*The Annals of Statistics*
2017, Vol. 45, No. 6, 2427–2454
https://doi.org/10.1214/16-AOS1529


# CONSISTENT PARAMETER ESTIMATION FOR LASSO AND APPROXIMATE MESSAGE PASSING

BY ALI MOUSAVI*, ARIAN MALEKI† AND RICHARD G. BARANIUK*

*Rice University** *and Columbia University*†

This paper studies the optimal tuning of the regularization parameter in LASSO or the threshold parameters in approximate message passing (AMP). Considering a model in which the design matrix and noise are zero-mean i.i.d. Gaussian, we propose a data-driven approach for estimating the regularization parameter of LASSO and the threshold parameters in AMP. Our estimates are consistent, that is, they converge to their asymptotically optimal values in probability as $n$, the number of observations, and $p$, the ambient dimension of the sparse vector, grow to infinity, while $n/p$ converges to a fixed number $\delta$. As a byproduct of our analysis, we will shed light on the asymptotic properties of the solution paths of LASSO and AMP.

## 1. Introduction.

1.1. *Motivation*. Consider the problem of estimating a vector $\beta_o \in \mathbb{R}^p$ from a set of undersampled random linear measurements $y = X\beta_o + w$, where $X \in \mathbb{R}^{n\times p}$ is the design matrix and $w \in \mathbb{R}^n$ denotes noise. One of the successful recovery algorithms, called the LASSO [12, 33], employs the following optimization problem to obtain an estimate of $\beta_o$:

$$\hat{\beta}^{\lambda} = \arg\min_{\beta} \frac{1}{2}\|y - X\beta\|_2^2 + \lambda\|\beta\|_1. \tag{1}$$

A rich literature has provided a detailed analysis of this algorithm [1, 3, 4, 6–9, 13, 14, 17–19, 23–27, 30, 31, 34–39, 41] in nonasymptotic and asymptotic regimes. The nonasymptotic studies consider $p$ and $n$ to be large but finite numbers and characterize the reconstruction error as a function of $p$ and $n$. These analyses provide qualitative guidelines on how to design compressive sensing (CS) and machine learning systems. However, they suffer from loose constants and are incapable of providing quantitative recommendations. Therefore, inspired by the seminal work of Donoho and Tanner [17], researchers have started performing asymptotic analyses of LASSO. In addition to providing sharp quantitative guidelines, these studies have led to new recovery algorithms such as *Approximate Message*

Received November 2015; revised November 2016.
*MSC2010 subject classifications.* 62G05, 62J05.
*Key words and phrases.* LASSO, estimation, sparsity, approximate message passing.

*Passing* (AMP) [15]. AMP finds the solution of LASSO through the following inexpensive iterations:

$$\begin{aligned} \beta^{t+1} &= \eta(\beta^t + X^* z^t; \tau^t), \\ z^t &= y - X\beta^t + \frac{|I^t|}{n} z^{t-1}. \end{aligned} \tag{2}$$

Here, $t$ is the index of iteration, $\beta^t$ is the estimate of $\beta_o$ at iteration $t$ and $I^t \triangleq \{i : \beta_i^t \neq 0\}$. $\eta$ is the soft thresholding function applied component-wise to the elements of the vector; for $a \in \mathbb{R}$, $\eta(a; \tau) \triangleq (|a| - \tau)_+ \operatorname{sign}(a)$. $\tau^t$ is called the threshold parameter.

Despite significant progress in the theoretical analysis of the estimates of LASSO and AMP, little is known about the practically important problem of the optimal tuning of the regularization parameter in LASSO or the threshold parameters in AMP. Our main objective in this paper is to study this problem under the assumption $X_{ij} \overset{i.i.d.}{\sim} N(0, 1/n)$ and $w_i \overset{\text{i.i.d.}}{\sim} N(0, \sigma_w^2)$. We propose a data-driven tuning scheme whose estimates converge to the optimal values of $\lambda$ for LASSO and $\tau^1, \tau^2, \ldots$ for AMP under the asymptotic setting $n \to \infty$, $p \to \infty$, while $n/p$ converges to a fixed number $\delta$. As a byproduct of our analysis, several intriguing asymptotic features of the solution paths of LASSO and AMP, such as the quasi-convexity of the mean square error of LASSO in terms of $\lambda$, will be discovered. Note that in certain applications, the i.i.d. model for $X$ is not appropriate, and hence our results cannot be applied.

