Fundamental Barriers to High-Dimensional Regression with Convex Penalties

Michael Celentano* and Andrea Montanari†

February 8, 2022

Abstract

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.

Contents

*Department of Statistics, University of California, Berkeley
†Department of Electrical Engineering and Department of Statistics, Stanford University

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 semi-continuous (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:

> *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 [BBEKY13, AG16]. 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 non-log-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 3-point prior

Consider the case of noiseless linear measurements, namely $\boldsymbol{w} = \boldsymbol{0}$ in Eq. (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

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

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 [CD95], 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 non-convex 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 $\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\|^2/\|\boldsymbol{\beta}_0\|_2^2$. 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 sub-optimal 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 non

	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

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).

log-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 [DI17]. However, the approach of [DI17] 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 non log-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, (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 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 [MM09]. For tutorials discussing it in the context of statistical estimation, see Sections III E and IV B of [ZK15]; Sections 4.2 and 4.3 of [BPW18]. For recent contributions mentioning this idea or analogous ones in the context of matrix estimation, see [BMDK17, LM19, BMR21].

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*

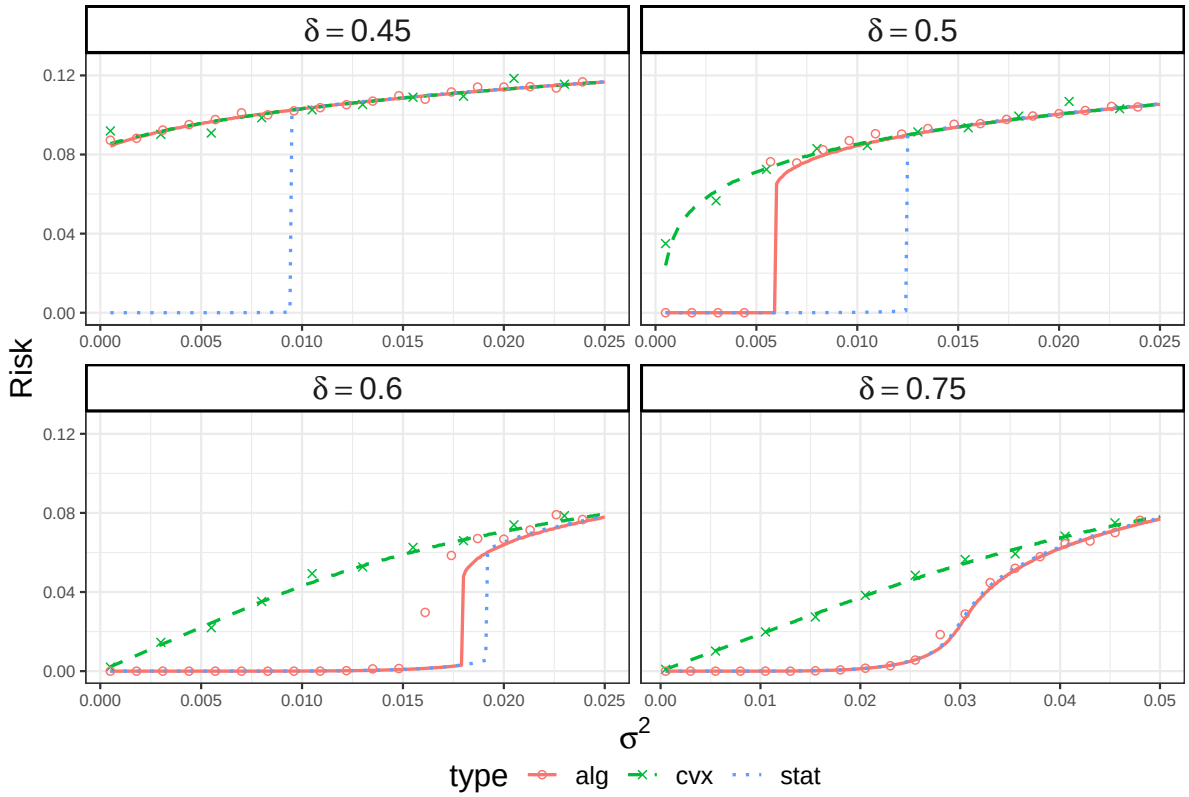

Figure 1: Median squared error of estimation in high-dimensional regression. Symbols refer to simulations for two different polynomial-time algorithms. Crosses: M-estimator (1.2) for a certain optimized penalty $\rho(\boldsymbol{\beta})$. Circles: Bayes-Approximate Message Passing. Dashed and solid lines 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.

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 [TAH18, BMDK17, BKM+19] (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 non log-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 non log-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, and 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, non-trivial phase diagrams occur exactly when the prior is non log-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 [KM11, BLM15, OT18] can be used to show that the same asymptotics hold for iid 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 [Tib96, CD95], 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 [CT05, CT07, BRT09, vdGB09]. 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 [Bol14, BM11, JM13]. In particular, the recent paper [BMN19] proves that state evolution holds for certain classes of non-separable nonlinearities. This is particularly relevant for the present setting, since we are interested in non-separable penalties $\rho(\boldsymbol{\beta})$.

The connection between M-estimation and AMP algorithms was first developed in [DMM09] and subsequently used in [BM12] 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 [DM16]. AMP algorithms were developed and analyzed for a number of statistical estimation problems, including generalized linear models [Ran11], phase retrieval [SR15, MXM19], and logistic regression [SC19].

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 [Sto13] and then developed by a number of authors in the context of regularized regression [TOH15], M-estimation [TAH18], generalized compressed sensing [CRPW12], binary compressed sensing [Sto10], the Lasso [MM18], 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 [EK13, EK18]. Closely related to the present work is the paper [BBEKY13], 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 [AG16].

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 [BDMK16, RP16]. Generalizations of this result were also obtained in [BKM+19] 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 [Wai09]. Strong lower bounds for compressed sensing reconstruction were proved in [BIPW10] using communication complexity ideas. Gamarnik and Zadik [DI17] 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 [GZ17]. 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 [MKL+20, TPT20, TPT21].

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 semi-definite matrices.

Subscripts under the expectation or probability sign, e.g. $\mathbb{E}_{\boldsymbol{\beta}_0,\boldsymbol{z}}$ 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 k-th moment. For a distribution $\pi \in \mathcal{P}_k(\mathbb{R})$, we will denote by $s_\ell(\pi)$ the ℓ-th 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

$$d_{\mathrm{W}}(\pi,\pi') = \inf_{X,X'} \left(\mathbb{E}_{X,X'}\left[(X-X')^2\right]\right)^{1/2}, \tag{1.4}$$

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 $\mathcal{P}_2(\mathbb{R})$ [San15]. Convergence in Wasserstein metric will be denoted $\overset{\mathrm{W}}{\to}$, and we use $\overset{\mathrm{p}}{\to}$, $\overset{\mathrm{as}}{\to}$, $\overset{\mathrm{d}}{\to}$ 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

$$\operatorname*{plim\,inf}_{p\to\infty} X_p = \sup\left\{t \in \mathbb{R} \,\middle|\, \lim_{p\to\infty} \mathbb{P}\left(X_p < t\right) = 0\right\},$$

and $\operatorname*{plim\,sup}_{p\to\infty} X_p = -\operatorname*{plim\,inf}_{p\to\infty}(-X_p)$. 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 $|X_p - Y_p| \overset{\mathrm{p}}{\to} 0$.

We adopt the convention that when the minimizing set in (1.2) is empty, $\widehat{\boldsymbol{\beta}}_{\mathsf{cvx}} = \boldsymbol{\infty}$ and $\|\boldsymbol{\infty} - \boldsymbol{x}\| = \infty$ 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

$$\|\varphi(\boldsymbol{x}_1, \dots, \boldsymbol{x}_\ell) - \varphi(\boldsymbol{y}_1, \dots, \boldsymbol{y}_\ell)\| \le C \left(1 + \sum_{i=1}^{\ell} \|\boldsymbol{x}_i\|^{k-1} + \sum_{i=1}^{\ell} \|\boldsymbol{y}_i\|^{k-1}\right) \sum_{i=1}^{\ell} \|\boldsymbol{x}_i - \boldsymbol{y}_i\|, \tag{1.5}$$

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 [RP16, BDMK16, BKM+19]:

- **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

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

While the second inequality holds by the statistical optimality of the Bayes risk, the first is non-trivial. Previous work established exactly when the second inequality is strict [BKM+19]. 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 [BBEKY13, AG16].

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 $\sigma^2 \ge 0$, these satisfy

$$\widehat{\pi}_{\boldsymbol{\beta}_0} := \frac{1}{p} \sum_{j=1}^{p} \delta_{\sqrt{p}\beta_{0j}} \overset{\text{W}}{\to} \pi \qquad \text{and} \qquad \frac{1}{n}\|\boldsymbol{w}\|^2 \to \sigma^2. \tag{2.1}$$

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

$$\beta_{0j} \overset{\text{iid}}{\sim} \pi/\sqrt{p}, \quad \boldsymbol{w} \sim \mathsf{N}(0, \sigma^2 \boldsymbol{I}_n), \tag{2.2}$$

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

$$\boldsymbol{y}_{\mathsf{seq}} = \boldsymbol{\beta}_0 + \tau \boldsymbol{z}, \tag{2.3}$$

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*:

$$\widehat{\boldsymbol{\beta}}_{\mathsf{seq}} := \arg\min_{\boldsymbol{\beta}} \frac{1}{2}\|\boldsymbol{y}_{\mathsf{seq}} - \boldsymbol{\beta}\|^2 + \lambda\rho(\boldsymbol{\beta}) =: \mathsf{prox}[\lambda\rho](\boldsymbol{y}_{\mathsf{seq}}). \tag{2.4}$$

By strong convexity, when ρ is lower semi-continuous and proper, the minimizer exists and is unique [PB13].

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 (2.4) in the sequence model [BM12, DM16,

EKBB+13, EK13, TOH15, TAH18]. A typical characterization takes the following form. For a sequence of penalties $\{\rho_p\}$, let (τ, λ) solve

$$\delta\tau^2 - \sigma^2 = \lim_{p\to\infty} \mathbb{E}_{\boldsymbol{z}}[\|\mathsf{prox}[\lambda\rho_p](\boldsymbol{\beta}_0 + \tau\boldsymbol{z}) - \boldsymbol{\beta}_0\|^2], \tag{2.5a}$$

$$2\lambda\left(1 - \frac{1}{\delta\tau}\lim_{p\to\infty} \mathbb{E}_{\boldsymbol{z}}[\langle \boldsymbol{z}, \mathsf{prox}[\lambda\rho_p](\boldsymbol{\beta}_0 + \tau\boldsymbol{z})\rangle]\right) = 1. \tag{2.5b}$$

Then under the HDA and DSN assumption,

$$\|\widehat{\boldsymbol{\beta}}_{\mathsf{cvx}} - \boldsymbol{\beta}_0\|^2 \xrightarrow{\mathrm{P}} \delta\tau^2 - \sigma^2 = \mathbb{E}_{\boldsymbol{z}}[\|\mathsf{prox}[\lambda\rho_p](\boldsymbol{\beta}_0 + \tau\boldsymbol{z}) - \boldsymbol{\beta}_0\|^2]. \tag{2.6}$$

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) [DM16, EK13]; other work involves analysis tailored to a specific penalty like the LASSO or SLOPE [BM12, BKRS21]. 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 Eq. (2.5) 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

$$\mathsf{R}^{\mathsf{opt}}_{\mathsf{seq,cvx}}(\tau; \pi, p) := \inf_{\rho\in\mathcal{C}_p} \mathbb{E}_{\boldsymbol{\beta}_0,\boldsymbol{z}}\left[\|\mathsf{prox}[\rho](\boldsymbol{\beta}_0 + \tau\boldsymbol{z}) - \boldsymbol{\beta}_0\|^2\right], \tag{2.7}$$

where $\boldsymbol{\beta}_0, \boldsymbol{z}$ are as in (2.3), and the optimal asymptotic risk using the sequences in $\mathcal{C}$ is

$$\mathsf{R}^{\mathsf{opt}}_{\mathsf{seq,cvx}}(\tau; \pi) = \liminf_{p\to\infty} \mathsf{R}^{\mathsf{opt}}_{\mathsf{seq,cvx}}(\tau; \pi, p) = \inf_{\{\rho_p\}\in\mathcal{C}} \liminf_{p\to\infty} \mathbb{E}_{\boldsymbol{\beta}_0,\boldsymbol{z}}\left[\|\mathsf{prox}[\rho_p](\boldsymbol{\beta}_0 + \tau\boldsymbol{z}) - \boldsymbol{\beta}_0\|^2\right]. \tag{2.8}$$

We will study a quantity similar to (2.8) 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:*

$$\begin{gathered}\textit{for all compact } T \subset (0,\infty), \textit{ there exists } \bar{\lambda} = \bar{\lambda}(T) < \infty \textit{ such that} \\ \limsup_{p\to\infty} \sup_{\lambda>\bar{\lambda},\tau\in T} \frac{1}{\tau}\mathbb{E}_{\boldsymbol{\beta}_0,\boldsymbol{z}}\left[\langle \boldsymbol{z}, \mathsf{prox}\left[\lambda\rho_p\right](\boldsymbol{\beta}_0 + \tau\boldsymbol{z})\rangle\right] < \delta.\end{gathered} \tag{2.9}$$

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 (2.9) with the Gaussian width of convex cones [CRPW12, ALMT14], 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, Eq. (O.11) of the Supplementary Material [CM21]). Thus,

$$\mathcal{C}_{\delta,\pi} = \mathcal{C} \quad \text{if } \delta > 1. \tag{2.10}$$

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 $\sigma \geq 0$. Define*

$$\tau^2_{\mathsf{reg},\mathsf{cvx}} = \sup\left\{\tau^2 \,\middle|\, \delta\tau^2 - \sigma^2 < \mathsf{R}^{\mathsf{opt}}_{\mathsf{seq},\mathsf{cvx}}(\tau;\pi)\right\}. \tag{2.11}$$

Under the HDA and RSN assumptions,[1]

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

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

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 Supplemenatary Material [CM21]. 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 [CM21] discusses the role of the restriction to $\mathcal{C}_{\delta,\pi}$. Because $\mathsf{R}^{\mathsf{opt}}_{\mathsf{seq},\mathsf{cvx}}(\tau;\pi)$ is continuous in τ whenever $\mathcal{C}$ is such that $\tau^2_{\mathsf{reg},\mathsf{cvx}}$ is finite (see Lemma C.2 of the Supplementary Material [CM21]), we will always have $\delta\tau^2_{\mathsf{reg},\mathsf{cvx}} - \sigma^2 = \mathsf{R}^{\mathsf{opt}}_{\mathsf{seq},\mathsf{cvx}}(\tau_{\mathsf{reg},\mathsf{cvx}};\pi)$ 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 $\tau^2_{\mathsf{reg},\mathsf{cvx}}$.*

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 (2.9). 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 (2.9). The convex lower bound in these two cases is the same.

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

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

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 (2.3). In particular, define

$$\mathsf{mmse}_\pi(\tau^2) = \mathbb{E}_{\beta_0,z}[(\mathbb{E}_{\beta_0,z}[\beta_0|\beta_0+\tau z]-\beta_0)^2], \tag{2.13}$$

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

$$\mathsf{mmse}_\pi(\tau^2) = \mathbb{E}_{\boldsymbol{\beta}_0,\boldsymbol{z}}\left[\left\|\mathbb{E}_{\boldsymbol{\beta}_0,\boldsymbol{z}}\left[\boldsymbol{\beta}_0|\sqrt{p}\boldsymbol{\beta}_0+\tau\boldsymbol{z}\right]-\boldsymbol{\beta}_0\right\|^2\right], \tag{2.14}$$

we see that $\mathsf{mmse}_\pi(\tau^2)$ is analogous to (2.7) except that the infimum is taken over all estimators, not just those in a restricted class. Finally, analogous to (2.11), define

$$\tau^2_{\mathsf{reg,amp}*} := \sup\left\{\tau^2 \,\middle|\, \delta\tau^2-\sigma^2 \le \mathsf{mmse}_\pi(\tau^2)\right\}. \tag{2.15}$$

Note that because $\mathsf{mmse}_\pi(\tau^2)$ is continuous in τ [DYSV11],

$$\delta\tau^2_{\mathsf{reg,amp}*}-\sigma^2 = \mathsf{mmse}_\pi(\tau^2_{\mathsf{reg,amp}*}). \tag{2.16}$$

As we will see, Bayes AMP asymptotically achieves estimation error arbitrary 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* $\pi \in \mathcal{P}_2(\mathbb{R})$,

$$\tau^2_{\mathsf{reg,cvx}} \ge \tau^2_{\mathsf{reg,amp}*} \tag{2.17}$$

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

$$\inf_{\{\rho_p\}\in\mathcal{C}_{\delta,\pi}} \operatorname*{plim\,inf}_{p\to\infty} \|\widehat{\boldsymbol{\beta}}_{\mathsf{cvx}} - \boldsymbol{\beta}_0\|^2 \geq \delta\tau^2_{\mathsf{reg,amp}*} - \sigma^2. \tag{2.18}$$

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

Proof of Corollary 2.3. Define

$$\tau^2_{\mathsf{reg,amp}} = \sup\left\{\tau^2 \,\middle|\, \delta\tau^2 - \sigma^2 < \mathsf{mmse}_\pi(\tau^2)\right\}. \tag{2.19}$$

In Section L of the Supplementary Material [CM21], 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 (2.7) and (2.14), we have $\mathsf{mmse}_\pi(\tau^2) \leq \mathsf{R}^{\mathsf{opt}}_{\mathsf{seq,cvx}}(\tau;\pi,p)$. By (2.8)), we obtain $\mathsf{mmse}_\pi(\tau^2) \leq \mathsf{R}^{\mathsf{opt}}_{\mathsf{seq,cvx}}\tau;\pi)$. Thus, the set $\left\{\tau^2 \mid \delta\tau^2 - \sigma^2 < \mathsf{mmse}_\pi(\tau^2)\right\} \subseteq \left\{\tau^2 \mid \delta\tau^2 - \sigma^2 < \mathsf{R}^{\mathsf{opt}}_{\mathsf{seq,cvx}}(\tau^2;\pi)\right\}$, and (2.17) follows from (2.11) and (2.19). Theorem 1 then gives (2.18). □

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 [DMM10, BKM+19]. Define the scalar iteration

$$\tau_0^2 = \frac{1}{\delta}\left(\sigma^2 + s_2(\pi)\right), \tag{2.20a}$$

$$\tau_{t+1}^2 = \frac{1}{\delta}\left(\sigma^2 + \mathsf{mmse}_\pi\left(\tau_t^2\right)\right). \tag{2.20b}$$

Moreover, let

$$\eta_t(y) = \mathbb{E}_{\beta_0,z}[\beta_0 | \beta_0 + \tau_t z = y] \tag{2.21a}$$

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

$$\mathsf{b}_t = \frac{1}{\delta}\mathbb{E}_{\beta_0,z}\left[\eta'_{t-1}\left(\beta_0 + \tau_{t-1}z\right)\right], \tag{2.22}$$

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

$$\eta_t(\boldsymbol{x})_j = \frac{1}{\sqrt{p}}\eta_t(\sqrt{p}x_j), \tag{2.23}$$

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

$$\begin{aligned} \boldsymbol{r}^t &= \frac{\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}^t}{n} + \mathsf{b}_t\boldsymbol{r}^{t-1}, \\ \widehat{\boldsymbol{\beta}}^{t+1} &= \eta_t\left(\widehat{\boldsymbol{\beta}}^t + \boldsymbol{X}^\mathsf{T}\boldsymbol{r}^t\right), \end{aligned} \tag{2.24}$$

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 (2.20) and $\widehat{\boldsymbol{\beta}}^t$ as defined by (2.24). Under the HDA and either the DSN or RSN assumptions, for any fixed t we have*

$$\lim_{p\to\infty}{}^{\mathrm{p}} \|\widehat{\boldsymbol{\beta}}^t - \boldsymbol{\beta}_0\|^2 = \mathsf{mmse}_\pi\left(\tau_t^2\right) \tag{2.25}$$

Further,

$$\lim_{t\to\infty} \tau_t^2 = \tau^2_{\mathsf{reg,amp}*}. \tag{2.26}$$

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

$$\lim_{p\to\infty}{}^{\mathrm{p}} \|\widehat{\boldsymbol{\beta}}^t - \boldsymbol{\beta}_0\|^2 \leq \delta\tau^2_{\mathsf{reg,amp}*} - \sigma^2 + \varepsilon. \tag{2.27}$$

Proposition 2.4 states that the state evolution (2.20) characterizes the large n, p behavior of Bayes AMP. It follows from standard results in the AMP literature [BM11]. A minor technical difficulty is that the main theorem of [BM11] requires Lipschitz non-linearities in the AMP iteration. The Bayes estimator η_t need not be Lipschitz. Thus, to apply the results of [BM11], 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 [CM21].

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

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

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

$$\phi(\tau^2;\pi,\delta,\sigma) = \frac{\sigma^2}{2\tau^2} - \frac{\delta}{2}\log\left(\frac{\sigma^2}{\tau^2}\right) + i(\tau^2), \tag{2.28}$$

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,

$$i(\tau^2) = \mathbb{E}_{\beta_0,z}\left[\log\frac{p(y|\beta_0)}{p(y)}\right] = -\frac{1}{2} - \mathbb{E}_{\beta_0,z}\log\left\{\int e^{-\frac{1}{2}(y-\beta/\tau)^2}\pi(\mathrm{d}\beta)\right\}. \tag{2.29}$$

Also define

$$\tau_{\mathsf{reg,stat}}(\pi;\delta,\sigma) = \arg\min_{\tau\geq 0}\phi(\tau^2;\pi,\delta,\sigma), \tag{2.30}$$

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

$$\frac{\mathrm{d}}{\mathrm{d}\tau^{-2}}\phi(\tau^2;\pi,\delta,\sigma) = \frac{1}{2}\left(\sigma^2 - \delta\tau^2 + \mathsf{mmse}_\pi(\tau^2)\right), \tag{2.31}$$

where we have used that $\frac{\mathrm{d}}{\mathrm{d}\tau^{-2}} i(\tau^2) = \frac{1}{2}\mathsf{mmse}_\pi(\tau^2)$ by [DYSV11, Corollary 1]. We see that if $\tau_{\mathsf{reg,stat}} > 0$, then

$$\delta\tau_{\mathsf{reg,stat}}^2 - \sigma^2 = \mathsf{mmse}_\pi(\tau_{\mathsf{reg,stat}}^2). \tag{2.32}$$

Equation (2.32) is closely related to (2.19). 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 [BKM+19])**.** *Fix $\pi \in \mathcal{P}_\infty(\mathbb{R})$, $\delta \in (0,\infty)$, and $\sigma > 0$. Under the HDA and RSN assumptions,*

$$\lim_{p\to\infty} \mathbb{E}_{\boldsymbol{\beta}_0,\boldsymbol{w},\boldsymbol{X}}\left[\|\mathbb{E}_{\boldsymbol{\beta}_0,\boldsymbol{w},\boldsymbol{X}}[\boldsymbol{\beta}_0|\boldsymbol{y},\boldsymbol{X}] - \boldsymbol{\beta}_0\|^2\right] = \mathsf{mmse}_\pi(\tau_{\mathsf{reg,stat}}^2) = \delta\tau_{\mathsf{reg,stat}}^2 - \sigma^2, \tag{2.33}$$

whenever the minimizer in (2.30) is unique. This occurs for almost every (δ,σ) (w.r.t. Lebesgue measure).

This is a specific case of Theorem 2 of [BKM+19]. We carry out the conversion from their notation to ours in Section L of the Supplementary Material [CM21]. This result was previously established under slightly less general conditions in [TAH18, BMDK17]. 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 $\tau_{\mathsf{reg,stat}}^2$.*

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 $\tau_{\mathsf{reg,cvx}}^2$, $\tau_{\mathsf{reg,amp}*}^2$, and $\tau_{\mathsf{reg,stat}}^2$. 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*

$$\tau_{\mathsf{reg,cvx}}^2 \geq \tau_{\mathsf{reg,stat}}^2. \tag{2.34}$$

Proof. The inequality $\tau_{\mathsf{reg,cvx}}^2 \geq \tau_{\mathsf{reg,amp}}^2$ holds because the supremum in (2.19) is taken over a subset of the supremum in (2.11). Thus, it suffices to show $\tau_{\mathsf{reg,amp}}^2 \geq \tau_{\mathsf{reg,stat}}^2$. For $\tau' < \tau_{\mathsf{reg,stat}}$,

$$\phi(\tau_{\mathsf{reg,stat}};\pi,\delta,\sigma) < \phi(\tau';\pi,\delta,\sigma) \tag{2.35}$$

$$= \phi(\tau_{\mathsf{reg,stat}};\pi,\delta,\sigma) + \frac{1}{2}\int_{\tau_{\mathsf{reg,stat}}^{-2}}^{\tau'^{-2}} \left(\sigma^2 - \delta\tau^2 + \mathsf{mmse}_\pi(\tau^2)\right) d\tau^{-2}. \tag{2.36}$$

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 (2.19), we have $\tau_{\mathsf{reg,amp}} \geq \tau_{\mathsf{reg,stat}}$, 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 (2.30) 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 $\delta \leq 1$ 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, non-trivial 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 $\sigma \geq 0$. 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 $\tau \geq 0$ 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

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

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

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

$$\inf_{\{\rho_p\} \in \mathcal{C}_{\delta,\pi}} \operatorname*{p\,liminf}_{p\to\infty} \|\widehat{\boldsymbol{\beta}}_{\mathsf{cvx}} - \boldsymbol{\beta}_0\|^2 > \delta\tau^2 - \sigma^2. \tag{3.2}$$

(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 [CM21].

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 (2.3), 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 [BBEKY13]. See Section J of the Supplementary Material [CM21] 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 [CM21]). If* $\mathcal{C}$ *contains all sequences of convex penalties, then under the HDA and RSN assumptions, inequality* (2.18) *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* (2.18).

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 (2.3) 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 [CM21] 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 (2.15), 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 $\pi * \mathsf{N}(0, \tau^2)$ [SW14, 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 (2.18) holds with the inequality reversed, so in fact holds with equality.

We now prove the "only if" direction. By (2.15) and the continuity of $\mathsf{mmse}_\pi(\tau^2)$ in τ^2 [DYSV11, Proposition 7], we have

$$\delta\tau^2_{\mathsf{reg,amp}*} - \sigma^2 = \mathsf{mmse}_\pi(\tau^2_{\mathsf{reg,amp}*}). \tag{3.3}$$

If $\pi * \mathsf{N}(0, \tau^2_{\mathsf{reg,amp}*})$ does not have log-concave density, by Proposition 3.1.(*ii*) Eq. (2.18) holds with strict inequality. By Lemma K.1 of the Supplementary Material [CM21], 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 (2.18) holds with equality.

(b) If $\sigma > 0$ and π does not have log-concave density, then there exist $0 \leq \delta_{\mathsf{alg}} < \infty$ such that inequality (2.18) 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 that 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 non-convex methods.

Proof of Corollary 3.2.

Part (a): By [SW14, 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 $\delta_{\mathsf{alg}} = \inf\{\delta \mid \pi * \mathsf{N}(0, \tau^2_{\mathsf{reg,amp}}) \text{ does not have log-concave density}\}$. By [SW14, 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 (2.19), $\tau_{\mathsf{reg,amp}}$ is non-increasing in δ. Combining these two facts, for $\delta > \delta_{\mathsf{alg}}$ we have $\mathsf{N}(0, \tau^2_{\mathsf{reg,amp}})$ does not have log-concave density, and for $\delta < \delta_{\mathsf{alg}}$ we have $\mathsf{N}(0, \tau^2_{\mathsf{reg,amp}})$ does have log-concave density. Then, by Theorem 2, inequality (2.18) 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 (2.16), $\tau^2_{\mathsf{reg,amp}} = \frac{1}{\delta}\big(\sigma^2 + \mathsf{mmse}_\pi(\tau^2_{\mathsf{reg,amp}})\big) \leq \frac{1}{\delta}\big(\sigma^2 + s_2(\pi)\big)$. Thus, $\lim_{\delta \to \infty} \tau^2_{\mathsf{reg,amp}} = 0$. Because log-concavity is preserved under convergence in distribution [SW14, 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 Eq. (2.28) 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*

$$\inf_{\{\rho_p\}_p \in \mathcal{C}_{\delta,\pi}} \liminf_{p \to \infty} \mathbb{E}_{\boldsymbol{\beta}_0, \boldsymbol{w}, \boldsymbol{X}} \left[\|\widehat{\boldsymbol{\beta}}_{\mathsf{cvx}} - \boldsymbol{\beta}_0\|^2 \right] = \lim_{p \to \infty} \mathbb{E}_{\boldsymbol{\beta}_0, \boldsymbol{w}, \boldsymbol{X}} \left[\|\mathbb{E}_{\boldsymbol{\beta}_0, \boldsymbol{w}, \boldsymbol{X}}[\boldsymbol{\beta}_0 | \boldsymbol{y}] - \boldsymbol{\beta}_0\|^2 \right], \tag{3.4}$$

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 (2.3) 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 [CM21] 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 (3.4) is $\delta\tau^2_{\mathsf{reg,stat}} - \sigma^2$ by Proposition 2.6 (this is where we use $\sigma > 0$). By (2.34), 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 [CM21], we have under the RSN assumption that (3.4) 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

$$\begin{aligned}\phi(\tau_{\mathsf{reg,stat}}; \pi, \delta, \sigma) &= \phi(\tau'; \pi, \delta, \sigma) + \frac{1}{2}\int_{\tau'^{-2}}^{\tau_{\mathsf{reg,stat}}^{-2}} \left(\sigma^2 - \delta\tau^2 + \mathsf{mmse}_\pi(\tau^2)\right) d\tau^{-2} \\ &> \phi(\tau_{\mathsf{reg,stat}}; \pi, \delta, \sigma) + \frac{1}{2}\int_{\tau'^{-2}}^{\tau_{\mathsf{reg,stat}}^{-2}} \left(\sigma^2 - \delta\tau^2 + \mathsf{mmse}_\pi(\tau^2)\right) d\tau^{-2}, \end{aligned} \tag{3.5}$$

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 [SW14, 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

$$\inf_{\{\rho_p\}_p \in \mathcal{C}_{\delta,\pi}} \lim_{p\to\infty} \mathbb{E}_{\boldsymbol{\beta}_0, \boldsymbol{w}, \boldsymbol{X}}\left[\|\widehat{\boldsymbol{\beta}}_{\mathsf{cvx}} - \boldsymbol{\beta}_0\|^2\right] \leq \delta\tau^2_{\mathsf{reg,stat}} - \sigma^2. \tag{3.6}$$

By (2.33), we have $\delta\tau^2_{\mathsf{reg,stat}} - \sigma^2$ equals the right-hand side of (3.4). The reverse inequality holds by the optimality of the Bayes risk, whence we conclude (3.4). □

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* (3.4) *holds.*

(b) If π does not have log-concave density with respect to Lebesgue measure, then there exist $0 \le \delta_{\mathsf{stat}} < \infty$ such that equality (3.4) holds for all $\delta < \delta_{\mathsf{stat}}$ for which ϕ has unique minimizer, and (3.4) holds with strict inequality replacing equality for all $\delta > \delta_{\mathsf{stat}}$ for which ϕ has unique minimizer. Moreover, $\delta_{\mathsf{stat}} \le \delta_{\mathsf{alg}}$.

Proof of Corollary 3.3.

Part (a): By [SW14, 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 (2.31) of ϕ with respect to τ^{-2} is strictly decreasing in δ, we have by (2.28) that $\tau_{\mathsf{reg,stat}}$ is strictly decreasing in δ. As in the proof of Corollary 3.2, this implies that for for $\delta > \delta_{\mathsf{stat}}$ we have $\mathsf{N}(0, \tau^2_{\mathsf{reg,stat}})$ does not have log-concave density and for $\delta < \delta_{\mathsf{stat}}$ we have $\mathsf{N}(0, \tau^2_{\mathsf{reg,stat}})$ 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 (3.4) 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 (2.30) and (2.31), we have $\tau^2_{\mathsf{reg,stat}} = \frac{1}{\delta}\left(\sigma^2 + \mathsf{mmse}_\pi(\tau^2_{\mathsf{reg,stat}})\right) \le \frac{1}{\delta}\left(\sigma^2 + s_2(\pi)\right)$, 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 [SW14, 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

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

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 (2.30) is unique, we have

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

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

$$\limsup_{\mathsf{snr}\to 0} \frac{\Delta(\pi,\delta,\sigma)}{\mathsf{snr}^2} \le s_2(\pi)\delta^2 \frac{s_3^2(\pi)}{2s_2^3(\pi)}, \tag{4.2}$$

where the lim sup *is taken over* σ *at which* (2.28) *has unique minimizer.*

The proof of this theorem is given in Section M of the Supplementary Material [CM21]. We believe its results provide some useful insight:

- The large δ regime of point (i) is most commonly analyzed in the statistics literature, because it ensures high-dimensional consistency. 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 $\pi * \mathsf{N}(0, \tau^2_{\mathsf{reg},\mathsf{stat}})$ will in some sense be "more log-concave" when $\tau^2_{\mathsf{reg},\mathsf{stat}}$ is larger. Because $\tau^2_{\mathsf{reg},\mathsf{stat}}$ 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 (2.5) 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* $\tau, \lambda, \tilde{\tau}, \tilde{\lambda}$ *are such that* $\tau, \lambda, \delta, \mathcal{T}$ *and* $\tilde{\tau}, \tilde{\lambda}, \delta, \tilde{\mathcal{T}}$ *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*

$$\widehat{\boldsymbol{\beta}}_{\mathsf{cvx}+} = \mathsf{prox}\left[\lambda \rho_p\right]\left(\widehat{\tilde{\boldsymbol{\beta}}}_{\mathsf{cvx}} + \frac{2\lambda}{n}\boldsymbol{X}^{\mathsf{T}}(\boldsymbol{y} - \boldsymbol{X}\widehat{\tilde{\boldsymbol{\beta}}}_{\mathsf{cvx}}) + \sqrt{\tau^2 - \tilde{\tau}^2}\boldsymbol{z}\right), \tag{5.1}$$

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*

$$\ell_p\left(\boldsymbol{\beta}_0, \widehat{\boldsymbol{\beta}}_{\mathsf{cvx}+}\right) \overset{\mathrm{p}}{\simeq} \ell_p\left(\boldsymbol{\beta}_0, \widehat{\boldsymbol{\beta}}_{\mathsf{cvx}}\right). \tag{5.2}$$

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

We prove Proposition 5.1 in Section H of the Supplementary Material [CM21]. 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*

$$\inf_{\{\rho_p\} \in \mathcal{C}_*} \operatorname*{plim\,inf}_{p\to\infty} \frac{1}{p} \sum_{j=1}^{p} \ell\left(\sqrt{p}\beta_{0j}, \sqrt{p}\widehat{\beta}_{\mathsf{cvx},j}\right) \geq \mathbb{E}_{\beta_0,z}[\ell(\beta_0, \eta(\beta_0 + \tau_{\mathsf{reg,cvx}} z)]. \tag{5.3}$$

When η is not the proximal operator of a convex function, inequality (5.3) is strict.

Further, when $\delta > 1$,

$$\inf_{\substack{\{\rho_p\} \in \mathcal{C}_* \\ \eta' \text{ Lipschitz}}} \operatorname*{plim}_{p\to\infty} \frac{1}{p} \sum_{j=1}^{p} \ell\left(\sqrt{p}\beta_{0j}, \eta'\left(\sqrt{p}\widehat{\boldsymbol{\beta}}_{\mathsf{cvx},j} + 2\lambda \frac{[\boldsymbol{X}^{\mathsf{T}}(\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}_{\mathsf{cvx}})]_j}{n}\right)\right)$$
$$= \mathbb{E}_{\beta_0,z}[\ell(\beta_0, \eta(\beta_0 + \tau_{\mathsf{reg,cvx}} z)]. \tag{5.4}$$

The sequences $\{\rho_p\}$ which minimize the ℓ_2 loss of $\widehat{\boldsymbol{\beta}}_{\mathsf{cvx}}$ also achieve the infimum in (5.4). (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 [CM21]. We expect inequality (5.3) 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 (5.4) to hold also when $\delta \leq 1$, but this requires establishing the tightness of the convex lower bound when $\delta \leq 1$, which are 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 [CM21], 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 (2.7)) and the Bayes estimator (as done in (2.14)). Theorem 5 states that by post-processing we may effectively replace the proximal operator in Eq. (H.1) of the Supplementary Material [CM21] by a non-proximal 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. [RSF14, RSF17]).

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 Section I of the Supplementary Material [CM21]. 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 [CM21]. 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

$$\mathcal{C}_* = \Big\{ \{\rho_p\} \in \mathcal{C} \mid \{\rho_p\} \text{ is uniformly strongly convex} \Big\}. \tag{6.1}$$

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

$$\rho_p(\boldsymbol{x}) = \mathbb{I}_{C_p}(\boldsymbol{x}) := \begin{cases} 0 & \boldsymbol{x} \in C_p \\ \infty & \text{otherwise,} \end{cases} \tag{6.2}$$

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

$$\widehat{\boldsymbol{\beta}}_{\mathsf{cvx}} = \arg\min_{\boldsymbol{\beta}} \Big\{ \frac{1}{n}\|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\|^2 \ : \ \boldsymbol{\beta} \in C_p \Big\}. \tag{6.3}$$

In this context, the condition (2.9) is closely related to bounding the Gaussian width of convex cones [CRPW12, ALMT14]. 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

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

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 $\boldsymbol{z} \sim \mathsf{N}(0, \boldsymbol{I}_p/p)$:

$$w(K) := \mathbb{E}_{\boldsymbol{z}}\left[\|\Pi_K(\boldsymbol{z})\|^2\right]. \tag{6.4}$$

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 (6.3) if and only if $\limsup_{p\to\infty} w(T_{C_p}(\boldsymbol{\beta}_0)) < \delta$ [ALMT14, CRPW12]. 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*

$$\lim_{p\to\infty} \mathbb{E}_{\boldsymbol{\beta}_0}\left[d\left(\boldsymbol{\beta}_0, C_p\right)\right] = 0. \tag{6.5}$$

Further assume that

$$\lim_{\varepsilon\to 0} \limsup_{p\to\infty} \mathbb{E}_{\boldsymbol{\beta}_0}\left[w(T_{C_p \cap B^c(\boldsymbol{\beta}_0, \varepsilon)}(\boldsymbol{\beta}_0))\right] < \delta, \tag{6.6}$$

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 (6.6) 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 (6.3). 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 (6.3) 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 (6.3) 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

$$\rho_p(\boldsymbol{x}) = \frac{1}{p}\sum_{j=1}^{p} \rho(\sqrt{p}x_j), \tag{6.7}$$

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 [BBEKY13, EKBB+13, DM16], and a few works have broken the separability assumption [TAH18]. While Theorem 1 is more general, it applies to separable penalties under a mild condition.

Proposition 6.4. *Consider ρ_p as in (6.7) 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 non-empty, we have*

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

if and only if $\{\rho_p\} \in C_{\delta,\pi}$.

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

$$\rho_p(\boldsymbol{x}) = \frac{1}{\sqrt{p}}\sum_{j=1}^{p} \kappa_j^{(p)} |x|_{(j)}, \tag{6.8}$$

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 (6.8) have been used for a few purposes. SLOPE has recently been proposed for sparse regression because it automatically adapts to sparsity level [BvdBS+15, SC16, BLT18]. More generally, the use of OWL norms has been argued to produce estimators which are more stable than LASSO under correlated designs [BR08, FN14].

Proposition 6.5. *Consider ρ_p as in (6.8). If for all $\varepsilon > 0$ there exists $\xi > 0$ such that $j \leq (1-\varepsilon)p$ implies $\kappa_j^{(p)} > \xi$, then $\{\rho_p\} \in \mathcal{C}_{\delta,\pi}$.*

Acknowledgements

MC was supported by the National Science Foundation Graduate Research Fellowship under Grant No. DGE – 1656518. AM was supported by NSF grants CCF-2006489 and the ONR grant N00014-18-1-2729.

References

[AG16] Madhu Advani and Surya Ganguli. Statistical mechanics of optimal convex inference in high dimensions. *Physical Review X*, 6(3):031034, 2016.

[AGZ10] Greg W. Anderson, Alice Guionnet, and Ofer Zeitouni. *An Introduction to Random Matrices.* Cambridge University Press, Cambridge, UK, 2010.

[ALMT14] Dennis Amelunxen, Martin Lotz, Michael B McCoy, and Joel A Tropp. Living on the edge: Phase transitions in convex programs with random data. *Information and Inference: A Journal of the IMA*, 3(3):224–294, 2014.

[BBC08] Dominique Bakry, Franck Barth, and Patrick Cattiaux. A simple proof of the Poincaré inequality for a large class of probability measures including the log-concave case. *Electronic Communications in Probability*, 13, 02 2008.

[BBEKY13] Derek Bean, Peter J. Bickel, Noureddine El Karoui, and Bin Yu. Optimal M-estimation in high-dimensional regression. *Proceedings of the National Academy of Sciences of the United States of America*, 110(36):14563–8, 9 2013.

[BDMK16] Jean Barbier, Mohamad Dia, Nicolas Macris, and Florent Krzakala. The mutual information in random linear estimation. In *2016 54th Annual Allerton Conference on Communication, Control, and Computing (Allerton)*, pages 625–632, 2016.

[BF81] Peter J. Bickel and David A. Freedman. Some asymptotic theory for the bootstrap. *The Annals of Statistics*, 9(6):1196–1217, 11 1981.

[Bil12] Patrick Billingsley. *Probability and Measure.* John Wiley & Sons, Inc., Hoboken, New Jersey, anniversary edition, 2012.

[BIPW10] Khanh Do Ba, Piotr Indyk, Eric Price, and David P. Woodruff. Lower bounds for sparse recovery. In *Proceedings of the twenty-first annual ACM-SIAM symposium on Discrete Algorithms*, pages 1190–1197. SIAM, 2010.

[BKM$^+$19] Jean Barbier, Florent Krzakala, Nicolas Macris, Léo Miolane, and Lenka Zdeborová. Optimal errors and phase transitions in high-dimensional generalized linear models. *Proceedings of the National Academy of Sciences*, 116(12):5451–5460, 2019.

[BKRS21] Zhiqi Bu, Jason M. Klusowski, Cynthia Rush, and Weijie J. Su. Algorithmic analysis and statistical estimation of slope via approximate message passing. *IEEE Transactions on Information Theory*, 67(1):506–537, 2021.

[BLM15] Mohsen Bayati, Marc Lelarge, and Andrea Montanari. Universality in polytope phase transitions and message passing algorithms. *The Annals of Applied Probability*, 25(2):753–822, 2015.

[BLM16] Stèphane Boucheron, Gàbor Lugosi, and Pascal Massart. *Concentration Inequalities: A Nonasymptotic Theory of Independence.* Oxford University Press, New York, NY, 2016.

[BLT18] Pierre C. Bellec, Guillaume Lecué, and Alexandre B. Tsybakov. Slope meets Lasso: Improved oracle bounds and optimality. *The Annals of Statistics*, 46(6B):3603 – 3642, 2018.

[BM11] Mohsen Bayati and Andrea Montanari. The dynamics of message passing on dense graphs, with applications to compressed sensing. *IEEE Trans. on Inform. Theory*, 57:764–785, 2011.

[BM12] Mohsen Bayati and Andrea Montanari. The LASSO risk for gaussian matrices. *IEEE Trans. on Inform. Theory*, 58:1997–2017, 2012.

[BMDK17] Jean Barbier, Nicolas Macris, Mohamad Dia, and Florent Krzakala. Mutual information and optimality of approximate message-passing in random linear estimation. *IEEE Transactions on Information Theory*, PP, 01 2017.

[BMN19] Raphaël Berthier, Andrea Montanari, and Phan-Minh Nguyen. State evolution for approximate message passing with non-separable functions. *Information and Inference: A Journal of the IMA*, 9(1):33–79, 01 2019.

[BMR21] Jess Banks, Sidhanth Mohanty, and Prasad Raghavendra. Local statistics, semidefinite programming, and community detection. In *Proceedings of the 2021 ACM-SIAM Symposium on Discrete Algorithms (SODA)*, pages 1298–1316. SIAM, 2021.

[Bol14] Erwin Bolthausen. An iterative construction of solutions of the TAP equations for the Sherrington–Kirkpatrick model. *Communications in Mathematical Physics*, 325(1):333–366, 2014.

[BPW18] Afonso S. Bandeira, Amelia Perry, and Alexander S. Wein. Notes on computational-to-statistical gaps: predictions using statistical physics. arXiv:1803.11132, 2018.

[BR08] Howard D. Bondell and Brian J. Reich. Simultaneous regression shrinkage, variable selection, and supervised clustering of predictors with OSCAR. *Biometrics*, 2008.

[Bro86] Lawrence D. Brown. *Fundamentals of Statistical Exponential Families with Applications to Statistical Decision Theory*. Institute of Mathematical Statistics, Hayward, CA, 1986.

[BRT09] Petet J. Bickel, Yacov Ritov, and Alexander B. Tsybakov. Simultaneous analysis of Lasso and Dantzig selector. *Amer. J. of Mathematics*, 37:1705–1732, 2009.

[BvdBS+15] Małgorzata Bogdan, Ewout van den Berg, Chiara Sabatti, Weijie Su, and Emmanuel Candès. SLOPE—adaptive variable selection via convex optimization. *The Annals of Applied Statistics*, 9(3):1103–1140, 9 2015.

[CD95] S.S. Chen and David Donoho. Examples of basis pursuit. In *Proceedings of Wavelet Applications in Signal and Image Processing III*, San Diego, CA, 1995.

[Cel21] Michael Celentano. Approximate separability of symmetrically penalized least squares in high dimensions: characterization and consequences. *Information and Inference: A Journal of the IMA*, 01 2021. iaaa037.

[CM21] Michael Celentano and Andrea Montanari. Supplement to "Fundamental barriers to high-dimensional regression with convex penalties.". 2021.

[CRPW12] Venkat Chandrasekaran, Benjamin Recht, Pablo A. Parrilo, and Alan S. Willsky. The convex geometry of linear inverse problems. *Foundations of Computational Mathematics*, 12(6):805–849, 12 2012.

[CT05] Emmanuel Candés and Terence Tao. Decoding by linear programming. *IEEE Trans. on Inform. Theory*, 51:4203–4215, 2005.

[CT07] Emmanuel Candés and Terence Tao. The Dantzig selector: statistical estimation when p is much larger than n. *Annals of Statistics*, 35:2313–2351, 2007.

[DI17] Gamarnik David and Zadik Ilias. High dimensional regression with binary coefficients. Estimating squared error and a phase transtition. In Satyen Kale and Ohad Shamir, editors, *Proceedings of the 2017 Conference on Learning Theory*, volume 65 of *Proceedings of Machine Learning Research*, pages 948–953. PMLR, 07–10 Jul 2017.

[DM16] David Donoho and Andrea Montanari. High dimensional robust M-estimation: asymptotic variance via approximate message passing. *Probability Theory and Related Fields*, 166(3-4):935–969, 12 2016.

[DMM09] David Donoho, Arian Maleki, and Andrea Montanari. Message passing algorithms for compressed sensing. *Proceedings of the National Academy of Sciences*, 106:18914–18919, 2009.

[DMM10] David Donoho, Arian Maleki, and Andrea Montanari. Message passing algorithms for compressed sensing: I. motivation and construction. In *2010 IEEE information theory workshop on information theory (ITW 2010, Cairo)*, pages 1–5. IEEE, 2010.

[Dur10] Rick Durrett. *Probability: Theory and Examples*. Cambridge University Press, New York, NY, fourth edition, 2010.

[DYSV11] Dongning Guo, Yihong Wu, Shlomo Shamai, and Sergio Verdú. Estimation in Gaussian noise: Properties of the minimum mean-square error. *IEEE Transactions on Information Theory*, 57(4):2371–2385, 4 2011.

[Efr11] Bradley Efron. Tweedie's formula and selection bias. *Journal of the American Statistical Association*, 106(496):1602–1614, 12 2011.

[EG15] Lawrence C. Evans and Ronald F. Gariepy. *Measure Theory and Fine Properties of Functions*. CRC Press, Taylor & Francis Group, Boca Raton, FL, revised edition, 2015.

[EK13] Noureddine El Karoui. Asymptotic behavior of unregularized and ridge-regularized high-dimensional robust regression estimators: rigorous results. arXiv:1311.2445, 2013.

[EK18] Noureddine El Karoui. On the impact of predictor geometry on the performance on high-dimensional ridge-regularized generalized robust regression estimators. *Probability Theory and Related Fields*, 170(1-2):95–175, 2018.

[EKBB+13] Noureddine El Karoui, Derek Bean, Peter J. Bickel, Chinghway Lim, and Bin Yu. On robust regression with high-dimensional predictors. *Proceedings of the National Academy of Sciences of the United States of America*, 110(36):14557–62, 9 2013.

[FN14] Mario A. T. Figueiredo and Robert D. Nowak. Sparse estimation with strongly correlated variables using ordered weighted L1 regularization. arXiv:1409.4005, 2014.

[Gar85] C.W. Gardiner. *Handbook of Stochastic Methods for Physics, Chemistry and the Natural Science*. Springer-Verlag, Berlin, Germany, second edition, 1985.

[GS84] Clark R. Givens and Rae Michael Shortt. A class of wasserstein metrics for probability distributions. *The Michigan Mathematical Journal*, 31(2):231–240, 1984.

[GZ17] David Gamarnik and Ilias Zadik. Sparse high-dimensional linear regression. Algorithmic barriers and a local search algorithm. arXiv:1711.04952, 2017.

[JM13] Adel Javanmard and Andrea Montanari. State evolution for general approximate message passing algorithms, with applications to spatial coupling. *Information and Inference: A Journal of the IMA*, 2(2):115–144, 2013.

[KM11] Satish Babu Korada and Andrea Montanari. Applications of the Lindeberg principle in communications and statistical learning. *IEEE transactions on information theory*, 57(4):2440–2450, 2011.

[Led99] Michel Ledoux. Concentration of measure and logarithmic Sobolev inequalities. In Jacques Azéma, Michel Émery, Michel Ledoux, and Marc Yor, editors, *Séminaire de Probabilités XXXIII*, pages 120–216, Berlin, Heidelberg, 1999. Springer Berlin Heidelberg.

[LM19] Marc Lelarge and Léo Miolane. Fundamental limits of symmetric low-rank matrix estimation. *Probability Theory and Related Fields*, 173(3):859–929, 2019.

[LR05] Erich L. Lehmann and Joseph P. Romano. *Testing Statistical Hypotheses*. Springer-Verlag New York, New York, NY, third edition, 2005.

[MKL+20] Francesca Mignacco, Florent Krzakala, Yue Lu, Pierfrancesco Urbani, and Lenka Zdeborová. The role of regularization in classification of high-dimensional noisy Gaussian mixture. In Hal Daumé III and Aarti Singh, editors, *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 6874–6883. PMLR, 13–18 Jul 2020.

[MM09] M. Mézard and A. Montanari. *Information, Physics, and Computation*. Oxford Graduate Texts. OUP Oxford, 2009.

[MM18] Léo Miolane and Andrea Montanari. The distribution of the Lasso: Uniform control over sparse balls and adaptive parameter tuning. arXiv:1811.01212, 2018.

[Moi77] Edwin E. Moise. *Geometric Topology in Dimensions 2 and 3*. Springer-Verlag, Flushing, N.Y., 1st edition, 1977.

[Mor65] Jean-Jacques Moreau. Proximité et Dualité dans un Espace Hilbertien. 93:278–299, 1965.

[MXM19] Junjie Ma, Ji Xu, and Arian Maleki. Optimization-based AMP for phase retrieval: The impact of initialization and ℓ_2 regularization. *IEEE Transactions on Information Theory*, 65(6):3600–3629, 2019.

[OT18] Samet Oymak and Joel A Tropp. Universality laws for randomized dimension reduction, with applications. *Information and Inference: A Journal of the IMA*, 7:753–822, 2018.

[PB13] Neal Parikh and Stephen Boyd. Proximal algorithms. *Foundations and Trends in Optimization*, 1(3):123–231, 2013.

[Ran11] Sundeep Rangan. Generalized approximate message passing for estimation with random linear mixing. In *Information Theory Proceedings (ISIT), 2011 IEEE International Symposium on*, pages 2168–2172. IEEE, 2011.

[Roc97] R. Tyrrell Rockafellar. *Convex Analysis*. Princeton University Press, Princeton, NJ, 1997.

[RP16] Galen Reeves and Henry D. Pfister. The replica-symmetric prediction for compressed sensing with gaussian matrices is exact. In *Information Theory (ISIT), 2016 IEEE International Symposium on*, pages 665–669. IEEE, 2016.

[RSF14] Sundeep Rangan, Philip Schniter, and Alyson K. Fletcher. On the convergence of approximate message passing with arbitrary matrices. In *Information Theory Proceedings (ISIT), 2014 IEEE International Symposium on*, pages 236–240. IEEE, 2014.

[RSF17] Sundeep Rangan, Philip Schniter, and Alyson K. Fletcher. Vector approximate message passing. In *Information Theory Proceedings (ISIT), 2017 IEEE International Symposium on*, pages 1588–1592. IEEE, 2017.

[San15] Filippo Santambrogio. *Optimal Transport for Applied Mathematicians*. Springer International Publishing Switzerland, New York, 2015.

[SC16] Weijie Su and Emmanuel Candès. SLOPE is adaptive to unknown sparsity and asymptotically minimax. *The Annals of Statistics*, 44(3):1038–1068, 6 2016.

[SC19] Pragya Sur and Emmanuel Candès. A modern maximum-likelihood theory for high-dimensional logistic regression. *Proceedings of the National Academy of Sciences*, 116(29):14516–14525, 2019.

[SR15] Philip Schniter and Sundeep Rangan. Compressive phase retrieval via generalized approximate message passing. *IEEE Transactions on Signal Processing*, 63(4):1043–1055, 2015.

[Ste81] Charles M. Stein. Estimation of the mean of a multivariate normal distribution. *The Annals of Statistics*, 9(6):1135–1151, 11 1981.

[Sto10] Mihailo Stojnic. Recovery thresholds for ℓ_1 optimization in binary compressed sensing. In *Information Theory Proceedings (ISIT), 2010 IEEE International Symposium on*, pages 1593–1597. IEEE, 2010.

[Sto13] Mihailo Stojnic. A framework to characterize performance of Lasso algorithms. arXiv:1303.7291, 2013.

[SW14] Adrien Saumard and Jon A. Wellner. Log-concavity and strong log-concavity: A review. *Statistics Surveys*, 8(0):45–114, 2014.

[TAH18] Christos Thrampoulidis, Ehsan Abbasi, and Babak Hassibi. Precise error analysis of regularized M-estimators in high dimensions. *IEEE Transactions on Information Theory*, 64(8):5592–5628, 2018.

[Tib96] Rob Tibshirani. Regression shrinkage and selection with the Lasso. *J. Royal. Statist. Soc B*, 58:267–288, 1996.

[TOH15] Christos Thrampoulidis, Samet Oymak, and Babak Hassibi. Regularized linear regression: A precise analysis of the estimation error. In *Conference on Learning Theory*, pages 1683–1709, 2015.

[TPT20] Hossein Taheri, Ramtin Pedarsani, and Christos Thrampoulidis. Sharp asymptotics and optimal performance for inference in binary models. In Silvia Chiappa and Roberto Calandra, editors, *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics*, volume 108 of *Proceedings of Machine Learning Research*, pages 3739–3749. PMLR, 26–28 Aug 2020.

[TPT21] Hossein Taheri, Ramtin Pedarsani, and Christos Thrampoulidis. Fundamental limits of ridge-regularized empirical risk minimization in high dimensions. In Arindam Banerjee and Kenji Fukumizu, editors, *Proceedings of The 24th International Conference on Artificial Intelligence and Statistics*, volume 130 of *Proceedings of Machine Learning Research*, pages 2773–2781. PMLR, 13–15 Apr 2021.

[vdGB09] Sara A. van de Geer and Peter Bühlmann. On the conditions used to prove oracle results for the Lasso. *Electron. J. Statist.*, 3:1360–1392, 2009.

[Ver12] Roman Vershynin. Introduction to the non-asymptotic analysis of random matrices. In Y. Eldar and G. Kutyniok, editors, *Compressed Sensing, Theory and Applications*, volume 23, chapter 5, pages 210–268. Cambridge University Press, 2012.

[Wai09] Martin J. Wainwright. Information-theoretic limits on sparsity recovery in the high-dimensional and noisy setting. *IEEE Transactions on Information Theory*, 55(12):5728–5741, 2009.

[ZK15] Lenka Zdeborová and Florent Krzakala. Statistical physics of inference: thresholds and algorithms. *Advances in Physics*, 65:453 – 552, 2015.

A Equivalence of lower bounds: proof of Proposition 2.2

In fact, for any finite p,

$$\mathsf{R}^{\mathsf{opt}}_{\mathsf{reg},\mathsf{cvx}}(\tau;\pi,p) = \inf_{\rho\in\mathcal{C}_1} \mathbb{E}_{\beta_0,z}[(\mathsf{prox}[\rho](\beta_0+\tau z)-\beta_0)^2],$$

both when $\mathcal{C}$ contains all lsc, proper, convex functions and when $\mathcal{C}$ contains all lsc, proper, convex functions. Here $\mathcal{C}_1$ is the set of all lsc, proper, convex functions on $\mathbb{R}$.

First, note that $\mathsf{R}^{\mathsf{opt}}_{\mathsf{reg},\mathsf{cvx}}(\tau;\pi,p) \le \inf_{\rho\in\mathcal{C}_1} \mathbb{E}_{\beta_0,z}[([\rho](\beta_0+\tau z)-\beta_0)^2]$ because $\mathbb{E}_{\beta_0,z}[(\mathsf{prox}[\rho](\beta_0+\tau z)-\beta_0)^2]$ is in fact the risk in the sequence model of dimension p of the procedure which uses separable penalty $\rho_p(\boldsymbol{x}) = \frac{1}{p}\sum_{j=1}^p \rho(\sqrt{p}x)$. Indeed, $\mathsf{prox}[\rho_p](\boldsymbol{y})_j = \mathsf{prox}[\rho](\sqrt{p}y_j)/\sqrt{p}$. (Note that ρ_p is separable and symmetric).

