbf{v}_0$ and for $k \geq 1$ repeat the iterations $\mathbf{v}_{k+1} = \mathcal{F}(\mathbf{v}_k)$ until convergence.

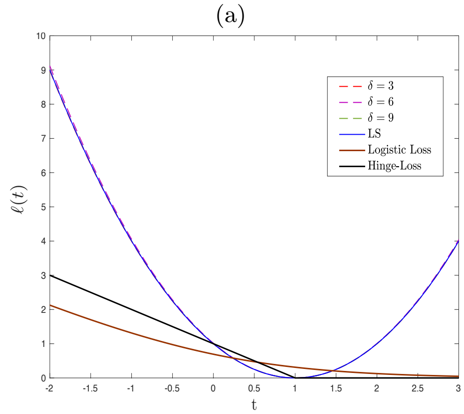
Logistic model. For the logistic model, comparison between the predicted values and the numerical results is illustrated in Figure 2a. Results are shown for LS, logistic and hinge loss functions. Note that minimizing the logistic loss corresponds to the maximum-likelihood estimator (MLE) for logistic model. An interesting observation in Figure 2a is that in the high-dimensional setting (finite δ) LS has comparable (if not slightly better) performance to MLE. Additionally we observe that in this model, performance of LS is almost the same as the best possible performance derived according to Theorem 3.1. This confirms the analytical conclusion of Section 4.1. The comparison between the optimal loss function as in Theorem 3.2 and other loss functions is illustrated in Figure 2b. We note the obvious similarity between the shapes of optimal loss functions and LS which further explains the similarity between their performance.

Optimal loss function. By putting together Theorems 3.1 and 3.2, we obtain a method on deriving the optimal loss function. This requires the following steps.
1. Find σ_{opt} by solving (10).
2. Compute the density of $W_{\text{opt}} = \sigma_{\text{opt}} G + SY$.
3. Compute ℓ_{opt} according to (12).

Note that computing σ_{opt} needs the density function p_W of the random variable $W = \sigma\, G + SY$. In principle p_W can be calculated as the convolution of the Gaussian density with the pdf p_{SY} of SY. Moreover, it follows from the recipe above that the optimal loss function depends on δ in general. This is because σ_{opt} itself depends on δ via (10).



(a)



(b)

Figure 2: Top: Comparison between analytical and empirical results for the performance of LS, Logistic loss, Hinge-loss and optimal loss function for Logistic model. The vertical dashed line represents $\delta_f^\star \approx 2.275$, as evaluated by (17). Bottom: Illustrations of optimal loss functions for different values of δ, derived according to Theorem 3.2 for Logistic model. In order to signify the similarity of optimal loss function to the LS loss, the optimal loss functions (hardly visible) are scaled such that $\ell(1) = 0$ and $\ell(2) = 1$.

6 CONCLUSION

This paper derives *sharp* asymptotic performance guarantees for a wide class of convex optimization based estimators for recovering a signal from binary observation models. We further provide a theoretical upper bound on the best achievable performance among all convex loss functions. Using this, we develop a procedure for computing the optimal loss function. Finally, we provide numerical studies that show tight agreement with our theoretical results. Interesting future directions include studying the generalized linear measurement model beyond binary observations and characterizing the optimal loss function for such general models.

Acknowledgment

This work was supported by NSF Grants CCF-1755808 and CCF-1909320 and an Academic Senate Research Grant from UCSB.

References

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

Andrew R Barron. Monotonic central limit theorem for densities. *Department of Statistics, Stanford University, California, Tech. Rep*, 50, 1984.

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

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

Derek Bean, Peter J Bickel, Noureddine El Karoui, and Bin Yu. Optimal m-estimation in high-dimensional regression. *Proceedings of the National Academy of Sciences*, 110(36):14563–14568, 2013.

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

Stephen Boyd and Lieven Vandenberghe. *Convex optimization*. Cambridge university press, 2009.

David R Brillinger. A generalized linear model with" gaussian" regressor variables. *A Festschrift For Erich L. Lehmann*, page 97, 1982.

Zhiqi Bu, Jason Klusowski, Cynthia Rush, and Weijie Su. Algorithmic analysis and statistical estimation of slope via approximate message passing. In *Advances in Neural Information Processing Systems*, pages 9361–9371, 2019.

Emmanuel J Candès and Pragya Sur. The phase transition for the existence of the maximum likelihood estimate in high-dimensional logistic regression. *arXiv preprint arXiv:1804.09753*, 2018.

Michael Celentano and Andrea Montanari. Fundamental barriers to high-dimensional regression with convex penalties. *arXiv preprint arXiv:1903.10603*, 2019.

Venkat Chandrasekaran, Benjamin Recht, Pablo A Parrilo, and Alan S Willsky. The convex geometry of linear inverse problems. *Foundations of Computational Mathematics*, 12(6):805–849, 2012.

Max H. M. Costa. A new entropy power inequality. *IEEE Trans. Information Theory*, 31:751–760, 1985.

Oussama Dhifallah, Christos Thrampoulidis, and Yue M Lu. Phase retrieval via polytope optimization: Geometry, phase transitions, and new algorithms. *arXiv preprint arXiv:1805.09555*, 2018.