1.2. *Related work on parameter tuning.* Both the tuning of the regularization parameter of LASSO and the threshold parameters of AMP have been studied in the literature. The proposed methods fall into the following three different categories:

(i) General model selection ideas such as cross validation are probably the most popular approach in practice [11]. For a review of these schemes, see Chapter 7 of [21]. While very useful in applications, these ideas have their own limitations. We summarize some of their limitations in the context of our paper: (i) AMP has many free parameters (the threshold parameters at every iteration) and if we blindly apply these techniques their estimate of the risk will suffer from high variance and will lead to poor estimates of the threshold parameters. (ii) There are very few papers that have studied the accuracy of these generic model selection techniques in the high dimensional settings. For the case of LASSO, there has been a few papers tackling this issue [11, 22]. These two papers have analyzed the performance of the cross validation (or methods inspired by cross validation) in the regime where $p$ and $n$ are both large but finite. Their results suffer from limitations similar to the ones we discussed in Section 1.1. In this paper, we employ one of the standard model selection techniques, namely Stein unbiased risk estimate

(SURE), in our asymptotic framework and show how the properties of the solution path of AMP and LASSO enable us to not only obtain an efficient parameter tuning scheme, but also prove the consistency of these schemes under the asymptotic setting.

(ii) The second approach employs the upper bounds derived in the literature on the risk of estimators, such as LASSO. For instance, [5, 10] suggest the regularization parameter $\lambda$ to be in the form of $c\sigma_w\sqrt{\log p}$ (when the sparsity level of $\beta_o$ is much smaller than $p$), where $c$ is a fixed number that does not depend on the dimension of the problem or $\sigma_w$. Such approaches suffer from the following limitations: (i) They are usually based on the minimax principle, and hence might be considered as a pessimistic approach for tuning. (ii) The constants of these calculations are loose (even though the bounds are usually order-optimal), and hence a tuning that is based on such bounds does not lead to good performance in practice.

(iii) The third approach, which is the closest to our paper, is based on asymptotic analyses of recovery algorithms. These methods employ asymptotic settings to obtain an accurate estimate of the reconstruction error of a recovery algorithm. Then this analysis is employed to obtain the optimal value of the parameters [15, 16]. The main drawback of this approach is that the signal model is assumed to be known. Since an accurate signal model is not available in practice, the least favorable signals are considered in the analyses which result in a pessimistic tuning of the parameters. Our tuning approach is data-driven and adapts itself to the statistics of the signal.

## 2. Asymptotic framework.

In this section, we review the asymptotic framework under which we study LASSO and AMP. Furthermore, we review some of the existing results that will be used later in our analysis.

### 2.1. *Notation*.

Capital letters denote both matrices and random variables. We sometimes denote $\beta$ with $\beta(p)$ to emphasize its dependency on the ambient dimension. For a matrix $X$, $X^*$, $\sigma_{\min}(X)$ and $\sigma_{\max}(X)$ denote the transpose of $X$, the minimum and the maximum singular values, respectively. Calligraphic letters such as $\mathcal{A}$ denote sets. For a vector $\beta \in \mathbb{R}^p$, $\beta_i$, $\|\beta\|_q \triangleq (\sum |\beta_i|^q)^{1/q}$ and $\|\beta\|_0 = |\{i : |\beta_i| \neq 0\}|$ represent the $i$th component, $\ell_q$ and $\ell_0$ norms, respectively. The notation $\mathbb{E}_B$ denotes the expected value with respect to the randomness in the random variable $B$. The two functions $\phi$ and $\Phi$ denote the probability density function and cumulative distribution function of the standard normal distribution. We will also use notation $\xrightarrow{p}$ and $\xrightarrow{\text{a.s.}}$ for the convergence in probability and almost sure, respectively. Finally, $\mathbb{I}(\cdot)$ denotes the indicator function.

### 2.2. *LASSO in the asymptotic framework*.

In this paper, we analyze the properties of the solution of LASSO and AMP when (i) the measurement matrix has

i.i.d. $N(0, 1/n)$ entries,[1] (ii) $w$ has i.i.d. $N(0, \sigma_w^2)$ elements and (iii) the ambient dimension and the number of measurements are large. Here is the formal definition of this framework [4, 16]: Let $n, p \to \infty$ while $\delta = \frac{n}{p}$ is fixed. We write the vectors and matrices as $\beta_o(p)$, $X(p)$, $y(p)$, and $w(p)$ to emphasize on the ambient dimension of the problem. Clearly, the number of rows of the matrix $X$ is equal to $\delta p$, but we assume that $\delta$ is fixed and, therefore, we do not include $n$ in our notation for $X$. The same argument is applied to $y(p)$ and $w(p)$. Now we define a specific type of a sequence.

DEFINITION 2.1. A sequence of instances $\{\beta_o(p), X(p), w(p)\}$ is called a standard converging sequence if the following conditions hold: (i) The empirical distribution of $\beta_o(p) \in \mathbb{R}^p$ converges weakly to a probability measure $p_\beta$ with bounded second moment. Further, $\frac{1}{p}\|\beta_o(p)\|_2^2$ converges to the second moment of $p_\beta$. (ii) The elements of $w$ are i.i.d. $N(0, \sigma_w^2)$. (iii) The elements of $X(p)$ are i.i.d. $N(0, \frac{1}{n})$.