Now note that for any lsc, proper, convex $\rho_p : \mathbb{R}^p \to \mathbb{R}\cup\{\infty\}$, fixing $\boldsymbol{y}_{-j}$ the function $y_j \mapsto \mathsf{prox}[\rho](\boldsymbol{y})$ is 1-Lipschitz, whence in fact $y_j \mapsto \mathsf{prox}[\rho](\boldsymbol{y})_j$ is 1-Lipschitz. Further, by the firm non-expansiveness of the proximal operator (Eq. (O.3)), we also have that $y_j \mapsto \mathsf{prox}[\rho](\boldsymbol{y})_j$ is non-decreasing. By Fact 2.1 of [Cel21], the set $\{\mathsf{prox}[\rho]\}$ as ρ varies over $\mathcal{C}!$ is exactly the set of 1-Lipschitz and non-decreasing functions on $\mathbb{R}$. Thus, we get $\mathbb{E}[(\mathsf{prox}[\rho_p](\boldsymbol{\beta}_0+\tau\boldsymbol{z})_j-\beta_{0j})^2|\boldsymbol{y}_{-j}] \ge \mathbb{E}_{\beta_0,z}[(\mathsf{prox}[\rho](\beta_0+\tau z)-\beta_0)^2]/p$ almost surely. We conclude that

$$\mathbb{E}[\|\mathsf{prox}[\rho_p](\boldsymbol{\beta}_0+\tau\boldsymbol{z})-\beta\|^2] = \sum_{j=1}^p \mathbb{E}[\mathbb{E}[(\mathsf{prox}[\rho_p](\boldsymbol{\beta}_0+\tau\boldsymbol{z})_j-\beta_{0j})^2|\boldsymbol{y}_{-j}]] \ge \mathbb{E}_{\beta_0,z}[(\mathsf{prox}[\rho](\beta_0+\tau z)-\beta_0)^2].$$

Having established both directions of the inequality completes the proof of Proposition 2.2.

B Exact asymptotics for the oracle estimator

As discussed in Section 2.1, our proof of the convex lower bound (Theorem 1) leverages exact asymptotics of the estimation error of penalized least squares estimators. Because we cannot provide exact asymptotics under only the δ-bounded width assumption, we will define an *oracle estimator* which performs at least as well as the original estimator (1.2) and to which we can apply exact asymptotic results. For any $\gamma \ge 0$, the oracle estimator is

$$\widehat{\boldsymbol{\beta}}^{(\gamma)}_{\mathsf{orc}} \in \arg\min_{\boldsymbol{\beta}} \left\{ \frac{1}{n}\|\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta}\|^2 + \rho(\boldsymbol{\beta}) + \frac{\gamma}{2}\|\boldsymbol{\beta}-\boldsymbol{\beta}_0\|^2 \right\}. \tag{B.1}$$

That is, we use the perturbed penalty

$$\rho^{(\gamma)}(\boldsymbol{\beta}) := \rho(\boldsymbol{\beta}) + \frac{\gamma}{2}\|\boldsymbol{\beta}-\boldsymbol{\beta}_0\|^2, \tag{B.2}$$

which includes a term which shrinks the estimate towards the true value $\boldsymbol{\beta}_0$. We remark that for $\gamma > 0$, *(i)* using this penalty in practice would require knowledge of the true parameter, so it cannot be implemented by the statistician, and *(ii)* because of its dependence on $\boldsymbol{\beta}_0$, the penalty defining the oracle estimator is itself random under the RSN assumption.

Previous work (e.g., [EK13]) has considered the addition of a small strongly-convex penalty in high-dimensional regression to permit rigorous exact asymptotics. The oracle term we add also serves this purpose, but is tailored to our goal of establishing estimation error lower bounds. Indeed, the oracle estimator performs at least as well as the original estimator for every realization of the data.

Lemma B.1. *For $\rho : \mathbb{R}^p \to \mathbb{R} \cup \{\infty\}$ an lsc, proper, convex function, $\boldsymbol{\beta}_0 \in \mathbb{R}^p$, $\boldsymbol{w} \in \mathbb{R}^n$, $\gamma > 0$, and all realizations of the design matrix $\boldsymbol{X} \in \mathbb{R}^{p\times n}$ and parameter $\boldsymbol{\beta}_0$, we have*

$$\|\widehat{\boldsymbol{\beta}}_{\mathsf{orc}}^{(\gamma)} - \boldsymbol{\beta}_0\|^2 \le \|\widehat{\boldsymbol{\beta}}_{\mathsf{cvx}} - \boldsymbol{\beta}_0\|^2, \tag{B.3}$$

for any $\boldsymbol{\beta}_{\mathsf{orc}}^{(\gamma)}$ satisfying (B.1). That is, the ℓ_2-loss of $\widehat{\boldsymbol{\beta}}_{\mathsf{orc}}^{(\gamma)}$ is no larger than the ℓ_2-loss of $\widehat{\boldsymbol{\beta}}_{\mathsf{cvx}}$.

Proof of Lemma B.1. If the minimizing set of (1.2) is empty, then the right-hand side of (B.3) is ∞ by convention, and there is nothing to show. Thus, assume $\widehat{\boldsymbol{\beta}}_{\mathsf{cvx}}$ satisfies (1.2). For any $\boldsymbol{\beta} \in \mathbb{R}^p$ with $\|\boldsymbol{\beta} - \boldsymbol{\beta}_0\| > \|\widehat{\boldsymbol{\beta}}_{\mathsf{cvx}} - \boldsymbol{\beta}_0\|$, we have

$$\begin{aligned}\frac{1}{n}\|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\|^2 + \rho(\boldsymbol{\beta}) + \frac{\gamma}{2}\|\boldsymbol{\beta} - \boldsymbol{\beta}_0\|^2 &> \frac{1}{n}\|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\|^2 + \rho(\boldsymbol{\beta}) + \frac{\gamma}{2}\|\widehat{\boldsymbol{\beta}}_{\mathsf{cvx}} - \boldsymbol{\beta}_0\|^2 \\ &\ge \frac{1}{n}\|\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}_{\mathsf{cvx}}\|^2 + \rho(\widehat{\boldsymbol{\beta}}_{\mathsf{cvx}}) + \frac{\gamma}{2}\|\widehat{\boldsymbol{\beta}}_{\mathsf{cvx}} - \boldsymbol{\beta}_0\|^2,\end{aligned}$$

where the second inequality follows from the definition of $\widehat{\boldsymbol{\beta}}_{\mathsf{cvx}}$ in (1.2). Thus, $\boldsymbol{\beta}$ cannot be a minimizer in (B.1)). Moreover, because $\rho^{(\gamma)}$ has strong-convexity parameter $\gamma > 0$, the minimizing set of (B.1)) is non-empty. Thus, we have (B.3). □

The exact asymptotic characterization of the oracle estimator requires several definitions. Denote by $\mathcal{T}_p$ a pair (π, ρ_p) where $\pi \in \mathcal{P}_2(\mathbb{R})$ and $\rho_p : \mathbb{R}^p \to \mathbb{R}\cup\{\infty\}$ is an lsc, proper, convex function. For any $\tau, \lambda \ge 0$ and $\boldsymbol{T} \in S_+^2$, define

$$\mathsf{R}_{\mathsf{reg,cvx}}(\tau, \lambda, \mathcal{T}_p) := \mathbb{E}_{\boldsymbol{\beta}_0, \boldsymbol{z}}\left[\|\mathsf{prox}[\lambda\rho_p](\boldsymbol{\beta}_0 + \tau\boldsymbol{z}) - \boldsymbol{\beta}_0\|^2\right], \tag{B.4a}$$

$$\mathsf{W}_{\mathsf{reg,cvx}}(\tau, \lambda, \mathcal{T}_p) := \frac{1}{\tau}\mathbb{E}_{\boldsymbol{\beta}_0, \boldsymbol{z}}\left[\langle \boldsymbol{z}, \mathsf{prox}[\lambda\rho_p](\boldsymbol{\beta}_0 + \tau\boldsymbol{z})\rangle\right], \tag{B.4b}$$

$$\mathsf{K}_{\mathsf{reg,cvx}}(\boldsymbol{T}, \lambda, \mathcal{T}_p) := \mathbb{E}_{\boldsymbol{\beta}_0, \boldsymbol{z}_1, \boldsymbol{z}_2}\left[\langle \mathsf{prox}\left[\lambda\rho_p\right](\boldsymbol{\beta}_0 + \boldsymbol{z}_1) - \boldsymbol{\beta}_0, \mathsf{prox}\left[\lambda\rho_p\right](\boldsymbol{\beta}_0 + \boldsymbol{z}_2) - \boldsymbol{\beta}_0\rangle\right], \tag{B.4c}$$

where $\beta_{0j} \overset{\text{iid}}{\sim} \pi/\sqrt{p}$, $\boldsymbol{z} \sim \mathsf{N}(\boldsymbol{0}, \boldsymbol{I}_p/p)$, and $(\boldsymbol{z}_1, \boldsymbol{z}_2) \sim \mathsf{N}(\boldsymbol{0}, \boldsymbol{T} \otimes \boldsymbol{I}_p/p)$. Consider a sequence of penalties $\{\rho_p : \mathbb{R}^p \to \mathbb{R}\cup\{\infty\}\}$. Let $\mathcal{T} = (\pi, \{\rho_p\})$. Define

$$\begin{aligned}\mathsf{R}^{\infty}_{\mathsf{reg,cvx}}(\tau, \lambda, \mathcal{T}) &:= \lim_{p\to\infty} \mathsf{R}_{\mathsf{reg,cvx}}(\tau, \lambda, \mathcal{T}_p), \\ \mathsf{W}^{\infty}_{\mathsf{reg,cvx}}(\tau, \lambda, \mathcal{T}) &:= \lim_{p\to\infty} \mathsf{W}_{\mathsf{reg,cvx}}(\tau, \lambda, \mathcal{T}_p), \\ \mathsf{K}^{\infty}_{\mathsf{reg,cvx}}(\boldsymbol{T}, \lambda, \mathcal{T}) &:= \lim_{p\to\infty} \mathsf{K}_{\mathsf{reg,cvx}}(\boldsymbol{T}, \lambda, \mathcal{T}_p),\end{aligned} \tag{B.5}$$

whenever these limits exist. Here $\mathcal{T}_p$ is related to $\mathcal{T}$ in the obvious way. Finally, denote

$$\begin{aligned}\tau_{\mathsf{orc}} = \tau_{\mathsf{orc}}(\tau, \lambda, \gamma) = \frac{\tau}{\lambda\gamma + 1}, \quad \lambda_{\mathsf{orc}} = \lambda_{\mathsf{orc}}(\lambda, \gamma) = \frac{\lambda}{\lambda\gamma + 1}, \\ \boldsymbol{T}_{\mathsf{orc}} = \boldsymbol{T}_{\mathsf{orc}}(\boldsymbol{T}, \lambda, \gamma) = \frac{\boldsymbol{T}}{(\lambda\gamma + 1)^2}.\end{aligned} \tag{B.6}$$

The exact asymptotic characterization is given by a solution (τ, λ) to the following system of equations.

$$\delta\tau^2 - \sigma^2 = \mathsf{R}^{\infty}_{\mathsf{reg,cvx}}(\tau_{\mathsf{orc}}, \lambda_{\mathsf{orc}}, \mathcal{T}), \tag{B.7a}$$

$$2\lambda\left(1 - \frac{1}{\delta(\lambda\gamma + 1)}\mathsf{W}^{\infty}_{\mathsf{reg,cvx}}(\tau_{\mathsf{orc}}, \lambda_{\mathsf{orc}}, \mathcal{T})\right) = 1. \tag{B.7b}$$

The following notion will be needed.

Definition B.2 (Strong stationarity). *For any $\tau \geq 0$, $\lambda > 0$, $\boldsymbol{T} \in S_+^2$, and $\gamma \geq 0$, we denote τ_{orc}, λ_{orc}, and $\boldsymbol{T}_{\mathsf{orc}}$ as in (B.6). We say the quintuplet $\tau, \lambda, \gamma, \delta, \mathcal{T}$ is* strongly stationary *if at λ_{orc} and at all $\tau' \geq 0$, $\boldsymbol{T}' \succeq \boldsymbol{0}$, the limits (B.5) exist, and at τ, λ, γ, the equations (B.7) are satisfied.*

We are ready to provide our exact characterization of oracle estimators.

Proposition B.3. *Consider $\pi \in \mathcal{P}_\infty(\mathbb{R})$, $\delta \in (0, \infty)$, and $\sigma \geq 0$. Consider a sequence of lsc, proper, convex functions $\rho_p : \mathbb{R}^p \to \mathbb{R} \cup \{\infty\}$. Let $\mathcal{T} = (\pi, \{\rho_p\})$. Assume $\tau, \lambda, \gamma \geq 0$ are such that $\tau, \lambda, \gamma, \delta, \mathcal{T}$ is strongly stationary. For each p, let $\widehat{\boldsymbol{\beta}}_{\mathsf{cvx}}^{(\gamma)}$ be a solution to (1.2). If either $\delta > 1$, $\gamma > 0$, or the ρ_p have uniform strong convexity parameter $\kappa > 0$, then*

(i) The solution to (B.1)) exists and is unique for all n large enough:

$$\mathbb{P}_{\boldsymbol{X}}(\textit{solution to (B.1)) exists and is unique}) = 1 \textit{ eventually.} \tag{B.8}$$

(ii) Under RSN assumption the loss obeys

$$\|\widehat{\boldsymbol{\beta}}_{\mathsf{cvx}}^{(\gamma)} - \boldsymbol{\beta}_0\|^2 \xrightarrow{\mathrm{p}} \mathsf{R}^{\infty}_{\mathsf{reg},\mathsf{cvx}}(\tau_{\mathsf{orc}}, \lambda_{\mathsf{orc}}, \mathcal{T}) = \delta\tau^2 - \sigma^2. \tag{B.9}$$

If the penalties are symmetric, then (B.9) holds also under the DSN assumption.

(iii) Consider the case that $\gamma = 0$ and either $\delta > 1$ or ρ_p are uniformly strongly convex. Consider any sequence of functions $\varphi_p : (\mathbb{R}^p)^2 \to \mathbb{R}$ which are uniformly pseudo-Lipschitz of order k for some k. Under the RSN assumption

$$\varphi_p\left(\boldsymbol{\beta}_0, \widehat{\boldsymbol{\beta}}_{\mathsf{cvx}} + 2\lambda \frac{\boldsymbol{X}^{\mathsf{T}}(\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}_{\mathsf{cvx}})}{n}\right) \overset{\mathrm{p}}{\simeq} \mathbb{E}_{\boldsymbol{z}}\left[\varphi_p(\boldsymbol{\beta}_0, \boldsymbol{\beta}_0 + \tau\boldsymbol{z})\right]. \tag{B.10}$$

If the penalties are symmetric, then (B.9) holds under the DSN assumption.

The proof of Proposition B.3 is provided in Appendix D.

Note that although $\mathsf{K}_{\mathsf{reg},\mathsf{cvx}}(\boldsymbol{T}, \lambda, \mathcal{T}_p)$ and $\mathsf{K}^{\infty}_{\mathsf{reg},\mathsf{cvx}}(\boldsymbol{T}, \lambda, \mathcal{T})$ do not appear in the equations (B.7), the existence of the limit (B.5) will play an essential role in the proof of Proposition B.3. In particular, it will allow us to control the convergence of the iterates of a certain AMP algorithm to the convex M-estimator, which will be important for establishing its characterization (see Section D for details).

In addition to its use in establishing the convex lower bound, Proposition B.3 will play a role in establishing the tightness of the convex lower bound under log-concavity assumptions or when $\delta > 1$. Proposition B.3(iii) plays a role in our consideration of non-quadratic losses and post-processing in Section 5.

Our proof follows closely the proof of the similar result in Theorem 1.2 of [DM16]. The authors of [DM16] establish an asymptotic characterization of the loss of M-estimators of the form $\widehat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} \sum_{i=1}^n \rho\left(y_i - [\boldsymbol{X}\boldsymbol{\beta}]_i\right)$ where ρ is strongly convex and $\delta > 1$. Our Proposition B.3 differs from their theorem in that (i) we impose a penalty on the parameters rather than an arbitrary penalty on the residuals, (ii) we permit non-separable penalties, and (iii) we consider $\delta \leq 1$. Nevertheless, our argument follows almost exactly theirs (see Appendix D). In handling non-separable penalties, we rely on recent results on approximate message passing algorithms with non-separable denoisers [BMN19], which the authors of [DM16] did not have access to.

A result similar to Proposition B.3 was also proved in [TAH18] using Gaussian comparison inequalities. The conditions in [TAH18] are not directly comparable to the ones of Proposition B.3. We prefer proving an independent statement, since checking the conditions of the general theorem in [TAH18] is non-trivial. As an advantage, Proposition B.3 gives access –via Eq. (B.10)– to the empirical distribution of the entries of $\widehat{\boldsymbol{\beta}}_{\mathsf{cvx}}$ and $\widehat{\boldsymbol{\beta}}_{\mathsf{cvx}} + 2\lambda\frac{\boldsymbol{X}^{\mathsf{T}}(\boldsymbol{y}-\boldsymbol{X}\widehat{\boldsymbol{\beta}}_{\mathsf{cvx}})}{n}$, which is not provided by [TAH18]. As stated above, this plays a role in our consideration of non-quadratic losses and post-processing.

Finally, Proposition explicitly describes the impact of the oracle term.

C Regularity lemmas

This appendix provides several lemmas controlling the regularity of various objects appearing in the exact characterization of Proposition B.3. These will be required in both the proof of Proposition B.3 and in its applications.

First, for $\tau > 0$,

$$\mathsf{W}_{\mathsf{reg},\mathsf{cvx}}(\tau,\lambda,\mathcal{T}_p) = \frac{1}{\tau}\mathbb{E}_{\boldsymbol{\beta}_0,\boldsymbol{z}}[\langle \boldsymbol{z}, \mathsf{prox}[\lambda\rho_p](\boldsymbol{\beta}_0+\tau\boldsymbol{z})\rangle] = \frac{1}{p}\mathbb{E}_{\boldsymbol{\beta}_0,\boldsymbol{z}}[\operatorname{div}\mathsf{prox}[\lambda\rho_p](\boldsymbol{\beta}_0+\tau\boldsymbol{z})] \le 1, \quad \text{(C.1)}$$

where in the first equality we have used the definition (B.4b), in the second equality we have used (O.11), and in the inequality we have used (O.7). Taking limits, we have (using (O.10))

$$\mathsf{W}^{\infty}_{\mathsf{reg},\mathsf{cvx}}(\tau,\lambda,\mathcal{T}) \le 1, \quad \text{and} \quad \mathsf{W}_{\mathsf{reg},\mathsf{cvx}}(\tau,\lambda,\mathcal{T}_p) \ge 0, \quad \text{and} \quad \mathsf{W}^{\infty}_{\mathsf{reg},\mathsf{cvx}}(\tau,\lambda,\mathcal{T}) \ge 0, \quad \text{(C.2)}$$

whenever these are defined.

Next, in this and other appendices we will sometimes need the following basic algebraic inequalities which hold for any vectors $\boldsymbol{a},\boldsymbol{a}',\boldsymbol{b},\boldsymbol{b}' \in \mathbb{R}^p$ and are straightforward to verify.

$$|\langle \boldsymbol{a},\boldsymbol{b}\rangle - \langle \boldsymbol{a}',\boldsymbol{b}'\rangle| \le 2\underbrace{\max\{\|\boldsymbol{a}\|,\|\boldsymbol{a}'\|,\|\boldsymbol{b}\|,\|\boldsymbol{b}'\|\}}_{(*)}\underbrace{(\|\boldsymbol{a}-\boldsymbol{a}'\| \vee \|\boldsymbol{b}-\boldsymbol{b}'\|)}_{(**)}, \quad \text{(C.3)}$$

$$\big|\|\boldsymbol{a}\|^2 - \|\boldsymbol{b}\|^2\big| \le 2\underbrace{(\|\boldsymbol{a}\| \vee \|\boldsymbol{b}\|)}_{(*)}\underbrace{\|\boldsymbol{a}-\boldsymbol{b}\|}_{(**)}. \quad \text{(C.4)}$$

We label the terms on the right-hand sides with $(*)$ and $(**)$ to facilitate future reference. The inequalities are straightforward to verify. In fact, (C.4) is a special case of (C.3).

We say a sequence of functions $\{\rho_p\} \in \mathcal{C}$ *does not shrink towards infinity* if

$$\sup_p \|\mathsf{prox}[\rho_p](\boldsymbol{0})\| < \infty. \quad \text{(C.5)}$$

We define the collection of penalty sequences which do not shrink towards infinity

$$\mathcal{B} = \Big\{\{\rho_p\} \in \mathcal{C} \;\Big|\; \text{(C.5) holds}\Big\}. \quad \text{(C.6)}$$

Finally, we provide a series of lemmas which we will need in later sections.

Lemma C.1. *Consider $\{\rho_p\} \in \mathcal{B}$ (see (C.6)). Then for any fixed $\tau, \lambda \geq 0$, the functions*

$$\varphi_{\mathsf{R}}^{(p)}(\boldsymbol{\beta}_0, \boldsymbol{z}) = \|\mathsf{prox}[\lambda\rho_p](\boldsymbol{\beta}_0 + \tau\boldsymbol{z}) - \boldsymbol{\beta}_0\|^2,$$
$$\varphi_{\mathsf{W}}^{(p)}(\boldsymbol{\beta}_0, \boldsymbol{z}) = \frac{1}{\tau}\left\langle \boldsymbol{z}, \mathsf{prox}[\lambda\rho_p](\boldsymbol{\beta}_0 + \tau\boldsymbol{z})\right\rangle,$$
$$\varphi_{\mathsf{K}}^{(p)}(\boldsymbol{\beta}_0, \boldsymbol{z}_1, \boldsymbol{z}_2) = \left\langle \mathsf{prox}\left[\lambda\rho_p\right](\boldsymbol{\beta}_0 + \boldsymbol{z}_1) - \boldsymbol{\beta}_0, \mathsf{prox}\left[\lambda\rho_p\right](\boldsymbol{\beta}_0 + \boldsymbol{z}_2) - \boldsymbol{\beta}_0\right\rangle,$$

are uniformly pseudo-Lipschitz of order 2.

Proof of C.1. Let $M := \sup_p \|\mathsf{prox}[\rho_p](\mathbf{0})\|$. We have $M < \infty$ because $\{\rho_p\} \in \mathcal{B}$. By (O.5), we have

$$\begin{aligned}\|\mathsf{prox}[\lambda\rho_p](\mathbf{0})\| &\leq \|\mathsf{prox}[\rho_p](\mathbf{0})\| + \|\mathsf{prox}[\lambda\rho_p](\mathbf{0}) - \mathsf{prox}[\rho_p](\mathbf{0})\| \\ &\leq \|\mathsf{prox}[\rho_p](\mathbf{0})\| + \|\mathsf{prox}[\lambda\rho_p](\mathbf{0})\|\,|\lambda - 1| \leq (2M+1)\lambda.\end{aligned}$$

Thus, $\|\mathsf{prox}[\lambda\rho_p](\mathbf{0})\|$ is bounded over p. Further, by (O.4), the functions $(\boldsymbol{\beta}_0, \boldsymbol{z}) \mapsto \mathsf{prox}[\lambda\rho_p](\boldsymbol{\beta}_0 + \tau\boldsymbol{z}) - \boldsymbol{\beta}_0$ and $(\boldsymbol{\beta}_0, \boldsymbol{z}) \mapsto \mathsf{prox}[\lambda\rho_p](\boldsymbol{\beta}_0 + \tau\boldsymbol{z})$ are uniformly pseudo-Lipschitz of order 1. Further, the function $(\boldsymbol{\beta}_0, \boldsymbol{z}) \mapsto \frac{1}{\tau}\boldsymbol{z}$ is trivially uniformly pseudo-Lipschitz of order 1. Applying Lemma P.2, the Lemma follows. □

Lemma C.2. *There exists universal constant C such that the functions $\tau \mapsto \mathsf{R}^{\mathsf{opt}}_{\mathsf{seq},\mathsf{cvx}}(\tau; \pi, p)$ and $\tau \mapsto \mathsf{R}^{\mathsf{opt}}_{\mathsf{seq},\mathsf{cvx}}(\tau; \pi)$ defined in (2.7) and (2.8) satisfy for any $\tau, \tau' \geq 0$ (using f to denote each function)*

$$|f(\tau') - f(\tau)| \leq C(1 + |\tau' - \tau| + f(\tau))|\tau' - \tau|. \tag{C.7}$$

This makes sense even when $\mathcal{C}$ is such that $\mathsf{R}^{\mathsf{opt}}_{\mathsf{seq},\mathsf{cvx}}(\tau; \pi)$ is infinite (note, f is always non-negative).

In particular, if $\mathsf{R}^{\mathsf{opt}}_{\mathsf{seq},\mathsf{cvx}}(\tau; \pi)$ is finite anywhere, it is finite and continuous everywhere.

Proof of Lemma C.2. First, we develop a bound on

$$\left|\mathbb{E}_{\boldsymbol{\beta}_0,\boldsymbol{z}}\left[\|\mathsf{prox}[\rho](\boldsymbol{\beta}_0 + \tau\boldsymbol{z}) - \boldsymbol{\beta}_0\|^2\right] - \mathbb{E}_{\boldsymbol{\beta}_0,\boldsymbol{z}}[\|\mathsf{prox}[\rho](\boldsymbol{\beta}_0 + \tau'\boldsymbol{z}) - \boldsymbol{\beta}_0\|^2]\right|,$$

for $\rho : \mathbb{R}^p \to \mathbb{R}^p$ an lsc, proper, convex function. We apply Jensen's inequality to get

$$\begin{aligned}&\left|\mathbb{E}_{\boldsymbol{\beta}_0,\boldsymbol{z}}\left[\|\mathsf{prox}[\rho](\boldsymbol{\beta}_0 + \tau\boldsymbol{z}) - \boldsymbol{\beta}_0\|^2\right] - \mathbb{E}_{\boldsymbol{\beta}_0,\boldsymbol{z}}[\|\mathsf{prox}[\rho](\boldsymbol{\beta}_0 + \tau'\boldsymbol{z}) - \boldsymbol{\beta}_0\|^2]\right| \\ &\qquad \leq \mathbb{E}_{\boldsymbol{\beta}_0,\boldsymbol{z}}\left[\left|\|\mathsf{prox}[\rho](\boldsymbol{\beta}_0 + \tau\boldsymbol{z}) - \boldsymbol{\beta}_0\|^2 - \|\mathsf{prox}[\rho](\boldsymbol{\beta}_0 + \tau'\boldsymbol{z}) - \boldsymbol{\beta}_0\|^2\right|\right].\end{aligned}$$

We bound the integrand by applying (C.4):

$$\begin{aligned}&\left|\|\mathsf{prox}[\rho](\boldsymbol{\beta}_0 + \tau\boldsymbol{z}) - \boldsymbol{\beta}_0\|^2 - \|\mathsf{prox}[\rho](\boldsymbol{\beta}_0 + \tau'\boldsymbol{z}) - \boldsymbol{\beta}_0\|^2\right| \\ &\qquad \leq 2\underbrace{\left(\|\mathsf{prox}[\rho](\boldsymbol{\beta}_0 + \tau\boldsymbol{z}) - \boldsymbol{\beta}_0\| + \|\tau\boldsymbol{z} - \tau'\boldsymbol{z}\|\right)}_{\text{bound on } (*)}\underbrace{\|\tau\boldsymbol{z} - \tau'\boldsymbol{z}\|}_{\text{bound on } (**)} \\ &\qquad \leq 2\|\mathsf{prox}[\rho](\boldsymbol{\beta}_0 + \tau\boldsymbol{z}) - \boldsymbol{\beta}_0\|\|\boldsymbol{z}\||\tau - \tau'| + 2\|\boldsymbol{z}\|^2(\tau - \tau')^2 \\ &\qquad \leq \left(\|\mathsf{prox}[\rho](\boldsymbol{\beta}_0 + \tau\boldsymbol{z}) - \boldsymbol{\beta}_0\|^2 + \|\boldsymbol{z}\|^2\right)|\tau - \tau'| + 2\|\boldsymbol{z}\|^2(\tau - \tau')^2.\end{aligned}$$

Combining the previous two displays,

$$\begin{aligned}&\left|\mathbb{E}_{\boldsymbol{\beta}_0,\boldsymbol{z}}\left[\|\mathsf{prox}[\rho](\boldsymbol{\beta}_0 + \tau\boldsymbol{z}) - \boldsymbol{\beta}_0\|^2\right] - \mathbb{E}_{\boldsymbol{\beta}_0,\boldsymbol{z}}[\|\mathsf{prox}[\rho](\boldsymbol{\beta}_0 + \tau'\boldsymbol{z}) - \boldsymbol{\beta}_0\|^2]\right| \\ &\qquad \leq \left(\mathbb{E}_{\boldsymbol{\beta}_0,\boldsymbol{z}}\left[\|\mathsf{prox}[\rho](\boldsymbol{\beta}_0 + \tau\boldsymbol{z}) - \boldsymbol{\beta}_0\|^2\right] + 1\right)|\tau - \tau'| + 2(\tau - \tau')^2.\end{aligned}$$

Thus, $\tau \mapsto \mathbb{E}_{\boldsymbol{\beta}_0,\boldsymbol{z}}\left[\|\mathsf{prox}[\rho](\boldsymbol{\beta}_0+\tau\boldsymbol{z})-\boldsymbol{\beta}_0\|^2\right]$ satisfies (C.7). To prove (C.7) for $\mathsf{R}^{\mathsf{opt}}_{\mathsf{seq},\mathsf{cvx}}(\tau;\pi,p)$ and $\mathsf{R}^{\mathsf{opt}}_{\mathsf{seq},\mathsf{cvx}}(\tau;\pi)$, we use that the property (C.7) is preserved by taking point-wise infima of collections of functions, as well as limit infima (provided infinite limit infima are permitted, with (C.7) interpreted in the natural way in this case). The finiteness and continuity of $\mathsf{R}^{\mathsf{opt}}_{\mathsf{seq},\mathsf{cvx}}(\tau;\pi,p)$ and $\mathsf{R}^{\mathsf{opt}}_{\mathsf{seq},\mathsf{cvx}}(\tau;\pi)$ in the case that these are finite anywhere is then automatic. □

Lemma C.3. *Let $\pi \in \mathcal{P}_2(\mathbb{R})$. Let $\varphi_p : \mathbb{R}^p \to \mathbb{R}$ be a sequence of functions which is uniformly pseudo-Lipschitz of order 2. Then*

$$\varphi_p(\boldsymbol{\beta}_0) \overset{\text{as}}{\simeq} \mathbb{E}_{\boldsymbol{\beta}_0}[\varphi_p(\boldsymbol{\beta}_0)],$$

where for each p, $\beta_{0j} \overset{\text{iid}}{\sim} \pi/\sqrt{p}$, and the $\boldsymbol{\beta}_0$ are independent across p.

In particular, if $\{\rho_p\} \in \mathcal{B}$ and the limits (B.5) exist (with $\boldsymbol{z} \sim \mathsf{N}(0, \boldsymbol{I}_p/p)$), then with respect to the randomness in $\boldsymbol{\beta}_0$,

$$\begin{aligned}
\mathbb{E}_{\boldsymbol{z}}\left[\|\mathsf{prox}[\lambda\rho_p](\boldsymbol{\beta}_0+\tau\boldsymbol{z})-\boldsymbol{\beta}_0\|^2\right] &\overset{\text{as}}{\simeq} \mathsf{R}_{\mathsf{reg},\mathsf{cvx}}(\tau,\lambda,\mathcal{T}),\\
\frac{1}{\tau}\mathbb{E}_{\boldsymbol{z}}\left[\langle \boldsymbol{z}, \mathsf{prox}[\lambda\rho_p](\boldsymbol{\beta}_0+\tau\boldsymbol{z})\rangle\right] &\overset{\text{as}}{\simeq} \mathsf{W}_{\mathsf{reg},\mathsf{cvx}}(\tau,\lambda,\mathcal{T}),\\
\mathbb{E}_{\boldsymbol{z}_1,\boldsymbol{z}_2}\left[\langle \mathsf{prox}\left[\lambda\rho_p\right](\boldsymbol{\beta}_0+\boldsymbol{z}_1)-\boldsymbol{\beta}_0, \mathsf{prox}\left[\lambda\rho_p\right](\boldsymbol{\beta}_0+\boldsymbol{z}_2)-\boldsymbol{\beta}_0\rangle\right] &\overset{\text{as}}{\simeq} \mathsf{K}_{\mathsf{reg},\mathsf{cvx}}(\boldsymbol{T},\lambda,\mathcal{T}).
\end{aligned}$$

Proof. Throughout the proof, β_0 will denote a random variable drawn from π. Let $s_2(\pi)^2 = \mathbb{E}[\beta_0^2]$. By assumption, the restriction of φ_p to $\{\|\boldsymbol{\beta}_0\|_2^2 \le s_2(\pi)^2+1\}$ is L-Lipschitz for some L which does not depend on p. Let $\bar{\varphi}_p$ be an L-Lipschitz extension of $\varphi_p|_{\{\|\boldsymbol{\beta}_0\|_2^2 \le s_2(\pi)^2 \le s_2(\pi)^2+1\}}$ to all of $\mathbb{R}^p$; that is, it is L-Lipschitz and agrees with φ_p on $\{\|\boldsymbol{\beta}_0\|_2^2 \le s_2(\pi)^2+1\}$. For example, one can check that $\bar{\varphi}_p(\boldsymbol{x}) = \sup_{\|\boldsymbol{x}'\|\le R}\{\varphi_p(\boldsymbol{x}') - \|\boldsymbol{x}-\boldsymbol{x}'\|\}$ is a valid Lipschitz extension.

Fix $1 > \epsilon > 0$. Pick $M > 0$ such that $\mathbb{E}[|\beta_0|^2\mathbf{1}_{|\beta_0|>M}] < \epsilon^2$. For each p, let $\boldsymbol{h}$ have coordinates drawn iid from the Laplace distribution with scale parameter $1/\sqrt{p}$ (i.e., density $\frac{\sqrt{p}}{2}e^{-\sqrt{p}|x|}$), independent of $\boldsymbol{\beta}_0$ and across p. Define $\boldsymbol{\beta}_0^\epsilon = (|\beta_{0j}|\mathbf{1}_{\sqrt{p}|\beta_{0j}|\le M} + h_j)_{j\in[p]}$. Then $\sqrt{p}\beta_{0j}^\epsilon$ satisfies a Poincaré inequality (this follows by Corollary 1.6 of [BBC08]); that is, for any weakly differentiable f, $\mathsf{Var}(f(\sqrt{p}\beta_{0j}^\epsilon)) \le C\mathbb{E}[f'(\sqrt{p}\beta_{0j}^\epsilon)^2]$ for some constant C which does not depend on p (but may depend on ϵ, π, M). Then the product measure $\pi^{\otimes p}$ satisfies a Poincaré inequality with the same constant C: $\mathsf{Var}(\bar{\varphi}_p(\boldsymbol{\beta})) \le C\mathbb{E}[\|\nabla\varphi_p(\boldsymbol{\beta}_0)\|^2]/p$. Then, by Corollary 4.6 of Ledoux [Led99], $\bar{\varphi}_p$ has exponential concentration. In particular, there exists a constant c, which does not depend on p, (but may depend on ϵ, π, M, L) such that

$$\mathbb{P}\left(\left|\bar{\varphi}_p(\boldsymbol{\beta}_0) - \mathbb{E}[\bar{\varphi}_p(\boldsymbol{\beta}_0^\epsilon)]\right| > t\right) \le 2e^{-c\min(\sqrt{p}t, pt^2)}.$$

Taking $t \to 0$ after $p \to \infty$ and using the Borel-Cantelli lemma, we get that $\bar{\varphi}_p(\boldsymbol{\beta}_0^\epsilon) \overset{\text{as}}{\simeq} \mathbb{E}[\bar{\varphi}_p(\boldsymbol{\beta}_0^\epsilon)]$.

By the strong law of large numbers and the definition of $\boldsymbol{\beta}_0^\epsilon$, we have $\|\boldsymbol{\beta}_0 - \boldsymbol{\beta}_0^\epsilon\|^2 \overset{\text{as}}{\to} \mathbb{E}[(\beta_0 - \beta_{0j}\mathbf{1}_{|\beta_0|\le M} - \epsilon h)^2] < 2\epsilon^2$, where h has Laplace distribution with scale parameter 1 and is independent of β_0. Thus, almost surely we have that for large enough p, $|\bar{\varphi}_p(\boldsymbol{\beta}_0) - \bar{\varphi}_p(\boldsymbol{\beta}_0^\epsilon)| < \sqrt{2}L\epsilon$. Also by the strong law of large numbers, we have $\|\boldsymbol{\beta}_0^2\| \to s_2(\pi)^2$ and $\|\boldsymbol{\beta}_0^\epsilon\|^2 \overset{\text{as}}{\to} \mathbb{E}[(\sqrt{p}\beta_{0j}^\epsilon)^2] < s_2(\pi)^2+1$, where the inequality holds because $\epsilon < 1$. Thus, almost surely we have that for large enough p, $\varphi_p(\boldsymbol{\beta}_0) = \bar{\varphi}_p(\boldsymbol{\beta}_0)$ and $\varphi_p(\boldsymbol{\beta}_0^\epsilon) = \bar{\varphi}_p(\boldsymbol{\beta}_0^\epsilon)$. Combining the preceding observations, almost surely we have that

for large enough p, $|\varphi_p(\boldsymbol{\beta}_0) - \mathbb{E}[\bar{\varphi}_p(\boldsymbol{\beta}_0^\epsilon)]| < \sqrt{2}L\epsilon$. Finally, $\mathbb{E}[\|\boldsymbol{\beta}_0 - \boldsymbol{\beta}_0^\epsilon\|] \le \mathbb{E}[\|\boldsymbol{\beta}_0 - \boldsymbol{\beta}_0^\epsilon\|^2]^{1/2} < \sqrt{2}\epsilon$. Thus, almost surely we have that for large enough p, $|\varphi_p(\boldsymbol{\beta}_0) - \mathbb{E}[\bar{\varphi}_p(\boldsymbol{\beta}_0)]| < 2\sqrt{2}L\epsilon$.

Because φ_p is pseudo-Lipschitz of order 2, $\bar{\varphi}_p$ is in fact $C(1 + s_2(\pi)^2 + 1)$-Lipschitz, so that $|\varphi(\boldsymbol{x}) - \bar{\varphi}(\boldsymbol{x})| \le |\varphi(\boldsymbol{x}) - \varphi(\boldsymbol{0})| + |\bar{\varphi}(\boldsymbol{x}) - \bar{\varphi}(\boldsymbol{0})| \le C(2 + s_2(\pi)^2)\|\boldsymbol{x}\| + C(1 + \|\boldsymbol{x}\|)\|\boldsymbol{x}\|$. Thus, there exists C which does not depend on p, ϵ, M such that $|\varphi_p(\boldsymbol{x}) - \bar{\varphi}_p(\boldsymbol{x})| \le C\|\boldsymbol{x}\|^2 \mathbf{1}_{\|\boldsymbol{x}\|^2 \ge s_2(\pi)^2+1}$. Thus, $|\mathbb{E}[\bar{\varphi}_p(\boldsymbol{\beta}_0)] - \mathbb{E}[\varphi_p(\boldsymbol{\beta}_0)]| < C\mathbb{E}[\|\boldsymbol{\beta}_0\|^2 \mathbf{1}_{\|\boldsymbol{\beta}_0\|^2 \ge s_2(\pi)^2+1}] \to 0$ because $\|\boldsymbol{\beta}_0\|^2$ is uniformly integrable and $\mathbb{P}(\|\boldsymbol{\beta}_0\|^2 \ge s_2(\pi)^2 + 1) \to 0$. We conclude that almost surely we have that for large enough p, $|\varphi_p(\boldsymbol{\beta}_0) - \mathbb{E}[\varphi_p(\boldsymbol{\beta}_0)]| < 2\sqrt{2}L\epsilon$. Because the left-hand side does not depend on ϵ, in fact $\varphi_p(\boldsymbol{\beta}_0) \overset{\text{as}}{\simeq} \mathbb{E}[\varphi_p(\boldsymbol{\beta}_0)]$ as desired.

The identities involving $\mathsf{R}_{\mathsf{reg},\mathsf{cvx}}(\tau, \lambda, \mathcal{T})$, $\mathsf{W}_{\mathsf{reg},\mathsf{cvx}}(\tau, \lambda, \mathcal{T})$, and $\mathsf{K}_{\mathsf{reg},\mathsf{cvx}}(\boldsymbol{T}, \lambda, \mathcal{T})$ now hold because $\boldsymbol{\beta}_0 \mapsto \mathbb{E}_{\boldsymbol{z}}\left[\|\mathsf{prox}[\lambda\rho_p](\boldsymbol{\beta}_0 + \tau\boldsymbol{z}) - \boldsymbol{\beta}_0\|^2\right]$ is uniformly pseudo-Lipschitz of order 2, and likewise for the remaining relevant functions. The proof is complete. □

Lemma C.4. *Consider a sequence $\{\varphi_p : (\mathbb{R}^p)^{k+1} \to \mathbb{R}\}$ of uniformly pseudo-Lipschitz functions of order 2. Moreover, assume φ are symmetric in the sense that for any $\sigma \in S_p$, the symmetric group on $[p]$, we have*

$$\varphi_p(\boldsymbol{x}_0^\sigma, \dots, \boldsymbol{x}_k^\sigma) = \varphi_p(\boldsymbol{x}_0, \dots, \boldsymbol{x}_k),$$

where $(\boldsymbol{x}^\sigma)_i := \boldsymbol{x}_{\sigma(i)}$. Fix deterministic sequence $\{\boldsymbol{x}_0(p)\}$ such that $p^{-1}\sum_{i=1}^p \delta_{\sqrt{p}x_{0,i}} \overset{\text{W}}{\to} \pi$ for some $\pi \in \mathcal{P}_2(\mathbb{R})$. Then

$$\lim_{p\to\infty} \mathbb{E}_{\boldsymbol{z}_1,\dots,\boldsymbol{z}_k}[\varphi_p(\boldsymbol{x}_0, \boldsymbol{z}_1, \dots, \boldsymbol{z}_k)] = \lim_{p\to\infty} \mathbb{E}_{\tilde{\boldsymbol{x}}_0,\boldsymbol{z}_1,\dots,\boldsymbol{z}_k}[\varphi_p(\tilde{\boldsymbol{x}}_0, \boldsymbol{z}_1, \dots, \boldsymbol{z}_k)] \tag{C.8}$$

whenever either of the limits exists, where on the right-hand side we take $\tilde{\boldsymbol{x}}_0$ with coordinates distributed iid from $\pi/\sqrt{p}$. In particular, both limits exist as soon as one of them exists.

Proof of Lemma C.4. We now drop index p from our notation to avoid clutter. Consider a probability space on which we have random vectors $\tilde{\boldsymbol{x}}_0, \boldsymbol{z}_1, \dots, \boldsymbol{z}_k \in \mathbb{R}^p$ for each p such that the coordinates of $\tilde{\boldsymbol{x}}_0$ are distributed iid from $\pi/\sqrt{p}$ and the $\tilde{\boldsymbol{x}}_0$ are independent for different values of p. By [BF81, Lemma 8.4],

$$d_{\text{W}}\left(\widehat{\pi}_{\boldsymbol{x}_0}, \widehat{\pi}_{\tilde{\boldsymbol{x}}_0}\right) \le d_{\text{W}}\left(\widehat{\pi}_{\boldsymbol{x}_0}, \pi\right) + d_{\text{W}}\left(\pi, \widehat{\pi}_{\tilde{\boldsymbol{x}}_0}\right) \overset{\text{a.s.}}{\to} 0, \tag{C.9}$$

where $\widehat{\pi}_{\boldsymbol{v}} \equiv p^{-1}\sum_{i=1}^p \delta_{\sqrt{p}v_i}$ denotes the empirical distributions of the entries of $\boldsymbol{v} \in \mathbb{R}^p$. For each p and realization $\tilde{\boldsymbol{x}}_0$, there is a permutation σ_p (depending on $\tilde{\boldsymbol{x}}_0$) such that $\left\|\boldsymbol{x}_0 - \tilde{\boldsymbol{x}}_0^{\sigma_p}\right\| = d_{\text{W}}\left(\widehat{\pi}_{\boldsymbol{x}_0}, \widehat{\pi}_{\tilde{\boldsymbol{x}}_0}\right)$. By the symmetry of φ_p, we have $\varphi_p(\tilde{\boldsymbol{x}}_0, \boldsymbol{z}_1, \dots, \boldsymbol{z}_k) = \varphi_p\left(\tilde{\boldsymbol{x}}_0^{\sigma_p}, \boldsymbol{z}_1^{\sigma_p}, \dots, \boldsymbol{z}_k^{\sigma_p}\right) \overset{\text{d}}{=} \varphi_p\left(\tilde{\boldsymbol{x}}_0^{\sigma_p}, \boldsymbol{z}_1, \dots, \boldsymbol{z}_k\right)$, where the equality of distribution follows because σ_p is independent of $\boldsymbol{z}_1, \dots, \boldsymbol{z}_k$, and the distribution of $\boldsymbol{z}_1, \dots, \boldsymbol{z}_k$ is invariant under permutation of the coordinates. We have

$$\begin{aligned}\left|\varphi_p\left(\boldsymbol{x}_0, \boldsymbol{z}_1, \dots, \boldsymbol{z}_k\right) - \varphi_p\left(\tilde{\boldsymbol{x}}_0^{\sigma_p}, \boldsymbol{z}_1, \dots, \boldsymbol{z}_k\right)\right| &\le L\left(1 + \|\boldsymbol{x}_0\| + \|\tilde{\boldsymbol{x}}_0^{\sigma_p}\| + 2\sum_{i=1}^k \|\boldsymbol{z}_i\|\right)\|\boldsymbol{x}_0 - \tilde{\boldsymbol{x}}_0^{\sigma}\| \\ &= L\left(1 + \|\boldsymbol{x}_0\| + \|\tilde{\boldsymbol{x}}_0\| + 2\sum_{i=1}^k \|\boldsymbol{z}_i\|\right) d_{\text{W}}\left(\widehat{\pi}_{\boldsymbol{x}_0}, \widehat{\pi}_{\tilde{\boldsymbol{x}}_0}\right) \overset{\text{p}}{\to} 0,\end{aligned} \tag{C.10}$$

where we have used (C.9) and that $\left(1 + \|\boldsymbol{x}_0\| + \|\tilde{\boldsymbol{x}}_0\| + 2\sum_{i=1}^k \|\boldsymbol{z}_i\|\right) = O_p(1)$. Further, we check uniform integrability. First,

$$\left|\varphi_p\left(\boldsymbol{x}_0, \boldsymbol{z}_1, \ldots, \boldsymbol{z}_k\right) - \varphi_p\left(\tilde{\boldsymbol{x}}_0^{\sigma_p}, \boldsymbol{z}_1, \ldots, \boldsymbol{z}_k\right)\right| \leq L\left(1 + \|\boldsymbol{x}_0\| + \|\tilde{\boldsymbol{x}}_0\| + \sum_{i=1}^k \|\boldsymbol{z}_i\|\right)\left(\|\boldsymbol{x}_0\| + \|\tilde{\boldsymbol{x}}_0\|\right).$$

Because $\|\boldsymbol{x}_0\|$ is bounded, we only need to check that $\|\tilde{\boldsymbol{x}}_0\|^2$ and $\|\boldsymbol{z}_i\|\|\tilde{\boldsymbol{x}}_0\|$ are uniformly integrable over p. Observe that $\|\tilde{\boldsymbol{x}}_0\|^2 = \frac{1}{p}\sum_{j=1}^p (\sqrt{p}\tilde{x}_{0j})^2$. The random variables $(\sqrt{p}\tilde{x}_{0j})^2$ are iid from an L_1 probability distribution (which does not depend on p), so that $\|\tilde{\boldsymbol{x}}_0\|^2$ are uniformly integrable. Also, $\|\boldsymbol{z}_i\|\|\tilde{\boldsymbol{x}}_0\| \leq \frac{1}{2}\left(\|\tilde{\boldsymbol{x}}_0\|^2 + \|\boldsymbol{z}_i\|^2\right)$, so these are uniformly integrable for the same reason. Thus, the probabilistic convergence (C.10) and Vitali's Convergence Theorem (see e.g. [Dur10, Theorem 5.5.2]) implies that

$$\begin{aligned}
&\left|\mathbb{E}_{\boldsymbol{z}_1,\ldots,\boldsymbol{z}_k}[\varphi_p(\boldsymbol{x}_0, \boldsymbol{z}_1, \ldots, \boldsymbol{z}_k)] - \mathbb{E}_{\tilde{\boldsymbol{x}}_0,\boldsymbol{z}_1,\ldots,\boldsymbol{z}_k}[\varphi_p(\tilde{\boldsymbol{x}}_0, \boldsymbol{z}_1, \ldots, \boldsymbol{z}_k)]\right| \\
&\qquad = \left|\mathbb{E}_{\boldsymbol{z}_1,\ldots,\boldsymbol{z}_k}\left[\varphi_p\left(\boldsymbol{x}_0, \boldsymbol{z}_1, \ldots, \boldsymbol{z}_k\right)\right] - \mathbb{E}_{\tilde{\boldsymbol{x}}_0,\boldsymbol{z}_1,\ldots,\boldsymbol{z}_k}\left[\varphi_p\left(\tilde{\boldsymbol{x}}_0^{\sigma_p}, \boldsymbol{z}_1, \ldots, \boldsymbol{z}_k\right)\right]\right| \\
&\qquad \leq \mathbb{E}_{\tilde{\boldsymbol{x}}_0,\boldsymbol{z}_1,\ldots,\boldsymbol{z}_k}\left[\left|\varphi_p\left(\boldsymbol{x}_0, \boldsymbol{z}_1, \ldots, \boldsymbol{z}_k\right) - \varphi_p\left(\tilde{\boldsymbol{x}}_0^{\sigma_p}, \boldsymbol{z}_1, \ldots, \boldsymbol{z}_k\right)\right|\right] \to 0. \qquad \text{(C.11)}
\end{aligned}$$

Thus, if $\mathbb{E}_{\boldsymbol{z}_1,\ldots,\boldsymbol{z}_k}[\varphi_p(\boldsymbol{x}_0, \boldsymbol{z}_1, \ldots, \boldsymbol{z}_k)]$ converges, then $\mathbb{E}_{\tilde{\boldsymbol{x}}_0,\boldsymbol{z}_1,\ldots,\boldsymbol{z}_k}[\varphi_p(\tilde{\boldsymbol{x}}_0, \boldsymbol{z}_1, \ldots, \boldsymbol{z}_k)]$ also converges and has the same limit, and conversely. □

Lemma C.5. *Consider $\pi \in \mathcal{P}_2(\mathbb{R})$ and $\{\rho_p\} \in \mathcal{B}$ (see (C.6)). For each p, let $\mathcal{T}_p = (\pi, \rho_p)$ and consider the functions $\mathsf{R}_{\mathsf{reg,cvx}}(\tau, \lambda, \mathcal{T}_p)$, $\mathsf{W}_{\mathsf{reg,cvx}}(\tau, \lambda, \mathcal{T}_p)$, and $\mathsf{K}_{\mathsf{reg,cvx}}(\boldsymbol{T}, \lambda, \mathcal{T}_p)$ defined by (B.4). Consider $0 < \tau_{\mathsf{min}} \leq \tau_{\mathsf{max}}$ and $0 < \lambda_{\mathsf{min}} \leq \lambda_{\mathsf{max}}$. We have the following:*

(i) $\mathsf{R}_{\mathsf{reg,cvx}}$ is uniformly (over p) Lipschitz continuous in τ and λ for $(\tau, \lambda) \in [0, \tau_{\mathsf{max}}] \times [\lambda_{\mathsf{min}}, \lambda_{\mathsf{max}}]$.

(ii) $\mathsf{W}_{\mathsf{reg,cvx}}$ is uniformly (over p) Lipschitz continuous in τ and λ for $(\tau, \lambda) \in [\tau_{\mathsf{min}}, \tau_{\mathsf{max}}] \times [\lambda_{\mathsf{min}}, \lambda_{\mathsf{max}}]$.

(iii) $\mathsf{K}_{\mathsf{reg,cvx}}$ is uniformly (over p) equicontinuous in $\boldsymbol{T}$ and uniformly Lipschitz continuous in λ for $\boldsymbol{0} \preceq \boldsymbol{T} \preceq \tau_{\mathsf{max}}^2 \boldsymbol{I}_2$ and $\lambda \in [\lambda_{\mathsf{min}}, \lambda_{\mathsf{max}}]$.

Proof of Lemma C.5. Let $M > \sup_p \|\mathsf{prox}[\rho_p](\boldsymbol{0})\|$ with $M < \infty$, which is permitted because $\{\rho_p\} \in \mathcal{B}$. Throughout the proof, we will denote by C a constant which may depend on M, π, τ_{max}, λ_{min}, or λ_{max}, but not on p or τ_{min}, and will denote by C_+ a constant which may depend also on τ_{min} but not on p. Both C and C_+ may differ at different appearances, even within the same chain of inequalities, as it absorbs terms.

Observe that for any λ we have

$$\begin{aligned}
\|\mathsf{prox}[\lambda\rho_p](\boldsymbol{\beta}_0 + \tau\boldsymbol{z})\| &\leq \|\mathsf{prox}[\rho_p](\boldsymbol{0})\| + \|\mathsf{prox}[\lambda\rho_p](\boldsymbol{0}) - \mathsf{prox}[\rho_p](\boldsymbol{0})\| + \|\mathsf{prox}[\lambda\rho_p](\boldsymbol{\beta}_0 + \tau\boldsymbol{z}) - \mathsf{prox}[\lambda\rho_p](\boldsymbol{0})\| \\
&\leq M + \|\mathsf{prox}[\rho_p](\boldsymbol{0})\||\lambda - 1| + \|\boldsymbol{\beta}_0\| + \tau\|\boldsymbol{z}\| \leq M(2 + \lambda_{\mathsf{max}}) + \|\boldsymbol{\beta}_0\| + \tau_{\mathsf{max}}\|\boldsymbol{z}\| \\
&\leq C\left(1 + \|\boldsymbol{\beta}_0\| + \|\boldsymbol{z}\|\right), \qquad \text{(C.12)}
\end{aligned}$$

where in the second inequality we have used (O.4) and (O.5). With one more application of the triangle inequality, we get

$$\|\mathsf{prox}[\lambda\rho_p](\boldsymbol{\beta}_0 + \tau\boldsymbol{z}) - \boldsymbol{\beta}_0\| \leq C\left(1 + \|\boldsymbol{\beta}_0\| + \|\boldsymbol{z}\|\right). \qquad \text{(C.13)}$$

Further, observe that for $\lambda, \lambda' \in [\lambda_{\min}, \lambda_{\max}]$ and $\tau \in [\tau_{\min}, \tau_{\max}]$, we have by applying (C.12) and the triangle inequality

$$\begin{aligned}\left\|\mathsf{prox}[\lambda\rho_p](\boldsymbol{\beta}_0 + \tau\boldsymbol{z}) - \mathsf{prox}[\lambda'\rho_p](\boldsymbol{\beta}_0 + \tau\boldsymbol{z})\right\| &\le \|\boldsymbol{\beta}_0 + \tau\boldsymbol{z} - \mathsf{prox}[\lambda\rho_p](\boldsymbol{\beta}_0 + \tau\boldsymbol{z})\| \left|\frac{\lambda'}{\lambda} - 1\right| \\ &\le C\left(1 + \|\boldsymbol{\beta}_0\| + \|\boldsymbol{z}\|\right)|\lambda - \lambda'|, \end{aligned} \tag{C.14}$$

where in the first inequality we have used (O.5) and in the second inequality we have usedC.13 and that $\left|\frac{\lambda'}{\lambda} - 1\right| = \frac{|\lambda - \lambda'|}{\lambda} \le \frac{|\lambda - \lambda'|}{\lambda \wedge \lambda'} \le C|\lambda - \lambda'|$.

We are ready to demonstrate the claimed continuity properties of $\mathsf{R}_{\mathsf{reg,cvx}}$, $\mathsf{W}_{\mathsf{reg,cvx}}$, and $\mathsf{K}_{\mathsf{reg,cvx}}$. Fix $\tau, \tau' \in [0, \tau_{\max}]$, $\lambda, \lambda' \in [\lambda_{\min}, \lambda_{\max}]$ and $\mathbf{0} \preceq \boldsymbol{T}, \boldsymbol{T}' \preceq \tau_{\max}^2 \boldsymbol{I}_2$. These will remain fixed throughout the remainder of the proof unless otherwise stated.

