Electronic Journal of Statistics
Vol. 10 (2016) 3894–3944
ISSN: 1935-7524
DOI: 10.1214/16-EJS1212

Robustness in sparse high-dimensional linear models: Relative efficiency and robust approximate message passing*

Jelena Bradic

Department of Mathematics
University of California, San Diego
e-mail: jbradic@math.ucsd.edu

Abstract: Understanding efficiency in high dimensional linear models is a longstanding problem of interest. Classical work with smaller dimensional problems dating back to Huber and Bickel has illustrated the clear benefits of efficient loss functions. When the number of parameters p is of the same order as the sample size n, $p \approx n$, an efficiency pattern different from the one of Huber was recently established. In this work, we study relative efficiency of sparsity linear models with $p \gg n$. In the interest of deriving the asymptotic mean squared error for l_1 regularized M-estimators, we propose a novel, robust and sparse approximate message passing algorithm (RAMP), that is adaptive to the error distribution. Our algorithm includes many non-quadratic and non-differentiable loss functions. We derive its asymptotic mean squared error and show its convergence, while allowing $p, n, s \to \infty$, with $n/p \in (0, 1)$ and $n/s \in (1, \infty)$. We identify new patterns of relative efficiency regarding l_1 penalized M estimators. We show that the classical information bound is no longer reachable, even for light–tailed error distributions. Moreover, we show new breakdown points regarding the asymptotic mean squared error. The asymptotic mean squared error of the l_1 penalized least absolute deviation estimator (P-LAD) breaks down at a critical ratio of the number of observations per number of sparse parameters in the case of light-tailed distributions; whereas, in the case of heavy-tailed distributions, the asymptotic mean squared error breaks down at a critical ratio of the optimal tuning parameter of P-LAD to the optimal tuning parameter of the l_1 penalized least square estimator.

MSC 2010 subject classifications: Primary, 62G35, 62J07; secondary 60F05.
Keywords and phrases: Lasso, LAD, efficiency, robustness, sparsity, AMP.

Received August 2015.

Contents

*The author is supported by the DMS NSF grant # 1205296

1. Introduction

In recent years, scientific communities face major challenge with the size and complexity of the data generated. The size of such contemporary datasets and the number of variables collected makes the search for, and exploitation of, sparsity vital to their statistical analysis. Moreover, the presence of heterogeneity, outliers and anomalous data in such samples is very common. However, statistical estimators that are not designed for both sparsity and robustness to the data irregularities simultaneously will give biased results, depending on the "magnitude" of the deviation and on the "sensitivity" of the method. An example of an early work on robust statistics is [12, 11]. Specifically, they argue that robust estimators based on a minimization of non-differentiable loss functions are insensitive to changes not involving the parameters. Subsequently, [39] [25], [24] and [9] laid the comprehensive foundations of a theory of robust statistics. In particular, Huber's seminal work on M-estimators [26] established asymptotic properties of a class of M-estimators in the situation where the number of parameters, p, is fixed and the number of samples, n, tends to infinity. Since then, numerous important steps have been taken toward analyzing and quantifying robust statistical methods – notably in the work of [40, 17, 44], among others. Even today, there exist several (related) mathematical concepts of robustness (see [35]). This illustrates diverse and rich aspects of robustness. However, its

intricate dependence on the dimensionality of the parameter space hasn't been explored much.

Modern dataset, where number of parameters is larger than the number of samples led statisticians to move away from the M-estimators and to consider the penalized M-estimators. To further the focus on penalized M-estimators, we consider a linear regression model:

$$Y = \mathbf{A}x_o + W \tag{1.1}$$

with $Y = (Y_1, ..., Y_n)^T \in \mathbb{R}^n$ a vector of responses, $\mathbf{A} \in \mathbb{R}^{n\times p}$ a known design matrix, $x_o \in \mathbb{R}^p$ a vector of parameters; the noise vector $W = (W_1, ..., W_n)^T \in \mathbb{R}^n$ having zero-mean components each with distribution $F = F_w$ and a density function f_w. When $p \geq n$ a form of sparsity is imposed on the model parameters $\mathbf{x_o}$, i.e., it is imposed that $\text{supp}(\mathbf{x}_o) = \{1 \leq j \leq p : x_{oj} \neq 0\}$ with $|\text{supp}(\mathbf{x}_o)| = s$. Early work on penalized estimators include least squares loss (LS) with l_1-penalty, Lasso, [38], concave penalty, SCAD [21], MCP [45], adaptive l_1 penalty [48], elastic net penalty [47], and many more. However, when the error distribution F_w deviates from the normal distribution, the l_2 loss function is typically changed to the $-\log f_w$. Unfortunately, in applications the error distribution F_w is unknown and a method that adapts to many different distributions is needed. Following classical literature on M-estimators, penalized robust methods such as penalized Quantile regression [7], penalized Least Absolute Deviation estimator [41], AR-Lasso estimator [22], robust adaptive Lasso [1] and many more, have been proposed. These methods penalize a convex loss function ρ

$$\hat{x}(\lambda) \equiv \underset{x\in\mathbf{R^p}}{\text{argmin}}\ \mathcal{L}(x) = \underset{x\in\mathbf{R^p}}{\text{argmin}}\ \sum_{i=1}^{n} \rho(Y_i - A_i^T x) + \lambda \sum_{j=1}^{p} |x_j|. \tag{1.2}$$

From hereon, we refer to $\hat{x}(\lambda)$ as the l_1 penalized M-estimator. Despite the substantial body of work on robust M-estimators, there is very little work on robust properties of l_1 penalized M-estimators. Robust assessments of penalized statistical estimators customarily are made ignoring model selection. Typical properties discussed are model selection consistency or tight upper bounds on the statistical estimation error (e.g., [14, 36, 22, 23, 31, 32, 30, 42, 16]). In particular, the existing work has been primarily reduced to the tools that are intrinsic to Huber's M-estimators. In order to do that, the authors establish a model selection consistency and then reduce the analysis to this selected model assuming that that is the true model. However, this analysis is dissatisfactory, as the necessary assumptions for the model selection consistency are far too restrictive. Hence, departures from such considerations are highly desirable. This is where our work makes progress as our robustness analysis does not assume restrictive Irrepresentable condition (and hence perfect model selection).

This enables us to answer question like: in high dimensional regime, which estimator is preferred? In the low-dimensional setting, several independent lines of work provide reasons for using distributionally robust estimators over their

least-squares alternatives [27]. However, in high dimensional setting, it remains an open question, what are the advantages of using a complicated loss function over a simple loss function such as the squared loss? Can we better understand how differences between probability distributions affect penalized M-estimators? One powerful justification exists, using the point of view of statistical efficiency. Huber's proposed measure of robustness [26] allows a comparison of estimators by comparing their asymptotic variance; one caveat is that the two estimators need to be consistent up to the same order. For cases with $p \geq n$ little or nothing is known about the asymptotic variance of the robust estimator (1.2) as $p \to \infty$ whenever $n \to \infty$. Moreover, the penalized M-estimator is biased and shrinks many coefficients to zero. For such estimators, the set of parameters for which Hodge's super–efficiency occurs is not of measure zero. Hence, asymptotic variance may not be the most optimal criterion for comparison. This suggest that a different criterion for comparison needs to be considered in the high dimensional asymptotic regime where $n \to \infty$, $p \to \infty$ and $n/p \to \delta \in (0, 1)$. We examine the asymptotic mean squared error (AMSE). AMSE is an effective measure of efficiency as it combines both the effect of the bias and of the variance [17]. However, in $p \gg n$ regime, it is not obvious that the asymptotic mean squared error will satisfy the classical formula.

AMSE was studied in [5, 29] for the case of ridge regularization, with the penalty $\|x\|_2^2$, and when $p \leq n$ but $p \approx n$. In this setting AMSE is equal to the asymptotic variance of $\hat{\mathbf{x}}(\lambda)$. They discovered a new Gaussian component in the AMSE of $\hat{x}(\lambda)$ that cannot be explained by the traditional Fisher Information Matrix. To analyze AMSE for the case of no-penalization, with $p \approx n$, [20] utilized the techniques of Approximate Massage Passing (AMP) and discovered the same Gaussian component. The advantage of the AMP framework is that it provides an exact asymptotic expression of the asymptotic mean squared error of the estimator instead of an upper bound. For the case of the least squares loss with $p \geq n$, [4] make a strong connection between the penalized least squares and the AMP algorithm of [19]. However, the AMP algorithm of [4] cannot recover the signal when the distribution of the noise is arbitrary. For this settings, we design a new, robust and sparse Approximate Message Passing (RAMP) algorithm.

Our proposed algorithm belongs to the general class of first-order approximate massage passing algorithms. However, in contrast to the existing methods it has three-steps. It has iterations that are based on gradient descent with an objective that is scaled and min regularized version of the original loss function ρ. Moreover, it allows non-differentiable loss functions. The three–step estimation method of RAMP is no longer a simple proxy for the one-step M estimation. Due to high dimensionality with $p \geq n$, such a step is no longer adequate. Our proof technique leverages the powerful technique of the AMP proposed in [3]; however, we require a more refined analysis here in order to extend the results to one involving non-differentiable and robust loss functions while simultaneously allowing $p \geq n$. We relate the proposed algorithm to the Lasso penalized M estimators when $p \gg n$ and show that a solution to one may lead to the solution to the other. We show its convergence while allowing non-differentiable loss

functions and $p, n, s \to \infty$, with $n/p \to \delta \in (0,1)$ and $n/s \to a \in (1,\infty)$. This enabled us to derive the AMSE of a general class of l_1 penalized M-estimators and to study their relative efficiency.

We show that the AMSE depends on the distribution of the *effective score* and that it takes a form much different than the classical one, in that it also depends on the sparsity s. Moreover, we present a new study of the relative efficiency of the penalized least squares method and the penalized least absolute deviation method. We discover regimes where one is more preferred than the other and that do not match classical findings of Huber. Several important insights follow immediately: relative efficiency is considerably affected by the model selection step; the most optimal loss function may no longer be the negative log likelihood function; the sparsest high dimensional estimators have an additional Gaussian component in their asymptotic mean squared error that does not disappear asymptotically. Moreover we find that the l_1 penalized least squares (P-LS) is preferred over the l_1 penalized least absolute deviations (P-LAD) when the error distribution is "light-tailed" with a new breakdown point for which the two methods are indistinguishable; furthermore, we find that P-LS is never preferred over P-LAD when the error distribution is "heavy-tailed".

We briefly describe the notation. We use $\langle u \rangle \equiv \frac{\sum_{i=1}^m u_i}{m}$ to denote the average of the vector $u \in \mathbb{R}^m$. Moreover, if given $f : \mathbb{R} \to \mathbb{R}$ and $v = (v_1, ..., v_n)^T \in \mathbb{R}^m$, we define $f(v) \in \mathbb{R}^m \equiv (f(v_1), ..., f(v_m))$. Its subgradient $f'(v)$ is taken coordinate-wise and is $(f'(v_1), \ldots, f'(v_m))$. For bivariate function $f(u,v)$, we define $\partial_1 f(u,v)$ to be the partial derivate with respect to the first argument; similarly $\partial_2 f(u,v)$, is the partial derivate with respect to the second argument. We use $\|\cdot\|_1$ to denote l_1 and $\|\cdot\|_2$ to denote the l_2 norm. We define the sign function as $\text{sign}(v) = \mathbb{1}\{v > 0\} - \mathbb{1}\{v < 0\}$, and zero whenever $v = 0$. We set $\delta = n/p$ and $\omega = \mathbb{E}\|X_0\|_0$ with a vector X_0 following a p_{x_0} distribution. We set $\omega = s/p$ and θ denotes the nonnegative thresholding parameter. Moreover, $\eta : \mathbb{R} \times \mathbb{R}_+ \to \mathbb{R}$ be the soft thresholding function $\eta(x,\theta) = x - \theta$ if $x > \theta$, $\eta(x,\theta) = 0$ if $-\theta \le x \le \theta$ and $\eta(x,\theta) = x + \theta$ if $x < -\theta$.

This paper investigates the effects of the l_1 penalization on robustness properties of the penalized estimators, in particular, how to incorporate bias induced by the penalization in the exploration of robustness. We present a new approximate massage passing algorithm (RAMP) that is adaptable to different loss functions and sparsity simultaneously; including the one of Least Absolute Deviation (LAD) and Quantile loss (see Section 2). Section 3 studies a number of important theoretical results concerning the RAMP algorithm as well as its convergence properties and its connections to the penalized M-estimators. Section 4 studies Relative Efficiency and establishes lower bounds for the AMSE. Moreover, this section also presents results on relative efficiency of P-LAD estimator with respect to P-LS estimator. Section 6 contains detailed numerical experiments on a number of RAMP losses, including LS, LAD, and a number of Quantile losses, and a number of error distributions, including normal, mixture of normals and student. In 6.1- 6.3, we demonstrate how to use RAMP method in practice, its convergence properties and the study of state-evolution equation

where we find that the RAMP works extremely well. In 6.4, we demonstrate properties of the RAMP algorithm with varying error distribution. Lastly, in 6.5 we present analysis and new patterns of relative efficiency between P-LS and P-LAD estimators where we consider $p \leq n$, $s < n$, $p \geq n$ and $s \approx n$.

2. Robust sparse approximate message passing (RAMP) algorithm

We propose an iterative algorithm called RAMP, for "robust approximate message passing" that begins from the initial estimate $x^0 = 0 \in \mathbb{R}^p$ and guarantees a sparse estimator at its final iteration. Let the loss function $\rho : \mathbb{R} \to \mathbb{R}_+$ to be a non-negative convex function with a subgradient

$$\rho'(x) = \{y | \rho(z) \geq \rho(x) + y(z - x), \text{ for all } z \in \mathbb{R}\}.$$

For $b > 0$ let $G(z, b)$ denote the rescaled, min regularized effective score function, i.e.,

$$G(z; b) = \frac{\delta}{\omega} b\rho'(Prox(z, b)), \tag{2.1}$$

with the proximal mapping operator $Prox(z, b)$ defined as:

$$Prox(z, b) = \arg\min_{x \in \mathbb{R}} \left\{ b\rho(x) + \frac{1}{2}(x - z)^2 \right\}. \tag{2.2}$$

Lemma 5 (see Supplemental Materials) shows the reason behind the use of the effective score $G(z, b)$ in the RAMP algorithm. In particular it shows that for every $\lambda = \theta\omega/(b\delta)$, the solution to the penalized M-estimator problem (1.2) corresponds to the fixed point solution of the RAMP algorithm described below. For all convex and closed losses ρ, the operator $Prox(z, b)$ exists for all b and is unique for big enough b and all z. The proximal mapping operator is widely used in non-differentiable convex optimization in defining proximal-gradient methods. The parameter b controls the extent to which the proximal operator maps points towards the minimum of ρ, with smaller values of b providing a smaller movement towards the minimum. Finally, the fixed points of the proximal operator of ρ are precisely the minimizers of ρ; for appropriate choice of b, the proximal minimization scheme converges to the optimum of ρ, with least geometric and possibly superlinear rates ([8]; [33]).

2.1. RAMP algorithm

The RAMP algorithm below applies to sparse estimation with $p \geq n$ and loss functions ρ that are not necessarily differentiable and those that do not necessarily satisfy restricted strong convexity condition [36] (a condition typically used in the literature). The extension is significantly complicated, as the set of fixed points of the proximal operator is no longer necessarily sparse. Important examples of such loss functions ρ are absolute deviation and quantile loss, as they

are neither differentiable nor do they satisfy restricted strong convexity condition. Each iteration $t = 1, 2, 3, \ldots$ is defined through a three-step procedure to update its estimate $x^t \in \mathbb{R}^p$. We name the iteration steps as the Adjusted Residuals, the Effective Score and the Estimation Step.

Adjusted Residuals: Using the previous estimate x^{t-1} and a current estimate x^t, compute the adjusted residuals $z^t \in \mathbb{R}^n$

$$z^t = Y - \mathbf{A}x^t + \frac{1}{\delta} G(z^{t-1}; b_{t-1}) \left\langle \partial_1 \eta\big(x^{t-1} + \mathbf{A}^T G(z^{t-1}; b_{t-1}); \theta_{t-1}\big) \right\rangle \quad (2.3)$$

where $\delta = n/p < 1$. We add a rescaled product to the ordinary residuals $Y - \mathbf{A}x^t$, that explicitly depends on n, p and s. This step can be recognized as proximal gradient descent [6] in the variable x of the function ρ using the stepsize $\left\langle \partial_1 \eta\big(x^{t-1} + \mathbf{A}^T G(z^{t-1}; b_{t-1}); \theta_{t-1}\big) \right\rangle / \omega$.

Effective Score: Choose the scalar b_t from the following equation, such that the empirical average of the effective score $G(z; b)$ has the slope 1,

$$1 = \frac{1}{n} \sum_{i=1}^{n} \partial_1 G(z_i^t; b_t). \quad (2.4)$$

As $n/s > 1$, for differentiable losses ρ previous equation has at least one solution, as $G(z; b)$ is continuous in b and takes values of both 0 and ∞. Whenever, $\partial_1 G$ is not continuous it can be defined uniquely in the form

$$b_t = \frac{1}{2}(b_t^+ + b_t^-)$$

where $b_t^+ = \sup\{d > 0 : \frac{1}{n}\sum_{i=1}^n \partial_1 G(z_i^t; d) > 1\}$ and $b_t^- = \sup\{d < 0 : \frac{1}{n}\sum_{i=1}^n \partial_1 G(z_i^t; d) < 1\}$. For non-differentiable losses ρ, we consider two adaptations. First, we allow parameter b_t, which controls the amount of min regularization of the robust loss ρ function, to be adaptive with each iteration t. Second, we consider a population equivalent of the (2.4) first, then design an estimator of it and solve the fixed point equation. In more details, for non-differentiable losses we propose to consider

$$1 = \hat{\nu}(b_t), \quad (2.5)$$

for a consistent estimator $\hat{\nu} = \hat{\nu}(b_t)$ of a population parameter ν defined as

$$\nu(b_t) = \partial_1 \mathbb{E}\left[G(z^t; b_t)\right].$$

A particular form of $\hat{\nu}$ depends on the choice of the loss function ρ and the density of the error term f_W.

Estimation: Using the regularization parameter b_t determined by the previous step, update the estimate x^t as follows,

$$x^{t+1} = \eta\left(x^t + A^T G(z^t; b_t); \theta_t\right), \quad (2.6)$$

with the soft thresholding function η.

Remark 1. *The estimation step of the algorithm introduces the thresholding step needed for inducing sparsity in the estimator. However, in contrast to the existing methods the estimation step is adjusted with the appropriately scaled score function G, (2.1). The three–step estimation method of RAMP is no longer a simple proxy for the one-step M estimation. Furthermore, the residuals require additional scaling, i.e, a factor proportional to the fraction of sparse elements of the current iterate, in other words,* $\langle \partial_1 \eta(x^{t-1} + A^T G(z^{t-1}; b_{t-1}); \theta_{t-1}) \rangle$ *(see Lemma 4 below). Unlike least squares problems [19], rescaling of* δ/ω *in the above term is absolutely crucial for the convergence of the proposed algorithm.*

To the best of our knowledge, RAMP algorithms is the first that simultaneously allows robustness in the loss function and shrinkage in the estimators simultaneously. Robust AMP of [20] merely applies to the $p \leq n$ case; when $p > n$, the second step of their algorithm fails to iterate and the other two stages do not match with (1.2). RAMP algorithm has a different Adjusted Residuals step that incorporates sparsity directly and a different Effective Score step to allow $\delta < 1$. One may attempt to apply the AMP of [20] to a modified proximal mapping operator (2.2) by including the l_1 norm (penalty) directly. However, such an algorithm would not be a generalized AMP algorithm and its solution can be shown doesn't converge to the penalized M-estimator (1.2).

2.2. Examples

In the following we present a few examples of RAMP algorithm for different choices of the loss function ρ. Let $\Phi(z, b) = \omega G(z; b)/\delta$.

