The Annals of Statistics
2022, Vol. 50, No. 1, 170–196
https://doi.org/10.1214/21-AOS2100

FUNDAMENTAL BARRIERS TO HIGH-DIMENSIONAL REGRESSION WITH CONVEX PENALTIES

BY MICHAEL CELENTANO[*] AND ANDREA MONTANARI[†]

Department of Statistics, Stanford University, [*]mcelen@stanford.edu; [†]montanar@stanford.edu

In high-dimensional regression, we attempt to estimate a parameter vector $\boldsymbol{\beta}_0 \in \mathbb{R}^p$ from $n \lesssim p$ observations $\{(y_i, \boldsymbol{x}_i)\}_{i \le n}$, where $\boldsymbol{x}_i \in \mathbb{R}^p$ is a vector of predictors and y_i is a response variable. A well-established approach uses convex regularizers to promote specific structures (e.g., sparsity) of the estimate $\widehat{\boldsymbol{\beta}}$ while allowing for practical algorithms. Theoretical analysis implies that convex penalization schemes have nearly optimal estimation properties in certain settings. However, in general the gaps between statistically optimal estimation (with unbounded computational resources) and convex methods are poorly understood.

We show that when the statistican has very simple structural information about the distribution of the entries of $\boldsymbol{\beta}_0$, a large gap frequently exists between the best performance achieved by *any convex regularizer* satisfying a mild technical condition and either: *(i)* the optimal statistical error or *(ii)* the statistical error achieved by optimal approximate message passing algorithms. Remarkably, a gap occurs at high enough signal-to-noise ratio if and only if the distribution of the coordinates of $\boldsymbol{\beta}_0$ is not log-concave. These conclusions follow from an analysis of standard Gaussian designs. Our lower bounds are expected to be generally tight, and we prove tightness under certain conditions.

1. Introduction.

Consider the classical linear regression model

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta}_0 + \boldsymbol{w}, \tag{1.1}$$

where $\boldsymbol{X} \in \mathbb{R}^{n \times p}$. The statistician observes $\boldsymbol{y}$ and $\boldsymbol{X}$ but not $\boldsymbol{\beta}_0$ or $\boldsymbol{w}$, and she seeks to estimate $\boldsymbol{\beta}_0$. We assume she approximately knows the ℓ_2-norm of the noise $\boldsymbol{w}$ and the empirical distribution of the coordinates of $\boldsymbol{\beta}_0$ in senses we will make precise below.

We are interested in the high-dimensional regime in which p is comparable to n, and both are large. In this regime, computational considerations are crucial: only estimators which can be implemented by polynomial-time algorithms are relevant to statistical practice.

This paper develops precise lower bounds that characterize a broad class of estimators which are attractive in large part for their computational tractability. These are penalized least-squares estimators of the form

$$\widehat{\boldsymbol{\beta}}_{\mathsf{cvx}} \in \arg\min_{\boldsymbol{\beta}} \left\{ \frac{1}{n} \|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\|^2 + \rho(\boldsymbol{\beta}) \right\}, \tag{1.2}$$

where $\rho : \mathbb{R}^p \to \mathbb{R} \cup \{\infty\}$ is a lower semicontinuous (lsc), proper, convex function. The penalty ρ is selected to incorporate prior knowledge on the structure of $\boldsymbol{\beta}_0$ into the estimation procedure. Convexity typically yields an estimator, which is efficiently computable. Concretely, we address the following question:

Received April 2019; revised June 2021.

MSC2020 subject classifications. Primary 62J05, 62J07; secondary 62F12.

Key words and phrases. High-dimensional regression, M-estimation, computational to statistical gaps, approximate message passing, convex, penalty.

How well can we hope estimator (1.2) to perform in the high-dimensional regime by optimally designing ρ? How does this performance compare to other polynomial-time algorithms and to conjectured computational lower bounds?

The design of optimal penalties or loss functions was considered only when the distribution of the noise or—in the case of Bayesian models—the prior had log-concave density with respect to Lebesgue measure [1, 12]. Log-concavity excludes important structural assumptions, like sparsity, and, as we will show, is exactly the condition which leads to gaps between convex procedures and important computational or information-theoretic benchmarks. Thus, the case of nonlog-concave priors is both practically important and algorithmically more subtle.

We will illustrate our conclusions with two small simulation studies.

1.1. *A surprise*: *Exact recovery of a vector from three-point prior.* Consider the case of noiseless linear measurements, namely, $\boldsymbol{w} = \boldsymbol{0}$ in equation (1.1). We assume that the empirical distribution of $\boldsymbol{\beta}_0$ is known and let S be the set of vectors with that empirical distribution (i.e., vectors obtained by permuting the entries of $\boldsymbol{\beta}_0$). If we had unbounded computational resources, we would attempt reconstruction by finding $\boldsymbol{\beta} \in S$ such that $\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta}$. If only one such vector exists, then we are sure it coincides $\boldsymbol{\beta}_0$. Otherwise, exact recovery is impossible.

What is the best we can achieve by convex procedures and practical (polynomial-time) algorithms? Most researchers with a knowledge of compressed sensing or high-dimensional statistics would consider the following convex relaxation:

$$\begin{aligned} &\text{find} && \boldsymbol{\beta} \in \mathsf{conv}(S), \\ &\text{subject to} && \boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta}. \end{aligned} \tag{1.3}$$

This is the tightest possible relaxation of the combinatorial constraint $\boldsymbol{\beta} \in S$. It can be written in the form (1.2), where, setting $C := \mathsf{conv}(S)$, the penalty is $\rho(\boldsymbol{\beta}) = \mathbb{I}_C(\boldsymbol{\beta})$, and $\mathbb{I}_C(\boldsymbol{\beta}) := 0$ if $\boldsymbol{\beta} \in C$, $\mathbb{I}_C(\boldsymbol{\beta}) := \infty$ otherwise.

Notice that the approach (1.3) is at least as effective as, for instance, basis pursuit [24] which minimizes $\|\boldsymbol{\beta}\|_1$ subject to $\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta}$. To see this, notice that (for a generic $\boldsymbol{X}$) the approach (1.3) fails if and only if there exists $\boldsymbol{\beta}_*$ in the interior of $\mathsf{conv}(S)$ such that $\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta}_*$. Since $S \subseteq \{\boldsymbol{\beta} : \|\boldsymbol{\beta}\|_1 \le \|\boldsymbol{\beta}_0\|_1\}$, this implies $\|\boldsymbol{\beta}_*\|_1 < \|\boldsymbol{\beta}\|_1$ and, therefore, basis pursuit fails as well.

Is replacing the combinatorial constraint S with its tightest convex relaxation $C \equiv \mathsf{conv}(S)$ the best we can do? We report the results of a simulation study, with $p = 2000$, $n = 0.4 \cdot p = 800$. We generate a parameter vector $\boldsymbol{\beta}_0$ in which $0.75 \cdot p = 1500$ coordinates are equal to 0, $0.15p = 300$ coordinates are equal to $0.2/\sqrt{p}$, and $0.1 \cdot p = 200$ coordinates are equal to $1/\sqrt{p}$. In particular, the empirical distribution of the coordinates of $\sqrt{p}\boldsymbol{\beta}_0$ is $\pi := .75 \cdot \delta_0 + .15 \cdot \delta_{0.2} + .1 \cdot \delta_1$, which is far from being log-concave. We generate Gaussian features $(X_{ij})_{i \le n, j \le p} \overset{\text{iid}}{\sim} \mathsf{N}(0, 1)$ and response $\boldsymbol{y}$ according to linear model (1.1) with $\boldsymbol{w} = \boldsymbol{0}$.

We attempt to recover $\boldsymbol{\beta}_0$ using two different methods: (i) an accelerated proximal gradient method to solve (1.3) and (ii) a Bayes-optimal approximate message passing (Bayes-AMP) algorithm at prior π (see Section 2.2). The former is a convex optimization method, while the latter is an efficient but nonconvex procedure. We generate 500 independent realizations of the data, and for each realization, we attempt to recover $\boldsymbol{\beta}_0$ by each method. In Table 1 we report the percentage of simulations in which full recovery was achieved by each method. For 498 of the 500 realizations of the data, Bayes-AMP achieved full recovery; that is, $\widehat{\boldsymbol{\beta}} = \boldsymbol{\beta}_0$ up to machine precision. In contrast, the convex procedure never fully recovered $\boldsymbol{\beta}_0$. We also report the median, minimal, and maximal value of the relative estimation error $\|\widehat{\boldsymbol{\beta}} -$

TABLE 1

Percentage of simulations in which full recovery is achieved by convex projection (*estimator* (1.3)) *and by Bayes-AMP as well as median, minimum, and maximum value of* $\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\|^2/\|\boldsymbol{\beta}_0\|_2^2$ *observed over* 500 *independent realization of the data. Full recovery for Bayes-AMP means* $\widehat{\boldsymbol{\beta}} = \boldsymbol{\beta}_0$ *up to machine precision. "Theory lower bounds" are high-probability asymptotic lower bounds on* $\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\|^2/\|\boldsymbol{\beta}_0\|_2^2$ *for any convex procedure* (*left*) *and for Bayes-AMP* (*right*)

	Projection denoising	Bayes-AMP
% Full Recovery	0.00	99.60
Median Est. Error	0.14	0.00
Min Est. Error	0.06	0.00
Max Est. Error	0.22	0.03
Theory Lower Bounds	0.06	0.00

$\boldsymbol{\beta}_0\|^2/\|\boldsymbol{\beta}_0\|_2^2$. The relative errors displayed indicate that projection denoising never comes close to achieving exact recovery of the true parameter vector.

This study supports the perhaps surprising conclusion that estimator (1.3) is suboptimal among polynomial-time estimators for the task of noiseless recovery of a parameter vector whose coordinates have known empirical distribution π. In fact, this paper rigorously establishes a substantially more powerful conclusions, namely, that *(i) any* convex estimator of the form (1.2) will with high probability not only fail to recover the true signal, but also have estimation error lower-bounded by a constant (we refer to Section 2 for precise asymptotic statements). This lower bound is reported in Table 1. Thus, in this case full recovery is possible both information theoretically and in polynomial-time but not via convex procedures. As we will see, this gap is driven by the nonlog-concavity of π. In fact, the convex estimator (1.3) is suboptimal with respect to ℓ_2-estimation error, even among convex procedures.

In contrast to convex procedures, Bayes-AMP achieves vanishingly small reconstruction error in the current setting with probability approaching 1. Let us mention that, for noiseless or nearly noiseless observations, an alternative polynomial-time algorithm that achieves exact recovery for discrete priors was recently developed in [32]. However, the approach of [32] does not apply when the signal-to-noise ratio is of order one, which is the main focus of the present paper.

1.2. *An example*: *Noisy estimation of a sparse vector.* Gaps between the performance of convex procedures and optimal polynomial-time algorithm persist in the presence of noise. They may also occur in regimes in which all known polynomial-time algorithms are suboptimal information theoretically. To illustrate these claims, in Figure 1 we report the results of a simulation study for $p = 2000$, $n = 2000\delta$. We generated Gaussian features $(X_{ij})_{i \le n, j \le p} \overset{\text{iid}}{\sim} \mathsf{N}(0, 1)$, noise $\boldsymbol{w} \sim \mathsf{Unif}(\sqrt{n}\sigma S^{n-1})$ the uniform distribution on the sphere of radius $\sqrt{n}\sigma$ in $\mathbb{R}^n$, and $\boldsymbol{\beta}_0$ such that $0.1p$ coefficients are $1/\sqrt{p}$, $0.1p$ coefficients are $-1/\sqrt{p}$, and $0.8p$ coefficients are 0. Observe that the empirical distribution of the coordinates of $\sqrt{p}\boldsymbol{\beta}_0$ is $\pi := (\varepsilon/2)\delta_{-1} + (1-\varepsilon)\delta_0 + (\varepsilon/2)\delta_1$ with $\varepsilon = 0.2$ which is, of course, nonlog-concave. We generated response variables $\boldsymbol{y}$ according to the linear model (1.1) and attempted to estimate the parameter vector $\boldsymbol{\beta}_0$ using two different methods: (i) a convex M-estimator of the form (1.2) with a penalty $\rho(\boldsymbol{\beta})$ which was carefully optimized for the prior π and (ii) an approximate message passing (AMP) algorithm called Bayes AMP (which is optimal among AMP algorithms for the prior π, but not always Bayes optimal).

The choice of Bayes-AMP as a reference algorithm is not arbitrary. It is, in fact, justified by the following conjecture, which is motivated by ideas in statistical physics and has appeared

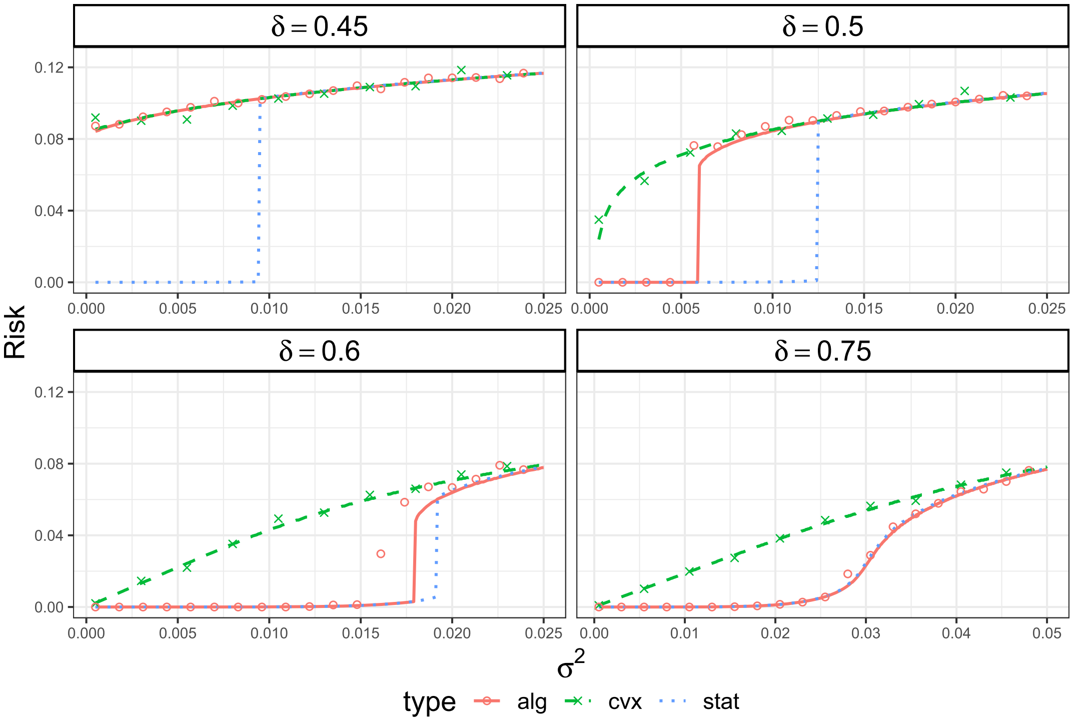

FIG. 1. *Median squared error of estimation in high-dimensional regression. Symbols refer to simulations for two different polynomial-time algorithms. Crosses: M-estimator (1.2) for a certain optimized penalty $\rho(\boldsymbol{\beta})$. Circles: Bayes-Approximate Message Passing. Dashed and solid lines correspond to our theoretical predictions for the asymptotic behavior of these algorithms. Dotted line corresponds to the asymptotics of the Bayes error; see main text for further details.*

informally several times in the literature. In the context of statistical estimation problems arising in information theory, this conjecture appears in Chapters 15 and 21 of [39]. For tutorials discussing it in the context of statistical estimation, see Sections III E and IV B of [62] and Sections 4.2 and 4.3 of [4]. For recent contributions mentioning this idea or analogous ones in the context of matrix estimation, see [5, 8, 37].