Uniform Lipschitz continuity of $\mathsf{R}_{\mathsf{reg,cvx}}$ in τ. We apply (C.4) identifying $\boldsymbol{a} = \mathsf{prox}[\lambda\rho_p](\boldsymbol{\beta}_0 + \tau\boldsymbol{z}) - \boldsymbol{\beta}_0$ and $\boldsymbol{b} = \mathsf{prox}[\lambda\rho_p](\boldsymbol{\beta}_0 + \tau'\boldsymbol{z}) - \boldsymbol{\beta}_0$. Using C.13 and (O.4) to bound $(*)$ and $(**)$ respectively, we get

$$\left|\|\boldsymbol{a}\|^2 - \|\boldsymbol{b}\|^2\right| \le \underbrace{C\left(1 + \|\boldsymbol{\beta}_0\| + \|\boldsymbol{z}\|\right)}_{\text{bound on } (*)} \cdot \underbrace{|\tau - \tau'|\|\boldsymbol{z}\|}_{\text{bound on } (**)} \le C\left(1 + \|\boldsymbol{\beta}_0\|^2 + \|\boldsymbol{z}\|^2\right)|\tau - \tau'|. \tag{C.15}$$

We have by Jensen's inequality

$$\begin{aligned}\left|\mathsf{R}_{\mathsf{reg,cvx}}(\tau, \lambda, \mathcal{T}_p) - \mathsf{R}_{\mathsf{reg,cvx}}(\tau', \lambda, \mathcal{T}_p)\right| &= \left|\mathbb{E}_{\boldsymbol{\beta}_0, \boldsymbol{z}}\left[\|\boldsymbol{a}\|^2\right] - \mathbb{E}_{\boldsymbol{\beta}_0, \boldsymbol{z}}\left[\|\boldsymbol{b}\|^2\right]\right| \le \mathbb{E}_{\boldsymbol{\beta}_0, \boldsymbol{z}}\left[\left|\|\boldsymbol{a}\|^2 - \|\boldsymbol{b}\|^2\right|\right] \\ &\le C\mathbb{E}_{\boldsymbol{\beta}_0, \boldsymbol{z}}\left[\left(1 + \|\boldsymbol{\beta}_0\|^2 + \|\boldsymbol{z}\|^2\right)\right]|\tau - \tau'| = C|\tau - \tau'|. \end{aligned} \tag{C.16}$$

Uniform Lipschitz continuity of $\mathsf{R}_{\mathsf{reg,cvx}}$ in λ. We apply (C.4) identifying $\boldsymbol{a} = \mathsf{prox}[\lambda\rho_p](\boldsymbol{\beta}_0 + \tau\boldsymbol{z}) - \boldsymbol{\beta}_0$ and $\boldsymbol{b} = \mathsf{prox}[\lambda'\rho_p](\boldsymbol{\beta}_0 + \tau\boldsymbol{z}) - \boldsymbol{\beta}_0$. Using C.13 and (C.14) to bound $(*)$ and $(**)$ respectively, we get

$$\begin{aligned}\left|\|\boldsymbol{a}\|^2 - \|\boldsymbol{b}\|^2\right| &\le \underbrace{C\left(1 + \|\boldsymbol{\beta}_0\| + \|\boldsymbol{z}\|\right)}_{\text{bound on } (*)} \cdot \underbrace{C\left(1 + \|\boldsymbol{\beta}_0\| + \|\boldsymbol{z}\|\right)|\lambda - \lambda'|}_{\text{bound on } (**)} \\ &\le C\left(1 + \|\boldsymbol{\beta}_0\|^2 + \|\boldsymbol{z}\|^2\right)|\lambda - \lambda'|. \end{aligned} \tag{C.17}$$

We have by Jensen's inequality

$$\begin{aligned}\left|\mathsf{R}_{\mathsf{reg,cvx}}(\tau, \lambda, \mathcal{T}_p) - \mathsf{R}_{\mathsf{reg,cvx}}(\tau, \lambda', \mathcal{T}_p)\right| &= \left|\mathbb{E}_{\boldsymbol{\beta}_0, \boldsymbol{z}}\left[\|\boldsymbol{a}\|^2\right] - \mathbb{E}_{\boldsymbol{\beta}_0, \boldsymbol{z}}\left[\|\boldsymbol{b}\|^2\right]\right| \le \mathbb{E}_{\boldsymbol{\beta}_0, \boldsymbol{z}}\left[\left|\|\boldsymbol{a}\|^2 - \|\boldsymbol{b}\|^2\right|\right] \\ &\le C\mathbb{E}_{\boldsymbol{\beta}_0, \boldsymbol{z}}\left[\left(1 + \|\boldsymbol{\beta}_0\|^2 + \|\boldsymbol{z}\|^2\right)\right]|\lambda - \lambda'| = C|\lambda - \lambda'|. \end{aligned} \tag{C.18}$$

Uniform Lipschitz continuity of $\mathsf{W}_{\mathsf{reg,cvx}}$ in τ. In this section only, we require also that $\tau, \tau' \ge \tau_{\min}$. We apply (C.3) identifying $\boldsymbol{a} = \frac{\boldsymbol{z}}{\tau}$, $\boldsymbol{b} = \mathsf{prox}[\lambda\rho_p](\boldsymbol{\beta}_0 + \tau\boldsymbol{z})$, $\boldsymbol{a}' = \frac{\boldsymbol{z}}{\tau'}$, and $\boldsymbol{b}' = \mathsf{prox}[\lambda\rho_p](\boldsymbol{\beta}_0 + \tau'\boldsymbol{z})$. Observe that $\|\boldsymbol{a} - \boldsymbol{a}'\| = |1/\tau - 1/\tau'|\|\boldsymbol{z}\| \le C_+\|\boldsymbol{z}\||\tau - \tau'|$, where the last inequality holds because $\tau, \tau' \ge \tau_{\min} > 0$. Using (C.12) and (O.4) to bound $\max\{\|\boldsymbol{a}\|, \|\boldsymbol{a}'\|, \|\boldsymbol{b}\|, \|\boldsymbol{b}'\|\}$ and $\|\boldsymbol{b} - \boldsymbol{b}'\|$ respectively, we get

$$\left|\langle \boldsymbol{a}, \boldsymbol{b}\rangle - \langle \boldsymbol{a}', \boldsymbol{b}'\rangle\right| \le \underbrace{C\left(1 + \|\boldsymbol{\beta}_0\| + \|\boldsymbol{z}\|\right)}_{\text{bound on } (*)} \cdot \underbrace{C_+\|\boldsymbol{z}\||\tau - \tau'|}_{\text{bound on } (**)} \le C_+\left(1 + \|\boldsymbol{\beta}_0\|^2 + \|\boldsymbol{z}\|^2\right)|\tau - \tau'|. \tag{C.19}$$

We have by Jensen's inequality

$$\begin{aligned}\left|\mathsf{W}_{\mathsf{reg,cvx}}(\tau,\lambda,\mathcal{T}_p)-\mathsf{W}_{\mathsf{reg,cvx}}(\tau',\lambda,\mathcal{T}_p)\right| &= \left|\mathbb{E}_{\boldsymbol{\beta}_0,\boldsymbol{z}}\left[\langle \boldsymbol{a},\boldsymbol{b}\rangle\right]-\mathbb{E}_{\boldsymbol{\beta}_0,\boldsymbol{z}}[\langle \boldsymbol{a}',\boldsymbol{b}'\rangle]\right| \le \mathbb{E}_{\boldsymbol{\beta}_0,\boldsymbol{z}}\left[\left|\langle \boldsymbol{a},\boldsymbol{b}\rangle-\langle \boldsymbol{a}',\boldsymbol{b}'\rangle\right|\right]\\ &\le C_+\mathbb{E}_{\boldsymbol{\beta}_0,\boldsymbol{z}}\left[\left(1+\|\boldsymbol{\beta}_0\|^2+\|\boldsymbol{z}\|^2\right)\right]|\tau-\tau'| = C_+|\tau-\tau'|.\end{aligned} \tag{C.20}$$

Uniform Lipschitz continuity of $\mathsf{W}_{\mathsf{reg,cvx}}$ in λ. We apply (C.3) identifying $\boldsymbol{a}=\frac{\boldsymbol{z}}{\tau}$, $\boldsymbol{b}=\mathsf{prox}[\lambda\rho_p](\boldsymbol{\beta}_0+\tau\boldsymbol{z})$, $\boldsymbol{a}'=\frac{\boldsymbol{z}}{\tau}$, and $\boldsymbol{b}'=\mathsf{prox}[\lambda'\rho_p](\boldsymbol{\beta}_0+\tau\boldsymbol{z})$. Using (C.12) and (C.14) to bound $(*)$ and $(**)$ respectively, we get

$$\left|\langle \boldsymbol{a},\boldsymbol{b}\rangle-\langle \boldsymbol{a}',\boldsymbol{b}'\rangle\right| \le \underbrace{C\left(1+\|\boldsymbol{\beta}_0\|+\|\boldsymbol{z}\|\right)}_{\text{bound on }(*)}\cdot\underbrace{C\left(1+\|\boldsymbol{\beta}_0\|+\|\boldsymbol{z}\|\right)|\lambda-\lambda'|}_{\text{bound on }(**)} \le C\left(1+\|\boldsymbol{\beta}_0\|^2+\|\boldsymbol{z}\|^2\right)|\lambda-\lambda'|. \tag{C.21}$$

We have by Jensen's inequality

$$\begin{aligned}\left|\mathsf{W}_{\mathsf{reg,cvx}}(\tau,\lambda,\mathcal{T}_p)-\mathsf{W}_{\mathsf{reg,cvx}}(\tau,\lambda',\mathcal{T}_p)\right| &= \left|\mathbb{E}_{\boldsymbol{\beta}_0,\boldsymbol{z}}\left[\langle \boldsymbol{a},\boldsymbol{b}\rangle\right]-\mathbb{E}_{\boldsymbol{\beta}_0,\boldsymbol{z}}[\langle \boldsymbol{a}',\boldsymbol{b}'\rangle]\right| \le \mathbb{E}_{\boldsymbol{\beta}_0,\boldsymbol{z}}\left[\left|\langle \boldsymbol{a},\boldsymbol{b}\rangle-\langle \boldsymbol{a}',\boldsymbol{b}'\rangle\right|\right]\\ &\le C\mathbb{E}_{\boldsymbol{\beta}_0,\boldsymbol{z}}\left[\left(1+\|\boldsymbol{\beta}_0\|^2+\|\boldsymbol{z}\|^2\right)\right]|\lambda-\lambda'| = C|\lambda-\lambda'|.\end{aligned} \tag{C.22}$$

Uniform equicontinuity of $\mathsf{K}_{\mathsf{reg,cvx}}$ in $\boldsymbol{T}$. Let $\boldsymbol{T},\boldsymbol{T}'\in S_+^2$. By [GS84, Proposition 7], we have

$$d_{\mathrm{W}}\left(\mathsf{N}(\boldsymbol{0},\boldsymbol{T}),\mathsf{N}(\boldsymbol{0},\boldsymbol{T}')\right)=\sqrt{\mathrm{Tr}\left(\boldsymbol{T}+\boldsymbol{T}'-2(\boldsymbol{T}^{1/2}\boldsymbol{T}'\boldsymbol{T}^{1/2})^{1/2}\right)}. \tag{C.23}$$

By [GS84, Proposition 1], there exists a coupling which achieves the infimum in (1.4). Let ν a probability distribution on $\mathbb{R}^4$ which implements the minimal coupling between $\mathsf{N}(\boldsymbol{0},\boldsymbol{T})$ and $\mathsf{N}(\boldsymbol{0},\boldsymbol{T}')$. Consider a probability space with random vectors $\boldsymbol{\beta}_0$ and $\boldsymbol{z}_1,\boldsymbol{z}_2,\boldsymbol{z}_1',\boldsymbol{z}_2'$ for all p such that $(z_{1j},z_{2j},z_{1j}',z_{2j}')\overset{\text{iid}}{\sim}\nu/\sqrt{p}$. Then

$$\mathbb{E}_{\boldsymbol{z}_1,\boldsymbol{z}_2,\boldsymbol{z}_1',\boldsymbol{z}_2'}\left[\|\boldsymbol{z}_1-\boldsymbol{z}_1'\|^2+\|\boldsymbol{z}_2-\boldsymbol{z}_2'\|^2\right]=\mathrm{Tr}\left(\boldsymbol{T}+\boldsymbol{T}'-2(\boldsymbol{T}^{1/2}\boldsymbol{T}'\boldsymbol{T}^{1/2})^{1/2}\right). \tag{C.24}$$

We apply (C.3) identifying $\boldsymbol{a}=\mathsf{prox}[\lambda\rho_p](\boldsymbol{\beta}_0+\boldsymbol{z}_1)-\boldsymbol{\beta}_0$, $\boldsymbol{b}=\mathsf{prox}[\lambda\rho_p](\boldsymbol{\beta}_0+\boldsymbol{z}_2)-\boldsymbol{\beta}_0$, $\boldsymbol{a}'=\mathsf{prox}[\lambda\rho_p](\boldsymbol{\beta}_0+\boldsymbol{z}_1')-\boldsymbol{\beta}_0$, and $\boldsymbol{b}'=\mathsf{prox}[\lambda\rho_p](\boldsymbol{\beta}_0+\boldsymbol{z}_2')-\boldsymbol{\beta}_0$. Using C.13 and (O.4) to bound $(*)$ and $(**)$ respectively, we get

$$\left|\langle \boldsymbol{a},\boldsymbol{b}\rangle-\langle \boldsymbol{a}',\boldsymbol{b}'\rangle\right| \le \underbrace{C\left(1+\|\boldsymbol{\beta}_0\|+\max(\|\boldsymbol{z}_1\|,\|\boldsymbol{z}_2\|,\|\boldsymbol{z}_1'\|,\|\boldsymbol{z}_2'\|)\right)}_{\text{bound on }(*)}\cdot\underbrace{\max(\|\boldsymbol{z}_1-\boldsymbol{z}_1'\|,\|\boldsymbol{z}_2-\boldsymbol{z}_2'\|)}_{\text{bound on }(**)}. \tag{C.25}$$

We have by Jensen's inequality and Cauchy-Schwartz

$$\begin{aligned}&\left|\mathsf{K}_{\mathsf{reg,cvx}}(\boldsymbol{T},\lambda,\mathcal{T}_p)-\mathsf{K}_{\mathsf{reg,cvx}}(\boldsymbol{T}',\lambda,\mathcal{T}_p)\right|\\ &\quad\le C\mathbb{E}_{\boldsymbol{\beta}_0,\boldsymbol{z}_1,\boldsymbol{z}_2,\boldsymbol{z}_1',\boldsymbol{z}_2'}\left[\left(1+\|\boldsymbol{\beta}_0\|+\max(\|\boldsymbol{z}_1\|,\|\boldsymbol{z}_2\|,\|\boldsymbol{z}_1'\|,\|\boldsymbol{z}_2'\|)\right)^2\right]^{1/2}\\ &\qquad\times\mathbb{E}_{\boldsymbol{z}_1,\boldsymbol{z}_2,\boldsymbol{z}_1',\boldsymbol{z}_2'}\left[\max(\|\boldsymbol{z}_1-\boldsymbol{z}_1'\|,\|\boldsymbol{z}_2-\boldsymbol{z}_2'\|)^2\right]^{1/2}.\end{aligned} \tag{C.26}$$

We have

$$\mathbb{E}_{\boldsymbol{\beta}_0,\boldsymbol{z}_1,\boldsymbol{z}_2,\boldsymbol{z}_1',\boldsymbol{z}_2'}\left[\left(1+\|\boldsymbol{\beta}_0\|+\max(\|\boldsymbol{z}_1\|,\|\boldsymbol{z}_2\|,\|\boldsymbol{z}_1'\|,\|\boldsymbol{z}_2'\|)\right)^2\right]^{1/2}$$
$$\le C\mathbb{E}_{\boldsymbol{\beta}_0,\boldsymbol{z}_1,\boldsymbol{z}_2,\boldsymbol{z}_1',\boldsymbol{z}_2'}\left[1+\|\boldsymbol{\beta}_0\|^2+\|\boldsymbol{z}_1\|^2+\|\boldsymbol{z}_2\|^2+\|\boldsymbol{z}_1'\|^2+\|\boldsymbol{z}_2'\|^2\right]^{1/2}\le C. \quad \text{(C.27)}$$

Further, by (C.24), we have

$$\mathbb{E}_{\boldsymbol{z}_1,\boldsymbol{z}_2,\boldsymbol{z}_1',\boldsymbol{z}_2'}\left[\max(\|\boldsymbol{z}_1-\boldsymbol{z}_1'\|,\|\boldsymbol{z}_2-\boldsymbol{z}_2'\|)^2\right]^{1/2}\le\sqrt{\mathrm{Tr}\left(\boldsymbol{T}+\boldsymbol{T}'-2(\boldsymbol{T}^{1/2}\boldsymbol{T}'\boldsymbol{T}^{1/2})^{1/2}\right)}. \quad \text{(C.28)}$$

Thus,

$$\left|\mathsf{K}_{\mathsf{reg,cvx}}(\boldsymbol{T},\lambda,\mathcal{T}_p)-\mathsf{K}_{\mathsf{reg,cvx}}(\boldsymbol{T}',\lambda,\mathcal{T}_p)\right|\le C\sqrt{\mathrm{Tr}\left(\boldsymbol{T}+\boldsymbol{T}'-2(\boldsymbol{T}^{1/2}\boldsymbol{T}'\boldsymbol{T}^{1/2})^{1/2}\right)}. \quad \text{(C.29)}$$

Now observe that $(\boldsymbol{T},\boldsymbol{T}')\mapsto\sqrt{\mathrm{Tr}\left(\boldsymbol{T}+\boldsymbol{T}'-2(\boldsymbol{T}^{1/2}\boldsymbol{T}'\boldsymbol{T}^{1/2})^{1/2}\right)}$ is continuous and is 0 when $\boldsymbol{T}=\boldsymbol{T}'$. Thus, it is uniformly continuous on the compact domain $\{(\boldsymbol{T},\boldsymbol{T}')\in(S_+^2)^2 \mid \boldsymbol{0}\preceq\boldsymbol{T},\boldsymbol{T}'\preceq\tau_{\mathsf{max}}^2\boldsymbol{I}_2\}$ (where because this is a finite dimensional Euclidean space, continuity holds with respect to any norm by equivalence of norms). Thus, for any $\varepsilon>0$, there exists $\delta>0$ such that if $\boldsymbol{0}\preceq\boldsymbol{T},\boldsymbol{T}'\preceq\tau_{\mathsf{max}}^2\boldsymbol{I}_2$ and $\|\boldsymbol{T}-\boldsymbol{T}\|_{\mathsf{op}}<\delta$, then $\sqrt{\mathrm{Tr}\left(\boldsymbol{T}+\boldsymbol{T}'-2(\boldsymbol{T}^{1/2}\boldsymbol{T}'\boldsymbol{T}^{1/2})^{1/2}\right)}<\varepsilon$. Because this modulus of continuity does not depend upon p, we have $\mathsf{K}_{\mathsf{reg,cvx}}(\boldsymbol{T},\lambda,\mathcal{T}_p)$ is uniformly (over p) equicontinuous in $\boldsymbol{T}$.

Uniform Lipschitz continuity of $\mathsf{K}_{\mathsf{reg,cvx}}$ in λ. Let $(\boldsymbol{z}_1,\boldsymbol{z}_2)\sim\mathsf{N}(0,\boldsymbol{T}\otimes\boldsymbol{I}_p/p)$. We apply (C.3) identifying $\boldsymbol{a}=\mathsf{prox}[\lambda\rho_p](\boldsymbol{\beta}_0+\tau\boldsymbol{z}_1)-\boldsymbol{\beta}_0$, $\boldsymbol{b}=\mathsf{prox}[\lambda\rho_p](\boldsymbol{\beta}_0+\tau\boldsymbol{z}_2)-\boldsymbol{\beta}_0$, $\boldsymbol{a}'=\mathsf{prox}[\lambda'\rho_p](\boldsymbol{\beta}_0+\tau\boldsymbol{z}_1)-\boldsymbol{\beta}_0$, and $\boldsymbol{b}'=\mathsf{prox}[\lambda'\rho_p](\boldsymbol{\beta}_0+\tau\boldsymbol{z}_2)-\boldsymbol{\beta}_0$. Using C.13 and (C.14) to bound $(*)$ and $(**)$ respectively, we get

$$\left|\langle\boldsymbol{a},\boldsymbol{b}\rangle-\langle\boldsymbol{a}',\boldsymbol{b}'\rangle\right|\le\underbrace{C\left(1+\|\boldsymbol{\beta}_0\|+\|\boldsymbol{z}_1\|\vee\|\boldsymbol{z}_2\|\right)}_{\text{bound on }(*)}\cdot\underbrace{C\left(1+\|\boldsymbol{\beta}_0\|+\|\boldsymbol{z}_1\|\vee\|\boldsymbol{z}_2\|\right)|\lambda-\lambda'|}_{\text{bound on }(**)}$$
$$\le C\left(1+\|\boldsymbol{\beta}_0\|^2+\|\boldsymbol{z}_1\|^2+\|\boldsymbol{z}_2\|^2\right)|\lambda-\lambda'|. \quad \text{(C.30)}$$

We have by Jensen's inequality

$$\left|\mathsf{K}_{\mathsf{reg,cvx}}(\boldsymbol{T},\lambda,\mathcal{T}_p)-\mathsf{K}_{\mathsf{reg,cvx}}(\boldsymbol{T},\lambda',\mathcal{T}_p)\right|=\left|\mathbb{E}_{\boldsymbol{\beta}_0,\boldsymbol{z}}\left[\langle\boldsymbol{a},\boldsymbol{b}\rangle\right]-\mathbb{E}_{\boldsymbol{\beta}_0,\boldsymbol{z}}[\langle\boldsymbol{a}',\boldsymbol{b}'\rangle]\right|\le\mathbb{E}_{\boldsymbol{\beta}_0,\boldsymbol{z}}\left[\left|\langle\boldsymbol{a},\boldsymbol{b}\rangle-\langle\boldsymbol{a}',\boldsymbol{b}'\rangle\right|\right]$$
$$\le C\mathbb{E}_{\boldsymbol{\beta}_0,\boldsymbol{z}}\left[\left(1+\|\boldsymbol{\beta}_0\|^2+\|\boldsymbol{z}\|^2\right)\right]|\lambda-\lambda'|=C|\lambda-\lambda'|. \quad \text{(C.31)}$$

This completes the proof. □

D Proof of Proposition B.3

This argument follows closely that of [DM16]. In contrast to [DM16], we consider penalized procedures and use non-separable penalties. Our analysis also establishes the impact of the oracle penalty on the fixed point equations (B.7). We find that using the recent results [BMN19] for AMP with non-separable denoisers, their argument goes through.

Throughout the argument, we will frequently (but not always) drop the index p from our notation. For sequences $\{X_p\}$ and $\{Y_p\}$ of real-valued random variables all defined on the same probability space, we use the notation $X_p\overset{\mathsf{as}}{\simeq}Y_p$ to denote $|X_p-Y_p|\overset{\mathsf{as}}{\to}0$.

D.1 Proof of part (i)

For each p, define

$$L(\boldsymbol{\beta}) := \frac{1}{n}\|\boldsymbol{y} - \boldsymbol{X}^\mathsf{T}\boldsymbol{\beta}\|^2 + \rho_p(\boldsymbol{\beta}_0) + \frac{\gamma}{2}\|\boldsymbol{\beta} - \boldsymbol{\beta}_0\|^2, \tag{D.1}$$

the objective in (B.1)). If $n > p$, $\gamma > 0$, or ρ_p is strongly convex, then L is strongly convex almost surely. Thus, $\mathbb{P}_{\boldsymbol{X}}(\text{solution to (1.2) exists and is unique}) = 1$ eventually. This justifies assuming existence and uniqueness of solutions to (1.2) for the remainder of the proof.

We now prove (ii) and (iii), which require substantially more work.

D.2 Pick a typical sequence of normal vectors

Without loss of generality, we may assume p is increasing. The remainder of the argument will occur conditional on the realization of the sequence of parameters $\{\boldsymbol{\beta}_0\}$. We will be able to carry out all steps under the DSN assumption if the penalties are symmetric, or almost surely under the RSN. (We will justify this as we go).

We construct a deterministic sequence of vectors $\{\boldsymbol{z}^0(p) \in \mathbb{R}^p\}$ such that for all $\tau' \geq 0$,

$$\begin{aligned}
\lim_{p\to\infty} \|\boldsymbol{z}^0\|^2 &= 1, \\
\lim_{p\to\infty} \|\mathsf{prox}[\lambda\rho_p^{(\gamma)}](\boldsymbol{\beta}_0 + \tau\boldsymbol{z}^0) - \boldsymbol{\beta}_0\|^2 &= \delta\tau^2 - \sigma^2, \\
\lim_{p\to\infty} \mathbb{E}_{\boldsymbol{z}}\left[\left\langle \mathsf{prox}[\lambda\rho_p^{(\gamma)}](\boldsymbol{\beta}_0 + \tau\boldsymbol{z}^0) - \boldsymbol{\beta}_0, \mathsf{prox}[\lambda\rho_p^{(\gamma)}](\boldsymbol{\beta}_0 + \tau'\boldsymbol{z}) - \boldsymbol{\beta}_0 \right\rangle\right] &= \mathsf{K}^\infty_{\mathsf{reg,cvx}}\left(\boldsymbol{T}_{\tau'}, \lambda, \mathcal{T}\right),
\end{aligned} \tag{D.2}$$

where $\boldsymbol{z} \sim \mathsf{N}(\boldsymbol{0}, \boldsymbol{I}_p/p)$ and $\boldsymbol{T}_{\tau'} := \begin{pmatrix} \tau^2 & 0 \\ 0 & \tau'^2 \end{pmatrix}$. Such a sequence exists because if we draw $\boldsymbol{z}^0(p) \sim \mathsf{N}(\boldsymbol{0}, \boldsymbol{I}_p/p)$ independently across p, then $\{\boldsymbol{z}^0\}$ satisfies the required properties (simultaneously over τ') almost surely, as we now show.

For such random $\boldsymbol{z}^0$, by Gaussian Lipschitz concentration (Lemma P.8), we have

$$\mathbb{P}_{\boldsymbol{z}_0}\left(\left|\|\boldsymbol{z}^0\| - \mathbb{E}_{\boldsymbol{z}^0}\left[\|\boldsymbol{z}^0\|\right]\right| > t/p^{1/4}\right) \leq 2e^{-\frac{p^{1/2}}{2}t^2}. \tag{D.3}$$

Because the right-hand side is summable over p (recall we assume p is increasing), we have by Borel-Cantelli that $\|\boldsymbol{z}^0\| \overset{\mathsf{as}}{\simeq} \mathbb{E}_{\boldsymbol{z}^0}\left[\|\boldsymbol{z}^0\|\right] \to 1$. Thus, the first identity of (D.2) holds almost surely.

For the remaining two identities, first note that by by (O.16), the second and third identities of (D.2) are equivalent to

$$\begin{aligned}
\lim_{p\to\infty} \|\mathsf{prox}[\lambda_{\mathsf{orc}}\rho_p](\boldsymbol{\beta}_0 + \tau_{\mathsf{orc}}\boldsymbol{z}^0) - \boldsymbol{\beta}_0\|^2 &= \delta\tau^2 - \sigma^2, \\
\lim_{p\to\infty} \mathbb{E}_{\boldsymbol{z}}\left[\left\langle \mathsf{prox}[\lambda_{\mathsf{orc}}\rho_p](\boldsymbol{\beta}_0 + \tau_{\mathsf{orc}}\boldsymbol{z}^0) - \boldsymbol{\beta}_0, \mathsf{prox}[\lambda_{\mathsf{orc}}\rho_p](\boldsymbol{\beta}_0 + \tau'_{\mathsf{orc}}\boldsymbol{z}) - \boldsymbol{\beta}_0 \right\rangle\right] &= \mathsf{K}^\infty_{\mathsf{reg,cvx}}\left(\boldsymbol{T}_{\tau'}, \lambda, \mathcal{T}\right).
\end{aligned}$$

If the ρ_n are symmetric, π has finite second moments, and the ρ_n are symmetric, then by Lemmas C.1 and C.4 and the symmetry of ρ_p, under the DSN assumption, for all $\tau' \geq 0$, $\lambda' \geq 0$, $\boldsymbol{T} \succeq 0$ fixed,

$$\begin{aligned}
\mathsf{R}^\infty_{\mathsf{reg,cvx}}(\tau', \lambda', \mathcal{T}_p) &= \lim_{p\to\infty} \mathbb{E}_{\boldsymbol{z}}\left[\|\mathsf{prox}[\lambda'\rho_p](\boldsymbol{\beta}_0 + \tau'\boldsymbol{z}) - \boldsymbol{\beta}_0\|^2\right], \\
\mathsf{W}^\infty_{\mathsf{reg,cvx}}(\tau', \lambda', \mathcal{T}_p) &= \lim_{p\to\infty} \frac{1}{\tau'}\mathbb{E}_{\boldsymbol{z}}\left[\langle \boldsymbol{z}, \mathsf{prox}[\lambda'\rho_p](\boldsymbol{\beta}_0 + \tau'\boldsymbol{z})\rangle\right], \\
\mathsf{K}^\infty_{\mathsf{reg,cvx}}(\boldsymbol{T}, \lambda', \mathcal{T}_p) &= \lim_{p\to\infty} \mathbb{E}_{\boldsymbol{z}_1,\boldsymbol{z}_2}\left[\left\langle \mathsf{prox}\left[\lambda'\rho_p\right](\boldsymbol{\beta}_0 + \boldsymbol{z}_1) - \boldsymbol{\beta}_0, \mathsf{prox}\left[\lambda'\rho_p\right](\boldsymbol{\beta}_0 + \boldsymbol{z}_2) - \boldsymbol{\beta}_0 \right\rangle\right],
\end{aligned} \tag{D.4}$$

(Note the expectations are only over the Gaussian random vectors and $\boldsymbol{\beta}_0$ is fixed). If the ρ_n are not necessarily symmetric, then under the RSN assumption, the previous display holds almost surely with respect to the realization of $\{\boldsymbol{\beta}_0\}$ by Lemma C.3.

Because $f_p(\boldsymbol{x}; \tau_{\mathsf{orc}}) := \mathsf{prox}[\lambda\rho_p](\boldsymbol{\beta}_0 + \tau_{\mathsf{orc}}\boldsymbol{x}) - \boldsymbol{\beta}_0$ is τ_{orc}-Lipschitz by (O.4), we have by Gaussian Lipschitz concentration (Lemma P.8) and Borel-Cantelli that

$$\left\| f_p(\boldsymbol{z}^0; \tau_{\mathsf{orc}}) \right\| \overset{\mathsf{as}}{\simeq} \mathbb{E}_{\boldsymbol{z}^0}\left[\left\| f_p(\boldsymbol{z}^0; \tau_{\mathsf{orc}}) \right\| \right]. \tag{D.5}$$

Now observe by Jensen's inequality that $\mathbb{E}_{\boldsymbol{z}^0}\left[\left\| f_p(\boldsymbol{z}^0; \tau_{\mathsf{orc}}) \right\|^2 \right] \geq \mathbb{E}_{\boldsymbol{z}^0}\left[\left\| f_p(\boldsymbol{z}^0; \tau_{\mathsf{orc}}) \right\| \right]^2$. By assumption, the left-hand side and hence the right-hand side is bounded. By exponential concentration of $\left\| f_p(\boldsymbol{z}^0; \tau_{\mathsf{orc}}) \right\|$, we conclude that $\left\| f_p(\boldsymbol{z}^0; \tau_{\mathsf{orc}}) \right\|^2$ is uniformly integrable. Because it concentrates on $\mathbb{E}_{\boldsymbol{z}^0}\left[\left\| f_p(\boldsymbol{z}^0; \tau_{\mathsf{orc}}) \right\| \right]^2$, we have

$$\lim_{p\to\infty} \mathbb{E}_{\boldsymbol{z}^0}\left[\left\| f_p(\boldsymbol{z}^0; \tau_{\mathsf{orc}}) \right\| \right]^2 = \lim_{p\to\infty} \mathbb{E}_{\boldsymbol{z}^0}\left[\left\| f_p(\boldsymbol{z}^0; \tau_{\mathsf{orc}}) \right\|^2 \right] = \delta\tau^2 - \sigma^2. \tag{D.6}$$

Then by (D.5), $\left\| f_p(\boldsymbol{z}^0; \tau_{\mathsf{orc}}) \right\|^2 \overset{\mathsf{as}}{\to} \delta\tau^2 - \sigma^2$. Thus, the second identity of (D.2) hold almost surely.

We now show that almost surely, the third identity (D.2) for all $\tau' \geq 0$. Recall by strong stationarity that the limit (B.5) holds for $\boldsymbol{T}_{\tau'} = \begin{pmatrix} \tau_{\mathsf{orc}}^2 & 0 \\ 0 & \tau'^2 \end{pmatrix}$ for all $\tau' \geq 0$. Now fix a particular $\tau' \geq 0$. Let $h_p(\boldsymbol{x}; \tau') = \mathbb{E}_{\boldsymbol{z}}[\langle f_p(\boldsymbol{x}; \tau_{\mathsf{orc}}), f_p(\boldsymbol{z}; \tau')\rangle]$, where $\boldsymbol{z} \sim \mathsf{N}(\boldsymbol{0}, \boldsymbol{I}_p/p)$. By Cauchy-Schwartz and because f_p is Lipschitz,

$$\left| h_p(\boldsymbol{x}_1; \tau') - h_p(\boldsymbol{x}_2; \tau') \right| \leq \tau_{\mathsf{orc}} \mathbb{E}_{\boldsymbol{z}}\left[\left\| f_p(\boldsymbol{z}; \tau') \right\| \right] \|\boldsymbol{x}_1 - \boldsymbol{x}_2\|. \tag{D.7}$$

By (D.6) and (O.4), we have $\mathbb{E}_{\boldsymbol{z}}[\|f_p(\boldsymbol{z}; \tau')\|] \leq \mathbb{E}_{\boldsymbol{z}}[\|f_p(\boldsymbol{z}; \tau_{\mathsf{orc}})\|] + |\tau_{\mathsf{orc}} - \tau'| \mathbb{E}_{\boldsymbol{z}}[\|\boldsymbol{z}\|] \leq \mathbb{E}_{\boldsymbol{z}}[\|f_p(\boldsymbol{z}; \tau_{\mathsf{orc}})\|] + |\tau_{\mathsf{orc}} - \tau'|$ is bounded in p. Thus, for a fixed τ', inequality (D.7) gives that $h_p(\cdot; \tau')$ is uniformly (over p) Lipschitz. Then, applying Lemma P.8 and Borel-Cantelli in the same way we did to establish (D.5), we have (using also (B.4c) and condition (B.5) of strong stationarity)

$$h_p(\boldsymbol{z}^0; \tau') \overset{\mathsf{as}}{\simeq} \mathbb{E}_{\boldsymbol{z}^0}[h_p(\boldsymbol{z}^0)] = \mathsf{K}_{\mathsf{reg,cvx}}(\boldsymbol{T}', \lambda, \mathcal{T}_p) \to \mathsf{K}^{\infty}_{\mathsf{reg,cvx}}(\boldsymbol{T}_{\tau'}, \lambda, \mathcal{T}). \tag{D.8}$$

This establishes the results for fixed $\tau' > 0$. We may extend to all of $\mathbb{R}_+$ by considering a countable dense subset of $\mathbb{R}_+$ and using continuity of h_p in τ'.

In summary, we have proved that if $\boldsymbol{z}^0 \sim \mathsf{N}(0, \boldsymbol{I}_p/p)$ for all p, then almost surely the limits (D.2) hold simultaneously for all $\tau' \geq 0$. Therefore, we may choose a deterministic sequence $\{\boldsymbol{z}^0\}$ such that these limits all hold.

D.3 The Approximate Message Passing (AMP) iteration

Let $\{\boldsymbol{z}^0\}$ be a deterministic sequence of vectors satisfying limits (D.2) for all $\tau' \geq 0$, as permitted by the previous section. For each p, define the sequence $\{\widehat{\boldsymbol{\beta}}^t\}_{t\geq 0}$ via the following iteration. Define

$$\mathsf{b} = 1 - \frac{1}{2\lambda}, \tag{D.9}$$

and for $t \geq 0$

$$\boldsymbol{r}^t = \frac{\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}^t}{n} + \mathsf{b}\boldsymbol{r}^{t-1}, \tag{D.10a}$$

$$\widehat{\boldsymbol{\beta}}^{t+1} = \mathsf{prox}[\lambda\rho_p^{(\gamma)}]\left(\widehat{\boldsymbol{\beta}}^t + \boldsymbol{X}^\mathsf{T}\boldsymbol{r}^t\right), \tag{D.10b}$$

$$\widehat{\boldsymbol{\beta}}^0 = \mathsf{prox}[\lambda\rho_p^{(\gamma)}]\left(\boldsymbol{\beta}_0 + \tau\boldsymbol{z}^0\right) \quad \text{and} \quad \boldsymbol{r}^{-1} = \boldsymbol{0}. \tag{D.10c}$$

This iteration is an approximate message passing (AMP) algorithm. Several papers (see [BMN19] and references therein) precisely characterize the iterates of such algorithms in the $p \to \infty$ limit, as we will see in Appendix D.5 below. Further, they satisfy certain identities relating them to ρ_p and L. First, by (O.1) and (D.10b), we have for $t \geq 0$ that

$$\widehat{\boldsymbol{\beta}}^t + \boldsymbol{X}^\mathsf{T}\boldsymbol{r}^t - \widehat{\boldsymbol{\beta}}^{t+1} \in \lambda\partial\rho_p^{(\gamma)}\left(\widehat{\boldsymbol{\beta}}^{t+1}\right). \tag{D.11}$$

Second, by (D.10a),

$$\nabla_{\boldsymbol{\beta}}\left(\frac{1}{n}\|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\|^2\right)\Big|_{\boldsymbol{\beta}=\widehat{\boldsymbol{\beta}}^{t+1}} = \frac{2}{n}\boldsymbol{X}^\mathsf{T}\left(\boldsymbol{X}\widehat{\boldsymbol{\beta}}^{t+1} - \boldsymbol{y}\right) = 2\boldsymbol{X}^\mathsf{T}\left(\mathsf{b}\boldsymbol{r}^t - \boldsymbol{r}^{t+1}\right). \tag{D.12}$$

Combining (D.11) and (D.12) with (D.1),

$$\begin{aligned}\partial L\left(\widehat{\boldsymbol{\beta}}^{t+1}\right) &\ni 2\boldsymbol{X}^\mathsf{T}\left(\mathsf{b}\boldsymbol{r}^t - \boldsymbol{r}^{t+1}\right) + \frac{\widehat{\boldsymbol{\beta}}^t + \boldsymbol{X}^\mathsf{T}\boldsymbol{r}^t - \widehat{\boldsymbol{\beta}}^{t+1}}{\lambda} \\ &= 2\mathsf{b}\boldsymbol{X}^\mathsf{T}\left(\boldsymbol{r}^t - \boldsymbol{r}^{t+1}\right) + \frac{(\widehat{\boldsymbol{\beta}}^t + \boldsymbol{X}^\mathsf{T}\boldsymbol{r}^t) - (\widehat{\boldsymbol{\beta}}^{t+1} + \boldsymbol{X}^\mathsf{T}\boldsymbol{r}^{t+1})}{\lambda},\end{aligned} \tag{D.13}$$

where in the equality we have used that $\frac{1}{\lambda} = 2(1-\mathsf{b})$. If L is κ-strong convex for some $\kappa > 0$ and $\boldsymbol{g} \in \partial L\left(\widehat{\boldsymbol{\beta}}^{t+1}\right)$, then for any $\boldsymbol{\beta} \in \mathbb{R}^p$

$$L(\boldsymbol{\beta}) \geq L\left(\widehat{\boldsymbol{\beta}}^{t+1}\right) + \left\langle \boldsymbol{g}, \boldsymbol{\beta} - \widehat{\boldsymbol{\beta}}^{t+1}\right\rangle + \frac{\kappa}{2}\left\|\boldsymbol{\beta} - \widehat{\boldsymbol{\beta}}^{t+1}\right\|^2 \geq L\left(\widehat{\boldsymbol{\beta}}^{t+1}\right) - \|\boldsymbol{g}\|\left\|\boldsymbol{\beta} - \widehat{\boldsymbol{\beta}}^{t+1}\right\| + \frac{\kappa}{2}\left\|\boldsymbol{\beta} - \widehat{\boldsymbol{\beta}}^{t+1}\right\|^2.$$

Because $L(\widehat{\boldsymbol{\beta}}_{\mathsf{cvx}}) \leq L(\widehat{\boldsymbol{\beta}}^{t+1})$ by (1.2), we have

$$\left\|\widehat{\boldsymbol{\beta}}_{\mathsf{cvx}} - \widehat{\boldsymbol{\beta}}^{t+1}\right\| \leq \frac{2\|\boldsymbol{g}\|}{\kappa}. \tag{D.14}$$

Combining (D.13) and (D.14), we have that if L is κ-strongly convex, then

$$\left\|\widehat{\boldsymbol{\beta}}_{\mathsf{cvx}} - \widehat{\boldsymbol{\beta}}^{t+1}\right\| \leq \frac{2}{\kappa}\left(2\mathsf{b}\left\|\boldsymbol{X}^\mathsf{T}\right\|_{\mathsf{op}}\left\|\boldsymbol{r}^t - \boldsymbol{r}^{t+1}\right\| + \frac{\left\|(\widehat{\boldsymbol{\beta}}^t + \boldsymbol{X}^\mathsf{T}\boldsymbol{r}^t) - (\widehat{\boldsymbol{\beta}}^{t+1} + \boldsymbol{X}^\mathsf{T}\boldsymbol{r}^{t+1})\right\|}{\lambda}\right). \tag{D.15}$$

Inequality (D.15) allows us to control the distance of the iterates $\widehat{\boldsymbol{\beta}}^t$ from the minimizer $\widehat{\boldsymbol{\beta}}_{\mathsf{cvx}}$ of L in terms of the rate at which the iterates are changing and the strong convexity parameter of L. We will control this distance in the $p \to \infty$, fixed t asymptotic regime by controlling the terms on the right-hand side of (D.15).

D.4 The state evolution

We now study a certain scalar iteration which, in the following sections, will allow us to characterize the $p \to \infty$, fixed t behavior of the AMP iteration (D.10). For $q \in [0,1]$, define $\boldsymbol{Q}_q = \begin{pmatrix} \tau_{\mathsf{orc}}^2 & q\tau_{\mathsf{orc}}^2 \\ q\tau_{\mathsf{orc}}^2 & \tau_{\mathsf{orc}}^2 \end{pmatrix}$, and observe that $\boldsymbol{Q}_q \succeq \boldsymbol{0}$. Define $\Psi : [0,1] \to \mathbb{R}$ by

$$\Psi(q) = \frac{1}{\delta\tau^2}\left(\sigma^2 + \mathsf{K}_{\mathsf{reg,cvx}}^{\infty}\left(\boldsymbol{Q}_q, \lambda_{\mathsf{orc}}, \mathcal{T}\right)\right). \tag{D.16}$$

Because $\tau, \lambda, \gamma, \delta, \mathcal{T}$ is strongly stationary, $\Psi(q)$ is well-defined for all $q \in [0,1]$ (recall strong stationarity requires the limit (B.5) exist for all $\boldsymbol{T} \in S_+^2$). Define the doubly-infinite symmetric matrix $Q = (q_{ij})_{i,j=1}^{\infty}$ via the following scalar iteration, referred to as the *state evolution*:

$$q_{1,1} = 1,\ q_{1,i} = q_{i,1} = 0 \text{ for } i > 1, \tag{D.17a}$$

$$q_{s+1,t+1} = \Psi(q_{s,t}). \tag{D.17b}$$

In order for (D.17b) to make sense, we must verify that $q_{s,t} \in [0,1]$ for all $s, t \geq 1$. By induction, it will suffice to show that $\Psi(q) \in [0,1]$ for all $q \in [0,1]$. In preparation for what is to come later in the proof, we will in fact show more.

Lemma D.1. *For any sequence of symmetric convex function ρ_p,*

$$\Psi(1) = 1, \tag{D.18a}$$

$$\Psi(q) \text{ is non-decreasing and convex for } q \in [0,1], \tag{D.18b}$$

$$\Psi(q) \geq 1 - \frac{1}{(\lambda\gamma+1)\vee\delta}(1-q) \text{ for } q \in [0,1]. \tag{D.18c}$$

Proof of properties (D.18a), (D.18b) of Lemma D.1. By (B.4a), (B.4c), and (B.5),

$$\mathsf{K}_{\mathsf{reg,cvx}}^{\infty}\left(\tau_{\mathsf{orc}}^2 \boldsymbol{I}_2, \lambda_{\mathsf{orc}}, \mathcal{T}\right) = \mathsf{R}_{\mathsf{reg,cvx}}^{\infty}(\tau_{\mathsf{orc}}, \lambda_{\mathsf{orc}}, \mathcal{T}). \tag{D.19}$$

Because $\tau, \lambda, \gamma, \delta, \mathcal{T}$ is strongly stationary, (B.7), (D.16), and (D.19) imply (D.18a).

For each p, define $\Psi_p : [0,1] \to \mathbb{R}$ by

$$\Psi_p(q) = \frac{1}{\delta\tau^2}\left(\sigma^2 + \mathsf{K}_{\mathsf{reg,cvx}}\left(\boldsymbol{Q}_q, \lambda_{\mathsf{orc}}, \mathcal{T}_p\right)\right), \tag{D.20}$$

where $\mathcal{T}_p$ is related to $\mathcal{T}$ in the obvious way. By strong stationarity condition (B.5), $\Psi(q) = \lim_{p\to\infty}\Psi_p(q)$ for every $q \in [0,1]$. It is straightforward to see that properties (D.18b), (D.18c) will hold if we can establish

$$\text{for all } p,\ \Psi_p(q) \text{ is increasing in } q \text{ for } q \in [0,1], \tag{D.21a}$$

$$\text{for all } p,\ \Psi_p(q) \text{ is convex in } q \text{ for } q \in [0,1], \tag{D.21b}$$

$$\limsup_{p\to\infty} \Psi_p'(1) \leq \frac{1}{(\lambda\gamma+1)\vee\delta}. \tag{D.21c}$$

To show (D.21a), (D.21b) we extend the argument of [DM16, Lemma 6.9] and [BM12, Lemma C.1] to multivariate maps. Fix p. Define $f : \mathbb{R}^p \to \mathbb{R}^p$, $\boldsymbol{x} \mapsto \frac{1}{\sqrt{\delta}\tau}\left(\mathsf{prox}\left[\lambda_{\mathsf{orc}}\rho_p\right](\boldsymbol{\beta}_0 + \tau_{\mathsf{orc}}\boldsymbol{x}) - \boldsymbol{\beta}_0\right)$.

Let $\{\boldsymbol{z}_t\}_{t\geq 0}$ be the p-dimensional Ornstein-Uhlenbeck process with mean $\mathbf{0}$ and covariance $\mathbb{E}\left[\boldsymbol{z}_s \boldsymbol{z}_t^{\mathsf{T}}\right] = e^{-|t-s|}\boldsymbol{I}_p/p$. By (D.20) and (B.4c), we may write $\Psi_p(q) = \frac{\sigma^2}{\delta\tau^2} + \mathbb{E}_{\boldsymbol{z}_0,\boldsymbol{z}_t}[\langle f(\boldsymbol{z}_0), f(\boldsymbol{z}_t)\rangle]$ for $t = \log(1/q)$. Denoting the i^{th} component of f by f_i, we have the spectral representation

$$f_i(\boldsymbol{x}) = \sum_{\boldsymbol{k}\in\mathbb{Z}^p_{\geq 0}} c_{i\boldsymbol{k}} \prod_{j=1}^p \phi_{k_j}(x_j),$$

where for each $k \geq 0$ we have ϕ_k is the eigenvector of the generator of the univariate Ornstein-Uhlenbeck process corresponding to eigenvalue k [Gar85, p. 134]. The equality is in L_2 with respect to base measure $\left(\frac{p}{2\pi}\right)^{p/2} e^{-\frac{p}{2}\|\boldsymbol{x}\|^2} d\boldsymbol{x}$. Then,

$$\begin{aligned}
\Psi_p(q) &= \sum_{i=1}^p \mathbb{E}_{\boldsymbol{z}_0,\boldsymbol{z}_t}[f_i(\boldsymbol{z}_0) f_i(\boldsymbol{z}_t)] = \sum_{i=1}^p \mathbb{E}_{\boldsymbol{z}_0}[f_i(\boldsymbol{z}_0)\mathbb{E}_{\boldsymbol{z}_0,\boldsymbol{z}_t}[f_i(\boldsymbol{z}_t)|\boldsymbol{z}_0]] \\
&= \sum_{\boldsymbol{k}\in\mathbb{Z}^p_{\geq 0}} c_{i\boldsymbol{k}} \mathbb{E}_{\boldsymbol{z}_0}\left[f_i(\boldsymbol{z}_0)\prod_{j=1}^p \phi_{k_j}(z_{0j}) e^{-k_j t}\right] = \sum_{i=1}^p \sum_{\boldsymbol{k}\in\mathbb{Z}^p_{\geq 0}} c_{i\boldsymbol{k}}^2 e^{-\left(\sum_{j=1}^p k_j\right)t} \\
&= \sum_{i=1}^p \sum_{\boldsymbol{k}\in\mathbb{Z}^p_{\geq 0}} c_{i\boldsymbol{k}}^2 q^{\sum_{j=1}^p k_j},
\end{aligned}$$

whence (D.21a), (D.21b) follow. □

The proof of property (D.18c) of Lemma D.1 requires the following technical lemma.

Lemma D.2. *If $h : \mathbb{R}^p \to \mathbb{R}$ is Lipschitz and $\boldsymbol{z}_1, \boldsymbol{z}_2 \sim \mathsf{N}(\mathbf{0}, \boldsymbol{I}_p/p)$ are independent, then*

$$\frac{\mathrm{d}}{\mathrm{d}q}\mathbb{E}_{\boldsymbol{z}_1,\boldsymbol{z}_2}\left[h(\boldsymbol{z}_1)\, h\left(q\,\boldsymbol{z}_1 + \sqrt{1-q^2}\boldsymbol{z}_2\right)\right]\Big|_{q=1} = -\frac{1}{p}\sum_{j=1}^p \mathbb{E}\left[(\partial_j h(\boldsymbol{z}))^2\right],$$

(where the derivative on the left-hand side is a left-derivative, and the derivatives on the right-hand side exist almost everywhere by [EG15, Theorem 3.2]).

Proof. This is an elementary fact, so we only sketch the proof idea. Denote by $F(q)$ the expectation on the left hand side, and $\boldsymbol{g}_1 := (\boldsymbol{z}_1 + (q\,\boldsymbol{z}_1 + \sqrt{1-q^2}\boldsymbol{z}_2))/2$, $\boldsymbol{g}_2 = (\boldsymbol{z}_1 - (q\,\boldsymbol{z}_1 + \sqrt{1-q^2}\boldsymbol{z}_2))/2$ assuming $\nabla^2 h$ bounded, Taylor's expansion implies

$$2(F(0) - F(q)) = 4\,\mathbb{E}\{\langle \nabla h(\boldsymbol{g}_1), \boldsymbol{g}_2\rangle^2\} + O(\mathbb{E}\{\|\boldsymbol{g}_2\|_2^4\}) = \frac{2}{p}\mathbb{E}\{\|\nabla h(\boldsymbol{g}_1)\|^2\}(1-q) + O((1-q)^2)\,.$$

and the claim follows by dominated convergence. This is extended to general Lipschitz h by a routine approximation argument. □

We are now ready to prove property (D.18c) of Lemma D.1.

Proof of property (D.18c) *of Lemma D.1.* As in the proof of properties (D.18a), (D.18b), define $f : \mathbb{R}^p \to \mathbb{R}^p$ the function which maps $\boldsymbol{x} \mapsto \frac{1}{\sqrt{\delta}\tau}\left(\mathsf{prox}\left[\lambda\rho_p\right](\boldsymbol{\beta}_0 + \tau\boldsymbol{x}) - \boldsymbol{\beta}_0\right)$ and let f_i be its ith coordinate. Applying Lemma D.2 to $h = f_i$ and summing over i, we have

$$\begin{aligned}
\Psi_p'(1) &= \frac{1}{p(\lambda\gamma+1)^2}\mathbb{E}_{\boldsymbol{z}}[\|\mathrm{D}f(\boldsymbol{z})\|_{\mathsf{F}}^2] = \frac{1}{\delta p(\lambda\gamma+1)^2}\mathbb{E}_{\boldsymbol{z}}\left[\|\mathrm{D}\,\mathsf{prox}\left[\lambda_{\mathsf{orc}}\rho_p\right](\boldsymbol{\beta}_0 + \tau_{\mathsf{orc}}\boldsymbol{z})]\|_{\mathsf{F}}^2\right] \\
&\le \frac{1}{\delta p(\lambda\gamma+1)^2}\mathbb{E}_{\boldsymbol{z}}\left[\|\mathrm{D}\,\mathsf{prox}\left[\lambda_{\mathsf{orc}}\rho_p\right](\boldsymbol{\beta}_0 + \tau_{\mathsf{orc}}\boldsymbol{z})\|_{\mathsf{op}}\|\mathrm{D}\,\mathsf{prox}\left[\lambda_{\mathsf{orc}}\rho_p\right](\boldsymbol{\beta}_0 + \tau_{\mathsf{orc}}\boldsymbol{z})\|_{\mathsf{nuc}}\right] \\
&\le \frac{1}{(\lambda\gamma+1)^2}\frac{1}{\delta p}\mathbb{E}_{\boldsymbol{z}}\left[\|\mathrm{D}\,\mathsf{prox}\left[\lambda_{\mathsf{orc}}\rho_p\right](\boldsymbol{\beta}_0 + \tau_{\mathsf{orc}}\boldsymbol{z})\|_{\mathsf{nuc}}\right] \\
&= \frac{1}{(\lambda\gamma+1)^2}\frac{1}{\delta p}\mathbb{E}_{\boldsymbol{z}}\left[\mathsf{div}\,\mathsf{prox}\left[\lambda_{\mathsf{orc}}\rho_p\right](\boldsymbol{\beta}_0 + \tau_{\mathsf{orc}}\boldsymbol{z})\right] \\
&= \frac{1}{(\lambda\gamma+1)^2}\cdot\frac{1}{\delta}\mathsf{W}_{\mathsf{reg,cvx}}(\tau_{\mathsf{orc}}, \lambda_{\mathsf{orc}}, \mathcal{T}_p). \qquad \text{(D.22)}
\end{aligned}$$

In the first equality, we have used that $\tau_{\mathsf{orc}}^2/\tau^2 = 1/(\lambda\gamma+1)^2$. In the first inequality, we have used that the operator and nuclear norms are dual with respect to the matrix inner product $\langle \boldsymbol{A}, \boldsymbol{B}\rangle = \mathrm{Tr}(\boldsymbol{A}^\mathsf{T}\boldsymbol{B})$, which induces the Frobenius norm. In the second inequality, we have applied (O.8). In the second-to-last line we have used that $\|\mathrm{D}\,\mathsf{prox}[\lambda\rho_p](\boldsymbol{\beta}_0+\tau\boldsymbol{z})\|_{\mathsf{nuc}} = \mathsf{div}\,\mathsf{prox}[\lambda\rho_p](\boldsymbol{\beta}_0+\tau\boldsymbol{z})$ because all eigenvalues of $\mathrm{D}\,\mathsf{prox}[\lambda\rho_p](\boldsymbol{\beta}_0 + \tau\boldsymbol{z})$ are non-negative by (O.9). In the last equality, we have used (B.4b) and (O.11). Because $\tau, \lambda, \gamma, \delta, \mathcal{T}$ is a strongly stationary quadruplet, by (B.7) we have $\lim_{p\to\infty}\frac{1}{\delta(\lambda\gamma+1)}\mathsf{W}_{\mathsf{reg,cvx}}(\tau_{\mathsf{orc}}, \lambda_{\mathsf{orc}}, \mathcal{T}_p) < 1$, whence $\limsup_{p\to\infty}\Psi_p'(1) \le \frac{1}{\lambda\gamma+1}$. Further, by (B.4b) and (O.12), we have $\mathsf{W}_{\mathsf{reg,cvx}}(\tau_{\mathsf{orc}}, \lambda_{\mathsf{orc}}, \mathcal{T}) \le 1$, whence we also have $\limsup_{p\to\infty}\Psi_p'(1) \le \frac{1}{\delta}$. Inequality (D.21c) follows. □

We are ready to verify that the recursion (D.16), (D.17) makes sense and establish some of its properties. By (D.18a) and (D.18b), we have $\Psi(q) \le 1$ for $q \in [0,1]$, and by (D.18c), we have $\Psi(q) \ge 0$ for $q \in [0,1]$. Then, inductively we have $q_{s,t} \in [0,1]$ for all s, t, so that (D.17) makes sense. Further, by (D.18c), we have for all $t \ge 1$ that $1 - q_{t+1,t+2} = 1 - \Psi(q_{t,t+1}) \le \frac{1}{(\lambda\gamma+1)\vee\delta}(1 - q_{t,t+1})$, so that inductively, with base case $1 - q_{1,2} = 1$, we have

$$1 - q_{t,t+1} \le \left(\frac{1}{(\lambda\gamma+1)\vee\delta}\right)^{t-1}. \qquad \text{(D.23)}$$

If either $\lambda\gamma > 0$ or $\delta > 1$, we have

$$q_{t,t+1} \xrightarrow[t\to\infty]{} 1. \qquad \text{(D.24)}$$

Further, by (D.17) and (D.18a), we get for all $t \ge 1$,

$$q_{t,t} = 1. \qquad \text{(D.25)}$$

D.5 Relating AMP and state evolution

We will show that for $t \ge 2$,

$$\operatorname*{p\,lim}_{p\to\infty}\sqrt{n}\left\|\boldsymbol{r}^t - \boldsymbol{r}^{t+1}\right\| = \sqrt{2(1 - q_{t+1,t+2})}\tau, \qquad \text{(D.26a)}$$

$$\operatorname*{p\,lim}_{p\to\infty}\left\|(\widehat{\boldsymbol{\beta}}^t + \boldsymbol{X}^\mathsf{T}\boldsymbol{r}^t) - (\widehat{\boldsymbol{\beta}}^{t+1} + \boldsymbol{X}^\mathsf{T}\boldsymbol{r}^{t+1})\right\| = \sqrt{2(1 - q_{t+1,t+2})}\tau. \qquad \text{(D.26b)}$$

These identities are a consequence of the characterization of the AMP iteration proved in [BMN19], as we now describe. The authors of [BMN19] study a more general AMP iteration given by

$$\boldsymbol{v}^t = \frac{1}{\sqrt{n}}\boldsymbol{X}e_t(\boldsymbol{u}^t) - \hat{\mathsf{b}}_t g_{t-1}(\boldsymbol{v}^{t-1}), \qquad \boldsymbol{u}^{t+1} = \frac{1}{\sqrt{n}}\boldsymbol{X}^{\mathsf{T}} g_t(\boldsymbol{v}^t) - \hat{\mathsf{d}}_t e_t(\boldsymbol{u}^t), \tag{D.27}$$

with initialization given by deterministic vector

$$\boldsymbol{u}^0 \in \mathbb{R}^p \quad \text{and} \quad g_{-1}(\cdot) = \boldsymbol{0}, \tag{D.28}$$

and for each $t \geq 0$ the functions $e_t : \mathbb{R}^p \to \mathbb{R}$ and $g_t : \mathbb{R}^n \to \mathbb{R}^n$ are uniformly (in p) pseudo-Lipschitz of order 1 (a.k.a. uniformly Lipschitz). In [BMN19], iteration (D.27) is written with respect to a random matrix $\boldsymbol{A}$ with entries $A_{ij} \overset{\text{iid}}{\sim} \mathsf{N}(0, 1/n)$. We have replaced this with $\boldsymbol{X}/\sqrt{n}$, which is distributed in this way. Theorem 1 and Corollary 2 of [BMN19] give that certain functionals of the iterates in (D.27) converge in probability to deterministic constants given by a scalar iteration called state evolution, of which the iteration in Appendix D.4 is, as we will see, a special case. In particular, we will show that iteration (D.10) is a special case of the iteration (D.27), the scalar recursion (D.16) and (D.17) is the corresponding state evolution, and the limits (D.26) are the result of Theorem 1 and Corollary 2 of [BMN19] applied to particular functions.[2] To avoid confusion with corollaries which appear in this paper, we will refer to Corollary 2 of [BMN19] as Corollary SE.