Example 1. [Absolute Deviation Loss] The Absolute Deviation loss function is defined as $\rho(x) = |x|$. We obtain

$$Prox(z, b) = \begin{cases} 0, & z \in (-b, b) \\ z - b \text{ sign}(z), & \text{otherwise} \end{cases}. \tag{2.7}$$

Observe that the form above is equivalent to the soft thresholding operator. Moreover, the Absolute Deviation effective score function becomes,

$$\Phi(z, b) = \begin{cases} z, & z \in (-b, b) \\ b \text{ sign}(z), & \text{otherwise} \end{cases}.$$

Since $\mathbb{E}\Phi(z, b) = \mathbb{E}[z\mathbb{1}(|z| \leq b)] + b\mathbb{P}(|z| > b)$, Condition (**R**), guarantees that

$$\nu\omega/\delta = F_z(b) - F_z(-b) - bf_z(b) + bf_z(-b),$$

for F_z, f_z denoting the distribution and density functions of z. Given a set of adjusted residuals $z_1^t, \ldots, z_n^t$, provided by (2.3) at any iteration t, b_t is a solution to an implicit function equation (2.5)

$$s/n = \hat{F}_z^t(b) - \hat{F}_z^t(-b) - b\hat{f}_z^t(b) + b\hat{f}_z^t(-b).$$

Example 2. [Quantile Loss] Let τ be a fixed quantile value and such that $\tau \in (0,1)$. The quantile loss function is defined as $\rho_\tau(x) = |x|\big((1-\tau)1\{x < 0\} + \tau 1\{x > 0\}\big) = \tau x_+ + (1-\tau)x_-$, for $x_+ = \max\{x, 0\}$ and $x_- = \min\{x, 0\}$. Hence,

$$Prox(z,b) = \begin{cases} z - b\tau, & z > b\tau \\ z - b(\tau-1), & z < b(\tau-1) \\ 0, & \text{otherwise} \end{cases}, \tag{2.8}$$

and with it that the Quantile score function becomes,

$$\Phi(z,b,\tau) = \begin{cases} z, & z \in (b\tau - b, b\tau) \\ b\tau, & z > b\tau \\ b(\tau-1), & z < b\tau - b \end{cases}.$$

For the case of the quantile loss $\nu\omega/\delta = \mathbb{E}\partial_1\Phi(z,b,\tau)$. Adding Condition (**R**) to the setup, we obtain $\nu = \partial_1\mathbb{E}\Phi(z,b,\tau)$. Narrowing the focus to $\mathbb{E}\Phi(z,b,\tau)$ we obtain $\mathbb{E}\Phi(z,b,\tau) = \mathbb{E}\left[z\mathbb{1}(z \le b\tau)\right] + \mathbb{E}\left[z\mathbb{1}(z \ge b(\tau-1))\right] + b\tau\mathbb{P}(z > b\tau) + b(\tau-1)\mathbb{P}(z < b(\tau-1))$. Now, refining the equation for ν we obtain

$$\nu = F_z(b\tau) - F_z(b(\tau-1)) - b\tau f_z(b\tau) + b(\tau-1)f_z(b(\tau-1)),$$

for F_z, f_z denoting the distribution and density functions of z. Given a set of adjusted residuals $z_1^t, \ldots, z_n^t$, provided by (2.3) at any iteration t, and a fixed $\tau \in (0,1)$, b_t is a solution to an implicit equation

$$s/n = \hat{F}_z^t(b\tau) - \hat{F}_z^t\big(b(\tau-1)\big) - b\tau\hat{f}_z^t(b\tau) + b(\tau-1)\hat{f}_z^t\big(b(\tau-1)\big).$$

In practice, $\hat{F}_z^t(b\tau)$ typically takes the form of an empirical cumulative distribution function $\hat{F}_z^t(b\tau) = \frac{1}{n}\sum_{i=1}^n \mathbb{1}\{z_i^t \le b\tau\}$. In contrast, there are numerous consistent estimators of $f_z(b\tau)$. For instance, by the asymptotic linearity results of Lemma 9, we consider

$$\frac{1}{h\sqrt{n}}\sum_{i=1}^n \left[\Phi(z_i^t + n^{-1/2}h; b\tau) - \Phi(z_i^t - n^{-1/2}h; b\tau)\right],$$

for a bandwidth parameter $h > 0$. In practice, it is difficult to obtain estimators $\hat{F}_z^t(b\tau)$ and $\hat{f}_z^t(b\tau)$ that are continuous functions of b. Hence, to solve the fixed point equations we implement a simple grid search and set b to be the average of the the first value of b on the grid for which the estimated function is bellow s/n and the the last value of b on the grid for which the estimated function is above s/n.

3. Theoretical considerations

In order to establish theoretical properties, we will impose a number of conditions on the density of the error term W, a class of robust loss functions ρ and a design matrix A. Although we assume that the error terms W_i's have bounded

density, we allow for densities with possibly unbounded moments and we do not assume any a-priori knowledge of the density f.

Condition (D): Let $W_1, \ldots, W_n$ be i.i.d. random variables with the distribution function F. Let F have two bounded derivatives f and f' and $f > 0$ in a neighborhood of $r_1, \cdots, r_k$, appearing in Condition (**R**)(i) below.

Condition (R): Let $i = 1, \ldots, n$. The loss function ρ is convex with sub-differential ρ'. Moreover, (i) for all $u \in \mathbb{R}$, $\rho'(u)$ is an absolutely continuous function which can be decomposed as

$$\rho'(u) = \upsilon_1(u) + \upsilon_2(u) + \upsilon_3(u)$$

where υ_1 has an absolutely continuous derivative υ_1', υ_2 is a continuous,piecewise linear continuous function, constant outside a bounded interval and υ_3 is a non-decreasing step function. In more details, $\upsilon_2'(u) = \alpha_\nu$, and $\upsilon_3(u) = \kappa_\nu$, for $r_\nu < u \leq r_{\nu+1}$, $\nu = 1, \cdots, k$, for $\alpha_0, \ldots, \alpha_k, \kappa_0, \ldots, \kappa_k \in \mathbb{R}$ with $\alpha_0 = \alpha_k = \kappa_0 = \kappa_k = 0$ and $-\infty = r_0 < r_1 < \cdots < r_k < r_{k+1} = \infty$, and $-\infty = \kappa_0 < \kappa_1 < \cdots < \kappa_k < \kappa_{k+1} = \infty$. Additionally, (ii) for all $u \in \mathbb{R}$, $|\rho'(u)| \leq k_0$, where k_0 is positive and bounded constant and (iii) the functional $h(t) = \int \rho(z-t)dF(z)$ has unique minimum at $t = 0$. Finally, (iv) for some $\delta > 0$ and $\eta > 1$, $\mathbb{E}\left[\sup_{|u|\leq\delta} |v_1^{''}(z+u)|\right]^\eta$ is finite; where, $v_1^{'}(z) = (d/dz)v_1(z)$ and $v_1^{''}(z) = (d^2/dz^2)v_1(z)$.

Condition (i) depict explicitly the trade–off between the smoothness of ϕ and smoothness of F. This assumption covers the classical Huber's and Hampel's loss functions. Although we allow for not necessarily differentiable loss functions, we consider a class of loss functions for which the sub-differential ρ' is bounded, a condition that is easily satisfied by many loss functions such are lad, quantile and Tukey's bi-squared loss. Condition (iii), is to assure uniqueness of the population parameter that we wish to estimate. Condition (iv) is essentially a moment condition that holds, for example, if $v_1^{''}$ is bounded and either $v_1^{''}(z) = 0$ for $z < a$ or $z > b$ with $-\infty < a < b < \infty$, or $\mathbb{E}|W|^{2+\epsilon} < \infty$ for some $\epsilon > 0$.

Condition (A): The design matrix $\mathbf{A}$ is such that A_{ij} are i.i.d and follow Normal distribution $\mathcal{N}(0, 1/n)$ for all $1 \leq i \leq n$ and $1 \leq j \leq p$.

While this setting is admittedly specific, the careful study of such matrix ensembles has a long tradition both in statistics and communications theory and is borrowed from the AMP formulation [4]. It simplifies the analysis significantly and can be relaxed if needed. In particular, it implies the Restricted Eigenvalue condition of [10]; that is, $\kappa(s,c) = \min_{J\subset\{1,\ldots p\},|J|\leq s} \min_{v\neq 0,\|v_{J^c}\|_1\leq c\|v_J\|_1} \frac{\|\mathbf{A}v\|_2}{\sqrt{n}\|v_j\|_2} > 0$ with high probability, as long as the sample size n satisfies $n > c'(1+8c)^2 s\log p/\kappa(s,c)^2$, for some universal constant c'. The integer s here plays the role of an upper bound on

the sparsity of a vector of coefficients X_0. Note that, with $c \geq 1$, the square submatrices of size $\leq 2s$ of the matrix $\frac{1}{n}\sum_{i=1}^{n} A_i^T A_i$ are necessarily positive definite.

3.1. State evolution of RAMP

In State Evolution (SE) formalism ([18],[19]), the asymptotic distribution of the residual and the asymptotic performance of the estimator can be measured while allowing $p \to \infty$. The parameter $\bar{\tau}_t^2$ can be considered as the state of the AMP algorithm. Moreover, the asymptotic mean squared error (AMSE), defined as

$$\text{AMSE} = \lim_{p\to\infty} \frac{1}{p}\sum_{i=1}^{p}(x_i^t - x_{0,i})^2,$$

is a function of a state evolution parameter $\bar{\tau}_t^2$. We show that RAMP converges and we offer how to compute $\bar{\tau}_t$, through a new iteration scheme adjusted for $p \geq n$ and not differentiable losses ρ.

Lemma 1. *Let Conditions* (**R**), (**D**) *and* (**A**) *hold. Then, the RAMP algorithm defined by the equations (2.3), (2.4) and (2.6) belongs to the general recursion of [3]. Let* $\bar{\sigma}_0^2 = \frac{1}{\delta}\mathbb{E}X_0^2$ *and let* X_0 *and* W *follow density* p_{X_0} *and* f_W *respectively, where* $\mathbb{E}W^2 = \sigma^2$. *Let* Z *be a standard normal random variable. Then, for all* $t \geq 0$ *the sequence* $\{\bar{\tau}_t^2\}_{t\geq 0}$ *is obtained by the following iterative system of equations:*

$$\bar{\tau}_t^2 = \mathbb{E}\left[G(W + \bar{\sigma}_t Z; b_t)\right]^2, \tag{3.1}$$

where

$$\bar{\sigma}_t^2 = \frac{1}{\delta}\mathbb{E}\left[\eta(X_0 + \bar{\tau}_{t-1}Z, \theta) - X_0)\right]^2, \tag{3.2}$$

is a state of the RAMP algorithm (2.3), (2.4) and (2.6).

In more details, define the sequence $\bar{\tau}_t^2$ by setting $\bar{\sigma}_0^2 = \frac{1}{\delta}\mathbb{E}[X_0^2]$ and with it $\bar{\tau}_0^2 = \frac{\delta^2}{\omega^2}\mathbb{E}[\Phi(W - \bar{\sigma}_0 Z; b(\bar{\tau}_0^2))]$. Then, the solution to the iterative equations (3.1) and (3.2), $\bar{\tau}_t^2$, can be defined as the solution to the iterative system of equations

$$\bar{\tau}_{t+1}^2 = \mathbb{V}(\bar{\tau}_t^2, b(\bar{\tau}_t^2), \theta(\bar{\tau}_t^2))$$

for

$$\mathbb{V}(\tau^2, b, \theta) = \frac{\delta^2}{\omega^2}\mathbb{E}[\Phi(W + \sigma Z; b)], \qquad \sigma^2 = \frac{1}{\delta}\mathbb{E}\left[\eta(X_0 + \tau Z, \theta) - X_0)\right]^2.$$

Lemma 2. *Let* ρ *be a convex function and let Conditions* (**R**), (**D**) *and* (**A**) *hold. For any* $\sigma^2 > 0$ *and* $\alpha > \alpha_{\min}$, *the fixed point equation*

$$\tau^2 = \mathbb{V}\left(\tau^2, b(\tau^2), \alpha\tau\right)$$

admits a unique solution $\tau^* = \tau^*(\alpha)$ *for all smooth loss functions* ρ. *Moreover,* $\lim_{t\to\infty}\tau_t = \tau^*(\alpha)$. *Further, the convergence takes place at any initial solution*

and is monotone. Additionaly, for all non-smooth loss functions the fixed point equation above, admits multiple solutions $\tau^* = \tau^*(\alpha)$. In such cases, the convergence take place but it depends on the initial solution and is monotone for each initialization.

Remark 2. *The display above offers explicit expressions of the additional Gaussian variable Z, its effects on the fixed points τ^* and σ^* and the loss function ρ. In the case of a simple Lasso estimator, with G being a rescaled least squares loss, $\bar{\tau}_t^2$ becomes $\sigma^2 + \delta^{-1}\mathbb{E}\left[\eta(X_0 + \bar{\tau}_{t-1}Z, \theta) - X_0\right]^2$ [3].*

Lemma 3. *Let Conditions Let Conditions* (**R**), (**D**) *and* (**A**) *hold. Let $\bar{\sigma}$ be a fixed point of the recursion* (3.1)-(3.2)*. For all twice differentiable losses ρ,*

$$\frac{\omega}{\delta} = \mathbb{E}\left[\partial_1 \Phi\left(W + \bar{\sigma}Z; b\right)\right],$$

where W and Z have F_W and $\mathcal{N}(0,1)$ distributions, respectively. Let $f_{C-\Phi(C;b)}$ denote the density of the random variable $C - \Phi(C;b)$ for $C = W - \bar{\sigma}_t Z$. Let the bandwidth, h, for the consistent estimator $\hat{\nu}$ be such that $h \to 0$ and $nh \to \infty$. Then, for the non-necessarily differentiable losses ρ,

$$\frac{\omega}{\delta} = \mathbb{E}\left[\partial_1 v_1\left(W + \bar{\sigma}Z; b\right)\right] + \sum_{\nu=1}^{k-1} \alpha_\nu b\left(f_{C-\Phi(C;b)}(r_{\nu+1}) - f_{C-\Phi(C;b)}(r_\nu)\right),$$

where v_1 is defined in Condition (**R**).

3.2. Asymptotic mean squared error

We say a function $\psi : \mathbb{R}^2 \to \mathbb{R}$ is pseudo-Lipschitz if there exist a constant $L > 0$ such that for all $x, y \in \mathbb{R}^2$: $|\psi(x) - \psi(y)| \leq L(1 + \|x\|_2 + \|y\|_2)\|x - y\|_2$.

Theorem 1. *Let Conditions* (**R**), (**D**) *and* (**A**) *hold and let $\psi : R \times R \to R$ be a pseudo-Lipschitz function. Moreover, X_0 follows p_{x_0}, which is a non-degenerate distribution. Then, almost surely*

$$\lim_{p\to\infty} \frac{1}{p} \sum_{i=1}^{p} \psi(x_i^{t+1}, x_{0,i}) = \mathbb{E}\{\psi(\eta(X_0 + \bar{\tau}_t Z; \theta_t), X_0)\},$$

for all $\bar{\tau}_t$ and $\bar{\sigma}_t$ defined by the recursion (3.1)-(3.2) *and $\theta_t = \alpha\bar{\tau}_t$.*

Next, we measure the L_2 norm distance between the RAMP iteration and the penalized estimator.

Theorem 2. *Let Conditions* (**R**), (**D**) *and* (**A**) *hold. Let $\hat{x}(\lambda)$ be the l_1 penalized M-estimator and let $\{x^t\}$ be the sequence of estimates produced by the RAMP algorithm with $\theta_t = \alpha(\lambda)\bar{\tau}_t$ with $\bar{\tau}_t$ satisfying* (3.1)-(3.2)*. Then,*

$$\lim_{t\to\infty}\lim_{p\to\infty} \frac{1}{p}||x^t - \hat{x}(\lambda)||_2^2 = 0,$$

for all $\lambda > 0$ *satisfying*

$$\lambda = \frac{\alpha\bar{\tau}}{b\delta}\mathbb{P}(|X_0 + \bar{\tau}Z| \geq \alpha\bar{\tau}), \tag{3.3}$$

with $\bar{\tau}$ *defined in Lemma 3 and* $\alpha = \alpha(\lambda)$.

Previous theorem extends the existing work on approximate message passing with shrinkage factors, as the latter only focuses on least squares losses. Our result above allows both non-differentiable but also loss functions that do not necessarily satisfy a restricted strong convexity condition - as both least absolute deviation and quantile losses don't. Per results above, we see that for an optimal value of λ there exists an optimal value of α so that the optimally tuned l_1 penalized estimator is approximated by an an optimally tuned RAMP solution.

Theorem 3. *Let Conditions* (**R**), (**D**) *and* (**A**) *hold. Denote with* $\hat{x}$ *the penalized M-estimator. Let* $\psi : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$ *be a pseudo-Lipschitz function. Then, we conclude*

$$\lim_{p\to\infty} \frac{1}{p}\sum_{i=1}^{p} \psi(\hat{x}_i(\lambda), x_{0,i}) = \mathbb{E}\left\{\psi\left(\eta\left(X_0 + \bar{\tau}Z; \alpha(\lambda)\bar{\tau}\right), X_0\right)\right\},$$

for $\alpha = \alpha(\lambda)$ *and all* λ *satisfying* (3.3) *and* $\bar{\tau}$ *defined in Lemma 3.*

Remark 3. *This result offers not only an upper bound on* $AMSE(\hat{x}, x_0)$*, but also an exact expression of it for an appropriate choices of the tuning parameters. Note the optimal choice of* λ *truly depends on the loss function (through* $\bar{\tau}$*). Choosing* $\psi(x, y) = (x - y)^2$*, we obtain the AMSE path of the RAMP*

$$AMSE(x^t, x_0) = \lim_{p\to\infty} \frac{1}{p}\sum_{i=1}^{p}(x_i^t - x_{0,i})^2 = \mathbb{E}\left[\eta\left(X_0 - \bar{\tau}_t Z; \alpha(\lambda)\bar{\tau}\right) - X_0\right]^2. \tag{3.4}$$

4. Relative efficiency

The robustness properties of sparse, high-dimensional estimators are difficult to quantify due to the shrinkage effects and the subsequent bias in estimation. Shrinkage is known to lead to the super-efficiency phenomena in the domain of classical efficiency studies. Hence, efficiency cannot distingush between two biased estimators. However, relative efficiency, can capture both the size of the bias and the variance together leading to a relevant robustness evaluation.

According to Theorem 3, the asymptotic mean squared error of penalized M-estimators is

$$\delta\mathbb{E}\left[\eta\left(X_0 + \bar{\tau}_t Z, \lambda\frac{\delta}{\omega}b_t\right) - X_0\right]^2, \tag{4.1}$$

with the expectation taken with respect to X_0 and Z and

$$\bar{\tau}_t^2 = \frac{\mathbb{E}\left[\Phi^2\left(W + \bar{\sigma}_t Z; b_t\right)\right]}{\left[\mathbb{E}\Phi'(W + \bar{\sigma}_t Z; b_t)\right]^2}. \tag{4.2}$$

Remark 4. *Observe that whenever $1/\delta = O(1/n)$ i.e., $p \leq n$ and p does not grow with n, $\bar{\sigma}_t^2 = \bar{\tau}_{t-1}^2$. Specifically, when $p = o(n)$, the bias in estimation disappears [20]. In contrast, we observe that Gaussian Z component or the estimation bias never disappear with $1/\delta = p/n \geq 1$. This indicates that efficiency, with $p \geq n$, never converges to the low-dimensional case unless perfect model selection is achieved. Whenever we allow deviations of model selection consistency the additional Z component has a substantial role in both the size of the asymptotic variance and the asymptotic bias, even if $s \ll n$.*

The following result computes the asymptotic lower bound of the variance and AMSE of the RAMP estimator.

Theorem 4. *Suppose that W has a well-defined Fisher information matrix $I(F_W)$. Let τ_t and σ_t be the state evolution parameters following equations (3.1) and (3.2). Assume X_0 is not identically zero, $P_{X_0}\{X_0 \neq 0\} > 0$ and let $\mathbb{E}\|X_0\|_0 = s/n$. Then, under conditions (**R**), (**D**) and (**A**) (i) for every iteration t of the RAMP algorithm (2.3), (2.4), (2.6), state variable τ_t satisfies*

$$\tau_t^2 \geq \frac{\omega}{\delta} \frac{1 + \sigma_t^2 I(F_W)}{I(F_W)};$$

(ii) for a fixed point solution (τ^*, σ^*) of the RAMP algorithm with all $\alpha \geq \alpha_{\min} > 0$

$$\tau^{*2} \geq \frac{s}{n-s} \frac{1}{I(F_W)};$$

(iii) for fixed values of $\alpha = \alpha(\lambda)$ and X_0, with $\theta = \alpha\tau$, there exist functions ν_1, ν_2 that are convex and increasing, respectively, and are such that the asymptotic mean squared error mapping for high dimensional problems satisfies:

$$AMSE(\tau^{*2}, b(\tau^{*2}), \alpha\tau^*) = \nu_1(\tau)\tau^2 + \nu_2(\tau),$$

with

$$\nu_1(\tau) = 1 + \alpha^2 - \mathbb{E}_{X_0}\left[\alpha^2\left(\Phi(\alpha - \frac{X_0}{\tau}) - \Phi(-\alpha - \frac{X_0}{\tau})\right)\right.$$
$$\left. - \left(\alpha + \frac{X_0}{\tau}\right)\phi(\alpha - \frac{X_0}{\tau}) - \left(\alpha - \frac{X_0}{\tau}\right)\phi(-\alpha - \frac{X_0}{\tau})\right]$$

and $\nu_2(\tau) = \mathbb{E}_{X_0}\left[X_0^2\left(\Phi(\alpha - \frac{X_0}{\tau}) - \Phi(-\alpha - \frac{X_0}{\tau})\right)\right]$.

Remark 5. *Recall that traditional lower bound of M-estimators with pfixed and $p \leq n$ and $n \to \infty$ is $1/I(F_W)$ and is such that asymptotic mean squares error is equal to the variance and is achievable. Theorem 4 implies that under diverging p and s and n, such that $p \gg n \geq s$, traditional lower bound is not achievable for all $s \geq n/2$, i.e., for all "dense" high dimensional problems. Hence, we observe a new phase transition regarding robustness in high dimensional and sparse problems. The effect of sparsity is extremely clear. If the problem is significantly sparse, with $n/s < \infty$, then the traditional information bound may be achieved, whereas for all other problems the traditional information bound cannot be achieved, as there is inflation in the variance.*

5. Relative efficiency of l_1-penalized least squares and l_1-penalized absolute deviations

Next, we study the relative efficiency of the l_1 penalized least squares (P-LS from hereon) estimator, with respect to the l_1 penalized least absolute deviation (P-LAD from hereon) estimator.