CONJECTURE 1.1. *Consider the problem of estimating $\boldsymbol{\beta}_0$ in the linear model (1.1) with standard Gaussian features $(X_{ij})_{i\le n, j\le p} \overset{\text{iid}}{\sim} \mathsf{N}(0,1)$, noise $(w_i)_{i\le n} \overset{\text{iid}}{\sim} \mathsf{N}(0,\sigma^2)$ with $\sigma > 0$, and coefficients such that $(\sqrt{p}\beta_{0,i})_{i\le p} \overset{\text{iid}}{\sim} \pi$ with π a distribution with finite second moment. Assume π is known to the statistician. Then, Bayes-AMP achieves the minimum mean square estimation error among all polynomial-time algorithms in the limit $n, p \to \infty$ with $n/p \to \delta$ fixed.*

We plot the median error under square loss achieved by these two estimators, as a function of the noise level, for four values of $\delta = n/p$. We also plot: (i') the asymptotic Bayes risk, as predicted by [7, 8, 57] (see Section 3.2); (ii') the predicted performance of Bayes-AMP (see Section 2.2); (iii') our lower bound on the risk of convex M-estimators (cf. Theorem 1). Three qualitatively different behaviors can be discerned:

- For $\delta = 0.45$, optimal convex M-estimators matches the performance of Bayes-AMP, and they are both substantially suboptimal with respect to Bayes estimation.
- For $\delta \in \{0.5, 0.6\}$, optimal convex M-estimation is suboptimal compared to Bayes AMP, and, in turn, they are both inferior to Bayes estimation.

- For $\delta = 0.75$, Bayes-AMP is Bayes optimal for all noise levels σ, and both Bayes-AMP and Bayes estimation are superior to optimal convex M-estimation.

We further note that our lower bound for convex M-estimation is nearly matched by the error achieved by the specific regularizer used in simulations. Our results rigorously establish the existence of these three qualitative behaviors and, as we will see, are driven by the nonlog-concavity of π convolved with various levels of Gaussian noise. Moreover, our convex lower bounds appear to be tight and are consistent with the conjectured computational lower bound achieved by Bayes AMP.

1.3. *Summary of contributions.*

The present paper establishes the scenario illustrated by Figure 1 and Table 1 in a precise way. Our results hold for the case of standard Gaussian features. Since convex regularizers are thought to perform well in this setting, establishing lower bounds in this case is particularly informative. Namely:

1. We prove that, for any given convex penalty, a solution to a certain system of equations provides a lower bound on the asymptotic estimation error achieved by this penalty. Further, this lower bound is tight—and hence precisely characterizes the asymptotic mean square error—if the penalty ρ is strongly convex.
2. We prove the lower bound on the error of any convex M-estimator plotted in Figure 1 and reported in Table 1. This lower bound applies to both log-concave and nonlog-concave priors for $\boldsymbol{\beta}_0$.
3. We prove that the three behaviors illustrated by Figure 1 are the only possible and that they indeed occur. Namely, the Bayes error is smaller than the Bayes-AMP error, sometimes strictly smaller, and the Bayes-AMP error is always smaller than the convex M-estimation error and sometimes strictly smaller.
4. The occurrence of these three phases is determined by the log-concavity or not of the prior convolved with Gaussian noise at a certain variance, which we specify. Importantly, nontrivial phase diagrams occur exactly when the prior is nonlog-concave. In particular, we provide a nearly complete characterization of when convex M-estimation achieves Bayes-optimal error and when it does not. In order get a quantitative understanding on the statistical-convex gap, we characterize it in the high and low signal-to-noise ratio regimes.
5. Finally, our general lower bound holds under a certain technical condition on the regularizers ρ, which we call *δ-bounded width.* We illustrate our results by considering a number of convex penalties introduced in the literature, including separable penalties, convex constraints, SLOPE, and OWL norms. We show that, in each of these cases, the bounded width condition holds.

Our work is consistent with Conjecture 1.1 in showing that no convex M-estimator of the form (1.1) can surpass the postulated lower bound on polynomial-time algorithms. Further, we believe that the characterization mentioned at the first point holds beyond strongly convex penalties: since we are mostly interested in the lower bound, we do not attempt to prove such general result.

The asymptotic characterization of Bayes-AMP is completely explicit and can be easily evaluated; hence, it can provide concrete guidance in specific problems. We expect that universality arguments [9, 36, 42] can be used to show that the same asymptotics hold for i.i.d. non-Gaussian features.

Finally, let us emphasize that we do not advocate the dismissal of convex penalization method in favor of other approaches, such as message passing algorithms. Convex algorithms present strong robustness properties that are practically important and not captured by our setting. At the same time, our work points at directions for improving their statistical properties. For instance, Section 5 shows that a suitable post-processing of a convex M-estimator can nearly bridge the gap to information-theoretically optimal performance in a large sample size regime (namely, for n/p large but of order one).

1.4. *Related literature.* By far the best-studied estimator of the form (1.2) is the Lasso [24, 59] which corresponds to the penalty $\rho(\boldsymbol{\beta}) = \lambda\|\boldsymbol{\beta}\|_1$. An impressive body of theoretical work supports the conclusion that the Lasso achieves nearly optimal performances when we know that the true vector $\boldsymbol{\beta}_0$ is sparse [15, 20, 21, 60]. Our main conclusion is that, if we attempt to exploit richer information about the empirical distribution of the coefficients $(\beta_{0,j})_{j\le p}$, then not only the Lasso, but also any convex estimator (1.2) is substantially suboptimal, as compared to the Bayes error or other polynomial-time algorithms. On the other hand, convex estimators are optimal if the coefficients distribution is log-concave.

Our work builds on a series of recent theoretical advances. First, we make use of the sharp analysis of AMP algorithms, using state evolution, which was developed in [10, 17, 35]. In particular, the recent paper [14] proves that state evolution holds for certain classes of nonseparable nonlinearities. This is particularly relevant for the present setting, since we are interested in nonseparable penalties $\rho(\boldsymbol{\beta})$.

The connection between M-estimation and AMP algorithms was first developed in [25] and subsequently used in [11] to characterize the asymptotic mean square error of the Lasso for standard Gaussian designs. The same approach was subsequently used in the context of robust regression in [27]. AMP algorithms were developed and analyzed for a number of statistical estimation problems, including generalized linear models [44], phase retrieval [38, 50], and logistic regression [54].

A different approach to sharp asymptotics in high-dimensional estimation problems makes use of Gaussian comparison inequalities. This line of work was pioneered by Stojnic [52] and then developed by a number of authors in the context of regularized regression [58], M-estimation [57], generalized compressed sensing [23], binary compressed sensing [51], the Lasso [41], and so on.

An independent approach to high-dimensional estimation, based on leave-one-out techniques, was developed by El Karoui in the context of ridge-regularized robust regression [28, 29]. Closely related to the present work is the paper [12] which considers convex M-estimation and constructs separable convex losses that match the Bayes optimal error in settings in which the noise distribution is log-concave and hence the gap between the two vanishes. Our work extends this analysis to cases in which log-concavity assumptions are violated so that the Bayes error cannot be achieved. In this paper we focus on the role of regularization rather than the loss function, though we suspect similar analyses should be possible for general convex losses. Optimal convex M-estimators were also studied, using tools from statistical physics—in [1].

As mentioned above, we compare the performance of convex M-estimators to the optimal Bayes error and conjectured computational lower bounds. The asymptotic value of the Bayes error for random designs was recently determined in [6, 47]. Generalizations of this result were also obtained in [7] for other regression problems.

Finally, the gap between polynomial-time algorithms and statistically optimal estimators has been studied from other points of view as well. It was noted early on that constrained least square methods (which exhaustively search over supports of given size) perform accurate regression under weaker conditions than required by the Lasso [61]. Strong lower bounds for compressed sensing reconstruction were proved in [3], using communication complexity ideas. Gamarnik and Zadik [32] study the case of binary coefficients, namely, $\boldsymbol{\beta}_0 \in \{0, 1\}^p$, and standard Gaussian designs $\boldsymbol{X}$. They prove existence of a gap between the maximum likelihood estimator (which requires exhaustive search over binary vectors) and the Lasso. They argue that the failure of polynomial-time algorithms originates in a certain "overlap gap property," which they also characterize. Further implications of this point of view are investigated in [33]. After a preprint of this paper appeared online, further work studied the design of optimal penalties and loss functions in classification models and analyzed the achievability of Bayes optimal performance [40, 55, 56].

1.5. *Notations.* The Euclidean norm of a vector $\boldsymbol{x} \in \mathbb{R}^p$ is denoted by $\|\boldsymbol{x}\| := \|\boldsymbol{x}\|_2$. The operator and nuclear norms of a matrix $\boldsymbol{X} \in \mathbb{R}^{n\times p}$ are denoted by $\|\boldsymbol{X}\|_{\mathsf{op}}$ and $\|\boldsymbol{X}\|_{\mathsf{nuc}}$, respectively. We denote by S_+^k the set of $k \times k$ positive semidefinite matrices.

Subscripts under the expectation or probability sign, for example, $\mathbb{E}_{\boldsymbol{\beta}_0,\boldsymbol{z}}$ and $\mathbb{P}_{\boldsymbol{\beta}_0,\boldsymbol{z}}$ indicate the variables which are random. We denote by $\mathcal{P}_k(\mathbb{R})$ the collection of Borel probability measures on $\mathbb{R}$ with finite kth moment. For a distribution $\pi \in \mathcal{P}_k(\mathbb{R})$, we will denote by $s_\ell(\pi)$ the ℓth moment of π. We will often extend a distribution $\pi \in \mathcal{P}_k(\mathbb{R})$ to a distribution on $\mathbb{R}^p$ by taking $\boldsymbol{\beta}_0 = (\beta_{0j})_{j\le p} \in \mathbb{R}^p$ with coordinates such that $(\sqrt{p}\beta_{0j})_{j\le p} \overset{\text{iid}}{\sim} \pi$. We will write this succinctly as $\beta_{0j} \overset{\text{iid}}{\sim} \pi/\sqrt{p}$. Under this normalization, $\mathbb{E}_{\boldsymbol{\beta}_0}[\|\boldsymbol{\beta}_0\|^2] = s_2(\pi)$ does not depend on p. We reserve z and $\boldsymbol{z}$ to denote Gaussian random variables and vectors, respectively. We will always take $z \sim \mathsf{N}(0,1)$ and $\boldsymbol{z} \sim \mathsf{N}(0, \boldsymbol{I}_p/p)$. Convolution of probability measures will be denoted by $*$.

We define the Wasserstein distance between two probability measures $\pi, \pi' \in \mathcal{P}_2(\mathbb{R})$ by

$$d_{\mathrm{W}}(\pi,\pi') = \inf_{X,X'} \left(\mathbb{E}_{X,X'}[(X-X')^2]\right)^{1/2},$$

where the infimum is taken over joint distributions of random variables (X, X') with marginal distributions $X \sim \pi$ and $X' \sim \pi'$. It is well known that this defines a metric on $\mathcal{P}_2(\mathbb{R})$ [48]. Convergence in Wasserstein metric will be denoted $\xrightarrow{\mathrm{W}}$, and we use $\xrightarrow{\mathrm{p}}$, $\xrightarrow{\mathrm{as}}$, $\xrightarrow{\mathrm{d}}$ for other standard notions of convergence. For any sequence of real-valued random variables $\{X_p\}$, not necessarily defined on the same probability space, we denote

$$\operatorname*{lim\,inf}^{\mathrm{p}}_{p\to\infty} X_p = \sup\Big\{t \in \mathbb{R} \mid \lim_{p\to\infty} \mathbb{P}(X_p < t) = 0\Big\}$$

and $\limsup^{\mathrm{p}}_{p\to\infty} X_p = -\liminf^{\mathrm{p}}_{p\to\infty}(-X_p)$. For sequences $\{X_p\}$ and $\{Y_p\}$ of real-valued random variables such that, for each p, X_p and Y_p are defined on the same probability space, we use the notation $X_p \overset{\mathrm{p}}{\simeq} Y_p$ to denote $|X_p - Y_p| \xrightarrow{\mathrm{p}} 0$.

We adopt the convention that when the minimizing set in (1.2) is empty, $\widehat{\boldsymbol{\beta}}_{\mathsf{cvx}} = \infty$ and $\|\infty - \boldsymbol{x}\| = \infty$ for any $\boldsymbol{x} \in \mathbb{R}^p$. Thus, the estimation error is infinite when no minimizer exists.

Finally, a collection of functions $\{\varphi : (\mathbb{R}^p)^\ell \to \mathbb{R}^m\}$, where p and m but not ℓ may vary, is said to be *uniformly pseudo-Lipschitz of order k* if for all φ and $\boldsymbol{x}_i, \boldsymbol{y}_i \in \mathbb{R}^p$, $i = 1,\dots,\ell$, we have

$$\|\varphi(\boldsymbol{x}_1,\dots,\boldsymbol{x}_\ell) - \varphi(\boldsymbol{y}_1,\dots,\boldsymbol{y}_\ell)\| \le C\left(1 + \sum_{i=1}^{\ell}\|\boldsymbol{x}_i\|^{k-1} + \sum_{i=1}^{\ell}\|\boldsymbol{y}_i\|^{k-1}\right)\sum_{i=1}^{\ell}\|\boldsymbol{x}_i - \boldsymbol{y}_i\|$$

for some C which does not depend on p, m.

2. The convex lower bound, the risk of Bayes-AMP, and the Bayes risk. In this section we present a rigorous lower bound on the ℓ_2 estimation error of convex M-estimators of the form (1.2) under proportional asymptotics, Gaussian noise, and structural assumptions on the unknown parameter $\boldsymbol{\beta}_0$. A primary focus will be comparing the convex lower bound to two important benchmarks which have been studied elsewhere [6, 7, 47]:

- *Risk of Bayes-AMP*: The ℓ_2-estimation error of a certain message passing algorithm conjectured to be optimal among all polynomial-time algorithms (see Conjecture 2.5).

- *Bayes risk*: The optimal risk over all (possibly computationally unbounded) estimators under a certain Bayesian model for the signal.

Before defining these quantities precisely, we may summarize the comparison we will establish by

$$\begin{matrix}\text{Convex}\\ \text{Lower Bound}\end{matrix} \geq \begin{matrix}\text{Risk of}\\ \text{Bayes AMP}\end{matrix} \geq \text{Bayes Risk.}$$

While the second inequality holds by the statistical optimality of the Bayes risk, the first is nontrivial. Previous work established exactly when the second inequality is strict [7]. We will likewise specify exactly when the first inequality is strict. Previous work has only considered optimal convex estimation in regimes in which strict inequality does not occur [1, 12].

Precisely, we study these three quantities under a certain high-dimensional proportional asymptotics for model (1.1):

High-dimensional asymptotics (HDA) The design matrix satisfies the following assumptions:

- The sample size and number of parameters $n, p \to \infty$ satisfy $n/p \to \delta \in (0, \infty)$, a fixed asymptotic aspect ratio.
- The matrix $\boldsymbol{X}$ has entries $X_{ij} \overset{\text{iid}}{\sim} \mathsf{N}(0, 1)$.

Further, we introduce two sets of assumptions on the unknown parameter $\boldsymbol{\beta}_0$ and the the noise $\boldsymbol{w}$.

Deterministic signal and noise (DSN) For each p and n, we have deterministic parameter vector $\boldsymbol{\beta}_0 \in \mathbb{R}^p$ and noise vector $\boldsymbol{w} \in \mathbb{R}^n$. For some $\pi \in \mathcal{P}_2(\mathbb{R})$ and $\sigma^2 \geq 0$, these satisfy

$$\widehat{\pi}_{\boldsymbol{\beta}_0} := \frac{1}{p}\sum_{j=1}^{p} \delta_{\sqrt{p}\beta_{0j}} \xrightarrow{\text{W}} \pi \quad \text{and} \quad \frac{1}{n}\|\boldsymbol{w}\|^2 \to \sigma^2.$$

Random signal and noise (RSN) assumption For each p and n, we have random parameter vector $\boldsymbol{\beta}_0 \in \mathbb{R}^p$ and noise vector $\boldsymbol{w} \in \mathbb{R}^n$ satisfying

$$\beta_{0j} \overset{\text{iid}}{\sim} \pi/\sqrt{p}, \qquad \boldsymbol{w} \sim \mathsf{N}(0, \sigma^2 \boldsymbol{I}_n),$$

where $\pi \in \mathcal{P}_2(\mathbb{R})$ and $\sigma^2 \geq 0$ do not depend on p.