Iteration (D.10) is equivalent to iteration (D.27) under the following change of variables.

$$\begin{aligned} \boldsymbol{v}^t &= \sqrt{p/n}\boldsymbol{w} - \sqrt{np}\boldsymbol{r}^t, & \boldsymbol{u}^{t+1} &= \sqrt{p}\left(\boldsymbol{\beta}_0 - \left(\boldsymbol{X}^{\mathsf{T}}\boldsymbol{r}^t + \widehat{\boldsymbol{\beta}}^t\right)\right), \\ e_t(\boldsymbol{u}) &= \sqrt{p}\big(\mathsf{prox}[\lambda\rho_p^{(\gamma)}]\left(\boldsymbol{\beta}_0 - \boldsymbol{u}/\sqrt{p}\right) - \boldsymbol{\beta}_0\big), \quad t \geq 0, & g_t(\boldsymbol{v}) &= \boldsymbol{v} - \sqrt{p/n}\boldsymbol{w},\ t \geq 0, \\ \boldsymbol{u}^0 &= \sqrt{p}\tau\boldsymbol{z}^0, & \hat{\mathsf{b}}_t &= -\mathsf{b} \quad \text{and} \quad \hat{\mathsf{d}}_t = 1. \end{aligned} \tag{D.29}$$

Due to their different choice of normalization, the authors of [BMN19] use a slightly different notion of a collection of functions' being uniformly pseudo-Lipschitz of order k than used in this paper. In particular, for them a collection of functions $\{\varphi : (\mathbb{R}^p)^\ell \to \mathbb{R}^m\}$, where p and m but not ℓ may vary, is uniformly pseudo-Lipschitz of order k if for all φ and $\boldsymbol{x}_i, \boldsymbol{y}_i \in \mathbb{R}^p$, $i = 1, \ldots, \ell$, we have

$$\frac{\|\varphi(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_\ell) - \varphi(\boldsymbol{y}_1, \ldots, \boldsymbol{y}_\ell)\|}{\sqrt{m}} \leq C\left(1 + \sum_{i=1}^{\ell}\left(\frac{\|\boldsymbol{x}_i\|}{\sqrt{p}}\right)^{k-1} + \sum_{i=1}^{\ell}\left(\frac{\|\boldsymbol{y}_i\|}{\sqrt{p}}\right)^{k-1}\right)\sum_{i=1}^{\ell}\frac{\|\boldsymbol{x}_i - \boldsymbol{y}_i\|}{\sqrt{p}}, \tag{D.30}$$

for some C which does not depend on p, m. We will refer to their notion as [BMN19]-uniformly pseudo-Lipschitz of order k. It is exactly equivalent to our notion under a change of normalization. In particular, the following claim is easy to check.

Claim D.3. *A collection of functions $\{\varphi\}$ is uniformly pseudo-Lipschitz of order k if and only if the collection of functions $\{\tilde{\varphi}\}$ defined by $\tilde{\varphi}(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_\ell) = \sqrt{m}\varphi\left(\boldsymbol{x}_1/\sqrt{p}, \ldots, \boldsymbol{x}_\ell/\sqrt{p}\right)$ is [BMN19]-uniformly pseudo-Lipschitz of order k.*

[2] In fact, most of this task has already been carried out by Theorem 14 of [BMN19]. Unfortunately, Theorem 14 of [BMN19] uses a different initialization than (D.10c) and does not address limits of the form (D.26). Thus, Theorem 14 gives us almost what we need, but not quite. To conclude (D.26), we perform the required change of variables and apply their more general theorem on the iteration (D.27) ourselves.

This will allow us to translate their conditions and results into our normalization. Corollary SE requires six assumptions on the iteration (D.27), which the authors label (B1) - (B6). In our setting, assumption (B1) holds by assumption; assumption (B2) holds by (D.29), (O.4), and inspection; assumption (B3), (B4), and (B5) hold by (D.29), (D.2), and the HDA assumption. Assumption (B6) holds by (D.29), strong stationarity (i.e. definition (B.4c) and the existence of the limit (B.5)) and the proximal operator identity (O.16), the HDA assumption $n/p \to \delta$, and the DSN assumption $\|\boldsymbol{w}\|^2/n \to \sigma^2$.

Finally, the authors of [BMN19] require that

$$\hat{\mathsf{d}}_t \overset{\mathrm{p}}{\simeq} \frac{1}{n}\mathbb{E}_{\boldsymbol{z}}\left[\operatorname{div} g_t(\Sigma_{t,t}\sqrt{n}\boldsymbol{z})\right], \quad \hat{\mathsf{b}}_t \overset{\mathrm{p}}{\simeq} \frac{1}{n}\mathbb{E}_{\boldsymbol{z}}\left[\operatorname{div} e_t(T_{t,t}\sqrt{p}\boldsymbol{z})\right], \tag{D.31}$$

where $\Sigma_{t,t}$ and $T_{t,t}$ are deterministic scalars which we now define. The authors of [BMN19] define the double infinite arrays $(\Sigma_{s,t})_{s,t\geq 0}$ and $(T_{s,t})_{s,t\geq 1}$ through the recursion

$$T_{s+1,t+1} = \lim_{p\to\infty}\frac{1}{n}\mathbb{E}_{\boldsymbol{z}_1,\boldsymbol{z}_2}\left[\left\langle g_s\left(\sqrt{n}\boldsymbol{z}_1\right), g_t\left(\sqrt{n}\boldsymbol{z}_2\right)\right\rangle\right], \tag{D.32a}$$

$$\Sigma_{s,t} = \lim_{p\to\infty}\frac{1}{n}\mathbb{E}_{\boldsymbol{z}_1,\boldsymbol{z}_2}\left[\left\langle e_s\left(\sqrt{p}\boldsymbol{z}_1\right), e_t\left(\sqrt{p}\boldsymbol{z}_2\right)\right\rangle\right], \tag{D.32b}$$

$$\Sigma_{0,0} = \lim_{p\to\infty}\frac{1}{n}\|e_0(\boldsymbol{u}^0)\|^2, \quad \Sigma_{0,i} = \Sigma_{i,0} = 0 \text{ for } i \geq 1, \tag{D.32c}$$

where in (D.32a) we take $(\boldsymbol{z}_1, \boldsymbol{z}_2) \sim \mathsf{N}\left(\boldsymbol{0}, \begin{pmatrix}\Sigma_{s,s} & \Sigma_{s,t}\\ \Sigma_{t,s} & \Sigma_{t,t}\end{pmatrix} \otimes \boldsymbol{I}_n/n\right)$ and in (D.32b) we take $(\boldsymbol{z}_1, \boldsymbol{z}_2) \sim \mathsf{N}\left(\boldsymbol{0}, \begin{pmatrix}T_{s,s} & T_{s,t}\\ T_{t,s} & T_{t,t}\end{pmatrix} \otimes \boldsymbol{I}_p/p\right)$. We claim that for all $s, t \geq 1$,

$$T_{s,t} = q_{s,t}\tau^2. \tag{D.33}$$

We establish this inductively. By (D.29), (D.2), and HDA assumption $n/p \to \delta$, we have

$$\Sigma_{0,0} = \lim_{p\to\infty}\frac{1}{n}\|e_0(\boldsymbol{u}^0)\|^2 = \lim_{p\to\infty}\frac{p}{n}\|\mathsf{prox}[\lambda\rho_p^{(\gamma)}](\boldsymbol{\beta}_0 + \tau\boldsymbol{z}^0) - \boldsymbol{\beta}_0\|^2 = \tau^2 - \frac{1}{\delta}\sigma^2. \tag{D.34}$$

Moreover, for any $s, t \geq 0$, we have by (D.29), (D.32a), the HDA assumption $n/p \to \delta$, and the DSN assumption $\|\boldsymbol{w}\|^2/n \to \sigma^2$, that

$$T_{s+1,t+1} = \lim_{p\to\infty}\frac{1}{n}\mathbb{E}_{\boldsymbol{z}_1,\boldsymbol{z}_2}\left[\left\langle\sqrt{n}\boldsymbol{z}_1 - \sqrt{p/n}\boldsymbol{w}, \sqrt{n}\boldsymbol{z}_2 - \sqrt{p/n}\boldsymbol{w}\right\rangle\right] = \frac{1}{\delta}\sigma^2 + \Sigma_{s,t}. \tag{D.35}$$

By (D.34) and (D.35), we have $T_{1,1} = \tau^2$, the base case. Now assume (D.33) holds for all $1 \leq s, t \leq l$. Fix $1 \leq s, t \leq l$. By (D.29), (D.32b), strong stationarity definition (B.4c) and condition (B.5), and HDA assumption $n/p \to \delta$, we have

$$\begin{aligned} T_{s+1,t+1} &= \frac{1}{\delta}\sigma^2 + \lim_{p\to\infty}\frac{p}{n}\mathbb{E}_{\boldsymbol{z}_1,\boldsymbol{z}_2}\left[\left\langle\mathsf{prox}[\lambda\rho_p^{(\gamma)}](\boldsymbol{\beta}_0 - \boldsymbol{z}_1) - \boldsymbol{\beta}_0, \mathsf{prox}[\lambda\rho_p^{(\gamma)}](\boldsymbol{\beta}_0 - \boldsymbol{z}_2) - \boldsymbol{\beta}_0\right\rangle\right] \\ &= \frac{1}{\delta}\sigma^2 + \frac{1}{\delta}\mathsf{K}^{\infty}_{\mathsf{reg},\mathsf{cvx}}\left(\boldsymbol{Q}_{q_{s,t}}, \lambda_{\mathsf{orc}}, \mathcal{T}\right) = \Psi(q_{s,t})\tau^2 = q_{s+1,t+1}\tau^2, \end{aligned} \tag{D.36}$$

where we have used in the second equality the HDA assumption $n/p \to \delta$, the inductive hypothesis that $(\boldsymbol{z}_1, \boldsymbol{z}_2) \sim \mathsf{N}\left(\boldsymbol{0}, \begin{pmatrix} T_{s,s} & T_{s,t} \\ T_{t,s} & T_{t,t} \end{pmatrix} \otimes \boldsymbol{I}_p/p\right) = \mathsf{N}(\boldsymbol{0}, \boldsymbol{Q}_{s,t} \otimes \boldsymbol{I}_p/p)$, and the oracle proximal identity (O.16); in the third equality, we have used definition (D.16); and in the last equality, we have used (D.17b). This confirms the inductive step. Thus, state evolution (D.16), (D.17b) exactly corresponds to the state evolution (D.32), (D.32c) of [BMN19].

Now we are able to verify assumptions (D.31), which are the final assumptions the authors of [BMN19] require for Corollary SE. By (D.29), we see that $\operatorname{div} g_t = n$, so that again by (D.29) we see the first identity in (D.31) holds with equality even in finite samples. By (D.33) and (D.25), we have $T_{t,t} = \tau^2$ for all $t \geq 1$. The second identity in (D.31) holds because

$$\begin{aligned} \frac{1}{n}\mathbb{E}_{\boldsymbol{z}}\left[(\operatorname{div} e_t)(\tau\sqrt{p}\boldsymbol{z})\right] &= -\frac{1}{n}\mathbb{E}_{\boldsymbol{z}}\left[(\operatorname{div} \mathsf{prox}[\lambda\rho_p^{(\gamma)}])(\boldsymbol{\beta}_0 - \tau\boldsymbol{z})\right] = \frac{p}{\tau_{\mathsf{orc}}(\lambda\gamma+1)n}\mathbb{E}_{\boldsymbol{z}}[\langle \boldsymbol{z}, \mathsf{prox}[\lambda_{\mathsf{orc}}\rho_p](\boldsymbol{\beta}_0 - \tau_{\mathsf{orc}}\boldsymbol{z})\rangle] \\ &\overset{\mathsf{p}}{\simeq} -\frac{1}{\delta(\lambda\gamma+1)}\mathsf{W}^{\infty}_{\mathsf{reg},\mathsf{cvx}}(\tau_{\mathsf{orc}}, \lambda_{\mathsf{orc}}, \mathcal{T}) = \frac{1}{2\lambda} - 1 = -\mathsf{b} = \widehat{\mathsf{b}}^t, \end{aligned} \tag{D.37}$$

where in the first equality we have used (D.29); in the second equality we have used (O.11); and in the fourth equality we have used strong stationarity condition (B.7) and (D.9). The third equality has two distinct justifications, depending on whether we are working under the DSN assumption, or under the RSN assumption conditional on the realization of $\{\boldsymbol{\beta}_0\}$. Under the DSN assumption and if the penalties are symmetric, we have used strong stationarity definition (B.4b), condition (B.5), and Lemma C.4. Under the RSN assumption and if the penalties are not necessarily symmetric, we have instead used Lemma C.3. Having verified (B1) - (B6) and (D.31), we have verified all assumptions required to apply Corollary SE.

Finally, we show that Corollary SE implies (D.26a), (D.26b). The collection of maps $(\mathbb{R}^p)^2 \ni (\boldsymbol{x}, \boldsymbol{x}') \mapsto \frac{\|\boldsymbol{x}-\boldsymbol{x}'\|}{\sqrt{p}}$ is [BMN19]-uniformly pseudo-Lipschitz of order 1. Thus, Corollary SE gives (because we have verified its assumptions) that $\lim_{p\to\infty} \frac{\|\boldsymbol{v}^{t-1}-\boldsymbol{v}^t\|}{\sqrt{p}} = \sqrt{\Sigma_{t-1,t-1} + \Sigma_{t,t} - 2\Sigma_{t-1,t}}$ and $\lim_{p\to\infty} \frac{\|\boldsymbol{u}^{t+1}-\boldsymbol{u}^{t+2}\|}{\sqrt{p}} = \sqrt{T_{t+1,t+1} + T_{t+2,t+2} - 2T_{t+1,t+2}}$ in probability. Under the change of variables (D.29) and using (D.25), (D.33), and (D.35), we get (D.26a), (D.26b).

D.6 Relating AMP and convex optimization

We now complete the proof of parts (ii) and (iii) of Proposition B.3. Observe that by (D.29) and (O.4),

$$\begin{aligned} \frac{\|e_t(\boldsymbol{0})\|^2}{p} &= \|\mathsf{prox}[\lambda\rho_p^{(\gamma)}](\boldsymbol{\beta}_0) - \boldsymbol{\beta}_0\|^2 \leq \left(\|\mathsf{prox}[\lambda\rho_p^{(\gamma)}](\boldsymbol{\beta}_0 - \tau\boldsymbol{z}) - \boldsymbol{\beta}_0\| + \tau\|\boldsymbol{z}\|\right)^2 \\ &\leq 2\|\mathsf{prox}[\lambda\rho_p^{(\gamma)}](\boldsymbol{\beta}_0 - \tau\boldsymbol{z}) - \boldsymbol{\beta}_0\|^2 + 2\tau^2\|\boldsymbol{z}\|^2. \end{aligned} \tag{D.38}$$

Considering $\boldsymbol{z} \sim \mathsf{N}(\boldsymbol{0}, \boldsymbol{I}_p/p)$, taking expectations on both sides, and using (B.4a) and (B.5), we get that $\frac{\|e_t(\boldsymbol{0})\|}{\sqrt{p}}$ is bounded. Moreover, by (O.4) and (D.29), we have $\boldsymbol{x} \mapsto \frac{e_t(\sqrt{p}\boldsymbol{x})}{\sqrt{p}}$ is uniformly (over p) pseudo-Lipschitz of order 1. By these two facts, Lemma P.5 gives that $\boldsymbol{x} \mapsto \frac{\|e_t(\sqrt{p}\boldsymbol{x})\|^2}{p}$ is uniformly pseudo-Lipschitz of order 2. By Claim D.3, we have $\boldsymbol{u} \mapsto \frac{\|e_t(\boldsymbol{u})\|^2}{p}$ is [BMN19]-uniformly pseudo-Lipschitz of order 2. Thus, by (D.10b), (D.29), Corollary SE, oracle proximal identity (O.16), and

strong stationarity condition (B.7), we have

$$\left\|\widehat{\boldsymbol{\beta}}^{t+1}-\boldsymbol{\beta}_0\right\|^2=\frac{\left\|e_t(\boldsymbol{u}^{t+1})\right\|^2}{p}\overset{\mathrm{p}}{\simeq}\mathbb{E}_{\boldsymbol{z}}\left[\frac{\left\|e_t(\tau\sqrt{p}\boldsymbol{z})\right\|^2}{p}\right]\overset{\mathrm{p}}{\simeq}\mathsf{R}^{\infty}_{\mathsf{reg},\mathsf{cvx}}(\tau_{\mathsf{orc}},\lambda_{\mathsf{orc}},\mathcal{T}). \tag{D.39}$$

By the triangle inequality,

$$\begin{aligned}&\left|\left\|\widehat{\boldsymbol{\beta}}_{\mathsf{cvx}}-\boldsymbol{\beta}_0\right\|-\sqrt{\mathsf{R}^{\infty}_{\mathsf{reg},\mathsf{cvx}}(\tau_{\mathsf{orc}},\lambda_{\mathsf{orc}},\mathcal{T})}\right|\\&\qquad\le\left\|\widehat{\boldsymbol{\beta}}^{t+1}-\widehat{\boldsymbol{\beta}}_{\mathsf{cvx}}\right\|+\left|\left\|\widehat{\boldsymbol{\beta}}^{t+1}-\boldsymbol{\beta}_0\right\|-\sqrt{\mathsf{R}^{\infty}_{\mathsf{reg},\mathsf{cvx}}(\tau_{\mathsf{orc}},\lambda_{\mathsf{orc}},\mathcal{T})}\right|.\end{aligned}$$

By (D.39) and (P.1), we have for fixed t

$$\begin{aligned}\limsup_{p\to\infty}^{\mathrm{p}}\left|\left\|\widehat{\boldsymbol{\beta}}_{\mathsf{cvx}}-\boldsymbol{\beta}_0\right\|-\sqrt{\mathsf{R}^{\infty}_{\mathsf{reg},\mathsf{cvx}}(\tau_{\mathsf{orc}},\lambda_{\mathsf{orc}},\mathcal{T})}\right|&\le\limsup_{p\to\infty}^{\mathrm{p}}\left\|\widehat{\boldsymbol{\beta}}^{t+1}-\widehat{\boldsymbol{\beta}}_{\mathsf{cvx}}\right\|\\&\le\limsup_{p\to\infty}^{\mathrm{p}}\frac{2}{\kappa}\left(2\mathsf{b}\left\|\boldsymbol{X}\right\|_{\mathsf{op}}\left\|\boldsymbol{r}^t-\boldsymbol{r}^{t+1}\right\|+\frac{\left\|(\widehat{\boldsymbol{\beta}}^t+\boldsymbol{X}^{\mathsf{T}}\boldsymbol{r}^t)-(\widehat{\boldsymbol{\beta}}^{t+1}+\boldsymbol{X}^{\mathsf{T}}\boldsymbol{r}^{t+1})\right\|}{\lambda}\right),\end{aligned} \tag{D.40}$$

where $\kappa>0$ is such that with probability going to 1 as $p\to\infty$, we have L is κ strongly convex and in the second inequality, we have used (D.15). . Such a κ exists whenever $\delta>0$, $\gamma>0$, or $\{\rho_p\}$ has positive uniform strong convexity parameter. By (D.24), (D.26a), and (P.2), we have

$$\lim_{t\to\infty}\limsup_{p\to\infty}^{\mathrm{p}}\left\|\boldsymbol{X}\right\|_{\mathsf{op}}\left\|\boldsymbol{r}^t-\boldsymbol{r}^{t+1}\right\|\le\lim_{t\to\infty}\limsup_{p\to\infty}^{\mathrm{p}}\frac{\left\|\boldsymbol{X}\right\|_{\mathsf{op}}}{\sqrt{n}}\sqrt{2(1-q_{t+1,t+2})\tau}=0, \tag{D.41}$$

where we have used that $\limsup_{p\to\infty}^{\mathrm{p}}\left\|\boldsymbol{X}\right\|_{\mathsf{op}}/\sqrt{n}<\infty$ (see [Ver12, Theorem 5.31]). Similarly, by (D.24) and (D.26b), we have

$$\lim_{t\to\infty}\lim_{p\to\infty}^{\mathrm{p}}\frac{\left\|(\widehat{\boldsymbol{\beta}}^t+\boldsymbol{X}^{\mathsf{T}}\boldsymbol{r}^t)-(\widehat{\boldsymbol{\beta}}^{t+1}+\boldsymbol{X}^{\mathsf{T}}\boldsymbol{r}^{t+1})\right\|}{\lambda}=\lim_{t\to\infty}\frac{\sqrt{2(1-q_{t+1,t+2})\tau}}{\lambda}=0. \tag{D.42}$$

We conclude

$$\lim_{t\to\infty}\limsup_{p\to\infty}^{\mathrm{p}}\left\|\widehat{\boldsymbol{\beta}}^{t+1}-\widehat{\boldsymbol{\beta}}_{\mathsf{cvx}}\right\|=0, \tag{D.43}$$

whence again by (D.40)

$$\lim_{t\to\infty}\limsup_{p\to\infty}^{\mathrm{p}}\left|\left\|\widehat{\boldsymbol{\beta}}_{\mathsf{cvx}}-\boldsymbol{\beta}_0\right\|-\sqrt{\mathsf{R}^{\infty}_{\mathsf{reg},\mathsf{cvx}}(\tau_{\mathsf{orc}},\lambda_{\mathsf{orc}},\mathcal{T})}\right|=0. \tag{D.44}$$

Thus, (B.9) holds, as desired. This complete the proof of part (ii) of Proposition B.3.

Now we complete the proof of part (iii) of Proposition B.3. Take φ_p as given in part (iii). By the DSN assumption, $\widehat{\pi}_{\boldsymbol{\beta}_0}\overset{\mathrm{W}}{\to}\pi\in\mathcal{P}_2(\mathbb{R})$, we have $\|\boldsymbol{\beta}_0\|$ is bounded (over p). Thus, by Lemmas P.3 and P.5, we have $\psi_p(\boldsymbol{x})=\varphi_p(\boldsymbol{\beta}_0,\boldsymbol{\beta}_0-\boldsymbol{x})$, is uniformly pseudo-Lipschitz of order k. Then by Claim D.3

$$\tilde{\varphi}_p(\boldsymbol{u})=\psi_p(\boldsymbol{u}/\sqrt{p}) \tag{D.45}$$

is [BMN19]-uniformly pseudo-Lipschitz of order k. Corollary SE and (D.29) then gives

$$\varphi_p\left(\boldsymbol{\beta}_0, \widehat{\boldsymbol{\beta}}^t + \boldsymbol{X}^\mathsf{T}\boldsymbol{r}^t\right) = \varphi_p\left(\boldsymbol{\beta}_0, \boldsymbol{\beta}_0 - \boldsymbol{u}^{t+1}/\sqrt{p}\right) = \psi_p\left(\boldsymbol{u}^{t+1}/\sqrt{p}\right) = \tilde{\varphi}_p\left(\boldsymbol{u}^{t+1}\right)$$
$$\overset{\mathrm{p}}{\simeq} \mathbb{E}_{\boldsymbol{z}}\left[\tilde{\varphi}_p(\tau\sqrt{p}\boldsymbol{z})\right] = \mathbb{E}_{\boldsymbol{z}}\left[\varphi_p(\boldsymbol{\beta}_0, \boldsymbol{\beta}_0 - \tau\boldsymbol{z})\right]. \tag{D.46}$$

By (D.10a), we have

$$\frac{\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}_{\mathsf{cvx}}}{(1-\mathsf{b})n} = \frac{\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}^t}{(1-\mathsf{b})n} + \frac{\boldsymbol{X}(\widehat{\boldsymbol{\beta}}^t - \widehat{\boldsymbol{\beta}}_{\mathsf{cvx}})}{(1-\mathsf{b})n} = \boldsymbol{r}^t + \frac{\mathsf{b}}{1-\mathsf{b}}(\boldsymbol{r}^t - \boldsymbol{r}^{t-1}) + \frac{\boldsymbol{X}(\widehat{\boldsymbol{\beta}}^t - \widehat{\boldsymbol{\beta}}_{\mathsf{cvx}})}{(1-\mathsf{b})n}. \tag{D.47}$$

Some algebra and the triangle inequality gives

$$\left\| \underbrace{\left(\widehat{\boldsymbol{\beta}}^t + \boldsymbol{X}^\mathsf{T}\boldsymbol{r}^t\right)}_{:=\boldsymbol{a}^t} - \underbrace{\left(\widehat{\boldsymbol{\beta}}_{\mathsf{cvx}} + \frac{\boldsymbol{X}^\mathsf{T}(\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}_{\mathsf{cvx}})}{(1-\mathsf{b})n}\right)}_{:\boldsymbol{b}^t} \right\| = \left\| \widehat{\boldsymbol{\beta}}^t - \widehat{\boldsymbol{\beta}}_{\mathsf{cvx}} - \boldsymbol{X}^\mathsf{T}\left(\frac{\mathsf{b}}{1-\mathsf{b}}(\boldsymbol{r}^t - \boldsymbol{r}^{t-1}) + \frac{\boldsymbol{X}(\widehat{\boldsymbol{\beta}}^t - \widehat{\boldsymbol{\beta}}_{\mathsf{cvx}})}{(1-\mathsf{b})n}\right)\right|$$
$$\leq \left\|\widehat{\boldsymbol{\beta}}^t - \widehat{\boldsymbol{\beta}}_{\mathsf{cvx}}\right\| + \|\boldsymbol{X}\|_{\mathsf{op}}\left\|\frac{\mathsf{b}}{1-\mathsf{b}}(\boldsymbol{r}^t - \boldsymbol{r}^{t-1}) + \frac{\boldsymbol{X}(\widehat{\boldsymbol{\beta}}^t - \widehat{\boldsymbol{\beta}}_{\mathsf{cvx}})}{(1-\mathsf{b})n}\right\|$$
$$\leq \left(1 + \frac{\|\boldsymbol{X}\|_{\mathsf{op}}^2}{(1-\mathsf{b})n}\right)\left\|\widehat{\boldsymbol{\beta}}^t - \widehat{\boldsymbol{\beta}}_{\mathsf{cvx}}\right\| + \frac{\|\boldsymbol{X}\|_{\mathsf{op}}\mathsf{b}}{1-\mathsf{b}}\|\boldsymbol{r}^t - \boldsymbol{r}^{t-1}\|, \tag{D.48}$$

where we have defined $\boldsymbol{a}^t, \boldsymbol{b}^t$ for future reference. Now combining (D.41), (D.43), and $\limsup_{p\to\infty}^{\mathrm{p}} \|\boldsymbol{X}\|_{\mathsf{op}}/\sqrt{n} < \infty$ (see [Ver12, Theorem 5.31]) using (P.1) and (P.2), we get

$$\lim_{t\to\infty} \limsup_{p\to\infty}^{\mathrm{p}} \|\boldsymbol{a}^t - \boldsymbol{b}^t\| = 0. \tag{D.49}$$

In the remainder of the argument, we let C be a constant which does not depend on p or t but which may change at each appearance. By (D.29), for each t

$$\|\boldsymbol{a}^t\| = \|\boldsymbol{\beta}_0 - \boldsymbol{u}^{t+1}/\sqrt{p}\| \leq \|\boldsymbol{\beta}_0\| + \|\boldsymbol{u}^{t+1}\|/\sqrt{p} \overset{\mathrm{p}}{\simeq} \|\boldsymbol{\beta}_0\| + \tau\mathbb{E}_{\boldsymbol{z}}[\|\boldsymbol{z}\|] \overset{\mathrm{p}}{\simeq} s_2^{1/2}(\pi) + \tau, \tag{D.50}$$

where for each p we let $\boldsymbol{z} \sim \mathsf{N}(\boldsymbol{0}, \boldsymbol{I}_p/p)$, and in the first probabilistic equality we have used Corollary SE and that $\boldsymbol{u} \mapsto \|\boldsymbol{u}\|/\sqrt{p}$ is [BMN19]-uniformly pseudo-Lipschitz of order 1, and in the second probabilistic equality we have used the DSN assumption (2.1). Because $\|\boldsymbol{b}^t\| \leq \|\boldsymbol{a}^t\| + \|\boldsymbol{a}^t - \boldsymbol{b}^t\|$, by (D.49), (D.50), and (P.1), we have for every t that

$$\limsup_{p\to\infty}^{\mathrm{p}} \|\boldsymbol{b}^t\| \leq s_2^{1/2}(\pi) + \tau. \tag{D.51}$$

Combining (D.50) and (D.51) using (P.1), we have for each t

$$\limsup_{p\to\infty}^{\mathrm{p}} \left(1 + \|\boldsymbol{\beta}_0\|^{k-1} + \|\boldsymbol{a}^t\|^{k-1} + \|\boldsymbol{b}^t\|^{k-1}\right) \leq C. \tag{D.52}$$

Because the φ_p are uniformly pseudo-Lipschitz of order k,

$$\lim_{t\to\infty} \limsup_{p\to\infty}^{\mathrm{p}} \left|\varphi_p\left(\boldsymbol{\beta}_0, \boldsymbol{a}^t\right) - \varphi_p\left(\boldsymbol{\beta}_0, \boldsymbol{b}^t\right)\right| \le C \lim_{t\to\infty} \limsup_{p\to\infty}^{\mathrm{p}} \left(1 + \|\boldsymbol{\beta}_0\|^{k-1} + \|\boldsymbol{a}^t\|^{k-1} + \|\boldsymbol{b}^t\|^{k-1}\right) \|\boldsymbol{a}^t - \boldsymbol{b}^t\|$$
$$\le C \lim_{t\to\infty} \limsup_{p\to\infty}^{\mathrm{p}} \|\boldsymbol{a}^t - \boldsymbol{b}^t\| = 0, \tag{D.53}$$

where in the first inequality we have used Definition 1.5, in the second inequality we have used (P.2) and (D.52), and in the equality we have used (D.49). By (D.46), (D.53), and the triangle inequality,

$$\left|\varphi_p(\boldsymbol{\beta}_0, \boldsymbol{b}^t) - \mathbb{E}_{\boldsymbol{z}}[\varphi(\boldsymbol{\beta}_0, \boldsymbol{\beta}_0 - \tau\boldsymbol{z})]\right| \le \left|\varphi_p(\boldsymbol{\beta}_0, \boldsymbol{b}^t) - \varphi_p(\boldsymbol{\beta}_0, \boldsymbol{a}^t)\right| + \left|\varphi_p(\boldsymbol{\beta}_0, \boldsymbol{a}^t) - \mathbb{E}_{\boldsymbol{z}}[\varphi(\boldsymbol{\beta}_0, \boldsymbol{\beta}_0 - \tau\boldsymbol{z})]\right|$$
$$\xrightarrow{\mathrm{p}} 0. \tag{D.54}$$

Plugging in for $\boldsymbol{b}^t$ yields (B.10), as desired.

Thus, we have shown part (iii) and completed the proof of Proposition B.3 □

E Proof of Theorem 1

The main technical challenge is that exact asymptotics for the estimation error of penalized least squares estimators rely on several technical assumptions we would like to avoid. We summarize the main technical hurdles below.

1. $\delta > 1$ **or strong convexity.** One set of technical assumptions under which exact asymptotics can be established in full generality is that either $\delta > 1$ or the penalties are strong-convexity. Proposition B.3 leverages this fact in establishing exact asymptotics for oracle estimators when either $\delta > 1$, $\gamma > 0$, or ρ_p is uniformly strongly convex. To establish our lower bound when $\delta \le 1$ and ρ_p need not be uniformly strongly convex, we construct an oracle estimator with oracle parameter $\gamma > 0$ which improves the estimation error of the original estimator. Its exact characterization then provides a lower bound on the estimation error of the original estimator.

2. **Strong stationarity.** Proposition B.3 requires the limits (B.5) exist and satisfy (B.7) (i.e., strong stationarity), but the δ-bounded width assumption $\{\rho_p\} \in \mathcal{C}_{\delta,\pi}$ does not require that (B.7) be satisfied or even that the limits (B.5) exist. To address this, we establish the existence of limits satisfying (B.7) along certain subsequences of penalties using a compactness argument. This is done in the proof of Lemma E.3.

3. **Solutions to fixed point equations are appropriately bounded.** We must show the oracle estimator does not have risk which is too small. For this we will use the δ-bounded width assumption. In fact, this is the only place the δ-bounded width assumption is used in our argument. This is done in the proof of Lemma E.3. For further discussion of the role of the δ-bounded width assumption, see Appendix I (though our proof does not use that Appendix).

The proof is organized as follows. In Section E.1, we argue that without loss of generality it is enough to consider penalty sequences $\{\rho_p\}$ which satisfy an additional technical assumption. This

technical assumption will be important for the compactness argument mentioned in item 2 above. In Section E.2, we carry out the main technical steps of our proof: defining the oracle estimator and showing that its risk is smaller than the risk of the original estimator but is not much smaller than the convex lower bound (2.12). The proof of the main technical lemma in that section, Lemma E.2, is deferred to Appendix F. It is here that the δ-bounded width assumption plays a role. In Section E.3, we combine the first two parts to finish the proof of the lower bound. In Section E.4, we show the lower bound is tight when $\delta > 1$.

E.1 Penalty sequences which do not shrink towards infinity

The following claim shows that it is enough to prove Theorem 1 under the additional assumption that the penalty sequence does not shrink towards infinity (see (C.5)).

Claim E.1. *To show (2.12) under the conditions of Theorem 1, it is enough to show*

$$\inf_{\{\rho_p\} \in \mathcal{C}_{\delta,\pi} \cap \mathcal{B}} \operatorname*{plim\,inf}_{p\to\infty} \|\widehat{\boldsymbol{\beta}}_{\mathsf{cvx}} - \boldsymbol{\beta}_0\|^2 \geq \delta \tau^2_{\mathsf{reg,cvx}} - \sigma^2, \tag{E.1}$$

under the conditions of Theorem 1. (If $\mathcal{C}_{\delta,\pi} \cap \mathcal{B}$ is empty, we take the infimum to be infinite).

The proof of Claim E.1 is based on the following lemma.

Lemma E.2. *Fix $\pi \in \mathcal{P}_2(\mathbb{R})$, $\delta \in (0,\infty)$, and $\sigma \geq 0$. For a sequence of convex functions $\{\rho_p\}$, we have under there exist constants $c_1 > 0$ and $c_2 \geq 0$ depending only on π, δ, σ such that*

$$\operatorname*{plim\,inf}_{p\to\infty} \|\widehat{\boldsymbol{\beta}}_{\mathsf{cvx}} - \boldsymbol{\beta}_0\|^2 \geq \liminf_{p\to\infty} (c_1 \|\mathsf{prox}[\rho_p](\mathbf{0})\| - c_2)^2. \tag{E.2}$$

(We use the same convention as in Theorem 1 when the minimizing set in (1.2) is empty).

Proof of Lemma E.2. For each p, observe that whenever the minimizing set in (1.2) is non-empty, $\frac{2}{n}\boldsymbol{X}^\mathsf{T}(\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}_{\mathsf{cvx}}) \in \partial\rho_p(\widehat{\boldsymbol{\beta}}_{\mathsf{cvx}})$, whence

$$\begin{aligned}
\frac{1}{2}\|\boldsymbol{\beta}\|^2 + \rho_p(\boldsymbol{\beta}) &\geq \frac{1}{2}\|\boldsymbol{\beta}\|^2 + \rho_p(\widehat{\boldsymbol{\beta}}_{\mathsf{cvx}}) + \frac{2}{n}(\boldsymbol{\beta} - \widehat{\boldsymbol{\beta}}_{\mathsf{cvx}})^\mathsf{T}\boldsymbol{X}^\mathsf{T}(\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}_{\mathsf{cvx}}) \\
&= \frac{1}{2}\|\widehat{\boldsymbol{\beta}}_{\mathsf{cvx}}\|^2 + \langle \widehat{\boldsymbol{\beta}}_{\mathsf{cvx}}, \boldsymbol{\beta} - \widehat{\boldsymbol{\beta}}_{\mathsf{cvx}}\rangle + \frac{1}{2}\|\boldsymbol{\beta} - \widehat{\boldsymbol{\beta}}_{\mathsf{cvx}}\|^2 + \rho_p(\widehat{\boldsymbol{\beta}}_{\mathsf{cvx}}) \\
&\qquad + \frac{2}{n}(\boldsymbol{\beta} - \widehat{\boldsymbol{\beta}}_{\mathsf{cvx}})^\mathsf{T}\boldsymbol{X}^\mathsf{T}(\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}_{\mathsf{cvx}}) \\
&\geq \frac{1}{2}\|\widehat{\boldsymbol{\beta}}_{\mathsf{cvx}}\|^2 + \rho_p(\widehat{\boldsymbol{\beta}}_{\mathsf{cvx}}) + \left(\frac{1}{2}\|\boldsymbol{\beta} - \widehat{\boldsymbol{\beta}}_{\mathsf{cvx}}\| - \|\widehat{\boldsymbol{\beta}}_{\mathsf{cvx}}\| - 2\frac{\|\boldsymbol{X}\|_{\mathsf{op}}}{\sqrt{n}}\frac{\|\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}_{\mathsf{cvx}}\|}{\sqrt{n}}\right)\|\boldsymbol{\beta} - \widehat{\boldsymbol{\beta}}_{\mathsf{cvx}}\| \\
&\geq \frac{1}{2}\|\widehat{\boldsymbol{\beta}}_{\mathsf{cvx}}\|^2 + \rho_p(\widehat{\boldsymbol{\beta}}_{\mathsf{cvx}}) + \left(\frac{1}{2}\|\boldsymbol{\beta}\| - \frac{3}{2}\|\widehat{\boldsymbol{\beta}}_{\mathsf{cvx}}\| - 2\frac{\|\boldsymbol{X}\|_{\mathsf{op}}}{\sqrt{n}}\frac{\|\boldsymbol{y}\|}{\sqrt{n}} - \frac{\|\boldsymbol{X}\|^2_{\mathsf{op}}}{n}\|\widehat{\boldsymbol{\beta}}_{\mathsf{cvx}}\|\right)\|\boldsymbol{\beta} - \widehat{\boldsymbol{\beta}}_{\mathsf{cvx}}\|.
\end{aligned}$$

By (2.4), $\frac{1}{2}\|\mathsf{prox}[\rho_p](\mathbf{0})\|^2 + \rho(\mathsf{prox}[\rho_p](\mathbf{0})) \leq \frac{1}{2}\|\widehat{\boldsymbol{\beta}}_{\mathsf{cvx}}\|^2 + \rho_p(\widehat{\boldsymbol{\beta}}_{\mathsf{cvx}})$. Thus, when evaluating the previous display at $\boldsymbol{\beta} = \mathsf{prox}[\rho_p](\mathbf{0})$, the expression in parentheses on the right-hand side is non-positive. That is,

$$\|\widehat{\boldsymbol{\beta}}_{\mathsf{cvx}}\| \geq \frac{\frac{1}{2}\|\mathsf{prox}[\rho_p](\mathbf{0})\| - 2\frac{\|\boldsymbol{X}\|_{\mathsf{op}}}{\sqrt{n}}\frac{\|\boldsymbol{y}\|}{\sqrt{n}}}{\frac{3}{2} + \frac{\|\boldsymbol{X}\|^2_{\mathsf{op}}}{n}} \geq \frac{\frac{1}{2}\|\mathsf{prox}[\rho_p](\mathbf{0})\| - 2\frac{\|\boldsymbol{X}\|_{\mathsf{op}}}{\sqrt{n}}\frac{\|\boldsymbol{w}\| + \|\boldsymbol{X}\|_{\mathsf{op}}\|\boldsymbol{\beta}_0\|}{\sqrt{n}}}{\frac{3}{2} + \frac{\|\boldsymbol{X}\|^2_{\mathsf{op}}}{n}}. \tag{E.3}$$

By [Ver12, Theorem 5.31] and the HDA assumption,

$$\|\boldsymbol{X}\|_{\mathsf{op}}/\sqrt{n} \xrightarrow{\mathrm{p}} 1 + \sqrt{1/\delta} =: c. \tag{E.4}$$

Then, by the DSN assumption and the Continuous Mapping Theorem, $4\frac{\|\boldsymbol{X}\|_{\mathsf{op}}}{\sqrt{n}}\frac{\|\boldsymbol{w}\|+\|\boldsymbol{X}\|_{\mathsf{op}}\|\boldsymbol{\beta}_0\|}{\sqrt{n}} \xrightarrow{\mathrm{p}} 4c(\sigma + cs_2(\pi))$. Let $c_1 = \frac{1}{3+2c^2}$ and $c_2 = 4cc_1(\sigma + cs_2(\pi)) + s_2(\pi)$. Then by (E.3) and Lemma P.1 from Appendix P whenever the minimizing set in (1.2) is non-empty, and the convention $\|\boldsymbol{\infty} - \boldsymbol{\beta}_0\|^2 = \infty$ otherwise, we have (E.2). □

We now establish Claim E.1.

Proof of Claim E.1. Assume we have shown (E.1) under the conditions of Theorem 1. Now, we assume the conditions of Theorem 1 and show the stronger (2.12).

Let $\{p(\ell)\}$ be the subsequence of $\{p\}$ containing exactly those p for which $(c_1\|\mathsf{prox}[\rho_p](\mathbf{0})\| - c_2)^2 \geq \delta\tau^2_{\mathsf{reg,cvx}} - \sigma^2 + 1$, and let $\{p'(\ell)\}$ its complement, that is, the subsequence of $\{p\}$ containing exactly those p for which $(c_1\|\mathsf{prox}[\rho_p](\mathbf{0})\| - c_2)^2 < \delta\tau^2_{\mathsf{reg,cvx}} - \sigma^2 + 1$. We permit that one of these subsequences be finite. It is straightforward to check that

$$\operatorname*{p\,lim\,inf}_{p\to\infty} \|\widehat{\boldsymbol{\beta}}_{\mathsf{cvx}}(p) - \boldsymbol{\beta}_0\|^2 \geq \min\left\{\operatorname*{p\,lim\,inf}_{\ell\to\infty} \|\widehat{\boldsymbol{\beta}}_{\mathsf{cvx}}(p(\ell)) - \boldsymbol{\beta}_0\|^2, \operatorname*{p\,lim\,inf}_{\ell\to\infty} \|\widehat{\boldsymbol{\beta}}_{\mathsf{cvx}}(p'(\ell)) - \boldsymbol{\beta}_0\|^2\right\}, \tag{E.5}$$

if we adopt the convention that when either of these sequences is finite, the corresponding $\operatorname*{p\,lim\,inf}$ is ∞. We now check that each expression in the minimum on the right-hand side of (E.5) is bounded below by $\delta\tau^2_{\mathsf{reg,cvx}} - \sigma^2$. First, we show that $\operatorname*{p\,lim\,inf}_{\ell\to\infty} \|\widehat{\boldsymbol{\beta}}_{\mathsf{cvx}}(p(\ell)) - \boldsymbol{\beta}_0\|^2 \geq \delta\tau^2_{\mathsf{reg,cvx}} - \sigma^2$. If $\{p(\ell)\}$ is finite, there is nothing to check. If $\{p(\ell)\}$ is infinite, then we apply Lemma E.2. Second, we show that $\operatorname*{p\,lim\,inf}_{\ell\to\infty} \|\widehat{\boldsymbol{\beta}}_{\mathsf{cvx}}(p'(\ell)) - \boldsymbol{\beta}_0\|^2 \geq \delta\tau^2_{\mathsf{reg,cvx}} - \sigma^2$. If $\{p'(\ell)\}$ is finite, there is nothing to check. If $\{p'(\ell)\}$ is infinite, then $\{\rho_{p'(\ell)}\} \in \mathcal{B}$ by construction, and we apply the assumption of the claim. Thus, for all $\{\rho_p\} \in \mathcal{C}_{\delta,\pi}$, the left-hand side of (E.5) is bounded below by (2.12), as desired. □

E.2 Constructing oracles with not-too-small effective noise

The bulk of the proof of Theorem 1 involves constructing a sequence of estimators to which we can apply the exact asymptotics of Proposition B.3 and whose asymptotic estimation error is not too much smaller than that of the original sequence of estimators. To do so, we take a subsequence of the $\{\rho_p\}$ and add a small but non-zero oracle term as in (B.2). The only place in our proof where the δ-bounded width assumption plays a role is in showing that small oracle penalties cannot improve the estimation error too much.

Lemma E.3. *Consider $\pi \in \mathcal{P}_2(\mathbb{R})$, $\delta \in (0,\infty)$, and $\sigma \geq 0$. Consider an increasing sequence of integers $\{p\}$ and a sequence $\{\rho_p\} \in \mathcal{C}_{\delta,\pi} \cap \mathcal{B}$.*

(i) Consider arbitrary δ and sequence $\{\rho_p\} \in \mathcal{C}_{\delta,\pi} \cap \mathcal{B}$. If $\tau_{\mathsf{lb}} > 0$ is such that $\delta\tau^2_{\mathsf{lb}} - \sigma^2 < \mathsf{R}^{\mathsf{opt}}_{\mathsf{seq,cvx}}(\tau_{\mathsf{lb}};\pi)$, then there exists sub-sequence $\{p(\ell)\}$, $\gamma > 0$, $\tau > \tau_{\mathsf{lb}}$, and $\lambda > 0$ such that the following is true: with $\mathcal{T} = (\pi, \{\rho_{p(\ell)}\})$, the quintuplet $\tau, \lambda, \delta, \gamma, \mathcal{T}$ is strongly stationary.

(ii) Consider sequence $\{\rho_p\}$ *and* $\mathcal{T}' = (\pi, \{\rho_p\})$. *If* $\delta > 1$, *and* $\tau \geq 0, \lambda' > 0$ *are such that*

$$\delta\tau^2 - \sigma^2 = \mathsf{R}^{\infty}_{\mathsf{reg,cvx}}(\tau, \lambda', \mathcal{T}'),$$

then there exists sub-sequence $\{p(\ell)\}$ *and* $\lambda > 0$ *such that the following holds: with* $\mathcal{T} = (\pi, \{\lambda'\rho_p/\lambda\})$, *the quintuplet* $\tau, \lambda, \delta, \gamma = 0, \mathcal{T}$ *is strongly stationary.*

Most of the technical machinery of the proof of Theorem 1 is contained in the proof of Lemma E.3. The proof of Lemma E.3 is provided in Appendix F. In part (i), the reader should have in mind taking $\tau_{\mathsf{lb}} \uparrow \tau_{\mathsf{reg,cvx}}$ and ε small, so that we may produce strongly stationary quintuplets with τ not too much smaller than $\tau_{\mathsf{reg,cvx}}$ in the case that $\tau_{\mathsf{reg,cvx}}$ is finite, or diverging in the case that $\tau_{\mathsf{reg,cvx}}$ is infinite. Part (ii) is only used in establishing tightness of the convex lower bound when $\delta > 1$.

E.3 Lower bounding the asymptotic loss

Assume the conditions of Theorem 1. If $\tau_{\mathsf{reg,cvx}} = 0$, then (2.12) is trivial. Thus, assume $\tau_{\mathsf{reg,cvx}} > 0$. We will show that for any $\{\rho_p\} \in \mathcal{C}_{\delta,\pi} \cap \mathcal{B}$ and any $\tau_{\mathsf{lb}} > 0$ such that $\delta\tau_{\mathsf{lb}}^2 - \sigma^2 > \mathsf{R}^{\mathsf{opt}}_{\mathsf{seq,cvx}}(\tau_{\mathsf{lb}}; \pi)$,

$$\lim_{p\to\infty} \mathbb{P}\left(\|\widehat{\boldsymbol{\beta}}_{\mathsf{cvx}} - \boldsymbol{\beta}_0\|^2 < \delta\tau_{\mathsf{lb}}^2 - \sigma^2\right) = 0. \tag{E.6}$$

We then take $\tau_{\mathsf{lb}} \uparrow \tau_{\mathsf{reg,cvx}}$ such that $\delta\tau_{\mathsf{lb}}^2 - \sigma^2 > \mathsf{R}^{\mathsf{opt}}_{\mathsf{seq,cvx}}(\tau_{\mathsf{lb}}; \pi)$ is satisfied along this sequence, which is permitted by the definition of $\tau_{\mathsf{reg,cvx}}$. By Claim E.1, this is enough.

Assume otherwise. Then for some $\xi > 0$, we may pick a subsequence $\{p(\ell)\}$ such that

$$\mathbb{P}\left(\|\widehat{\boldsymbol{\beta}}_{\mathsf{cvx}}(p(\ell)) - \boldsymbol{\beta}_0(p(\ell))\|^2 < \delta\tau_{\mathsf{lb}}^2 - \sigma^2\right) > \xi \tag{E.7}$$

for all ℓ. Observe $\{\rho_{p(\ell)}\} \in \mathcal{C}_{\delta,\pi} \cap \mathcal{B}$ because conditions (2.9) and (C.5) are closed under taking subsequences. By Lemma E.3(i), we may choose $\gamma > 0$, a further subsequence $\{p'(\ell)\}$, $\tau > \tau_{\mathsf{lb}}$, and $\lambda > 0$ such that, with $\mathcal{T} = (\pi, \{\rho^{(\gamma)}_{p'(\ell)}\})$, we have that $\tau, \lambda, \gamma, \delta, \mathcal{T}$ is strongly stationary (here, we have used $\tau_{\mathsf{lb}} > 0$). By Proposition B.3, we have $\|\widehat{\boldsymbol{\beta}}^{(\gamma)}_{\mathsf{orc}}(p'(\ell)) - \boldsymbol{\beta}_0(p'(\ell))\|^2 \xrightarrow[\ell\to\infty]{\mathrm{p}} \delta\tau^2 - \sigma^2$, whence

$$\lim_{\ell\to\infty} \mathbb{P}\left(\|\widehat{\boldsymbol{\beta}}^{(\gamma)}_{\mathsf{orc}}(p'(\ell)) - \boldsymbol{\beta}_0(p'(\ell))\|^2 < \delta\tau_{\mathsf{lb}}^2 - \sigma^2\right) = 0. \tag{E.8}$$

By Lemma B.1, $\|\widehat{\boldsymbol{\beta}}^{(\gamma)}_{\mathsf{orc}}(p'(\ell)) - \boldsymbol{\beta}_0(p'(\ell))\|^2 \leq \|\widehat{\boldsymbol{\beta}}_{\mathsf{cvx}}(p'(\ell)) - \boldsymbol{\beta}_0(p'(\ell))\|^2$ for all ℓ and all realizations of $\boldsymbol{X}$, whence

$$\lim_{\ell\to\infty} \mathbb{P}\left(\|\widehat{\boldsymbol{\beta}}_{\mathsf{cvx}}(p'(\ell)) - \boldsymbol{\beta}_0(p'(\ell))\|^2 < \delta\tau_{\mathsf{lb}}^2 - \sigma^2\right) = 0, \tag{E.9}$$

contradicting (E.7). We conclude (E.6).

E.4 Tightness for $\delta > 1$

Tightness is trivial when both the left and right-hand side of (2.12) is infinite, so we assume $\tau^2_{\mathsf{reg,cvx}} < \infty$. In particular, $\mathsf{R}^{\mathsf{opt}}_{\mathsf{seq,cvx}}(\tau; \pi)$ is finite for some τ.

Then, by Lemma C.2, $\mathsf{R}^{\mathsf{opt}}_{\mathsf{seq,cvx}}(\tau; \pi)$ is finite for all τ and continuous. Thus, by the definition of $\tau^2_{\mathsf{reg,cvx}}$, we have $\delta\tau^2_{\mathsf{reg,cvx}} - \sigma^2 = \mathsf{R}^{\mathsf{opt}}_{\mathsf{seq,cvx}}(\tau_{\mathsf{reg,cvx}}; \pi)$. The infimum in (2.8) can always be achieved

by taking a sequence of $\{\{\rho_p^{(k)}\}_p\}_k$ approaching the infimum, and then taking a sequence $\{\rho_p^{(k(p))}\}_p$ where $k(p)$ goes to infinity appropriately as a function of p. By passing to a subsequence, we can assume that the limit infimum is a limit. Thus, we may assume we have a sequence $\{\rho_p\}_p$ such that

$$\delta\tau^2_{\mathsf{reg,cvx}} - \sigma^2 = \lim_{p\to\infty} \mathbb{E}_{\boldsymbol{\beta}_0,\boldsymbol{z}}\left[\|\mathsf{prox}[\rho_p](\boldsymbol{\beta}_0 + \tau\boldsymbol{z}) - \boldsymbol{\beta}_0\|^2\right] = \mathsf{R}_{\mathsf{reg,cvx}}(\tau_{\mathsf{reg,cvx}}, 1, \mathcal{T}'),$$

where $\mathcal{T}' = (\pi, \{\rho_p\})$. Because $\|\mathsf{prox}[\rho_p](\boldsymbol{\beta}_0 + \tau\boldsymbol{z}) - \boldsymbol{\beta}_0\| \geq \|\mathsf{prox}[\rho_p](\mathbf{0})\| - \tau\|\boldsymbol{z}\| - 2\|\boldsymbol{\beta}_0\|$, we may conclude that $\{\rho_p\} \in \mathcal{B}$. By Lemma E.3(ii), we may find $\lambda > 0$ and a subsequence $\{p(\ell)\}$ such that $\tau_{\mathsf{reg,cvx}}, \lambda, \delta, \gamma = 0, \mathcal{T} = (\pi, \{\rho_{p(\ell)}/\lambda\})$ is strongly stationary. By Proposition B.3, under the penalty sequence $\{\rho_{p(\ell)}/\lambda\}$ we have $\|\widehat{\boldsymbol{\beta}}_{\mathsf{cvx}} - \boldsymbol{\beta}_0\|^2 \xrightarrow{\mathrm{p}} \delta\tau^2_{\mathsf{reg,cvx}} - \sigma^2$.

The proof of Theorem 1 is complete. □

F Proof of Lemma E.3

Lemma E.3 contains most of the technical machinery of our proof. The argument relies on several lemmas, some of whose proofs are deferred to Appendix G.

To guide the reader, we first provide a high-level overview of the argument.

1. *Finite sample version of fixed-point equations.* We begin by defining finite-sample versions of the fixed point equations (B.7) (see (F.1) below). The solutions to (B.7) which we construct will be limits to solutions of (F.1).

2. *Bounds on possible solutions to finite sample fixed-point equations.* By comparing the right and left-hand sides of (F.1a), we place an upper (Lemma F.1) and lower (Lemma F.2) bound on the noise variance τ^2 at a solution to the finite-sample fixed point equations. The lower bound is the only location in our argument where the δ-bounded width assumption is used, corresponding to the statement that the estimation error of the oracle estimator is not too much smaller than the convex lower bound.

3. *Existence of solutions to finite sample fixed-point equations.* Using a topological argument, we show that solutions to the finite sample fixed-point equations must exist (Lemma F.3). Although the Lemma is quite intuitive (having the flavor of a two-dimensional intermediate value theorem), we could not find a statement of the required result in the literature, and the proof is a bit involved.

4. *From finite-sample fixed points to strongly stationary quintuplets.* Using the existence of and bounds on the solutions to (F.1), we apply a compactness arguments to find a subsequence and construct the required strongly stationary quintuplet (see Section F.2).

F.1 Solutions to finite-sample version of fixed point equations

We first consider the following finite-sample versions of the fixed point equations (B.7):

$$\delta\tau^2 - \sigma^2 = \mathsf{R}_{\mathsf{reg,cvx}}(\tau_{\mathsf{orc}}, \lambda_{\mathsf{orc}}, \mathcal{T}_p), \tag{F.1a}$$

$$2\lambda\left(1 - \frac{1}{\delta(\lambda\gamma+1)}\mathsf{W}_{\mathsf{reg,cvx}}(\tau_{\mathsf{orc}}, \lambda_{\mathsf{orc}}, \mathcal{T}_p)\right) = 1. \tag{F.1b}$$

The first step in proving Lemma E.3 is to establish the existence of solutions to these finite-sample equations and to control their size. This is achieved by the following series of lemmas, whose proofs are provided in Appendix G.

Lemma F.1. *Consider an lsc, proper, convex, function $\rho : \mathbb{R}^p \to \mathbb{R} \cup \{\infty\}$. Let $M \geq \|\mathsf{prox}[\rho](\mathbf{0})\|$. Let $\mathcal{T}_p = (\pi, \rho)$. If $\delta > 1$ or $\gamma > 0$ or ρ is κ-strongly convex with $\kappa > 0$, there exists some τ_{max} depending only on $\pi, M, \delta, \gamma, \kappa$ (and not on p) such that if τ, λ is a solution of (F.1b) at γ with $\tau \geq \tau_{\mathsf{max}}$, then*

$$\delta\tau^2 - \sigma^2 > \mathsf{R}_{\mathsf{reg,cvx}}(\tau_{\mathsf{orc}}, \lambda_{\mathsf{orc}}, \mathcal{T}_p). \tag{F.2}$$

The inequality of Lemma F.1 says solutions to (F.1) cannot be too big when either $\delta > 1$, $\gamma > 0$, or ρ is strongly convex. The next lemma establishes –under certain additional restrictions– the reverse inequality at a value of τ which is not too small.

Lemma F.2. *Consider $\{\rho_p\} \in \mathcal{C}_{\delta,\pi}$ and $\tau_{\mathsf{lb}} > 0$ such that $\delta\tau_{\mathsf{lb}}^2 - \sigma^2 < \mathsf{R}^{\mathsf{opt}}_{\mathsf{seq,cvx}}(\tau_{\mathsf{lb}}; \pi)$. For each p, let $\mathcal{T}_p = (\pi, \rho_p)$. Then we can find $\gamma > 0$, $\tau_{\mathsf{min}} \geq \tau_{\mathsf{lb}}$, and a subsequence $\{p(\ell)\}$ such that for all p in the subsequence we have the following: for all λ which solves (F.1b) at $\tau_{\mathsf{min}}, \gamma$,*

$$\delta\tau_{\mathsf{min}}^2 - \sigma^2 < \mathsf{R}_{\mathsf{reg,cvx}}(\tau_{\mathsf{min,orc}}, \lambda_{\mathsf{orc}}, \mathcal{T}_p), \tag{F.3}$$

where $\tau_{\mathsf{min,orc}} = \frac{\tau_{\mathsf{min}}}{\lambda\gamma+1}$ and $\lambda_{\mathsf{orc}} = \frac{\lambda}{\lambda\gamma+1}$.

Combining Lemmas F.1 and F.2, the next lemma allows us to choose an oracle parameter such that, along a subsequence of $\{p\}$, there exist solutions to (F.1) with effective noise parameters τ which are neither too large or too small.