Remark 6. *From the results above, we can clearly compute the asymptotic mean squared error of the P-LS and P-LAD as the recursive equations*

$$\tau^2_{P\text{-}LS} = \sigma^2_W + \sigma^2_{P\text{-}LS}, \tag{5.1}$$

$$\tau^2_{P\text{-}LAD} = \frac{\mathbb{E}\left[(W+\sigma_{P\text{-}LAD}Z)^2\,\mathbb{1}\left\{|W+\sigma_{P\text{-}LAD}Z|\le b\right\}\right] + b^2\mathbb{P}\left(|W+\sigma_{P\text{-}LAD}Z|>b\right)}{\mathbb{P}^2\left(|W+\sigma_{P\text{-}LAD}Z|\le b\right)}. \tag{5.2}$$

Here, both $\sigma_{P\text{-}LAD}$ and $\sigma_{P\text{-}LS}$ satisfy the equation of (3.2) with $\tau_{P\text{-}LAD}$ and $\tau_{P\text{-}LS}$, respectively.

Notice that in, sparse, high dimensional setting, the distribution of the X_0 can be represented as a convex combination of the Dirac measure at 0 and a measure that doesn't have mass at zero. Let us denote with Δ and U two random variables, each having the two measures above. Then, the asymptotic mean squared error satisfies

$$\delta\sigma^2 = \delta\tau^2\left((1-\omega)(\mathbb{E}_Z\eta(Z,\alpha))^2 + \omega\mathbb{E}_{(U,Z)}\left[\eta\left(\frac{U}{\tau}+Z;\alpha\right)-\frac{U}{\tau}\right]^2\right).$$

We will explore this representation to study the relative efficiency of P-LS and P-LAD estimators. The relative efficiency of P-LS vs. P-LAD is defined as the quotient of their asymptotic mean squared errors. By results of previous sections, this amounts to the quotient of $\sigma^2_{\text{P-LS}}/\sigma^2_{\text{P-LAD}}$. To evaluate this quotient, we study the behavior of $\sigma^2_{\text{P-LS}}/\sigma^2_W$ and $\sigma^2_{\text{P-LAD}}/\sigma^2_W$ independently. In order to do so, we need a preparatory result below.

Theorem 5. *Let Conditions* (**R**), (**D**) *and* (**A**) *hold. Let $\bar\sigma^2_{P\text{-}LAD}$ be a fixed point solution to the state-evolution system of equations (3.1) and (3.2), with a loss $\rho(x)=|x|$. Let σ^2_W be a variance of the error term W (1.1). Then, $\tau^2_{P\text{-}LAD}\to 0$ and $\sigma^2_{P\text{-}LAD}\to 0$, whenever $\sigma^2_W\to 0$ and $\tau^2_{P\text{-}LAD}\to\infty$ and $\sigma^2_{P\text{-}LAD}\to\infty$, whenever $\sigma^2_W\to\infty$.*

Next, we consider a class of distributions f_W such that σ^2_W exists and consider state variable $\sigma^2_{\text{P-LAD}}$ as a function of σ^2_W. We provide limiting behavior when both $p,n\to\infty$ of both P-LS and P-LAD. We separate the analysis further into two cases: the case of "light tailed distributions" and the case of "heavy-tailed distribution."

Theorem 6. *Let Conditions* (**R**), (**D**) *and* (**A**) *hold. Let $\bar\sigma^2_{P\text{-}LAD}$ and $\bar\sigma^2_{P\text{-}LS}$ be a fixed point solution to the state-evolution system of equations (3.1) and (3.2) with a loss $\rho(x)=|x|$ and a loss $\rho(x)=(x)^2$, respectively and an optimal choice*

of the tuning parameters λ_{LAD} and λ_{LS}. Let σ_W^2 be a variance of the error term W (1.1). In turn, if $M(\omega) < \delta$,

$$\lim_{\sigma_W^2 \to 0} \lim_{p \to \infty} \frac{\bar{\sigma}^2_{P\text{-}LS}}{\sigma_W^2} = \frac{1}{1 - M(\omega)/\delta}, \qquad \lim_{\sigma_W^2 \to 0} \lim_{p \to \infty} \frac{\bar{\sigma}^2_{P\text{-}LAD}}{\sigma_W^2} = \infty,$$

with $M(\omega) = \inf_\tau \left\{ (1-\omega)\mathbb{E}\eta^2(Z;\tau) + \omega \sup_{\mu \geq 0} \mathbb{E}\left(\eta(\mu + Z;\tau) - \mu\right)^2 \right\}$, where $\mathbb{E}$ is with respect to Z.

Remark 7. *A recent work [46] proved that $M(\omega)/\delta \geq \omega/\delta$. Together with the results of Lemma 6, we can see that the LAD method has less efficiency than the LS method for all of $\omega < \delta$. In other situations where $\omega \to \delta$, both limits on the right hand side of Lemma 6 are infinity and the two methods are inseparable. Classically, the LS method is more efficient than the LAD method. However, with high dimensional asymptotic, where $s \to n$, the breakdown point is where $M(\omega) = \delta$, that is,*

$$\sup \left\{ \omega : \inf_\tau \left[(1-\omega)\mathbb{E}\eta^2(Z;\tau) + \omega \sup_{\mu \geq 0} \mathbb{E}\left(\eta(\mu + Z;\tau) - \mu\right)^2 \right] < \delta \right\}.$$

Next, we provide limiting behavior of both P-LS and P-LAD in cases where $\sigma_W^2 \to \infty$; that is, in the case of "heavy tailed distributions."

Theorem 7. *Let Conditions* (**R**), (**D**) *and* (**A**) *hold. Let $\bar{\sigma}^2_{P\text{-}LAD}$ and $\bar{\sigma}^2_{P\text{-}LS}$ be a fixed point solution to the state-evolution system of equations (3.1) and (3.2) with a loss $\rho(x) = |x|$ and a loss $\rho(x) = (x)^2$, respectively and an optimal choice of the tuning parameters λ_{LAD} and λ_{LS}. Let σ_W^2 be a variance of the error term W (1.1). Then, if $\Gamma < \delta$,*

$$\lim_{\sigma_W^2 \to \infty} \lim_{p \to \infty} \frac{\bar{\sigma}^2_{P\text{-}LS}}{\sigma_W^2} = \frac{1}{1 - \Gamma(\alpha_{LS})/\delta}, \qquad \lim_{\sigma_W^2 \to \infty} \lim_{p \to \infty} \frac{\bar{\sigma}^2_{P\text{-}LAD}}{\sigma_W^2} = \frac{\Gamma(\alpha_{LAD})}{\delta},$$

with $\alpha = \alpha(\lambda)$, (3.3) and $\Gamma(\alpha) = \mathbb{E}\eta^2(Z;\alpha)$, with $\mathbb{E}$ is with respect to Z.

The result above is a path-dependent result, in the sense that it holds for every value of α as well; that is, for a sequence of λ values we can find an accompanying sequence of α values and the limits above would still apply (note that the right hand sides depend on α explicitly).

Remark 8. *As $1/(1 - \Gamma(\alpha)/\delta) \geq \Gamma(\alpha)/\delta$ and Γ is an increasing function, optimal P-LAD is more efficient than the optimal P-LS if α_{LS} and α_{LAD} are such that $\alpha_{LS} \geq \alpha_{LAD}$. However, the size of the optimal tuning parameter are unknown in general, hence further studies need to be developed.*

6. Numerical examples

Within this section, we would like to show the finite sample performance of RAMP.

6.1. Tuning parameter selection & implementation

The policy to choose for thresholds θ_t is based on [19], which sets $\theta_t = \alpha\bar{\tau}_t$, where α is taken to be fixed. We choose a grid of α within an interval $[\alpha_{min}, \alpha_{max}]$. For each α, we get the RAMP estimator x^t and SE iterative parameters $\bar{\tau}_t$ and $\bar{\sigma}_t$. We use these parameters to evaluate the AMSE(x^t, x_0) and then tune the optimal α by minimizing AMSE(x^t, x_0). In other words, $\bar{\tau}_t$ is calculated by the recursion $\bar{\tau}_t^2 = \mathbb{V}(\bar{\sigma}^2, \alpha\bar{\tau}_t)$, where $\mathbb{V}$ is the right hand side of equation (3.1) and $\bar{\sigma}$ is calculated from equation (3.2). A number of simulation sections substitute $\theta(\alpha)$ to be λ as a tuning parameter based on Lemma 5, in order to do a path-wise comparison between RAMP and penalized estimator. However, relative efficiency is always studies for only optimally tuned RAMP and optimally tuned penalized estimator. In the simulations each element of A is i.i.d. and follows $N(0, 1/n)$. Unless otherwise stated we consider a fixed ratio $\delta = 0.64$. The distribution of the true parameter is set as $\mathbb{P}(x_0 = 1) = \mathbb{P}(x_0 = -1) = 0.064$ and $\mathbb{P}(x_0 = 0) = 0.872$.

6.2. Existence and uniqueness of state evolution parameters

In this section only we work with $\alpha = 2$ to illustrate the worst case behavior of the RAMP alpgorithm. We fix $p = 500$ and focus on Gaussian distribution $\mathcal{N}(0, 0.2)$ for the errors W. Results of the state evolution equations are presented in Figures 1 and 2 below, where in the Gaussian setting above, we consider the least absolute deviation loss and the quantile losses with $\tau = 0.7$ and $\tau = 0.3$. We observe that the unique value of the state-evolution recursions is easily found even for the non-differentiable losses, under the recommendations of Section 2. Figure 1 shows how $\bar{\tau}_t^2$ evolves to the fixed point near 2.264, 2.933, 3.378 for the case of the least absolute deviations and quantile losses, respectively. Simultaneously the mapping $\mathbb{V}(\bar{\tau}^2, b, \theta)$ evolves to the fixed points near 2.260, 2.926, 3.359 for all non-differentiable losses. Moreover, Figure 2 illustrates that the loss is not great, even when we start from the randomly chosen starting α value.

6.3. Limit behavior of the parameters of RAMP

We assess the limit behaviors of parameters of different loss functions to express the iterations of the RAMP algorithm. The error W follows $N(0, 0.2)$ and the sample size is 320. We use $\omega = s/p = 0.128$ based on the setting of the p_{x_0} into equation (2.3) to generate b. We generate a series of α, and regard the threshold $\theta_t = \alpha * \bar{\tau}_t$. Then, we use the iteration of $\bar{\sigma}_t$, $\bar{\tau}_t$ from Lemma 1 to find the stable point $\bar{\tau}^*$ with stopping at $|\bar{\tau}_t - \bar{\tau}_{t-1}| < tol$, where tol is a small positive number and is taken to be 10^{-6} here. Lastly, we use the expression of $\lambda = \frac{\alpha\bar{\tau}^*\omega}{b\delta}$ and the expression of AMSE in Theorem 2 to find the AMSE(x^t, x_0). The penalized M-estimators theory suggest cross-validation for the optimal values of λ. For such value we find its corresponding AMSE(x^t, x_0) and present it in Table 1 below.

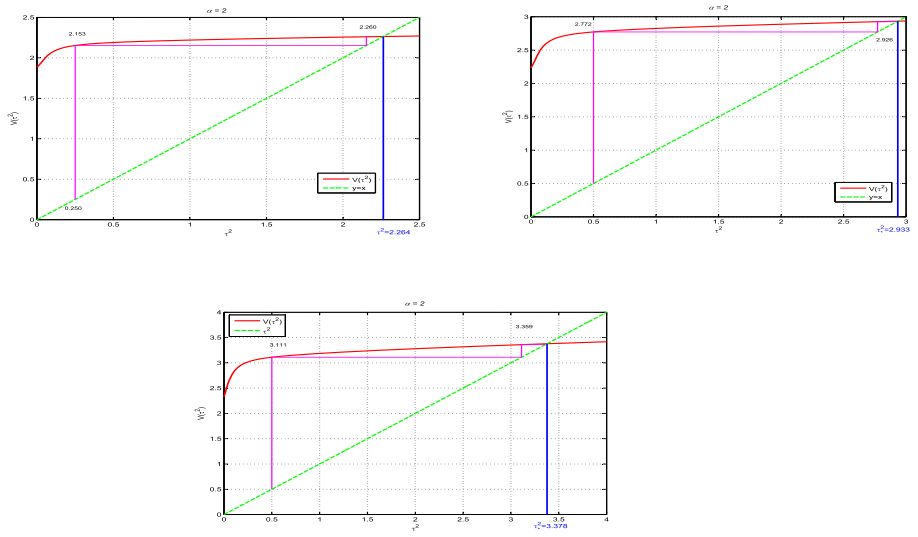

FIG 1. *Existence and uniqueness of* $\bar{\tau}^2$ *with non-differentiable loss functions: the absolute loss and the quantile losses with* $\tau = 0.7$ *and* $\tau = 0.3$*, respectively.*

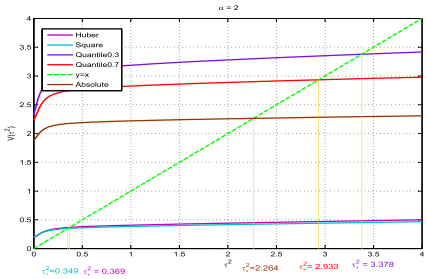

FIG 2. *Comparative plot of the mapping* $\mathbb{V}(\bar{\tau}^2)$ *with different loss functions with the error* W *following standard normal distribution and with fixed* $\delta = 0.64$ *and* $p = 500$*.*

TABLE 1
Convergence of RAMP iteration with different loss function

Loss Function	b	optimal λ	iteration steps	$\bar{\tau}^{*2}$	AMSE
Square Loss	0.2711864	0.6970546	8	0.3265822	0.0810126
Huber Loss	0.2714135	0.6261463	12	0.3431436	0.09150527
Absolute Loss	0.4990769	1.91523	8	2.0276825	0.0943257
Quantile Loss	0.7319994	1.402867	11	2.821827	0.1177329

Table 1 compares several necessary parameters in the iteration of the RAMP algorithm. We contrast four different loss functions: Least Squares loss, Huber loss, Least Absolute Deviation loss and Quantile Loss. The results presented in the table are averages over 100 repetitions. We notice that within only twenty

iteration steps, the RAMP algorithm becomes stable no matter of the loss function considered. Furthermore, we present values of a number of parameters of the RAMP algorithm: min-regularization b, regularization λ and state evolution $\bar{\tau}^*$. We observe that they all differ according to the loss function considered, illustrating that there is no universal choice of the above parameters that works uniformly well for all loss function.

Additionally, we present Figure 3 and show the empirical convergence of asymptotic variance $\text{AVAR}(x^t, x_0)$ with respect to the tuning parameter λ and different loss functions. The plots illustrate the bias-variance decomposition. For example, that when λ becomes larger, the $\text{AVAR}(x^t, x_0)$ of RAMP decreases dramatically and stabilizes around 0.136 for the case of Least Squares loss and Normal errors W. The reason $\text{AVAR}(x^t, x_0)$ becomes fixed on 0.12 is because the RAMP algorithm shrinks the estimator x^t to be the zero vector for which $\text{AMSE}(x^t, x_0) = ||x^t||_2^2 = 0.064 + 0.064 = 0.128$ and asymptotic $\text{Bias}^2 = 0.008$, when λ is large enough. We also see that for the optimal value of λ the AVAR of P-LS and P-LAD changes depending on the error distribution: for Normal errors, the optimal AVAR of P-LS $\leq$ than that of P-LAD (notice that the tuning is done independently and that the scale of λ is different); for Student errors we observe that the optimal AVAR of P-LS is $\geq$ than that of P-LAD.

6.4. Robustness of RAMP with respect to the error distribution

Further, we know that using square loss to solve problem (1.2) is very sensitive with respect to the error distribution, which is the reason we release the loss function from the least squares loss to the general convex loss function satisfying Condition **(R)**. We consider the robustness of the solution when the tail of error in model varies. We considered $n = 640$ observations and compared five scenarios for the error vector w: (a) light–tailed distribution: Normal $\mathcal{N}(0, 0.2)$, Mixnormal $0.5\mathcal{N}(0, 0.3) + 0.5\mathcal{N}(0, 1)$ and (b) heavy–tailed distribution: t_8, t_4, MixNormal $0.7\mathcal{N}(0, 1) + 0.3\mathcal{N}(0, 3)$ and Cauchy$(0, 1)$. The Mixture of Normals distribution generates samples from different normal distributions with corresponding probability and samples are centered to have mean zero.

Results of this experiment are presented in Figure 4. A few observations immediately follow. The Lasso estimator is sensitive to the heavy tail error distribution whereas, the Huber loss and the Least Absolute Deviation loss perform better as the tail of the error distribution becomes heavier. Moreover, with larger tails the Least Absolute Deviation loss is clearly preferred over both the Huber and the Least Squares loss, whereas situation reverses when the tails are light. The Mixture of Normals errors are particularly difficult due to the bimodality of the error distribution. We see that in both light and heavy tales cases of Mixture distribution, Huber Loss is preferred over the Least Squares loss. Lastly, as the tails becomes even heavier, all estimators face the problem of estimating the unknown parameter accurately.

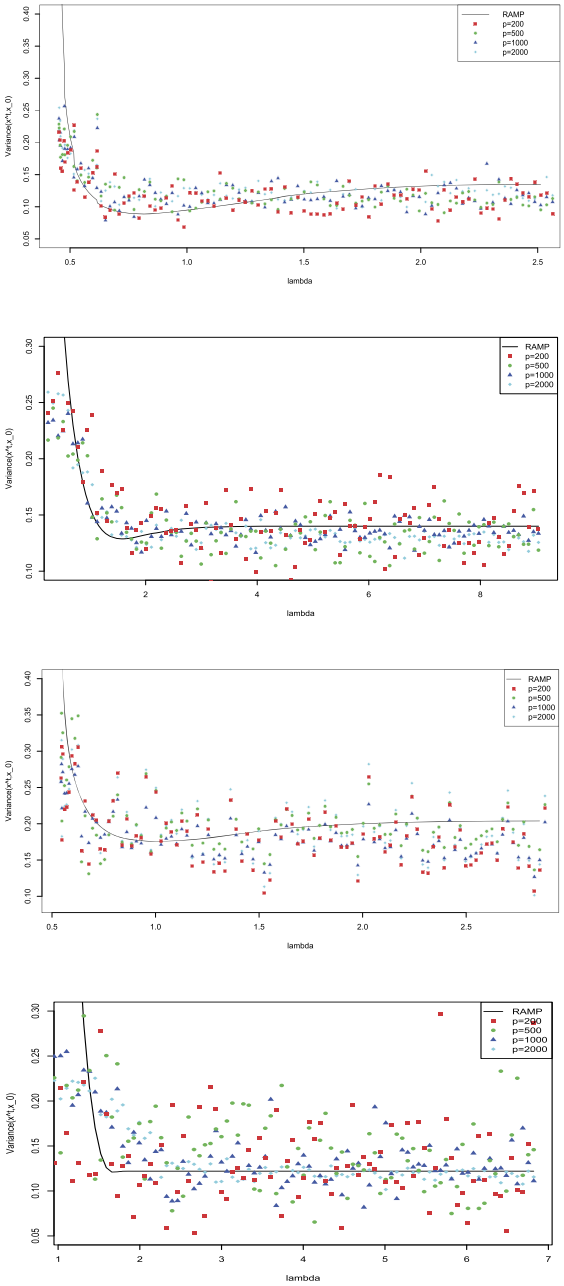

FIG 3. *AVAR*(x^t, x_0) *compared to AVAR*$(\hat{x}(\lambda_{opt}), x_0)$ *under Normal (first row) and Student* t_4 *(second row) error distributions and two loss functions: least square (first column) and least absolute deviation (second column)*

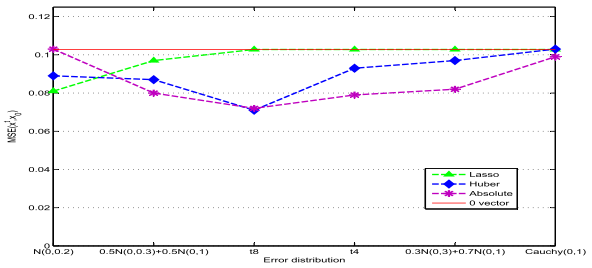

FIG 4. *$AMSE(x^t, x_0)$ under various error distribution settings*

6.5. Relative efficiency

We use RAMP iteration to calculate the relative efficiency of the Least square estimator versus the Least absolute estimator. It is known that the least square estimator is preferable in normal error assumption, but the least absolute estimator beats the least square estimator in double-exponential error assumption under classical low-dimensional setting.

In Table 2, we fix $p = 50$ and discuss the comparison of relative efficiency between the low-dimensional case (where $p < n$) and the high-dimensional dense case (where $p \approx n$). We discuss the $\text{AMSE}(x^t, x_0)$ with a different ratio of $\frac{p}{n}$ (10, 8, 3, 1.6, 1.4, 1.2) under two error settings (which are $N(0, 0.2)$ and double exponential $(0, 1)$). When we implement the equations (3.1), (3.2) and (3.4), we consider η function to be an identity function and ω is 1, because neither the penalty nor the sparsity is needed.

From the first two rows of Table 2, we see that in a Normal error setting, the Least Square estimator is preferable and the relative efficiency of the Least Square estimator vs. the Least Deviation estimator is around $\frac{2}{\pi}$. Further, we can see that in the Double exponential error setting, the Least Square estimator performs worse. This result matches the classical inference. From the last two rows, we can see that the Least Squares estimator is preferable whenever error is Normal or Double Exponential. This result is foreseen by [20] and [29].

Remarkably, in Table 3, we discuss a high-dimensional and sparse case ($p > n$). We fix $\delta = 0.64$ and $p = 500$ and consider optimally tuned (cross-validation) P-LS and P-LAD. For the number of the non-zeros in true parameter, s, we choose a variety of options which range from low-sparsity, 25, to high sparsity, 300. From the first two rows of Table 3, we see that in a Normal error setting, P-LS estimator is no longer preferred in all settings. When the sparsity, s, is high and reaches n, P-LAD estimator is preferred, whereas when the sparsity, s, is low, P-LS estimator is preferred. However, from the last two rows, in the setting of the Laplace distribution, we see that P-LAD estimator is always preferred no matter of the size of s. This contradicts the findings of Table 2 and shows that model selection affects the choice of the optimal loss function.