When necessary to indicate where $\boldsymbol{\beta}_0$ $\boldsymbol{w}$ fall in the sequence of realizations with growing dimensions, we include indices as $\boldsymbol{\beta}_0(p)$ and $\boldsymbol{w}(p)$.

Under the DSN assumption we will establish a convex lower bound for *symmetric* convex penalties, that is, penalties which are invariant to permutation of the coordinates of their argument. The DSN assumption specifies the limiting empirical distribution of the coordinates of $\boldsymbol{\beta}_0$, which captures structural information, like sparsity, which is permutation invariant. Nevertheless, the lower bound applies also to models in which additional information about the order in which the coordinates appear is available: for example, the statistician may know that the coordinates are monotone, have sparse first differences, or satisfy other smoothness conditions. The lower bound, which applies only to symmetric convex penalties, describes a limitation of convex procedures which fail to exploit such information.

In contrast, under the RSN assumption we will establish a convex lower bound for *arbitrary* convex penalties. Here, the statistician can exploit all available information. But because she has no prior knowledge about the ordering of the coordinates of $\boldsymbol{\beta}_0$, she cannot benefit from asymmetric procedures.

The two sets of assumptions are complementary, differing in how they impose symmetry on the problem—either through the method or through the model. It turns out that the lower bound on the estimation error under the two sets assumptions is the same.

We only make comparisons to information theoretic lower bounds, that is, the Bayes risk, under the RSN assumption. Indeed, the RSN assumption is needed for the Bayes risk to be meaningful.

2.1. *The convex lower bound.* The convex lower bound is defined via a comparison of the linear model (1.1) to a simpler Gaussian sequence model. In the sequence model we observe

$$\boldsymbol{y}_{\mathsf{seq}} = \boldsymbol{\beta}_0 + \tau \boldsymbol{z}, \tag{2.1}$$

where $\beta_{0j} \overset{\text{iid}}{\sim} \pi/\sqrt{p}$, $\boldsymbol{z} \sim \mathsf{N}(\boldsymbol{0}, \boldsymbol{I}_p/p)$ independent and $\tau^2 \geq 0$. Analogously to (1.2), we consider convex M-estimators in the sequence model, also known as *proximal operators*,

$$\widehat{\boldsymbol{\beta}}_{\mathsf{seq}} := \arg\min_{\boldsymbol{\beta}} \frac{1}{2}\|\boldsymbol{y}_{\mathsf{seq}} - \boldsymbol{\beta}\|^2 + \lambda\rho(\boldsymbol{\beta}) =: \mathsf{prox}[\lambda\rho](\boldsymbol{y}_{\mathsf{seq}}). \tag{2.2}$$

By strong convexity, when ρ is lower semicontinuous and proper, the minimizer exists and is unique [43].

A large body of work exactly characterizes the estimation error of the estimators (1.2) in the linear model in terms of the behavior of the estimators (2.2) in the sequence model [11, 27, 28, 30, 57, 58]. A typical characterization takes the following form. For a sequence of penalties $\{\rho_p\}$, let (τ, λ) solve

$$\begin{aligned} &\delta\tau^2 - \sigma^2 = \lim_{p\to\infty} \mathbb{E}_{\boldsymbol{z}}\big[\big\|\mathsf{prox}[\lambda\rho_p](\boldsymbol{\beta}_0 + \tau\boldsymbol{z}) - \boldsymbol{\beta}_0\big\|^2\big], \\ &2\lambda\bigg(1 - \frac{1}{\delta\tau}\lim_{p\to\infty} \mathbb{E}_{\boldsymbol{z}}\big[\big\langle \boldsymbol{z}, \mathsf{prox}[\lambda\rho_p](\boldsymbol{\beta}_0 + \tau\boldsymbol{z})\big\rangle\big]\bigg) = 1. \end{aligned} \tag{2.3}$$

Then, under the HDA and DSN assumption,

$$\|\widehat{\boldsymbol{\beta}}_{\mathsf{cvx}} - \boldsymbol{\beta}_0\|^2 \xrightarrow{\mathrm{p}} \delta\tau^2 - \sigma^2 = \mathbb{E}_{\boldsymbol{z}}\big[\big\|\mathsf{prox}[\lambda\rho_p](\boldsymbol{\beta}_0 + \tau\boldsymbol{z}) - \boldsymbol{\beta}_0\big\|^2\big].$$

In words, the ℓ_2 estimation error in the linear model asymptotically agrees with the ℓ_2 risk in the sequence model at noise variance τ^2 and regularization λ. Substantial effort is required to make this rigorous, and many technical assumptions are required. For example, some work requires strong-convexity assumptions on the cost function (1.2) [27, 28]; other work involves analysis tailored to a specific penalty, like the LASSO or SLOPE [11, 19]. We instead provide a lower bound on the estimation error of estimators (1.2) which holds simultaneously for a large class of penalties. We rely on weak assumptions—weaker than what is needed for exact characterizations using existing techniques. At a high level the lower bound follows from controlling the possible solutions to equation (2.3) and applying exact characterization results.

Denote by $\mathcal{C}_p \subseteq \{\rho : \mathbb{R}^p \to \mathbb{R} \cup \{\infty\}\}$ any collection of lsc, proper, and convex functions which is closed under scaling; that is, $\rho_p \in \mathcal{C}_p$ implies $\lambda\rho_p \in \mathcal{C}_p$ for all $\lambda > 0$. Denote by $\mathcal{C}$ the collection of sequences $\{\rho_p\}_p$ such that $\rho_p \in \mathcal{C}_p$ for all p. We will mostly be interested in two cases: either $\mathcal{C}$ consists of all the sequences of convex functions or it consists of all convex symmetric functions.

The optimal risk of convex M-estimation using collection $\mathcal{C}_p$ in the sequence model is

$$\mathsf{R}^{\mathsf{opt}}_{\mathsf{seq,cvx}}(\tau; \pi, p) := \inf_{\rho\in\mathcal{C}_p} \mathbb{E}_{\boldsymbol{\beta}_0,\boldsymbol{z}}\big[\big\|\mathsf{prox}[\rho](\boldsymbol{\beta}_0 + \tau\boldsymbol{z}) - \boldsymbol{\beta}_0\big\|^2\big], \tag{2.4}$$

where $\boldsymbol{\beta}_0$, $\boldsymbol{z}$ are as in (2.1), and the optimal asymptotic risk using the sequences in $\mathcal{C}$ is

$$\begin{aligned}\mathsf{R}^{\mathsf{opt}}_{\mathsf{seq,cvx}}(\tau;\pi) &= \liminf_{p\to\infty} \mathsf{R}^{\mathsf{opt}}_{\mathsf{seq,cvx}}(\tau;\pi,p) \\ &= \inf_{\{\rho_p\}\in\mathcal{C}} \liminf_{p\to\infty} \mathbb{E}_{\boldsymbol{\beta}_0,\boldsymbol{z}}\big[\|\mathsf{prox}[\rho_p](\boldsymbol{\beta}_0+\tau\boldsymbol{z})-\boldsymbol{\beta}_0\|^2\big].\end{aligned}\tag{2.5}$$

We will study a quantity similar to (2.5) in the linear model (1.1), except that the infimum is taken over a slightly more restrictive collection, which we now define.

DEFINITION 2.1. For $\pi\in\mathcal{P}_2(\mathbb{R})$ and $\delta\in(0,\infty)$, we say a sequence of lsc, proper, convex functions $\{\rho_p\}$ has δ-*bounded width* at prior π, if the following holds:

$$\begin{gathered}\text{for all compact } T\subset(0,\infty), \text{ there exists } \bar\lambda=\bar\lambda(T)<\infty \text{ such that} \\ \limsup_{p\to\infty}\sup_{\lambda>\bar\lambda,\tau\in T}\frac{1}{\tau}\mathbb{E}_{\boldsymbol{\beta}_0,\boldsymbol{z}}\big[\langle\boldsymbol{z},\mathsf{prox}[\lambda\rho_p](\boldsymbol{\beta}_0+\tau\boldsymbol{z})\rangle\big]<\delta.\end{gathered}\tag{2.6}$$

For a collection of penalty sequences $\mathcal{C}$, we denote by $\mathcal{C}_{\delta,\pi}$ the subset of sequences that satisfy this condition.

The terminology here is motivated by the resemblance of condition (2.6) with the Gaussian width of convex cones [2, 23]; see Section 6.2. It is straightforward to show that, for $\delta>1$ and any $\pi\in\mathcal{P}_2(\mathbb{R})$, all sequences of penalties have δ-bounded width at π (see Section O, equation (O.11) of the Supplementary Material [22]). Thus,

$$\mathcal{C}_{\delta,\pi}=\mathcal{C}\quad\text{if } \delta>1.$$

The convex lower bound we establish in the next theorem applies to sequences of penalties in $\mathcal{C}_{\delta,\pi}$.

THEOREM 1. *Fix* $\pi\in\mathcal{P}_2(\mathbb{R})$, $\delta\in(0,\infty)$, *and* $\sigma\ge0$. *Define*

$$\tau^2_{\mathsf{reg,cvx}}=\sup\{\tau^2\mid\delta\tau^2-\sigma^2<\mathsf{R}^{\mathsf{opt}}_{\mathsf{seq,cvx}}(\tau;\pi)\}.\tag{2.7}$$

Under the HDA and RSN assumptions,[1]

$$\inf_{\{\rho_p\}\in\mathcal{C}_{\delta,\pi}}\underset{p\to\infty}{\overset{\mathrm{p}}{\liminf}}\|\widehat{\boldsymbol{\beta}}_{\mathsf{cvx}}-\boldsymbol{\beta}_0\|^2\ge\delta\tau^2_{\mathsf{reg,cvx}}-\sigma^2.$$

If $\mathcal{C}$ *contains only symmetric penalties, then the preceding display holds also under DSN assumption.* (*Note that we may have* $\tau^2_{\mathsf{reg,cvx}}=\infty$.)

In both cases, for $\delta>1$, *the infimum can be taken over the full collection* $\mathcal{C}$ (*instead of* $\mathcal{C}_{\delta,\pi}$), *and the lower bound is tight.*

The proof of Theorem 1 is provided in Section E of the Supplementary Material [22]. In Section 6 we argue through examples that $\mathcal{C}_{\delta,\pi}$ includes most, if not all, reasonable penalty sequences. Section I of the Supplementary Material [22] discusses the role of the restriction to $\mathcal{C}_{\delta,\pi}$. Because $\mathsf{R}^{\mathsf{opt}}_{\mathsf{seq,cvx}}(\tau;\pi)$ is continuous in τ whenever $\mathcal{C}$ is such that $\tau^2_{\mathsf{reg,cvx}}$ is finite (see Lemma C.2 of the Supplementary Material [22]), we will always have $\delta\tau^2_{\mathsf{reg,cvx}}-\sigma^2=\mathsf{R}^{\mathsf{opt}}_{\mathsf{seq,cvx}}(\tau_{\mathsf{reg,cvx}};\pi)$ in this case. Thus, Theorem 1 should be interpreted as stating,

> *Optimal convex M-estimation in the linear model is no better than optimal convex M-estimation in the sequence model at noise variance* $\tau^2_{\mathsf{reg,cvx}}$.

[1] When the minimizing set has multiple elements, we make no assumption on the mechanism used to break ties.

Importantly, the convex lower bound applies even when π is not log-concave.

Although Theorem 1 applies to any potentially restricted collection $\mathcal{C}$ of convex penalty sequences, our main interest is to apply it to the largest possible collections. This is because we are interested in studying *fundamental* barriers to regression with any convex estimators of the form (1.2). Thus, for the remainder of the paper we will consider only two cases: under the RSN assumption, we will consider $\mathcal{C}$ to contain all sequences of convex penalties. In this case, $\{\rho_p\} \in \mathcal{C}_{\delta,\pi}$ contains any sequence of penalties satisfying (2.6). Under the DSN assumption we will consider $\mathcal{C}$ to contain all sequences of symmetric convex penalties. In this case, $\{\rho_p\} \in \mathcal{C}_{\delta,\pi}$ contains any sequence of symmetric penalties satisfying (2.6). The convex lower bound in these two cases is the same.

PROPOSITION 2.2. *The parameter $\tau^2_{\mathsf{reg,cvx}}$ defined with $\mathcal{C}$ all sequences of convex penalties or with $\mathcal{C}$ all sequences of symmetric convex penalties agree.*

Although we consider two cases throughout the remainder of the paper, there is only one fundamental convex lower bound, and it applies to both cases. In the first case—that described by the RSN assumption—the statistician has no information about the order in which the coordinates of the unknown parameter occur, and the convex lower bound applies to any convex procedure. In the second case—that described by the DSN assumption—the statistician may have information about the order in which the coordinates of the unknown parameter occur, and the convex lower bound applies only to symmetric convex procedures. Thus, the convex lower bound applies either to settings in which information about the order of the coordinates is not available or to settings where such information is not exploited.

2.2. *The risk of Bayes AMP.* Bayes AMP, which we define below, is a fast iterative scheme for performing estimation in model (1.1). Analogously to the convex lower bound, its estimation error is defined via a comparison of the linear model (1.1) to the sequence model (2.1). In particular, define

$$\mathsf{mmse}_\pi(\tau^2) = \mathbb{E}_{\beta_0,z}\big[\big(\mathbb{E}_{\beta_0,z}[\beta_0 \mid \beta_0 + \tau z] - \beta_0\big)^2\big],$$

for random scalars $\beta_0 \sim \pi$, $z \sim \mathsf{N}(0,1)$ independent. Because

$$\mathsf{mmse}_\pi(\tau^2) = \mathbb{E}_{\boldsymbol{\beta}_0,z}\big[\big\|\mathbb{E}_{\boldsymbol{\beta}_0,z}[\boldsymbol{\beta}_0 \mid \sqrt{p}\boldsymbol{\beta}_0 + \tau z] - \boldsymbol{\beta}_0\big\|^2\big], \tag{2.8}$$

we see that $\mathsf{mmse}_\pi(\tau^2)$ is analogous to (2.4), except that the infimum is taken over all estimators, not just those in a restricted class. Finally, analogous to (2.7), define

$$\tau^2_{\mathsf{reg,amp*}} := \sup\{\tau^2 \mid \delta\tau^2 - \sigma^2 \le \mathsf{mmse}_\pi(\tau^2)\}. \tag{2.9}$$

Note that because $\mathsf{mmse}_\pi(\tau^2)$ is continuous in τ [34],

$$\delta\tau^2_{\mathsf{reg,amp*}} - \sigma^2 = \mathsf{mmse}_\pi(\tau^2_{\mathsf{reg,amp*}}). \tag{2.10}$$

As we will see, Bayes AMP asymptotically achieves estimation error arbitrarily close to $\delta\tau^2_{\mathsf{reg,amp*}} - \sigma^2 = \mathsf{mmse}_\pi(\tau^2_{\mathsf{reg,amp*}})$ in time $O(np)$. That is,

> *Bayes AMP in the linear model is exactly as good as Bayesian estimation in the sequence model at noise variance $\tau^2_{\mathsf{reg,amp*}}$.*

Thus, a comparison of the convex lower bound and the risk of Bayes AMP reduces to a comparison of the parameters $\tau^2_{\mathsf{reg,cvx}}$ and $\tau^2_{\mathsf{reg,amp*}}$. The following corollary of Theorem 1 establishes under generic conditions, the convex lower bound is no smaller than the estimation error of Bayes AMP, consistent with conjectured optimality of Bayes AMP among polynomial time algorithms.