Lemma F.3. *We have the following.*

(i) Assume conditions of Lemma F.2. Assume additionally that $\{\rho_p\} \in \mathcal{B}$. Then we can find $\gamma > 0$, $\tau_{\mathsf{max}} < \infty$, $\lambda_{\mathsf{max}} < \infty$, and a subsequence $\{p(\ell)\}$ such that for all p in the subsequence, there exists solution τ, λ to (F.1) with $(\tau, \lambda) \in [\tau_{\mathsf{lb}}, \tau_{\mathsf{max}}] \times [1/2, \lambda_{\mathsf{max}}]$.

(ii) If $\delta > 1$ or $\{\rho_p\}$ are κ-strongly convex with $\kappa > 0$, part (i) holds except we may also take $\gamma = 0$.

The needed characterization of solutions to (F.1) is complete.

F.2 From finite-sample fixed points to strongly stationary quintuplets

We now apply the lemmas above to prove Lemma E.3.

Proof of Lemma E.3(i). By Lemma F.3(i), we can (and do) choose $\gamma > 0$, $\tau_{\mathsf{max}} < \infty$, $\lambda_{\mathsf{max}} < \infty$, and a subsequence $\{p(\ell)\}$ such that for all p in the subsequence there exists a solution τ_p, λ_p to (F.1) with $(\tau_p, \lambda_p) \in [\tau_{\mathsf{lb}}, \tau_{\mathsf{max}}] \times [1, \lambda_{\mathsf{max}}]$. By Bolzano-Weierstrass, we can find a further subsequence $\{p'(\ell)\}$ and $(\tau, \lambda) \in [\tau_{\mathsf{lb}} - \varepsilon, \tau_{\mathsf{max}}] \times [1, \lambda_{\mathsf{max}}]$ such that

$$(\tau_{p'(\ell)}, \lambda_{p'(\ell)}) \to (\tau, \lambda) \in [\tau_{\mathsf{lb}}, \tau_{\mathsf{max}}] \times [1, \lambda_{\mathsf{max}}]. \tag{F.4}$$

To simplify notation, we write the subsequence as $\{p\}$. By (F.1a) and (F.4), we get

$$\mathsf{R}_{\mathsf{reg,cvx}}(\tau_{p,\mathsf{orc}}, \lambda_{p,\mathsf{orc}}, \mathcal{T}_p) \underset{p\to\infty}{\longrightarrow} \delta\tau^2 - \sigma^2. \tag{F.5}$$

We also have by the definition of τ_{orc} and λ_{orc} and (F.4) that

$$(\tau_{p,\mathsf{orc}}, \lambda_{p,\mathsf{orc}}) \to (\tau_{\mathsf{orc}}, \lambda_{\mathsf{orc}}). \tag{F.6}$$

Because $\{\rho_p\} \in \mathcal{B}$, by Lemma C.5, $\mathsf{R}_{\mathsf{reg,cvx}}(\tau', \lambda', \mathcal{T}_p)$ are uniformly Lipschitz continuous on compact sets. Thus, by (F.5) and (F.6),

$$\mathsf{R}_{\mathsf{reg,cvx}}(\tau_{\mathsf{orc}}, \lambda_{\mathsf{orc}}, \mathcal{T}_p) \underset{p\to\infty}{\longrightarrow} \delta\tau^2 - \sigma^2. \tag{F.7}$$

That is, the limit (B.5) exists at $\tau_{\mathsf{orc}}, \lambda_{\mathsf{orc}}$, and the limiting value solves (B.7a).

Similarly, by (F.1b), for each p we have $\mathsf{W}_{\mathsf{reg,cvx}}(\tau_{p,\mathsf{orc}}, \lambda_{p,\mathsf{orc}}, \mathcal{T}_p) = \delta(\lambda_p\gamma+1)\left(1 - \frac{1}{2\lambda_p}\right)$, so that

$$\mathsf{W}_{\mathsf{reg,cvx}}(\tau_{p,\mathsf{orc}}, \lambda_{p,\mathsf{orc}}, \mathcal{T}_p) \underset{p\to\infty}{\longrightarrow} \delta(\lambda\gamma+1)\left(1 - \frac{1}{2\lambda}\right). \tag{F.8}$$

By Lemma C.5, $\mathsf{W}_{\mathsf{reg,cvx}}(\tau', \lambda', \mathcal{T}_p)$ is uniformly Lipschitz continuous in (τ', λ') on compact sets. Thus, by (F.8) and (F.6),

$$\mathsf{W}_{\mathsf{reg,cvx}}(\tau_{\mathsf{orc}}, \lambda_{\mathsf{orc}}, \mathcal{T}_p) \underset{p\to\infty}{\longrightarrow} \delta(\lambda\gamma+1)\left(1 - \frac{1}{2\lambda}\right). \tag{F.9}$$

That is, the limit (B.5) exists at $\tau_{\mathsf{orc}}, \lambda_{\mathsf{orc}}$, and the limiting value solves (B.7b).

Finally, by Lemma C.5, the functions $\mathsf{R}_{\mathsf{reg,cvx}}(\tau', \lambda', \mathcal{T}_p)$, $\mathsf{W}_{\mathsf{reg,cvx}}(\tau', \lambda', \mathcal{T}_p)$, and $\mathsf{K}_{\mathsf{reg,cvx}}(\boldsymbol{T}', \lambda', \mathcal{T}_p)$ are uniformly equicontinuous in τ', λ', and $\boldsymbol{T}'$ on bounded sets. Further, the convergence (F.7) and (F.9) gives us that $\mathsf{R}_{\mathsf{reg,cvx}}(\tau, \lambda, \mathcal{T}_p)$ and $\mathsf{R}_{\mathsf{reg,cvx}}(\tau, \lambda, \mathcal{T}_p)$ are uniformly bounded over p. Further, for $\boldsymbol{T} = \tau^2\boldsymbol{I}_2$, by (B.4a) and (B.4c), we have $\mathsf{K}_{\mathsf{reg,cvx}}(\boldsymbol{T}, \lambda, \mathcal{T}_p) = \mathsf{R}_{\mathsf{reg,cvx}}(\tau, \lambda, \mathcal{T}_p)$, so that $\mathsf{K}_{\mathsf{reg,cvx}}(\boldsymbol{T}, \lambda, \mathcal{T}_p)$ are uniformly bounded over p. Thus, by the Arzelá-Ascoli theorem, we may take a further subsequence $\{p(\ell)\}$ along which the limits (B.5) exist for all $\tau', \lambda', \boldsymbol{T}'$. We have now established that with $\mathcal{T} = (\pi, \{\rho_{p(\ell)}\})$, the quintuplet $\tau, \lambda, \gamma, \delta, \mathcal{T} = (\pi, \{\rho_{p(\ell)}\})$ is strongly stationary. □

Proof of Lemma E.3(ii). Because $\mathsf{W}_{\mathsf{reg,cvx}}(\tau, \lambda', \mathcal{T}_p') \in [0,1]$ for all p by (C.1) and (C.2), there is a subsequence $\{p(\ell)\}$ such that $\mathsf{W}_{\mathsf{reg,cvx}}(\tau, \lambda', \mathcal{T}_{p(\ell)}')$ converges to a limit w. Let $\lambda = \frac{1}{2(1-w/\delta)}$, so that $1 = 2\lambda(1 - w/\delta)$. Note $\mathsf{R}^{\infty}_{\mathsf{reg,cvx}}(\tau, \lambda, (\pi, \{\lambda'\rho_{p(\ell)}/\lambda\})) = \mathsf{R}^{\infty}_{\mathsf{reg,cvx}}(\tau, \lambda', (\pi, \{\rho_{p(\ell)}\}))$ and $\mathsf{W}^{\infty}_{\mathsf{reg,cvx}}(\tau, \lambda, (\pi, \{\lambda'\rho_{p(\ell)}/\lambda\})) = \mathsf{W}^{\infty}_{\mathsf{reg,cvx}}(\tau, \lambda', (\pi, \{\rho_{p(\ell)}\}))$. Thus, (B.7) are satisfied for $\mathcal{T} = (\pi, \{\lambda'\rho_p/\lambda\})$ at $\tau, \lambda, \gamma = 0, \delta$. Now we may take a further subsequence such that the limits (B.5) exist for all $\tau', \boldsymbol{T}'$ by the same argument used in the proof of Lemma E.3(i). □

G Proofs of Appendix F Lemmas

G.1 Proof of Lemma F.1

We prove Lemma F.1 by controlling the size of $\mathsf{R}_{\mathsf{reg,cvx}}(\tau, \lambda, \mathcal{T}_p)$ for large τ. The following claim is what we need.

Claim G.1. *For any lsc, proper, convex $\rho : \mathbb{R}^p \to \mathbb{R} \cup \{\infty\}$ and $\boldsymbol{z} \sim \mathsf{N}(0, \boldsymbol{I}_p/p)$ independent of $\boldsymbol{\beta}_0$,*

$$\mathbb{E}_{\boldsymbol{\beta}_0,\boldsymbol{z}}[\|\mathsf{prox}[\rho](\boldsymbol{\beta}_0+\tau\boldsymbol{z})-\boldsymbol{\beta}_0\|^2] \le \left(\sqrt{\mathbb{E}_{\boldsymbol{\beta}_0,\boldsymbol{z}}[\langle\tau\boldsymbol{z}, \mathsf{prox}[\rho](\boldsymbol{\beta}_0+\tau\boldsymbol{z})\rangle]} + \sqrt{\mathbb{E}_{\boldsymbol{\beta}_0,\boldsymbol{z}}[\|\mathsf{prox}[\rho](\boldsymbol{\beta}_0) - \boldsymbol{\beta}_0\|^2]}\right)^2. \tag{G.1}$$

Proof of Claim G.1. Write

$$\|\mathsf{prox}\left[\rho\right](\boldsymbol{\beta}_0+\tau\boldsymbol{z})-\boldsymbol{\beta}_0\|^2 \le \|\mathsf{prox}\left[\rho\right](\boldsymbol{\beta}_0+\tau\boldsymbol{z})-\mathsf{prox}\left[\rho\right](\boldsymbol{\beta}_0)\|^2 \tag{G.2a}$$

$$+2\|\mathsf{prox}\left[\rho\right](\boldsymbol{\beta}_0+\tau\boldsymbol{z})-\mathsf{prox}\left[\rho\right](\boldsymbol{\beta}_0)\|\,\|\mathsf{prox}\left[\rho\right](\boldsymbol{\beta}_0)-\boldsymbol{\beta}_0\| \tag{G.2b}$$

$$+\|\mathsf{prox}\left[\rho\right](\boldsymbol{\beta}_0)-\boldsymbol{\beta}_0\|^2. \tag{G.2c}$$

First, we bound the first term on the right-hand side.

By (O.3), we have

$$\|\mathsf{prox}[\rho](\boldsymbol{\beta}_0+\tau\boldsymbol{z})-\mathsf{prox}[\rho](\boldsymbol{\beta}_0)\|^2 \le \langle \tau\boldsymbol{z}, \mathsf{prox}[\rho](\boldsymbol{\beta}_0+\tau\boldsymbol{z})-\mathsf{prox}[\rho](\boldsymbol{\beta}_0)\rangle. \tag{G.3}$$

Taking expectations of both sides and using that $\mathbb{E}_{\boldsymbol{\beta}_0,\boldsymbol{z}}[\langle \tau\boldsymbol{z}, \mathsf{prox}[\rho](\boldsymbol{\beta}_0)\rangle]=0$ by the independence of $\boldsymbol{\beta}_0$ and $\boldsymbol{z}$ and the fact that $\mathbb{E}_{\boldsymbol{z}}[\boldsymbol{z}]=\mathbf{0}$, we get

$$\mathbb{E}_{\boldsymbol{\beta}_0,\boldsymbol{z}}[\|\mathsf{prox}[\rho](\boldsymbol{\beta}_0+\tau\boldsymbol{z})-\mathsf{prox}[\rho](\boldsymbol{\beta})\|^2] \le \mathbb{E}_{\boldsymbol{\beta}_0,\boldsymbol{z}}[\langle \tau\boldsymbol{z}, \mathsf{prox}[\rho](\boldsymbol{\beta}_0+\tau\boldsymbol{z})\rangle]. \tag{G.4}$$

We bound the expectation of (G.2b) by Cauchy-Schwartz.

$$\begin{aligned}
&\mathbb{E}_{\boldsymbol{\beta}_0,\boldsymbol{z}}[\|\mathsf{prox}\left[\rho\right](\boldsymbol{\beta}_0+\tau\boldsymbol{z})-\mathsf{prox}\left[\rho\right](\boldsymbol{\beta}_0)\|\,\|\mathsf{prox}\left[\rho\right](\boldsymbol{\beta}_0)-\boldsymbol{\beta}_0\|] \\
&\qquad \le \sqrt{\mathbb{E}_{\boldsymbol{\beta}_0,\boldsymbol{z}}[\|\mathsf{prox}\left[\rho\right](\boldsymbol{\beta}_0+\tau\boldsymbol{z})-\mathsf{prox}\left[\rho\right](\boldsymbol{\beta}_0)\|^2]}\sqrt{\mathbb{E}_{\boldsymbol{\beta}_0,\boldsymbol{z}}[\|\mathsf{prox}[\rho](\boldsymbol{\beta}_0)-\boldsymbol{\beta}_0\|^2]} \\
&\qquad \le \sqrt{\mathbb{E}_{\boldsymbol{\beta}_0,\boldsymbol{z}}[\langle \tau\boldsymbol{z}, \mathsf{prox}[\rho](\boldsymbol{\beta}_0+\tau\boldsymbol{z})\rangle]}\sqrt{\mathbb{E}_{\boldsymbol{\beta}_0,\boldsymbol{z}}[\|\mathsf{prox}[\rho](\boldsymbol{\beta}_0)-\boldsymbol{\beta}_0\|^2]},
\end{aligned} \tag{G.5}$$

where in the third line we have used (G.4). Taking the expectation of (G.2) and applying bounds (G.4), (G.5) gives (G.1). □

We are ready to prove Lemma F.1. Fix $\gamma \ge 0$ and $\kappa \ge 0$, so that ρ_p is κ-strongly convex (note, when $\kappa = 0$ we make no strong convexity assumption). Consider solutions τ, λ to (F.1b) at γ. To simplify notation, we denote $\rho_{\mathsf{orc}} = \lambda_{\mathsf{orc}}\rho$. By (B.4a) and Claim G.1,

$$\begin{aligned}
\frac{1}{\delta\tau^2}\mathsf{R}_{\mathsf{reg,cvx}}(\tau_{\mathsf{orc}},\lambda_{\mathsf{orc}},\mathcal{T}_p) &= \frac{1}{\delta\tau^2}\mathbb{E}_{\boldsymbol{\beta}_0,\boldsymbol{z}}[\|\mathsf{prox}[\rho_{\mathsf{orc}}](\boldsymbol{\beta}_0+\tau_{\mathsf{orc}}\boldsymbol{z})-\boldsymbol{\beta}_0\|^2] \\
&\le \left(\sqrt{\frac{1}{\delta\tau^2}\mathbb{E}_{\boldsymbol{\beta}_0,\boldsymbol{z}}[\langle \tau_{\mathsf{orc}}\boldsymbol{z}, \mathsf{prox}[\rho_{\mathsf{orc}}](\boldsymbol{\beta}_0+\tau_{\mathsf{orc}}\boldsymbol{z})\rangle]}+\sqrt{\frac{1}{\delta\tau^2}\mathbb{E}_{\boldsymbol{\beta}_0,\boldsymbol{z}}[\|\mathsf{prox}[\rho_{\mathsf{orc}}](\boldsymbol{\beta}_0)-\boldsymbol{\beta}_0\|^2]}\right)^2.
\end{aligned} \tag{G.6}$$

First we bound the second term on the right-hand side of (G.6). By (C.1) and (F.1b), $1 \ge 2\lambda\left(1-\frac{1}{\delta(\lambda\gamma+1)}\frac{1}{\lambda_{\mathsf{orc}}\kappa+1}\right) = 2\lambda\left(1-\frac{1}{\delta}\frac{1}{\lambda\kappa+\lambda\gamma+1}\right)$. If either $\delta > 1$, $\gamma > 0$, or $\kappa > 0$, the right-hand side diverges to ∞ for $\lambda \to \infty$. Thus, there exists λ_{max} depending only on δ, γ, κ such that all solutions τ, λ to (F.1b) at γ satisfy $\lambda \le \lambda_{\mathsf{max}}$. Then we have

$$\begin{aligned}
\|\mathsf{prox}[\rho_{\mathsf{orc}}](\boldsymbol{\beta}_0)-\boldsymbol{\beta}_0\| &\le \|\mathsf{prox}[\rho](\mathbf{0})\| + \|\mathsf{prox}[\lambda_{\mathsf{orc}}\rho](\mathbf{0})-\mathsf{prox}[\rho](\mathbf{0})\| \\
&\qquad + \|\mathsf{prox}[\lambda_{\mathsf{orc}}\rho](\boldsymbol{\beta}_0)-\mathsf{prox}[\lambda_{\mathsf{orc}}\rho](\mathbf{0})\| + \|\boldsymbol{\beta}_0\| \\
&\le M + M|\lambda_{\mathsf{orc}}-1| + 2\|\boldsymbol{\beta}_0\| \le M(\lambda_{\mathsf{max}}+2) + 2\|\boldsymbol{\beta}_0\|,
\end{aligned}$$

where in the second inequality, we have used (O.5) and (O.4), and in the third inequality, we have used $\lambda_{\mathsf{orc}} \le \lambda \le \lambda_{\mathsf{max}}$. Thus,

$$\mathbb{E}_{\boldsymbol{\beta}_0,\boldsymbol{z}}[\|\mathsf{prox}[\rho_{\mathsf{orc}}](\boldsymbol{\beta}_0) - \boldsymbol{\beta}_0\|^2] \le 2M^2(\lambda_{\mathsf{max}}+2)^2 + 8s_2(\pi), \tag{G.7}$$

where $s_2(\pi)$ is the second moment of π.

Second we bound the first term on the right-hand side of (G.6). We bound the first term by using the fact that τ, λ, γ solve (F.1b). In particular, by (F.1b) we have that

$$\frac{1}{\delta(\lambda\gamma+1)}\mathsf{W}_{\mathsf{reg,cvx}}(\tau_{\mathsf{orc}}, \lambda_{\mathsf{orc}}, \mathcal{T}_p) = 1 - \frac{1}{2\lambda} \le 1 - \frac{1}{2\lambda_{\mathsf{max}}} \tag{G.8}$$

Then, applying (B.4b), we get

$$\begin{aligned}\frac{1}{\delta\tau^2}\mathbb{E}_{\boldsymbol{\beta}_0,\boldsymbol{z}}[\langle \tau_{\mathsf{orc}}\boldsymbol{z}, \mathsf{prox}[\rho_{\mathsf{orc}}](\boldsymbol{\beta}_0 + \tau_{\mathsf{orc}}\boldsymbol{z})\rangle] &= \frac{1}{\delta(\lambda\gamma+1)^2}\mathsf{W}_{\mathsf{reg,cvx}}(\tau_{\mathsf{orc}}, \lambda_{\mathsf{orc}}, \mathcal{T}_p) \\ &\le 1 - \frac{1}{2\lambda_{\mathsf{max}}}.\end{aligned}$$

Plugging this and (G.7) into (G.6), we get

$$\frac{1}{\delta\tau^2}\mathsf{R}_{\mathsf{reg,cvx}}(\tau_{\mathsf{orc}}, \lambda_{\mathsf{orc}}, \mathcal{T}_p) \le \left(\sqrt{1 - \frac{1}{2\lambda_{\mathsf{max}}}} + \sqrt{\frac{2M^2(\lambda_{\mathsf{max}}+2)^2 + 8s_2(\pi)}{\delta\tau^2}}\right)^2. \tag{G.9}$$

Choose τ_{max} such that

$$1 > \frac{\sigma^2}{\delta\tau_{\mathsf{max}}^2} + \left(\sqrt{1 - \frac{1}{2\lambda_{\mathsf{max}}}} + \sqrt{\frac{2M^2(\lambda_{\mathsf{max}}+2)^2 + 8s_2(\pi)}{\tau_{\mathsf{max}}^2\delta}}\right)^2, \tag{G.10}$$

which is possibly because $1 - \frac{1}{2\lambda_{\mathsf{max}}} < 1$. This choice depends only on $\pi, M, \gamma, \kappa, \delta$. This inequality also holds for any $\tau \ge \tau_{\mathsf{max}}$. Chaining (G.9) and (G.10) and performing some rearrangement, we get that F.2 holds, as desired. □

G.2 Proof of Lemma F.2

The proof proceeds in three steps. The only place where the δ-bounded width assumption is used is in Case 1 of Step 2.

Step 1: Construct interval on which $\delta\tau^2 - \sigma^2 < \mathsf{R}^{\mathsf{opt}}_{\mathsf{seq,cvx}}(\tau;\pi)$.

By the definition of $\mathsf{R}^{\mathsf{opt}}_{\mathsf{seq,cvx}}(\tau;\pi)$, there exists $\zeta > 0$ such that

$$\delta\tau_{\mathsf{lb}}^2 - \sigma^2 < \mathsf{R}^{\mathsf{opt}}_{\mathsf{seq,cvx}}(\tau_{\mathsf{lb}};\pi,p) - \zeta$$

eventually. By the regularity property established in Lemma C.2, we may pick $\Delta > 0$ such that $\mathsf{R}^{\mathsf{opt}}_{\mathsf{seq,cvx}}(\tau;\pi,p) > \mathsf{R}^{\mathsf{opt}}_{\mathsf{seq,cvx}}(\tau_{\mathsf{lb}};\pi,p) - \zeta/3$ and $\delta\tau^2 - \sigma^2 < \delta\tau_{\mathsf{lb}}^2 - \sigma^2 + \zeta/3$ for all $\tau \in [\tau_{\mathsf{lb}}, \tau_{\mathsf{lb}} + \Delta]$. In particular, for all such τ

$$\delta\tau^2 - \sigma^2 \le \delta\tau_{\mathsf{lb}}^2 - \sigma^2 + \zeta/3 < \mathsf{R}^{\mathsf{opt}}_{\mathsf{seq,cvx}}(\tau_{\mathsf{lb}};\pi,p) - 2\zeta/3 < \mathsf{R}^{\mathsf{opt}}_{\mathsf{seq,cvx}}(\tau;\pi,p) - \zeta/3. \tag{G.11}$$

Step 2: Choose oracle parameter with not-too-small oracle effective noise.

The meaning of the preceding statement will become clear shortly. Let

$$\tau_{\mathsf{min}} = \tau_{\mathsf{lb}} + \Delta. \tag{G.12}$$

Denote

$$\frac{\tau_{\mathsf{lb}}}{\tau_{\mathsf{lb}} + \Delta} = 1 - \theta. \tag{G.13}$$

For simplicity, for the remainder of the proof, we denote the subsequence $\{p(\ell)\}$ as $\{p\}$. We will show how to choose $\gamma > 0$ such that, for each p, any solution λ to (F.1b) at τ_{min}, γ satisfies

$$\tau_{\mathsf{min,orc}} \geq \tau_{\mathsf{lb}}, \tag{G.14}$$

where we have denoted $\tau_{\mathsf{min,orc}} = \frac{\tau_{\mathsf{min}}}{\lambda\gamma+1}$. This is what we mean by "choose oracle parameter with not-too-small oracle effective noise." There are two cases.

- **Case 1:** $\delta \leq 1$**.**

  Because $\{\rho_p\} \in \mathcal{C}_{\delta,\pi}$, by (2.9), we can (and do) choose $\bar{\lambda} > 0$ and $\xi > 0$ such that

  $$\limsup_{p\to\infty} \sup_{\lambda>\bar{\lambda},\tau'\in[\delta\tau_{\mathsf{min}}/2,\tau_{\mathsf{min}}]} \frac{1}{\tau'}\mathbb{E}_{\boldsymbol{\beta}_0,\boldsymbol{z}}[\langle \boldsymbol{z}, \mathsf{prox}[\lambda\rho_p](\boldsymbol{\beta}_0 + \tau'\boldsymbol{z})\rangle] < \delta(1-\xi), \tag{G.15}$$

  (note that by assumption, $\delta/2 < 1$, so the interval is non-empty). Let $\{p(\ell)\}$ be a subsequence of $\{p\}$ such that

  $$\sup_{\lambda>\bar{\lambda},\tau'\in[\delta\tau_{\mathsf{min}}/2,\tau_{\mathsf{min}}]} \frac{1}{\tau'}\mathbb{E}_{\boldsymbol{\beta}_0,\boldsymbol{z}}[\langle \boldsymbol{z}, \mathsf{prox}[\lambda\rho_{p(\ell)}](\boldsymbol{\beta}_0 + \tau'\boldsymbol{z})\rangle] < \delta(1-\xi) \tag{G.16}$$

  for all ℓ. Now choose

  $$0 < \gamma < \min\left\{\frac{2}{\delta} - 1,\ \frac{\theta}{\bar{\lambda}},\ \frac{\theta}{1-\theta}2\xi\right\}. \tag{G.17}$$

  It is straightforward to check that the right-hand side is positive, so that such γ exist. Now consider any solution λ to (F.1b) at $\tau_{\mathsf{min}}, \gamma$. Thus,

  $$\begin{aligned}\frac{2}{\delta} - 1 > \gamma &= 2\lambda\gamma\left(1 - \frac{1}{\delta(\lambda\gamma+1)}\mathsf{W}_{\mathsf{reg,cvx}}(\tau_{\mathsf{min,orc}}, \lambda_{\mathsf{orc}}, \mathcal{T}_p)\right) \\ &\geq 2\lambda\gamma\left(1 - \frac{1}{\delta(\lambda\gamma+1)}\right) = 2\left(\left(\frac{1}{\lambda\gamma+1}\right)^{-1} - 1\right)\left(1 - \frac{1}{\delta(\lambda\gamma+1)}\right),\end{aligned}$$

  where in the first inequality we have used (G.17), in the first equality we have used (F.1b), and in the second inequality we have used (C.1). The right-hand side is strictly decreasing in $\frac{1}{\lambda\gamma+1}$. Moreover, the right-hand side equals $\frac{2}{\delta} - 1$ when $\frac{1}{\lambda\gamma+1} = \frac{\delta}{2}$. We conclude that $\frac{1}{\lambda\gamma+1} \geq \frac{\delta}{2}$, whence

  $$\tau_{\mathsf{min}} \geq \tau_{\mathsf{min,orc}} \geq \frac{\delta\tau_{\mathsf{min}}}{2}, \tag{G.18}$$

where the first inequality holds because trivially $1 \geq \frac{1}{\lambda\gamma+1}$. We now use the crude lower bound of (G.18) to generate the lower bound (G.14). Either $\lambda > \frac{1}{2\xi}$ or $\lambda \leq \frac{1}{2\xi}$. If $\lambda > \frac{1}{2\xi}$, then

$$\frac{1}{\tau_{\mathsf{min,orc}}}\mathbb{E}_{\boldsymbol{\beta}_0,\boldsymbol{z}}\left[\langle \boldsymbol{z}, \mathsf{prox}\left[\lambda_{\mathsf{orc}}\rho\right](\boldsymbol{\beta}_0 + \tau_{\mathsf{min,orc}}\boldsymbol{z})\rangle\right] = \mathsf{W}_{\mathsf{reg,cvx}}(\tau_{\mathsf{min,orc}}, \lambda_{\mathsf{orc}}, \mathcal{T}_p)$$
$$= \delta(\lambda\gamma+1)\left(1 - \frac{1}{2\lambda}\right) > \delta\left(1-\xi\right), \quad \text{(G.19)}$$

where in the first line, we have used (B.4b), and in the second line, we have used (F.1b). Combining this with (G.16) and (G.18), we conclude $\bar{\lambda} \geq \lambda_{\mathsf{orc}}$. Thus, $\frac{1}{\lambda\gamma+1} = 1 - \frac{\lambda\gamma}{\lambda\gamma+1} = 1 - \lambda_{\mathsf{orc}}\gamma \geq 1 - \bar{\lambda}\gamma$. By (G.12), (G.17) and (G.13),

$$\tau_{\mathsf{min,orc}} = \frac{\tau_{\mathsf{lb}} + \Delta}{\lambda\gamma+1} \geq (\tau_{\mathsf{lb}} + \Delta)\left(1 - \bar{\lambda}\gamma\right) \geq (\tau_{\mathsf{lb}} + \Delta)(1-\theta) = \tau_{\mathsf{lb}},$$

so we have (G.14). On the other hand, if $\lambda \leq \frac{1}{2\xi}$, then by (G.17)

$$\tau_{\mathsf{min,orc}} = \frac{\tau_{\mathsf{lb}} + \Delta}{\lambda\gamma+1} \geq \frac{\tau_{\mathsf{lb}} + \Delta}{\gamma/(2\xi)+1} \geq \frac{\tau_{\mathsf{lb}} + \Delta}{\frac{\theta}{1-\theta}+1} = (\tau_{\mathsf{lb}} + \Delta)(1-\theta) = \tau_{\mathsf{lb}},$$

so we also have (G.14). Thus, if we choose γ to satisfy (G.17), then (G.14) holds at any solution λ to (F.1b) at $\tau_{\mathsf{min}}, \gamma$.

- **Case 2:** $\delta > 1$.

  Choose
  $$0 \leq \gamma < \frac{2\theta(\delta-1)}{(1-\theta)\delta}. \quad \text{(G.20)}$$
  Now consider any solution λ to (F.1b) at $\tau_{\mathsf{min}}, \gamma$. By (C.1),
  $$1 = 2\lambda\left(1 - \frac{1}{\delta(\lambda\gamma+1)}\mathsf{W}_{\mathsf{reg,cvx}}(\tau_{\mathsf{min,orc}}, \lambda_{\mathsf{orc}}, \mathcal{T}_p)\right) \geq 2\lambda\left(1 - \frac{1}{\delta}\right).$$
  We conclude that $\lambda \leq \frac{\delta}{2(\delta-1)}$. Thus, by (G.20) and (G.13),
  $$\tau_{\mathsf{min,orc}} = \frac{\tau_{\mathsf{lb}} + \Delta}{\lambda\gamma+1} > \frac{\tau_{\mathsf{lb}} + \Delta}{\frac{\delta}{2(\delta-1)}\frac{2\theta(\delta-1)}{(1-\theta)\delta}+1} = (\tau_{\mathsf{lb}} + \Delta)(1-\theta) = \tau_{\mathsf{lb}},$$
  so we have (G.14). Thus, if we choose γ to satisfy (G.20), then (G.14) holds at any solution λ to (F.1b) at $\tau_{\mathsf{min}}, \gamma$.

Step 3: Combine steps 1 and 2.
We now provide the construction required by the lemma. We choose $\gamma, \tau_{\mathsf{min}}$, and subsequence $\{p(\ell)\}$ as in Step 2. We showed that, along this sequence, for any λ which solves (F.1b), we have (G.14). Because $\frac{1}{\lambda\gamma+1} \leq 1$, we also have $\tau_{\mathsf{min,orc}} \leq \tau_{\mathsf{lb}} + \Delta$. Thus, $\tau_{\mathsf{min,orc}} \in [\tau_{\mathsf{lb}}, \tau_{\mathsf{lb}} + \Delta]$, and by (G.11), we have $\delta\tau_{\mathsf{min}}^2 - \sigma^2 < \delta\tau_{\mathsf{lb}}^2 - \sigma^2 + \zeta/3 < \mathsf{R}^{\mathsf{opt}}_{\mathsf{seq,cvx}}(\tau_{\mathsf{min,orc}}; \pi, p)$. We conclude (F.3). □

G.3 Proof of Lemma F.3

Proof of Lemma F.3. We prove parts (i) and (ii) in parallel.

Under the conditions of part (i), by Lemma F.2, we can (and do) choose $\gamma > 0, \tau_{\mathsf{min}} \geq \tau_{\mathsf{lb}}$, and a subsequence $\{p(\ell)\}$ of $\{p\}$ such that for all p in the subsequence and all λ which solves (F.1b) at $\tau_{\mathsf{min}}, \gamma$, (F.3) holds. Under the conditions of part (ii), we take $\tau_{\mathsf{min}} = \tau_{\mathsf{lb}}$ and $\gamma = 0$. Now there exists a subsequence such that (F.3) holds for any λ by the definition of $\mathsf{R}^{\mathsf{opt}}_{\mathsf{reg,cvx}}$ and τ_{lb} (and in particular, it holds for those λ solving (F.1b)).

Because $\{\rho_{p(\ell)}\} \in \mathcal{B}$ (indeed, property (C.5) is closed under taking subsequences), we may choose M such that $M \geq \|\mathsf{prox}[\rho_{p(\ell)}](\mathbf{0})\|$ for all ℓ. By Lemma F.1, under the conditions of parts (i) and (ii) and the respective choices of γ, we can (and do) choose τ_{max} such that if τ, λ is a solution of (F.1b) at γ with $\tau \geq \tau_{\mathsf{max}}$, then F.2 holds.

Choose $\lambda_{\mathsf{max}} > 0$ such that

$$2\lambda_{\mathsf{max}} \left(1 - \frac{1}{\delta(\lambda_{\mathsf{max}}\gamma + \lambda_{\mathsf{max}}\kappa + 1)}\right) > 1, \tag{G.21}$$

where $\kappa = 0$ when $\{\rho_p\}$ is not uniformly strongly convex. Note that this is possible in part (i) because $\gamma > 0$, and in part (ii) because either $\delta > 1$ or $\kappa > 0$. Finally, choose $\lambda_{\mathsf{min}} > 0$ such that

$$2\lambda_{\mathsf{min}} < 1. \tag{G.22}$$

For simplicity, we denote the subsequence $\{p(\ell)\}$ by $\{p\}$ for the remainder of the proof.

For each p, denote

$$r_p(\tau, \lambda) = \delta\tau^2 - \sigma^2 - \mathsf{R}_{\mathsf{cvx,cvx}}(\tau_{\mathsf{orc}}, \lambda_{\mathsf{orc}}, \mathcal{T}_p), \tag{G.23}$$

$$w_p(\tau, \lambda) = 2\lambda \left(1 - \frac{1}{\delta(\lambda\gamma + 1)} \mathsf{W}_{\mathsf{reg,cvx}}(\tau_{\mathsf{orc}}, \lambda_{\mathsf{orc}}, \mathcal{T}_p)\right). \tag{G.24}$$

By Lemma C.5 and the continuity of the map $(\tau, \lambda) \mapsto \left(\frac{\tau}{\lambda\gamma+1}, \frac{\lambda}{\lambda\gamma+1}\right)$ on $(\tau, \lambda) \in [\tau_{\mathsf{min}}, \tau_{\mathsf{max}}] \times [\lambda_{\mathsf{min}}, \lambda_{\mathsf{max}}]$, we have that w_p and r_p are continuous on $[\tau_{\mathsf{min}}, \tau_{\mathsf{max}}] \times [\lambda_{\mathsf{min}}, \lambda_{\mathsf{max}}]$. To simplify notation in the argument that follows, we will work under the change of variables implemented by the linear bijection

$$\iota : [0, 1] \times [0, 2] \to [\tau_{\mathsf{min}}, \tau_{\mathsf{max}}] \times [\lambda_{\mathsf{min}}, \lambda_{\mathsf{max}}], \tag{G.25}$$

$$(a, b) \mapsto \left((1 - a)\tau_{\mathsf{min}} + a\tau_{\mathsf{max}}, \left(1 - \frac{b}{2}\right)\lambda_{\mathsf{min}} + \frac{b}{2}\lambda_{\mathsf{max}}\right). \tag{G.26}$$

The functions $r_p \circ \iota$ and $w_p \circ \iota$ are continuous on $[0, 1] \times [0, 2]$. By (C.1), (C.2), and (O.12), we have for all τ, λ that $2\lambda \geq w_p(\tau, \lambda) \geq 2\lambda \left(1 - \frac{1}{\delta(\lambda\gamma+1)(1+\lambda_{\mathsf{orc}}\kappa)}\right) = 2\lambda \left(1 - \frac{1}{\delta(1+\lambda\gamma+\lambda\kappa)}\right)$. Thus, by (G.21), (G.22), and (G.24),

$$w_p \circ \iota(a, 0) < 1, \quad w_p \circ \iota(a, 2) > 1 \quad \text{for all } a \in [0, 1]. \tag{G.27}$$

We seek $(a, b) \in [0, 1] \times [0, 2]$ such that

$$r_p \circ \iota(a, b) = 0 \quad \text{and} \quad w_p \circ \iota(a, b) = 1. \tag{G.28}$$

The next several paragraphs provide the construction, which essentially amounts to a type of two-dimensional intermediate value theorem.

Let $D_0 = [0,1] \times \{0\}$ and $D_2 = [0,1] \times \{2\}$. Let $S = \{(a,b) \in [0,1] \times [0,2] \mid w_p \circ \iota(a,b) \leq 1\}$. Note that D_0 is a connected subset of S by (G.27). Let $C_0 = \bigcup C$, where the union is taken over connected sets $C \subset S$ which contain D_0. The set C_0 is connected [Moi77, Theorem 1.14], so we are justified in calling C_0 "the connected component of S which contains D_0." The set C_0 is also closed because S is closed and the closure of any connected set is still connected. Thus, it is compact. By (G.27), $D_2 \cap S = \emptyset$, so that C_0 and D_2 are disjoint. Because C_0 and D_2 are disjoint and compact, they are separated by some Euclidean distance $\xi > 0$.

For any $\theta > 0$, define

$$C_{0,\theta} = \{(a,b) \in [0,1] \times [0,2] \mid d((a,b), C_0) \leq \theta\}, \tag{G.29}$$

where d denotes Euclidean distance. Clearly, $C_{0,\theta}$ is closed. For $\theta < \xi/3$, $C_{0,\theta}$ is distance at least $2\xi/3$ from D_2. We consider the lattice on $[0,1] \times [0,2]$ consisting of points $\left(\frac{i}{N}, \frac{j}{N}\right)$ for $i \in \{0,1,\ldots,N\}$ and $j \in \{0,1,\ldots,2N\}$, where N is chosen to be large enough so that

$$\theta_N := \frac{\sqrt{5}}{N} = \mathsf{diam}\left(\left[\frac{i}{N}, \frac{i+2}{N}\right] \times \left[\frac{j}{N}, \frac{j+1}{N}\right]\right) < \frac{\xi}{3}. \tag{G.30}$$

Here, diam denotes the supremal distance between two points contained in a set. We define a set of points $\mathcal{V}$ and line segments $\mathcal{E}$ as follows. The vertex set $\mathcal{V}$ is

$$\mathcal{V} = \left\{ \boldsymbol{v}_{ij} := \left(\frac{i}{N}, \frac{j}{N}\right) \;\middle|\; i \in \{0,1,\ldots,N\},\, j \in \{0,1,\ldots,2N\} \right\}. \tag{G.31}$$

The edge set $\mathcal{E}$ contains "horizontal" edges $E^H_{ij} := \left\{\left(\frac{i}{N}, \frac{j}{N}\right), \left(\frac{i+1}{N}, \frac{j}{N}\right)\right\}$ and "vertical" edges $E^V_{ij} := \left\{\left(\frac{i}{N}, \frac{j}{N}\right), \left(\frac{i}{N}, \frac{j+1}{N}\right)\right\}$ for certain values of i, j, as we now specify.

Horizontal edges. The edge $E^H_{ij} \in \mathcal{E}$ if and only if the following are all true.

(i) $i \in \{0,\ldots,N-1\}$ and $j \in \{1,\ldots,2N-1\}$ (ie. we exclude edges along the bottom or top edge of $[0,1] \times [0,2]$).

(ii) Either (i) $j - i$ is even and exactly one of the open rectangles $\left(\frac{i}{N}, \frac{i+2}{N}\right) \times \left(\frac{j}{N}, \frac{j+1}{N}\right)$ and $\left(\frac{i-1}{N}, \frac{i+1}{N}\right) \times \left(\frac{j-1}{N}, \frac{j}{N}\right)$ has non-empty intersection with $C_{1,\xi/3}$, or (ii) $j - i$ is odd and exactly one of the open rectangles $\left(\frac{i-1}{N}, \frac{i+1}{N}\right) \times \left(\frac{j}{N}, \frac{j+1}{N}\right)$ and $\left(\frac{i}{N}, \frac{i+2}{N}\right) \times \left(\frac{j-1}{N}, \frac{j}{N}\right)$ has non-empty intersection with $C_{1,\xi/3}$.

Vertical edges. The edge $E^V_{ij} \in \mathcal{E}$ if and only if the following are all true.

(i) $i \in \{0,\ldots,2N-1\}$ and $j \in \{1,\ldots,N-1\}$ (ie. we exclude edges along the left or right edge of $[0,1] \times [0,2]$).

(ii) $j - i$ is even and exactly one of the open rectangles $\left(\frac{i-2}{N}, \frac{i}{N}\right) \times \left(\frac{j}{N}, \frac{j+1}{N}\right)$ and $\left(\frac{i}{N}, \frac{i+2}{N}\right) \times \left(\frac{j}{N}, \frac{j+1}{N}\right)$ has non-empty intersection with $C_{1,\xi/3}$.

Remark G.1. To interpret the preceding definitions, the reader should have in mind the following picture. We tile the rectangle $[0,1] \times [0,2]$ with "bricks" of width 2 and height 1 whose alignment is offset by 1 in neighboring rows (as is done in [Moi77, Theorem 4.4]). The collection of edges we have specified delineates the outer-boundary of the union of bricks in the tiling which intersect $C_{0,\xi/3}$ (excluding the shared boundary with $[0,1] \times [0,2]$ itself). We should think of think of this as a more topologically well-behaved approximation to the boundary of C_0 itself.

We establish the following series of claims.

Claim G.2. *For all edges $E \in \mathcal{E}$, all points $\boldsymbol{p} \in E$ are distance at least θ_N and at most $2\theta_N$ from C_0 and at least $\xi/3$ from D_2.*

Proof of Claim G.2. Note each edge is contained in the boundary of each of the rectangles invoked in its definition. That is, for horizontal edges E^H_{ij} with $j-i$ even, we have $E^H_{ij} \in \left[\frac{i}{N}, \frac{i+2}{N}\right] \times \left[\frac{j}{N}, \frac{j+1}{N}\right]$ and $\left[\frac{i-1}{N}, \frac{i+1}{N}\right] \times \left[\frac{j-1}{N}, \frac{j}{N}\right]$, and for $j-i$ odd we have $E^H_{ij} \in \left[\frac{i-1}{N}, \frac{i+1}{N}\right] \times \left[\frac{j}{N}, \frac{j+1}{N}\right]$ and $\left[\frac{i}{N}, \frac{i+2}{N}\right] \times \left[\frac{j-1}{N}, \frac{j}{N}\right]$. For vertical edges, we have $E^V_{ij} \in \left[\frac{i-2}{N}, \frac{i}{N}\right] \times \left[\frac{j}{N}, \frac{j+1}{N}\right]$ and $\left[\frac{i}{N}, \frac{i+2}{N}\right] \times \left[\frac{j}{N}, \frac{j+1}{N}\right]$. Thus, all edges $E \in \mathcal{E}$ are contained in the boundary of a rectangle which does not intersect $C_{0,\xi/3}$, so that all $\boldsymbol{p} \in E$ are distance at least $\xi/3 > \theta_N$ from C_0. Also, all edges $E \in \mathcal{E}$ are contained in the boundary of a rectangle of diameter $< \theta_N$ with non-empty intersection with $C_{0,\xi/3}$. Because every point of $C_{0,\xi/3}$ is distance at most $\xi/3$ from C_0, we see that all $\boldsymbol{p} \in E$ are distance at most $\xi/3 + \theta_N < 2\xi/3$ from C_0. Because C_0 and D_2 are separated by distance ξ, all $\boldsymbol{p} \in E$ are distance at least $\xi/3$ from D_2. We have established Claim G.2. □

Claim G.3. *For $i \neq 0$ or N and $j \neq 0$ or $2N$, the vertex $\boldsymbol{v}_{ij}$ is the endpoint of either 0 or 2 edges in $\mathcal{E}$. (That is, this applies to vertices not on the boundary of $[0,1] \times [0,2]$).*

Proof of G.3. The only edges which possibly have endpoint $\boldsymbol{v}_{ij}$ are vertical edges $E^V_{ij}, E^V_{i(j-1)}$ and horizontal edges $E^H_{ij}, E^H_{(i-1)j}$. First, consider that $j-i$ is even. Then $E^V_{i(j-1)} \notin \mathcal{E}$ because $j-1-i$ is not even. There are three rectangles whose intersection with C_{0,θ_N} determine the membership of the remaining three edges, E^V_{ij}, E^H_{ij}, and $E^H_{(i-1)j}$, in $\mathcal{E}$. They are $\left(\frac{i-2}{N}, \frac{i}{N}\right) \times \left(\frac{j}{N}, \frac{j+1}{N}\right)$, $\left(\frac{i}{N}, \frac{i+2}{N}\right) \times \left(\frac{j}{N}, \frac{j+1}{N}\right)$, and $\left(\frac{i-1}{N}, \frac{i+1}{N}\right) \times \left(\frac{j-1}{N}, \frac{j}{N}\right)$. The edge E^V_{ij} is in $\mathcal{E}$ if exactly one of the first two rectangles has non-empty intersection with C_{0,θ_N}; the edge E^H_{ij} is in $\mathcal{E}$ if exactly one of the last two has non-empty intersection with C_{0,θ_N}; and the edge $E^H_{(i-1)j}$ is in $\mathcal{E}$ if exactly one of the first and last rectangle has has non-empty intersection with C_{0,θ_N}. Thus, if exactly one or two of the three rectangles has non-empty intersection with C_{0,θ_N}, then two of the edges $E^V_{ij}, E^H_{ij}, E^H_{(i-1)j}$ is in $\mathcal{E}$; otherwise, none of these edges are in $\mathcal{E}$. The case $j-i$ odd is similar. This establishes Claim G.3. □

Claim G.4. *If $i = 0$ or N or $j = 0$ or $2N$, then the vertex $\boldsymbol{v}_{ij}$ is the endpoint of either 0 or 1 edges in $\mathcal{E}$.*

Proof of Claim G.4. For $i = 0$, it is easy to check that the only edge which could be in $\mathcal{E}$ without violating conditions (i) is E^H_{0j}. The other cases are similar, establishing Claim G.4. □

Though we have defined $\mathcal{V}$ and $\mathcal{E}$ as sets of points and line segments in the plane, we may think of them as vertices and edges in a graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$. Claims G.3 and G.4 establish by elementary graph theory that the graph is partitioned into connected components, each of which is a path whose endpoints are on the boundary of $[0,1] \times [0,2]$ and whose other vertices are in the interior of $[0,1] \times [0,2]$. These paths contain each of the vertices in the path exactly once.

Claim G.5. *There is a path $\boldsymbol{p}_0, \dots, \boldsymbol{p}_K$ in the graph $\mathcal{G}$ such that $\boldsymbol{p}_0 \in \{0\} \times [0,2]$ and $\boldsymbol{p}_K \in \{1\} \times [0,2]$, the left and right boundary of $[0,1] \times [0,2]$.*

Proof of Claim G.5. Observe that $\left(0, \frac{2}{N}\right) \times \left(0, \frac{1}{N}\right)$ intersects C_{0,θ_N} because it is distance 0 from $[0,1] \times \{0\} = D_0 \subset C_0$. Also, $\left(-\frac{1}{N}, \frac{1}{N}\right) \times \left(\frac{2N-1}{N}, 2\right)$ does not intersect C_{0,θ_N} because it has diameter $\theta_N < \xi/3$ and intersects D_2, which has distance at least ξ from C_0. Thus, there is a j_{max} the maximal value of j such that $\left(\frac{i(j)}{N}, \frac{i(j)+2}{N}\right) \times \left(\frac{j-1}{N}, \frac{j}{N}\right)$ has non-empty intersection with C_{0,θ_N}, where we have denoted $i(j) = -1$ if j is even and $i(j) = 0$ if j is odd. By the definition of $\mathcal{E}$, we see that $E^H_{0j_{\mathsf{max}}} \in \mathcal{E}$ and j_{max} is the maximal j for which this is true. Let $\boldsymbol{p}_0 = \boldsymbol{v}_{0j_{\mathsf{max}}}$ and $\boldsymbol{p}_0, \boldsymbol{p}_1, \dots, \boldsymbol{p}_K$ be the connected path in $\mathcal{G}$ to which $\boldsymbol{p}_0$ belongs. We claim $\boldsymbol{p}_K \in \{1\} \times [0,2]$. We have already established that $\boldsymbol{p}_K$ is on the boundary of $[0,1] \times [0,2]$, so we only need to eliminate the possibility that it belongs to the top, bottom, or left boundaries. Because $\boldsymbol{p}_K$ is contained in an edge $E \in \mathcal{E}$, we have $\boldsymbol{p}_K \notin D_2$, the top boundary, by Claim G.2. Similarly, $\boldsymbol{p}_K \notin D_0$, the bottom boundary, because $D_0 \subset C_0$ and, by Claim G.2, $\boldsymbol{p}_K$ is distance at least θ_N from C_0. Finally, consider that $\boldsymbol{p}_K$ were in $\{0\} \times [0,2]$, the left boundary. Then the final edge in the path is E^V_{0j} for some $j \neq j_{\mathsf{max}}$. By the definition of j_{max}, we in fact have $j < j_{\mathsf{max}}$. Also, $j > 0$ because otherwise $\boldsymbol{p}_{K-1}$ is also on the boundary of $[0,1] \times [0,2]$. If we connect $\boldsymbol{p}_K = \left(0, \frac{j}{N}\right)$ and $\boldsymbol{p}_0 = \left(0, \frac{j_{\mathsf{max}}}{N}\right)$ by a line-segment, then $\boldsymbol{p}_0, \boldsymbol{p}_1, \dots, \boldsymbol{p}_K$ are the vertices of a polygon P (formally, the union of line segments connecting the adjacent vertices and $\boldsymbol{p}_0, \boldsymbol{p}_K$). By [Moi77, Theorem 2.1], $\mathbb{R}^2 \setminus P$ has two connected components which are disconnected from each other, one of which is bounded and one of which is unbounded. It is straightforward to check that the open rectangle $\left(0, \frac{1}{N}\right) \times \left(\frac{j_{\mathsf{max}}-1}{N}, \frac{j_{\mathsf{max}}}{N}\right)$ is in the bounded component,[3] and D_0 is in the unbounded component (because $j > 0$). But $\left(\frac{i(j_{\mathsf{max}})}{N}, \frac{i(j_{\mathsf{max}})+2}{N}\right) \times \left(\frac{j_{\mathsf{max}}-1}{N}, \frac{j_{\mathsf{max}}}{N}\right)$ intersects $C_{0,\xi/3}$ but not the polygon, and $D_0 \subset C_{0,\xi/3}$ and $C_{0,\xi/3}$ is connected, which contradicts that $\left(0, \frac{1}{N}\right) \times \left(\frac{j_{\mathsf{max}}-1}{N}, \frac{j_{\mathsf{max}}}{N}\right)$ and D_0 are contained in disconnected components of $\mathbb{R}^2 \setminus P$. Thus, we conclude $\boldsymbol{p}_K \notin \{0\} \times [0,2]$, the left boundary. We have established Claim G.5. □

Now we construct such a path for a sequence $N \to \infty$. That is, for each N we have a path $\boldsymbol{p}^{(N)}_0, \dots, \boldsymbol{p}^{(N)}_{K_N}$ such that $\boldsymbol{p}^{(N)}_0 \in \{0\} \times [0,2]$, $\boldsymbol{p}^{(N)}_{K_N} \in \{1\} \times [0,2]$, and whose edges satisfy Claim G.2. By compactness, we may take a subsequence $\{N(\ell)\}$ of $\{N\}$ such that $\boldsymbol{p}_{N(\ell)0} \to \boldsymbol{p}_{\mathsf{left}}$ and $\boldsymbol{p}_{N(\ell)K_{N(\ell)}} \to \boldsymbol{p}_{\mathsf{right}}$ for some $\boldsymbol{p}_{\mathsf{left}}, \boldsymbol{p}_{\mathsf{right}}$. Because by Claim G.2 the points $\boldsymbol{p}_{N(\ell)0}$ and $\boldsymbol{p}_{N(\ell)K_{N(\ell)}}$ are between distance θ_N and $2\theta_N$ from C_0, we have that $\boldsymbol{p}_{\mathsf{left}}, \boldsymbol{p}_{\mathsf{right}} \in \partial C_0$. Thus, $w_p \circ \iota(\boldsymbol{p}_{\mathsf{left}}) = w_p \circ \iota(\boldsymbol{p}_{\mathsf{right}}) = 1$. Thus, by Lemmas F.1 and F.2, we have $r_p \circ \iota(\boldsymbol{p}_{\mathsf{left}}) < 0$ and $r_p \circ \iota(\boldsymbol{p}_{\mathsf{right}}) > 0$. By the continuity of $r_p \circ \iota$, we have for sufficiently large ℓ that $r_p \circ \iota(\boldsymbol{p}_{N(\ell)0}) > 0$ and $r_p \circ \iota(\boldsymbol{p}_{N(\ell)K_{N(\ell)}}) < 0$.

[3] This can be established rigorously by computing the "index" in the sense of [Moi77, Lemma 2.2] of a point $\boldsymbol{p}$ in its interior. Compute the index via a horizontal ray which starts at $\boldsymbol{p}$ and points left. This ray intersect the polygon in 1 point, so has index 1. See [Moi77] for details.