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

David L Donoho. Compressed sensing. *Information Theory, IEEE Transactions on*, 52(4):1289–1306, 2006.

David L Donoho, Arian Maleki, and Andrea Montanari. Message-passing algorithms for compressed sensing. *Proceedings of the National Academy of Sciences*, 106 (45):18914–18919, 2009.

David L Donoho, Arian Maleki, and Andrea Montanari. The noise-sensitivity phase transition in compressed sensing. *Information Theory, IEEE Transactions on*, 57(10):6920–6941, 2011.

Martin Genzel. High-dimensional estimation of structured signals from non-linear observations with general convex loss functions. *IEEE Transactions on Information Theory*, 63(3):1601–1619, 2017.

Martin Genzel and Peter Jung. Recovering structured data from superimposed non-linear measurements. *arXiv preprint arXiv:1708.07451*, 2017.

Larry Goldstein, Stanislav Minsker, and Xiaohan Wei. Structured signal recovery from non-linear and heavy-tailed measurements. *IEEE Transactions on Information Theory*, 64(8):5513–5530, 2018.

Yehoram Gordon. *On Milman's inequality and random subspaces which escape through a mesh in $\mathbb{R}^n$*. Springer, 1988.

Noureddine El Karoui. Asymptotic behavior of unregularized and ridge-regularized high-dimensional robust regression estimators: rigorous results. *arXiv preprint arXiv:1311.2445*, 2013.

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

Léo Miolane and Andrea Montanari. The distribution of the lasso: Uniform control over sparse balls and adaptive parameter tuning. *arXiv preprint arXiv:1811.01212*, 2018.

Ali Mousavi, Arian Maleki, Richard G Baraniuk, et al. Consistent parameter estimation for lasso and approximate message passing. *The Annals of Statistics*, 46(1):119–148, 2018.

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

Samet Oymak, Christos Thrampoulidis, and Babak Hassibi. The squared-error of generalized lasso: A precise analysis. *arXiv preprint arXiv:1311.0830*, 2013.

Yaniv Plan and Roman Vershynin. The generalized lasso with non-linear observations. *IEEE Transactions on information theory*, 62(3):1528–1537, 2016.

R Tyrrell Rockafellar. Convex analysis princeton university press. *Princeton, NJ*, 1970.

R Tyrrell Rockafellar and Roger J-B Wets. *Variational analysis*, volume 317. Springer Science & Business Media, 2009.

Fariborz Salehi, Ehsan Abbasi, and Babak Hassibi. The impact of regularization on high-dimensional logistic regression. *arXiv preprint arXiv:1906.03761*, 2019.

Mihailo Stojnic. Various thresholds for ℓ_1-optimization in compressed sensing. *arXiv preprint arXiv:0907.3666*, 2009.

Mihailo Stojnic. A framework to characterize performance of lasso algorithms. *arXiv preprint arXiv:1303.7291*, 2013.

Pragya Sur and Emmanuel J Candès. A modern maximum-likelihood theory for high-dimensional logistic regression. *Proceedings of the National Academy of Sciences*, page 201810420, 2019.

Hossein Taheri, Ramtin Pedarsani, and Christos Thrampoulidis. 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. IEEE, 2019.

Hossein Taheri, Ramtin Pedarsani, and Christos Thrampoulidis. Sharp asymptotics and optimal performance for inference in binary models. *arXiv preprint arXiv:2002.07284*, 2020.

Christos Thrampoulidis and Ankit Singh Rawat. The generalized lasso for sub-gaussian measurements with dithered quantization. *arXiv preprint arXiv:1807.06976*, 2018.

Christos Thrampoulidis, Ehsan Abbasi, and Babak Hassibi. Lasso with non-linear measurements is equivalent to one with linear measurements. In *Advances in Neural Information Processing Systems*, pages 3420–3428, 2015a.

Christos Thrampoulidis, Samet Oymak, and Babak Hassibi. Regularized linear regression: A precise analysis of the estimation error. In *Proceedings of The 28th Conference on Learning Theory*, pages 1683–1709, 2015b.

Christos Thrampoulidis, Ehsan Abbasi, and Babak Hassibi. Precise error analysis of regularized m-estimators in high dimensions. *IEEE Transactions on Information Theory*, 64(8):5592–5628, 2018a.

Christos Thrampoulidis, Weiyu Xu, and Babak Hassibi. Symbol error rate performance of box-relaxation decoders in massive mimo. *IEEE Transactions on Signal Processing*, 66(13):3377–3392, 2018b.

Joel A Tropp. Convex recovery of a structured signal from independent random linear measurements. *arXiv preprint arXiv:1405.1102*, 2014.

Haolei Weng, Arian Maleki, Le Zheng, et al. Overcoming the limitations of phase transition by higher order analysis of regularization techniques. *The Annals of Statistics*, 46(6A):3099–3129, 2018.

Ji Xu, Arian Maleki, Kamiar Rahnama Rad, and Daniel Hsu. Consistent risk estimation in high-dimensional linear regression. *arXiv preprint arXiv:1902.01753*, 2019.