COROLLARY 2.3. *For any $\pi \in \mathcal{P}_2(\mathbb{R})$,*

$$\tau^2_{\mathsf{reg,cvx}} \geq \tau^2_{\mathsf{reg,amp*}} \tag{2.11}$$

holds for almost every value of δ, σ (w.r.t. Lebesgue measure). In fact, for any fixed σ, it holds for almost all values of δ, and for any fixed δ, for almost all values of σ.

For such values δ, σ, under the HDA and RSN assumptions, then

$$\inf_{\{\rho_p\} \in \mathcal{C}_{\delta,\pi}} \overset{\mathrm{p}}{\liminf_{p\to\infty}} \|\widehat{\boldsymbol{\beta}}_{\mathsf{cvx}} - \boldsymbol{\beta}_0\|^2 \geq \delta \tau^2_{\mathsf{reg,amp*}} - \sigma^2. \tag{2.12}$$

If $\mathcal{C}$ contains only symmetric penalties, then the preceding display holds instead under DSN assumption.

PROOF OF COROLLARY 2.3. Define

$$\tau^2_{\mathsf{reg,amp}} = \sup\{\tau^2 \mid \delta\tau^2 - \sigma^2 < \mathsf{mmse}_\pi(\tau^2)\}. \tag{2.13}$$

In Section L of the Supplementary Material [22], we show that, for any $\pi \in \mathcal{P}_2(\mathbb{R})$, the equality $\tau^2_{\mathsf{reg,amp}} = \tau^2_{\mathsf{reg,amp*}}$ holds for almost every value of δ, σ (w.r.t. Lebesgue measure). In fact, for any fixed σ, it holds for almost all values of δ, and for any fixed δ, for almost all values of σ. Thus, we only need to establish the result for $\tau^2_{\mathsf{reg,amp}}$ in place of $\tau^2_{\mathsf{reg,amp*}}$.

By (2.4) and (2.8), we have $\mathsf{mmse}_\pi(\tau^2) \leq \mathsf{R}^{\mathsf{opt}}_{\mathsf{seq,cvx}}(\tau; \pi, p)$. By (2.5)), we obtain $\mathsf{mmse}_\pi(\tau^2) \leq \mathsf{R}^{\mathsf{opt}}_{\mathsf{seq,cvx}}(\tau; \pi)$. Thus, the set $\{\tau^2 \mid \delta\tau^2 - \sigma^2 < \mathsf{mmse}_\pi(\tau^2)\} \subseteq \{\tau^2 \mid \delta\tau^2 - \sigma^2 < \mathsf{R}^{\mathsf{opt}}_{\mathsf{seq,cvx}}(\tau^2; \pi)\}$, and (2.11) follows from (2.7) and (2.13). Theorem 1 then gives (2.12). □

In the remainder of this section, we describe the Bayes AMP algorithm and formally characterize its risk. Bayes AMP and its characterization via state evolution has been derived elsewhere [7, 26]. Define the scalar iteration

$$\begin{aligned} \tau_0^2 &= \frac{1}{\delta}\big(\sigma^2 + s_2(\pi)\big), \\ \tau_{t+1}^2 &= \frac{1}{\delta}\big(\sigma^2 + \mathsf{mmse}_\pi(\tau_t^2)\big). \end{aligned} \tag{2.14}$$

Moreover, let

$$\eta_t(y) = \mathbb{E}_{\beta_0,z}[\beta_0 \mid \beta_0 + \tau_t z = y], \tag{2.15}$$

where $\beta_0 \sim \pi$, $z \sim \mathsf{N}(0,1)$ are independent. Define

$$\mathsf{b}_t = \frac{1}{\delta}\mathbb{E}_{\beta_0,z}\big[\eta'_{t-1}(\beta_0 + \tau_{t-1} z)\big],$$

where η'_t a weak derivative of η_t. For each p, define $\eta_t : \mathbb{R}^p \to \mathbb{R}^p$ by

$$\eta_t(\boldsymbol{x})_j = \frac{1}{\sqrt{p}}\eta_t(\sqrt{p}x_j),$$

where, for convenience, we use the same notation η_t for the multivariate and scalar functions. They are distinguished by the nature of their argument. The Bayes-AMP iteration is

$$\begin{aligned} \boldsymbol{r}^t &= \frac{\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}^t}{n} + \mathsf{b}_t \boldsymbol{r}^{t-1}, \\ \widehat{\boldsymbol{\beta}}^{t+1} &= \eta_t\big(\widehat{\boldsymbol{\beta}}^t + \boldsymbol{X}^\mathsf{T}\boldsymbol{r}^t\big), \end{aligned} \tag{2.16}$$

with initialization $\widehat{\boldsymbol{\beta}}^0 = \boldsymbol{0}$, $\boldsymbol{r}^{-1} = \boldsymbol{0}$. For any fixed t, we may compute $\widehat{\boldsymbol{\beta}}^t$ in $O(np)$ time. The following proposition characterizes the asymptotic loss of $\widehat{\boldsymbol{\beta}}^t$ as an estimator of $\boldsymbol{\beta}_0$.

PROPOSITION 2.4. *Fix* $\pi \in \mathcal{P}_2(\mathbb{R})$, $\delta \in (0, \infty)$, *and* $\sigma \geq 0$. *Assume* $s_2(\pi) > 0$. *Consider* τ_t *as defined by* (2.14) *and* $\widehat{\boldsymbol{\beta}}^t$ *as defined by* (2.16). *Under the HDA and either the DSN or RSN assumptions, for any fixed t we have*

$$\operatorname*{\overset{p}{\lim}}_{p\to\infty} \|\widehat{\boldsymbol{\beta}}^t - \boldsymbol{\beta}_0\|^2 = \mathsf{mmse}_\pi(\tau_t^2).$$

Further,

$$\lim_{t\to\infty} \tau_t^2 = \tau^2_{\mathsf{reg,amp}*}.$$

In particular, for all $\varepsilon > 0$, *there exists t fixed such that*

$$\operatorname*{\overset{p}{\lim}}_{p\to\infty} \|\widehat{\boldsymbol{\beta}}^t - \boldsymbol{\beta}_0\|^2 \leq \delta\tau^2_{\mathsf{reg,amp}*} - \sigma^2 + \varepsilon.$$

Proposition 2.4 states that the state evolution (2.14) characterizes the large n, p behavior of Bayes AMP. It follows from standard results in the AMP literature [10]. A minor technical difficulty is that the main theorem of [10] requires Lipschitz nonlinearities in the AMP iteration. The Bayes estimator η_t need not be Lipschitz. Thus, to apply the results of [10], we must use a truncation trick. Though this is a routine proof, we are unaware of a result that immediately implies Proposition 2.4. For completeness, we provide this argument in Section L of the Supplementary Material [22].

Proposition 2.4 shows that a polynomial-time (in fact, linear time) algorithm exists which achieves asymptotic loss arbitrarily close to $\delta\tau^2_{\mathsf{reg,amp}*} - \sigma^2$. As discussed in the Introduction, we do not know of any polynomial-time algorithm that achieves asymptotic risk below $\delta\tau^2_{\mathsf{reg,amp}*} - \sigma^2$. Below is a more precise restatement of Conjecture 1.1.

CONJECTURE 2.5. *Fix* $\pi \in \mathcal{P}_2(\mathbb{R})$, $\delta \in (0, \infty)$, *and* $\sigma > 0$. *Under the HDA and RSN assumptions at* π, δ, σ, *no polynomial-time algorithm achieves asymptotic risk smaller than* $\delta\tau^2_{\mathsf{reg,amp}*} - \sigma^2$.

2.3. *The Bayes risk.* The information theoretic lower bound under the RSN assumption is the Bayes risk

$$\mathbb{E}_{\boldsymbol{\beta}_0,\boldsymbol{w},\boldsymbol{X}}\big[\|\mathbb{E}_{\boldsymbol{\beta}_0,\boldsymbol{w},\boldsymbol{X}}[\boldsymbol{\beta}_0 \mid \boldsymbol{y}, \boldsymbol{X}] - \boldsymbol{\beta}_0\|^2\big],$$

which cannot be outperformed, even in finite samples. In this section we recall recent results on the asymptotic value of the Bayes risk on the HDA and RSN assumptions.

Define the potential

$$\phi(\tau^2; \pi, \delta, \sigma) = \frac{\sigma^2}{2\tau^2} - \frac{\delta}{2}\log\left(\frac{\sigma^2}{\tau^2}\right) + i(\tau^2), \tag{2.17}$$

where $i(\tau^2)$ is the base-e mutual information between β_0 and y in the univariate model $y = \beta_0 + \tau z$ when $\beta_0 \sim \pi$, $z \sim \mathsf{N}(0, 1)$ independent. That is,

$$i(\tau^2) = \mathbb{E}_{\beta_0,z}\left[\log \frac{p(y \mid \beta_0)}{p(y)}\right] = -\frac{1}{2} - \mathbb{E}_{\beta_0,z}\log\left\{\int e^{-\frac{1}{2}(y-\beta/\tau)^2}\pi(\mathrm{d}\beta)\right\}.$$

Also, define

$$\tau_{\mathsf{reg,stat}}(\pi; \delta, \sigma) = \arg\min_{\tau\geq 0} \phi(\tau^2; \pi, \delta, \sigma), \tag{2.18}$$

whenever π, δ, and σ are such that the minimizer is unique. The derivative of ϕ will be useful in what follows. It is

$$\frac{\mathrm{d}}{\mathrm{d}\tau^{-2}}\phi(\tau^2; \pi, \delta, \sigma) = \frac{1}{2}(\sigma^2 - \delta\tau^2 + \mathsf{mmse}_\pi(\tau^2)), \tag{2.19}$$

where we have used that $\frac{\mathrm{d}}{\mathrm{d}\tau^{-2}} i(\tau^2) = \frac{1}{2}\mathsf{mmse}_\pi(\tau^2)$ by [34], Corollary 1. We see that if $\tau_{\mathsf{reg,stat}} > 0$, then

$$\delta\tau^2_{\mathsf{reg,stat}} - \sigma^2 = \mathsf{mmse}_\pi(\tau^2_{\mathsf{reg,stat}}). \tag{2.20}$$

Equation (2.20) is closely related to (2.13). The next result relates the effective noise parameter $\tau_{\mathsf{reg,stat}}$ to the asymptotic Bayes risk in model (1.1) under the RSN assumption.

PROPOSITION 2.6 (Theorem 2 of [7]). *Fix $\pi \in \mathcal{P}_\infty(\mathbb{R})$, $\delta \in (0,\infty)$, and $\sigma > 0$. Under the HDA and RSN assumptions,*

$$\lim_{p\to\infty} \mathbb{E}_{\boldsymbol{\beta}_0,\boldsymbol{w},\boldsymbol{X}}\big[\|\mathbb{E}_{\boldsymbol{\beta}_0,\boldsymbol{w},\boldsymbol{X}}[\boldsymbol{\beta}_0 \mid \boldsymbol{y},\boldsymbol{X}] - \boldsymbol{\beta}_0\|^2\big] = \mathsf{mmse}_\pi(\tau^2_{\mathsf{reg,stat}}) = \delta\tau^2_{\mathsf{reg,stat}} - \sigma^2, \tag{2.21}$$

whenever the minimizer in (2.18) is unique. This occurs for almost every (δ,σ) (w.r.t. Lebesgue measure).

This is a specific case of Theorem 2 of [7]. We carry out the conversion from their notation to ours in Section L of the Supplementary Material [22]. This result was previously established under slightly less general conditions in [8, 57]. In particular, Proposition 2.6 states that,

> *Bayesian estimation in the linear model is exactly as good as Bayesian estimation in the sequence model at noise variance $\tau^2_{\mathsf{reg,stat}}$.*

Thus, a comparison of the convex lower bound, the risk of Bayes AMP, and the Bayes risk reduces to a comparison of the noise variances $\tau^2_{\mathsf{reg,cvx}}$, $\tau^2_{\mathsf{reg,amp*}}$, and $\tau^2_{\mathsf{reg,stat}}$. Because it is simply a lower bound, the convex lower bound could plausibly sometimes be smaller than the Bayes risk. Fortunately, this does not occur.

COROLLARY 2.7. *For all π, δ, σ, we have*

$$\tau^2_{\mathsf{reg,cvx}} \ge \tau^2_{\mathsf{reg,stat}}. \tag{2.22}$$

PROOF. The inequality $\tau^2_{\mathsf{reg,cvx}} \ge \tau^2_{\mathsf{reg,amp}}$ holds because the supremum in (2.13) is taken over a subset of the supremum in (2.7). Thus, it suffices to show $\tau^2_{\mathsf{reg,amp}} \ge \tau^2_{\mathsf{reg,stat}}$. For $\tau' < \tau_{\mathsf{reg,stat}}$,

$$\begin{aligned}\phi(\tau_{\mathsf{reg,stat}};\pi,\delta,\sigma) &< \phi(\tau';\pi,\delta,\sigma)\\ &= \phi(\tau_{\mathsf{reg,stat}};\pi,\delta,\sigma) + \frac{1}{2}\int_{\tau^{-2}_{\mathsf{reg,stat}}}^{\tau'^{-2}} \big(\sigma^2 - \delta\tau^2 + \mathsf{mmse}_\pi(\tau^2)\big)\, d\tau^{-2}.\end{aligned}$$

Thus, the integral in the previous display must be positive for all $\tau' < \tau_{\mathsf{reg,stat}}$ which implies there exists $\tau' < \tau_{\mathsf{reg,stat}}$ arbitrarily close to $\tau_{\mathsf{reg,stat}}$ for which $\delta\tau'^2 - \sigma^2 < \mathsf{mmse}_\pi(\tau'^2)$. By (2.13), we have $\tau_{\mathsf{reg,amp}} \ge \tau_{\mathsf{reg,stat}}$, as desired. □

3. Log-concavity and convex-algorithmic-statistical gaps.

The results in the preceding section establish that: *(i)* if $\tau^2_{\mathsf{reg,cvx}} > \tau^2_{\mathsf{reg,amp*}}$, there is a gap between the asymptotic estimation error achieved by convex M-estimators (1.2) and that achieved by Bayes AMP, and *(ii)* for generic (δ,σ) (i.e., those for which the minimizer in (2.18) is unique), if $\tau^2_{\mathsf{reg,cvx}} > \tau^2_{\mathsf{reg,stat}}$, there is a gap between the asymptotic estimation error achieved by convex M-estimators (1.2) and that achieved by information theoretically optimal estimation. Two important questions remain:

1. Is the converse true? Namely, if $\tau^2_{\mathsf{reg,cvx}} = \tau^2_{\mathsf{reg,amp*}}$ or $\tau^2_{\mathsf{reg,cvx}} = \tau^2_{\mathsf{reg,stat}}$, is convex M-estimation as good as Bayes AMP or Bayesian estimation?

2. Can we provide more interpretable conditions which determine whether the strict inequalities $\tau^2_{\mathsf{reg,cvx}} > \tau^2_{\mathsf{reg,amp*}}$ and $\tau^2_{\mathsf{reg,cvx}} > \tau^2_{\mathsf{reg,stat}}$ occur?