Then, by the Intermediate Value Theorem along the path $\boldsymbol{p}_{N(\ell)0}, \ldots, \boldsymbol{p}_{N(\ell)K_{N(\ell)}}$, we have for each ℓ sufficiently large a point $\boldsymbol{p}_{N(\ell)}$ on the path such that $r_p \circ \iota(\boldsymbol{p}_{N(\ell)}) = 0$. By compactness, there exists a further subsequence $\{N'(\ell)\}$ of $\{N(\ell)\}$ such that $\boldsymbol{p}_{N'(\ell)} \to \boldsymbol{p}^*$. By continuity, we have $r_p \circ \iota(\boldsymbol{p}^*) = 0$. Further, because $\boldsymbol{p}_{N'(\ell)}$ is between distance $\theta_{N'(\ell)}$ and $2\theta_{N'(\ell)}$ from C_0, we have in fact that $\boldsymbol{p}^* \in \partial C_0$, whence $w_p \circ \iota(\boldsymbol{p}^*) = 1$. With $(\tau, \lambda) = \iota(\boldsymbol{p}^*)$, we have that $(\tau, \lambda) \in [\tau_{\mathsf{reg,cvx}} - \varepsilon, \tau_{\mathsf{max}}] \times [1/2, \lambda_{\mathsf{max}}]$ and τ, λ solves (F.1a), (F.1b), as desired. □

H Proofs for Section 5: beyond mean square error

Proof of Proposition 5.1. By the strong stationarity of $\tau, \lambda, \delta, \mathcal{T}$, we have by (B.4a), (B.5), (B.7), that $\mathbb{E}_{\tilde{\boldsymbol{\beta}}_0, \boldsymbol{z}}\left[\|\mathsf{prox}[\lambda\rho_p](\tilde{\boldsymbol{\beta}}_0 + \tau\boldsymbol{z}) - \tilde{\boldsymbol{\beta}}_0\|^2\right]$ is bounded, where $\tilde{\beta}_{0j} \overset{\text{iid}}{\sim} \pi/\sqrt{p}$. By Jensen's, also $\mathbb{E}_{\tilde{\boldsymbol{\beta}}_0, \boldsymbol{z}}\left[\|\mathsf{prox}[\lambda\rho_p](\tilde{\boldsymbol{\beta}}_0 + \tau$ is bounded. By the triangle inequality that

$$\begin{aligned}\|\mathsf{prox}[\lambda\rho_p](\mathbf{0})\| &\le \|\mathsf{prox}[\lambda\rho_p](\tilde{\boldsymbol{\beta}}_0 + \tau\boldsymbol{z}) - \tilde{\boldsymbol{\beta}}_0\| + \|\tilde{\boldsymbol{\beta}}_0\| \\ &\quad + \|\mathsf{prox}[\lambda\rho_p](\mathbf{0}) - \mathsf{prox}[\lambda\rho_p](\tilde{\boldsymbol{\beta}}_0 + \tau\boldsymbol{z})\| \\ &\le \|\mathsf{prox}[\lambda\rho_p](\tilde{\boldsymbol{\beta}}_0 + \tau\boldsymbol{z}) - \tilde{\boldsymbol{\beta}}_0\| + \|\tilde{\boldsymbol{\beta}}_0\| + \|\tilde{\boldsymbol{\beta}}_0 + \tau\boldsymbol{z}\|,\end{aligned}$$

where in the second inequality we have applied (O.4) from Appendix O. Taking expectations on both sides, we get that $\mathsf{prox}[\lambda\rho_p](\mathbf{0})$ is bounded. Further, again by (O.4) from Appendix O, we have that the sequence (in p) of functions $(\boldsymbol{x}_1, \boldsymbol{x}_2) \mapsto (\boldsymbol{x}_1, \mathsf{prox}[\lambda\rho_p](\boldsymbol{x}_2))$ is uniformly pseudo-Lipschitz of order 1 and bounded at $(\mathbf{0}, \mathbf{0})$. Then, by Lemma P.5 from Appendix P, we have the sequence of functions $(\boldsymbol{x}_1, \boldsymbol{x}_2) \mapsto \ell_p(\boldsymbol{x}_1, \mathsf{prox}[\lambda\rho_p](\boldsymbol{x}_2))$ is uniformly pseudo-Lipschitz of order k. By Proposition B.3(iii) applied to $\tilde{\tau}, \tilde{\lambda}, \delta, \tilde{\mathcal{T}}$, we then have

$$\ell_p\left(\boldsymbol{\beta}_0, \mathsf{prox}[\lambda\rho_p]\left(\widehat{\boldsymbol{\beta}}_{\mathsf{cvx}} + 2\lambda\frac{\boldsymbol{X}^\mathsf{T}(\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}_{\mathsf{cvx}})}{n}\right)\right) \overset{\text{p}}{\simeq} \mathbb{E}_{\boldsymbol{z}}\left[\ell_p\left(\boldsymbol{\beta}_0, \boldsymbol{\beta}_0 + \tau\boldsymbol{z}\right)\right], \tag{H.1}$$

where we have used that either $\delta > 1$ or $\{\rho_p\} \in \mathcal{C}_*$. Further, by Lemma P.4, the sequence of functions $(\boldsymbol{x}_1, \boldsymbol{x}_2) \mapsto \mathbb{E}_{\boldsymbol{z}}\left[\ell_p(\boldsymbol{x}_1, \mathsf{prox}[\lambda\rho_p](\boldsymbol{x}_2 + \sqrt{\tau^2 - \tilde{\tau}^2}\boldsymbol{z})\right]$ is uniformly pseudo-Lipschitz of order k. By Proposition B.3(iii) applied to $\tilde{\tau}, \tilde{\lambda}, \delta, \tilde{\mathcal{T}}$, we then have under either (i) the HDA and RSN assumption, or (ii) the HDA and DSN assumptions if $\tilde{\rho}_p$ are symmetric, that

$$\begin{aligned}&\mathbb{E}_{\boldsymbol{z}}\left[\ell_p\left(\boldsymbol{\beta}_0, \mathsf{prox}\left[\lambda\rho_p\right]\left(\widehat{\tilde{\boldsymbol{\beta}}}_{\mathsf{cvx}} + 2\lambda\frac{\boldsymbol{X}^\mathsf{T}(\boldsymbol{y} - \boldsymbol{X}\widehat{\tilde{\boldsymbol{\beta}}}_{\mathsf{cvx}})}{n} + \sqrt{\tau^2 - \tilde{\tau}^2}\boldsymbol{z}\right)\right)\right] \\ &\qquad \overset{\text{p}}{\simeq} \mathbb{E}_{\boldsymbol{z}_1, \boldsymbol{z}_2}\left[\ell_p\left(\boldsymbol{\beta}_0, \boldsymbol{\beta}_0 + \tilde{\tau}\boldsymbol{z}_1 + \sqrt{\tau^2 - \tilde{\tau}^2}\boldsymbol{z}_2\right)\right] \\ &\qquad = \mathbb{E}_{\boldsymbol{z}}\left[\ell_p\left(\boldsymbol{\beta}_0, \boldsymbol{\beta}_0 + \tau\boldsymbol{z}\right)\right],\end{aligned} \tag{H.2}$$

where $\boldsymbol{z}_1, \boldsymbol{z}_2 \sim \mathsf{N}(\mathbf{0}, \boldsymbol{I}_p/p)$ are independent and we have used that either $\delta > 1$ or $\{\rho_p\} \in \mathcal{C}_*$. By Lemma C.3, we have that

$$\begin{aligned}&\mathbb{E}_{\boldsymbol{z}}\left[\ell_p\left(\boldsymbol{\beta}_0, \mathsf{prox}\left[\lambda\rho_p\right]\left(\widehat{\tilde{\boldsymbol{\beta}}}_{\mathsf{cvx}} + 2\lambda\frac{\boldsymbol{X}^\mathsf{T}(\boldsymbol{y} - \boldsymbol{X}\widehat{\tilde{\boldsymbol{\beta}}}_{\mathsf{cvx}})}{n} + \sqrt{\tau^2 - \tilde{\tau}^2}\boldsymbol{z}\right)\right)\right] \\ &\qquad \overset{\text{p}}{\simeq} \ell_p\left(\boldsymbol{\beta}_0, \mathsf{prox}\left[\lambda\rho_p\right]\left(\widehat{\tilde{\boldsymbol{\beta}}}_{\mathsf{cvx}} + 2\lambda\frac{\boldsymbol{X}^\mathsf{T}(\boldsymbol{y} - \boldsymbol{X}\widehat{\tilde{\boldsymbol{\beta}}}_{\mathsf{cvx}})}{n} + \sqrt{\tau^2 - \tilde{\tau}^2}\boldsymbol{z}\right)\right).\end{aligned} \tag{H.3}$$

Combining (H.1), (H.2), and (H.3), and using the definition (5.1), we get (5.2), as desired. □

Proof of Theorem 5. By the same argument as in Claim E.1, it is enough to show (5.3) for $\{\rho_p\} \in \mathcal{C}_* \cap \mathcal{B}$. Assume for the sake of contradiction that the left-hand side of (5.3) is less than the right-hand side. By passing to a subsequence, we may assume we have $\{\rho_p\} \in \mathcal{C}_*$ such that

$$\operatorname*{p\lim}_{p\to\infty} \frac{1}{p}\sum_{j=1}^{p} \ell\left(\sqrt{p}\beta_{0j}, \sqrt{p}\widehat{\beta}_{\mathsf{cvx},j}\right) < \mathbb{E}_{\beta_0,z}[\ell(\beta_0, \eta(\beta_0 + \tau_{\mathsf{reg,cvx}} z)].$$

By Lemma E.3, we may find a further subsequence $\{p(\ell)\}$, $\tau \geq \tau_{\mathsf{reg,cvx}}$, and $\lambda > 0$ such that with $\mathcal{T} = (\pi, \{\rho_{p(\ell)}\})$, the quintuplet $\tau, \lambda, \delta, \gamma = 0, \mathcal{T}$ is strongly stationary. By the KKT conditions for (1.2),

$$\widehat{\boldsymbol{\beta}}_{\mathsf{cvx}} = \mathsf{prox}[\lambda\rho]\left(\widehat{\boldsymbol{\beta}}_{\mathsf{cvx}} + 2\lambda\frac{\boldsymbol{X}^\top(\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}_{\mathsf{cvx}})}{n}\right),$$

whence Proposition B.3(iii) implies (either under the HDA and RSN assumptions, or, if the penalties are symmetric, with the RSN assumption replaced by the DSN assumption)

$$\begin{aligned}
\operatorname*{p\lim}_{p\to\infty} \frac{1}{p}\sum_{j=1}^{p} \ell\left(\sqrt{p}\beta_{0j}, \sqrt{p}\widehat{\beta}_{\mathsf{cvx},j}\right) &\overset{\mathrm{p}}{\simeq} \mathbb{E}_{\boldsymbol{z}}\left[\frac{1}{p}\sum_{j=1}^{p} \ell\left(\sqrt{p}\beta_{0j}, \sqrt{p}\mathsf{prox}[\lambda\rho_p](\boldsymbol{\beta}_0 + \tau_{\mathsf{reg,cvx}}\boldsymbol{z})_j\right)\right] \\
&\overset{\mathrm{p}}{\simeq} \mathbb{E}_{\tilde{\boldsymbol{\beta}}_0,\boldsymbol{z}}\left[\frac{1}{p}\sum_{j=1}^{p} \ell\left(\sqrt{p}\tilde{\beta}_{0j}, \sqrt{p}\mathsf{prox}[\lambda\rho_p](\tilde{\boldsymbol{\beta}}_0 + \tau_{\mathsf{reg,cvx}}\boldsymbol{z})_j\right)\right] \\
&\geq \mathbb{E}_{\beta_0,z}[\ell(\beta_0, \eta(\beta_0 + \tau_{\mathsf{reg,cvx}}))],
\end{aligned}$$

where the second inequality holds by Lemma C.3 under the DSN and RSAN assumption or by Lemma C.4 when ρ_p are symmetric and the DSN assumption holds (here $\tilde{\beta}_{0j} \overset{\mathrm{iid}}{\sim} \pi/\sqrt{p}$; and the final inequality holds by the optimality of η with respect to the loss ℓ. Moreover, if $\eta \neq \mathsf{prox}[\lambda\rho_p]$, which occurs when η is not a proximal operator, this inequality is strict.

By Theorem 1, when $\delta > 1$ the convex lower bound is strict. As we saw in its proof in Section E.4, tightness holds because there exists $\{\rho_p\} \in \mathcal{C}_*$ and $\lambda \geq 0$ such that with $\mathcal{T} = (\pi, \{\rho_p\})$ we have that $\tau_{\mathsf{reg,cvx}}, \lambda, \gamma = 0, \delta, \mathcal{T}$ is strongly stationary. Thus, for any Lipschitz η', by Proposition B.3 we have

$$\operatorname*{p\lim}_{p\to\infty} \frac{1}{p}\sum_{j=1}^{p} \ell\left(\sqrt{p}\beta_{0j}, \eta'\left(\sqrt{p}\widehat{\boldsymbol{\beta}}_{\mathsf{cvx},j} + 2\lambda\frac{[\boldsymbol{X}^\mathsf{T}(\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}_{\mathsf{cvx}})]_j}{n}\right)\right) = \mathbb{E}_{\beta_0,z}[\ell(\beta_0, \eta'(\beta_0 + \tau_{\mathsf{reg,cvx}} z))].$$

Because the set of Lipschitz functions is dense in $L_2(\pi * \mathsf{N}(0, \tau^2_{\mathsf{reg,cvx}}))$, taking the infimum over η' gives $\mathbb{E}_{\beta_0,z}[\ell(\beta_0, \eta(\beta_0, \tau_{\mathsf{reg,cvx}} z))]$ on the right-hand side. This completes the proof. □

I The role of the δ-bounded width assumption

The primary weakness of Theorem 1 is its restriction to sequences of convex functions in $\mathcal{C}_{\delta,\pi}$. For $\delta > 1$, this is no restriction at all. In this section, we provide some reflection on the nature of the

restriction for $\delta < 1$ and the role it plays in Theorem 1. No other sections or appendices depend upon the results in this appendix.

First, we observe that for $\delta < 1$, Theorem 1 does not hold if we instead take the infimum in (2.12) over $\{\rho_p\} \in \mathcal{C}$, the collection of all sequences of convex penalties.

Claim I.1. *Take $\rho_p = 0$ (so $\{\rho_p\} \notin \mathcal{C}_{\delta,\pi}$). Under the RSN assumption, if $\delta < 1$, there exists a random sequence $\widehat{\boldsymbol{\beta}}_{\mathsf{cvx}}$ such that for each p we have $\widehat{\boldsymbol{\beta}}_{\mathsf{cvx}} \in \arg\min_{\boldsymbol{\beta}} \frac{1}{n}\|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\|^2 + \rho_p(\boldsymbol{\beta})$ with probability 1 but*

$$\lim_{p\to\infty}{}^{\mathrm{p}} \|\widehat{\boldsymbol{\beta}}_{\mathsf{cvx}} - \boldsymbol{\beta}_0\|^2 = \frac{\delta\sigma^2}{1-\delta}.$$

For some such values of π, δ, σ, we have $\frac{\delta\sigma^2}{1-\delta} < \delta\tau^2_{\mathsf{reg,stat}} - \sigma^2 \le \delta\tau^2_{\mathsf{reg,cvx}} - \sigma^2$.

Proof. For sufficiently large p, we have $p > n$ because $n/p \to \delta < 1$. Take such sufficiently large p. Define

$$\widehat{\boldsymbol{\beta}}_{\mathsf{cvx}} = \arg\min_{\boldsymbol{\beta}} \left\{ \|\boldsymbol{\beta} - \boldsymbol{\beta}_0\|^2 \;\middle|\; \boldsymbol{\beta} \in \arg\min_{\boldsymbol{\beta}'} \left\{ \frac{1}{n}\|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\|^2 \right\} \right\}. \tag{I.1}$$

Clearly $\widehat{\boldsymbol{\beta}}_{\mathsf{cvx}} \in \arg\min_{\boldsymbol{\beta}} \frac{1}{n}\|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\|^2 + \rho_p(\boldsymbol{\beta})$. Let the singular value decomposition of $\boldsymbol{X}$ be $\boldsymbol{U}\boldsymbol{S}\boldsymbol{V}^\mathsf{T}$ where $\boldsymbol{U} \in \mathbb{R}^{n\times n}$ is orthonormal, $\boldsymbol{S} \in \mathbb{R}^{n\times n}$ is diagonal, and $\boldsymbol{V} \in \mathbb{R}^{p\times n}$ has orthonormal columns. Let $\boldsymbol{V}_\perp \in \mathbb{R}^{p\times(p-n)}$ have orthonormal columns orthogonal to those of $\boldsymbol{V}$. Because $p > n$, this makes sense, and moreover, $\boldsymbol{X}$ is full-rank with probability 1, whence $\boldsymbol{S}$ is non-singular. We parameterize $\boldsymbol{\beta}$ as $\boldsymbol{V}\boldsymbol{b} + \boldsymbol{V}_\perp\boldsymbol{b}_\perp$ for $\boldsymbol{b} \in \mathbb{R}^n$ and $\boldsymbol{b}_\perp \in \mathbb{R}^{p-n}$. Then

$$\frac{1}{n}\|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\|^2 = \frac{1}{n}\left\|\boldsymbol{y} - \boldsymbol{U}\boldsymbol{S}\boldsymbol{V}^\mathsf{T}(\boldsymbol{V}\boldsymbol{b} + \boldsymbol{V}_\perp\boldsymbol{b}_\perp)\right\|^2 = \frac{1}{n}\|\boldsymbol{y} - \boldsymbol{U}\boldsymbol{S}\boldsymbol{b}\|^2 = \frac{1}{n}\left\|\boldsymbol{U}^\mathsf{T}\boldsymbol{y} - \boldsymbol{S}\boldsymbol{b}\right\|.$$

Because $\boldsymbol{S}$ is non-singular,

$$\arg\min_{\boldsymbol{\beta}} \left\{ \frac{1}{n}\|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\|^2 \right\} = \left\{ \boldsymbol{V}\boldsymbol{S}^{-1}\boldsymbol{U}^\mathsf{T}\boldsymbol{y} + \boldsymbol{V}_\perp\boldsymbol{b}_\perp \mid \boldsymbol{b}_\perp \in \mathbb{R}^{p-n} \right\}.$$

Observe

$$\begin{aligned}\left\|\boldsymbol{V}\boldsymbol{S}^{-1}\boldsymbol{U}^\mathsf{T}\boldsymbol{y} + \boldsymbol{V}_\perp\boldsymbol{b}_\perp - \boldsymbol{\beta}_0\right\|^2 &= \left\|\boldsymbol{V}\boldsymbol{S}^{-1}\boldsymbol{U}^\mathsf{T}\boldsymbol{y} + \boldsymbol{V}_\perp\boldsymbol{b}_\perp - \boldsymbol{V}\boldsymbol{V}^\mathsf{T}\boldsymbol{\beta}_0 - \boldsymbol{V}_\perp\boldsymbol{V}_\perp^\mathsf{T}\boldsymbol{\beta}_0\right\|^2 \\ &= \left\|\boldsymbol{S}^{-1}\boldsymbol{U}^\mathsf{T}\boldsymbol{y} - \boldsymbol{V}^\mathsf{T}\boldsymbol{\beta}_0\right\|^2 + \left\|\boldsymbol{b}_\perp - \boldsymbol{V}_\perp^\mathsf{T}\boldsymbol{\beta}_0\right\|^2. \end{aligned} \tag{I.2}$$

This is minimized at $\boldsymbol{b}_\perp = \boldsymbol{V}_\perp^\mathsf{T}\boldsymbol{\beta}_0$, whence

$$\widehat{\boldsymbol{\beta}}_{\mathsf{cvx}} = \boldsymbol{V}\boldsymbol{S}^{-1}\boldsymbol{U}^\mathsf{T}\boldsymbol{y} + \boldsymbol{V}_\perp\boldsymbol{V}_\perp^\mathsf{T}\boldsymbol{\beta}_0. \tag{I.3}$$

Now consider the oracle estimator with parameter γ. The objective we must minimize is

$$\begin{aligned}\frac{1}{n}\|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\|^2 + \frac{\gamma}{2}\|\boldsymbol{\beta} - \boldsymbol{\beta}_0\|^2 &= \frac{1}{n}\left\|\boldsymbol{U}^\mathsf{T}\boldsymbol{y} - \boldsymbol{S}\boldsymbol{b}\right\|^2 + \frac{\gamma}{2}\left\|\boldsymbol{b} - \boldsymbol{V}^\mathsf{T}\boldsymbol{\beta}_0\right\|^2 + \frac{\gamma}{2}\left\|\boldsymbol{b}_\perp - \boldsymbol{V}_\perp^\mathsf{T}\boldsymbol{\beta}_0\right\|^2 \\ &= (\boldsymbol{b} - \boldsymbol{a})^\mathsf{T}\left(\boldsymbol{S}^2/n + \gamma\boldsymbol{I}_n/2\right)(\boldsymbol{b} - \boldsymbol{a}) + \frac{\gamma}{2}\left\|\boldsymbol{b}_\perp - \boldsymbol{V}_\perp^\mathsf{T}\boldsymbol{\beta}_0\right\|^2, \end{aligned}$$

where $\boldsymbol{a} = \left(\boldsymbol{S}^2/n + \gamma\boldsymbol{I}_n/2\right)^{-1}\left(\boldsymbol{S}\boldsymbol{U}^\mathsf{T}\boldsymbol{y}/n + \gamma\boldsymbol{V}^\mathsf{T}\boldsymbol{\beta}_0/2\right)$. Thus,

$$\widehat{\boldsymbol{\beta}}^{(\gamma)}_{\mathsf{cvx}} = \boldsymbol{V}\left(\boldsymbol{S}^2/n + \gamma\boldsymbol{I}_n/2\right)^{-1}\left(\boldsymbol{S}\boldsymbol{U}^\mathsf{T}\boldsymbol{y}/n + \gamma\boldsymbol{V}^\mathsf{T}\boldsymbol{\beta}_0/2\right) + \boldsymbol{V}_\perp\boldsymbol{V}_\perp^\mathsf{T}\boldsymbol{\beta}_0. \tag{I.4}$$

We get

$$\begin{aligned}
\|\widehat{\boldsymbol{\beta}}_{\mathsf{cvx}} - \widehat{\boldsymbol{\beta}}_{\mathsf{cvx}}^{(\gamma)}\| &= \left\|\boldsymbol{V}(\boldsymbol{S}^2/n)^{-1}\boldsymbol{S}\boldsymbol{U}^\mathsf{T}\boldsymbol{y}/n - \boldsymbol{V}\left(\boldsymbol{S}^2/n + \gamma\boldsymbol{I}_n/2\right)^{-1}\left(\boldsymbol{S}\boldsymbol{U}^\mathsf{T}\boldsymbol{y}/n + \gamma\boldsymbol{V}^\mathsf{T}\boldsymbol{\beta}_0/2\right)\right\| \\
&= \left\|(\boldsymbol{S}^2/n)^{-1}\boldsymbol{S}\boldsymbol{U}^\mathsf{T}\boldsymbol{y}/n - \left(\boldsymbol{S}^2/n + \gamma\boldsymbol{I}_n/2\right)^{-1}\left(\boldsymbol{S}\boldsymbol{U}^\mathsf{T}\boldsymbol{y}/n + \gamma\boldsymbol{V}^\mathsf{T}\boldsymbol{\beta}_0/2\right)\right\| \\
&\le \left\|\left((\boldsymbol{S}^2/n)^{-1} - \left(\boldsymbol{S}^2/n + \gamma\boldsymbol{I}_n/2\right)^{-1}\right)\boldsymbol{S}\boldsymbol{U}^\mathsf{T}\boldsymbol{y}/n\right\| \\
&\qquad + \left\|\left(\boldsymbol{S}^2/n + \gamma\boldsymbol{I}_n/2\right)^{-1}\gamma\boldsymbol{V}^\mathsf{T}\boldsymbol{\beta}_0/2\right\| \\
&\le \left\|(\boldsymbol{S}^2/n)^{-1} - \left(\boldsymbol{S}^2/n + \gamma\boldsymbol{I}_n/2\right)^{-1}\right\|_{\mathsf{op}} \frac{\|\boldsymbol{S}\|_{\mathsf{op}}}{\sqrt{n}}\frac{\|\boldsymbol{y}\|}{\sqrt{n}} \\
&\qquad + \frac{\gamma}{2}\left\|\left(\boldsymbol{S}^2/n + \gamma\boldsymbol{I}_n/2\right)^{-1}\right\|_{\mathsf{op}}\left\|\boldsymbol{V}^\mathsf{T}\boldsymbol{\beta}_0\right\| \\
&= \left|\frac{1}{\sigma_{\mathsf{min}}(\boldsymbol{X})^2/n} - \frac{1}{\sigma_{\mathsf{min}}(\boldsymbol{X})^2/n + \gamma/2}\right|\frac{\|\boldsymbol{X}\|_{\mathsf{op}}}{\sqrt{n}}\frac{\|\boldsymbol{y}\|}{\sqrt{n}} + \frac{\gamma}{2}\frac{1}{\sigma_{\mathsf{min}}(\boldsymbol{X})^2/n + \gamma/2}\left\|\boldsymbol{V}^\mathsf{T}\boldsymbol{\beta}_0\right\| \\
&= \varepsilon(\gamma)O_p(1),
\end{aligned} \tag{I.5}$$

for some deterministic function $\varepsilon(\gamma) \downarrow 0$ as $\gamma \to 0$ and $O_p(1)$ tight over both p and γ, where $\sigma_{\mathsf{min}}(\boldsymbol{X})$ is the minimal non-zero singular value of $\boldsymbol{X}$ and we have used that $\|\boldsymbol{X}\|_{\mathsf{op}}/\sqrt{n}$ and $\sigma_{\mathsf{min}}(\boldsymbol{X})/\sqrt{n}$ both converge in probability to constants by [Ver12, Theorem 5.31].

Let $\mathcal{T} = (\pi, \{\rho_p = 0\})$. Because $\rho_p = 0$, for all τ, λ, $\mathsf{prox}[\lambda\rho_p](\boldsymbol{\beta}_0 + \tau\boldsymbol{z}) - \boldsymbol{\beta}_0 = \boldsymbol{\beta}_0 + \tau\boldsymbol{z} - \boldsymbol{\beta}_0 = \tau\boldsymbol{z}$. Then by (B.4a) and (B.4b), we have that $\mathsf{R}^{\infty}_{\mathsf{reg,cvx}}(\tau, \lambda, \mathcal{T}) = \tau^2$ and $\mathsf{W}^{\infty}_{\mathsf{reg,cvx}}(\tau, \lambda, \mathcal{T}) = 1$. Thus, at oracle parameter γ, the fixed point equations (B.7) are equivalent to

$$\delta\tau^2 - \sigma^2 = \frac{\tau^2}{(\lambda\gamma+1)^2} \quad \text{and} \quad 2\lambda\left(1 - \frac{1}{\delta}\frac{1}{\lambda\gamma+1}\right) = 1. \tag{I.6}$$

It is straightforward to see that such a solution exists: we may choose non-negative λ to solve the second equation in (I.6) by the intermediate value theorem; at this value of λ we have $\frac{1}{\lambda\gamma+1} < \delta$, whence setting $\tau^2 = \frac{\sigma^2}{\delta - (\lambda\gamma+1)^{-2}}$ solves the first equation in (I.6). Then $\tau, \lambda, \gamma, \delta, \mathcal{T}$ is strongly stationary.

Equation (I.6) implies that $\frac{1}{\lambda\gamma+1} < \delta$, which implies $\lambda > \frac{\delta^{-1}-1}{\gamma} \to \infty$ as $\gamma \to 0$ because $\delta < 1$. We conclude that $\frac{1}{\lambda\gamma+1} = \delta\left(1 - \frac{1}{2\lambda}\right) \to \delta$ as $\gamma \to 0$. Then, writing equation (I.6) as $\delta(\lambda\gamma+1)^2\frac{\tau^2}{(\lambda\gamma+1)^2} - \sigma^2 = \frac{\tau^2}{(\lambda\gamma+1)^2}$, we get $\frac{\tau^2}{(\lambda\gamma+1)^2} = \frac{\sigma^2}{\delta(\lambda\gamma+1)^2 - 1} \to \frac{\sigma^2}{\delta^{-1}-1} = \frac{\delta\sigma^2}{1-\delta}$. In particular, by Proposition B.3, we have

$$\lim_{\gamma\to 0}\operatorname*{plim}_{p\to\infty}\|\widehat{\boldsymbol{\beta}}_{\mathsf{cvx}}^{(\gamma)} - \boldsymbol{\beta}_0\|^2 = \frac{\delta\sigma^2}{1-\delta}. \tag{I.7}$$

By (I.5), we have $\|\widehat{\boldsymbol{\beta}}_{\mathsf{cvx}} - \boldsymbol{\beta}_0\|^2 = \|\widehat{\boldsymbol{\beta}}_{\mathsf{cvx}}^{(\gamma)} - \boldsymbol{\beta}_0\|^2 + \varepsilon(\gamma)O_p(1)$. Taking $\gamma \to 0$ and applying (I.7), we get

$$\operatorname*{plim}_{p\to\infty}\|\widehat{\boldsymbol{\beta}}_{\mathsf{cvx}} - \boldsymbol{\beta}_0\|^2 = \frac{\delta\sigma^2}{1-\delta},$$

as desired.

It is easy to construct examples in which this is smaller than $\delta\tau^2_{\mathsf{reg,stat}} - \sigma^2$ and $\tau_{\mathsf{reg,stat}} \le \tau_{\mathsf{reg,cvx}}$. Here is one construction. Observe that all solutions $\tau^2_{\mathsf{reg,stat}}$ to (2.32) must satisfy $\tau^2_{\mathsf{reg,stat}} \ge \sigma^2/\delta$. Thus, for fixed σ, we have $\lim_{\delta\to 0}(\delta\tau^2_{\mathsf{reg,stat}} - \sigma^2) = \lim_{\delta\to 0}\mathsf{mmse}_\pi(\tau^2_{\mathsf{reg,stat}}) = \lim_{\tau\to\infty}\mathsf{mmse}_\pi(\tau^2) =$

$s_2(\pi)$ [DYSV11, Eq. (61)]. Moreover, $\lim_{\delta\to 0} \frac{\delta\sigma^2}{1-\delta} = 0$. Thus, if $s_2(\pi) > 0$ (which is true unless π is a point mass at 0), then for sufficiently small δ we have $\frac{\delta\sigma^2}{1-\delta} < \delta\tau^2_{\mathsf{reg,stat}} - \sigma^2$. When the minimizer of (2.30) is unique, we have $\delta\tau^2_{\mathsf{reg,stat}} - \sigma^2 \le \delta\tau^2_{\mathsf{reg,cvx}} - \sigma^2$ by Theorem 2. Because the minimizer of (2.30) is unique for almost every (δ, σ) (w.r.t. Lebegesgue measure), for some σ there are arbitrarily large δ at which the minimizer of (2.30) is unique. This completes the construction. □

Of course, Claim I.1 does not –indeed, could not– imply that we can achieve smaller than Bayes risk using convex M-estimation. The construction of $\widehat{\boldsymbol{\beta}}_{\mathsf{cvx}}$ given in (I.1) uses knowledge of $\boldsymbol{\beta}_0$, so is information theoretically inaccessible to the statistician. Indeed, even though our measurements and our penalty are completely uninformative along directions parallel to the null space of $\boldsymbol{X}$, the estimator $\widehat{\boldsymbol{\beta}}_{\mathsf{cvx}}$ in (I.1) achieves perfect estimation along these directions, as captured by the term $\boldsymbol{V}_\perp \boldsymbol{V}_\perp^\mathsf{T} \boldsymbol{\beta}_0$ in (I.3).

The counterexample of Claim I.1 demonstrates that the conclusion of Theorem 1 is too strong to remove all restrictions on the penalty sequence in (2.12). This is because Theorem 1 applies to all mechanisms for breaking ties between members of the minimizing set, even those which rely on knowledge of $\boldsymbol{\beta}_0$. The counterexample of Claim I.1 uses an uninformative penalty. When $\rho_p = 0$, the set of minimizers is large, and we have much to gain from breaking ties by looking at $\boldsymbol{\beta}_0$, something which the conditions of Theorem 1 do not prohibit.

This discussion is perhaps unsurprising given the way in which the δ-bounded width assumption is used in the proof of Theorem 1. The δ-bounded width assumption is used only in the proof of Lemma E.3. This Lemma shows that the oracle estimator which exploit knowledge of $\boldsymbol{\beta}_0$ does so weakly enough that it achieves loss at best negligibly smaller than the convex lower bound (see Lemma E.3). It is not hard to show that when $\rho_p = 0$, even arbitrarily weak oracles can dramatically improve the performance of the M-estimator by allowing us to estimating $\boldsymbol{\beta}_0$ exactly correctly along the null space of $\boldsymbol{X}$. Said more generally (but more heuristically), when the minimizing set of the original M-estimator is large –as it is when $\delta < 1$ and $\rho_p = 0$– arbitrarily weak oracles break ties in the way that best exploits knowledge of $\boldsymbol{\beta}_0$. Thus, arbitrarily weak oracles achieve a non-negligible improvement over the convex lower bound. Indeed, this is essentially what we have used in the proof of Claim I.1.

This is not to say that the statistician can do better by choosing penalty sequence from $\mathcal{C}$ rather than $\mathcal{C}_{\delta,\pi}$. Without making statements which are fully precise, we conjecture that (i) no convex M-estimator which with high-probability returns a singleton minimizing set (or perhaps even a minimizing set which is "small" in an appropriate sense) can achieve asymptotic loss smaller than $\delta\tau^2_{\mathsf{reg,cvx}} - \sigma^2$, and (ii) no convex procedure which breaks ties among members of the minimizing set with a polynomial-time algorithm can achieve asymptotic loss smaller than $\delta\tau^2_{\mathsf{reg,amp}*} - \sigma^2$.[4] If (i) is true, then it is possible to expand, at least slightly, the set over which we take the infimum in Theorem 1. We suspect that the restriction $\{\rho_p\} \in \mathcal{C}_{\delta,\pi}$ corresponds closely, though not exactly, to the restriction that the minimizing set be "small" in the appropriate sense. Resolving (i) would require identifying the appropriate weaker condition. Successfully resolving statement (ii) would require addressing some of the deepest and most insurmountable problems in the theory of computational complexity. Exploring whether and in what sense any of these speculations is true is beyond the scope of the current work.

[4] Perhaps the lower bound is even larger than this, because we are requiring that we use convex M-estimation for at least a part of the procedure.

J Proof of Proposition 3.1

Proof of Proposition 3.1(i). In fact, we prove (3.1) when the infimum is taken over $\{\rho_p\} \in \mathcal{C}_*$, the sequences of uniformly strongly convex penalties:

$$\inf_{\{\rho_p\}\in\mathcal{C}_*} \operatorname*{p\,lim}_{p\to\infty} \|\widehat{\boldsymbol{\beta}}_{\mathsf{cvx}} - \boldsymbol{\beta}_0\|^2 \le \delta\tau^2 - \sigma^2. \tag{J.1}$$

This is stronger than (3.1). If we show (J.1), we can conclude that under RSN assumption (3.1) holds also when the limit in probability is replaced by $\lim_{p\to\infty} \mathbb{E}_{\boldsymbol{\beta}_0,\boldsymbol{w},\boldsymbol{X}}\left[\|\hat{\boldsymbol{\beta}}_{\mathsf{cvx}} - \boldsymbol{\beta}_0\|_2^2\right]$ by applying Lemma K.2 and using that $\mathcal{C}_* \subset \mathcal{C}_{\delta,\pi}$ (see Proposition 6.2).

The proof of (J.1) proceeds in three steps.

Step 1: Construct lsc, proper, convex $\rho : \mathbb{R} \to \mathbb{R}$ such that $\mathsf{prox}[\rho]$ is the Bayes estimator. This construction is provided in [BBEKY13, pg. 14567]. We provide most of the details for completeness. Let $p_Y(x)$ be the density of $\pi * \mathsf{N}(0,\tau^2)$ (recall, $\tau > 0$, so that this exists). Let $m(y) = -\tau^2 \log p_Y(y)$ and $p_2(x) = \frac{1}{2}x^2$. By assumption, m is convex. Observe that up to the additive constant $\tau^2 \log\left(\sqrt{2\pi}\tau\right)$

$$m(y) = -\tau^2 \log \int e^{-\frac{1}{2\tau^2}(y-x)^2} \pi(\mathrm{d}x) = \frac{1}{2}y^2 - \tau^2 \log \int e^{\frac{1}{\tau^2}yx - \frac{1}{2\tau^2}x^2} \pi(\mathrm{d}x). \tag{J.2}$$

We identify the second term on the right-hand side –up to a multiplicative and additive constant– as the cumulant generating function of the probability distribution with density proportional to $e^{-\frac{1}{2\tau^2}x^2}$ with respect to π. This term can be written as $p_2(y) - m(y)$. Because, for all y, $e^{\frac{1}{\tau^2}yx - \frac{1}{2\tau^2}x^2}$ is bounded over x, this term is finite for all y, and by [Bro86, Theorem 1.13], it is infinitely differentiable and lsc, proper, and convex in y.

For an lsc, proper, convex $f : \mathbb{R} \to \mathbb{R} \cup \{\infty\}$, the *Fenchel-Legendre conjugate* f^* is defined by $f^*(g) = \sup_{x\in\mathbb{R}}\{gx - f(x)\}$. Define

$$\rho = (p_2 - m)^* - p_2. \tag{J.3}$$

This makes sense because we have argued that $p_2 - m$ is lsc, proper, and convex. Moreover, as argued in [BBEKY13, pg. 14567] by appeal to [Mor65, Proposition 9.b], ρ so defined is convex.[5] Define the *Moreau envelope* of ρ by $\mathsf{M}[\rho](y) = \inf_x \left\{\frac{1}{2}(y-x)^2 + \rho(x)\right\}$. Repeating the argument of [BBEKY13, pg. 14567], we have

$$\begin{aligned}\mathsf{M}[\rho](y) &= \inf_x \left\{\frac{1}{2}(y-x)^2 + \rho(x)\right\} = p_2(y) - \sup_x \{yx - (\rho(x) + p_2(x))\} \\ &= \left(p_2 - (\rho + p_2)^*\right)(y) = m(y),\end{aligned} \tag{J.4}$$

where we use that for any lsc, proper, convex f, we have $f^{**} = f$ [Roc97, Theorem 12.2]. By a fundamental identity for Moreau envelopes (see [BBEKY13, pg. 14567] and references therein), we get

$$\frac{\mathrm{d}}{\mathrm{d}y}\mathsf{M}[\rho](y) = \mathsf{prox}[\rho^*](y) = y - \mathsf{prox}[\rho](y). \tag{J.5}$$

[5] Roughly, this is because $p_2 - m$ is "less convex" than p_2, so $(p_2 - m)^*$ is "more convex" than p_2.

Let $\eta : \mathbb{R} \to \mathbb{R}$ be the Bayes estimator with respect to ℓ_2-loss in the scalar model $y = \beta_0 + \tau z$ where $\beta_0 \sim \pi$ and $z \sim \mathsf{N}(0,1)$ independent of β_0. That is $\eta(y) = \mathbb{E}_{\beta_0,z}[\beta_0|y]$. Recall $\tau > 0$. Thus, by Tweedie's formula (Lemma P.7), $\eta(y) = y - m'(y)$. By comparison with (J.5), we conclude

$$\eta(y) = \mathsf{prox}[\rho](y). \tag{J.6}$$

Step 2: Strongly stationary $\tau, \lambda, \gamma = 0, \delta, \mathcal{T}$ with uniformly strongly convex penalty. We have $\mathsf{mmse}_\pi(\tau^2) = \mathbb{E}_{\beta_0,z}[(\eta(y) - \beta_0)^2]$, whence by (J.6) and the assumption of the proposition

$$\delta\tau^2 - \sigma^2 > \mathsf{mmse}_\pi(\tau^2) = \mathbb{E}_{\beta_0,z}\left[(\mathsf{prox}[\rho](\beta_0 + \tau z) - \beta_0)^2\right]. \tag{J.7}$$

Observe also that for any $f : \mathbb{R} \to \mathbb{R}$ measurable for which the following expectations exist and are finite,

$$\begin{aligned}\mathbb{E}_{\beta_0,z}\left[f(y)(\mathbb{E}_{\beta_0,z}[\beta_0|y] - \beta_0)\right] &= \mathbb{E}_{\beta_0,z}\left[\mathbb{E}_{\beta_0,z}[f(y)(\mathbb{E}_{\beta_0,z}[\beta_0|y] - \beta_0) \mid y]\right] \\ &= \mathbb{E}_{\beta_0,z}\left[f(y)\mathbb{E}_{\beta_0,z}[\mathbb{E}_{\beta_0,z}[\beta_0|y] - \beta_0 \mid y]\right] = 0.\end{aligned}$$

Let $f(y) = y - \mathsf{prox}[\rho](y)$ in the previous display and recall $y = \beta_0 + \tau z$. After rearrangement and using the $\mathbb{E}_{\beta_0,z}[z\beta_0] = 0$,

$$\frac{1}{\tau}\mathbb{E}_{\beta_0,z}[z\,\mathsf{prox}[\rho](y)] = \frac{1}{\tau^2}\mathbb{E}_{\beta_0,z}\left[(\mathsf{prox}(y) - \beta_0)^2\right] < \delta - \frac{\sigma^2}{\tau^2} \le \delta. \tag{J.8}$$

Now consider $\kappa > 0$ and define

$$\rho^{(\kappa)}(x) = \rho(x) + \frac{\kappa}{2}x^2. \tag{J.9}$$

Then by (2.4)

$$\begin{aligned}\mathsf{prox}[\rho^{(\kappa)}](y) &= \arg\min_x \left\{\frac{1}{2}(y-x)^2 + \rho(x) + \frac{\kappa}{2}x^2\right\} = \arg\min_x \left\{\frac{1}{2}\left(\frac{1}{1+\kappa}y - x\right)^2 + \frac{1}{1+\kappa}\rho(x)\right\} \\ &= \mathsf{prox}\left[\frac{1}{1+\kappa}\rho\right]\left(\frac{1}{1+\kappa}y\right).\end{aligned} \tag{J.10}$$

First, we will choose $\kappa > 0$ sufficiently small such that (J.7) and (J.8) still hold with ρ replaced by $\rho^{(\kappa)}$. To make our notation more compact, we let $c_\kappa = \frac{1}{1+\kappa}$. Let $a = \mathsf{prox}[\rho](y) - \beta_0$ and $b = \mathsf{prox}\left[\rho^{(\kappa)}\right](y) - \beta_0$. We have

$$|a| \le |\mathsf{prox}[\rho](0)| + |\mathsf{prox}[\rho](y) - \mathsf{prox}[\rho](0)| + |\beta_0| \le |\mathsf{prox}[\rho](0)| + |y| + |\beta_0|, \tag{J.11}$$

$$\begin{aligned}|b| &\le |\mathsf{prox}[\rho](0)| + |\mathsf{prox}\left[c_\kappa\rho\right](0) - \mathsf{prox}[\rho](0)| + |\mathsf{prox}\left[c_\kappa\rho\right](c_\kappa y) - \mathsf{prox}\left[c_\kappa\rho\right](0)| + |\beta_0| \\ &\le |\mathsf{prox}[\rho](0)| + |\mathsf{prox}[\rho](0)|\,|c_\kappa - 1| + |c_\kappa y| + \beta_0 \\ &\le (c_\kappa + 2)|\mathsf{prox}[\rho](0)| + |c_\kappa y| + |\beta_0|,\end{aligned} \tag{J.12}$$

$$\begin{aligned}|a - b| &\le |y - \mathsf{prox}[\rho](y)|\,|c_\kappa - 1| \le (|y| + |\mathsf{prox}[\rho](y) - \mathsf{prox}[\rho](0)| + |\mathsf{prox}[\rho](0)|)\,|c_\kappa - 1| \\ &\le (|\mathsf{prox}[\rho](0)| + 2|y|)\,|c_\kappa - 1|,\end{aligned} \tag{J.13}$$

where in (J.11), we have used (O.4), and in both (J.12) and (J.13), we have used (O.4) and (O.5). Then by (J.11) and (J.12), we have $|a| \vee |b| \leq (c_\kappa + 2)|\mathsf{prox}[\rho](0)| + |y| + |\beta_0|$. Applying this bound, Jensen's inequality, (C.4), and (J.13), we conclude

$$\begin{aligned}\left|\mathbb{E}_{\beta_0,z}\left[(\mathsf{prox}[\rho](y)-\beta_0)^2\right] - \mathbb{E}_{\beta_0,z}\left[(\mathsf{prox}[\rho^{(\kappa)}](y)-\beta_0)^2\right]\right| &\leq \mathbb{E}_{\beta_0,z}\left[|a^2-b^2|\right] \\ &\leq 2\mathbb{E}_{\beta_0,z}\Big[\Big((c_\kappa+2)|\mathsf{prox}[\rho](0)|+|y|+|\beta_0|\Big)\Big(|\mathsf{prox}[\rho](0)|+2|y|\Big)\Big]|c_\kappa - 1| \xrightarrow[\kappa\to 0]{} 0, \end{aligned} \tag{J.14}$$

because $c_\kappa - 1 \to 0$ as $\kappa \to 0$, and the expectation is bounded. Also, by Jensen's inequality, Cauchy-Schwartz, and (J.13),

$$\begin{aligned}\left|\frac{1}{\tau}\mathbb{E}_{\beta_0,z}\left[z\mathsf{prox}[\rho](y)\right] - \frac{1}{\tau}\mathbb{E}_{\beta_0,z}[z\mathsf{prox}[\rho^{(\kappa)}](y)]\right| &\leq \frac{1}{\tau}\mathbb{E}_{\beta_0,z}\left[|z(a-b)|\right] \leq \frac{1}{\tau}\mathbb{E}_{\beta_0,z}[(a-b)^2]^{1/2} \\ &\leq \frac{1}{\tau}\mathbb{E}_{\beta_0,z}\big[(|\mathsf{prox}[\rho](0)|+2|y|)^2\big](c_\kappa-1)^2 \xrightarrow[\kappa\to 0]{} 0. \end{aligned} \tag{J.15}$$

By (J.7), (J.8), (J.14), and (J.15), we can (and do) choose κ sufficiently small that

$$\mathbb{E}_{\beta_0,z}\left[(\mathsf{prox}[\rho^{(\kappa)}](y)-\beta_0)^2\right] < \delta\tau^2 - \sigma^2, \tag{J.16}$$

$$\frac{1}{\tau}\mathbb{E}_{\beta_0,z}\left[z\mathsf{prox}[\rho^{(\kappa)}](y)\right] < \delta. \tag{J.17}$$

We now will define an lsc, proper, convex function $\tilde{\rho} : \mathbb{R} \to \mathbb{R}$ such that (J.16) holds with equality and (J.17) holds with the same strict inequality when $\rho^{(\kappa)}$ is replaced by $\tilde{\rho}$. By (J.16), we may choose $c \in \mathbb{R}$ such that

$$\mathbb{E}_{\beta_0,z}\left[(\mathsf{prox}[\rho^{(\kappa)}](y)+c-\beta_0)^2\right] = \delta\tau^2 - \sigma^2. \tag{J.18}$$

Define the lsc, proper, convex function $\tilde{\rho} : \mathbb{R} \to \mathbb{R}$ by $\tilde{\rho}(x) = -cx + \rho^{(\kappa)}(x-c)$. Then by (O.14), we have $\mathsf{prox}[\tilde{\rho}](y) = \mathsf{prox}[\rho^{(\kappa)}](y) + c$. Then (J.18) can be written

$$\mathbb{E}_{\beta_0,z}\left[(\mathsf{prox}\left[\tilde{\rho}\right](y)-\beta_0)^2\right] = \delta\tau^2 - \sigma^2 \tag{J.19}$$

Further, by (J.17) and because $\mathbb{E}_z[zc] = 0$, we have

$$\frac{1}{\tau}\mathbb{E}_{\beta_0,z}\left[z\mathsf{prox}[\tilde{\rho}](y)\right] = \frac{1}{\tau}\mathbb{E}_{\beta_0,z}\left[z\mathsf{prox}[\rho^{(\kappa)}](y)\right] < \delta. \tag{J.20}$$

Let $\lambda = \frac{1}{2}\left(1 - \frac{1}{\delta\tau}\mathbb{E}_{\beta_0,z}\left[z\mathsf{prox}[\tilde{\rho}](y)\right]\right)^{-1}$, where $\lambda > 0$ by (J.20). For each p, define the lsc, proper, symmetric, convex function $\rho_p : \mathbb{R}^p \to \mathbb{R} \cup \{\infty\}$ by

$$\rho_p(\boldsymbol{x}) = \frac{1}{\lambda p}\sum_{j=1}^{p}\tilde{\rho}(\sqrt{p}x_j). \tag{J.21}$$

By (O.13) and (O.15), we have for each $1 \leq j \leq p$ and $\boldsymbol{y} \in \mathbb{R}^p$ that $\mathsf{prox}[\lambda\rho_p](\boldsymbol{y})_j = \frac{1}{\sqrt{p}}\mathsf{prox}[\tilde{\rho}](\sqrt{p}y_j)$. For each p, let $\tilde{\boldsymbol{\beta}}_0 \in \mathbb{R}^p$ be random with coordinates distributed iid from $\pi/\sqrt{p}$ and $\boldsymbol{z} \sim \mathsf{N}(\boldsymbol{0}, \boldsymbol{I}_p/p)$.

These coordinates are denoted $\tilde{\beta}_{0j}$ and z_j. As above, β_0, z denote independent random variables distributed from π and $\mathsf{N}(0,1)$, respectively. For any $\tau' \geq 0$ and $\boldsymbol{T}' \in S_+^2$, we have

$$\begin{aligned}
\mathbb{E}_{\tilde{\boldsymbol{\beta}}_0, \boldsymbol{z}}\left[\left\|\mathsf{prox}\left[\lambda\rho_p\right](\tilde{\boldsymbol{\beta}}_0 + \tau'\boldsymbol{z}) - \tilde{\boldsymbol{\beta}}_0\right\|^2\right] &= \frac{1}{p}\sum_{j=1}^{p}\mathbb{E}_{\beta_{0j}, z_j}\left[(\mathsf{prox}[\tilde{\rho}](\sqrt{p}(\tilde{\beta}_{0j} + \tau' z_j)) - \sqrt{p}\tilde{\beta}_{0j})^2\right] \\
&= \mathbb{E}_{\beta_0, z}\left[(\mathsf{prox}[\tilde{\rho}](\beta_0 + \tau' z) - \beta_0)^2\right], \qquad \text{(J.22a)} \\
\frac{1}{\tau}\mathbb{E}_{\tilde{\boldsymbol{\beta}}_0, \boldsymbol{z}}\left[\langle \boldsymbol{z}, \mathsf{prox}[\lambda\rho_p](\tilde{\boldsymbol{\beta}}_0 + \tau'\boldsymbol{z})\rangle\right] &= \frac{1}{\tau p}\sum_{j=1}^{p}\mathbb{E}_{\tilde{\beta}_{0j}, z_j}\left[\sqrt{p} z_j \mathsf{prox}[\tilde{\rho}](\sqrt{p}(\tilde{\beta}_{0j} + \tau' z_j)\right] \\
&= \frac{1}{\tau}\mathbb{E}_{\beta_0, z}\left[z\mathsf{prox}[\tilde{\rho}](\beta_0 + \tau' z)\right], \qquad \text{(J.22b)}
\end{aligned}$$

$$\begin{aligned}
&\mathbb{E}_{\tilde{\boldsymbol{\beta}}_0 \boldsymbol{z}_1, \boldsymbol{z}_2}\left[\left\langle \mathsf{prox}\left[\lambda\rho_p\right]\left(\tilde{\boldsymbol{\beta}}_0 + \boldsymbol{z}_1\right) - \tilde{\boldsymbol{\beta}}_0, \mathsf{prox}\left[\lambda\rho_p\right]\left(\tilde{\boldsymbol{\beta}}_0 + \boldsymbol{z}_2\right) - \tilde{\boldsymbol{\beta}}_0\right\rangle\right] \\
&\quad = \frac{1}{p}\sum_{j=1}^{p}\mathbb{E}_{\tilde{\beta}_{0j}, z_{1j}, z_{2j}}\left[\left(\mathsf{prox}\left[\tilde{\rho}\right]\left(\sqrt{p}(\tilde{\beta}_{0j} + z_{1j})\right) - \sqrt{p}\tilde{\beta}_{0j}\right)\left(\mathsf{prox}\left[\tilde{\rho}\right]\left(\sqrt{p}(\tilde{\beta}_{0j} + z_{2j})\right) - \sqrt{p}\tilde{\beta}_{0j}\right)\right] \\
&\quad = \mathbb{E}_{\beta_0, z_1, z_2}\left[\left(\mathsf{prox}\left[\tilde{\rho}\right](\beta_0 + z_1) - \beta_0\right)\left(\mathsf{prox}\left[\tilde{\rho}\right](\beta_0 + z_2) - \beta_0\right)\right]. \qquad \text{(J.22c)}
\end{aligned}$$

Let $\mathcal{T} = (\pi, \{\rho_p\})$. We see the limits (B.5) exist for all $\tau' \geq 0$, $\boldsymbol{T}' \succeq \boldsymbol{0}$ at the λ we have defined. By (J.19), (J.22a), (J.22b), and the definition of λ, we see that equations (B.7) are satisfied at $\tau, \lambda, \gamma = 0, \delta, \mathcal{T}$. Thus, $\tau, \lambda, \gamma = 0, \delta, \mathcal{T}$ is strongly stationary.

Step 3: Exactly characterize the asymptotic risk.

By (J.9) and (J.21), observe that ρ_p has uniform strong convexity parameter $\kappa > 0$. Because $\tau, \lambda, \gamma = 0, \delta, \mathcal{T}$ is strongly stationary, by Proposition B.3 we have $\|\widehat{\boldsymbol{\beta}}_{\mathsf{cvx}} - \boldsymbol{\beta}_0\|^2 \xrightarrow{\mathrm{p}} \delta\tau^2 - \sigma^2$ where $\widehat{\boldsymbol{\beta}}_{\mathsf{cvx}}$ is defined as with respect to the penalties (J.21). This holds under the HDA and either the RSN or DSN assumptions because the penalties are symmetric.

Thus, by construction, we see that the risk $\delta\tau^2 - \sigma^2$ is achieved on the class $\mathcal{C}_*$ of uniformly strongly convex sequences of estimators, whence (J.1) follows. □

To prove Proposition 3.1.(ii), we will need the following lemma.

Lemma J.1. *Consider $\pi \in \mathcal{P}_2(\mathbb{R})$ and $\tau > 0$ such that $\pi * \mathsf{N}(0, \tau^2)$ does not have log-concave density with respect to Lebesgue measure on $\mathbb{R}$. Then*

$$\mathsf{R}^{\mathsf{opt}}_{\mathsf{seq,cvx}}(\tau; \pi) > \mathsf{mmse}_\pi(\tau^2). \qquad \text{(J.23)}$$

Proof of Lemma J.1. Throughout this proof, we will let $Y \sim \pi * \mathsf{N}(0, \tau^2)$ be a random variable. Because $\tau > 0$, $\pi * \mathsf{N}(0, \tau^2)$ has density with respect to Lebesgue measure which is infinitely continuously differentiable. Call this density p_Y. Let $\eta : \mathbb{R} \to \mathbb{R}$ be the Bayes estimator of β_0 given observation $\beta_0 + \tau z$ where $\beta_0 \sim \pi$ and $z \sim \mathsf{N}(0,1)$ independent of β_0. By Tweedie's formula (Lemma P.7),

$$\eta(y) = y + \tau^2 \frac{\mathrm{d}}{\mathrm{d}y}\log p_Y(y). \qquad \text{(J.24)}$$

Because $\pi * \mathsf{N}(0, \tau^2)$ is not log-concave and has infinitely continuously differentiable density, there exists $v \in \mathbb{R}$ and $\xi, \varepsilon > 0$ such that $\frac{\mathrm{d}^2}{\mathrm{d}y^2}\log p_Y(y) > \xi/\tau^2$ on $[v - 2\varepsilon, v + 2\varepsilon]$. Thus,

$$\eta'(y) > 1 + \xi \quad \text{ on } \quad [v - 2\varepsilon, v + 2\varepsilon].$$

Then, for any 1-Lipschitz function $\eta_{\rm Lip} : \mathbb{R} \to \mathbb{R}$, either $|\eta(y) - \eta_{\rm Lip}(y)| \geq \xi\varepsilon$ for $y \in [v + \varepsilon, v + 2\varepsilon]$ (if $\eta_{\rm Lip}(v) \leq \eta(v)$), or $|\eta(y) - \eta_{\rm Lip}(y)| \geq \xi\varepsilon$ for $y \in [v - 2\varepsilon, v - \varepsilon]$ (if $\eta_{\rm Lip}(v) \geq \eta(v)$). Thus, for any 1-Lipshictz function,

$$\mathbb{E}_{Y\sim p_Y}[(\eta_{\rm Lip}(Y) - \eta(Y))^2] \geq \xi\varepsilon \min\{p_Y([v + \varepsilon, v + 2\varepsilon]), p_Y([v - 2\varepsilon, v - \varepsilon])\} =: \Delta > 0. \tag{J.25}$$

Consider $\boldsymbol{\beta}_0 \in \mathbb{R}^p$ with coordinates distributed iid from $\pi/\sqrt{p}$ and $\boldsymbol{z} \sim \mathsf{N}(0, \boldsymbol{I}_p/p)$ independent of $\boldsymbol{\beta}_0$. Let $\boldsymbol{y} = \boldsymbol{\beta}_0 + \tau\boldsymbol{z}$. Clearly, $\sqrt{p}\boldsymbol{y}$ has coordinates distributed iid from $\pi * \mathsf{N}(0, \tau^2)$. We define the application of η to a vector by $\eta(\boldsymbol{y})_j = \frac{1}{\sqrt{p}}\eta(\sqrt{p}y_j)$. This agrees with (P.4) when $p = 1$, so no confusion should result. Observe that $\eta(\boldsymbol{y}) = \mathbb{E}_{\boldsymbol{\beta}_0,\boldsymbol{z}}[\boldsymbol{\beta}_0|\boldsymbol{y}]$. Because $\mathsf{prox}[\rho](\boldsymbol{y})_j - \eta(y_j)$ is uncorrelated with $\eta(y_j) - \beta_{0j}$ conditional on $\boldsymbol{y}_{-j}, \boldsymbol{\beta}_{0,-j}$ (where these denote the coordinates of $\boldsymbol{y}, \boldsymbol{\beta}_0$ excluding coordinate j), we have

$$\begin{aligned}\mathbb{E}_{\boldsymbol{\beta}_0,\boldsymbol{z}}\left[(\mathsf{prox}[\rho](\boldsymbol{y})_j - \beta_{0j})^2|\boldsymbol{y}_{-j}, \boldsymbol{\beta}_{0,-j}\right] &= \mathbb{E}_{\boldsymbol{\beta}_0,\boldsymbol{z}}\left[(\mathsf{prox}[\rho](\boldsymbol{y})_j - \eta(y_j))^2|\boldsymbol{y}_{-j}, \boldsymbol{\beta}_{0,-j}\right] \\ &\quad + \mathbb{E}_{\boldsymbol{\beta}_0,\boldsymbol{z}}\left[(\eta(y_j) - \beta_{0j})^2|\boldsymbol{y}_{-j}, \boldsymbol{\beta}_{0,-j}\right] \\ &= \mathbb{E}_{\boldsymbol{\beta}_0,\boldsymbol{z}}\left[(\mathsf{prox}[\rho](\boldsymbol{y})_j - \eta(y_j))^2|\boldsymbol{y}_{-j}, \boldsymbol{\beta}_{0,-j}\right] + \mathsf{mmse}_\pi(\tau^2)/p.\end{aligned} \tag{J.26}$$

For any lsc, proper, convex $\rho : \mathbb{R}^p \to \mathbb{R} \cup \{\infty\}$, fixing $\boldsymbol{y}_{-j}$ the function $y_j \mapsto \mathsf{prox}[\rho](\boldsymbol{y})$ is 1-Lipschitz, whence in fact $y_j \mapsto \mathsf{prox}[\rho](\boldsymbol{y})_j$ is 1-Lipschitz. Then by (J.25),

$$\mathbb{E}_{\boldsymbol{\beta}_0,\boldsymbol{z}}\left[(\mathsf{prox}[\rho](\boldsymbol{y})_j - \eta(y_j))^2|\boldsymbol{y}_{-j}, \boldsymbol{\beta}_{0,-j}\right] \geq \Delta/p, \text{ almost surely.}$$

We conclude

$$\mathbb{E}_{\boldsymbol{\beta}_0,\boldsymbol{z}}[\|\mathsf{prox}[\rho](\boldsymbol{y}) - \boldsymbol{\beta}_0\|^2] = \sum_{j=1}^p \mathbb{E}_{\boldsymbol{\beta}_0,\boldsymbol{z}}[\mathbb{E}_{\boldsymbol{\beta}_0,\boldsymbol{z}}[(\mathsf{prox}[\rho](\boldsymbol{y})_j - \beta_{0j})^2|\boldsymbol{y}_{-j}, \boldsymbol{\beta}_{0,-j}]] \geq \mathsf{mmse}_\pi(\tau^2) + \Delta.$$

The proof is complete. □

We are ready to prove the second part of Proposition 3.1.

Proof of Proposition 3.1.(ii). By Lemma J.1, we have $\mathsf{R}^{\mathsf{opt}}_{\mathsf{seq},\mathsf{cvx}}(\tau;\pi) > \mathsf{mmse}_\pi(\tau^2)$. By assumption, $\mathsf{mmse}_\pi(\tau^2) \geq \delta\tau^2 - \sigma^2$. Thus, $\mathsf{R}^{\mathsf{opt}}_{\mathsf{seq},\mathsf{cvx}}(\tau;\pi) > \delta\tau^2 - \sigma^2$. By Lemma C.2, the left and right-hand sides are continuous in τ, so that there exists $\tau' > \tau$ with $\mathsf{R}^{\mathsf{opt}}_{\mathsf{seq},\mathsf{cvx}}(\tau';\pi) > \delta\tau'^2 - \sigma^2$. Then by (2.11), $\tau_{\mathsf{reg},\mathsf{cvx}} \geq \tau' > \tau$. Proposition 3.1(ii) then follows from Theorem 1. □

Finally, we prove the third part of Proposition 3.1.

Proof of Proposition 3.1.(iii). If $\pi * \mathsf{N}(0, \tau^2_{\mathsf{reg},\mathsf{amp}*})$ has log concave density, so too does $\pi * \mathsf{N}(0, \tau^2)$ for all $\tau > \tau_{\mathsf{reg},\mathsf{amp}*}$. By the definition of $\tau_{\mathsf{reg},\mathsf{amp}*}$ (Eq. (2.15)), we have $\delta\tau^2 - \sigma^2 > \mathsf{mmse}_\pi(\tau^2)$ for all such τ. Then by Proposition 3.1(i), Theorem 1, and the fact that $\mathcal{C}_* \subset \mathcal{C}_{\delta,\pi}$, we have

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

Taking $\tau \downarrow \tau_{\mathsf{reg},\mathsf{amp}*}$ gives $\tau^2_{\mathsf{reg},\mathsf{amp}*} = \tau^2_{\mathsf{reg},\mathsf{cvx}}$.

If $\pi * \mathsf{N}(0, \tau^2_{\mathsf{reg},\mathsf{amp}*})$ does not have log concave density, then because $\delta\tau^2_{\mathsf{reg},\mathsf{amp}*} - \sigma^2 = \mathsf{mmse}_\pi(\tau^2_{\mathsf{reg},\mathsf{amp}*})$, we have $\tau^2_{\mathsf{reg},\mathsf{cvx}} > \tau^2_{\mathsf{reg},\mathsf{amp}*}$ by Proposition 3.1.