TABLE 2
Relative efficiency of Square Loss estimator v.s. Absolute Loss estimator under various low dimensional setting

Relative Efficiency	Least Squares			Least Deviations		
	$p < n$, with fixed $p = 50$ and varying n					
	$\delta = 10$	$\delta = 8$	$\delta = 3$	$\delta = 10$	$\delta = 8$	$\delta = 3$
Normal	0.204	0.234	0.308	0.395	0.439	0.568
Laplace	2.362	2.376	3.119	1.415	1.792	1.578
	$p \approx n$, with fixed $p = 50$ and varying n					
	$\delta = 1.6$	$\delta = 1.4$	$\delta = 1.2$	$\delta = 1.6$	$\delta = 1.4$	$\delta = 1.2$
Normal	0.489	0. 643	1.102	0.946	0.962	1.192
Laplace	5.544	7.276	12.475	7.014	11.351	17.929

TABLE 3
Relative efficiency of penalized Square Loss estimator with $\lambda_{1,opt}$ v.s. penalized Absolute Loss estimator $\lambda_{2,opt}$ under various high dimensional and sparsity setting

Relative Efficiency	Least Squares			Least Deviations		
	$p > n$ and $s < n$, with fixed $p = 500$ and $\delta = 0.64$ and varying s/n					
	$\omega = 0.05$	$\omega = 0.1$	$\omega = 0.2$	$\omega = 0.05$	$\omega = 0.1$	$\omega = 0.2$
Normal	0.042	0.0839	0.139	0.0458	0.113	0.183
Laplace	0.0437	0.0914	0.192	0.0322	0.0745	0.177
	$p > n$ and $s \approx n$, with $p = 500$ and fixed $\delta = 0.64$					
	$\omega = 0.5$	$\omega = 0.55$	$\omega = 0.6$	$\omega = 0.5$	$\omega = 0.55$	$\omega = 0.6$
Normal	0.394	0.458	0.468	0.385	0.432	0.477
Laplace	0.522	0.531	0.584	0.207	0.245	0.289

7. Proofs

In this section we collect all the detailed proofs of the main and auxiliary results.

7.1. Preliminaries

The last term $\langle \partial_1 \eta(A^T G(z^{t-1}; b_{t-1}) + x^{t-1}; \theta) \rangle$ in step 1 of RAMP iteration (equation (2.3)) is a correction of the residual, called Onsager reaction term.

This term is generated from the theory of belief propagation in factor graphical models. Adding the Onsager reaction term in each iteration is the main difference from AMP iteration and soft thresholding iteration. The intuition of this term in each step is considering undersampling and sparsity simultaneously. The following Lemma 4 shows the relationship between the Onsager reaction term and in Donoho's [19] term the undersampling–sparsity.

Similarly to [20] and [29] we use min regularization to regularize the squared loss with the robust loss ρ. This introduces the family of regularizations of the robust loss ρ as follows: $\rho(b,z) = \min_{x\in\mathbb{R}} \left\{b\rho(x) + \frac{1}{2}(x-z)^2\right\}$. Moreover, the proximal operator $Prox(z,b)$ admits the subgradient characterization: if $Prox(z,b) = u$ then $z - u \in b\rho'(u)$.

Lemma 4. *Let $(z_*, b_*, \hat{x}_*)$ be a fixed point of the RAMP equations (2.3), (2.4) and (2.6) having $b_* > 0$. According to the definition of $\eta(x)$, the correction term $\langle \partial_1 \eta(A^T G(z;b) + x; \theta)\rangle$ evaluated at the fixed point $(z_*, b_*, \hat{x}_*)$ is equal to $\|\hat{x}_*\|_0/p$, i.e.,to ω.*

7.2. Proofs of the main results

Proof of Theorem 4. Let $I(F_W)$ be a well defined information matrix of the errors, W_i. If the distribution of the errors W_i is a convolution $D = F_W \circ N(0,\sigma^2)$, then,

$$\mathbb{E}_D \Psi' = \omega/\delta.$$

Let the score function for the location of D be denoted with L_D. Then, the information matrix of D can be represented as $I(D) = \mathbb{E}[L_D^2]$ and $\mathbb{E}_D\Psi' = \mathbb{E}_G \Psi L_D$. In turn, simple Cauchy-Swartz inequality provides

$$\tau_t^2 = \frac{\omega}{\delta}\mathbb{E}_G\Psi^2 \geq \frac{\omega}{\delta}\frac{|\mathbb{E}_G\Psi L_D|^2}{\mathbb{E}[L_D^2]} = \frac{\omega}{\delta}\frac{(\mathbb{E}_D\Psi')^2}{I(D)} = \frac{\omega}{\delta}\frac{1}{I(F_W \circ N(0,\sigma_t^2))}.$$

By Lemma 3.5 of [20], the lower bound can be further reduced to

$$\tau_t^2 \geq \frac{\omega}{\delta}\frac{1+\sigma_t^2 I(F_W)}{I(F_W)}.$$

The proof is finalized by obtaining a lower bound of σ_t.

$$\sigma_t^2 = \frac{1}{\delta}\mathbb{E}\left[\eta(X_0 + \tau_{t-1}Z, \theta) - X_0)\right]^2.$$

For $\theta = \alpha\tau_{t-1}$, Proposition 1.3 of [4] shows that σ_t^2 is a strictly concave function for $\alpha > \alpha_{\min} > 0$ and $X_0 \neq 0$ and an increasing function of τ^2. Hence, $\sigma_t^2 > \tau_{t-1}^2$ for small τ_{t-1}^2 and $\sigma_t^2 < \tau_{t-1}^2$ for large τ_{t-1}^2. Hence,

$$\tau_t^2 \geq \frac{\omega}{\delta}\frac{1+\tau_{t-1}^2 I(F_W)}{I(F_W)} \geq \frac{\omega}{\delta}\frac{1+\omega/\delta}{I(F_W)}.$$

Iterating previous equation k times, we obtain that for $t > k$

$$\tau_t^2 \geq \frac{\omega}{\delta} \frac{1 + \omega/\delta + (\omega/\delta)^2 + \cdots + (\omega/\delta)^k}{I(F_W)}.$$

When $k \to \infty$, $\tau_t^2 \to \tau^{*2}$, we obtain

$$\tau^{*2} \geq \frac{\omega/\delta}{1 - \omega/\delta} \frac{1}{I(F_W)} = \frac{s}{n-s} \frac{1}{I(F_W)}.$$

Part (*iii*). Utilizing the scale-invariance property of the soft-thresholding function η, we obtain that

$$\begin{aligned}
\nu_1(\tau) = \alpha^2 \mathbb{P}\left(\left|Z + \frac{X_0}{\tau}\right| \geq \alpha\right) - 2\alpha \mathbb{E}\left[Z \text{sign}(Z) \mathbb{1}\left\{\left|Z + \frac{X_0}{\tau}\right| \geq \alpha\right\}\right] \\
+ \mathbb{E}\left[Z^2 \mathbb{1}\left\{\left|Z + \frac{X_0}{\tau}\right| \geq \alpha\right\}\right], \\
\nu_2(\tau) = \mathbb{E}\left[X_0^2 \mathbb{1}\left\{\left|Z + \frac{X_0}{\tau}\right| \leq \alpha\right\}\right].
\end{aligned}$$

Let us first focus on the second component, i.e., $\nu_2(\tau)$. The derivative of $\nu_2(\tau)$ is

$$\frac{\partial \nu_2(\tau)}{\partial \tau} = \mathbb{E}_{X_0}\left[\frac{X_0^3}{\tau^2}\left(\phi(\alpha - \frac{X_0}{\tau}) - \phi(-\alpha - \frac{X_0}{\tau})\right)\right].$$

By observing that the last term on the RHS is non-negative for all $X_0 > 0$ and negative for all $X_0 < 0$, we conclude that $\nu_2(\tau)$ is an increasing function.

We conclude the proof with the analysis of the first term, $\nu_1(\tau)$. The displays above imply that the first and the last term of $\nu_1(\tau)$ together lead to $\mathbb{E}\left[(Z^2 + \alpha^2)\mathbb{1}\left\{\left|Z + \frac{X_0}{\tau}\right| \geq \alpha\right\}\right]$, whereas the middle term can be written as

$$2\alpha \mathbb{E}\left[Z \mathbb{1}\left\{Z \leq \alpha - \frac{X_0}{\tau}\right\}\right] + 2\alpha \mathbb{E}\left[Z \mathbb{1}\left\{Z \leq -\alpha - \frac{X_0}{\tau}\right\}\right].$$

By Stein's lemma we know that the previous expression is equal to

$$2\alpha \mathbb{E}_{X_0}\left[\phi(\alpha - \frac{X_0}{\tau}) - \phi(-\alpha - \frac{X_0}{\tau})\right].$$

Furthermore, utilizing the variance computation of a truncated random variable, conditional on X_0, it is easy to check that

$$\mathbb{E}\left[Z^2 \mathbb{1}\left\{Z + \frac{X_0}{\tau} \leq \alpha\right\}\right] = 1 - \left(\alpha - \frac{X_0}{\tau}\right)\phi\left(\alpha - \frac{X_0}{\tau}\right).$$

The rest of terms can be computed similarly. Combining all of the above we obtain

$$\nu_1(\tau) = \mathbb{E}_{X_0}\left[\alpha^2 + 1 - \alpha^2\left[\Phi(\alpha - \frac{X_0}{\tau}) - \Phi(-\alpha - \frac{X_0}{\tau})\right]\right.$$

$$-\alpha\left[\phi(\alpha-\frac{X_0}{\tau})+\phi(-\alpha-\frac{X_0}{\tau})\right]-\frac{X_0}{\tau}\left[\phi(\alpha-\frac{X_0}{\tau})-\phi(-\alpha-\frac{X_0}{\tau})\right]\Big].$$

Evaluating the derivative of $\nu_1(\tau)$, we obtain

$$\frac{\partial\nu_1(\tau)}{\partial\tau}=\mathbb{E}_{X_0}\left[\frac{X_0}{\tau^2}\left(1-\frac{X_0^2}{\tau^2}\right)\left(\phi(\alpha-\frac{X_0}{\tau})-\phi(-\alpha-\frac{X_0}{\tau})\right)\right].$$

Hence, for small τ^2 the expression above is negative and for large values of τ^2 it is positive. It follows that, $\nu_1(\tau)$ is a convex function of τ^2. □

Proof of Theorem 5. Notice that in sparse, high dimensional setting, the distribution of the X_0 can be represented as a convex combination of the Dirac measure at 0 and a measure that doesn't have mass at zero. Let us denote with Δ and U two random variables, each having the two measures above. Let

$$\Psi_\alpha(\tau)=\frac{1-\omega}{\delta}\mathbb{E}\eta^2(Z,\alpha)+\frac{\omega}{\delta}\mathbb{E}\left[\eta\left(\frac{U}{\tau}+Z;\alpha\right)-\frac{U}{\tau}\right]^2.$$

First, we prove that whenever $\sigma_W^2\to 0$ then $\tau^2_{\text{P-LAD}}\to 0$ as long as $\lim_{\tau\to 0}\Psi_\alpha(\tau)\neq 0$. To accomplish this, let's prove that $\lim_{\tau\to 0}\Psi_\alpha(\tau)\neq 0$ and look at the relationship between $\tau_{\text{P-LAD}}$ and σ_W.

Notice that by the result of Theorem 4 [46], we conclude

$$\lim_{\tau\to 0}\Psi_\alpha(\tau)=\frac{\omega}{\delta},$$

which is different from 0 whenever $s\neq 0$.

Observe that whenever $\sigma_W^2\to 0$, it holds that $Y\to\sigma^2_{\text{P-LAD}}Z$ and $f(W;\tau^2_{\text{P-LAD}})\to 0$. In this case

$$\tau^2_{\text{P-LAD}}\left(1-\Psi_\alpha^{-1}\tilde{g}(\tau^2_{\text{P-LAD}})\right)=b^2\frac{\mathbb{P}\left(\left|\tau^2_{\text{P-LAD}}\Psi_\alpha Z\right|>b\right)}{\mathbb{P}^2\left(|\tau^2_{\text{P-LAD}}\Psi_\alpha Z|\leq b\right)},\tag{7.1}$$

where

$$\begin{aligned}&\tilde{g}(\tau^2_{\text{P-LAD}})\mathbb{P}^2\left(\left|\tau^2_{\text{P-LAD}}\Psi_\alpha Z\right|\leq b\right)\\&=\mathbb{E}_Z\left[Z^2\left(F_W(b-\tau_{\text{P-LAD}}\Psi_\alpha^{1/2}Z)-F_W(-b-\tau_{\text{P-LAD}}\Psi_\alpha^{1/2}Z)\right)\right].\end{aligned}$$

Hence, $\tilde{g}(0)=\mathbb{E}_Z\left[Z^2\left(F_W(b)-F_W(-b)\right)\right]=F_W(b)-F_W(-b)<\infty$. In turn, by plugging in $\tau_{\text{P-LAD}}=0$ it satisfies both sides of the equation (7.1). □

Proof of Theorem 6. Notice that in sparse, high dimensional setting, the distribution of the X_0 can be represented as a convex combination of the Dirac measure at 0 and a measure that doesn't have mass at zero. Let us denote with Δ and U two random variables, each having the two measures above. Let

$$\Psi_\alpha=\Psi_\alpha(\tau_{\text{P-LAD}})=\frac{1-\omega}{\delta}\mathbb{E}\eta^2(Z,\alpha)+\frac{\omega}{\delta}\mathbb{E}\left[\eta\left(\frac{U}{\tau_{\text{P-LAD}}}+Z;\alpha\right)-\frac{U}{\tau_{\text{P-LAD}}}\right]^2.$$

We first discuss the P-LAD estimator. By the state-evolution recursion, (3.2)

$$\tau^2_{\text{P-LAD}}\Psi_\alpha = \sigma^2_{\text{P-LAD}}. \tag{7.2}$$

Let $Y = W + \sigma^2_{\text{P-LAD}}Z$. According to (4.2),

$$\tau^2_{\text{P-LAD}} = \frac{\mathbb{E}[Y^2\mathbb{1}|Y| \le b] + b^2\mathbb{P}(|Y| > b)}{\mathbb{P}^2(|Y| \le b)}. \tag{7.3}$$

Next observe that $\mathbb{E}[Y^2\mathbb{1}|Y| \le b] = \mathbb{E}[W^2\mathbb{1}|Y| \le b] + \sigma^2_{\text{P-LAD}}\mathbb{E}[Z\mathbb{1}|Y| \le b]$; moreover, $\mathbb{E}[Y^2\mathbb{1}|Y| \le b] = \sigma^2_W - \mathbb{E}[W^2\mathbb{1}|Y| > b] + \sigma^2_{\text{P-LAD}}\mathbb{E}[Z\mathbb{1}|Y| \le b]$. Plugging into (7.3) we obtain

$$\tau^2_{\text{P-LAD}} = \sigma^2_{\text{P-LAD}}\frac{\mathbb{E}[Z\mathbb{1}|Y| \le b]}{\mathbb{P}^2(|Y| \le b)} + \frac{\sigma^2_W - \mathbb{E}[W^2\mathbb{1}|Y| > b]}{\mathbb{P}^2(|Y| \le b)} + \xi(b) \tag{7.4}$$

for

$$\xi(b) = b^2\frac{\mathbb{P}\left(\left|W + \sigma^2_{\text{P-LAD}}Z\right| > b\right)}{\mathbb{P}^2\left(|W + \sigma^2_{\text{P-LAD}}Z| \le b\right)}.$$

Let

$$g(\tau^2_{\text{P-LAD}}) = \frac{\mathbb{E}[Z\mathbb{1}|Y| \le b]}{\mathbb{P}^2(|Y| \le b)}, \qquad f(W;\tau^2_{\text{P-LAD}}) = \frac{\sigma^2_W - \mathbb{E}[W^2\mathbb{1}|Y| > b]}{\mathbb{P}^2(|Y| \le b)},$$

then

$$\tau^2_{\text{P-LAD}} = \sigma^2_{\text{P-LAD}}g(\tau^2_{\text{P-LAD}}) + f(W;\tau^2_{\text{P-LAD}}) + \xi(b). \tag{7.5}$$

Substituting (7.1) in (7.2) we obtain

$$\frac{\sigma^2_{\text{P-LAD}}}{\sigma^2_W} = \frac{\Psi_\alpha}{1 - g(\tau^2_{\text{P-LAD}})\Psi_\alpha}\left[\frac{f(W;\tau^2_{\text{P-LAD}})}{\sigma^2_W} + \frac{\xi(b)}{\sigma^2_W}\right]. \tag{7.6}$$

By Stein's lemma and some algebra we arrive at the representation of $g(\tau^2_{\text{P-LAD}})$ and $f(W;\tau^2_{\text{P-LAD}})$, as

$$g(\tau^2_{\text{P-LAD}}) = \mathbb{E}_Z\left[Z^2\left(F_W(b - \sigma_{\text{P-LAD}}Z) - F_W(-b - \sigma_{\text{P-LAD}}Z)\right)\right]/\mathbb{P}^2(|Y| \le b),$$
$$f(W;\tau^2_{\text{P-LAD}}) = \mathbb{E}_W\left[W^2\left(\Phi\left(\frac{b - W}{\sigma_{\text{P-LAD}}}\right) - \Phi\left(\frac{-b - W}{\sigma_{\text{P-LAD}}}\right)\right)\right]/\mathbb{P}^2(|Y| \le b).$$

Let us first focus on the case of $\sigma^2_W \to 0$. By Lemma 5 we conclude that $\tau^2_{\text{P-LAD}} \to 0$ and $\sigma^2_{\text{P-LAD}} \to 0$. Hence,

$$\lim_{\sigma^2_W \to 0}\frac{\sigma^2_{\text{P-LAD}}}{\sigma^2_W} = \lim_{\tau \to 0, \sigma^2_W \to 0}\frac{\Psi_\alpha(\tau)}{1 - g(\tau)\Psi_\alpha(\tau)}\left[\frac{f(W;\tau)}{\sigma^2_W} + \frac{\xi(b)}{\sigma^2_W}\right].$$

We proceed to show that the last term in the display above is converging to ∞. Observe that whenever $\sigma^2_W \to 0$, it holds that $Y \to \sigma^2_{\text{P-LAD}}Z$ and

$$\xi(b) \to b^2\frac{\mathbb{P}\left(\left|\tau^2\Psi_\alpha(\tau)Z\right| > b\right)}{\mathbb{P}^2\left(|\tau^2\Psi_\alpha(\tau)Z| \le b\right)}.$$

Furthermore, with $\sigma_{\text{P-LAD}} \to 0$ and $b > 0$, it holds that $\xi(b) \to 0$. For ϕ denoting the density of the standard normal, the application of L'Hõpital's rules guarantees

$$\lim_{\sigma\to 0,\sigma_W^2\to 0} \frac{\xi(b)}{\sigma_W^2} = b^2 \lim_{\sigma\to 0,\sigma_W^2\to 0} \frac{\phi(b/\sigma)\sigma^{-2} + \phi(-b/\sigma)\sigma^{-2}}{4\sigma_W},$$

which implies $\frac{\xi(b)}{\sigma_W^2} \to \infty$ as $\sigma_W \to 0$.

We finish the proof by discussing the P-LS estimator. By Lemma 1 we see that the special case of the RAMP algorithm, when the loss function $\rho(x) = (x)^2$ is the approximate message passing algorithm of [4]. Hence, results that apply to the algorithm in [4] apply. In particular, a recent work [46] discusses the properties of $\lim_{\sigma_W^2\to 0} \frac{\bar{\sigma}^2_{\text{P-LS}}}{\sigma_W^2}$ in their Theorem 7. □

Proof of Theorem 7. We will use the notation defined in the proof of Lemma 6. We first discuss the Penalized LAD estimator. Based on the representation proved in Lemma 6

$$\lim_{\sigma_W^2\to\infty} \frac{\sigma^2_{\text{P-LAD}}}{\sigma_W^2} = \lim_{\tau\to\infty,\sigma_W^2\to\infty} \frac{\Psi_\alpha(\tau)}{1 - g(\tau)\Psi_\alpha(\tau)} \left[\frac{f(W;\tau)}{\sigma_W^2} + \frac{\xi(b)}{\sigma_W^2}\right].$$

It suffices to discuss the limiting properties of the first, second and the third term in the right hand side above. Let us discuss the last term first. Observe that we can rewrite

$$\lim_{\tau\to\infty,\sigma_W^2\to\infty} \frac{\xi(b)}{\sigma_W^2} = \lim_{\sigma\to\infty} \frac{\xi(b)}{\sigma^2} = \lim_{\sigma\to\infty} \frac{b^2 \frac{\mathbb{P}\left(\left|W+\sigma^2_{\text{P-LAD}}Z\right|>b\right)}{\mathbb{P}^2\left(\left|W+\sigma^2_{\text{P-LAD}}Z\right|\le b\right)}}{\frac{\mathbb{E}\Phi^2(W+\sigma^2_{\text{P-LAD}}Z,b)}{\mathbb{P}^2\left(\left|W+\sigma^2_{\text{P-LAD}}Z\right|\le b\right)}}$$
$$= b^2 \lim_{\sigma\to\infty} \frac{b^2\mathbb{P}\left(\left|W+\sigma^2_{\text{P-LAD}}Z\right|>b\right)}{\mathbb{E}\Phi^2(W+\sigma^2_{\text{P-LAD}}Z,b)} = 1,$$

where in the last step we used the fact that when $\tau \to \infty$, $W + \sigma^2_{\text{P-LAD}}Z \to \infty$

$$\mathbb{E}\Phi^2(\infty, b) = b^2 \lim_{\sigma\to\infty} \mathbb{1}\{W + \sigma^2 Z \ge b\} = b^2.$$

Next, we discuss the limit of $\Psi_\alpha(\tau)$. Corollary 6 of [46] guarantees that $\lim_{\tau\to\infty} \Psi_\alpha(\tau) = \mathbb{E}\eta^2(Z;\alpha)/\delta$, that is, $\Psi_\alpha(\infty) = \Gamma/\delta$.