It turns out that the condition we provide to answer the second question will provide an affirmative answer to the first question. In particular, we will show that $\tau^2_{\mathsf{reg,cvx}} = \tau^2_{\mathsf{reg,amp*}}$ (resp., $\tau^2_{\mathsf{reg,cvx}} = \tau^2_{\mathsf{reg,stat}}$) if and only if $\pi * \mathsf{N}(0, \tau^2_{\mathsf{reg,amp*}})$ (resp., $\pi * \mathsf{N}(0, \tau^2_{\mathsf{reg,stat}})$) is log-concave. Moreover, while when $\delta \le 1$ we do not guarantee the tightness of the convex lower bound generally, we will guarantee its tightness in the case that $\pi * \mathsf{N}(0, \tau^2_{\mathsf{reg,cvx}})$ is log-concave. Because $\tau^2_{\mathsf{reg,cvx}} = \tau^2_{\mathsf{reg,amp*}}$ implies $\pi * \mathsf{N}(0, \tau^2_{\mathsf{reg,amp*}})$ and hence $\pi * \mathsf{N}(0, \tau^2_{\mathsf{reg,cvx}})$, is log-concave, it also implies that convex M-estimation is as good as Bayes AMP. A similar line of reasoning follows when $\tau^2_{\mathsf{reg,cvx}} = \tau^2_{\mathsf{reg,stat}}$. Thus, the converse described in the first question indeed holds.

Before describing this argument in detail, we remark that when π itself is log-concave, $\pi * \mathsf{N}(0, \tau^2)$ is log-concave for all τ^2. In this case the convex lower bound, the risk of Bayes AMP, and the Bayes risk agree for all values of σ, δ. Moreover, in this case the convex lower bound is always tight so that convex M-estimators (1.2) always achieve information theoretically optimal performance. In contrast, we will show that when π is not log-concave, there exist values of σ, δ for which the convex lower bound is strictly larger than the the risk of Bayes AMP and the Bayes risk. Thus, nontrivial performance of convex M-estimation relative to computational and information-theoretic benchmarks occurs exactly when π is not log-concave.

PROPOSITION 3.1. *Consider $\pi \in \mathcal{P}_\infty(\mathbb{R})$, $\delta \in (0, \infty)$, and $\sigma \ge 0$. If $\mathcal{C}$ consists of all sequences of convex penalties, the following statements hold under the HDA and RSN assumptions; if $\mathcal{C}$ consists of all sequences of symmetric convex penalties, we may replace the RSN by the DSN assumption:*

(i) *If $\tau \ge 0$ is such that $\pi * \mathsf{N}(0, \tau^2)$ has log-concave density (w.r.t. Lebesgue measure) and $\delta\tau^2 - \sigma^2 > \mathsf{mmse}_\pi(\tau^2)$, then*

$$\inf_{\{\rho_p\} \in \mathcal{C}_{\delta,\pi}} \operatorname*{p-lim}_{p \to \infty} \|\widehat{\boldsymbol{\beta}}_{\mathsf{cvx}} - \boldsymbol{\beta}_0\|^2 \le \delta\tau^2 - \sigma^2. \tag{3.1}$$

We may replace the limit in probability with $\lim_{p\to\infty} \mathbb{E}_{\boldsymbol{\beta}_0, \boldsymbol{w}, \boldsymbol{X}}[\|\widehat{\boldsymbol{\beta}}_{\mathsf{cvx}} - \boldsymbol{\beta}_0\|^2]$ under the RSN assumption. (We set these limits to ∞ when they do not exist.)

(ii) *If $\tau \ge 0$ is such that $\pi * \mathsf{N}(0, \tau^2)$ does not have log-concave density (w.r.t. Lebesgue measure) and $\delta\tau^2 - \sigma^2 \le \mathsf{mmse}_\pi(\tau^2)$, then $\tau^2_{\mathsf{reg,cvx}} > \tau^2$ whence*

$$\inf_{\{\rho_p\} \in \mathcal{C}_{\delta,\pi}} \operatorname*{p-liminf}_{p \to \infty} \|\widehat{\boldsymbol{\beta}}_{\mathsf{cvx}} - \boldsymbol{\beta}_0\|^2 > \delta\tau^2 - \sigma^2.$$

(iii) *We have $\tau^2_{\mathsf{reg,cvx}} = \tau^2_{\mathsf{reg,stat}}$ if and only if $\pi * \mathsf{N}(0, \tau^2_{\mathsf{reg,stat}})$ is log-concave. In the (generic) case that $\tau^2_{\mathsf{reg,amp}} = \tau^2_{\mathsf{reg,amp*}}$, we have $\tau^2_{\mathsf{reg,cvx}} = \tau^2_{\mathsf{reg,amp*}}$ if and only if $\pi * \mathsf{N}(0, \tau^2_{\mathsf{reg,amp*}})$.*

The proof of Proposition 3.1 is provided in Section J of the Supplementary Material [22].

While the relevance of the log-concavity of the convolutional density $\pi * \mathsf{N}(0, \tau^2)$ may seem surprising, it is related to the following fact: in the Gaussian sequence model (2.1), the Bayes estimator is the proximal operator of some convex function if and only if $\pi * \mathsf{N}(0, \tau^2)$ is log-concave. This is a remarkable consequence of Tweedie's formula. Our construction of penalties achieving (3.1) involves identifying the penalty whose proximal operator is the Bayes estimator at noise variance τ^2 in the sequence model. This is related to the construction of [12]; see Section J of the Supplementary Material [22] for details of this fact and its use in proving Proposition 3.1.

3.1. *Gaps between convex M-estimators and Bayes AMP.* Under generic conditions, convex M-estimators achieve the risk of Bayes AMP if and only if $\pi * \mathsf{N}(0, \tau^2_{\mathsf{reg,amp}*})$ has log-concave density.

THEOREM 2. *Consider* $\pi \in \mathcal{P}_2(\mathbb{R})$, $\delta \in (0, \infty)$, $\sigma \geq 0$. *Assume* $\tau_{\mathsf{reg,amp}} = \tau_{\mathsf{reg,amp}*}$ (*which holds generically, see the proof of Corollary* 2.3 *as well as Section L of the Supplementary Material* [22]). *If* $\mathcal{C}$ *contains all sequences of convex penalties, then, under the HDA and RSN assumptions, inequality* (2.12) *holds with equality* if and only if $\pi * \mathsf{N}(0, \tau^2_{\mathsf{reg,amp}*})$ *has log-concave density* (*w.r.t. Lebesgue measure*) *which occurs if and only if* $\tau^2_{\mathsf{reg,cvx}} = \tau^2_{\mathsf{reg,amp}*}$. *The same holds if we replace the limits in probability with the limits of expectations in* (2.12).

If $\mathcal{C}$ *contains all sequences of symmetric convex penalties, the preceding statements hold also under the DSN assumption.*

When equality occurs in Theorem 3, the penalty achieving the convex lower bound is (up to a small strong convexity term added for technical reasons) given by the convex function whose proximal operator is the Bayes estimator in the sequence model (2.1) at noise variance $\tau^2_{\mathsf{reg,amp}*}$. The existence of such a penalty is a consequence of the log-concavity of $\pi * \mathsf{N}(0, \tau^2_{\mathsf{reg,amp}})$; see the remark following Proposition 3.1 and the proof of that proposition in Section J of the Supplementary Material [22] for further details.

PROOF OF THEOREM 2. The equivalence of $\pi * \mathsf{N}(0, \tau^2_{\mathsf{reg,amp}*})$, having log-concave density and $\tau^2_{\mathsf{reg,cvx}} = \tau^2_{\mathsf{reg,amp}*}$, holds by Proposition 3.1.(*iii*). We now focus on the remaining parts of the theorem.

We first prove the "if" direction. By (2.9) we have, for $\tau > \tau_{\mathsf{reg,amp}*}$, that $\delta\tau^2 - \sigma^2 > \mathsf{mmse}_\pi(\tau^2)$. Further, because $\pi * \mathsf{N}(0, \tau^2_{\mathsf{reg,amp}*})$ has log-concave density, so too does $\pi * \mathsf{N}(0, \tau^2)$ ([49], Proposition 3.5). By Proposition 3.1.(*i*) we have that (3.1) holds with this choice of τ. Taking $\tau \downarrow \tau_{\mathsf{reg,amp}*} = \tau_{\mathsf{reg,amp}}$, we conclude that (2.12) holds with the inequality reversed, so, in fact, holds with equality.

We now prove the "only if" direction. By (2.9) and the continuity of $\mathsf{mmse}_\pi(\tau^2)$ in τ^2 ([34], Proposition 7), we have

$$\delta\tau^2_{\mathsf{reg,amp}*} - \sigma^2 = \mathsf{mmse}_\pi(\tau^2_{\mathsf{reg,amp}*}).$$

If $\pi * \mathsf{N}(0, \tau^2_{\mathsf{reg,amp}*})$ does not have log-concave density, by Proposition 3.1.(*ii*) equation (2.12) holds with strict inequality. By Lemma K.1 of the Supplementary Material [22], the same holds when replace limits in probability with limits of expectations. □

A corollary of Theorem 1 is that, when π has log-concave density, gaps between convex M-estimation and the risk of Bayes AMP do not occur, whereas when π does not have log-concave density, they do occur at large enough signal-to-noise ratios.

COROLLARY 3.2. *Consider* $\pi \in \mathcal{P}_2(\mathbb{R})$ *and* $\sigma \geq 0$. *Let* $\mathcal{B} \subseteq \mathbb{R}$ *be the set of* $\delta > 0$ *for which* $\tau_{\mathsf{reg,amp}} < \tau_{\mathsf{reg,amp}*}$ *holds* (*recall that, by the proof of Corollary* 2.3, $\mathcal{B}$ *has zero Lebesgue measure*). *We have the following*:

(a) *If π has log-concave density, then for all $\delta \in \mathbb{R}_{>0} \setminus \mathcal{B}$, inequality* (2.12) *holds with equality.*

(b) *If $\sigma > 0$ and π do not have log-concave density, then there exist $0 \leq \delta_{\mathsf{alg}} < \infty$ such that inequality* (2.12) *holds with equality for $\delta \in (0, \delta_{\mathsf{alg}}) \setminus \mathcal{B}$ and with strict inequality for all $\delta \in (\delta_{\mathsf{alg}}, \infty) \setminus \mathcal{B}$.*

Part (*b*) states that, if π is not log-concave, then either: (*i*) there is always a gap between convex M-estimation and the best algorithm we know of, or (*ii*) for small δ, the algorithmic lower bound is achieved by a convex procedure, while for large δ there is a gap between convex M-estimation and the best algorithm we know of. This might seem counterintuitive, because large δ corresponds to larger sample size and, therefore, easier estimation. An intuitive explanation of this result is that, for large δ, we can exploit more of the structure of the prior π, and this requires nonconvex methods.

PROOF OF COROLLARY 3.2. *Part (a)*: By [49], Proposition 3.5, $\pi * \mathsf{N}(0, \tau^2_{\mathsf{reg,amp}})$ has log-concave density. The result follows by Theorem 2.

Part (b): Define $\delta_{\mathsf{alg}} = \inf\{\delta \mid \pi * \mathsf{N}(0, \tau^2_{\mathsf{reg,amp}})$ does not have log-concave density$\}$. By [49], Proposition 3.5, if $\tau < \tau'$ and $\pi * \mathsf{N}(0, \tau^2)$ has log-concave density, then so too does $\pi * \mathsf{N}(0, \tau'^2)$. By (2.13), $\tau_{\mathsf{reg,amp}}$ is nonincreasing in δ. Combining these two facts, for $\delta > \delta_{\mathsf{alg}}$ we have $\mathsf{N}(0, \tau^2_{\mathsf{reg,amp}})$, which does not have log-concave density, and for $\delta < \delta_{\mathsf{alg}}$ we have $\mathsf{N}(0, \tau^2_{\mathsf{reg,amp}})$ which does have log-concave density. Then, by Theorem 2, inequality (2.12) holds with equality for $\mathcal{B} \ni \delta < \delta_{\mathsf{alg}}$ and with strict inequality when $\mathcal{B} \ni \delta > \delta_{\mathsf{alg}}$. We need only check that $\delta_{\mathsf{alg}} < \infty$. By (2.10), $\tau^2_{\mathsf{reg,amp}} = \frac{1}{\delta}(\sigma^2 + \mathsf{mmse}_\pi(\tau^2_{\mathsf{reg,amp}})) \leq \frac{1}{\delta}(\sigma^2 + s_2(\pi))$. Thus, $\lim_{\delta\to\infty} \tau^2_{\mathsf{reg,amp}} = 0$. Because log-concavity is preserved under convergence in distribution ([49], Proposition 3.6) and $\pi * \mathsf{N}(0, \tau^2) \xrightarrow[\tau\to 0]{\mathrm{d}} \pi$, we conclude that, for δ sufficiently large, $\pi * \mathsf{N}(0, \tau^2_{\mathsf{reg,amp}})$ does not have log-concave density, as desired. □

3.2. *Gaps between convex M-estimators and the Bayes risk.* Under generic conditions, convex M-estimators achieve the Bayes risk exactly when the convex lower bound is equal to the Bayes risk which, in turn, occurs exactly when $\pi * \mathsf{N}(0, \tau^2_{\mathsf{reg,stat}})$ has log-concave density.

THEOREM 3. *Consider $\pi \in \mathcal{P}_\infty(\mathbb{R})$, $\delta \in (0, \infty)$, and $\sigma > 0$. Assume the potential ϕ, defined in equation* (2.17). *has a unique minimizer. If $\mathcal{C}$ cosists of all sequences of convex penalties, then, under the HDA and RSN assumptions, $\tau^2_{\mathsf{reg,cvx}} = \tau^2_{\mathsf{reg,stat}}$ if and only if*

$$(3.2) \qquad \inf_{\{\rho_p\}_p \in \mathcal{C}_{\delta,\pi}} \liminf_{p\to\infty} \mathbb{E}_{\boldsymbol{\beta}_0,\boldsymbol{w},\boldsymbol{X}}\big[\|\widehat{\boldsymbol{\beta}}_{\mathsf{cvx}} - \boldsymbol{\beta}_0\|^2\big] = \lim_{p\to\infty} \mathbb{E}_{\boldsymbol{\beta}_0,\boldsymbol{w},\boldsymbol{X}}\big[\big\|\mathbb{E}_{\boldsymbol{\beta}_0,\boldsymbol{w},\boldsymbol{X}}[\boldsymbol{\beta}_0 \mid \boldsymbol{y}] - \boldsymbol{\beta}_0\big\|^2\big]$$

which, in turn, occurs if and only if $\pi * \mathsf{N}(0, \tau^2_{\mathsf{reg,stat}})$ has log-concave density with respect to Lebesgue measure on $\mathbb{R}$.

Analogously to Theorem 2, when equality occurs in Theorem 3, the penalty achieving the convex lower bound is (up to a small strong convexity term added for technical reasons) given by the convex function whose proximal operator is the Bayes estimator in the sequence model (2.1) at noise variance $\tau^2_{\mathsf{reg,stat}}$; see the remark following Proposition 3.1 and the proof of that proposition in Section J of the Supplementary Material [22] for further details. The condition that the minimizer of ϕ is unique holds—by analyticity considerations—for all (δ, σ), except a set of Lebesgue measure zero.

PROOF OF THEOREM 3. The equivalence of $\pi * \mathsf{N}(0, \tau^2_{\mathsf{reg,stat}})$, having log-concave density and $\tau^2_{\mathsf{reg,cvx}} = \tau^2_{\mathsf{reg,stat}}$, holds by Proposition 3.1(iii). We now focus on the remaining parts of the Theorem.

The right-hand side of (3.2) is $\delta\tau^2_{\mathsf{reg,stat}} - \sigma^2$ by Proposition 2.6 (this is where we use $\sigma > 0$). By (2.22), if $\tau^2_{\mathsf{reg,cvx}} \neq \tau^2_{\mathsf{reg,stat}}$, then $\tau^2_{\mathsf{reg,cvx}} > \tau^2_{\mathsf{reg,stat}}$. Then, by Theorem 1 as well as Lemma K.1 of the Supplementary Material [22], we have under the RSN assumption that (3.2) holds with equality replace by strict inequality.