The argument for $\tau^2_{\mathsf{reg},\mathsf{stat}}$ is completely analogous. □

K Connection with the random signal and noise model

In this appendix we state and prove two lemmas that provide explicit connection between the deterministic and random signal and noise models. The first lemma will allow us to extend lower bounds on the lim inf of sequences of estimation errors.

Lemma K.1. *Fix $\pi \in \mathcal{P}_2(\mathbb{R})$, $\delta \in (0,\infty)$, and $\sigma \geq 0$. Consider any sequence of estimators $\{\widehat{\boldsymbol{\beta}}\}$ (ie. measurable functions of $\boldsymbol{y}, \boldsymbol{X}$ and potentially some auxiliary noise). Assume that the HDA and DSN assumptions imply that for some constant c we have*

$$\liminf_{p\to\infty}^{\mathrm{p}} \|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\|^2 \geq c. \tag{K.1}$$

Then under the HDA and RSN assumption (where the randomness in $\boldsymbol{\beta}_0$ and $\boldsymbol{w}$ is independent of the auxiliary noise used to construction $\widehat{\boldsymbol{\beta}}$), we have

$$\liminf_{p\to\infty} \mathbb{E}_{\boldsymbol{\beta}_0,\boldsymbol{w},\boldsymbol{X}}\left[\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\|^2\right] \geq c. \tag{K.2}$$

Proof of Lemma K.1. By [BF81, Lemma 8.4], if $\beta_{0j} \overset{\text{iid}}{\sim} \pi/\sqrt{p}$ for $\pi \in \mathcal{P}_2(\mathbb{R})$, then

$$d_{\mathrm{W}}(\widehat{\pi}_{\boldsymbol{\beta}_0}, \pi) \overset{\text{as}}{\to} 0, \tag{K.3}$$

where $\widehat{\pi}_{\boldsymbol{\beta}_0}$ is as in (2.1). Further, under assumption RSN, by the strong law of large numbers, $\frac{1}{n}\|\boldsymbol{w}\|^2 \overset{\text{as}}{\to} \sigma^2$. Thus, under the RSN assumption the sequences $\{\boldsymbol{\beta}_0\}$, $\{\boldsymbol{w}\}$ satisfy the DSN assumption with probability 1. Thus, if (K.1) holds under the DSN assumption, we have under the RSN assumption that for all $\varepsilon > 0$

$$\mathbb{P}_{\boldsymbol{\beta}_0,\boldsymbol{w},\boldsymbol{X}}\left(\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\|^2 > c - \varepsilon \;\middle|\; \boldsymbol{\beta}_0, \boldsymbol{w}\right) \xrightarrow[p\to\infty]{\text{as}} 1. \tag{K.4}$$

Observe by bounded convergence

$$\begin{aligned}\mathbb{E}_{\boldsymbol{\beta}_0,\boldsymbol{w}}[\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\|^2] &= \mathbb{E}_{\boldsymbol{\beta}_0,\boldsymbol{w},\boldsymbol{X}}\left[\mathbb{E}_{\boldsymbol{\beta}_0,\boldsymbol{w},\boldsymbol{X}}\left[\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\|^2 \;\middle|\; \boldsymbol{\beta}_0, \boldsymbol{w}\right]\right] \\ &\geq (c-\varepsilon)\mathbb{E}_{\boldsymbol{\beta}_0,\boldsymbol{w},\boldsymbol{X}}\left[\mathbb{P}_{\boldsymbol{\beta}_0,\boldsymbol{w},\boldsymbol{X}}\left(\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\|^2 > c - \varepsilon \;\middle|\; \boldsymbol{\beta}_0, \boldsymbol{w}\right)\right] \to c - \varepsilon.\end{aligned}$$

Taking $\varepsilon \downarrow 0$ gives (K.2). □

Observe that Lemma K.1 applies to any sequence of estimators $\{\widehat{\boldsymbol{\beta}}\}$ defined in any way. In particular, the estimators need not be defined via convex M-estimation. The second lemma will allow us to extend the exact loss characterization of Proposition B.3.

Lemma K.2. *Fix $\pi \in \mathcal{P}_2(\mathbb{R})$, $\delta \in (0,\infty)$, and $\sigma \geq 0$. Consider a sequence $\{\rho_p\} \in \mathcal{C}_*$ and the corresponding M-estimators (1.2) (which always exist and are unique by strong convexity). Assume that the HDA and DSN assumptions imply that for some constant c we have*

$$\|\widehat{\boldsymbol{\beta}}_{\mathsf{cvx}} - \boldsymbol{\beta}_0\|^2 \overset{\mathrm{p}}{\to} c. \tag{K.5}$$

Then under the HDA and RSN assumption

$$\lim_{p\to\infty} \mathbb{E}_{\boldsymbol{\beta}_0,\boldsymbol{w},\boldsymbol{X}}\left[\|\widehat{\boldsymbol{\beta}}_{\mathsf{cvx}} - \boldsymbol{\beta}_0\|^2\right] = c. \tag{K.6}$$

Proof of Lemma K.2. Let $\gamma > 0$ be such that ρ_p is strongly convex with parameter γ for all p. Because ρ_p is strongly convex, it has a unique minimizer, which we will denote by $\boldsymbol{m}_p$. First we show that $\|\boldsymbol{m}_p\|$ is bounded in p. Without loss of generality, we may assume $\rho_p(\boldsymbol{m}_p) = 0$ for all p. Thus, $\rho_p(\boldsymbol{\beta}) \geq \frac{\gamma}{2}\|\boldsymbol{\beta} - \boldsymbol{m}_p\|^2$ for all p and all $\boldsymbol{\beta} \in \mathbb{R}^p$. By (1.2),

$$\frac{1}{n}\|\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}_{\mathsf{cvx}}\|^2 + \frac{\gamma}{2}\|\widehat{\boldsymbol{\beta}}_{\mathsf{cvx}} - \boldsymbol{m}_p\|^2 \leq \frac{1}{n}\|\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}_{\mathsf{cvx}}\|^2 + \rho_p(\widehat{\boldsymbol{\beta}}_{\mathsf{cvx}}) \leq \frac{1}{n}\|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{m}_p\|^2.$$

By optimality, we have that $\frac{2}{n}\boldsymbol{X}^\mathsf{T}(\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}_{\mathsf{cvx}}) \in \partial\rho_p(\widehat{\boldsymbol{\beta}}_{\mathsf{cvx}})$. Thus,

$$\begin{aligned}\rho_p(\widehat{\boldsymbol{\beta}}_{\mathsf{cvx}}) \geq \rho_p(\boldsymbol{m}_p) &\geq \rho_p(\widehat{\boldsymbol{\beta}}_{\mathsf{cvx}}) + \frac{2}{n}(\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}_{\mathsf{cvx}})^\mathsf{T}\boldsymbol{X}(\boldsymbol{m}_p - \widehat{\boldsymbol{\beta}}_{\mathsf{cvx}}) + \frac{\gamma}{2}\|\boldsymbol{m}_p - \widehat{\boldsymbol{\beta}}_{\mathsf{cvx}}\|^2 \\ &\geq \rho_p(\widehat{\boldsymbol{\beta}}_{\mathsf{cvx}}) - \frac{2}{n}\|\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}_{\mathsf{cvx}}\|\|\boldsymbol{X}\|_{\mathsf{op}}\|\boldsymbol{m}_p - \widehat{\boldsymbol{\beta}}_{\mathsf{cvx}}\| + \frac{\gamma}{2}\|\boldsymbol{m}_p - \widehat{\boldsymbol{\beta}}_{\mathsf{cvx}}\|^2.\end{aligned}$$

Also,

$$\|\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}_{\mathsf{cvx}}\| \leq \|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}_0\| + \|\boldsymbol{X}(\widehat{\boldsymbol{\beta}}_{\mathsf{cvx}} - \boldsymbol{\beta}_0)\| \leq \|\boldsymbol{w}\| + \|\boldsymbol{X}\|_{\mathsf{op}}\|\widehat{\boldsymbol{\beta}}_{\mathsf{cvx}} - \boldsymbol{\beta}_0\|.$$

Combining the previous two displays,

$$\|\boldsymbol{m}_p - \widehat{\boldsymbol{\beta}}_{\mathsf{cvx}}\| \leq \frac{4}{\gamma}\frac{\|\boldsymbol{X}\|_{\mathsf{op}}}{\sqrt{n}}\left(\frac{\|\boldsymbol{w}\|}{\sqrt{n}} + \frac{\|\boldsymbol{X}\|_{\mathsf{op}}}{\sqrt{n}}\|\widehat{\boldsymbol{\beta}}_{\mathsf{cvx}} - \boldsymbol{\beta}_0\|\right).$$

In particular,

$$\begin{aligned}\|\boldsymbol{m}_p\| &\leq \|\boldsymbol{\beta}_0\| + \|\widehat{\boldsymbol{\beta}}_{\mathsf{cvx}} - \boldsymbol{\beta}_0\| + \|\boldsymbol{m}_p - \widehat{\boldsymbol{\beta}}_{\mathsf{cvx}}\| \\ &\leq \|\boldsymbol{\beta}_0\| + \|\widehat{\boldsymbol{\beta}}_{\mathsf{cvx}} - \boldsymbol{\beta}_0\| + \frac{4}{\gamma}\frac{\|\boldsymbol{X}\|_{\mathsf{op}}}{\sqrt{n}}\left(\frac{\|\boldsymbol{w}\|}{\sqrt{n}} + \frac{\|\boldsymbol{X}\|_{\mathsf{op}}}{\sqrt{n}}\|\widehat{\boldsymbol{\beta}}_{\mathsf{cvx}} - \boldsymbol{\beta}_0\|\right).\end{aligned}$$

The random variable $\|\boldsymbol{X}\|_{\mathsf{op}}/\sqrt{n}$ is tight by [AGZ10], the random variable $\|\boldsymbol{w}\|/\sqrt{n}$ is tight by the law of large numbers, and $\|\widehat{\boldsymbol{\beta}}_{\mathsf{cvx}} - \boldsymbol{\beta}_0\|$ is tight under the DSN assumption by assumption. Because $\|\boldsymbol{m}_p\|$ is deterministic, it must be bounded in p. Let M be such that $\|\boldsymbol{m}_p\| \leq M$ for all p.

Now we turn to proving (K.6) under the RSN assumption. As in the proof of Lemma K.1, we have that the sequences $\{\boldsymbol{\beta}_0\}, \{\boldsymbol{w}\}$ satisfy the DSN assumption with probability 1. Thus, we have (K.5). By Vitali's convergence theorem, we only need to verify that $\|\widehat{\boldsymbol{\beta}}_{\mathsf{cvx}} - \boldsymbol{\beta}_0\|^2$ is uniformly integrable over p [Bil12, Theorem 16.14]. Observe that for any $\boldsymbol{\beta} \in \mathbb{R}^p$,

$$\begin{aligned}\frac{1}{n}\|\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}_{\mathsf{cvx}}\|^2 + \rho_p(\widehat{\boldsymbol{\beta}}_{\mathsf{cvx}}) &\geq \frac{1}{n}\|\boldsymbol{y}\|^2 - \frac{2}{n}\boldsymbol{y}^\mathsf{T}\boldsymbol{X}\widehat{\boldsymbol{\beta}}_{\mathsf{cvx}} + \frac{\gamma}{2}\|\widehat{\boldsymbol{\beta}}_{\mathsf{cvx}} - \boldsymbol{m}_p\|^2 \\ &\geq \frac{1}{n}\|\boldsymbol{y}\|^2 - 2\frac{\|\boldsymbol{X}^\mathsf{T}\boldsymbol{y}\|}{n}\|\widehat{\boldsymbol{\beta}}_{\mathsf{cvx}}\| + \frac{\gamma}{2}\|\widehat{\boldsymbol{\beta}}_{\mathsf{cvx}}\|^2 - \gamma M\|\widehat{\boldsymbol{\beta}}_{\mathsf{cvx}}\| \\ &\geq \frac{1}{n}\|\boldsymbol{y}\|^2 + \frac{\gamma}{4}\|\widehat{\boldsymbol{\beta}}_{\mathsf{cvx}}\|^2 - \frac{1}{\gamma}\left(2\frac{\|\boldsymbol{X}^\mathsf{T}\boldsymbol{y}\|}{n} + \gamma M\right)^2 \\ &\geq \frac{1}{n}\|\boldsymbol{y}\|^2 + \frac{\gamma}{4}\|\widehat{\boldsymbol{\beta}}_{\mathsf{cvx}}\|^2 - \frac{8\|\boldsymbol{X}^\mathsf{T}\boldsymbol{y}\|^2}{\gamma n^2} - 2\gamma M^2.\end{aligned}$$

Further, recalling $\rho_p(\boldsymbol{m}_p) = 0$, by (1.2) and the triangle inequality

$$\begin{aligned}\frac{1}{n}\|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{m}_p\|^2 + \rho_p(\boldsymbol{m}_p) = \frac{1}{n}\|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{m}_p\|^2 &\leq \frac{2}{n}\|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}_0\|^2 + \frac{2}{n}\|\boldsymbol{X}(\boldsymbol{m}_p - \boldsymbol{\beta}_0)\|^2 \\ &= \frac{2}{n}\|\boldsymbol{w}\|^2 + \frac{2}{n}\|\boldsymbol{X}(\boldsymbol{m}_p - \boldsymbol{\beta}_0)\|^2.\end{aligned}$$

Combining the previous two displays, we get

$$\begin{aligned}
\|\widehat{\boldsymbol{\beta}}_{\mathsf{cvx}}\|^2 &\le \frac{4}{\gamma}\left(\frac{1}{n}\|\boldsymbol{y}-\boldsymbol{X}\widehat{\boldsymbol{\beta}}_{\mathsf{cvx}}\|^2+\rho(\widehat{\boldsymbol{\beta}}_{\mathsf{cvx}})-\frac{1}{n}\|\boldsymbol{y}\|^2+\frac{8\|\boldsymbol{X}^\mathsf{T}\boldsymbol{y}\|^2}{\gamma n^2}+2\gamma M^2\right)\\
&\le \frac{4}{\gamma}\left(\frac{1}{n}\|\boldsymbol{y}-\boldsymbol{X}\boldsymbol{m}_p\|^2+\rho(\boldsymbol{m}_p)+\frac{8\|\boldsymbol{X}^\mathsf{T}\boldsymbol{y}\|^2}{\gamma n^2}+2\gamma M^2\right)\\
&\le \frac{4}{\gamma}\left(\frac{2}{n}\|\boldsymbol{w}\|^2+\frac{2}{n}\|\boldsymbol{X}(\boldsymbol{m}_p-\boldsymbol{\beta}_0)\|^2+\frac{8}{\gamma}\left(\frac{\|\boldsymbol{X}^\mathsf{T}\boldsymbol{w}\|}{n}+\frac{\|\boldsymbol{X}^\mathsf{T}\boldsymbol{X}\boldsymbol{\beta}_0\|}{n}\right)^2+2\gamma M^2\right)\\
&\le \frac{4}{\gamma}\left(\frac{2}{n}\|\boldsymbol{w}\|^2+\frac{4}{n}\|\boldsymbol{X}\boldsymbol{m}_p\|^2+\frac{4}{n}\|\boldsymbol{X}\boldsymbol{\beta}_0\|^2+\frac{16}{\gamma n^2}\|\boldsymbol{X}^\mathsf{T}\boldsymbol{w}\|^2+\frac{16}{\gamma n^2}\|\boldsymbol{X}^\mathsf{T}\boldsymbol{X}\boldsymbol{\beta}_0\|^2+2\gamma M^2\right).
\end{aligned} \tag{K.7}$$

We show the right-hand side is uniformly integrable one term at a time. First, we recall two well-known facts about uniform integrability, which we state without proof.

Claim K.3. *If the collection (over j) $\{A_j\}$ is uniformly integrable, then the collection (over p) $\left\{\frac{1}{p}\sum_{i=1}^p A_i\right\}_p$ is uniformly integrable.*

Claim K.4. *If $\{A_p\}$ and $\{B_p\}$ are uniformly integrable and for each p the random variables A_p and B_p are defined on the same probability space and are independent, then $\{A_pB_p\}$ are uniformly integrable.*

First, the $\frac{2}{n}\|\boldsymbol{w}\|^2$ are uniformly integrable by Claim K.3 because the w_j^2 are integrable from the same distribution. Second, $\frac{4}{n}\|\boldsymbol{X}\boldsymbol{m}_p\|^2 \sim 4\|\boldsymbol{m}_p\|^2\chi_n^2/n \overset{\mathrm{d}}{=} \frac{4\|\boldsymbol{m}_p\|^2}{n}\sum_{i=1}^n Z_i^2$, where $Z_i \overset{\mathrm{iid}}{\sim} \mathsf{N}(0,1)$. By Claim K.3, the $\frac{1}{n}\sum_{i=1}^n Z_i^2$ are uniformly integrable, and because the $\|\boldsymbol{m}_p\|^2$ are bounded, the $\frac{4\|\boldsymbol{m}_p\|^2}{n}\sum_{i=1}^n Z_i^2$, and hence the $\frac{4}{n}\|\boldsymbol{X}\boldsymbol{m}_p\|^2$, are the uniformly integrable by Claim K.4. Third, $\frac{4}{n}\|\boldsymbol{X}\boldsymbol{\beta}_0\|^2 = \frac{4}{n}\sum_{i=1}^n\left(\sum_{j=1}^p X_{ij}\beta_{0j}\right)^2$. Observe that conditional on $\boldsymbol{\beta}_0$, the random variable $\sum_{j=1}^p X_{ij}\beta_{0j}$ has distribution $\mathsf{N}(0,\|\boldsymbol{\beta}_0\|^2)$, so that $\left(\sum_{j=1}^p X_{ij}\beta_{0j}\right)^2 \overset{\mathrm{d}}{=} Z^2\|\boldsymbol{\beta}_0\|^2$ for $Z\sim\mathsf{N}(0,1)$ independent of $\boldsymbol{\beta}_0$. Observe that the $\|\boldsymbol{\beta}_0\|^2 = \frac{1}{p}\sum_{j=1}^p(\sqrt{p}\beta_{0j})^2$ are uniformly integrable (over p) by Claim K.3 because $\sqrt{p}\beta_{0j}\sim\pi\in\mathcal{P}_2(\mathbb{R})$ for all p. Then, by Claim K.4, the $Z^2\|\boldsymbol{\beta}_0\|^2$, and hence the $\left(\sum_{j=1}^p X_{ij}\beta_{0j}\right)^2$, are uniformly integrable. Then, by Claim K.3, the $\frac{4}{n}\|\boldsymbol{X}\boldsymbol{\beta}_0\|^2$ are uniformly integrable. Fourth, the $\frac{16}{\gamma n^2}\|\boldsymbol{X}^\mathsf{T}\boldsymbol{w}\|^2$ are uniformly integrable by the same argument (just replace $\boldsymbol{\beta}_0$ with $\boldsymbol{w}/\sqrt{n}$ and switch i,j and n,p). Fifth, and lastly, we show the $\frac{16}{\gamma n^2}\|\boldsymbol{X}^\mathsf{T}\boldsymbol{X}\boldsymbol{\beta}_0\|^2$ are uniformly integrable. We have

$$\begin{aligned}
\frac{\|\boldsymbol{X}^\mathsf{T}\boldsymbol{X}\boldsymbol{\beta}_0\|^2}{n^2} &= \frac{1}{n}\sum_{j=1}^p\left(\frac{1}{\sqrt{n}}\sum_{i=1}^n X_{ij}[\boldsymbol{X}\boldsymbol{\beta}_0]_i\right)^2 = \frac{1}{n}\sum_{j=1}^p\left(\frac{1}{\sqrt{n}}\sum_{i=1}^n X_{ij}^2\beta_{0j}+\frac{1}{\sqrt{n}}\sum_{i=1}^n\sum_{l\neq j}^p X_{ij}X_{il}\beta_{0l}\right)^2\\
&\le \underbrace{2\sum_{j=1}^p\left(\frac{1}{n}\sum_{i=1}^n X_{ij}^2\beta_{0j}\right)^2}_{:=a}+\underbrace{\frac{2}{n}\sum_{j=1}^p\left(\frac{1}{\sqrt{n}}\sum_{i=1}^n\sum_{l\neq j}^p X_{ij}X_{il}\beta_{0l}\right)^2}_{:=b}.
\end{aligned}$$

We write a as $\frac{2}{n^2 p}\sum_{j=1}^{p}\sum_{i_1,i_2=1}^{n} X_{i_1 j}^2 X_{i_2 j}^2(\sqrt{p}\beta_{0j})^2$, which are uniformly integrable by Claim K.3 because the $X_{i_1 j}^2 X_{i_2 j}^2(\sqrt{p}\beta_{0j})^2$ are integrable and have one of only two possible distributions (depending on whether $i_1 = i_2$ or $i_1 \neq i_2$) which do not depend on n, p. Now we consider b. We denote the columns of $\boldsymbol{X}$ by $\boldsymbol{X}_j$. Observe that conditional on $\boldsymbol{X}_j, \boldsymbol{\beta}_0$, we have $\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\sum_{l\neq j}^{p} X_{ij}X_{il}\beta_{0l} \sim \mathsf{N}\left(0, \left(\sum_{l\neq j}^{p}\beta_{0l}^2\right)\frac{\|\boldsymbol{X}_j\|^2}{n}\right)$, so that in fact, $\left(\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\sum_{l\neq j}^{p} X_{ij}X_{il}\beta_{0l}\right)^2 \overset{\mathrm{d}}{=} Z^2\left(\sum_{l\neq j}^{p}\beta_{0l}^2\right)\frac{\|\boldsymbol{X}_j\|^2}{n}$ for $Z \sim \mathsf{N}(0,1)$ independent of $\boldsymbol{\beta}_0, \boldsymbol{X}$. Observe that the $\left(\sum_{l\neq j}^{p}\beta_{0l}^2\right)$ are uniformly integrable because they are dominated by $\|\boldsymbol{\beta}_0\|^2$, whose uniform integrability we already established. Further, the $\frac{\|\boldsymbol{X}_j\|^2}{n} = \frac{1}{n}\sum_{i=1}^{n} X_{ij}^2$ are uniformly integrable by Claim K.3. Then the $Z^2\left(\sum_{l\neq j}^{p}\beta_{0l}^2\right)\frac{\|\boldsymbol{X}_j\|^2}{n}$ are uniformly integrable by two applications of Claim K.4, because they are the product of three independent and uniformly integrable terms. Thus, the b's are uniformly integrable by Claim K.3, and the $\frac{\|\boldsymbol{X}^{\mathsf{T}}\boldsymbol{X}\boldsymbol{\beta}_0\|^2}{n^2}$ are uniformly integrable by the uniform integrability of the a's and b's. We conclude that the right-hand side of (K.7) is uniformly integrable, whence the $\|\widehat{\boldsymbol{\beta}}_{\mathsf{cvx}}\|^2$ are uniformly integrable.

Because

$$\|\widehat{\boldsymbol{\beta}}_{\mathsf{cvx}} - \boldsymbol{\beta}_0\|^2 \le 2\left(\|\widehat{\boldsymbol{\beta}}_{\mathsf{cvx}}\|^2 + \|\boldsymbol{\beta}_0\|^2\right),$$

and $\|\boldsymbol{\beta}_0\|^2$ are uniformly integrable, we also have the $\|\widehat{\boldsymbol{\beta}}_{\mathsf{cvx}} - \boldsymbol{\beta}_0\|^2$ are uniformly integrable, completing the proof. □

L Proof of Proposition 2.6, Proposition 2.4, and equivalence of $\tau_{\mathsf{reg,amp}*}$ and $\tau_{\mathsf{reg,amp}}$

Proof of Proposition 2.6. Model [BKM+19, eq. (1)] is equivalent to our model (1.1) under the following change of variable (with the notation of [BKM+19] on the left).

$$\begin{aligned}
&\boldsymbol{\Phi} \leftarrow \boldsymbol{X}, && \boldsymbol{X}^* \leftarrow \sqrt{p}\boldsymbol{\beta}_0, && Y_\mu \leftarrow y_i, && A_\mu \leftarrow w_i, && (m,n) \leftarrow (n,p),\\
&\varphi(x,a) = x + a, && P_0 \leftarrow \pi, && \alpha \leftarrow \delta, && r \leftarrow 1/\tau^2, && \rho \leftarrow s_2(\pi),
\end{aligned}$$

where we have used an equal sign for any quantity which we do not have our own notation for. The authors of [BKM+19] denote by X_0, Z_0 independent random scalars distributed from P_0 and $\mathsf{N}(0,1)$ respectively. In our notation, we denote by β_0, z independent random scalars distributed from π and $\mathsf{N}(0,1)$ respectively. We will also denote the random scalar $y = \beta_0/\tau + z$. To avoid clutter, we will write s_2 in place of $s_2(\pi)$ for the remainder of the proof. The authors of [BKM+19] define in Eq. (5) (where we have already converted to our notation)

$$\begin{aligned}
\psi_\pi(1/\tau^2) &= \mathbb{E}_{\beta_0,z}\log\int e^{\beta_0\beta/\tau^2 + \beta z/\tau - \beta^2/2\tau^2}\pi(\mathrm{d}\beta) = \mathbb{E}_{\beta_0,z}\log\left(e^{\frac{1}{2}(\beta_0/\tau+z)^2}\int e^{-\frac{1}{2}(\beta_0/\tau+z-\beta/\tau)^2}\right)\pi(\mathrm{d}\beta)\\
&= \frac{s_2}{2\tau^2} + \frac{1}{2} + \mathbb{E}_{\beta_0,z}\int e^{-\frac{1}{2}(y-\beta/\tau)^2}\pi(\mathrm{d}\beta) = \frac{s_2}{2\tau^2} - i(\tau^2), \qquad \text{(L.1)}
\end{aligned}$$

where the last line follows by (2.29). Their $P_{\mathsf{out}}\left(Y_\mu \middle| \frac{1}{\sqrt{n}}[\boldsymbol{\Phi}\boldsymbol{X}^*]_\mu\right)$ is the conditional density (w.r.t. Lebesgue measure) of $Y_\mu \mid \frac{1}{\sqrt{n}}[\boldsymbol{\Phi}\boldsymbol{X}^*]_\mu$ (in their notation), which in our notation is $P_{\mathsf{out}}(y|x) =$

$\frac{1}{\sqrt{2\pi}\sigma}\exp\left(-\frac{1}{2\sigma^2}(y-x)^2\right)$. The authors of [BKM+19] denote by V, W independent random scalars distributed from $\mathsf{N}(0,1)$ and by $\tilde{Y}_0$ a random scalar distributed from $P_{\mathsf{out}}(\cdot|\sqrt{q}V+\sqrt{s_2-q}W)$. In our notation and with our choice of P_{out}, we denote by z_1, z_2 independent random scalars distributed from $\mathsf{N}(0,1)$ (corresponding to V, W respectively) and observe that $\sqrt{q}z_1+\sqrt{s_2-q}z_2+\sigma z_3 \sim P_{\mathsf{out}}(\cdot|\sqrt{q}z_1+\sqrt{s_2-q}z_2)$ where $z_3 \sim \mathsf{N}(0,1)$ independent of z_1, z_2. The authors of [BKM+19] define in Eq. (6) (where we have already converted to our notation)

$$\begin{aligned}
\Psi_{P_{\mathsf{out}}}(q;s_2) &= \mathbb{E}_{z_1,z_2,z_3}\log\int\frac{1}{\sqrt{2\pi}}e^{-\frac{1}{2}w^2}P_{\mathsf{out}}\left(\sqrt{q}z_1+\sqrt{s_2-q}z_2+\sigma z_3|\sqrt{q}z_1+\sqrt{s_2-q}w\right)\mathrm{d}w \\
&= \mathbb{E}_{z_1,z_2,z_3}\log\int\frac{1}{\sqrt{2\pi}}e^{-\frac{1}{2}w^2}\frac{1}{\sqrt{2\pi}\sigma}\exp\left(-\frac{1}{2\sigma^2}\left(\sqrt{s_2-q}z_2+\sigma z_3-\sqrt{s_2-q}w\right)^2\right)\mathrm{d}w \\
&= \mathbb{E}_{z_2,z_3}\log\left(\frac{1}{\sqrt{2\pi(\sigma^2+s_2-q)}}\exp\left(-\frac{1}{2(\sigma^2+s_2-q)}\left(\sqrt{s_2-q}z_2+\sigma z_3\right)^2\right)\right) \\
&= -\frac{1}{2}\log 2\pi-\frac{1}{2}\log(\sigma^2+s_2-q)-\frac{1}{2}. \qquad \text{(L.2)}
\end{aligned}$$

The authors of [BKM+19] define in Eq. (4)

$$f_{\mathsf{RS}}(q,1/\tau^2;s_2)=\psi_\pi(1/\tau^2)+\delta\Psi_{P_{\mathsf{out}}}(q;s_2)-\frac{q}{2\tau^2}. \qquad \text{(L.3)}$$

In Eq. (3) they define a parameter q^* via a variational formula which in Theorem 1 they show to be equivalent to defining q^* as the first coordinate of

$$\arg\max_{(q,\tau)\in\Gamma} f_{\mathsf{RS}}(q,1/\tau^2;s_2), \qquad \text{(L.4)}$$

whenever maximizers exist and the first coordinate of maximizing pairs is unique, where

$$\Gamma=\left\{(q,\tau)\in[0,s_2]\times[0,\infty]\;\middle|\;\frac{\mathrm{d}}{\mathrm{d}q}f_{\mathsf{RS}}(q,1/\tau^2;s_2)=\frac{\mathrm{d}}{\mathrm{d}\tau^{-2}}f_{\mathsf{RS}}(q,1/\tau^2;s_2)=0\right\}.$$

Some calculus applied to (L.1), (L.2), and (L.3) shows that $\frac{\mathrm{d}}{\mathrm{d}q}f_{\mathsf{RS}}(q,1/\tau^2;s_2)=0$ if and only if $s_2-q=\delta\tau^2-\sigma^2$. That is,

$$(q,\tau)\in\Gamma \Rightarrow q=s_2-\delta\tau^2+\sigma^2. \qquad \text{(L.5)}$$

We claim that maximizing f_{RS} over Γ is equivalent to maximizing f_{RS} over the larger set $q=s_2-\delta\tau^2+\sigma^2$, as we now show. By (L.1), (L.2), and (L.3), we have

$$\begin{aligned}
f_{\mathsf{RS}}(s_2-\delta\tau^2+\sigma^2,1/\tau^2;s_2) &= \frac{s_2}{2\tau^2}-i(\tau^2)-\frac{\delta}{2}\log 2\pi-\frac{\delta}{2}\log(\delta\tau^2)-\frac{\delta}{2}-\frac{s_2-\delta\tau^2+\sigma^2}{2\tau^2} \\
&= -\left(\frac{\sigma^2}{2\tau^2}-\frac{\delta}{2}\log\left(\frac{\sigma^2}{\tau^2}\right)+i(\tau^2)\right)+C=-\phi(\tau^2)+C, \qquad \text{(L.6)}
\end{aligned}$$

where C is a constant which depends only on δ, σ^2 and numerical constants. For $\tau\to 0$ and $\tau\to\infty$, we see from (L.6) that $f_{\mathsf{RS}}(s_2-\delta\tau^2+\sigma^2,1/\tau^2;s_2)$ goes to $-\infty$, so that it is maximized at a point for which $\frac{\mathrm{d}}{\mathrm{d}\tau^{-2}}f_{\mathsf{RS}}(\delta\tau^2-\sigma^2,1/\tau^2;s_2)=0$. Because $\frac{\mathrm{d}}{\mathrm{d}q}f_{\mathsf{RS}}(q,1/\tau^2;s_2)\Big|_{q=s_2-\delta\tau^2+\sigma^2}=0$, we have that $f_{\mathsf{RS}}(s_2-\delta\tau^2+\sigma^2,1/\tau^2;s_2)$ is maximized at a τ for which $\frac{\mathrm{d}}{\mathrm{d}\tau^{-2}}f_{\mathsf{RS}}(q,1/\tau^2;s_2)\Big|_{q=s_2-\delta\tau^2+\sigma^2}=0$.

That is, any maximizer (q, τ) of f_{RS} which satisfies $q = s_2 - \delta\tau^2 + \sigma^2$ must lie in Γ, as claimed. In particular, maximizing f_{RS} over the weaker constraint $q = s_2 - \delta\tau^2 + \sigma^2$ yields the same maximizers as maximizing f_{RS} over the stronger constraint $(q, \tau) \in \Gamma$.

To summarize, all solutions to (L.4) are constructed in the following way: let $\tau^* \in \arg\max_\tau\{-\phi(\tau^2)\} = \arg\min_\tau \phi(\tau^2)$ and let $q^* = s_2 - \delta\tau^2 + \sigma^2$. Further, by (L.5), when τ^* is the unique minimizer of ϕ, we have q^* is the unique first coordinate of maximizers of (L.4). We see that $\tau^* = \tau_{\mathsf{reg,stat}}$. After converting into our notation, Theorem 2 of [BKM+19] and their Eq. (8) state that under certain assumptions which we will list, $\lim_{p\to\infty} \mathbb{E}_{\boldsymbol{\beta}_0,\boldsymbol{w},\boldsymbol{X}}\left[\|\mathbb{E}_{\boldsymbol{\beta}_0,\boldsymbol{w},\boldsymbol{X}}[\boldsymbol{\beta}_0|\boldsymbol{y},\boldsymbol{X}] - \boldsymbol{\beta}_0\|^2\right] = s_2 - q^*$. The assumptions they require are that $\pi \in \mathcal{P}_\infty(\mathbb{R})$, $\mathbb{E}_{\boldsymbol{\beta}_0,\boldsymbol{w},\boldsymbol{X}}\left[\left|\sum_{j=1}^p X_{1j}\beta_{0j} + w_1\right|^{2+\gamma}\right]$ is bounded for some $\gamma > 0$ (for us, it is bounded for all $\gamma > 0$), the function φ is continuous, $\sigma > 0$, and the minimizer q^* is unique. These are all satisfied in our setting when the minimizer of ϕ is unique. By equation (2.32) and because $s_2 - q^* = \delta\tau^2_{\mathsf{reg,stat}} - \sigma^2$, equation (2.33) follows.

Finally, by [BKM+19, Theorm 2], we have for fixed σ^2 that the maximizer q^* is unique for almost every δ (w.r.t. Lebesuge measure). By Fubini's theorem, this holds for almost every (δ, σ) (w.r.t. Lebesgue measure). □

In the proof of Corollary 2.3 in Section 2, we use the following claim.

Claim L.1. *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).*

Proof of Claim L.1. Comparing (2.19) and (2.15), we see $\tau^2_{\mathsf{reg,amp}*} \geq \tau^2_{\mathsf{reg,amp}}$ always. Consider the case that $\tau_{\mathsf{reg,amp}} < \tau_{\mathsf{reg,amp}*}$. By (2.19), for all $\tau \in (\tau_{\mathsf{reg,amp}}, \tau_{\mathsf{reg,amp}*}]$ we have $\delta\tau^2 - \sigma^2 \geq \mathsf{mmse}_\pi(\tau^2)$. By (2.15), for all $\tau > \tau_{\mathsf{reg,amp}*}$, we have $\delta\tau^2 - \sigma^2 > \mathsf{mmse}_\pi(\tau^2)$. By the continuity of $\mathsf{mmse}_\pi(\tau^2)$ [DYSV11], we have $\delta\tau^2_{\mathsf{reg,amp}*} - \sigma^2 = \mathsf{mmse}_\pi(\tau^2_{\mathsf{reg,amp}*})$. Combining these three observations, we conclude by the differentiability of $\mathsf{mmse}_\pi(\tau^2)$ at $\tau_{\mathsf{reg,amp}*} > 0$ [DYSV11] that $\delta = \frac{\mathrm{d}}{\mathrm{d}\tau^2}\mathsf{mmse}_\pi(\tau^2)\Big|_{\tau=\tau^2_{\mathsf{reg,amp}*}}$ and $\sigma^2 = \delta\tau^2_{\mathsf{reg,amp}*} - \mathsf{mmse}_\pi(\tau^2_{\mathsf{reg,amp}*})$. Thus, the set of δ, σ^2 for which $\tau^2_{\mathsf{reg,amp}} = \tau^2_{\mathsf{reg,amp}*}$ holds is contained within the set

$$\left\{\left(\frac{\mathrm{d}}{\mathrm{d}\tau^2}\mathsf{mmse}_\pi(\tau^2), \tau^2\frac{\mathrm{d}}{\mathrm{d}\tau^2}\mathsf{mmse}_\pi(\tau^2) - \mathsf{mmse}_\pi(\tau^2)\right) \;\middle|\; \tau^2 > 0\right\},$$

which has Lebesgue measure 0 because mmse_π is infinitely differentiable [DYSV11]. □

Proof of Proposition 2.4. We will prove the proposition under the DSN assumption. Because the DSN assumption holds almost surely under the RSN assumption, the proposition also holds under the RSN assumption.

The proposition is nearly an instance of Theorem 14 of [BMN19], except that η_t as we have defined it need not be Lipschitz continuous, which is required by [BMN19]. Versions of Proposition 2.4 appear elsewhere in the literature (e.g., [BMN19, BMDK17]), though often without proof, citing works in which state evolution for AMP is established for Lipschitz denoisers [BM11, JM13, BMN19]. For the sake of completeness, we address here the minor technical difficulty that arises when η_t is not Lipschitz using a truncation technique which is standard in the AMP literature.

The truncation argument requires the following lemma.

Lemma L.2. *There exist constants $C_t > 0$ such that for each t, $|\eta_t(y)| \leq C_t(1 + |y|)$.*

Proof. Assume K is such that $\pi([-K,K]) \geq 1/2$. For $y > K$, we have

$$\begin{aligned}\mathbb{E}_{\beta_0,z}[\beta_0|\beta_0+\tau_t z = y] &\leq y + \int_0^\infty \mathbb{P}_{\beta_0,z}(\beta_0 \geq y+t|\beta_0+\tau_t z = y)\mathrm{d}t \\ &\leq 2y + K + \int_0^\infty \frac{\int_{2y+K+t}^\infty \exp(-(y-s)^2/(2\tau_t^2))\pi(\mathrm{d}s)}{\int_{-\infty}^\infty \exp(-(y-s)^2/(2\tau_t^2))\pi(\mathrm{d}s)}\mathrm{d}t \\ &\leq 2y + K + \frac{\exp(-(y+K)^2/(2\tau_t^2))}{\exp(-(y+K)^2/(2\tau_t^2))/2} = 2y+K+2.\end{aligned}$$

A similar argument shows that for $y < -K$, $\mathbb{E}_{\beta_0,z}[\beta_0|\beta_0+\tau_t z = y] \geq 2y - K - 2$. This establishes the lemma. □

Define $\eta_{M,t}(y) = \eta_t(y)\mathbf{1}_{|y|\leq M} + \eta_t(M)\mathbf{1}_{y>M} + \eta_t(-M)\mathbf{1}_{y<M}$. The reason for defining this truncation is that, because η_t has continuous first derivative, $\eta_{M,t}$ is Lipschitz continuous.

Define $\tau_{M,0}^2 = \tau_0^2$ and

$$\tau_{M,t+1}^2 = \frac{1}{\delta}(\sigma^2 + \mathbb{E}_{\beta_0,z}[(\eta_{M,t}(\beta_0+\tau_{M,t}z) - \beta_0)^2]), \quad t \geq 0, \tag{L.7}$$

$$\mathsf{b}_{M,t} = \frac{1}{\delta}\mathbb{E}_{\beta_0,z}[\eta'_{M,t-1}(\beta_0+\tau_{M,t-1}z)]. \tag{L.8}$$

The truncated Bayes AMP iteration is

$$\begin{aligned}\boldsymbol{r}_M^t &= \frac{\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}^t}{n} + \mathsf{b}_{M,t}\boldsymbol{r}_M^{t-1}, \\ \widehat{\boldsymbol{\beta}}_M^{t+1} &= \eta_{M,t}(\widehat{\beta}_{M,t}^t + \boldsymbol{X}^\top \boldsymbol{r}_M^t).\end{aligned} \tag{L.9}$$

To prove Proposition 2.4, we will use Theorem 14 of [BMN19] to establish state evolution for the iteration (L.9) and then show that for large M, iteration (L.9) approximates iteration (2.24).

State evolution for truncated Bayes AMP.

We claim for any fixed t,

$$\lim_{p\to\infty}{}^{\mathrm{p}} \|\widehat{\boldsymbol{\beta}}_M^t - \boldsymbol{\beta}_0\|^2 = \mathbb{E}_{\beta_0,z}[(\eta_{M,t}(\beta_0+\tau_{M,t}z)^2 - \beta_0^2)]. \tag{L.10}$$

This statement follow directly from Theorem 14 of [BMN19] because, due to the truncation, all the technical conditions of that theorem are satisfied, as we now show.

First, Theorem 14 of [BMN19] is related to our setting by the following change of variables (with the notation of [BMN19] on the left).

$$\begin{aligned}&\boldsymbol{A} \leftarrow \frac{1}{\sqrt{n}}\boldsymbol{X}, \qquad \boldsymbol{\theta}_0 \leftarrow \sqrt{p}\boldsymbol{\beta}_0, \qquad \boldsymbol{y} \leftarrow \sqrt{\frac{p}{n}}\boldsymbol{y}, \qquad \boldsymbol{w} \leftarrow \sqrt{\frac{p}{n}}\boldsymbol{w}, \quad \widehat{\boldsymbol{\theta}}^t \leftarrow \sqrt{p}\widehat{\boldsymbol{\beta}}_M^t, \\ &\boldsymbol{r}^t \leftarrow \sqrt{np}\boldsymbol{r}_M^t, \quad (m,n) \leftarrow (n,p), \quad \eta_t(\boldsymbol{x})_j \leftarrow \eta_{M,t}(\sqrt{p}x_j), \quad \sigma_w^2 \leftarrow \sigma^2/\delta, \quad \mathsf{b}_t \leftarrow \mathsf{b}_{M,t}.\end{aligned} \tag{L.11}$$

We must check conditions (C1) - (C6) of [BMN19] and one more condition which we list as equation (L.13) below. (C1) holds by assumption; (C2) holds because the posterior mean is continuously differentiable to all orders,[6], so it is Lipschitz on compact intervals, and $\eta_{M,t}$ defined by $\eta_{M,t}(\boldsymbol{x})_j = \eta_{M,t}(\sqrt{p}x_j)/\sqrt{p}$ is uniformly Lipschitz; and (C3) and (C4) hold by the DSN assumption.

[6]See [LR05, Theorem 2.7.1]. Because the posterior mean as a function of y under Gaussian noise is the mean of an exponential family with natural parameter y/τ^2, this theorem applies.

For (C5), we must check that $\lim_{p\to\infty} \mathbb{E}_{\boldsymbol{z}}\left[\langle \boldsymbol{\beta}_0, \eta_{M,t}(\boldsymbol{\beta}_0 + \tau \boldsymbol{z})\rangle\right]$ exists and is finite. Note the functions $(\boldsymbol{x}_1, \boldsymbol{x}_2) \mapsto \boldsymbol{x}_1$ and $(\boldsymbol{x}_1, \boldsymbol{x}_2) \mapsto \eta_{M,t}(\boldsymbol{x}_1 + \tau \boldsymbol{x}_2)$ are uniformly pseudo-Lipschitz of order 1 (the first trivially, the second by (C2)), and their norm evaluated at $\mathbf{0}$ is bounded over p. Because these functions are symmetric, Lemma C.4, which gives

$$\begin{aligned}\lim_{p\to\infty} \mathbb{E}_{\boldsymbol{z}}[\langle \boldsymbol{\beta}_0, \eta_{M,t}(\boldsymbol{\beta}_0 + \tau \boldsymbol{z})\rangle] &= \lim_{p\to\infty} \mathbb{E}_{\tilde{\boldsymbol{\beta}}_0, \boldsymbol{z}}\left[\left\langle \tilde{\boldsymbol{\beta}}_0, \eta_{M,t}(\tilde{\boldsymbol{\beta}}_0 + \tau \boldsymbol{z})\right\rangle\right] \\ &= \lim_{p\to\infty} \frac{1}{p}\sum_{j=1}^{p} \mathbb{E}_{\tilde{\boldsymbol{\beta}}_0, \boldsymbol{z}}\left[\sqrt{p}\tilde{\beta}_{0j}\eta_{M,t}(\sqrt{p}\tilde{\beta}_{0j} + \tau\sqrt{p} z_j)\right] = \mathbb{E}_{\beta_0, z}[\beta_0 \eta_{M,t}(\beta_0 + \tau z)], \quad \text{(L.12)}\end{aligned}$$

where $\tilde{\boldsymbol{\beta}}_0$ has coordinates distributed iid from $\pi/\sqrt{p}$, and $\beta_0 \sim \pi$, $z \sim \mathsf{N}(0,1)$ independent. Because $\eta_{M,t}$ is bounded, the expectation on the right-hand side is finite, and (C5) is established.

For (C6), we must show that or any s, t and any 2×2 covariance matrix $\boldsymbol{T} \in S_+^2$, the limit $\lim_{p\to\infty} \mathbb{E}_{\boldsymbol{z}_1, \boldsymbol{z}_2}\left[\langle \eta_{M,s}(\boldsymbol{\beta}_0 + \boldsymbol{z}_1), \eta_{M,t}(\boldsymbol{\beta}_0 + \boldsymbol{z}_2)\rangle\right]$ exists and is finite, where $(\boldsymbol{z}_1, \boldsymbol{z}_2) \sim \mathsf{N}(\mathbf{0}, \boldsymbol{T} \otimes \boldsymbol{I}_p/p)$. This is shown in the same way we establisehd (C5).

Under the change of variables (with the notation of [BMN19] on the left) $\tau_t^2 \leftarrow \tau_{M,t}^2$, iteration (206), (207) of [BMN19] becomes the scalar iteration (L.7). Under this change of variables, the condition given by equation (208) of [BMN19] becomes

$$\mathsf{b}_{M,t} \overset{\text{p}}{\simeq} \frac{1}{n}\mathbb{E}_{\boldsymbol{z}}\left[\operatorname{div} \eta_{M,t-1}\left(\boldsymbol{\beta}_0 + \tau_{M,t-1}\boldsymbol{z}\right)\right], \quad \text{(L.13)}$$

where $\boldsymbol{z} \sim \mathsf{N}(\mathbf{0}, \boldsymbol{I}_p/p)$. Note $\tau_{M,t-1} > 0$ by induction: for the base case, use $s_2(\pi) > 0$, and then use that the right-hand side of (L.7) must be positive whenever $\tau_{M,t} > 0$ because perfect recovery with a non-trivial prior is impossible under Guassian corruption. Then by Gaussian integration by parts (Lemma P.6), we have

$$\frac{1}{n}\mathbb{E}_{\boldsymbol{z}}\left[\operatorname{div} \eta_{M,t-1}\left(\boldsymbol{\beta}_0 + \tau_{M,t-1}\boldsymbol{z}\right)\right] = \frac{p}{\tau_{M,t-1} n}\mathbb{E}_{\boldsymbol{z}}\left[\langle \boldsymbol{z}, \eta_{M,t-1}\left(\boldsymbol{\beta}_0 + \tau_{M,t-1}\boldsymbol{z}\right)\rangle\right]. \quad \text{(L.14)}$$

As the dimension p varies, the functions $(\boldsymbol{x}_0, \boldsymbol{x}_1) \mapsto \boldsymbol{x}_1$ and $(\boldsymbol{x}_0, \boldsymbol{x}_1) \mapsto \eta_{M,t-1}(\boldsymbol{x}_0 + \tau_{M,t-1}\boldsymbol{x}_1)$ are uniformly pseudo-Lipschitz of order 1 (the first trivially, the second by (C2)). By the same argument as in (C5), their norm when evaluated at $\mathbf{0}$ is bounded (over p). Thus, by Lemma P.2, the functions $(\boldsymbol{x}_0, \boldsymbol{x}_1) \mapsto \langle \boldsymbol{x}_1, \eta_{M,t-1}(\boldsymbol{x}_0 + \tau_{M,t-1}\boldsymbol{x}_1)\rangle$ are uniformly pseudo-Lipschitz of order 2. Because these functions are also symmetric and $\{\boldsymbol{\beta}_0\}$ satisfies the DSN assumption (2.1), we may apply Lemma C.4, which gives

$$\begin{aligned}&\mathbb{E}_{\boldsymbol{z}}\left[\langle \boldsymbol{z}, \eta_{M,t-1}\left(\boldsymbol{\beta}_0 + \tau_{M,t-1}\boldsymbol{z}\right)\rangle\right] \overset{\text{p}}{\simeq} \mathbb{E}_{\tilde{\boldsymbol{\beta}}_0, \boldsymbol{z}}\left[\left\langle \boldsymbol{z}, \eta_{M,t-1}(\tilde{\boldsymbol{\beta}}_0 + \tau_{M,t-1}\boldsymbol{z})\right\rangle\right] \\ &= \frac{1}{p}\sum_{j=1}^{p} \mathbb{E}_{\tilde{\boldsymbol{\beta}}_0, \boldsymbol{z}}\left[\left\langle \sqrt{p} z_{0j}\eta_{M,t-1}(\sqrt{p}\tilde{\beta}_{0j} + \sqrt{p} z_{0j})\right\rangle\right] = \frac{1}{p}\sum_{j=1}^{p} \mathbb{E}_{\beta_0, z}\left[\langle z\eta_{M,t-1}\left(\beta_0 + z_0\right)\rangle\right] \\ &= \tau_{M,t-1}\mathbb{E}_{\beta_0, z}\left[\eta'_{M,t-1}(\beta_0 + \tau_{M,t-1} z)\right], \quad \text{(L.15)}\end{aligned}$$

where we have taken $\tilde{\boldsymbol{\beta}}_0$ with coordinates distributed iid from $\pi/\sqrt{p}$, in the second equality we have used (2.23), in the third line we have taken $\beta_0 \sim \pi$, $z \sim \mathsf{N}(0,1)$ independent, and in the fourth equality we have used Lemma P.6 and the fact that $\eta_{M,t-1} : \mathbb{R} \mapsto \mathbb{R}$ is Lipschitz (see (C2)). By the HDA assumption, $n/p \to \delta$, whence (L.14) and (L.15) yield (L.13).

Estimation error of truncated Bayes AMP

Having checked the above conditions, we may apply Theorem 14 of [BMN19]. Because $\eta_{M,t} : \mathbb{R}^p \to \mathbb{R}^p$ are uniformly pseudo-Lipschitz of order 1 by (C2) and $\|\eta_{M,t}(\mathbf{0}) - \mathbf{0}\| = \|\eta_{M,t}(\mathbf{0})\|$ is uniformly (over p) bounded by the argument in (C5), we have by Lemma P.2 that the functions $(\boldsymbol{x}_0, \boldsymbol{x}_1) \mapsto \|\eta_{M,t}(\boldsymbol{x}_1) - \boldsymbol{x}_0\|^2$ are uniformly pseudo-Lipschitz of order 2. Thus, by Claim D.3, the functions $\Psi_p(\boldsymbol{x}_0, \boldsymbol{x}_1) := \left\|\eta_{M,t}\left(\boldsymbol{x}_1/\sqrt{p}\right) - \boldsymbol{x}_0/\sqrt{p}\right\|^2$ are [BMN19]-uniformly pseudo-Lipschitz of order 2.[7] Under the change of variables (L.11), we have $\widehat{\boldsymbol{\theta}}^t + \boldsymbol{A}^\mathsf{T}\boldsymbol{r}^t \leftarrow \sqrt{p}(\widehat{\boldsymbol{\beta}}_M^t + \boldsymbol{X}\boldsymbol{r}_M^t)$ and $\boldsymbol{\theta}_0 \leftarrow \sqrt{p}\boldsymbol{\beta}_0$. Then (justification follows equations)

$$\begin{aligned}\left\|\widehat{\boldsymbol{\beta}}_M^{t+1} - \boldsymbol{\beta}_0\right\|^2 &= \left\|\eta_{M,t}\left(\widehat{\boldsymbol{\beta}}_M^t + \boldsymbol{X}^\mathsf{T}\boldsymbol{r}_M^t\right) - \boldsymbol{\beta}_0\right\|^2 \overset{\mathrm{p}}{\simeq} \mathbb{E}_{\boldsymbol{z}}\left[\left\|\eta_{M,t}\left(\boldsymbol{\beta}_0 + \tau_{M,t}\boldsymbol{z}\right) - \boldsymbol{\beta}_0\right\|^2\right] \\ &\overset{\mathrm{p}}{\simeq} \mathbb{E}_{\tilde{\boldsymbol{\beta}}_0,\boldsymbol{z}}\left[\left\|\eta_{M,t}(\tilde{\boldsymbol{\beta}}_0 + \tau_{M,t}\boldsymbol{z}) - \boldsymbol{\beta}_0\right\|^2\right] = \frac{1}{p}\sum_{j=1}^p \mathbb{E}_{\tilde{\boldsymbol{\beta}}_0,\boldsymbol{z}}\left[\left(\eta_{M,t}(\sqrt{p}\tilde{\beta}_{0j} + \tau_{M,t}\sqrt{p}z_j) - \sqrt{p}\beta_{0j}\right)^2\right] \\ &= \mathbb{E}_{\beta_0,z}\left[\left(\eta_{M,t}\left(\beta_0 + \tau_{M,t}z\right) - \beta_0\right)^2\right], \end{aligned} \tag{L.16}$$

where in the first equality we have used (2.24); in the first line we have taken $\boldsymbol{z} \sim \mathsf{N}(\mathbf{0}, \boldsymbol{I}_p/p)$; in the first probabilistic equality we have used Theorem 14 of [BMN19] (in particular, Eq. (210) applied to ψ_p); in the second probabilistic equality we have used Lemma C.4; in the second line we have taken $\tilde{\boldsymbol{\beta}}_0$ with coordinates distributed iid from $\pi/\sqrt{p}$; in the third line we have used (2.23); and in the fourth line we have taken $\beta_0 \sim \pi$, $z \sim \mathsf{N}(0,1)$ independent.

The truncated and untruncated state evolutions are close

By induction, for each $t \geq 0$, we have

$$\lim_{M\to\infty} \tau_{M,t} = \tau_t \quad \text{and} \quad \lim_{M\to\infty} \mathsf{b}_{M,t} = \mathsf{b}_t. \tag{L.17}$$

Indeed, the base case ($t = 0$) holds by definition. For the induction step, assume $\tau_{M,t} \to \tau_t$. Denote $\eta(\cdot;\tau)$ the Bayes estimator at noise level τ. That is, $\eta(y;\tau) = \mathbb{E}_{\beta_0,z}[\beta_0|\beta_0 + \tau z = y]$. By the same argument we used in (C2), the Bayes estimator η is continuous in τ for $\tau > 0$. Then, by the inductive hypothesis, $\eta(y;\tau_{M,t}) \xrightarrow[M\to\infty]{} \eta(y;\tau_t)$ for all $y \in \mathbb{R}$. Further, because for β_0, z fixed we have $M > \beta_0 + \tau_t z$ for sufficiently large M, we have

$$\eta_{M,t}(\beta_0 + \tau_{M,t}z) \xrightarrow[M\to\infty]{} \eta(\beta_0 + \tau_t z;\tau_t) \text{ pointwise.} \tag{L.18}$$

Also, $|\eta_{M,t}(\beta_0+\tau_{M,t}z)| < |\eta(\beta_0+\tau_{M,t}z;\tau_{M,t})|$ and the collection of random variables $\{\eta(\beta_0+\tau z;\tau)^2 \mid \tau \geq 0\}$ is uniformly integrable because $\eta(\beta_0 + \tau z;\tau)^2 = \mathbb{E}_{\beta_0,z}[\beta_0|\beta_0 + \tau z]^2 \leq \mathbb{E}_{\beta_0,z}[\beta_0^2|\beta_0 + \tau z]$, and $\mathbb{E}_{\beta_0,z}[\mathbb{E}_{\beta_0,z}[\beta_0^2|\beta_0 + \tau z]\mathbf{1}_{\mathbb{E}_{\beta_0,z}[\beta_0^2|\beta_0+\tau z]>C}] = \mathbb{E}_{\beta_0,z}[\beta_0^2\mathbf{1}_{\mathbb{E}_{\beta_0,z}[\beta_0^2|\beta_0+\tau z]>C}]$ becomes uniformly small for sufficiently large C because $\mathbb{P}_{\beta_0,z}(\mathbb{E}_{\beta_0,z}[\beta_0^2|\beta_0 + \tau z] > C) \leq \frac{s_2(\pi)}{C}$ by Markov's inequality. Thus, in fact, the collection $\{(\eta_{M,t}(\beta_0 + \tau_{M,t}z) - \beta_0)^2\}$ over all values of M and t is uniformly integrable. By Vitali's Convergence Theorem (see e.g. [Dur10, Theorem 5.5.2]) and (L.18), we have

$$\mathbb{E}_{\beta_0,z}\left[\left(\eta_{M,t}\left(\beta_0 + \tau_{M,t}z\right) - \beta_0\right)^2\right] \xrightarrow[M\to\infty]{} \mathbb{E}_{\beta_0,z}\left[\left(\eta\left(\beta_0 + \tau_t z, \tau_t\right) - \beta_0\right)^2\right] = \mathsf{mmse}_\pi\left(\tau_t^2\right). \tag{L.19}$$

[7] See (D.30). This terminology just refers to the use of the notion of being uniformly pseudo-Lipschitz under the different choice of normalization used by [BMN19]. Thus, it tells us the functions to which we are able to apply their theorem.

By (2.20b), (L.7) and (L.19), we have (L.17). The induction is complete, so in fact $\tau_{M,t} \to \tau_t$ as $M \to \infty$ holds for all t. A similar argument shows that convergence of $\mathsf{b}_{M,t}$.

The truncated and untruncated state evolutions are close

We claim

$$\lim_{M\to\infty} \limsup_{p\to\infty}{}^{\mathrm{p}} \|\widehat{\boldsymbol{\beta}}_M^t - \widehat{\boldsymbol{\beta}}^t\|^2 = 0.$$

This follows inductively by combining $|\eta_t(y) - \eta_{M,t}(y)| \le C_t(1+|y|)\mathbf{1}_{|y|\ge M}$ (Lemma L.2), $\mathsf{b}_{M,t} \to \mathsf{b}_t$, and the boundedness (in probability) of $\|\widehat{\boldsymbol{\beta}}_M^t\|^2$. Thus, by (L.10) and (L.17), we conclude (2.25).