In the following, we analyze the limit of

$$g(\tau) = \frac{\mathbb{E}_Z\left[Z^2 F_W(b-\sigma Z) - Z^2 F_W(-b-\sigma Z)\right]}{\mathbb{P}^2\left(|W+\sigma Z| \le b\right)}$$

as $\tau \to \infty$. In view of the fact that, both the numerator and denominator of $g(\tau)$ converge to 0 when $\tau \to \infty$, we use the L'Hõpital's rule in determining its limit. Therefore,

$$\lim_{\tau\to\infty} g(\tau) = \lim_{\tau\to\infty} \frac{\mathbb{E}_Z\left[-Z^3 f_W(b-\sigma Z) + Z^3 f_W(-b-\sigma Z)\right]}{2\mathbb{P}\left(W+\sigma Z \le b\right)\left(F_W(b-\sigma Z) + 1 - F_W(-b-\sigma Z)\right)}.$$

Moreover, the last expression still needs L'Hõpital's rule. Hence,

$$\lim_{\tau\to\infty} g(\tau) = \lim_{\tau\to\infty} \frac{\mathbb{E}_Z\left[Z^4 f'_W(b-\sigma Z) - Z^4 f'_W(-b-\sigma Z)\right]}{2(F_W(b-\sigma Z)+1-F_W(-b-\sigma Z))} = 0.$$

The proof is finalized by the analysis of $f(W;\tau)/\sigma_W^2$, when $\sigma_W^2 \to \infty$ and $\tau \to \infty$. We begin with the following representation of $f(W;\tau)$,

$$f(W;\tau) = \frac{\mathbb{E}_W\left[W^2\Phi_Z(\frac{b-W}{\sigma}) + W^2 - W^2\Phi_Z(\frac{-b-W}{\sigma})\right]}{\mathbb{P}^2\left(|W+\sigma Z| \le b\right)}.$$

We observe that in the limit when $\tau \to \infty$, of the above expression takes the form $0/0$; hence, we apply the L'Hõpital's rule to obtain

$$\begin{aligned}
&\lim_{\sigma\to\infty,\sigma_W^2\to\infty} \frac{f(W;\tau)}{\sigma_W^2} \\
&= \lim_{\sigma\to\infty,\sigma_W^2\to\infty} \\
&\frac{\left(\mathbb{E}_W\left[-W^2\phi_Z(\frac{b-W}{\sigma})(b-W)/\sigma^2 - W^2\phi_Z(\frac{-b-W}{\sigma})(b+W)/\sigma^2\right]\right)/2\sigma_W}{2\mathbb{P}\left(|W+\sigma Z|\le b\right)\mathbb{E}_W\left[-\phi_Z(\frac{b-W}{\sigma})(b-W)/\sigma^2 - \phi_Z(\frac{-b-W}{\sigma})(b+W)/\sigma^2\right]} \\
&= \lim_{\sigma\to\infty,\sigma_W^2\to\infty} \frac{1}{4}\frac{\mathbb{E}_W\left[W^2\phi_Z(\frac{b-W}{\sigma})(b-W)^2 - W^2\phi_Z(\frac{-b-W}{\sigma})(b+W)^2\right]/\sigma}{\mathbb{E}_W^2\left[-\phi_Z(\frac{b-W}{\sigma})(b-W) - \phi_Z(\frac{-b-W}{\sigma})(b+W)\right]} \\
&+ o(1) \\
= &\lim_{\sigma\to\infty,\sigma_W^2\to\infty} \frac{1}{64b^2}\frac{\mathbb{E}_W\left[W^2\phi_Z(\frac{b-W}{\sigma})(b-W)^2 - W^2\phi_Z(\frac{-b-W}{\sigma})(b+W)^2\right]/\sigma}{\mathbb{E}_Z^2\left[-\sigma Z f_W(\sigma Z)\right]} \\
&+ o(1)
\end{aligned}$$

where in the last step we used the change of variables to go from $\mathbb{E}_W$ to $\mathbb{E}_Z$. The last expression converges to zero as both $\sigma \to \infty, \sigma_W^2 \to \infty$. □

Lemma 5. *Let $(z_*, b_*, \hat{x}_*)$ be a fixed point of the RAMP equations (2.3), (2.4) and (2.6), having $b_* > 0$. Then, $\hat{x}_*$ is a solution to the penalized M-estimator problem (1.2) with $\lambda = \frac{\theta_*\omega}{b_*\delta}$. Vice versa, any minimizer $\hat{x}(\lambda)$ of the problem (1.2) corresponds to one (or more) RAMP fixed points of the form $\left(z_*, \frac{\theta_*\omega}{\lambda\delta}, \hat{x}_*\right)$.*

7.3. Proofs for examples

7.3.1. Equation (2.7)

According to (2.2), we observe that proximal mapping operator satisfies $b\rho'(Prox(z,b)) + Prox(z,b) - z \in 0$. We consider $Prox(z,b) \neq 0$ first. We observe that $Prox(z,b) < 0$, when $z < -b$ and $Prox(z,b) > 0$ when $z > b$. This indicates that $\text{sign}(Prox(z,b)) = \text{sign}(z)$. Substituting it in the previous equation, we get $Prox(z,b) = z - b\,\text{sign}(z)$. Next, we observe that when $Prox(z,b) = 0$ we have $\partial(b|x|)/\partial x = b\xi$, where $\xi \in (-1,1)$. Substituting it in the proximal mapping equation, we get $z \in (-b,b)$. Above all,

7.3.2. Equation (2.8)

The family of min regularized loss function is then defined as follows

$$\rho(z,b,\tau) \equiv \min_{x\in\mathbb{R}} \left\{ b\rho_\tau(x) + \frac{1}{2}(x-z)^2 \right\}.$$

Similarly, as before, $b\rho'(Prox(z,b)) + Prox(z,b) - z \in 0$. Now, we first consider $Prox(z,b) \neq 0$, in which case we obtain

$$\rho'(Prox(z,b)) = \text{sign}\,(Prox(z,b))\Big((1-\tau)1\{Prox(z,b)<0\} + \tau 1\{Prox(z,b)>0\}\Big).$$

Next, we observe that when $Prox(z,b) = 0$ we have $\partial(\rho_\tau(x))/\partial x = b\xi((1-\tau)1\{x<0\} + \tau 1\{x>0\})$, where $\xi \in (-1,1)$. Analyzing the positive and negative parts separately, we see that $\partial(\rho_\tau(x))/\partial x = b\tau\xi$ and $\partial(\rho_\tau(x))/\partial x = b(1-\tau)\xi$, respectively.

7.4. Proofs of preliminary statements

Proof of Lemma 4. Let (x,z) be a fixed point of the RAMP algorithm iteration. Then the fixed point conditions at x read as

$$x = \eta(A^TG(z;b)+x;\theta) = \begin{cases} x + A^TG(z,b) - \theta, & \text{if } x + A^TG(z,b) > \theta \\ 0, & \text{if } -\theta \leq x + A^TG(z,b) \leq \theta \\ x + A^TG(z,b) + \theta, & \text{if } x + A^TG(z,b) < -\theta \end{cases}.$$

They imply that for all $x + A^TG(z,b) > \theta$, $x = x + A^TG(z,b) - \theta$, or in other terms that $A^TG(z,b) = \theta$. Similarly, $x + A^TG(z,b) < \theta$, $x = x + A^TG(z,b) + \theta$, or using different terms, that $A^TG(z,b) = -\theta$. For the middle term, we observe that $x = 0$, if and only if $-\theta < A^TG(z,b) < \theta$. Hence,

$$A^TG(z,b) = \theta v, \tag{7.7}$$

where $v \in \mathbb{R}^p$ with each element $v_i = \begin{cases} \text{sign}(x_i) & \text{if } x_i \neq 0 \\ (-1,1) & \text{if } x_i = 0 \end{cases}$. Therefore, the correction term defined as the average of the first derivative of $\eta(A^TG(z,b) + x;\theta)$, becomes:

$$\langle \partial_1 \eta(A^TG(z;b)+x;\theta)\rangle = \langle \mathbb{1}\{|A^TG(z,b)| \ \neq \theta\}\rangle = \langle \mathbb{1}\{x \neq 0\}\rangle = \frac{||x||_0}{p} = \omega.$$

□

Proof of Lemma 5. The fixed point condition at z reads

$$z = Y - Ax + \frac{1}{\delta}G(z;b)\langle \partial_1 \eta(A^TG(z;b)+x;\theta)\rangle.$$

Moreover, from Lemma 4 we conclude $\langle \partial_1 \eta(A^T G(z;b) + x; \theta)\rangle = \omega$, and hence $z = Y - Ax + \frac{1}{\delta}\omega G(z;b)$. By definition of the rescaled effective score G, we conclude $z = Y - Ax + \Phi(z;b)$, which shows that $Y - Ax = z - \Phi(z;b)$. Then, we have that the left hand side of the KKT condition becomes

$$A^T \rho'(Y - Ax) = A^T \rho'(z - \Phi(z;b)) \overset{(i)}{=} A^T \rho'(Prox(z,b)) \overset{(ii)}{=} A^T \Phi(z,b)/b \overset{(iii)}{=} \frac{\theta v \omega}{\delta b}. \tag{7.8}$$

The equations (i) and (ii) are derived from the definition of $\Phi(z;b)$, equation (iii) is based upon the proof of Lemma 4, equation (7.7). Hence,

$$A^T G(z,b) = A^T \Phi(z,b)\delta/\omega = \theta v$$

so that $A^T \Phi(z,b)/b = \frac{\theta v \omega}{\delta b}$. Plugging $\theta = \frac{\lambda b \delta}{\omega}$ into equation (7.8), we have $A^T \rho'(Y - Ax) = \lambda v$. □

7.5. Proofs of section 3.1

Proof of Lemma 1. This is an immediate application of state evolution as defined in [3], which considers general recursions. Hence, it suffices to show that the proposed algorithm is a special case of it. In the original notation of [3], the generalized recursions studied are

$$b^t = Aq^t - \lambda_t m^{t-1} \tag{7.9}$$

$$h^{t+1} = A^T m^t - \xi_t q^t \tag{7.10}$$

where

$$q^t = f_t(h^t), \qquad m^t = g^t(b^t, w). \tag{7.11}$$

The two scalars ξ_t and λ_t are defined as

$$\xi_t = \langle g_t'(b^t, w)\rangle \tag{7.12}$$

and

$$\lambda_t = \frac{1}{\delta}\langle f_t'(h^t)\rangle, \tag{7.13}$$

where$\langle\cdot\rangle$ denotes an empirical mean over the entries in a vector and derivatives are with respect to the first argument. According to [3], the state evolution recursion involves two variables: $\bar{\tau}_t^2 = \mathbb{E} g_t^2(\bar{\sigma}_t Z, W)$ and $\bar{\sigma}_t = \frac{1}{\delta}\mathbb{E} f_t^2(\bar{\tau}_{t-1} Z)$. To see that the RAMP algorithm in (2.6), (2.4) and (2.3) is a special case of this recursion, we specify the above components of the general recursion to be

$$h^{t+1} = x_0 - A^T G(z^t, b_t) - x^t \tag{7.14}$$

$$q^t = x^t - x_0 \tag{7.15}$$

$$z^t = w - b^t \tag{7.16}$$

$$m^t = -G(z^t, b_t) \tag{7.17}$$
$$g_t(s, w) = -G(w - s, b_t) \tag{7.18}$$
$$f_t(s) = \eta(x_0 - s; \theta) - x_0, \tag{7.19}$$

with the initial condition being $q^0 = x_0$. Now, we verify that the simplification of the above series of equations (7.14)-(7.19) offer the RAMP algorithm iterations. We discuss the first step of the algorithm and then the third, whereas we leave the discussion of the second step as the last. We observe

$$\begin{aligned} x^t &\overset{(7.15)}{=} q^t + x_0 \overset{(7.11)}{=} f_t(h^t) + x_0 \overset{(7.19)}{=} \eta(x_0 - h^t; \theta) - x_0 + x_0 \\ &\overset{(7.14)}{=} \eta(x_0 - x_0 + (A^T G(z^{t-1}, b_{t-1})) + x^{t-1}; \theta) \\ &= \eta(x^{t-1} + A^T G(z^{t-1}, b_{t-1}); \theta), \end{aligned}$$

which is the first step of our algorithm. Also,

$$\begin{aligned} z^t &\overset{(7.16)}{=} w - b^t \overset{(7.9)}{=} w - Aq^t + \lambda_t m^{t-1} \overset{(7.15)}{=} w - A(x^t - x_0) + \lambda_t m^{t-1} \\ &\overset{(1.1)}{=} Y - Ax_0 - A(x^t - x_0) + \lambda_t m^{t-1} = Y - Ax^t + \lambda_t m^{t-1} \\ &\overset{(7.17)(7.13)}{=} Y - Ax^t + \frac{1}{\delta}\langle f_t'(h^t)\rangle(-G(z^t, b_t)) \\ &= Y - Ax^t + \frac{1}{\delta} G(z^t, b_t)\langle -\eta'(x_0 - h^t; \theta)\rangle, \ \text{(Since } \langle f_t'(h^t)\rangle = \langle -\eta'(x_0 - h^t; \theta)\rangle) \end{aligned}$$

which is the third step of our algorithm. Further, we need to show that h^{t+1} in the above special recursion satisfies the equation of h^{t+1} in general AMP, which means we need $h^{t+1} = A^T m^t - \xi_t q^t = x_0 - (A^T G(z^t, b_t)) - x^t$. Therefore,

$$\begin{aligned} h^{t+1} = A^T m^t - \xi_t q^t &\overset{(7.17)}{=} -A^T G(z^t, b_t) - \xi_t q^t \\ &\overset{(7.15)}{=} -A^T G(z^t, b_t) - \xi_t(x^t - x_0) = x_0 - (A^T G(z^t, b_t)) - x^t. \end{aligned}$$

This equation is only true when $\xi_t = 1$. Moreover, by the definition of G, we conclude that

$$\xi_t \overset{(7.12)}{=} \langle G'(z^t, b_t)\rangle = \frac{\delta}{\omega}\langle \Phi'(z^t, b_t)\rangle = 1$$

Therefore, we showed that the RAMP algorithm is a special case of general recursion, and we can conclude that the Theorem 2 of [3] applies and provides

$$\begin{aligned} \bar{\sigma}_t^2 &\overset{(3.2)}{=} \frac{1}{\delta}\mathbb{E}(\eta(X_0 - \bar{\tau}_{t-1} Z, \theta) - X_0))^2 \\ \bar{\tau}_t^2 &\overset{(3.1)}{=} \mathbb{E}(G(W - \bar{\sigma}_t Z); b_t)^2. \end{aligned}$$

The proof is then completed by a simple observation that Z and $-Z$ have the same distribution. □

Proof of Lemma 2. The statement of the lemma follows if we successfully show that (a) the total first derivative of $\mathbb{V}(\tau^2, b(\tau), \alpha\tau)$ is strictly positive for τ^2 large enough; (b) the function $\mathbb{V}$ is concave for all smooth loss functions ρ and not for non-smooth loss functions ρ; and (c) the $\lim_{\tau\to\infty} \mathbb{V}'(\tau^2, b(\tau), \alpha\tau^2)$ is a strictly decreasing function of α.

Part (a).

According to the definition of Φ and Condition (**R**), we can represent $\frac{\partial \mathbb{V}(\tau^2,b,\theta)}{\partial(\tau^2,\theta)}$ as

$$\begin{aligned}
&\frac{\delta^2}{\omega^2}\frac{\partial}{\partial(\tau^2,\theta)}\mathbb{E}[\Phi(W+\sigma Z;b)]\\
&=\frac{\delta^2}{\omega^2}\frac{\partial}{\partial(\tau^2,\theta)}\left[b\mathbb{E}\upsilon_1(W+\sigma Z)+b\sum_{\nu=1}^{k}\alpha_\nu\mathbb{P}\left\{r_\nu\le Prox(W+\sigma Z,b)\le r_{\nu+1}\right\}\right]\\
&=\frac{\delta^2}{\omega^2}\left[b\mathbb{E}\left[\upsilon_1'(W+\sigma Z)\frac{\partial\sigma}{\partial(\tau^2,\theta)}Z\right]\right.\\
&\qquad\left.+b\sum_{\nu=1}^{k}\alpha_\nu\mathbb{E}_{X_0,Z}\left(f(\bar r_{\nu+1})\frac{\partial\bar r_{\nu+1}}{\partial(\tau^2,\theta)}-f(\bar r_\nu)\frac{\partial\bar r_\nu}{\partial(\tau^2,\theta)}\right)\right] \qquad (7.20)
\end{aligned}$$

where $Prox'$ is derivative of the $Prox$ function with respect to its first argument, f is the density of W and $\bar r_{\nu+1}$ is such that

$$\bar r_{\nu+1}-\Phi(\bar r_{\nu+1}+\sigma Z;b)=r_{\nu+1}-\sigma Z.$$

By integrating the implicit relation above we obtain

$$\frac{\partial\bar r_{\nu+1}}{\partial(\tau^2,\theta)}=-Z\frac{\partial\sigma}{\partial(\tau^2,\theta)}\frac{\partial\bar Prox(\bar r_{\nu+1}+\sigma Z;b)}{\partial(\tau^2,\theta)}. \qquad (7.21)$$

Observe that $Prox$ is a strongly convex function with bounded level sets. [4] derive σ to be concave and for large τ^2 strictly increasing. Hence, $\bar r_{\nu+1}+\sigma Z$ can be made large and positive for large values of τ^2. In turn, $\mathbb{E}\frac{\partial\bar r_{\nu+1}}{\partial(\tau^2,\theta)}$ can be made strictly positive for large values of τ^2. Together with Condition (**D**) and convexity of ρ we are ready to conclude that $\frac{\partial}{\partial(\tau^2,\theta)}\mathbb{E}[\Phi(W+\sigma Z;b)]$ is strictly positive for large τ^2.

By changing the order of differentiation and expectation (allowed by boundedness of functions considered given by Condition (**R**)), we obtain that b is defined as a solution to the equation $\partial_1\mathbb{E}[\Phi(W+\sigma Z;b)]=\frac{\omega}{\delta}$ (see Lemma 3 for details). More specifically,

$$\begin{aligned}
&b\mathbb{E}\left[\upsilon_1(W+\sigma Z)\frac{\partial\sigma(\tau^2,\theta)}{\partial\tau^2}Z\right]\\
&\qquad+b\sum_{\nu=1}^{k}\alpha_\nu\mathbb{E}_{X_0,Z}\left(f(\bar r_{\nu+1})\frac{\partial\bar r_{\nu+1}}{\partial\tau^2}-f(\bar r_\nu)\frac{\partial\bar r_\nu}{\partial\tau^2}\right)=\frac{\omega}{\delta}. \qquad (7.22)
\end{aligned}$$

Moreover, as before, in Equation (7.20)

$$\begin{aligned}&\frac{\delta^2}{\omega^2}\frac{\partial}{\partial b}\mathbb{E}[\Phi(W+\sigma Z;b)]\\&=\frac{\delta^2}{\omega^2}\left[\mathbb{E}\left[\upsilon_1(W+\sigma Z)\right]+\sum_{\nu=1}^{k}\alpha_\nu\mathbb{E}_{X_0,Z}\left(f(\bar{r}_{\nu+1})\frac{\partial\bar{r}_{\nu+1}}{\partial\tau^2}-f(\bar{r}_\nu)\frac{\partial\bar{r}_\nu}{\partial\tau^2}\right)\right]\frac{\partial b(\tau^2)}{\partial\tau^2}.\end{aligned}\tag{7.23}$$

We focus on the last part of the above display. The total derivative of (7.22) provides the implicit equation for $\frac{\partial b}{\partial\tau^2}$,

$$\begin{aligned}&b\mathbb{E}\left[\upsilon_1'(W+\sigma Z)\frac{\partial^2\sigma(\tau^2,\theta)}{\partial(\tau^2)^2}Z\right]\\&+b\sum_{\nu=1}^{k}\alpha_\nu\mathbb{E}_{X_0,Z}\left(f'(\bar{r}_{\nu+1})\frac{\partial\bar{r}_{\nu+1}}{\partial\tau^2}-f'(\bar{r}_\nu)\frac{\partial\bar{r}_\nu}{\partial\tau^2}+f(\bar{r}_{\nu+1})\frac{\partial^2\bar{r}_{\nu+1}}{\partial(\tau^2)^2}-f(\bar{r}_\nu)\frac{\partial^2\bar{r}_\nu}{\partial(\tau^2)^2}\right)\\&+\left[\mathbb{E}\left[\upsilon_1(W+\sigma Z)\frac{\partial\sigma(\tau^2,\theta)}{\partial\tau^2}Z\right]\right.\\&\qquad\left.+\sum_{\nu=1}^{k}\alpha_\nu\mathbb{E}_{X_0,Z}\left(f(\bar{r}_{\nu+1})\frac{\partial\bar{r}_{\nu+1}}{\partial\tau^2}-f(\bar{r}_\nu)\frac{\partial\bar{r}_\nu}{\partial\tau^2}\right)\right]\frac{\partial b(\tau^2)}{\partial\tau^2}=0.\end{aligned}\tag{7.24}$$

Next, we observe that υ_1 can be made positive for large τ^2 and that υ_1' is positive. By (7.21) we can see that $\frac{\partial\bar{r}_\nu}{\partial\tau^2}>0$, $\frac{\partial^2\bar{r}_\nu}{\partial(\tau^2)^2}\geq 0$, and $\frac{\partial^2\bar{r}_\nu}{\partial(\tau^2)^2}<\frac{\partial^2\bar{r}_{\nu+1}}{\partial(\tau^2)^2}$ (curvature of a convex function decays away from the origin). Moreover, [4] prove that σ is strictly concave for $\alpha>0$. All of the above implies that $\frac{\partial\sigma(\tau^2,\theta)}{\partial\tau^2}>0$ for large τ^2; $\frac{\partial^2\sigma(\tau^2,\theta)}{\partial(\tau^2)^2}\geq 0$. Condition (**D**) guarantees $f>0$. Hence, $\frac{\partial b}{\partial\tau^2}$ can be made positive for large τ^2. Next, it suffices to observe that the total derivative of $\mathbb{V}(\tau^2,b,\theta)$ is given by the sum of the above marginal derivatives, all of which can be made positive.