Now, consider that $\tau^2_{\mathsf{reg,cvx}} = \tau^2_{\mathsf{reg,stat}}$ or, equivalently, that $\pi * \mathsf{N}(0, \tau^2_{\mathsf{reg,stat}})$ has log-concave density. Assume $\mathsf{N}(0, \tau^2_{\mathsf{reg,stat}})$ has log-concave density, $\sigma > 0$, and ϕ has unique minimizer. For $\tau' > \tau_{\mathsf{reg,stat}}$ we have

$$\begin{aligned}\phi(\tau_{\mathsf{reg,stat}}; \pi, \delta, \sigma) &= \phi(\tau'; \pi, \delta, \sigma) + \frac{1}{2}\int_{\tau'^{-2}}^{\tau^{-2}_{\mathsf{reg,stat}}} \left(\sigma^2 - \delta\tau^2 + \mathsf{mmse}_\pi(\tau^2)\right) d\tau^{-2} \\ &> \phi(\tau_{\mathsf{reg,stat}}; \pi, \delta, \sigma) + \frac{1}{2}\int_{\tau'^{-2}}^{\tau^{-2}_{\mathsf{reg,stat}}} \left(\sigma^2 - \delta\tau^2 + \mathsf{mmse}_\pi(\tau^2)\right) d\tau^{-2},\end{aligned}$$

where in the inequality we use that the minimizer of ϕ is unique. Thus, the integral is negative for all $\tau' > \tau_{\mathsf{reg,stat}}$, so there exists $\tau' > \tau_{\mathsf{reg,stat}}$ arbitrarily close to $\tau_{\mathsf{reg,stat}}$ for which $\delta\tau'^2 - \sigma^2 > \mathsf{mmse}_\pi(\tau'^2)$. By [49], Proposition 3.5, we have, for all such τ', that $\pi * \mathsf{N}(0, \tau'^2)$ has log-concave density. Taking $\tau' \downarrow \tau_{\mathsf{reg,stat}}$ along τ' for which $\delta\tau'^2 - \sigma^2 > \mathsf{mmse}_\pi(\tau'^2)$ and applying Proposition 3.1.(i), we have under the RSN assumption that

$$\inf_{\{\rho_p\}_p \in \mathcal{C}_{\delta,\pi}} \lim_{p\to\infty} \mathbb{E}_{\boldsymbol{\beta}_0,\boldsymbol{w},\boldsymbol{X}}\left[\|\widehat{\boldsymbol{\beta}}_{\mathsf{cvx}} - \boldsymbol{\beta}_0\|^2\right] \leq \delta\tau^2_{\mathsf{reg,stat}} - \sigma^2.$$

By (2.21) we have $\delta\tau^2_{\mathsf{reg,stat}} - \sigma^2$ equals the right-hand side of (3.2). The reverse inequality holds by the optimality of the Bayes risk, whence we conclude (3.2). □

A corollary of Theorem 1 is that, when π has log-concave density, gaps between convex M-estimation and the Bayes risk do not occur, whereas when π does have log-concave density, they do occur at large enough signal-to-noise ratios.

COROLLARY 3.3. *Consider $\pi \in \mathcal{P}_\infty(\mathbb{R})$ and $\sigma > 0$. We have the following*:

(a) *If π has log-concave density with respect to Lebesgue measure*, *then*, *for all $\delta > 0$ for which ϕ has unique minimizer*, *equality* (3.2) *holds.*

(b) *If π does not have log-concave density with respect to Lebesgue measure*, *then there exist $0 \leq \delta_{\mathsf{stat}} < \infty$ such that equality* (3.2) *holds for all $\delta < \delta_{\mathsf{stat}}$ for which ϕ has unique minimizer*, *and* (3.2) *holds with strict inequality replacing equality for all $\delta > \delta_{\mathsf{stat}}$ for which ϕ has unique minimizer. Moreover*, $\delta_{\mathsf{stat}} \leq \delta_{\mathsf{alg}}$.

PROOF OF COROLLARY 3.3. *Part (a)*: By [49], Proposition 3.5, we have $\pi * \mathsf{N}(0, \tau^2_{\mathsf{reg,stat}})$ has log-concave density with respect to Lebesgue measure. The result follws by Theorem 3.

Part (b): Define $\delta_{\mathsf{stat}} = \inf\{\delta \mid \pi * \mathsf{N}(0, \tau^2_{\mathsf{reg,stat}})$ does not have log-concave density$\}$. Because the derivative (2.19) of ϕ with respect to τ^{-2} is strictly decreasing in δ, we have by (2.17) that $\tau_{\mathsf{reg,stat}}$ is strictly decreasing in δ. As in the proof of Corollary 3.2, this implies that for $\delta > \delta_{\mathsf{stat}}$ we have $\mathsf{N}(0, \tau^2_{\mathsf{reg,stat}})$ which does not have log-concave density and for $\delta < \delta_{\mathsf{stat}}$ we have $\mathsf{N}(0, \tau^2_{\mathsf{reg,stat}})$ which does have log-concave density. Then, by Theorem 3, if ϕ has unique minimizer and $\delta > \delta_{\mathsf{stat}}$, then the left-hand side of (3.2) is strictly larger than the right-hand side, and if ϕ has unique minimizer and $\delta < \delta_{\mathsf{stat}}$, equality holds. We need only check that $\delta_{\mathsf{stat}} < \infty$. By (2.18) and (2.19), we have $\tau^2_{\mathsf{reg,stat}} = \frac{1}{\delta}(\sigma^2 + \mathsf{mmse}_\pi(\tau^2_{\mathsf{reg,stat}})) \le \frac{1}{\delta}(\sigma^2 + s_2(\pi))$, where $s_2(\pi)$ is the second moment of π. Thus, $\lim_{\delta\to\infty}\tau^2_{\mathsf{reg,stat}} = 0$. Because log-concavity is preserved under convergence in distribution ([49], Proposition 3.6) and $\pi * \mathsf{N}(0,\tau^2) \xrightarrow[\tau\to 0]{\mathrm{d}} \pi$, we conclude that, for sufficiently large δ, $\pi * \mathsf{N}(0,\tau^2_{\mathsf{reg,stat}})$ is not log-concave, as desired. □

4. Quantifying the gap: High and low signal-to-noise ratio (SNR) regimes.

We now provide quantitative estimates of the gap between convex M-estimation and the Bayes risk when such gaps occur. Consider $\pi \in \mathcal{P}_\infty(\mathbb{R})$, $\delta \in (0,\infty)$, $\sigma > 0$, and let $\mathcal{C}$ contain all sequences of convex penalties. Define the asymptotic gap between convex M-estimation and Bayes error

$$\begin{aligned}\Delta(\pi,\delta,\sigma) \\ \equiv \Big(\inf_{\{\rho_p\}_p\in\mathcal{C}_{\delta,\pi}} \liminf_{p\to\infty} \mathbb{E}_{\boldsymbol{\beta}_0,\boldsymbol{w},\boldsymbol{X}}\big[\|\widehat{\boldsymbol{\beta}}_{\mathsf{cvx}} - \boldsymbol{\beta}_0\|^2\big]\Big) \\ - \Big(\lim_{p\to\infty} \mathbb{E}_{\boldsymbol{\beta}_0,\boldsymbol{w},\boldsymbol{X}}\big[\big\|\mathbb{E}_{\boldsymbol{\beta}_0,\boldsymbol{w},\boldsymbol{X}}[\boldsymbol{\beta}_0 \mid \boldsymbol{y},\boldsymbol{X}] - \boldsymbol{\beta}_0\big\|^2\big]\Big),\end{aligned}$$

where the limits are taken under the HDA and RSN assumptions. The results of Section 3.2 characterize whether $\Delta(\pi,\delta,\sigma) = 0$ or $\Delta(\pi,\delta,\sigma) > 0$. Here, we provide a more quantitative estimate of its size for large δ (high SNR) and for large σ (low SNR).

THEOREM 4. *Fix $\pi \in \mathcal{P}_\infty(\mathbb{R})$, and let $\mathcal{C}$ contain all sequences of convex penalties*:

(i) *Restricting ourselves to $\delta, \sigma > 0$ for which the minimizer of* (2.18) *is unique*, *we have*

$$\Delta(\pi,\delta,\sigma) \ge \mathsf{R}^{\mathsf{opt}}_{\mathsf{seq,cvx}}(\sigma/\sqrt{\delta};\pi) - \mathsf{mmse}_\pi(\sigma^2/\delta) + O(1/\sqrt{\delta}), \tag{4.1}$$

where O hides constants depending only on the moments of π.

(ii) *Let* $\mathsf{snr} = \frac{s_2(\pi)}{\sigma^2}$ *denote the signal-to-noise ratio for the sequence model. For any fixed δ, we have $\Delta(\pi,\delta,\sigma) = O(\mathsf{snr}^2)$ as $\mathsf{snr} \to 0$. More precisely,*

$$\limsup_{\mathsf{snr}\to 0} \frac{\Delta(\pi,\delta,\sigma)}{\mathsf{snr}^2} \le s_2(\pi)\delta^2 \frac{s_3^2(\pi)}{2s_2^3(\pi)},$$

where the lim sup *is taken over σ at which* (2.17) *has unique minimizer.*

The proof of this theorem is given in Section M of the Supplementary Material [22]. We believe its results provide some useful insight:

- Because it ensures high-dimensional consistency, the large δ regime of point (i) is most commonly analyzed in the statistics literature. In this regime, Theorem 4 establishes that the gap between convex M-estimation and Bayes error is essentially determined by the analogous gap in the sequence model for noise level $\sigma/\sqrt{\delta}$. As will be discussed in the next section, in this regime it makes sense to refine the M-estimate by post-processing.
- In the low SNR regime (large σ), the structure of the signal $\boldsymbol{\beta}_0$ (and, in particular, the distribution of the coefficients β_{0j}) is blurred by the Gaussian noise, and the gap vanishes. This should be compared with the results of Corollary 3.3, which state that gaps, when they occur, occur for small values of δ, which also corresponds to a low SNR regime. Both of these results can be traced to the fact that the measure $\pi * \mathsf{N}(0,\tau^2_{\mathsf{reg,stat}})$ will in some sense be "more log-concave" when $\tau^2_{\mathsf{reg,stat}}$ is larger. Because $\tau^2_{\mathsf{reg,stat}}$ quantifies, in a certain sense, the intrinsic noisiness of the problem, we see that convex M-estimation comes closer to achieving (or exactly achieves) information theoretic limits at low SNR.

5. Beyond mean square error. A natural concern with the optimality theory we have presented is that it only addresses ℓ_2 loss. With a certain type of efficient post-processing, the optimality theory for general continuous losses is essentially unchanged. In particular, if we consider two-step procedures in which we first compute a penalized least squares estimator $\widehat{\boldsymbol{\beta}}_{\mathsf{cvx}}$ and second implement simple post-processing detailed below, the optimal choice of penalty in the first step should not depend on the loss ℓ. The main reason for this is captured by the following result. (This proposition relies on the notion of strong stationarity introduced in Section B which formalizes the notion of solving the fixed point equations (2.3) and includes a few more technical conditions. It also uses the collection of penalty sequences $\mathcal{C}_*$, which are *uniformly strongly convex*, defined below in Definition 6.1. This is a subset of the collection of convex penalty sequences.)

PROPOSITION 5.1. *Consider $\pi \in \mathcal{P}_\infty(\mathbb{R})$, $\delta \in (0, \infty)$, and $\sigma \geq 0$. Let $\{\rho_p\}$, $\{\tilde{\rho}_p\}$ be sequences of lsc, proper, convex penalties. Let $\mathcal{T} = (\pi, \{\rho_p\})$ and $\tilde{\mathcal{T}} = (\pi, \{\tilde{\rho}_p\})$, and assume τ, λ, $\tilde{\tau}$, $\tilde{\lambda}$ are such that τ, λ, δ, $\mathcal{T}$ and $\tilde{\tau}$, $\tilde{\lambda}$, δ, $\tilde{\mathcal{T}}$ are strongly stationary. Without loss of generality, consider $\tilde{\tau} \leq \tau$. Assume either $\delta > 1$ or $\{\rho_p\}, \{\tilde{\rho}_p\} \in \mathcal{C}_*$ (see Definition 6.1 below). Let $\widehat{\boldsymbol{\beta}}_{\mathsf{cvx}}$ and $\widehat{\tilde{\boldsymbol{\beta}}}_{\mathsf{cvx}}$ be defined by (1.2) with penalties ρ_p and $\tilde{\rho}_p$, respectively. For such sufficiently large p, let*

$$\widehat{\boldsymbol{\beta}}_{\mathsf{cvx}+} = \mathsf{prox}[\lambda\rho_p]\biggl(\widehat{\tilde{\boldsymbol{\beta}}}_{\mathsf{cvx}} + \frac{2\lambda}{n}\boldsymbol{X}^{\mathsf{T}}(\boldsymbol{y} - \boldsymbol{X}\widehat{\tilde{\boldsymbol{\beta}}}_{\mathsf{cvx}}) + \sqrt{\tau^2 - \tilde{\tau}^2}\boldsymbol{z}\biggr),$$

where, for each p, $\boldsymbol{z} \sim \mathsf{N}(\boldsymbol{0}, \boldsymbol{I}_p/p)$ is independent of $\boldsymbol{X}$.

Under the HDA and RSN assumptions, for any sequence of symmetric, uniformly pseudo-Lipschitz sequence of losses $\ell_p : (\mathbb{R}^p)^2 \to \mathbb{R}$ of order k for some k, we have

$$\ell_p(\boldsymbol{\beta}_0, \widehat{\boldsymbol{\beta}}_{\mathsf{cvx}+}) \overset{\mathrm{p}}{\simeq} \ell_p(\boldsymbol{\beta}_0, \widehat{\boldsymbol{\beta}}_{\mathsf{cvx}}).$$

If the penalties ρ_p, $\tilde{\rho}_p$ are symmetric, then the preceding display holds also under the DSN assumption.

We prove Proposition 5.1 in Section H of the Supplementary Material [22]. Proposition 5.1 establishes that, when $\tilde{\tau} \leq \tau$, we can always post-process $\widehat{\tilde{\boldsymbol{\beta}}}_{\mathsf{cvx}}$ to construct an estimator $\widehat{\boldsymbol{\beta}}_{\mathsf{cvx}+}$ whose performance matches that of $\widehat{\boldsymbol{\beta}}_{\mathsf{cvx}}$ with respect to loss ℓ. Proposition 5.1 suggests that, for any loss, the optimal choice of penalty in the M-estimation step in this two-step procedure is that which minimizes the effective noise parameter τ. It turns out this is equivalent to choosing a penalty which minimizes ℓ_2 loss.

A formalization of this discussion is provided in the next theorem.

THEOREM 5. *Assume $\eta : \mathbb{R} \to \mathbb{R}$ is the Bayes estimator of β_0 in the scalar model $y = \beta_0 + \tau_{\mathsf{reg,cvx}} z$ with respect to loss ℓ. If $\mathcal{C}$ contains all sequences of convex penalties, then, under the HDA and RSN assumption,*

$$\inf_{\{\rho_p\} \in \mathcal{C}_*} \ \operatorname*{p-lim\,inf}_{p \to \infty} \frac{1}{p}\sum_{j=1}^{p} \ell(\sqrt{p}\beta_{0j}, \sqrt{p}\widehat{\beta}_{\mathsf{cvx},j}) \geq \mathbb{E}_{\beta_0,z}\bigl[\ell(\beta_0, \eta(\beta_0 + \tau_{\mathsf{reg,cvx}} z))\bigr]. \tag{5.1}$$

When η is not the proximal operator of a convex function, inequality (5.1) is strict.