Bayes AMP achieves noise variance $\tau^2_{\mathsf{reg,amp}*}$

Now we prove (2.26). Because $s_2(\pi) > 0$, for all $\tau > 0$, we have $\mathsf{mmse}_\pi(\tau^2) < s_2(\pi)$. Thus, for $\tau^2 \ge \frac{1}{\delta}(\sigma^2 + s_2(\pi))$, we have $\delta\tau^2 - \sigma^2 \ge s_2(\pi) > \mathsf{mmse}_\pi(\tau^2)$. Thus, by (2.15) and (2.20a) and the continuity of $\mathsf{mmse}_\pi(\tau^2)$ in τ^2, we have $\tau_0 > \tau_{\mathsf{reg,amp}*}$. Further, if $\tau > \tau_{\mathsf{reg,amp}*}$, because $\mathsf{mmse}_\pi(\tau^2)$ is strictly increasing in τ (see [DYSV11, Eq. (65)]), we have $\frac{1}{\delta}\left(\sigma^2 + \mathsf{mmse}_\pi(\tau^2)\right) > \frac{1}{\delta}\left(\sigma^2 + \mathsf{mmse}_\pi(\tau^2_{\mathsf{reg,amp}*})\right) = \tau^2_{\mathsf{reg,amp}*}$. Thus, by (2.20b) and induction, we have $\tau_{t-1} > \tau_t > \tau_{\mathsf{reg,amp}*}$ for all t. Further, because $\mathsf{mmse}_\pi(\tau^2)$ is continuous in τ^2 [DYSV11], for all $\varepsilon > 0$ such that $\tau_{\mathsf{reg,amp}*} + \varepsilon < \tau_0$, we have $\inf_{\tau\in[\tau_{\mathsf{reg,amp}*}+\varepsilon,\tau_0]}\left\{\tau^2 - \frac{1}{\delta}\left(\sigma^2 + \mathsf{mmse}_\pi(\tau^2)\right)\right\} > 0$. Thus, for all t such that $\tau_t > \tau_{\mathsf{reg,amp}*} + \varepsilon$, we have $\tau_t - \tau_{t+1}$ is bounded below by a positive constant. Thus, for t sufficiently large we must have $\tau_t \le \tau_{\mathsf{reg,amp}*} + \varepsilon$. Because this is true for all sufficiently small $\varepsilon > 0$, we have $\limsup_{t\to\infty} \tau_t \le \tau_{\mathsf{reg,amp}*}$. Because we also have $\tau_t > \tau_{\mathsf{reg,amp}*}$ for all t, we in fact have (2.26).

Eq. (2.27) now follows by (2.25) and taking t sufficiently large. □

M Proof of Theorem 4

Proof of Theorem 4.(i). Throughout the proof, we will drop π from our notation for the moments of π. That is, we write s_k in place of $s_k(\pi)$. Observe that $\mathsf{mmse}_\pi(\tau^2) \le s_2$. Also, $\mathsf{R}^{\mathsf{opt}}_{\mathsf{seq,cvx}}(\tau;\pi) \le s_2$ because at each p we may take in (2.7) the function $\rho_p(\boldsymbol{x}) = \mathbb{I}_{\boldsymbol{x}=\mathbf{0}}$ which is 0 when $\boldsymbol{x} = \mathbf{0}$ and ∞ otherwise. Thus,

$$\frac{\sigma^2}{\delta} \le \frac{\sigma^2 + \mathsf{mmse}_\pi(\tau^2)}{\delta} \le \frac{\sigma^2 + s_2}{\delta}, \tag{M.1}$$

$$\frac{\sigma^2}{\delta} \le \frac{\sigma^2 + \mathsf{R}^{\mathsf{opt}}_{\mathsf{seq,cvx}}(\tau;\pi)}{\delta} \le \frac{\sigma^2 + s_2}{\delta}. \tag{M.2}$$

By (M.1) and (2.32), $\frac{\sigma^2}{\delta} \le \tau^2_{\mathsf{reg,stat}} \le \frac{\sigma^2+s_2}{\delta}$, whence $\tau_{\mathsf{reg,stat}} = \frac{\sigma}{\sqrt{\delta}} + O\left(\frac{1}{\sqrt{\delta}}\right)$, where we have used the inequality that for $a, b \ge 0$ we have $\sqrt{a+b} \le \sqrt{a} + \sqrt{b}$. By Lemma C.2, $\mathsf{R}^{\mathsf{opt}}_{\mathsf{seq,cvx}}(\tau;\pi)$ is continuous in τ, whence by (2.11), we have $\delta\tau^2_{\mathsf{reg,cvx}} - \sigma^2 = \mathsf{R}^{\mathsf{opt}}_{\mathsf{seq,cvx}}(\tau_{\mathsf{reg,cvx}};\pi)$. Combined with (M.2), we have $\frac{\sigma^2}{\delta} \le \tau^2_{\mathsf{reg,cvx}} \le \frac{\sigma^2+s_2}{\delta}$, whence $\tau_{\mathsf{reg,cvx}} = \frac{\sigma}{\sqrt{\delta}} + O\left(\frac{1}{\sqrt{\delta}}\right)$ as well.

With $\beta_0 \sim \pi$, $z \sim \mathsf{N}(0,1)$ independent and $y = \beta_0 + \tau z$, we have (justification to follow)

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}\tau}\mathsf{mmse}_\pi(\tau^2) &= -\frac{2}{\tau^3}\frac{\mathrm{d}}{\mathrm{d}\tau^{-2}}\mathsf{mmse}_\pi(\tau^2) = \frac{2}{\tau^3}\mathbb{E}_{\beta_0,z}\left[\mathbb{E}_{\beta_0,z}\left[(\beta_0 - \mathbb{E}_{\beta_0,z}[\beta_0|y])^2\middle|y\right]^2\right] \\ &\le \frac{2}{\tau^3}\mathbb{E}_{\beta_0,z}\left[(\beta_0 - \mathbb{E}_{\beta_0,z}[\beta_0|y])^4\right] \le 32\sqrt{24}\tau,\end{aligned} \tag{M.3}$$

where in the second equality we have used [DYSV11, Proposition 9], in the first inequality we have used Jensen's inequality, and in the second inequality we have used [DYSV11, Proposition 5]. Further, because $\pi \in \mathcal{P}_6(\mathbb{R})$, $\tau^{-2} \mapsto \mathsf{mmse}_\pi(\tau^2)$ is continuously differentiable to second order on $[0,\infty)$ [DYSV11, Proposition 7], whence $\frac{\mathrm{d}}{\mathrm{d}\tau^{-2}}\mathsf{mmse}_\pi(\tau^2)$ is bounded for $\tau \geq C$ for any $C > 0$. Combined with (M.3) (which bounds the derivative for small τ), we get that $\mathsf{mmse}_\pi(\tau^2)$ is Lipschitz in τ on the entirety of its domain $[0,\infty)$. Because $\mathsf{R}^{\mathsf{opt}}_{\mathsf{seq,cvx}}(\tau;\pi)$ is also Lipschitz in τ, we have by Theorem 1

$$\Delta(\pi,\delta,\sigma) \geq \mathsf{R}^{\mathsf{opt}}_{\mathsf{seq,cvx}}(\tau_{\mathsf{reg,cvx}};\pi) - \mathsf{mmse}_\pi(\tau^2_{\mathsf{reg,stat}}) = \mathsf{R}^{\mathsf{opt}}_{\mathsf{seq,cvx}}(\sigma/\sqrt{\delta};\pi) - \mathsf{mmse}_\pi(\sigma^2/\delta) + O(1/\sqrt{\delta}).$$

Thus, we have (4.1). □

Proof of Theorem 4.(ii). As in the proof of part (i), throughout the proof, we will drop π from our notation for the moments of π. That is, we write s_k in place of $s_k(\pi)$. By [DYSV11, Eq. (61)], we have

$$\mathsf{mmse}_\pi(\tau^2) = s_2 - s_2^2\frac{1}{\tau^2} + \frac{1}{2}\left(2s_2^3 - s_3^2\right)\frac{1}{\tau^4} - \frac{1}{6}\left(15s_2^4 - 12s_2s_3^2 - 6s_2^2s_4 + s_4^2\right)\frac{1}{\tau^6} + O\left(\frac{1}{\tau^8}\right), \tag{M.4}$$

where $O\left(\frac{1}{\tau^8}\right)$ hides constants depending only on the moments of π. Define $\kappa^2 = \frac{\sigma^2}{\delta}\left(1 + \frac{s_2}{\sigma^2}\right)$ and $\Delta = \kappa^2 - \tau^2_{\mathsf{reg,stat}}$. By (2.32) and some rearrangement, we have

$$s_2 - \delta\Delta = \mathsf{mmse}_\pi\left(\kappa^2 - \Delta\right). \tag{M.5}$$

For the remainder of the proof, O will also hide constants depending δ (in addition to the moments of π), but will not depend on σ^2 and likewise on κ^2 or Δ. We see that $\Delta \leq \frac{s_2}{\delta} = O(1)$, so that by (M.4) we have

$$\begin{aligned}\mathsf{mmse}_\pi\left(\kappa^2 - \Delta\right) = s_2 &- \frac{s_2^2}{\kappa^2}\left(1 + \frac{\Delta}{\kappa^2} + \frac{\Delta^2}{\kappa^4} + O\left(\frac{1}{\kappa^6}\right)\right) + \frac{1}{2}\frac{2s_2^3 - s_3^2}{\kappa^4}\left(1 + O\left(\frac{\Delta}{\kappa^2}\right)\right) \\ &- \frac{1}{6}\frac{15s_2^4 - 12s_2s_3^2 - 6s_2^2s_4 + s_4^2}{\kappa^6}\left(1 + O\left(\frac{\Delta}{\kappa^2}\right)\right) + O\left(\frac{1}{\kappa^8}\right).\end{aligned} \tag{M.6}$$

Comparing with (M.5) and using $\Delta = O(1)$, we see that $\Delta = \frac{s_2^2}{\delta\kappa^2} + O\left(\frac{1}{\kappa^4}\right)$. Thus, we have

$$\begin{aligned}\frac{s_2^2}{\kappa^2}\left(1 + \frac{\Delta}{\kappa^2} + \frac{\Delta^2}{\kappa^4} + O\left(\frac{1}{\kappa^6}\right)\right) &= \frac{s_2^2}{\kappa^2} + \frac{s_2^4}{\delta\kappa^6} + O\left(\frac{1}{\kappa^8}\right), \\ \frac{1}{2}\frac{2s_2^3 - s_3^2}{\kappa^4}\left(1 + O\left(\frac{\Delta}{\kappa^2}\right)\right) &= \frac{1}{2}\frac{2s_2^3 - s_3^2}{\kappa^4} + O\left(\frac{1}{\kappa^8}\right).\end{aligned}$$

Plugging into (M.6), we have

$$\begin{aligned}\mathsf{mmse}_\pi(\tau^2_{\mathsf{reg,stat}}) = \mathsf{mmse}_\pi(\kappa^2 - \Delta) = s_2 &- \frac{s_2^2}{\kappa^2} + \frac{1}{2}\frac{2s_2^3 - s_3^2}{\kappa^4} \\ &- \frac{s_2^4}{\delta\kappa^6} - \frac{1}{6}\frac{15s_2^4 - 12s_2s_3^2 - 6s_2^2s_4 + s_4^2}{\kappa^6} + O\left(\frac{1}{\kappa^8}\right).\end{aligned} \tag{M.7}$$

We now write this expansion in terms of the signal-to-noise parameter snr. Applying the definition of κ^2, we have $\frac{s_2}{\kappa^2} = \delta\,\mathsf{snr}(1-\mathsf{snr}+\mathsf{snr}^2)+O(\mathsf{snr}^4)$, $\frac{s_2^2}{\kappa^4} = \delta^2\mathsf{snr}^2(1-2\,\mathsf{snr})+O(\mathsf{snr}^4)$, and $\frac{s_2^3}{\kappa^6} = \delta^3\,\mathsf{snr}^3+O(\mathsf{snr}^4)$. Plugging into (M.7) and rearranging, we get

$$\begin{aligned}\mathsf{mmse}_\pi(\tau^2_{\mathsf{reg,stat}}) = s_2 - s_2\delta\,\mathsf{snr} + s_2\left(\delta+\delta^2\left(1-\frac{s_3^2}{2s_2^3}\right)\right)\mathsf{snr}^2 \\ + s_2\left(-\delta-\delta^2\left(3-\frac{s_3^2}{s_2^3}\right)-\delta^3\left(\frac{5}{2}-2\frac{s_3^2}{s_2^3}-\frac{s_4}{s_2^2}+\frac{s_4^2}{6s_2^4}\right)\right)\mathsf{snr}^3+O(\mathsf{snr}^4). \quad \text{(M.8)}\end{aligned}$$

Now let τ_{ridge} solve

$$\delta\tau^2-\sigma^2 = \mathsf{mmse}_{\mathsf{N}(0,s_2)}(\tau^2). \quad \text{(M.9)}$$

We will show that ridge regression with appropriately chosen regularization achieves risk $\mathsf{mmse}_{\mathsf{N}(0,s_2)}(\tau^2_{\mathsf{ridge}})$. Let $\rho_p(\boldsymbol{x}) = \frac{\sigma^2}{\delta s_2}\|\boldsymbol{x}\|^2$. The risk of this estimator has been studied previously by [EK13]. We repeat the analysis here for completeness. Let

$$\lambda_{\mathsf{ridge}} = \frac{\delta\tau^2_{\mathsf{ridge}}}{2\sigma^2}. \quad \text{(M.10)}$$

Observe then that $\mathsf{prox}[\lambda_{\mathsf{ridge}}\rho_p](\boldsymbol{y}) = \frac{s_2}{s_2+\tau^2_{\mathsf{ridge}}}\boldsymbol{y}$. Let $\mathcal{T} = (\pi,\{\rho_p\})$. Observe that $\mathsf{R}^\infty_{\mathsf{reg,cvx}}(\tau_{\mathsf{ridge}},\lambda_{\mathsf{ridge}},\mathcal{T}) = \frac{s_2\tau^2}{s_2+\tau^2} = \mathsf{mmse}_{\mathsf{N}(0,s_2)}(\tau^2_{\mathsf{ridge}})$ and $\mathsf{W}^\infty_{\mathsf{reg,cvx}}(\tau_{\mathsf{ridge}},\lambda_{\mathsf{ridge}},\mathcal{T}) = \frac{s_2}{s_2+\tau^2_{\mathsf{ridge}}}$. One can then check using (M.9) and (M.10) that $\tau_{\mathsf{ridge}},\lambda_{\mathsf{ridge}}$ solve (B.7) at $\gamma=0$. Because ρ_p are uniformly strongly convex, by Proposition B.3.(ii) and Lemma K.2, we have

$$\lim_{p\to\infty}\mathbb{E}_{\boldsymbol{\beta}_0,\boldsymbol{w},\boldsymbol{X}}\left[\|\widehat{\boldsymbol{\beta}}_{\mathsf{cvx}}-\boldsymbol{\beta}_0\|^2\right] = \mathsf{R}^\infty_{\mathsf{orc,cvx}}(\tau_{\mathsf{ridge}},\lambda_{\mathsf{ridge}},\mathcal{T}) = \mathsf{mmse}_{\mathsf{N}(0,s_2)}(\tau^2_{\mathsf{ridge}}). \quad \text{(M.11)}$$

Because τ_{ridge} solves (M.9), we in fact have that formula (M.8) holds for $\mathsf{mmse}_{\mathsf{N}(0,s_2)}(\tau^2_{\mathsf{ridge}})$ after replacing the moments with those of $\mathsf{N}(0,s_2)$. That is,

$$\mathsf{mmse}_{\mathsf{N}(0,s_2)}(\tau^2_{\mathsf{ridge}}) = s_2 - s_2\delta\,\mathsf{snr} + s_2\left(\delta+\delta^2\right)\mathsf{snr}^2 + s_2\left(-\delta-3\delta^2-\delta^3\right)\mathsf{snr}^3+O(\mathsf{snr}^4), \quad \text{(M.12)}$$

where we have used that the third moment of $\mathsf{N}(0,s_2)$ is 0 and fourth moment of $\mathsf{N}(0,s_2)$ is $3s_2^2$.

By [DYSV11, Eq. (65)], we have

$$\begin{aligned}-\frac{\mathrm{d}}{\mathrm{d}\tau^{-2}}\mathsf{mmse}_\pi(\tau^2) &= \mathbb{E}_{\beta_0,z}\left[(\mathbb{E}_{\beta_0,z}\left[(\mathbb{E}_{\beta_0,z}[\beta_0|y]-\beta_0)^2|y\right])^2\right] \le \mathbb{E}_{\beta_0,z}\left[\mathbb{E}_{\beta_0,z}\left[(\mathbb{E}_{\beta_0,z}[\beta_0|y]-\beta_0)^4|y\right]\right] \\ &\le 8\mathbb{E}_{\beta_0,z}\left[\mathbb{E}_{\beta_0,z}[\beta_0|y]^4+\beta_0^4\right] \le 16\mathbb{E}_{\beta_0,\boldsymbol{z}}[\beta_0^4] = 16s_4,\end{aligned}$$

where y denotes $\beta_0+\tau z$. Thus, $\frac{\mathrm{d}}{\mathrm{d}\tau^2}\mathsf{mmse}_\pi(\tau^2) = -\frac{1}{\tau^4}\frac{\mathrm{d}}{\mathrm{d}\tau^{-2}}\mathsf{mmse}_\pi(\tau^2) \le \frac{16s_4}{\tau^4}$. Thus, for sufficiently large τ, the derivative of the right-hand side of (2.31) is strictly negative, so that for sufficiently large σ there can be at most one solution to (2.32) in the region $[\sigma^2/\delta,\infty)$. But all solutions $\tau^2_{\mathsf{reg,stat}}$ must satisfy $\tau^2_{\mathsf{reg,stat}} \ge \sigma^2/\delta$, whence the minimizer of (2.30) is unique for sufficiently large σ. Then, by Proposition 2.6 and Eq. (M.11), for sufficiently large σ we have $\Delta(\pi,\delta,\sigma) \le \mathsf{mmse}_{\mathsf{N}(0,s_2)}(\tau^2_{\mathsf{ridge}}) - \mathsf{mmse}_\pi(\tau^2_{\mathsf{reg,stat}})$. Combining (M.8) and (M.12), we get (4.2), as desired. □

N Proofs for Section 6: examples

N.1 Proof of Proposition 6.2

This follows from inequality (O.12) proved in Appendix O, which gives

$$\frac{1}{\tau}\mathbb{E}_{\boldsymbol{\beta}_0,\boldsymbol{z}}\left[\langle \boldsymbol{z}, \mathsf{prox}\left[\lambda\rho_p\right](\boldsymbol{\beta}_0+\tau\boldsymbol{z})\rangle\right] \le \frac{1}{\lambda\gamma+1}, \tag{N.1}$$

because $\lambda\rho_p$ has strong convexity parameter $\lambda\gamma$. The right-hand side of (N.1) does not depend upon τ or p. Thus, choosing $1/(\bar{\lambda}\gamma+1) < \delta$ yields the proposition.

N.2 Proof of Proposition 6.3

claim Observe $\mathsf{prox}[\lambda\rho_p](\boldsymbol{x}) = \Pi_{C_p}(\boldsymbol{x})$ for all λ, where Π_{C_p} denotes projection onto the set C_p. Further, observe that $\mathbb{E}_{\boldsymbol{\beta}_0,\boldsymbol{z}}[\langle \boldsymbol{z}, \boldsymbol{\beta}_0\rangle] = 0$. Thus, we must show

$$\limsup_{p\to\infty}\sup_{\tau\in T}\frac{1}{\tau}\mathbb{E}_{\boldsymbol{\beta}_0,\boldsymbol{z}}\left[\left\langle \boldsymbol{z}, \Pi_{C_p}\left(\boldsymbol{\beta}_0+\tau\boldsymbol{z}\right)-\boldsymbol{\beta}_0\right\rangle\right] < \delta. \tag{N.2}$$

First, we argue conditionally on $\boldsymbol{\beta}_0$, which for now we treat as fixed. To simplify notation, we translate our problem—both $\boldsymbol{\beta}_0$ and C_p—by $-\boldsymbol{\beta}_0$, so that we may without loss of generality consider $\boldsymbol{\beta}_0 = \mathbf{0}$. In the translated problem, denote $\boldsymbol{b} = \Pi_{C_p}(\mathbf{0})$. Then

$$\left\langle \tau\boldsymbol{z}, \Pi_{C_p}\left(\tau\boldsymbol{z}\right)\right\rangle = \underbrace{\left\langle \tau\boldsymbol{z}-\boldsymbol{b}, \Pi_{C_p}\left(\tau\boldsymbol{z}\right)-\boldsymbol{b}\right\rangle}_{(*)} + \underbrace{\left\langle \boldsymbol{b}, \Pi_{C_p}\left(\tau\boldsymbol{z}\right)-\boldsymbol{b}\right\rangle + \left\langle \tau\boldsymbol{z}, \boldsymbol{b}\right\rangle}_{(**)}. \tag{N.3}$$

For $t\in[0,1]$, we have $t\boldsymbol{b}+(1-t)\Pi_{C_p}\left(\tau\boldsymbol{z}\right)\in C_p$. Thus,

$$\frac{\mathrm{d}}{\mathrm{d}t}\left\|\tau\boldsymbol{z}-\left(t\boldsymbol{b}+(1-t)\Pi_{C_p}\left(\tau\boldsymbol{z}\right)\right)\right\|^2\Bigg|_{t=0} \ge 0.$$

Some rearrangement gives

$$(*) \ge \|\Pi_{C_p}(\tau\boldsymbol{z})-\boldsymbol{b}\|^2 \ge \|\Pi_{C_p}(\tau\boldsymbol{z})\|^2 - 2\|\boldsymbol{b}\|\|\tau\boldsymbol{z}\|. \tag{N.4}$$

Cauchy-Schwartz gives

$$(**) \ge -\|\boldsymbol{b}\|\Big(\|\Pi_{C_p}(\tau\boldsymbol{z})-\boldsymbol{b}\| + \|\tau\boldsymbol{z}\|\Big) \ge -2\|\boldsymbol{b}\|\|\tau\boldsymbol{z}\|, \tag{N.5}$$

where in the second inequality we have used that projections onto convex sets are 1-Lipschitz. Also, if $\|\Pi_{C_p}(\tau\boldsymbol{z})\| > \varepsilon$, then $\Pi_{C_p}(\tau\boldsymbol{z}) \in T_{C_p\cap B^c(\mathbf{0},\varepsilon)}(\mathbf{0})$. Thus,

$$\|\tau\boldsymbol{z}-\Pi_{C_p}(\tau\boldsymbol{z})\| \ge \|\tau\boldsymbol{z}-\Pi_{T_{C_p\cap B^c(\mathbf{0},\varepsilon)}}(\tau\boldsymbol{z})\| \quad \text{if} \quad \|\Pi_{C_p}(\tau\boldsymbol{z})\| > \varepsilon. \tag{N.6}$$

Thus, if $\|\Pi_{C_p}(\tau\boldsymbol{z})\| > \varepsilon$,

$$\begin{aligned}
\|\Pi_{T_{C_p\cap B^c(\mathbf{0},\varepsilon)}}(\tau\boldsymbol{z})\|^2 &= \|\tau\boldsymbol{z}\|^2 - \|\tau\boldsymbol{z}-\Pi_{T_{C_p\cap B^c(\mathbf{0},\varepsilon)}}(\tilde{\boldsymbol{z}})\|^2 \\
&\ge \|\tau\boldsymbol{z}\|^2 - \|\tau\boldsymbol{z}-\Pi_{C_p}(\tau\boldsymbol{z})\|^2 \\
&= 2\langle\tau\boldsymbol{z}, \Pi_{C_p}(\tau\boldsymbol{z})\rangle - \|\Pi_{C_p}(\tau\boldsymbol{z})\|^2 \\
&\ge \langle\tau\boldsymbol{z}, \Pi_{C_p}(\tau\boldsymbol{z})\rangle - 4\|\boldsymbol{b}\|\|\tau\boldsymbol{z}\|,
\end{aligned}$$

where in the second line, we have used (N.6), and in the last line, we have used (N.3), (N.4), and (N.5). We conclude that

$$\begin{aligned}\langle \tau \boldsymbol{z}, \Pi_{C_p}(\tau \boldsymbol{z})\rangle &\leq \left(\|\Pi_{T_{C_p\cap B^c(\mathbf{0},\varepsilon)}}(\tau\boldsymbol{z})\|^2 + 4\|\boldsymbol{b}\|\|\tau\boldsymbol{z}\|\right)\mathbf{1}_{\|\Pi_{C_p}(\tau\boldsymbol{z})\|>\varepsilon} + \|\tau\boldsymbol{z}\|\varepsilon\mathbf{1}_{\|\Pi_{C_p}(\tau\boldsymbol{z})\|\leq\varepsilon}\\ &\leq \|\Pi_{T_{C_p\cap B^c(\mathbf{0},\varepsilon)}}(\tau\boldsymbol{z})\|^2 + (4\|\boldsymbol{b}\|+\varepsilon)\|\tau\boldsymbol{z}\|.\end{aligned}$$

Substituting the value of $\boldsymbol{b}$, undoing the translation, and averaging over $\boldsymbol{\beta}_0$ and $\boldsymbol{z}$, we get

$$\mathbb{E}_{\boldsymbol{\beta}_0,\boldsymbol{z}}\left[\left\langle \tau\boldsymbol{z}, \Pi_{C_p}\left(\boldsymbol{\beta}_0+\tau\boldsymbol{z}\right)-\boldsymbol{\beta}_0\right\rangle\right] \leq \mathbb{E}_{\boldsymbol{\beta}_0,\boldsymbol{z}}\left[\left\|\Pi_{T_{C_p\cap B^c(\boldsymbol{\beta}_0,\varepsilon)}}\left(\tau\boldsymbol{z}\right)\right\|^2\right] + \mathbb{E}_{\boldsymbol{\beta}_0,\boldsymbol{z}}\left[\left(4d\left(\boldsymbol{\beta}_0,C_p\right)+\varepsilon\right)\|\tau\boldsymbol{z}\|\right] \tag{N.7}$$

$$= \tau^2\mathbb{E}_{\boldsymbol{\beta}_0}\left[w\left(T_{C_p\cap B^c(\boldsymbol{\beta}_0,\varepsilon)}\right)\right] + \tau\mathbb{E}_{\boldsymbol{\beta}_0,\boldsymbol{z}}\left[\left(4d\left(\boldsymbol{\beta}_0,C_p\right)+\varepsilon\right)\|\boldsymbol{z}\|\right]. \tag{N.8}$$

By independence of $\boldsymbol{\beta}_0$ and $\boldsymbol{z}$, and (6.5), we get

$$\limsup_{p\to\infty}\mathbb{E}_{\boldsymbol{\beta}_0,\boldsymbol{z}}\left[\left(4d\left(\boldsymbol{\beta}_0,C_p\right)+\varepsilon\right)\|\tau\boldsymbol{z}\|\right] \leq \varepsilon\tau\mathbb{E}_{\boldsymbol{\beta}_0,\boldsymbol{z}}\left[\|\boldsymbol{z}\|\right] \leq \varepsilon\tau. \tag{N.9}$$

Fix compact $[\tau_{\mathsf{min}},\tau_{\mathsf{max}}]\subset(0,\infty)$.

$$\limsup_{p\to\infty}\sup_{\tau\in T}\frac{1}{\tau}\mathbb{E}_{\boldsymbol{\beta}_0,\boldsymbol{z}}\left[\left\langle \boldsymbol{z}, \Pi_{C_p}\left(\boldsymbol{\beta}_0+\tau\boldsymbol{z}\right)-\boldsymbol{\beta}_0\right\rangle\right] \leq \limsup_{p\to\infty}\mathbb{E}_{\boldsymbol{\beta}_0}\left[w\left(T_{C_p\cap B^c(\boldsymbol{\beta}_0,\varepsilon)}\right)\right] + \frac{\varepsilon}{\tau_{\mathsf{min}}} = \overline{\delta}(\varepsilon) + \frac{\varepsilon}{\tau_{\mathsf{min}}}, \tag{N.10}$$

where we defined $\overline{\delta}(\varepsilon) := \limsup_{p\to\infty}\mathbb{E}_{\boldsymbol{\beta}_0}\left[w\left(T_{C_p\cap B^c(\boldsymbol{\beta}_0,\varepsilon)}\right)\right]$. The claim (N.2) follows from taking the limit $\varepsilon\to 0$ and using Eq. (6.6).

N.3 Proof of Proposition 6.4

Applying the change of scaling identity for proximal operators (see Appendix O, Eq. (O.13)), we get

$$\mathsf{prox}[\lambda\rho_p](\boldsymbol{\beta}_0+\tau\boldsymbol{z})_j = \mathsf{prox}[\lambda\rho]\left(\sqrt{p}(\beta_{0j}+\tau z_j)\right)/\sqrt{p}, \tag{N.11}$$

so that

$$\frac{1}{\tau}\mathbb{E}_{\boldsymbol{\beta}_0,\boldsymbol{z}}[\langle \boldsymbol{z}, \mathsf{prox}[\lambda\rho_p](\boldsymbol{\beta}_0+\tau\boldsymbol{z})\rangle] = \frac{1}{\tau}\mathbb{E}_{\beta_0,z}[z\mathsf{prox}[\lambda\rho](\beta_0+\tau z)].$$

Having removed the dependence on p, the left-hand side of (2.9) becomes

$$\sup_{\lambda>\bar{\lambda},\tau\in T}\frac{1}{\tau}\mathbb{E}_{\beta_0,z}[z\mathsf{prox}[\lambda\rho](\beta_0+\tau z)].$$

It is easy to check using (2.4) that for any fixed β_0, z, τ, we have $\lim_{\lambda\to\infty}\mathsf{prox}[\lambda\rho](\beta_0+\tau z) = \Pi_C(\beta_0+\tau z)$. Further, if $m\in C$, we have by the 1-Lipschitz property of the proximal operator (O.4) that $|\mathsf{prox}[\lambda\rho](\beta_0+\tau z)| < |m| + |\beta_0| + \tau|z|$. By dominated convergence, we have

$$\lim_{\lambda\to\infty}\frac{1}{\tau}\mathbb{E}_{\beta_0,z}[z\mathsf{prox}[\lambda\rho](\beta_0+\tau z)] = \frac{1}{\tau}\mathbb{E}_{\beta_0,z}[z\Pi_C(\beta_0+\tau z)]. \tag{N.12}$$

By Gaussian integration by parts (Appendix O, Eq. (O.11)),

$$\frac{1}{\tau}\mathbb{E}_{\beta_0,z}[z\Pi_C(\beta_0+\tau z)] = \mathbb{E}_{\beta_0,z}[\mathbf{1}_{\beta_0+\tau z\in C}] = \mathbb{P}_{\beta_0,z}(\beta_0+\tau z\in C). \tag{N.13}$$

First assume the δ-bounded width assumption is satisfied. Observe $\sup_{\lambda\geq\bar\lambda}\frac{1}{\tau}\mathbb{E}_{\beta_0,z}[z\mathsf{prox}[\lambda\rho](\beta_0+\tau z)] \geq \mathbb{P}_{\beta_0,z}(\beta_0+\tau z\in C)$, whence for any compact $T\subset(0,\infty)$, we have $\delta > \sup_{\tau\in T}\mathbb{P}_{\beta_0,z}(\beta_0+\tau z\in C)$. Because $\tau\mapsto\mathbb{P}_{\beta_0,z}(\beta_0+\tau z\in C)$ is continuous and converges to 0 as $\tau\to\infty$, we have $\sup_{\tau>\varepsilon}\mathbb{P}_{\beta_0,z}(\beta_0+\tau z\in C) = \sup_{\tau\in[\varepsilon,M]}\mathbb{P}_{\beta_0,z}(\beta_0+\tau z\in C)$ for some finite M. We conclude $\sup_{\tau>\varepsilon}\mathbb{P}_{\beta_0,z}(\beta_0+\tau z\in C)<\delta$.

Conversely, assume $\sup_{\tau>\varepsilon}\mathbb{P}_{\beta_0,z}(\beta_0+\tau z\in C)<\delta$. Because $\mathsf{prox}[\lambda\rho]$ is 1-Lipschitz, we have that $\tau\to\frac{1}{\tau}\mathbb{E}_{\beta_0,z}[z\mathsf{prox}[\lambda\rho](\beta_0+\tau z)]$ is L-Lipschitz on compact $T=[\tau_{\mathsf{min}},\tau_{\mathsf{max}}]\subset(0,\infty)$ for some sufficiently large L depending on T (a complete argument of this fact occurs in a more general setting in the proof of Lemma C.5 in Appendix C). Thus, in order to pick $\bar\lambda$ such that $\sup_{\lambda\geq\bar\lambda,\tau\in T}\frac{1}{\tau}\mathbb{E}_{\beta_0,z}[z\mathsf{prox}[\lambda\rho](\beta_0+\tau z)]<\delta$, we can choose a $\frac{\delta-\sup_\tau\mathbb{P}(\beta_0+\tau z\in C)}{4L}$-cover of T and a $\bar\lambda$ sufficiently large such that $\frac{1}{\tau}\mathbb{E}_{\beta_0,z}[z\mathsf{prox}[\lambda\rho](\beta_0+\tau z)] < \frac{\delta+\sup_\tau\mathbb{P}(\beta_0+\tau z\in C)}{2}$ for all $\lambda>\bar\lambda$ and all τ in the cover.

N.4 Proof of Proposition 6.5

Fix any compact interval $T=[\tau_{\mathsf{min}},\tau_{\mathsf{max}}]\subset(0,\infty)$. First, we describe how to choose a $\bar\lambda$ for which (2.9) holds, and then we will prove that our choice works. Pick ε with

$$\frac{\delta^2\tau_{\mathsf{min}}^2}{4(s_2(\pi)+\tau_{\mathsf{max}}^2)} > \varepsilon > 0, \tag{N.14}$$

such that, for any $A\subseteq\mathbb{R}$,

$$\mathbb{P}(A)\leq\varepsilon \quad\Longrightarrow\quad \mathbb{E}_{\beta_0}[\beta_0^2\mathbf{1}_A]<\delta^2\tau_{\mathsf{min}}^2/32 \quad\text{and}\quad \mathbb{E}_z[\tau_{\mathsf{max}}^2z^2\mathbf{1}_A]<\delta^2\tau_{\mathsf{min}}^2/32\,. \tag{N.15}$$

Pick t such that

$$\mathbb{P}_{\beta_0}(|\beta_0|>t)<\varepsilon \quad\text{and}\quad \mathbb{P}_z(|\tau z|>t)<\varepsilon. \tag{N.16}$$

Finally pick $\xi>0$ such that $j\leq(1-\varepsilon)p$ implies $\kappa_j^{(p)}>\xi$. We claim that (2.9) is satisfied for $\bar\lambda=2t/\xi$.

First we recall that the proximal operator for the OWL penalty satisfies (e.g. see Lemma 3.1 of [SC16])

$$\|\mathsf{prox}[\lambda\rho_p](\boldsymbol{x})\|\leq\|(|\boldsymbol{x}|-\lambda\boldsymbol{\kappa}^{(p)}/\sqrt{p})_+\|, \tag{N.17}$$

where $|\boldsymbol{x}|$ denotes the coordinate-wise absolute value and $(\cdot)_+$ denotes the coordinate-wise positive part. By Cauchy-Schwartz, for any τ,λ,

$$\begin{aligned}\frac{1}{\tau}\mathbb{E}_{\boldsymbol{\beta}_0,\boldsymbol{z}}[\langle\boldsymbol{z},\mathsf{prox}[\lambda\rho_p](\boldsymbol{\beta}_0+\tau\boldsymbol{z})\rangle] &\leq \frac{1}{\tau}\sqrt{\mathbb{E}_{\boldsymbol{z}}[\|\boldsymbol{z}\|^2]\mathbb{E}_{\boldsymbol{\beta}_0,\boldsymbol{z}}[\|\mathsf{prox}[\rho_p](\boldsymbol{\beta}_0+\tau\boldsymbol{z})\|^2]}\\ &\leq \frac{1}{\tau}\sqrt{\mathbb{E}_{\boldsymbol{\beta}_0,\boldsymbol{z}}[\|(|\boldsymbol{\beta}_0+\tau\boldsymbol{z}|-\lambda\boldsymbol{\kappa}^{(p)}/\sqrt{p})_+\|^2]},\end{aligned} \tag{N.18}$$

where the last inequality holds by (N.17). For $\lambda > \bar{\lambda}$,

$$\begin{aligned}
\left\| \left(|\boldsymbol{\beta}_0 + \tau \boldsymbol{z}| - \frac{\lambda \boldsymbol{\kappa}^{(p)}}{\sqrt{p}} \right)_+ \right\|^2 &= \sum_{j=1}^{\lfloor (1-\varepsilon)p \rfloor} \left(|\beta_{0j} + \tau z_j| - \frac{\lambda \kappa_j^{(p)}}{\sqrt{p}} \right)_+^2 + \sum_{j=\lfloor (1-\varepsilon)p \rfloor + 1}^{p} \left(|\beta_{0j} + \tau z_j| - \frac{\lambda \kappa_j^{(p)}}{\sqrt{p}} \right)_+^2 \\
&\le \sum_{j=1}^{\lfloor (1-\varepsilon)p \rfloor} (\beta_{0j} + \tau z_j)^2 \, \mathbf{1}_{|\beta_{0j} + \tau z_j| > \frac{2t}{\sqrt{p}}} + \sum_{j=\lfloor (1-\varepsilon)p \rfloor + 1}^{p} (\beta_{0j} + \tau z_j)^2 \\
&\le 2 \sum_{j=1}^{\lfloor (1-\varepsilon)p \rfloor} \left(\beta_{0j}^2 + \tau^2 z_j^2 \right) \left(\mathbf{1}_{|\beta_{0j}| > \frac{t}{\sqrt{p}}} + \mathbf{1}_{|\tau z_j| > \frac{t}{\sqrt{p}}} \right) + \sum_{j=\lfloor (1-\varepsilon)p \rfloor + 1}^{p} (\beta_{0j} + \tau z_j)^2 ,
\end{aligned}$$

where in the first inequality, we have used that $\lambda \kappa_j^{(p)} / \sqrt{p} \ge 2t/\sqrt{p}$ because $\lambda > 2t/\xi$ and $\kappa_j^{(p)} > \xi$ for $j \le (1-\varepsilon)p$ and that for any $x, y \in \mathbb{R}$ we have $(|x| - y)_+^2 \le x^2 \mathbf{1}_{|x| > y}$; and in the second inequality, we have used that $(\beta_{0j} + \tau z_j)^2 \le 2\beta_{0j}^2 + 2\tau^2 z_j^2$ and $\mathbf{1}_{|\beta_{0j} + \tau z_j| > \frac{2t}{\sqrt{p}}} \le \mathbf{1}_{|\beta_{0j}| > \frac{t}{\sqrt{p}}} + \mathbf{1}_{|\tau z_j| > \frac{t}{\sqrt{p}}}$. Taking expectations, inequality (N.18) becomes

$$\begin{aligned}
\frac{1}{\tau} \mathbb{E}_{\boldsymbol{\beta}_0, \boldsymbol{z}} [\langle \boldsymbol{z}, \mathsf{prox}[\lambda \rho_p](\boldsymbol{\beta}_0 + \tau \boldsymbol{z}) \rangle] &\le \frac{1}{\tau} \sqrt{2 \mathbb{E}_{\beta_0, z} \left[\left(\beta_0^2 + \tau^2 z^2 \right) \left(\mathbf{1}_{|\beta_0| > t} + \mathbf{1}_{|\tau z| > t} \right) \right] + \varepsilon (\mathbb{E}_{\beta_0}[\beta_0^2] + \tau^2)} \\
&\le \frac{1}{\tau} \sqrt{\delta^2 \tau_{\mathsf{min}}^2 / 4 + \delta^2 \tau_{\mathsf{min}}^2 / 4} \\
&\le \delta / \sqrt{2} < \delta,
\end{aligned}$$

where in the second inequality we have bounded the first term under the square-root by (N.15) and the second term under the square-root by (N.14), and in the third inequality, we have used that $\tau_{\mathsf{min}} \le \tau$. This completes the proof.

O Proximal operator identities

We collect here various identities and properties of proximal operators, defined in (2.4). Many arguments are included because they are not well-known, others for the reader's convenience. Throughout this section, $\rho : \mathbb{R}^p \to \mathbb{R} \cup \{\infty\}$ is an lsc, proper, convex function which is γ-strongly convex for $\gamma \ge 0$ (if ρ is not strongly convex, we take $\gamma = 0$).

- We have the following sub-gradient identity, which follows by the KKT conditions applied to (2.4).

$$\boldsymbol{y} - \mathsf{prox}[\rho](\boldsymbol{y}) \in \partial \rho(\mathsf{prox}[\rho](\boldsymbol{y})). \tag{O.1}$$

- We have the following fixed point identity, which follows from (O.1).

$$\boldsymbol{y} = \mathsf{prox}[\rho](\boldsymbol{y}) \iff \boldsymbol{y} \text{ minimizes } \rho. \tag{O.2}$$

- $\mathsf{prox}[\rho]$ is firmly non-expansive [PB13, p. 131]. That is,

$$\langle \boldsymbol{y} - \boldsymbol{y}', \mathsf{prox}[\rho](\boldsymbol{y}) - \mathsf{prox}[\rho](\boldsymbol{y}') \rangle \ge (1 + \gamma) \| \mathsf{prox}[\rho](\boldsymbol{y}) - \mathsf{prox}[\rho](\boldsymbol{y}') \|^2. \tag{O.3}$$

- $\mathsf{prox}[\rho]$ is $(1+\gamma)^{-1}$-Lipschitz [PB13]. That is, for $\boldsymbol{y}, \boldsymbol{y}' \in \mathbb{R}^p$,

$$\|\mathsf{prox}[\rho](\boldsymbol{y}) - \mathsf{prox}[\rho](\boldsymbol{y}')\| \le (1+\gamma)^{-1}\|\boldsymbol{y} - \boldsymbol{y}'\|. \tag{O.4}$$

  This follows by applying Cauchy-Schwartz to the left-hand side of (O.3) and rearrangement.

- $\mathsf{prox}[\lambda\rho]$ satisfies the following continuity property in regularization parameter λ. For $\lambda > 0$, $\lambda' \ge 0$, we have

$$\|\mathsf{prox}[\lambda\rho](\boldsymbol{y}) - \mathsf{prox}[\lambda'\rho](\boldsymbol{y})\| \le \|\boldsymbol{y} - \mathsf{prox}[\lambda\rho](\boldsymbol{y})\| \left|\frac{\lambda'}{\lambda} - 1\right|, \tag{O.5}$$

  as we now argue. For simplicity, denote $\boldsymbol{a} = \mathsf{prox}[\lambda\rho](\boldsymbol{y})$. By (O.1), we have $\boldsymbol{y} - \boldsymbol{a} \in \lambda\partial\rho(\boldsymbol{a})$. Scaling by $\frac{\lambda'}{\lambda}$, we have $\frac{\lambda'}{\lambda}(\boldsymbol{y} - \boldsymbol{a}) \in \lambda'\partial\rho(\boldsymbol{a})$. Thus,

$$\left(\frac{\lambda'}{\lambda} - 1\right)(\boldsymbol{y} - \boldsymbol{a}) \in \partial\left(\frac{1}{2}\|\boldsymbol{y} - \boldsymbol{x}\|^2 + \lambda'\rho(\boldsymbol{x})\right)\Big|_{\boldsymbol{x}=\boldsymbol{a}}. \tag{O.6}$$

  Denote $\boldsymbol{a}' = \mathsf{prox}[\lambda'\rho](\boldsymbol{y})$. We have

$$\begin{aligned}\frac{1}{2}\|\boldsymbol{y} - \boldsymbol{a}\|^2 + \lambda'\rho(\boldsymbol{a}) &\ge \frac{1}{2}\|\boldsymbol{y} - \boldsymbol{a}'\|^2 + \lambda'\rho(\boldsymbol{a}') + \frac{1}{2}\|\boldsymbol{a} - \boldsymbol{a}'\|^2 \\ &\ge \frac{1}{2}\|\boldsymbol{y} - \boldsymbol{a}\|^2 + \lambda'\rho(\boldsymbol{a}) + \left\langle\left(\frac{\lambda'}{\lambda} - 1\right)(\boldsymbol{y} - \boldsymbol{a}), \boldsymbol{a}' - \boldsymbol{a}\right\rangle + \|\boldsymbol{a}' - \boldsymbol{a}\|^2,\end{aligned}$$

  where in both inequalities we have used the strong convexity of $\boldsymbol{x} \mapsto \frac{1}{2}\|\boldsymbol{y} - \boldsymbol{x}\|^2 + \lambda'\rho(\boldsymbol{x})$, in the first inequality we have used that this function has sub-gradient $\boldsymbol{0}$ at $\boldsymbol{a}'$ by optimality, and in the second inequality we have used (O.6). By Cauchy-Schwartz, rearrangement, and substitution for the values of $\boldsymbol{a}$ and $\boldsymbol{a}'$, we get (O.5).

- $\mathsf{prox}[\rho]$ is almost everywhere differentiable. This follows because $\mathsf{prox}[\rho]$ is Lipschitz [EG15]. Whenever we write the divergence $\operatorname{div}\mathsf{prox}[\rho]$ and the Jacobian $\mathrm{D}\,\mathsf{prox}[\rho]$, they are understood to be defined almost everywhere. For all $\boldsymbol{y}$ for which the left-hand sides are defined, we have by (O.4),

$$\operatorname{div}\mathsf{prox}[\rho](\boldsymbol{y}) \le \frac{p}{1+\gamma}, \tag{O.7}$$

$$\|\mathrm{D}\,\mathsf{prox}[\rho](\boldsymbol{y})\|_{\mathsf{op}} \le \frac{1}{1+\gamma}, \tag{O.8}$$

  and by (O.3),

$$\mathrm{D}\,\mathsf{prox}[\rho](\boldsymbol{y}) \succeq \boldsymbol{0}. \tag{O.9}$$

  By (O.9), we have

$$\operatorname{div}\mathsf{prox}[\rho](\boldsymbol{y}) \ge 0. \tag{O.10}$$

- We may apply Stein's Lemma (i.e. Gaussian integration by parts) to proximal operators. That is, for any $\tau \ge 0$,

$$\mathbb{E}_{\boldsymbol{z}}\left[\langle \boldsymbol{z}, \mathsf{prox}[\rho](\boldsymbol{\beta}_0 + \tau\boldsymbol{z})\rangle\right] = \frac{\tau}{p}\mathbb{E}_{\boldsymbol{z}}[\operatorname{div}\mathsf{prox}[\rho](\boldsymbol{\beta}_0 + \tau\boldsymbol{z})], \tag{O.11}$$

where $\boldsymbol{z} \sim \mathsf{N}(\boldsymbol{0}, \boldsymbol{I}_p/p)$. To see this, observe that real-valued function $z_i \mapsto \mathsf{prox}[\rho](\boldsymbol{\beta}_0 + \tau\boldsymbol{z})_i$, holding the other z_j's fixed, is Lipschitz continuous. Thus, it is the indefinite integral of its almost-everywhere derivative. Applying Stein's lemma [Ste81], averaging over the other z_j's, and summing over i yields (O.11).

- By (O.8) and (O.11), we have for any $\tau \geq 0$ and $\boldsymbol{b} \in \mathbb{R}^p$,

$$\mathbb{E}_{\boldsymbol{z}}\left[\langle \boldsymbol{z}, \mathsf{prox}[\rho](\boldsymbol{b} + \tau\boldsymbol{z})\rangle\right] \leq \frac{\tau}{1+\gamma}, \tag{O.12}$$

  where $\boldsymbol{z} \sim \mathsf{N}(\boldsymbol{0}, \boldsymbol{I}_p/p)$.

- Proximal operators obey the following identity under the change of scaling of ρ [PB13, p. 130]. Let $\tilde{\rho}(\boldsymbol{x}) = a\rho(b\boldsymbol{x})$. Then

$$\mathsf{prox}[\tilde{\rho}](\boldsymbol{y}) = \mathsf{prox}\left[ab^2\rho\right](b\boldsymbol{y})/b. \tag{O.13}$$

- Proximal operators shift by a constant under the following perturbation. For $\boldsymbol{c} \in \mathbb{R}^p$ fixed, let $\tilde{\rho}(\boldsymbol{x}) = -\langle \boldsymbol{c}, \boldsymbol{x}\rangle + \rho(\boldsymbol{x} - \boldsymbol{c})$. We have,

$$\begin{aligned}
\mathsf{prox}[\tilde{\rho}](\boldsymbol{y}) &= \arg\min_{\boldsymbol{x}}\left\{\frac{1}{2}\|\boldsymbol{y} - \boldsymbol{x}\|^2 - \langle \boldsymbol{c}, \boldsymbol{x}\rangle + \rho(\boldsymbol{x} - \boldsymbol{c})\right\} \\
&= \arg\min_{\boldsymbol{x}}\left\{\frac{1}{2}\|\boldsymbol{y} - (\boldsymbol{x} - \boldsymbol{c})\|^2 + \rho(\boldsymbol{x} - \boldsymbol{c})\right\} \\
&= \mathsf{prox}[\rho](\boldsymbol{y}) + \boldsymbol{c}.
\end{aligned} \tag{O.14}$$

- Proximal operators of separable functions are separable. In particular, if $\rho(\boldsymbol{x}) = \sum_{j=1}^p f(x_j)$ for some lsc, proper, convex $f : \mathbb{R} \to \mathbb{R} \cup \{\infty\}$, then for all j,

$$\mathsf{prox}[\rho](\boldsymbol{y})_j = \mathsf{prox}[f](y_j). \tag{O.15}$$

- The oracle penalty correspons to a decrease in both noise level and regularization. Precisely,

$$\mathsf{prox}[\lambda\rho^{(\gamma)}]\left(\boldsymbol{\beta}_0 + \tau\boldsymbol{z}\right) = \mathsf{prox}\left[\frac{\lambda}{\lambda\gamma+1}\rho\right]\left(\boldsymbol{\beta}_0 + \frac{\tau}{\lambda\gamma+1}\boldsymbol{z}\right). \tag{O.16}$$

  Indeed, for any $\boldsymbol{y}$

$$\begin{aligned}
\mathsf{prox}[\lambda\rho^{(\gamma)}](\boldsymbol{y}) &= \arg\min_{\boldsymbol{x}}\left\{\frac{1}{2}\|\boldsymbol{y} - \boldsymbol{x}\|^2 + \lambda\rho(\boldsymbol{x}) + \frac{\lambda\gamma}{2}\|\boldsymbol{x} - \boldsymbol{\beta}_0\|^2\right\} \\
&= \arg\min_{\boldsymbol{x}}\left\{\frac{1}{2}\left\|\frac{\lambda\gamma\boldsymbol{\beta}_0 + \boldsymbol{y}}{\lambda\gamma+1} - \boldsymbol{x}\right\|^2 + \frac{\lambda}{\lambda\gamma+1}\rho(\boldsymbol{x})\right\} = \mathsf{prox}\left[\frac{\lambda}{\lambda\gamma+1}\rho\right]\left(\frac{\lambda\gamma\boldsymbol{\beta}_0 + \boldsymbol{y}}{\lambda\gamma+1}\right).
\end{aligned}$$

  Eq. (O.16) corresponds to $\boldsymbol{y} = \boldsymbol{\beta}_0 + \tau\boldsymbol{z}$.

P Useful tools

We omit proof for the first five lemmas, which are easy to verify. Lemmas P.2, P.3, and P.4 appear as Lemmas 20, 21, and 22 of [BMN19]. The remaining lemmas in this section are well-known, and we provide citations for each.

Lemma P.1. *The probabilistic limit supremum satisfies the following. For any real valued random variables X_p, Y_p such that, for each p, X_p and Y_p are defined on the same probability space,*

$$\operatorname*{p\,lim\,sup}_{p\to\infty} X_p + Y_p \le \left(\operatorname*{p\,lim\,sup}_{p\to\infty} X_p\right) + \left(\operatorname*{p\,lim\,sup}_{p\to\infty} Y_p\right). \tag{P.1}$$

If $X_p \ge 0$ and $\mathrm{p}-\limsup_{p\to\infty} X_p < \infty$, then $X_p = O_{\mathrm{p}}(1)$. If $X_p, Y_p \ge 0$, then

$$\operatorname*{p\,lim\,sup}_{p\to\infty} X_p Y_p \le \left(\operatorname*{p\,lim\,sup}_{p\to\infty} X_p\right)\left(\operatorname*{p\,lim\,sup}_{p\to\infty} Y_p\right). \tag{P.2}$$

Lemma P.2 (Lemma 20 in [BMN19])**.** *Consider two sequences $f : (\mathbb{R}^p)^{\ell_1} \to \mathbb{R}^p$ and $g : (\mathbb{R}^p)^{\ell_2} \to \mathbb{R}^p$, $p \ge 1$, of uniformly pseudo-Lipschitz functions of order k such that $\|f(\mathbf{0})\|$, $\|g(\mathbf{0})\|$ are bounded over p. The sequence of functions $\varphi : (\mathbb{R}^p)^{\ell_1} \times (\mathbb{R}^p)^{\ell_2} \to \mathbb{R}$, $p \ge 1$ defined by $\varphi(\boldsymbol{x}, \boldsymbol{y}) = \langle f(\boldsymbol{x}), g(\boldsymbol{y})\rangle$ is uniformly pseudo-Lipschitz of order $2k$.*

Lemma P.3 (Lemma 21 in [BMN19])**.** *Let t, s, and k be any three positive integers. Consider a sequence (in p) of $\boldsymbol{x}_1, \dots, \boldsymbol{x}_s \in \mathbb{R}^p$ such that $\|\boldsymbol{x}_j\| \le c_j$ for some constants c_j independent of p, for $j = 1, \dots, s$, and a sequence (in p) of uniformly pseudo-Lipschitz functions $\varphi_p : (\mathbb{R}^p)^{t+s} \to \mathbb{R}$. Then the sequence of functions $\phi_p(\cdot) := \varphi_p(\cdot, \boldsymbol{x}_1, \dots, \boldsymbol{x}_s)$ is also uniformly pseudo-Lipschitz of order k.*

Lemma P.4 (Lemma 22 in [BMN19])**.** *Let t be any positive integer. Consider a sequence (in p) of uniformly pseudo-Lipschitz functions $\varphi_p : (\mathbb{R}^p)^t \to \mathbb{R}$ of order k. The sequence of functions $\phi_p : (\mathbb{R}^p)^t \to \mathbb{R}$ such that $\phi_p(\boldsymbol{x}_1, \dots, \boldsymbol{x}_t) = \mathbb{E}_{\boldsymbol{z}}\left[\varphi_p(\boldsymbol{x}_1, \dots, \boldsymbol{x}_{t-1}, \boldsymbol{x}_t + \tau\boldsymbol{z})\right]$ where $\boldsymbol{z} \sim \mathsf{N}(\mathbf{0}, \boldsymbol{I}_p/p)$ and $\tau \ge 0$ does not depend on p, is also uniformly pseudo-Lipschitz of order k.*

Lemma P.5. *If $f : \mathbb{R}^p \to \mathbb{R}^n$ is a sequence (in p) of uniformly pseudo-Lipschitz functions of order k and $g : \mathbb{R}^n \to \mathbb{R}^m$ is a sequence (in p) of uniformly pseudo-Lipschitz functions of order l such that $\|g(\mathbf{0})\|$ is bounded, then $g \circ f : \mathbb{R}^p \to \mathbb{R}^m$ is a sequence of uniformly pseudo-Lipschitz functions of order kl.*

Lemma P.6 (Stein's lemma [Ste81])**.** *Let $f : \mathbb{R}^p \to \mathbb{R}$ be "almost differentiable" in the sense that there exists measurable $\nabla f : \mathbb{R}^p \to \mathbb{R}^p$ such that, for all $\boldsymbol{\delta} \in \mathbb{R}^p$,*

$$f(\boldsymbol{x} + \boldsymbol{\delta}) - f(\boldsymbol{x}) = \int_0^1 \langle \boldsymbol{\delta}, \nabla f(\boldsymbol{x} + t\boldsymbol{\delta})\rangle \mathrm{d}t,$$

for almost every $\boldsymbol{x} \in \mathbb{R}^p$. In particular, this is satisfied if f is pseudo-Lipschitz. If $(\boldsymbol{z}_1, \boldsymbol{z}_2) \sim \mathsf{N}(\mathbf{0}, \boldsymbol{T} \otimes \boldsymbol{I}_p/p)$ for some $\boldsymbol{T} \in S_+^2$, then

$$\mathbb{E}_{\boldsymbol{z}_1, \boldsymbol{z}_2}\left[\langle \boldsymbol{z}_1, f(\boldsymbol{z}_2)\rangle\right] = \frac{T_{12}}{p}\mathbb{E}[\operatorname{div} f(\boldsymbol{z}_2)]. \tag{P.3}$$

Lemma P.7 (Tweedie's Formula. Eq. (2.8) in [Efr11]). *Fix $\tau > 0$. Fix π any probability measure on $\mathbb{R}$. Let $y = \beta_0 + \tau z$ where $\beta_0 \sim \pi$, $z \sim \mathsf{N}(0,1)$ independent. Let p_Y denote the density of y with respect to Lebesgue measure. Then*

$$\mathbb{E}_{\beta_0,z}\left[\beta_0 | y\right] = y + \tau^2 \frac{\mathrm{d}}{\mathrm{d}y} \log p_Y(y). \tag{P.4}$$

(See [Efr11] for a discussion of earlier references for this remarkable formula.)

Lemma P.8 (Gaussian concentration of Lipschitz functions. Theorem 5.6 in [BLM16]). *If $\boldsymbol{z} \sim \mathsf{N}(\mathbf{0}, \boldsymbol{I}_p/p)$ and $f : \mathbb{R}^p \to \mathbb{R}$ is an L-Lipschitz function, then for all $t > 0$,*

$$\mathbb{P}_{\boldsymbol{z}}\left(|f(\boldsymbol{z}) - \mathbb{E}_{\boldsymbol{z}}[f(\boldsymbol{z})]| \geq t\right) \leq e^{-\frac{p}{2L^2}t^2}. \tag{P.5}$$

Lemma P.9 (Gaussian Poincaré inequality. Theorem 3.20 in [BLM16]). *Let $\boldsymbol{z} \sim \mathsf{N}(\mathbf{0}, \boldsymbol{I}_p/p)$ and $\varphi : \mathbb{R}^p \to \mathbb{R}$ be continuous and weakly differentiable. Then, for some universal constant c,*

$$\mathsf{Var}[\varphi(\boldsymbol{z})] \leq \frac{c}{p}\mathbb{E}_{\boldsymbol{z}}\left[\|\nabla\varphi(\boldsymbol{z})\|^2\right]. \tag{P.6}$$