Part(b).

Careful inspection of the second derivative of $\mathbb{V}(\tau^2,b,\theta)$ provides details (by the same arguments above) that the second derivative is negative, i.e., that the function $\mathbb{V}$ is concave for all smooth ρ and not necessarily negative for all non-smooth loss functions ρ. We show the analysis for one of the marginals as the analysis for the rest is done equivalently.

$$\begin{aligned}&\frac{\omega^2}{\delta^2}\frac{\partial^2\mathbb{V}(\tau^2,b,\theta)}{\partial(\tau^2)^2}\\&=\frac{\partial}{\partial\tau^2}\left[b\mathbb{E}\left[\upsilon_1'(W+\sigma Z)\frac{\partial\sigma}{\partial\tau^2}Z\right]+b\sum_{\nu=1}^{k}\alpha_\nu\mathbb{E}_{X_0,Z}\left(f(\bar{r}_{\nu+1})\frac{\partial\bar{r}_{\nu+1}}{\partial\tau^2}-f(\bar{r}_\nu)\frac{\partial\bar{r}_\nu}{\partial\tau^2}\right)\right]\\&=\underbrace{b\mathbb{E}\left[\upsilon_1''(W+\sigma Z;b)Z^2\left(\frac{\partial\sigma(\tau^2,\theta)}{\partial\tau^2}\right)^2+\upsilon_1'(W+\sigma Z)Z\frac{\partial^2\sigma(\tau^2,\theta)}{\partial(\tau^2)^2}\right]}_{T_1}\end{aligned}$$

$$\underbrace{\begin{array}{l} +b\sum_{\nu=1}^{k}\alpha_\nu \mathbb{E}_{X_0,Z}\left(f'(\bar{r}_{\nu+1})\left(\frac{\partial \bar{r}_{\nu+1}}{\partial \tau^2}\right)^2\right. \\ \left. -f'(\bar{r}_\nu)\left(\frac{\partial \bar{r}_\nu}{\partial \tau^2}\right)^2 + f(\bar{r}_{\nu+1})\frac{\partial^2 \bar{r}_{\nu+1}}{\partial(\tau^2)^2} - f(\bar{r}_\nu)\frac{\partial^2 \bar{r}_\nu}{\partial(\tau^2)^2}\right) \end{array}}_{T_2}$$

Next, we show that the above display is negative for all smooth ρ. Observe that for all smooth losses ρ, $T_2 = 0$ and otherwise $T_2 \neq 0$. Hence, for the smooth losses, it suffices to show that $T_1 \leq 0$. Condition (**R**) provides that $\mathbb{E}\upsilon_1''$ and υ_1' is negative. Furthermore, Z has a symmetric density and σ is concave [3]; hence, $T_1 < 0$.

Let us know focus on non-smooth loss functions. As f is a continuous density, $f(\bar{r}_{\nu+1}) < f(\bar{r}_\nu)$ for all $\bar{r}_\nu, \bar{r}_{\nu+1} \geq 0$ and $f(\bar{r}_{\nu+1}) > f(\bar{r}_\nu)$ otherwise. Moreover, for symmetric densities $f'(\bar{r}_{\nu+1}) > 0$ for all $\bar{r}_{\nu+1} < 0$. Moreover, $f'(\bar{r}_{\nu+1}) > f'(\bar{r}_\nu)$ for all $\bar{r}_{\nu+1} > \bar{r}_\nu$ and $\bar{r}_{\nu+1}, \bar{r}_\nu < 0$. Opposite inequalities will hold on the positive axis with $f'(\bar{r}_{\nu+1}) < 0$ for all $\bar{r}_{\nu+1} > 0$. Additionally, as $\bar{r}_\nu$ is a proxy for a $Prox^{-1}$, it is concave with a negative second derivative ($Prox$ is a convex function). Therefore, the marginal derivative above is necessarily negative. Hence the sign of T_2 will alternate between negative and positive.

Part(c).

For part (c), the result of [4] provides that

$$\lim_{\tau\to\infty} \sigma'(\tau^2, \alpha\tau) = f(\alpha).$$

Moreover, they show that σ is strictly concave for $\alpha > 0$. Hence, σ will converge to some $\sigma_{\min}$ when $\tau \to \infty$.

Hence, $\partial\mathbb{V}(\tau^2, b(\tau), \alpha(\tau))$ will converge to

$$(\delta/\omega)^2 f(\alpha)\mathbb{E}2\Big[\Phi(W + \sigma_{\min}Z; b)\partial_1\Phi(W + \sigma_{\min}Z; b)\Big] \times \left[1 - \frac{\partial_{11}\mathbb{E}\left[\ \Phi(W + \sigma_{\min}Z; b)Z\right]}{\partial_{21}\mathbb{E}\left[\Phi(W + \sigma_{\min}Z; b)\right]}\right].$$

[4] show that $f(\alpha)$ is decreasing function of α. Hence, the above limit is as well. □

Proof of Lemma 3. This proof relies on Lemma 1 and a simple modification of Theorem 2 of [3]. This theorem provides a state evolution equation for a general recursion algorithm. As Lemma 1 establishes a connections between our algorithm and general recursion, the proof is then a simple application of Theorem 2 of [3], with a simple relaxation of its conditions.

Let $\bar{\tau}_t$ and $\bar{\sigma}_t$ be defined by recursion (3.1)-(3.2). By Lemma 1 and with b_i^t defined therein (7.9), Theorem 2 of [3] states

$$\lim_{n\to\infty} \frac{1}{n}\sum_{i=1}^{n} \psi(b_i^t, W_i) = \mathbb{E}\left[\psi(\bar{\sigma}_t Z, W)\right] \tag{7.25}$$

for any pseudo-Lipschitz function $\psi : \mathbb{R}^2 \to \mathbb{R}$ of order k and for all W_i with bounded $2k-2$ moments. Careful inspection of the proof of Lemma 5 of [3] shows that if ψ is a function that is uniformly bounded, the restriction on the moments of W_i is unnecessary. A version of Hoeffding's inequality suffices, as applied to independent and not-necessarily equally distributed random variables (see Theorem 12.1 in [13]).

Next, we split the analysis into two cases: Φ is differentiable and Φ is not differentiable. For the first case, it suffices to observe that by Lemma 1 we have $b_i^t = W_i - z_i^t$, with z_i^t defined in (2.3). Next, we choose ψ to be

$$\psi(s,t) = \partial_1 \Phi(t-s;b).$$

Then, $\psi(b_i^t, W_i) = \partial_1 \Phi(W_i - W_i + z_i^t; b)$ and by Condition (**R**) ψ is a uniformly bounded function. Thus, application of the result above provides

$$\lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} \partial_1 \Phi(z_i^t; b) = \mathbb{E}\left[\partial_1 \Phi(W - \bar{\sigma}_t Z; b)\right].$$

The proof then follows by observing that the right hand side is equal to ω/δ by (2.4).

Next, we discuss the case of non-differentiable losses ρ. Let h be a bandwidth parameter of an estimator of $\nu(b)$. We define

$$S_n(h) = \sum_{i=1}^{n} \left[\Phi(z_i^t + n^{-1/2}h; b) - \Phi(z_i^t; b)\right]$$

for $h \in [0, C]$ for some constant C: $0 < C < \infty$. We set $S_n^o(h) = S_n(h) - \mathbb{E}S_n(h)$, for $h \in [0, C]$. Moreover, by Condition (**R**) (i) $\Phi(z_i^t; b) = bv_1(Prox(z_i^t, b)) + bv_2(Prox(z_i^t, b))$. Absolutely-continuous term ν_1 can be handled as the above case; hence, without loss of generality we can assume it is equal to zero. Hence,

$$S_n(h) = \sum_{i=1}^{n} \sum_{\nu=1}^{k-1} b \Big[\mathbb{1}\left\{Prox(z_i^t + n^{-1/2}h, b) \in (r_\nu, r_{\nu+1})\right\} - \mathbb{1}\left\{Prox(z_i^t, b) \in (r_\nu, r_{\nu+1})\right\}\Big]$$

and

$$\mathbb{E}S_n(h) = b \sum_{\nu=1}^{k-1} \Big[\mathbb{P}\left\{Prox(z_i^t + n^{-1/2}h, b) \in (r_\nu, r_{\nu+1})\right\} - \mathbb{P}\left\{Prox(z_i^t, b) \in (r_\nu, r_{\nu+1})\right\}\Big].$$

As $Prox(z, b) = z - \Phi(z, b)$, we know the term above can be further written as

$$\mathbb{E}S_n(h) = \sum_{i=1}^{n} b \sum_{\nu=1}^{k-1} \Big[\mathbb{P}\left\{z_i^t + n^{-1/2}h - \Phi(z_i^t + n^{-1/2}h, b) \in (r_\nu, r_{\nu+1})\right\} - \mathbb{P}\left\{z_i^t - \Phi(z_i^t, b) \in (r_\nu, r_{\nu+1})\right\}\Big].$$

Then, by the same arguments as for (7.25) we obtain $n^{-1/2}S_n^o(h) \to \mathcal{N}(0, \gamma^2(h))$, for $h \in [0, C]$, in distribution, where for each $h \in [0, C]$, and

$$\gamma^2(h) = b \sum_{\nu=1}^{k-1} \alpha_\nu \lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} h \left[f_{z_i^t - \Phi(z_i^t, b)}(r_{\nu+1}) - f_{z_i^t - \Phi(z_i^t, b)}(r_\nu) \right].$$

Then, by the arguments of (7.25) we conclude

$$\gamma^2(h) = bh \sum_{\nu=1}^{k-1} \alpha_\nu \left[f_{W - \bar{\sigma}_t Z}(r_{\nu+1}) - f_{W - \bar{\sigma}_t Z}(r_\nu) \right].$$

The right hand side of the equality above is finite by the Condition (**R**). To establish a uniform statement, we need to establish the compactness or tightness of the sequence $n^{-1/2}S_n(h)$ for $h \in [0, C]$. This follows by noticing that the sequence is a sequence of differences of two, univariate, empirical distribution functions, both of which weakly converge to a Wiener function (see Lemma 5.5.1 in [28]). Hence,

$$\sup_{|h|\le C} n^{-1/2} \sum_{i=1}^{n} \left[\Phi(z_i^t + h; b) - \Phi(z_i^t; b) + hb\gamma^* \right] = O_P(n^{-\tau}) \tag{7.26}$$

where $\tau = 1/2$ for continuous ψ and $\tau = 1/4$ for discontinuous ψ. In the display above $\gamma^* = \sum_{\nu=1}^{k} (\alpha_\nu - \alpha_{\nu-1}) f_{W - \bar{\sigma}_t Z}(r_\nu)$. By the definition ω/δ is the derivative of a consistent estimator of $\nu(b) = \partial_1 \mathbb{E}\Phi(z^t, b_t)$. Because of the equation above, we see that

$$\omega/\delta = b \sum_{\nu=1}^{k} (\alpha_\nu - \alpha_{\nu-1}) f_{W - \bar{\sigma}_t Z}(r_\nu),$$

for all consistent estimators of $\nu(b)$ with a bandwidth choice of $h \to 0$ and $nh \to \infty$. □

7.6. Proofs for section 3.2

Proof of Theorem 1. The proof is split into two parts. In the first step, we show that the proposed algorithm belongs to the class of generalized recursions as defined in [3]. The result is presented in Lemma 1.

In the second step, we utilize conditioning technique and the result of Theorem 2 of [3] designed for generalized recursions. For an appropriate sequence of vectors h_i^t of generalized recursions and a x_0 the true regression coefficient, they show

$$\lim_{p\to\infty} \frac{1}{p} \sum_{i=1}^{p} \psi(h_i^{t+1}, x_{0,i}) = \mathbb{E}\left[\psi(\bar{\tau}_t^* Z, X_0)\right] \tag{7.27}$$

for a pseudo-Lipschitz function ψ. We now proceed to identify x^t for a suitable h_i^t of the proposed RAMP algorithm. By definition of RAMP,

$$x^{t+1} \overset{(i)}{=} \eta(x^t + A^T G(z^t, b_t); \theta) \overset{(ii)}{=} \eta(x_0 - x_0 + x^t + A^T G(z^t, b_t); \theta)$$
$$\overset{(iii)}{=} \eta(x_0 - h^{t+1}), \tag{7.28}$$

where equation (i) is because of the iteration RAMP, the equation (ii) is plus and minus a same term and the equation (iii) is the special choice of h^{t+1} in equation(7.14). Therefore, combining x^t in equation (7.28) and equation (7.27), we obtain

$$\lim_{p\to\infty} \frac{1}{p} \sum_{i=1}^{p} \psi(x_i^t, x_{i,0}) = \lim_{p\to\infty} \frac{1}{p} \sum_{i=1}^{p} \psi(\eta(x_{0,i} - h_i^t; \theta), x_{0,i})$$
$$= \mathbb{E}[\psi(\eta(X_0 - \bar{\tau}_t^* Z; \theta), X_0)]$$

□

Proof of Theorem 2. In order to prove this result we designed a series of Lemmas $4-15$ provided in the Appendix . The main part of the proof is provided by the results of Lemma 6. In the next steps we apply Lemma 6 to the specific choice of vectors $x = x^t$ and $r = |\hat{x} - x^t|$. We show there exist constants $c_1, ..., c_5 > 0$, such that for each $\epsilon > 0$ and some iteration t, Conditions $(C1)-(C6)$ of Lemma 6 hold with probability going to 1 as $p \to \infty$.

Condition $(C1)$. We need to show $|||x^t - \hat{x}|||_2 \leq c_1\sqrt{p}$. Lemma 1 proves that the RAMP algorithm is a special case of a general iterative and recursive scheme, as defined in [3]. From (7.27) we choose $\psi(a, b) = a^2$ and obtain

$$\lim_{t\to\infty}\lim_{p\to\infty} \frac{||x^t||^2}{p} = \mathbb{E}\{\eta(X_0 + \bar{\tau}^* Z; \theta^*)\}^2 < \infty.$$

Moreover, we observe that $\frac{||\hat{x}||_2}{p} < \infty$ by assumptions of the Theorem.

Condition $(C2)$. By the definition of $\hat{x}$ as the minimizer of the $\mathcal{L}$, we conclude that $\mathcal{L}(\hat{x}) < \mathcal{L}(x)$ for any $x \neq \hat{x}$ and this applies for $x = x^t$.

Condition $(C3)$. We need to show $||sg(\mathcal{L}, x^t)||_2 \leq \epsilon\sqrt{p}$. By the definition of the RAMP iteration

$$x^t = \begin{cases} A^T G(z^{t-1}, b_{t-1}) + x^{t-1} + \theta_{t-1}, & \text{if } A^T G(z^{t-1}, b_{t-1}) + x^{t-1} \geq \theta_{t-1} \\ A^T G(z^{t-1}, b_{t-1}) + x^{t-1} - \theta_{t-1}, & \text{if } A^T G(z^{t-1}, b_{t-1}) + x^{t-1} \leq -\theta_{t-1} \\ 0 & \text{otherwise} \end{cases}.$$

This indicates that when $x^t = 0$

$$\frac{|A^T G(z^{t-1}, b_{t-1}) + x^{t-1}|}{\theta_{t-1}} \leq 1,$$

and that in cases of $x^t \neq 0$

$$A^T G(z^{t-1}, b_{t-1}) + x^{t-1} = x^t + \text{sign}(x^t)\theta_{t-1}.$$

Therefore, the subgradient $sg(\mathcal{L}, x^t)$ must satisfy

$$sg(\mathcal{L}, x^t) \equiv \begin{cases} \lambda \text{sign}(x^t) - A^T\rho'(Y - Ax^t), & \text{if } x^t \neq 0 \\ \lambda \frac{A^T G(z^{t-1}, b_{t-1}) + x^{t-1}}{\theta_{t-1}} - A^T\rho'(Y - Ax^t), & \text{if } x^t = 0 \end{cases} . \quad (7.29)$$

Moreover, by equation (2.3) and Lemma 1

$$Y - Ax^t = z^t - \Phi(z^{t-1}, b_{t-1}).$$

Then,

$$\begin{aligned} A^T\rho'(Y - Ax^t) &= A^T\rho'(z^t - \Phi(z^{t-1}, b_{t-1})) \\ &= A^T\rho'(Prox(z^t, b_t) + \Phi(z^t, b_t) - \Phi(z^{t-1}, b_{t-1})), \quad (7.30) \end{aligned}$$

where we used the fact that $z^t - \Phi(z^{t-1}, b_{t-1}) = Prox(z^t, b_t)$. Adding equation (7.30) to (7.29) and the expression of $sg(\mathcal{L}, x^t)$, we conclude

$$sg(\mathcal{L}, x^t) = \lambda s^t - A^T\rho'(Prox(z^t, b_t) + \Phi(z^t, b_t) - \Phi(z^{t-1}, b_{t-1})), \quad (7.31)$$

where

$$s^t = \begin{cases} \text{sign}(x^t) & \text{, if } x^t \neq 0 \\ \frac{A^T G(z^{t-1}, b_{t-1}) + x^{t-1}}{\theta_{t-1}} & \text{, if } x^t = 0 \end{cases} .$$

Now, we rewrite $sg(\mathcal{L}, x^t)$ as follows

$$\begin{aligned} sg(\mathcal{L}, x^t) &= \frac{1}{\theta_{t-1}} \left[\lambda\theta_{t-1} s^t - \frac{\lambda\delta b}{\omega} A^T\rho'(Prox(z^t, b_t) + \Phi(z^t, b_t) - \Phi(z^{t-1}, b_{t-1})) \right] \\ &+ \frac{1}{\theta_{t-1}} \left[\frac{\lambda\delta b}{\omega} - \theta_{t-1} \right] A^T\rho'(Prox(z^t, b_t) + \Phi(z^t, b_t) - \Phi(z^{t-1}, b_{t-1})). \end{aligned}$$

Then, by the non-negativity of θ_{t-1} and triangular inequality

$$\begin{aligned} &\frac{1}{\sqrt{p}} \left\| sg(\mathcal{L}, x^t) \right\|_2 \\ \leq\ & \underbrace{\frac{\lambda}{\theta_{t-1}\sqrt{p}} \left\| \lambda\theta_{t-1} s^t - \frac{\lambda\delta b}{\omega} A^T\rho'(Prox(z^t, b_t) + \Phi(z^t, b_t) - \Phi(z^{t-1}, b_{t-1})) \right\|_2}_{A} \\ &+ \underbrace{\frac{\left| \frac{\lambda\delta b}{\omega} - \theta_{t-1} \right|}{\theta_{t-1}} \left\| A^T\rho'(Prox(z^t, b_t) + \Phi(z^t, b_t) - \Phi(z^{t-1}, b_{t-1})) \right\|_2}_{B} . \end{aligned}$$

We consider the bound of B first.

Observe that $Prox(z^t, b_t) = z^t - \Phi(z^t, b_t)$. Then, utilizing (7.25) and (7.26), we observe that there exists a $0 < q < \sqrt{p}$ such that $\|Prox(z^t, b_t) + \Phi(z^t, b_t) - \Phi(z^{t-1}, b_{t-1})\|_2 \leq q$. We define $M \equiv \sup_{\|z\|_2 \leq q} \rho''(z)$, where $\rho''(z) = v_1'(z)$. Then, by

Condition (**R**) (i)-(ii) and (iv) we know that $M < \infty$. Then by Taylor expansion and Triangle inequality, we conclude

$$B \leq \frac{|\frac{\lambda\delta b}{\omega} - \theta_{t-1}|}{\theta_{t-1}} \times \frac{1}{\sqrt{p}}\Big[\| + A^T\rho'(Prox(z^t, b_t))\|_2 + M\|A\|_2\|\Phi(z^t, b_t) - \Phi(z^{t-1}, b_{t-1})\|_2\Big].$$

Moreover,

$$\frac{|\frac{\lambda\delta b}{\omega} - \theta_{t-1}|}{\theta_{t-1}}\frac{1}{\sqrt{p}}||A^T\rho'(Prox(z^t, b_t))||_2 = \frac{|\frac{\lambda\delta b}{\omega} - \theta_{t-1}|}{\theta_{t-1}b_t}\frac{1}{\sqrt{p}}||A^T\Phi(z^t, b_t)||_2 \leq \frac{|\frac{\lambda\delta b}{\omega} - \theta_{t-1}|}{\theta_{t-1}b_t}\frac{1}{\sqrt{p}}\sigma_{max}(A)||\Phi(z^t, b_t)||_2$$

Next, we observe that the state evolution (by Lemma 3 and Theorem 1) guarantees,

$$\lim_{p\to\infty}\frac{||\Phi(z^t, b_t)||_2}{p} < \infty.$$

Moreover, $\sigma_{max}(A)$ is almost surely bounded as $p \to \infty$ []. Hence, we conclude

$$\lim_{t\to\infty}\lim_{p\to\infty}\frac{\sigma_{max}(A)\|\Phi(z^t, b_t) - \Phi(z)\|_2}{\sqrt{p}} = 0.$$

Furthermore, using Lemma 5 we obtain

$$\lim_{t\to\infty}\lim_{p\to\infty}\frac{|\frac{\lambda\delta b_t}{\omega} - \theta_{t-1}|}{\theta_{t-1}b_t} = \frac{\frac{\lambda\delta b^*}{\omega} - \theta^*}{\theta^* b^*} = 0.$$

Therefore, B converges to 0 when $p \to \infty$.