Further, when $\delta > 1$,

$$\begin{aligned} \inf_{\substack{\{\rho_p\} \in \mathcal{C}_* \\ \eta' \textit{ Lipschitz}}} \ \operatorname*{p-lim}_{p \to \infty} \frac{1}{p}\sum_{j=1}^{p} \ell\biggl(\sqrt{p}\beta_{0j}, \eta'\biggl(\sqrt{p}\widehat{\boldsymbol{\beta}}_{\mathsf{cvx},j} + 2\lambda\frac{[\boldsymbol{X}^{\mathsf{T}}(\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}_{\mathsf{cvx}})]_j}{n}\biggr)\biggr) \\ = \mathbb{E}_{\beta_0,z}\bigl[\ell(\beta_0, \eta(\beta_0 + \tau_{\mathsf{reg,cvx}} z))\bigr]. \end{aligned} \tag{5.2}$$

The sequences $\{\rho_p\}$, which minimize the ℓ_2 loss of $\widehat{\boldsymbol{\beta}}_{\mathsf{cvx}}$, also achieve the infimum in (5.2). (*Note that the infimum over η' is taken* after *the limit $p \to \infty$ and*, *in particular*, *η' does not depend on p*.)

If $\mathcal{C}$ contains all sequences of symmetric convex penalties, *the preceding statements hold also under the DSN assumption.*

We prove Theorem 5 in Section H of the Supplementary Material [22]. We expect inequality (5.1) to hold also when the infimum is taken over $\mathcal{C}_{\delta,\pi}$, but we are not aware how to control the estimation error with respect to arbitrary pseudo-Lipschitz losses for $\{\rho_p\} \in \mathcal{C}_{\delta,\pi}$. We expect equality (5.2) to hold also when $\delta \le 1$, but this requires establishing the tightness of the convex lower bound when $\delta \le 1$, which we are unable to do (see discussion following Theorem 1). We believe these extensions may be possible using currently available tools but leave it for future work.

For large δ, post-processing nearly closes the gap between convex M-estimation and Bayes AMP. Indeed, as is shown in Section M of the Supplementary Material [22], when δ is large (high SNR)—so that (4.1) provides a good approximation of the gap $\Delta(\pi, \delta, \sigma)$—we have $\tau_{\mathsf{reg,cvx}} \approx \tau_{\mathsf{reg,amp}*} \approx \sigma/\sqrt{\delta}$. Thus, the gap between the convex lower bound and the Bayes risk in this case is driven not by the difference between $\tau_{\mathsf{reg,cvx}}$ and $\tau_{\mathsf{reg,amp}*}$ but rather by the difference between estimation at that noise level using the optimal proximal operator (as done in (2.4)) and the Bayes estimator (as done in (2.8)). Theorem 5 states that, by post-processing, we may effectively replace the proximal operator in equation (H.1) of the Supplementary Material [22] by a nonproximal denoiser, which we may take to be the Bayes estimator (or a Lipschitz approximation of it) with respect to ℓ_2 loss. This is an important insight because we suspect that the behavior of M-estimation with one step of post-processing is more robust to model misspecification than is the behavior of Bayes AMP, whose finite sample convergence has been observed to be highly sensitive to distributional assumptions on the design matrix $\boldsymbol{X}$ (see, e.g., [45, 46]).

6. Examples.

Recall that, for $\delta > 1$, the assumption that ρ has δ-bounded width does not pose any restriction. For $\delta \le 1$, our proof requires $\rho \in \mathcal{C}_{\delta,\pi}$ for technical reasons which are discussed in Section I of the Supplementary Material [22]. We believe the conclusion of Theorem 1 should hold more generally. Nevertheless, as illustrated in the present section, the assumption $\rho \in \mathcal{C}_{\delta,\pi}$ is quite weak and is satisfied by broad classes of penalties.

Most proofs are omitted from this section and can be found in Section N of the Supplementary Material [22]. Through this section we take $\mathcal{C}$ to contain all sequences of convex penalties so that $\mathcal{C}_{\delta,\pi}$ contains all sequences with δ-bounded width.

6.1. *Strongly convex penalties.*

We introduce the notion of uniform strong convexity.

DEFINITION 6.1 (Uniform strong convexity). A sequence $\rho_p : \mathbb{R}^p \to \mathbb{R} \cup \{\infty\}$ of lsc, proper, convex functions *has uniform strong-convexity parameter* $\gamma \ge 0$ if $\boldsymbol{x} \mapsto \rho_p(\boldsymbol{x}) - \frac{\gamma}{2}\|\boldsymbol{x}\|^2$ is convex for all p. We say that $\{\rho_p\}$ is *uniformly strongly convex* if this holds for some $\gamma > 0$.

We define

$$\mathcal{C}_* = \{\{\rho_p\} \in \mathcal{C} \mid \{\rho_p\} \text{ is uniformly strongly convex}\}.$$

When the penalties are uniformly strongly convex, the situation is particularly nice.

PROPOSITION 6.2. *For all $\pi \in \mathcal{P}_2(\mathbb{R})$ and $\delta \in (0, \infty)$*, *we have $\mathcal{C}_* \subset \mathcal{C}_{\delta,\pi}$*.

6.2. *Convex constraints.* Consider

$$\rho_p(\boldsymbol{x}) = \mathbb{I}_{C_p}(\boldsymbol{x}) := \begin{cases} 0 & \boldsymbol{x} \in C_p, \\ \infty & \text{otherwise,} \end{cases}$$

where C_p is a closed convex set. Convex M-estimation, using this penalty, is equivalent to defining $\widehat{\boldsymbol{\beta}}_{\mathsf{cvx}}$ via the constrained optimization problem

$$\widehat{\boldsymbol{\beta}}_{\mathsf{cvx}} = \arg\min_{\boldsymbol{\beta}} \left\{ \frac{1}{n} \|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\|^2 : \boldsymbol{\beta} \in C_p \right\}. \tag{6.1}$$

In this context, the condition (2.6) is closely related to bounding the Gaussian width of convex cones [2, 23]. We briefly recall the relevant notions.

Given a closed convex set K, we denote by Π_K the orthogonal projector onto K. Namely, $\Pi_K(\boldsymbol{y}) := \arg\min_{\boldsymbol{x}\in K} \|\boldsymbol{y} - \boldsymbol{x}\|_2$. Recall that K is a convex cone if K is convex and, for every $\alpha > 0$, $K = \{\alpha \boldsymbol{x} \mid \boldsymbol{x} \in K\}$. For any set $A \subseteq \mathbb{R}^p$, we define the closed, conic hull of A centered at $\boldsymbol{b} \in \mathbb{R}^p$ by

$$T_A(\boldsymbol{b}) := \mathsf{cone}(\{\boldsymbol{x} - \boldsymbol{b} \mid \boldsymbol{x} \in A\}) := \overline{\mathsf{conv}(\{\alpha(\boldsymbol{x} - \boldsymbol{b}) \mid \boldsymbol{x} \in A, \alpha \geq 0\})},$$

where the overline denotes closure and conv denotes the convex hull. There are several equivalent definitions of the Gaussian width of a closed, convex cone K. The following translates most readily into our setup (recall that $\boldsymbol{z} \sim \mathsf{N}(0, \boldsymbol{I}_p/p)$:

$$w(K) := \mathbb{E}_{\boldsymbol{z}}\big[\|\Pi_K(\boldsymbol{z})\|^2\big].$$

The Gaussian width is closely related to the geometry of high-dimensional linear inverse problems. In particular, under the HDA and DSN assumptions, exact recovery $\widehat{\boldsymbol{\beta}}_{\mathsf{cvx}} = \boldsymbol{\beta}_0$ in the noiseless setting (i.e., $\boldsymbol{w} = \boldsymbol{0}$) is achieved with high probability by (6.1) if and only if $\limsup_{p\to\infty} w(T_{C_p}(\boldsymbol{\beta}_0)) < \delta$ [2, 23]. The same condition which guarantees *stable recovery* under noisy measurements, namely, that the error $\|\widehat{\boldsymbol{\beta}}_{\mathsf{cvx}} - \boldsymbol{\beta}_0\|$ is bounded, up to a constant, by the norm of the noise $\|\boldsymbol{w}\|$. Thus, when $w(T_{C_p}(\boldsymbol{\beta}_0)) > \delta$, we expect the estimation error of $\widehat{\boldsymbol{\beta}}_{\mathsf{cvx}}$ to be uncontrolled. It is, therefore, reasonable to focus on the case $w(T_{C_p}(\boldsymbol{\beta}_0)) < \delta$.

In the case of convex constraints, the δ-bounded width assumption reduces to a slightly weaker condition than $w(T_{C_p}(\boldsymbol{\beta}_0)) < \delta$. This is perhaps not surprising in light of the fact that, for $\rho_p = \mathbb{I}_{C_p}(\boldsymbol{x})$, the proximal operator $\mathsf{prox}[\lambda\rho_p](\boldsymbol{\beta}_0 + \tau\boldsymbol{z}) = \Pi_{C_p}(\boldsymbol{\beta}_0 + \tau\boldsymbol{z})$ and $\lim_{\tau\to 0} \frac{1}{\tau}\mathbb{E}_{\boldsymbol{z}}[\langle \boldsymbol{z}, \mathsf{prox}[\lambda\rho_p](\boldsymbol{\beta}_0 + \tau\boldsymbol{z}) = \Pi_{C_p}(\boldsymbol{\beta}_0 + \tau\boldsymbol{z})\rangle] = \mathbb{E}_{\boldsymbol{z}}[\|\Pi_{T_{C_p}(\boldsymbol{\beta}_0)}(\boldsymbol{z})\|^2]$. The following proposition makes the relationship between Gaussian widths and the δ-bounded width assumption precise.

PROPOSITION 6.3. *Consider C_p closed, symmetric, convex sets, $\pi \in \mathcal{P}_2(\mathbb{R})$, and $\delta \in (0, \infty)$. Assume that*

$$\lim_{p\to\infty} \mathbb{E}_{\boldsymbol{\beta}_0}\big[d(\boldsymbol{\beta}_0, C_p)\big] = 0.$$

Further assume that

$$\lim_{\varepsilon\to 0} \limsup_{p\to\infty} \mathbb{E}_{\boldsymbol{\beta}_0}\big[w\big(T_{C_p\cap B^c(\boldsymbol{\beta}_0,\varepsilon)}(\boldsymbol{\beta}_0)\big)\big] < \delta, \tag{6.2}$$

where $B^c(\boldsymbol{\beta}_0, \varepsilon)$ denotes the complement of the ball of radius ε centered at $\boldsymbol{\beta}_0$. Then, $\{\mathbb{I}_{C_p}\} \in \mathcal{C}_{\delta,\pi}$.

The quantity $\lim_{\varepsilon\to 0} w(T_{C_p\cap B^c(\boldsymbol{\beta}_0,\varepsilon)}(\boldsymbol{\beta}_0))$ agrees with $w(T_{C_p}(\boldsymbol{\beta}))$ when $\boldsymbol{\beta}_0 \in \partial C_p$. Thus, when $\boldsymbol{\beta}_0 \in \partial C_p$ almost surely, assumption (6.2) of Proposition 6.3 is exactly that $\limsup_{p\to\infty} w(T_{C_p}(\boldsymbol{\beta}_0)) < \delta$. This condition guarantees exact and stable recovery for the convex program (6.1). Thus, Proposition 6.3 implies that if constraint sets $\{C_p\}$ guarantee exact and stable recovery, then $\{\mathbb{I}_{C_p}\} \in \mathcal{C}_{\delta,\pi}$.

In the definition of the δ-bounded width assumption (or under the RSN assumption), $\boldsymbol{\beta}_0$ is random. Thus, it will, in general, be close to but not exactly on the boundary of C_p. For $\boldsymbol{\beta}_0$ in an ε-neighborhood of the boundary but not on the boundary, the quantity $w(T_{C_p\cap B^c(\boldsymbol{\beta}_0,\varepsilon)}(\boldsymbol{\beta}_0))$ describes the behavior of the convex program (6.1) and the quantity $w(T_{C_p}(\boldsymbol{\beta}))$ does not. Indeed, $w(T_{C_p}(\boldsymbol{\beta}))$ is highly sensitive to small perturbations of $\boldsymbol{\beta}_0$: it jumps to 1 when $\boldsymbol{\beta}_0$ is in the interior of C_p. In contrast, the behavior of the convex program (6.1) is not sensitive to such small perturbations. When $\boldsymbol{\beta}_0$ is asymptotically arbitrarily close to but not necessarily exactly on the boundary of C_p, the condition of Proposition 6.3 is the correct extension of the condition $\limsup_{p\to\infty} w(T_{C_p}(\boldsymbol{\beta}_0)) < \delta$. It guarantees recovery with asymptotically vanishing error $\|\widehat{\boldsymbol{\beta}}_{\mathsf{cvx}} - \boldsymbol{\beta}_0\|^2 \to 0$ when $d(\boldsymbol{\beta}_0, \partial C_p) \to 0$. For such $\boldsymbol{\beta}_0$, this is the natural replacement of the more stringent notion of exact recovery which will not occur if $\boldsymbol{\beta}_0 \notin \partial C_p$.

6.3. *Separable penalties.*

A common class of penalties considered in high-dimensional regression are the separable penalties

$$\rho_p(\boldsymbol{x}) = \frac{1}{p}\sum_{j=1}^{p} \rho(\sqrt{p}x_j), \tag{6.3}$$

for an lsc, proper, convex function $\rho : \mathbb{R} \to \mathbb{R} \cup \{\infty\}$ which does not depend on p. Much previous work has analyzed the asymptotic properties of M-estimators, which use separable penalties [12, 27, 30], and a few works have broken the separability assumption [57]. While Theorem 1 is more general, it applies to separable penalties under a mild condition.

PROPOSITION 6.4. *Consider ρ_p, as in (6.3), for some lsc, proper, convex $\rho : \mathbb{R} \to \mathbb{R} \cup \{\infty\}$. Let $C \subseteq \mathbb{R}$ be the set of minimizers of ρ (which is necessarily a closed interval). If C is nonempty, we have*

$$\sup_{\tau>\varepsilon} \mathbb{P}_{\beta_0,z}(\beta_0 + \tau z \in C) < \delta \quad \textit{for all } \varepsilon > 0,$$

if and only if $\{\rho_p\} \in \mathcal{C}_{\delta,\pi}$.

REMARK 6.1. Proposition 6.4 applies whenever C is a singleton set because in this case $\mathbb{P}(\beta_0 + \tau z \in C) = 0$ for all $\tau > 0$. Thus, Proposition 6.4 covers most, if not all, separable penalties commonly considered in practice (and many more).

6.4. *SLOPE and OWL norms.*

Here, we consider the ordered weighted ℓ_1 (OWL) norms, defined by

$$\rho_p(\boldsymbol{x}) = \frac{1}{\sqrt{p}}\sum_{j=1}^{p} \kappa_j^{(p)} |x|_{(j)}, \tag{6.4}$$

where $\kappa_1^{(p)} \geq \kappa_2^{(p)} \geq \cdots \geq \kappa_p^{(p)} \geq 0$ are the coordinates of $\boldsymbol{\kappa}^{(p)} \in \mathbb{R}^p$ and $|x|_{(j)}$ are the decreasing order statistics of the absolute values of the coordinates of $\boldsymbol{x}$. When $\kappa_j^{(p)} = \Phi^{-1}(1 - jq/(2p))$ for some $q \in (0, 1)$ and Φ^{-1} the standard normal cdf, the estimator (1.2) is referred to as SLOPE. Penalties of the form (6.4) have been used for a few purposes.

SLOPE has recently been proposed for sparse regression because it automatically adapts to sparsity level [13, 16, 53]. More generally, the use of OWL norms has been argued to produce estimators which are more stable than LASSO under correlated designs [18, 31].

PROPOSITION 6.5. *Consider ρ_p, as in* (6.4). *If for all $\varepsilon > 0$ there exists $\xi > 0$ such that $j \leq (1-\varepsilon)p$ implies $\kappa_j^{(p)} > \xi$, then $\{\rho_p\} \in \mathcal{C}_{\delta,\pi}$.*

Funding. The first author was supported in part by NSF Grants DGE – 1656518, CCF – 1714305, IIS – 1741162, and ONR N00014-18-1-2729.