Now we consider A. From equation (7.29), we conclude that

$$\theta_{t-1}s^t - \frac{\delta b^t}{\omega}A^T\rho'(Prox(z^t, b_t)) = \theta_{t-1}s^t s^t - \frac{\delta}{\omega}A^T\Phi(z^t, b_t) = x^t - x^{t-1}.$$

Plugging into A, we obtain

$$A \leq \frac{\lambda}{\theta_{t-1}\sqrt{p}}\|x^t - x^{t-1}\|_2 + \frac{\lambda\delta b_t}{\omega\theta_{t-1}\sqrt{p}}M\sigma_{max}(A)\left\|\Phi(z^t, b_t) - \Phi(z^{t-1}, b_{t-1})\right\|_2.$$

The convergence of the second term is by the convergence of the term B and the first term is converging to 0 by the convergence of the RAMP algorithm – that is the result of Theorem 1 holds. Therefore, A converges to 0 when $p \to \infty$. This finishes the proof of Condition ($C3$).

Condition ($C4$). This result follows from Lemma 12 provided in the Appendix.

Condition ($C5$). Let $A \in R^{n\times p}$ be a matrix with i.i.d. entries such that $\mathbb{E}\{A_{ij}\} = 0$, $\mathbb{E}\{A_{ij}^2\} = 1/n$, and $n = p\delta$. Let $\sigma_{max}(A)$ be the largest singular

value of A and $\sigma_{min}(A)$ be its smallest non-zero singular value. Then, [2] provide a general result that claims

$$\lim_{p\to\infty}\sigma_{max}(A)=\frac{1+\sqrt{\delta}}{\sqrt{\delta}},\qquad \lim_{p\to\infty}\sigma_{min}(A)=\frac{1-\sqrt{\delta}}{\sqrt{\delta}}. \tag{7.32}$$

Condition $(C6)$. Assumption $(\mathbf{R})$ is guaranteeing the validity of $(C6)$. Conditions $(C1)$-$(C6)$ are checked and the proof is completed. □

7.7. *Auxiliary results*

This section gathers results used throughout the proofs. They are of secondary interest, so we present them in this Appendix section.

Lemma 6. *Let r and x be vectors in $\mathbb{R}^p$, $\mathcal{L}$ defined in Problem (1.2) and $sg(\mathcal{L},x)\in\partial\mathcal{L}(x)$ the subgradient of $\mathcal{L}$ with respect to x. For any $c_1,...c_5>0$, if the following Conditions 1-5 hold, then there exists a function $\xi(\epsilon,c_1,...,c_5)\to 0$ as $\epsilon\to 0$ such that $||r||_2\leq\sqrt{p}\xi(\epsilon,c_1,...,c_5)$.*

The conditions are: (C1) $\|r\|_2\leq c_1\sqrt{p}$; (C2) $\mathcal{L}(x+r)\leq\mathcal{L}(x)$; (C3) There exists a $sg(\mathcal{L},x)\in\partial\mathcal{L}(x)$ with $||sg(\mathcal{L},x)||_2\leq\sqrt{p}\epsilon$; (C4) Let $v\equiv\frac{1}{\lambda}[\sum_{i=1}^p\rho'(Y_i-A_i^Tx)A_i+sg(\mathcal{L},x)]\in\partial||x||_1$, and $S(c_2)\equiv\{i\in[p]:|v_i|\geq 1-c_2\}$. Then, for any $S'\subseteq[p],|S'|\leq c_3p$, we have $\sigma_{min}(\{A\}_{S(c_2)\cup S'})\geq c_4$; (C5) The maximum and minimum non-zero singular value of A satisfy $\frac{1}{c_5}\leq\sigma_{min}(A)^2\leq\sigma_{max}(A)^2\leq c_5$; (C6) For all such vectors r, the loss function ρ satisfies $E_i=\mathbb{E}(v_1'(W_i))\geq k_1$ for a constant $k_1>0$.

Proof of Lemma 6. The proof follows the strategy of Lemma 3.1. of [4], with nontrivial adaptation to a class of general loss functions.

Let $S=supp(x)\subseteq[p]$, where $supp(x)\equiv\{i|x_i\neq 0\}$ and $[p]=\{1,2,...,p\}$ and let $\bar{S}$ be its complement. Let r be the vector that satisfies Conditions $(C1)$ and $(C2)$, i.e., it is such that $\|r\|_2\leq\sqrt{p}$ and $\mathcal{L}(x+r)-\mathcal{L}(x)\geq 0$. Observe that we can decompose the Lasso penalty as follows

$$\|x+r\|_1-\|x\|_1=\|x_S+r_S\|_1-\|x_S\|_1+\|r_{\bar{S}}\|_1, \tag{7.33}$$

as $x_S=x$ and $r=r_S+r_{\bar{S}}$.

Let us define a vector v as

$$v\equiv\frac{1}{\lambda}[\sum_{i=1}^p\rho'(Y_i-A_i^Tx)A_i+sg(\mathcal{L},x)] \tag{7.34}$$

By observing that the subgradients of $\mathcal{L}(x)$ satisfy $sg(\mathcal{L},x)=\lambda\partial||x||_1-\sum_{i=1}^n\rho'(Y_i-A_i^Tx)A_i$, we obtain that $v_S=\partial||x_S||_1$. Moreover, by adding and subtracting $\langle v,r\rangle$

$$\|r_{\bar{S}}\|_1\geq-p\langle\partial\|x_S\|_1,r_S\rangle+(\|r_{\bar{S}}\|_1-p\langle v_{\bar{S}},r_{\bar{S}}\rangle)+p\langle v,r\rangle) \tag{7.35}$$

where $\langle u,v\rangle\equiv\frac{1}{m}\sum_{i=1}^m u_iv_i$, denotes the scalar product for $u,v\in\mathbb{R}^m$.

By observing that $\mathcal{L}(x+r)-\mathcal{L}(x)\geq 0$ and by plugging in all of the above inequalities, we conclude

$$\begin{aligned} 0 \quad &\overset{(iii)}{\geq} \quad \lambda\left(\frac{\|x_S+r_S\|_1-\|x_S\|_1}{p}-\langle\partial\|x_S\|_1,r_S\rangle\right)+\lambda\left(\frac{\|r_{\bar{S}}\|_1}{p}-\langle v_{\bar{S}},r_{\bar{S}}\rangle\right) \\ &+ \quad \lambda\langle v,r\rangle-\Delta_n \end{aligned} \tag{7.36}$$

where (iii) follows from plugging equations (7.33) and (7.35) in $\mathcal{L}(x+r)-\mathcal{L}(x)\geq 0$ and where

$$p\Delta_n=\sum_{i=1}^{n}\left[\rho(Y_i-A_i^Tx-A_i^Tr)-\rho(Y_i-A_i^Tx)\right].$$

Next, we observe that

$$\lambda\langle v,r\rangle=\langle sg(\mathcal{L},x),r\rangle+p^{-1}\sum_{i=1}^{n}\rho'(Y_i-A_i^Tx)(A_i^Tr). \tag{7.37}$$

Let γ_n be a sequence of positive numbers. We define the following event

$$\mathcal{E}_n=\left\{\left|\frac{\sum_{i=1}^{n}\rho'(Y_i-A_i^Tx)(A_i^Tr)}{p}\right|\leq\gamma_n:\forall\|r\|_2\leq\sqrt{p}\right\}. \tag{7.38}$$

Then, conditionally on $\mathcal{E}$ we have

$$\begin{aligned} \gamma_n \quad &\overset{(iii)}{\geq} \quad \lambda\left(\frac{\|x_S+r_S\|_1-\|x_S\|_1}{p}-\langle\partial\|x_S\|_1,r_S\rangle\right)+\lambda\left(\frac{\|r_{\bar{S}}\|_1}{p}-\langle v_{\bar{S}},r_{\bar{S}}\rangle\right) \\ &+ \quad \lambda\langle sg(\mathcal{L},x),r\rangle-\Delta_n. \end{aligned} \tag{7.39}$$

We discuss the last term first. We rewrite $p\Delta_n$ as

$$p\Delta_n=\mathbb{V}_n(r)+n\mathbb{E}\mathrm{v}_1(r),$$

where $\mathbb{V}_n(r)=\sum_{i=1}^{n}[\mathrm{v}_i(r)-\mathbb{E}\mathrm{v}_i(r)]$, with $\mathrm{v}_i(r)=\rho(Y_i-A_i^Tx-A_i^Tr)-\rho(Y_i-A_i^Tx)$.

Let η_n be a sequence of positive numbers. Then, we consider the following event

$$\mathcal{V}_n=\left\{|\mathbb{V}_n(r)|\leq\eta_n:\|r\|_2^2\leq p\right\}.$$

Conditioning on this event, the inequality (7.36) becomes

$$\begin{aligned} \gamma_n+\eta_n \quad &\overset{(iii)}{\geq} \quad \lambda\left(\frac{\|x_S+r_S\|_1-\|x_S\|_1}{p}-\langle\partial\|x_S\|_1,r_S\rangle\right) \\ &+\lambda\left(\frac{\|r_{\bar{S}}\|_1}{p}-\langle v_{\bar{S}},r_{\bar{S}}\rangle\right)+\lambda\langle sg(\mathcal{L},x),r\rangle+\frac{n}{p}\mathbb{E}\mathrm{v}_1(r). \end{aligned} \tag{7.40}$$

Moreover, Cauchy Schwartz Inequality tells that:

$$-\frac{||sg(\mathcal{L},x)||_2||r||_2}{p} \leq -\langle sg(\mathcal{L},x), r\rangle \leq \frac{||sg(\mathcal{L},x)||_2||r||_2}{p}.$$

Using Conditions $(C1)$ and $(C3)$, inequality (7.40) becomes

$$\lambda\left(\frac{\|x_S + r_S\|_1 - \|x_S\|_1}{p} - \langle \text{sign}(x_S), r_S\rangle\right) + \lambda\left(\frac{\|r_{\bar{S}}\|_1}{p} - \langle v_{\bar{S}}, r_{\bar{S}}\rangle\right) + \frac{n}{p}\mathbb{E}\mathrm{v}_1(r) \leq c_1\epsilon + \gamma_n + \eta_n.$$

The first two terms of the above right hand side are non-negative (proven by arguments identical to the Lemma 3.1 in [4]). For the last term we employ results of Lemma 9 to obtain

$$\begin{aligned}&\frac{n}{p}\mathbb{E}[\rho(Y_i - A_i^T x - A_i^T r) - \rho(Y_i - A_i^T x)]\\ &\geq -\frac{n}{p}\mathbb{E}[\psi(W_i)A_i^T r] + \frac{n}{p}\kappa[A_i^T r]^2 - o_P(1).\end{aligned}$$

In the display above, the first term disappears; for the second one

$$2\kappa = \mathbb{E}v_1'(W_i) + \gamma,$$

for γ defined in Lemma 9. According to Lemma 9 and Condition (C6), we conclude that γ is strictly positive. Therefore, there exists a constant $C > 0$ such that

$$\frac{1}{p}C\|Ar\|_2^2 \leq c_1\epsilon + \gamma_n + \eta_n + o_P(1) := \xi_1(\epsilon). \tag{7.41}$$

To complete the proof we need to show that $\xi_1(\epsilon) \to 0$ and then employ arguments similar to Lemma 3.1 in [4]. This can be done by effectively bounding the size of the events $\mathcal{E}_n$ and $\mathcal{V}_n$.

The size of η_n can be found by choosing appropriate sequence u_n of Lemma 7. For $u_n = \sqrt{(\log p)^2/(pn)}$ we obtain that $\eta_n = nu_n = (\log p)\sqrt{n}/\sqrt{p}$ is sufficient to guarantee that $P(\mathcal{V}_n) \geq 1 - \exp\{-2\log p/\kappa^2\}$.

Similarly, the size of γ_n can be found by choosing appropriate sequence u_n of Lemma 8. For $u_n = \sqrt{(\log p)^2/(pn)}$ we obtain that $\eta_n = nu_n = (\log p)\sqrt{n}/\sqrt{p}$ is sufficient to guarantee that $P(\mathcal{E}_n) \geq 1 - \exp\{-2\log p/\kappa^2\}$.

□

Lemma 7. *Let $|\rho'(u)| \leq \kappa$ for all $u \in \mathbb{R}$ and some constant $\kappa < \infty$. Then, for all vectors $\mathbf{r}$, such that $\|\mathbf{r}\|_2 \leq \sqrt{p}$ and for any sequence of positive numbers $u_n \geq 0$ we have*

$$\mathbb{P}\left(\left|\sum_{i=1}^n \mathrm{v}_i(\mathbf{r}) - \mathbb{E}\mathrm{v}_i(\mathbf{r})\right| \geq npu_n\right) \leq \exp\left\{-2\frac{n^2pu_n^2}{\kappa^2\log p}\right\}, \tag{7.42}$$

for $\mathrm{v}_i(\mathbf{r}) = \rho(Y_i - A_i^T\mathbf{x}_o - A_i^T\mathbf{r}) - \rho(Y_i - A_i^T\mathbf{x}_o)$.

Proof of Lemma 7. Let $\mathbb{V}_n(\mathbf{r}) = \sum_{i=1}^n [\mathrm{v}_i(\mathbf{r}) - \mathbb{E}\mathrm{v}_i(\mathbf{r})]$. We begin by observing

$$p^{-1}\mathbb{V}_n(\mathbf{r}) \leq \sum_{i=1}^n p^{-1} \left|\mathrm{v}_i(\mathbf{r}) - \mathbb{E}\mathrm{v}_i(\mathbf{r})\right|,$$

for $\mathrm{v}_i(\mathbf{r}) = \rho(Y_i - A_i^T\mathbf{x}_o - A_i^T\mathbf{r}) - \rho(Y_i - A_i^T x)$. Then, by a Taylor expansion of the loss function ρ around, we conclude

$$|\rho(Y_i - A_i^T\mathbf{x}_o - A_i^T\mathbf{r}) - \rho(Y_i - A_i^T\mathbf{x}_o)| \leq |H_i(c)A_i^T\mathbf{r}|$$

for $H_i(c) = \sup_{|u|\leq c} \rho'(W_i - u)$. By Hoelder's inequality we conclude

$$|\mathrm{v}_i(\mathbf{r}) - \mathbb{E}\mathrm{v}_i(\mathbf{r})| \leq |H_i(c) - \mathbb{E}H_i(c)| \left|\langle A_i^T\mathbf{r}\rangle\right|.$$

We proceed to bound each term in the RHS above, independently. For the first term, we observe that for a positive, bounded constant κ, the boundedness of the sub-gradient provides $|H_i(c) - \mathbb{E}H_i(c)| \leq \kappa$. For the second term, as A_{ij} are Gaussian with variance $1/n$, by the weighted Bernstein inequality

$$\begin{aligned}
\mathbb{P}\left(1/p\left|\langle A_i^T\mathbf{r}\rangle\right| \geq a_n\right) &\leq \mathbb{P}\left(\sum_{j=1}^p |A_{ij}r_j| \geq pa_n\right) \\
&\leq \exp\left\{-\frac{p^2a_n^2}{4\sum_{j=1}^p A_{ij}^2r_{ij}^2 + 2Cp\max_j|A_{ij}|/3}\right\} \\
&\leq \exp\left\{-\frac{p^2na_n^2}{4\|\mathbf{r}\|_2^2}\right\}. \qquad (7.43)
\end{aligned}$$

For all r such that $\|\mathbf{r}\|_2 \leq \sqrt{p}c_1$, the right hand side is smaller than $\exp\{-pna_n^2/4c_1\}$. Hence, a choice of $a_n = \sqrt{\log p/(np)}$ leads that

$$p^{-1}|\mathrm{v}_i(\mathbf{r}) - \mathbb{E}\mathrm{v}_i(\mathbf{r})| = O_P\left(\sqrt{\frac{\log p}{np}}\right).$$

This, in turn, guarantees that $\frac{1}{p}\mathbb{V}_n(\mathbf{r})$ is a sum of n terms, each of which is $o_P(1)$. By Hoefding's inequality for bounded random variables, for any positive sequence of u_n

$$\mathbb{P}\left(\frac{1}{p}|\mathbb{V}_n(\mathbf{r})| \geq nu_n\right) \leq \exp\left\{-2\frac{n^2u_n^2}{\kappa^2\sum_{i=1}^n \frac{\log p}{np}}\right\} \leq \exp\left\{-2\frac{pn^2u_n^2}{\kappa^2\log p}\right\}. \quad \square$$

Lemma 8. *Let $|\rho'(u)| \leq \kappa$ for all $u \in \mathbb{R}$ and some constant $\kappa < \infty$. Then, for a positive sequence of $u_n \geq 0$ we have*

$$\mathbb{P}\Big(|\langle \bigtriangledown\mathcal{L}(\mathbf{x}_o), \mathbf{r}\rangle| \geq u_n\Big) \leq \exp\left\{-\frac{npu_n^2}{2\kappa^2(\log p)}\right\}.$$

Proof of Lemma 8. Let

$$\bigtriangledown \mathcal{L}(\mathbf{x}_o) = \sum_{i=1}^{n} \rho'(Y_i - A_i^T \mathbf{x}_o) A_i^T$$

and observe that $\mathbb{E} \bigtriangledown \mathcal{L}(\mathbf{x}_o) = 0$ by the vanishing property of the true score function $\mathbb{E}\rho'(Y_i - A_i^T \mathbf{x}_o) = 0$. Hence,

$$\langle \bigtriangledown \mathcal{L}(\mathbf{x}_o), r \rangle = p^{-1} \sum_{i=1}^{n} \rho'(Y_i - A_i^T \mathbf{x}_o)(A_i^T \mathbf{r})$$

and is such that $\mathbb{E}\langle \bigtriangledown \mathcal{L}(\mathbf{x}_o), \mathbf{r} \rangle = 0$. By a triangular inequality and the bounded sub-gradient assumption

$$|\langle \bigtriangledown \mathcal{L}(\mathbf{x}_o), \mathbf{r} \rangle| \leq \kappa \sum_{i=1}^{n} q_i(\mathbf{r})$$

with

$$q_i(\mathbf{r}) = p^{-1} \sum_{j=1}^{p} |A_{ij} r_j|.$$

Then, $\mathbb{E}q_i(\mathbf{r}) = 0$ as A is a mean zero design matrix. From Lemma 7, equation (7.43), we conclude that

$$\mathbb{P}\left(\sum_{i=1}^{n} q_i(\mathbf{r}) \geq u_n \right) \leq \exp\left\{ -\frac{u_n^2}{2(\log p)/(np)} \right\}. \qquad \square$$

Lemma 9. *Consider the model* (1.1) *with Conditions* (**R**), (**D**) *and* (**A**) *satisfied. Let* $\mathbf{r}$ *be a vector in* $\mathbb{R}^p$ *such that* $\|\mathbf{r}\|_2^2 \leq Cp$ *for a constant* C*:* $0 < C < \infty$*. Then,*

$$\sup_{\|\mathbf{r}\|_2 \leq \sqrt{p}} p^{-1} \left| \sum_{i=1}^{n} \left[\rho(Y_i - A_i^T x - A_i^T \mathbf{r}) - \rho(Y_i - A_i^T x) \right] + \mathbf{r}^T \sum_{i=1}^{n} A_i \rho'(Y_i - A_i^T x) - \gamma \mathbf{r}^T \sum_{i=1}^{n} A_i^T A_i \mathbf{r} \right| = o_P(1), \quad (7.44)$$

as n *and* $p \to \infty$*, with* $2\gamma = \mathbb{E}v_1'(W) + \sum_{\nu=1}^{k} (\alpha_\nu - \alpha_{\nu-1}) f_W(r_\nu)$.

Proof of Lemma 9. It suffices to prove

$$\sup_{\|\mathbf{r}\|_2 \leq \sqrt{p}} \left\| \sum_{i=1}^{n} A_i \left[\psi(Y_i - A_i^T x - A_i^T \mathbf{r}) - \psi(Y_i - A_i^T x) \right] + \gamma \sum_{i=1}^{n} A_i^T A_i \mathbf{r} \right\|_\infty = O_P(\sqrt{p \log p / n}) \quad (7.45)$$

where $2\gamma = \mathbb{E}v_1'(W) + \sum_{\nu=1}^{k}(\alpha_\nu - \alpha_{\nu-1})f_W(r_\nu)$, together with $|\sum_{i=1}^{n} A_{ij}\Psi(W_i)| = O_P(1)$, for all $j = 1, \ldots, p$. The above, in turn, implies through integration over r the statement (7.44).

Let $j = 1, \ldots, p$. We first argue that $|\sum_{i=1}^{n} A_{ij}\Psi(W_i)| = O_P(1)$: by Condition (D), $A_{ij} = O_P(\sqrt{1/n})$ and by Condition (R), $|n^{-1/2}\sum_{i=1}^{n}\Psi(W_i)| = O_P(1)$ (bounded random variables no matter of the size of W_i – consequence of Theorem 12.1 [13]).

Next, we prove (7.45). For that end, define a stochastic process

$$S_n(r) = \sum_{i=1}^{n} A_i\left[\psi(Y_i - A_i^T x - A_i^T \mathbf{r}) - \psi(Y_i - A_i^T x)\right]$$

for $r \in [-C\sqrt{p}, C\sqrt{p}]^p$, for some C: $0 < C < \infty$. We let $\psi = v_1 + v_2$ and denote the absolute continuous and step-function components by v_1 and v_2, respectively.

Case I: $\psi = v_2$ (i.e., $v_1 = 0$).

Without loss of generality, we assume that there is a single jump-point. We set $v_2(y)$ to be 0 or 1 according to y being ≤ 0 or > 0. By the vector structure in (7.45), it suffices to show that for each coordinate of $S_n(r)$ the uniform asymptotic linearity result holds for $r \in [-C\sqrt{p}, C\sqrt{p}]^p$. To simplify the notation, we consider only the first coordinate and drop the subscript 1 in $S_{n1}(r)$:

$$S_n^0(r) = S_n(r) - \mathbb{E}S_n(r),$$

where $\mathbb{E}S_n(r) = \sum_{i=1}^{n} A_{i1}\left[F_W(0) - F_W(A_i^T\mathbf{r})\right]$. By Taylor expansion, we have $F_W(0) - F_W(A_i^T\mathbf{r}) = f_w(0)A_i^T\mathbf{r} + f_W'(\xi)[A_i^T\mathbf{r}]^2$, for $\xi \in (0, A_i^T\mathbf{r})$. Moreover, by (7.43), $|A_i^T\mathbf{r}| = O_P(\sqrt{p\log p/n})$ and by Condition (D), $A_{ij} = O_P(\sqrt{1/n})$. Therefore,

$$\left|\sum_{i=1}^{n} A_{i1}f_W'(\xi)[A_i^T\mathbf{r}]^2\right| = O_P((p\log p)/n).$$

Hence, by Hoefding's inequality, Theorem 12.1 of [13], we have

$$|S_n^0(r)| = O_P(\sqrt{p\log p/n}),$$

for $\mathbf{r} \in [-C\sqrt{p}, C\sqrt{p}]^p$. To prove uniform asymptotic linearity we resort to the known weak convergence properties of the empirical cumulative distribution functions to the Brownian motion [28] or by uniform decompositions of the work of [7].

Case II: $\psi = v_1$ (i.e., $v_2 = 0$). Note that for every $\mathbf{r} \in [-C\sqrt{p}, C\sqrt{p}]^p$, by a second-order Taylor's expansion,

$$\psi(Y_i - A_i^T x - A_i^T\mathbf{r}) - \psi(Y_i - A_i^T x) = v_1'(Y_i - A_i^T x)[-A_i^T\mathbf{r}] + R$$

where the remainder term

$$R = 1/2\int_{Y_i - A_i^T x}^{Y_i - A_i^T x - A_i^T\mathbf{r}} (Y_i - A_i^T x - A_i^T\mathbf{r} - t)^2 v_1''(t)dt \leq \frac{1}{2!}[A_i^T\mathbf{r}]^2$$

as $v_1''(t) \leq C$ for all $t \in [Y_i - A_i^T x, Y_i - A_i^T x - A_i^T \mathbf{r}]$ and by (7.43), is of the order $O_P(p(\log p/n))$.

Now, it can be easily shown that for any r_1 and r_2 of distinct points

$$\mathrm{Var}\left(S_n(r_1) - S_n(r_2)\right) \leq \sum_{i=1}^{n} A_{i1}^2 \mathbb{E}\left[\psi(Y_i - A_i^T x - A_i^T \mathbf{r}) - \psi(Y_i - A_i^T x)\right]^2 \tag{7.46}$$

$$\leq K\|r_1 - r_2\|_2^2 \tag{7.47}$$

uniformly in r_1, r_2, for a constant K: $0 < K < \infty$. Also, the boundedness of v_1' we have

$$\mathbb{E}\left[S_n(r_1) - S_n(r_2) - \sum_{i=1}^{n} A_{i1} A_i^T (r_1 - r_2)\right] \tag{7.48}$$

$$\leq \sum_{i=1}^{n} A_{i1}^2 \mathbb{E}\left[\psi(Y_i - A_i^T x - A_i^T \mathbf{r}) - \psi(Y_i - A_i^T x)\right]^2 \tag{7.49}$$

$$\leq K_1\|r_1 - r_2\|_2 \tag{7.50}$$

uniformly in r_1, r_2, for a constant K_1: $0 < K_1 < \infty$. With all of the above we conclude

$$S_n(r_1) - S_n(r_2) - \sum_{i=1}^{n} A_{i1} A_i^T (r_1 - r_2) = O_P(\sqrt{p}). \tag{7.51}$$

To prove the compactness, we shell consider increments of $S_n(r)$ over small blocks. For $r_2 > r_1$, the increments of $S_n(\cdot)$ over the block $B = B(r_1, r_2)$ is

$$S_n(B) = S_n(r_2) - S_n(r_1) = \sum_{i=1}^{n} A_i \psi_i(W_i; B)$$

for $i = 1, \ldots, n$ and

$$\psi_i(W_i; B) = \psi_i(W_i - A_i^T r_2) - \psi_i(W_i - A_i^T r_1).$$

As $A_i^T r_2$ and $A_i^T r_1$ are of the order of $O_P(\sqrt{p \log p/n})$ and ψ is a bounded function, we have $\psi_i(W_i; B) = O_P(\sqrt{p \log p/n})$. Moreover, $\psi_i(W_i; B) = 0$ if any of the arguments lay in the same interval. Hence,

$$\sup\left\{|S_n(B(-K, r_2))| : -K \leq r_2 \leq K\right\} \leq \sum_{i=1}^{n} |A_{i1}| K \left(\|A_i r_1\|_1 + \|A_i r_2\|_1\right) I_i$$

where I_i are independent, non-negative indicator variables with

$$\mathbb{E} I_i \leq K_1 \left(\|A_i r_1\|_1 + \|A_i r_2\|_1\right)$$

for a constant K_1: $0 < K_1 < \infty$. Hence,

$$\mathrm{Var}\left\{\sup\left\{|S_n(B(-K, r_2))| : -K \leq r_2 \leq K\right\}\right\} = O(\sqrt{p \log p/n}).$$

□

The following lemma is a simple modification of Lemma 5.3 of [4]; hence, we omit the proof.

Lemma 10. *Let $S \subseteq [p]$ be measurable on the $\sigma-$algebra σ_t generated by $\{z^0, ..., z^{t-1}\}$ and $\{x^0 + A^T G(z^0, b_0), ..., x^t + A^T G(z^t, b_t)\}$; assume $|S| \leq p(\delta - c)$ for some $c > 0$. Then, there exists $a_1 = a_1(c) > 0$ and $a_2 = a_2(c, t) > 0$, such that $\min_{S'}\{\sigma_{min}(A_{S\cup S'}) : S' \subseteq [p], |S'| \leq a_1 p\} \geq a_2$ with probability converging to 1 as $p \to \infty$.*

We apply this lemma to a specific choice of the set S. Defining

$$v^t \equiv \frac{1}{\theta_{t-1}}(x^{t-1} + A^T G(z^{t-1}, b_{t-1}) - x^t).$$

Lemma 11. *Fix $\gamma \in (0, 1)$ and let $S_t(\gamma) \equiv \{i \in [p] : |v_i^t| \geq 1 - \gamma\}$ for $\gamma \in (0, 1)$. For any $\xi > 0$ there exists $t_*(\xi, \gamma)$ such that for all $t_2 > t_1 > t_*$,*

$$\lim_{p\to\infty} \mathbb{P}\{|S_{t_2} \backslash S_{t_1}| \geq p\xi\} = 0.$$

Proof of the Lemma 11 follows exact steps as Lemma 3.5 in [4]. The change is in the definition of the appropriate set $S_t(\gamma)$.

The Lemma 10 and Lemma 11 imply the following important result.

Lemma 12. *There exist constants $\gamma_1 \in (0, 1)$, $\gamma_2 = a_1(c)/2, \gamma_3 = a_2(c, t_{min}) > 0$ and $t_{min} < \infty$ such that, for any $t \geq t_{min}$,*

$$\min\{\sigma_{min}(A_{S_t(\gamma_1)\cup S'}) : S' \subseteq [p], |S'| \leq \gamma_2 p\} \geq \gamma_3,$$

with probability converging to 1 when $p \to \infty$.

Proof of Lemma 12. Observe that the σ algebra σ_t, contains$\{x^0, ..., x^t\}$ by design of the RAMP algorithms. Therefore, it contains the vector v^t. By Lemma 3, the empirical distribution of $(x_0 - A^T G(z^{t-1}, b_t) - x^{t-1}, x_0)$ converges weakly to $(\bar{\tau}_{t-1} Z, x_0)$. Now we need to check if $S_t(\gamma) \leq p(\delta - c)$

$$\begin{aligned}
\lim_{p\to\infty} \frac{|S_t(\gamma)|}{p} &= \lim_{p\to\infty} \frac{1}{p} \sum_{i=1}^{p} \mathbb{1}_{\{\frac{1}{\theta_{t-1}}|x_i^{t-1} + [A^T G(z^{t-1}, b_{t-1})]_i - x_i^t| \geq 1-\gamma\}} \\
&= \lim_{p\to\infty} \frac{1}{p} \mathbb{1}_{\{\frac{1}{\theta_{t-1}}|x_0 - h^t - \eta(x_0 - h^t, \theta_{t-1})| \geq 1-\gamma\}} \\
&= \mathbb{P}\left\{\frac{1}{\theta_{t-1}}|X_0 + \bar{\tau}_{t-1} Z - \eta(X_0 + \tau_{t-1} Z, \theta_{t-1})| \geq 1 - \gamma\right\}. \quad (7.52)
\end{aligned}$$

Because

$$|X_0 + \bar{\tau}_{t-1} Z - \eta(X_0 + \tau_{t-1} Z, \theta_{t-1})| = \begin{cases} \theta_{t-1} & |X_0 + \bar{\tau}_{t-1} Z| \geq \theta_{t-1} \\ |X_0 + \bar{\tau}_{t-1} Z| & others \end{cases},$$

from the equation (7.52), we conclude

$$\lim_{p\to\infty} \frac{|S_t(\gamma)|}{p} = \mathbb{E}\{\eta'(X_0 + \bar{\tau}_{t-1} Z, \theta_{t-1})\} + \mathbb{P}\left\{(1 - \gamma) \leq \frac{1}{\theta_{t-1}}|X_0 + \bar{\tau}_{t-1} Z| \leq 1\right\}.$$

The fact that $\omega < \delta$, the first term will be strictly smaller than δ for large enough t. And the second term converges to 0. Therefore, we can choose constants $\gamma_1 \to (0,1)$ and $c > 0$ such that

$$\lim_{p\to\infty} \mathbb{P}\{|S_t(\gamma_1)| < p(\delta - c)\} = 1.$$

for all t larger than some t_{min}. For any $t \geq t_{min}$, apply Lemma 10 for some a_1 and a_2. Fix $c>0$ and a_1. Let $\xi = a_1/2$ in Lemma 11, $t_{min} = max(t_{min}, t_*(a_1/2, \gamma_1))$. We have

$$\min\{\sigma_{min}(A_{S_t(\gamma_1)\cup S'}) : S' \subseteq [p], |S'| \leq a_1 p\} \geq a_2,$$

together with $\lim_{p\to\infty} \mathbb{P}\{|S_t \backslash S_{t_{min}}| \geq p a_1/2\} = 0$. □

Proof of Theorem 3. The result of Theorem 3 follows the same arguments as those of [4]; we observe that $||x^{t+1}||_2^2/p$, $||\hat{x}||_2^2/p$ are bounded and that

$$\lim_{p\to\infty} \frac{1}{p} \sum_{i=1}^{p} \psi(\hat{x}_i, x_{0,i}) = \lim_{t\to\infty} \lim_{p\to\infty} \sum_{i=1}^{p} \psi(x_i^{t+1}, x_{0,i}).$$

By Theorem 2, we have $||x^{t+1}||_2^2/p$ is bounded. Moreover, an upper bound on $||\hat{x}||_2^2/p$ is guaranteed by the conditions. □

Acknowledgements

The author would like to thank the Editor, Associate Editor and two reviewers for their helpful comments and questions and Jiao Chen for helpful discussion and help with implementation.

References

[1] Avella Medina, M. A., and Ronchetti, E. (2014). Robust and consistent variable selection for generalized linear and additive models. (310). Retrieved from http://archive-ouverte.unige.ch/unige:36961

[2] Bai, Z. D. and Yin,Y. Q. (1993) Limit of the Smallest Eigenvalue of a Large Dimensional Sample Covariance Matrix, *The Annals of Probability*, **21**, 1275–1294. MR1235416

[3] Bayati, M. and Montanari, A. (2011), The dynamics of message passing on dense graphs, with applications to compressed sensing, *IEEE Trans. on Inform. Theory*, **57** (2), 764–785. MR2810285

[4] Bayati, M. and Montanari, A. (2012), The LASSO risk for Gaussian matrices, *IEEE Trans. on Inform. Theory*, **58** (4), 1997–2017. MR2951312

[5] Bean, D., Bickel, P.J., Karoui, N.E. and Yu B. (2013), Optimal M-estimation in highdimensional regression, *Proceedings of the National Academy of Sciences*, **110** (36), 14563–14568.

[6] Beck, A. and Teboulle, M. (2009), A Fast Iterative Shrinkage-Thresholding Algorithm for Linear Inverse Problems, *SIAM Journal on Imaging Sciences*, **2** (1), 183–202. MR2486527

[7] Belloni, A. and Chernozhukov, V. (2011), l_1-penalized quantile regression in high-dimensional sparse models, *The Annals of Statistics*, **39** (1), 82–130. MR2797841

[8] Bertsekas, D.P. and Tsitsiklis, J.N. (1999), Gradient Convergence in Gradient methods with Errors. *SIAM J. on Optimization*, **10** (3), 627–642. MR1741189

[9] Bickel, P.J. (1975), One-step Huber estimates in the linear model. *Journal of the American Statistical Association*, **70** (350), 428–434. MR0386168

[10] Bickel, P.J., Ritov, Y. and Tsybakov, A.B. (2009), Simultaneous analysis of Lasso and Dantzig selector, *The Annals of Statistics*, **37** (4), 1705–1732. MR2533469

[11] Box, G.E.P. (1953), Non-normality and tests on variances. *Biometrika*, **40** (3–4), 318–335. MR0058937

[12] Box, G.E.P. and Andersen, S.L. (1955), Permutation Theory in the Derivation of Robust Criteria and the Study of Departures from Assumption, *Journal of the Royal Statistical Society. Series B (Methodological)* **17** (1), 1–34.

[13] Boucheron, S. and Lugosi, G. and Massart, P. (2013), Concentration Inequalities: A nonasymptotic theory of independence, Oxford University Press, Oxford, 481. MR3185193

[14] Bradic, J., Fan, J. and Wang, W. (2011), Penalized composite quasi-likelihood for ultrahigh dimensional variable selection. *Journal of the Royal Statistical Society: Series B*, **73** (3), 325-349. MR2815779

[15] Bühlmann, P. and van de Geer, S. (2011), Statistics for High-Dimensional Data: Methods, Theory and Applications, Springer Series in Statistics, 550. MR2807761

[16] Chen Z., Tang M.-L., Wei Gao and Shi N.-Z. (2014), New Robust Variable Selection Methods for Linear Regression Models, *Scandinavian Journal of Statistics*, **41**, 725–741. MR3249425

[17] Donoho, D. L. and Liu, R. C. (1988), The "Automatic" Robustness of Minimum Distance Functionals. *Ann. Statist.*, **16** (2), 552–586. MR0947562

[18] Donoho, D., Maleki, A. and Montanari, A.(2010), The Noise-Sensitivity Phase Transition in Compressed Sensing, *IEEE Trans. on Inform. Theory*, **57** (10), 6920–6941. MR2882271

[19] Donoho, D., Maleki, A. and Montanari, A. (2010) Message passing algorithms for compressed sensing: I. motivation and construction, *Information Theory* (ITW 2010, Cairo), 2010 IEEE Information Theory Workshop on, vol. 1, no. 5, pp. 6–8.

[20] Donoho, D. and Montanari, A. (2013), High Dimensional Robust M-Estimation: Asymptotic Variance via Approximate Message Passing, http://arxiv.org/pdf/1310.7320v3.pdf MR3568043

[21] Fan, J. and Li, R. (2001), Variable selection via non concave penalized

likelihood and its oracle properties, *Journal of the American Statistical Association*, **96** (456), 1348–1360. MR1946581

[22] Fan, J., Fan, Y. and Barut, E. (2014), Adaptive Robust Variable Selection, *The Annals of Statistics*, **42** (1), 324–351. MR3189488

[23] Fan, J., Li, Q. and Wang, Y. (2014), Robust Estimation of High-Dimensional Mean Regression, http://arxiv.org/pdf/1410.2150v1.pdf

[24] Hampel, F.R. (1968), Contributions to the Theory of Robust Estimation, Ph.D. Thesis, University of California, Berkeley. MR2617979

[25] Huber, P.J. (1964), Robust estimation of a location parameter, *Ann. Math. Statist.*, **35**, 73–101. MR0161415

[26] Huber, P.J. (1973), Robust regression: Asymptotics, conjectures and Monte Carlo, *Annals of Statistics*, **1** (5), 799–821. MR0356373

[27] Huber, P. (1981), Robust statistics, Wiley, J. & InterScience, New York. MR0606374

[28] Jurečkova, J. and Sen, P.K. (1996), Robust Statistical Procedures: Asymptotics and Interrelations, Wiley Series in Probability and statistics, New York. MR1387346

[29] Karoui, N. (2013), Asymptotic behavior of unregularized and ridge-regularized high-dimensional robust regression estimators: rigorous results, http://arxiv.org/pdf/1311.2445v1.pdf

[30] Lambert-Lacroix, S. and Zwald, L. (2011), Robust regression through the Huber's criterion and adaptive lasso penalty, *Electron. J. Statist*, **5**, 1015–1053. MR2836768

[31] Lerman, G., McCoy, M., Tropp, J.A. and Zhang, T. (2015), Robust computation of linear models via convex relaxation, *Found. Comput. Math*, **15** (2), 363–410 . MR3320929

[32] Loh, P.-L. (2015), Statistical consistency and asymptotic normality for high-dimensional robust M-estimators, http://arxiv.org/pdf/arXiv:1501.00312

[33] Iusem, A.N. and Teboulle, M. (1995), Convergence Rate Analysis of Nonquadratic Proximal Methods for Convex and Linear Programming, *Mathematics of Operations Research*, **20** (3), 657–677 MR1354775

[34] Mammen, E. (1989), Asymptotics with increasing dimension for robust regression with applications to the bootstrap, *Annals of Statistics*, **17** (1), 382–400. MR0981457

[35] Maronna, R.A., Martin, R.D. and Yohai, V.J. (2006), Robust Statistics: Theory and Methods, J. Wiley. MR2238141

[36] Negahban, S., Ravikumar, P., Wainwright, M.J. and Yu B. (2012), A unified framework for high-dimensional analysis of M-estimators with decomposable regularizers. *Statistical Science*, **27** (4), 538–557. MR3025133

[37] Portnoy. S. (1985), Asymptotic behavior of M estimators of p regression parameters when p^2/n is large; II. Normal approximation. *Annals of Statistics*, **13** (4), 1403–1417. MR0811499

[38] Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society, B,* **58**, 267–288. MR1379242

[39] Tukey, J.W. (1960), A survey of sampling from contaminated distributions.

In: Contributions to Prob. and Statist. (Olkin, I., Ed.), Stanford Univ. Press, Stanford, 448–485. MR0120720

[40] Rousseeuw, P.J. (1984), Least Median of Squares Regression, *Journal of the American Statistical Association*, **79** (388), 871–880. MR0770281

[41] Wang, L. (2013), L_1 penalized LAD estimator for high dimensional linear regression, *Journal of Multivariate Analysis*, **120**, 135–151. MR3072722

[42] Wang, X., and YJiang, M., Mian Huang, M. and Heping Zhang, H. (2013), Robust Variable Selection With Exponential Squared Loss, *Journal of the American Statistical Association*, **108** (502), 632–643 MR3174647

[43] Wu, Y. and Liu, Y. (2009), Variable selection in quantile regression, *Statistica Sinica*, **19**, 801–817. MR2514189

[44] Yohai, V.J. (1987), High breakdown-point and high efficiency robust estimates for regression. *Annals of Statistics*, **15** (2), 642–656. MR0888431

[45] Zhang, C.-H. (2010), Nearly unbiased variable selection under minimax concave penalty. *Ann. Statist.*, **38** (2), 894–942. MR2604701

[46] Zheng, L., Maleki, A., Wang, X., and Long, T. (2015), Does ℓ_p-minimization outperform ℓ_1-minimization?, `http://arxiv.org/pdf/1501.03704v1.pdf`

[47] Zou, H. and Hastie, T. (2005), Regularization and variable selection via the elastic net *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **67** (2), 301–320. MR2137327

[48] Zou, H. (2006), The Adaptive Lasso and Its Oracle Properties *Journal of the American Statistical Association*, **101** (476), 1418–1429. MR2279469