SUPPLEMENTARY MATERIAL

Supplement A: Supplement to 'Fundamental barriers to high-dimensional regression with convex penalties' (DOI: 10.1214/21-AOS2100SUPP; .pdf). The supplement contains proofs and technical details that were omitted from the main text. It further provide discussion on the role the δ-bounded width assumption plays in the theory.

REFERENCES

[1] ADVANI, M. and GANGULI, S. (2016). Statistical mechanics of optimal convex inference in high dimensions. *Phys. Rev. X* **6** 031034.

[2] AMELUNXEN, D., LOTZ, M., MCCOY, M. B. and TROPP, J. A. (2014). Living on the edge: Phase transitions in convex programs with random data. *Inf. Inference* **3** 224–294. MR3311453 https://doi.org/10.1093/imaiai/iau005

[3] BA, K. D., INDYK, P., PRICE, E. and WOODRUFF, D. P. (2010). Lower bounds for sparse recovery. In *Proceedings of the Twenty-First Annual ACM-SIAM Symposium on Discrete Algorithms* 1190–1197. SIAM, Philadelphia, PA. MR2809736

[4] BANDEIRA, A. S., PERRY, A. and WEIN, A. S. (2018). Notes on computational-to-statistical gaps: Predictions using statistical physics. *Port. Math.* **75** 159–186. MR3892753 https://doi.org/10.4171/PM/2014

[5] BANKS, J., MOHANTY, S. and RAGHAVENDRA, P. (2021). Local statistics, semidefinite programming, and community detection. In *Proceedings of the* 2021 *ACM-SIAM Symposium on Discrete Algorithms* (*SODA*) 1298–1316. SIAM, Philadelphia, PA. MR4262512 https://doi.org/10.1137/1.9781611976465.79

[6] BARBIER, J., DIA, M., MACRIS, N. and KRZAKALA, F. (2016). The mutual information in random linear estimation. In 2016 54*th Annual Allerton Conference on Communication*, *Control*, *and Computing* (*Allerton*) 625–632.

[7] BARBIER, J., KRZAKALA, F., MACRIS, N., MIOLANE, L. and ZDEBOROVÁ, L. (2019). Optimal errors and phase transitions in high-dimensional generalized linear models. *Proc. Natl. Acad. Sci. USA* **116** 5451–5460. MR3939767 https://doi.org/10.1073/pnas.1802705116

[8] BARBIER, J., MACRIS, N., DIA, M. and KRZAKALA, F. (2020). Mutual information and optimality of approximate message-passing in random linear estimation. *IEEE Trans. Inf. Theory* **66** 4270–4303. MR4130617 https://doi.org/10.1109/TIT.2020.2990880

[9] BAYATI, M., LELARGE, M. and MONTANARI, A. (2015). Universality in polytope phase transitions and message passing algorithms. *Ann. Appl. Probab.* **25** 753–822. MR3313755 https://doi.org/10.1214/14-AAP1010

[10] BAYATI, M. and MONTANARI, A. (2011). The dynamics of message passing on dense graphs, with applications to compressed sensing. *IEEE Trans. Inf. Theory* **57** 764–785. MR2810285 https://doi.org/10.1109/TIT.2010.2094817

[11] BAYATI, M. and MONTANARI, A. (2012). The LASSO risk for Gaussian matrices. *IEEE Trans. Inf. Theory* **58** 1997–2017. MR2951312 https://doi.org/10.1109/TIT.2011.2174612

[12] BEAN, D., BICKEL, P. J., EL KAROUI, N. and YU, B. (2013). Optimal M-estimation in high-dimensional regression. *Proc. Natl. Acad. Sci. USA* **110** 14563–14568.

[13] BELLEC, P. C., LECUÉ, G. and TSYBAKOV, A. B. (2018). Slope meets Lasso: Improved oracle bounds and optimality. *Ann. Statist.* **46** 3603–3642. MR3852663 https://doi.org/10.1214/17-AOS1670

[14] BERTHIER, R., MONTANARI, A. and NGUYEN, P.-M. (2020). State evolution for approximate message passing with non-separable functions. *Inf. Inference* **9** 33–79. MR4079177 https://doi.org/10.1093/imaiai/iay021

[15] BICKEL, P. J., RITOV, Y. and TSYBAKOV, A. B. (2009). Simultaneous analysis of lasso and Dantzig selector. *Ann. Statist.* **37** 1705–1732. MR2533469 https://doi.org/10.1214/08-AOS620

[16] BOGDAN, M., VAN DEN BERG, E., SABATTI, C., SU, W. and CANDÈS, E. J. (2015). SLOPE—adaptive variable selection via convex optimization. *Ann. Appl. Stat.* **9** 1103–1140. MR3418717 https://doi.org/10.1214/15-AOAS842

[17] BOLTHAUSEN, E. (2014). An iterative construction of solutions of the TAP equations for the Sherrington–Kirkpatrick model. *Comm. Math. Phys.* **325** 333–366. MR3147441 https://doi.org/10.1007/s00220-013-1862-3

[18] BONDELL, H. D. and REICH, B. J. (2008). Simultaneous regression shrinkage, variable selection, and supervised clustering of predictors with OSCAR. *Biometrics* **64** 115–123. MR2422825 https://doi.org/10.1111/j.1541-0420.2007.00843.x

[19] BU, Z., KLUSOWSKI, J. M., RUSH, C. and SU, W. J. (2021). Algorithmic analysis and statistical estimation of SLOPE via approximate message passing. *IEEE Trans. Inf. Theory* **67** 506–537. MR4231969 https://doi.org/10.1109/TIT.2020.3025272

[20] CANDES, E. and TAO, T. (2007). The Dantzig selector: Statistical estimation when p is much larger than n. *Ann. Statist.* **35** 2313–2351. MR2382644 https://doi.org/10.1214/009053606000001523

[21] CANDES, E. J. and TAO, T. (2005). Decoding by linear programming. *IEEE Trans. Inf. Theory* **51** 4203–4215. MR2243152 https://doi.org/10.1109/TIT.2005.858979

[22] CELENTANO, M. and MONTANARI, A. (2022). Supplement to "Fundamental barriers to high-dimensional regression with convex penalties." https://doi.org/10.1214/21-AOS2100SUPP

[23] CHANDRASEKARAN, V., RECHT, B., PARRILO, P. A. and WILLSKY, A. S. (2012). The convex geometry of linear inverse problems. *Found. Comput. Math.* **12** 805–849. MR2989474 https://doi.org/10.1007/s10208-012-9135-7

[24] CHEN, S. S. and DONOHO, D. (1995). Examples of basis pursuit. In *Proceedings of Wavelet Applications in Signal and Image Processing III*, San Diego, CA.

[25] DONOHO, D., MALEKI, A. and MONTANARI, A. (2009). Message-passing algorithms for compressed sensing. *Proc. Natl. Acad. Sci. USA* **106** 18914–18919. https://doi.org/10.1073/pnas.0909892106

[26] DONOHO, D., MALEKI, A. and MONTANARI, A. (2010). Message passing algorithms for compressed sensing: I. Motivation and construction. In 2010 *IEEE Information Theory Workshop on Information Theory* (*ITW* 2010, *Cairo*) 1–5. IEEE.

[27] DONOHO, D. and MONTANARI, A. (2016). High dimensional robust M-estimation: Asymptotic variance via approximate message passing. *Probab. Theory Related Fields* **166** 935–969. MR3568043 https://doi.org/10.1007/s00440-015-0675-z

[28] EL KAROUI, N. (2013). Asymptotic behavior of unregularized and ridge-regularized high-dimensional robust regression estimators: Rigorous results. Available at arXiv:1311.2445.

[29] EL KAROUI, N. (2018). On the impact of predictor geometry on the performance on high-dimensional ridge-regularized generalized robust regression estimators. *Probab. Theory Related Fields* **170** 95–175. MR3748322 https://doi.org/10.1007/s00440-016-0754-9

[30] EL KAROUI, N., BEAN, D., BICKEL, P. J., LIM, C. and YU, B. (2013). On robust regression with high-dimensional predictors. *Proc. Natl. Acad. Sci. USA* **110** 14557–14562.

[31] FIGUEIREDO, M. A. T. and NOWAK, R. D. (2014). Sparse estimation with strongly correlated variables using ordered weighted L1 regularization. Available at arXiv:1409.4005.

[32] GAMARNIK, D. and ILIAS, Z. High dimensional regression with binary coefficients. Estimating squared error and a phase transtition. In *Proceedings of the* 2017 *Conference on Learning Theory* (S. Kale and O. Shamir, eds.). *Proceedings of Machine Learning Research* **65** 948–953.

[33] GAMARNIK, D. and ZADIK, I. (2017). Sparse high-dimensional linear regression. Algorithmic barriers and a local search algorithm. Available at arXiv:1711.04952.

[34] GUO, D., WU, Y., SHAMAI, S. and VERDÚ, S. (2011). Estimation in Gaussian noise: Properties of the minimum mean-square error. *IEEE Trans. Inf. Theory* **57** 2371–2385. MR2809096 https://doi.org/10.1109/TIT.2011.2111010

[35] JAVANMARD, A. and MONTANARI, A. (2013). State evolution for general approximate message passing algorithms, with applications to spatial coupling. *Inf. Inference* **2** 115–144. MR3311445 https://doi.org/10.1093/imaiai/iat004

[36] KORADA, S. B. and MONTANARI, A. (2011). Applications of the Lindeberg principle in communications and statistical learning. *IEEE Trans. Inf. Theory* **57** 2440–2450. MR2809100 https://doi.org/10.1109/TIT.2011.2112231

[37] LELARGE, M. and MIOLANE, L. (2019). Fundamental limits of symmetric low-rank matrix estimation. *Probab. Theory Related Fields* **173** 859–929. MR3936148 https://doi.org/10.1007/s00440-018-0845-x

[38] MA, J., XU, J. and MALEKI, A. (2019). Optimization-based AMP for phase retrieval: The impact of initialization and ℓ_2 regularization. *IEEE Trans. Inf. Theory* **65** 3600–3629. MR3959008 https://doi.org/10.1109/TIT.2019.2893254

[39] MÉZARD, M. and MONTANARI, A. (2009). *Information, Physics, and Computation. Oxford Graduate Texts*. Oxford Univ. Press, Oxford. MR2518205 https://doi.org/10.1093/acprof:oso/9780198570837.001.0001

[40] MIGNACCO, F., KRZAKALA, F., LU, Y., URBANI, P. and ZDEBOROVÁ, L. The role of regularization in classification of high-dimensional noisy Gaussian mixture. In *Proceedings of the* 37*th International Conference on Machine Learning* (H. Daumé, III and A. Singh, eds.). *Proceedings of Machine Learning Research* **119** 6874–6883.

[41] MIOLANE, L. and MONTANARI, A. (2021). The distribution of the Lasso: Uniform control over sparse balls and adaptive parameter tuning. *Ann. Statist.* **49** 2313–2335. MR4319252 https://doi.org/10.1214/20-aos2038

[42] OYMAK, S. and TROPP, J. A. (2018). Universality laws for randomized dimension reduction, with applications. *Inf. Inference* **7** 337–446. MR3858331 https://doi.org/10.1093/imaiai/iax011

[43] PARIKH, N. and BOYD, S. (2013). Proximal algorithms. *Found. Trends Optim.* **1** 123–231.

[44] RANGAN, S. (2011). Generalized approximate message passing for estimation with random linear mixing. In *Information Theory Proceedings* (*ISIT*), 2011 *IEEE International Symposium on* 2168–2172. IEEE.

[45] RANGAN, S., SCHNITER, P. and FLETCHER, A. K. (2014). On the convergence of approximate message passing with arbitrary matrices. In *Information Theory Proceedings* (*ISIT*), 2014 *IEEE International Symposium on* 236–240. IEEE.

[46] RANGAN, S., SCHNITER, P. and FLETCHER, A. K. (2019). Vector approximate message passing. In *Information Theory Proceedings* (*ISIT*), 2017 *IEEE International Symposium on* 1588–1592. IEEE.

[47] REEVES, G. and PFISTER, H. D. (2016). The replica-symmetric prediction for compressed sensing with Gaussian matrices is exact. In *Information Theory* (*ISIT*), 2016 *IEEE International Symposium on* 665–669. IEEE.

[48] SANTAMBROGIO, F. (2015). *Optimal Transport for Applied Mathematicians: Calculus of Variations, PDEs, and Modeling. Progress in Nonlinear Differential Equations and Their Applications* **87**. Birkhäuser/Springer, Cham. MR3409718 https://doi.org/10.1007/978-3-319-20828-2

[49] SAUMARD, A. and WELLNER, J. A. (2014). Log-concavity and strong log-concavity: A review. *Stat. Surv.* **8** 45–114. MR3290441 https://doi.org/10.1214/14-SS107

[50] SCHNITER, P. and RANGAN, S. (2015). Compressive phase retrieval via generalized approximate message passing. *IEEE Trans. Signal Process.* **63** 1043–1055. MR3311635 https://doi.org/10.1109/TSP.2014.2386294

[51] STOJNIC, M. (2010). Recovery thresholds for ℓ_1 optimization in binary compressed sensing. In *Information Theory Proceedings* (*ISIT*), 2010 *IEEE International Symposium on* 1593–1597. IEEE.

[52] STOJNIC, M. (2013). A framework to characterize performance of Lasso algorithms. Available at arXiv:1303.7291.

[53] SU, W. and CANDÈS, E. (2016). SLOPE is adaptive to unknown sparsity and asymptotically minimax. *Ann. Statist.* **44** 1038–1068. MR3485953 https://doi.org/10.1214/15-AOS1397

[54] SUR, P. and CANDÈS, E. J. (2019). A modern maximum-likelihood theory for high-dimensional logistic regression. *Proc. Natl. Acad. Sci. USA* **116** 14516–14525. MR3984492 https://doi.org/10.1073/pnas.1810420116

[55] TAHERI, H., PEDARSANI, R. and THRAMPOULIDIS, C. Sharp asymptotics and optimal performance for inference in binary models. In *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics* (S. Chiappa and R. Calandra, eds.). *Proceedings of Machine Learning Research* **108** 3739–3749.

[56] TAHERI, H., PEDARSANI, R. and THRAMPOULIDIS, C. Fundamental limits of ridge-regularized empirical risk minimization in high dimensions. In *Proceedings of the* 24*th International Conference on Artificial Intelligence and Statistics* (A. Banerjee and K. Fukumizu, eds.). *Proceedings of Machine Learning Research* **130** 2773–2781.

[57] THRAMPOULIDIS, C., ABBASI, E. and HASSIBI, B. (2018). Precise error analysis of regularized M-estimators in high dimensions. *IEEE Trans. Inf. Theory* **64** 5592–5628. MR3832326 https://doi.org/10.1109/TIT.2018.2840720

[58] THRAMPOULIDIS, C., OYMAK, S. and HASSIBI, B. (2015). Regularized linear regression: A precise analysis of the estimation error. In *Conference on Learning Theory* 1683–1709.

[59] TIBSHIRANI, R. (1996). Regression shrinkage and selection via the lasso. *J. Roy. Statist. Soc. Ser. B* **58** 267–288. MR1379242

[60] VAN DE GEER, S. A. and BÜHLMANN, P. (2009). On the conditions used to prove oracle results for the Lasso. *Electron. J. Stat.* **3** 1360–1392. MR2576316 https://doi.org/10.1214/09-EJS506

[61] WAINWRIGHT, M. J. (2009). Information-theoretic limits on sparsity recovery in the high-dimensional and noisy setting. *IEEE Trans. Inf. Theory* **55** 5728–5741. MR2597190 https://doi.org/10.1109/TIT.2009.2032816

[62] ZDEBOROVÁ, L. and KRZAKALA, F. (2015). Statistical physics of inference: Thresholds and algorithms. *Adv. Phys*. **65** 453–552.