For the purposes of this paper, $p_\beta$ is not necessarily a sparsity promoting prior. For each problem instance $\beta_o(p)$, $X(p)$ and $w(p)$, we solve the LASSO and obtain $\hat{\beta}^\lambda(p)$ as the estimate. We now evaluate certain measures of performance for this estimate, such as the MSE. The following theorem, conjectured in [16] and proved in [3], plays a pivotal role in our analysis.

THEOREM 2.2. *Consider a standard converging sequence* $\{\beta_o(p), X(p), w(p)\}$. *Suppose that* $\hat{\beta}^\lambda(p)$ *is the solution of the LASSO problem. Then for any pseudo-Lipschitz function*[2] $\psi : \mathbb{R}^2 \to \mathbb{R}$, *almost surely*

$$\lim_{p\to\infty} \frac{1}{p} \sum_i \psi\big(\hat{\beta}_i^\lambda(p), \beta_{o,i}(p)\big) = \mathbb{E}_{B,W}\big[\psi\big(\eta(B + \hat{\sigma} W; \chi\hat{\sigma}), B\big)\big], \tag{3}$$

*where* $B$ *and* $W$ *are two independent random variables with distributions* $p_\beta$ *and* $N(0, 1)$, *respectively.* $\eta$ *is the soft thresholding operator, and* $\hat{\sigma}$ *and* $\chi$ *satisfy the following equations with* $\sigma_w$ *being the variance of the input noise*:

$$\hat{\sigma}^2 = \sigma_w^2 + \frac{1}{\delta}\mathbb{E}_{B,W}\big[\big(\eta(B + \hat{\sigma} W; \chi\hat{\sigma}) - B\big)^2\big], \tag{4}$$

$$\lambda = \chi\hat{\sigma}\bigg(1 - \frac{1}{\delta}\mathbb{P}\big(|B + \hat{\sigma} W| > \chi\hat{\sigma}\big)\bigg). \tag{5}$$

[1] With the recent advances in high dimensional statistics [2], our results can be easily extended to sub-Gaussian matrices. However, for notational simplicity we focus on the Gaussian setting here.

[2] A function $\psi : \mathbb{R}^2 \to \mathbb{R}$ is pseudo-Lipschitz of order $k$ if there exists a constant $L > 0$ such that for all $x, y \in \mathbb{R}^2$ we have $|\psi(x) - \psi(y)| \le L(1 + \|x\|_2 + \|y\|_2)^k \|x - y\|_2$.

Theorem 2.2 will provide the first step in our analysis of the LASSO's solution path. Before we proceed to the implications of this theorem, let us explain some of its interesting features. Suppose that $\hat{\beta}^{\lambda}$ has i.i.d. elements, and each element is in law equal to $\eta(B+\hat{\sigma}W;\chi\hat{\sigma})$, where $B\sim p_{\beta}$ and $W\sim N(0,1)$. Also, assume that $\beta_{o,i}\overset{\text{i.i.d.}}{\sim} p_{\beta}$. If these two assumptions were true, then we could use strong law of large numbers (SLLN) and argue that (3) were true under some mild conditions (as required for the SLLN). While this heuristic is not quite correct, and the elements of $\hat{\beta}_i^{\lambda}$ are not necessarily independent, at the level of calculating $\lim_{p\to\infty}\frac{1}{p}\sum_{i=1}^{N}\psi(\beta_{o,i}(p),\hat{\beta}_i^{\lambda}(p))$ ($\psi$ being pseudo-Lipschitz) this theorem confirms the heuristic. Note that the key elements that have led to this heuristic is the randomness in the $X$ and the large size of the problem.

REMARK 2.3. We are also interested in $\lim_{p\to\infty}\frac{\|\hat{\beta}_{\lambda}\|_0}{p}$ that is $\lim_{p\to\infty}\frac{1}{p}\sum_{i=1}^{N}\psi(\beta_{o,i}(p),\hat{\beta}_i^{\lambda}(p))$ when $\psi(u,v)=\mathbb{I}(v\neq 0)$. However, the $\psi$ function is not pseudo-Lipschitz, and hence Theorem 2.2 does not apply. However, as conjectured in [16] and proved in [3], we can still claim that if $\hat{\beta}^{\lambda}(p)$ denotes the sequence of solutions of the LASSO problem for a standard converging sequence of instances $\{\beta_o(p),X(p),w(p)\}$, then

$$\lim_{p\to\infty}\frac{1}{p}\sum_{i}\mathbb{I}\big(\hat{\beta}_i^{\lambda}(p)\neq 0\big)=\mathbb{P}\big(\big|\eta(B+\hat{\sigma}W;\chi\hat{\sigma})\big|>0\big),$$

where $\chi$, $\tau$ and $\hat{\sigma}$ satisfy (4) and (5).

2.3. *AMP in the asymptotic setting*. In this section, we review some background on the asymptotic analysis of AMP. This section is mainly based on the results in [4, 15, 16], and the interested reader is referred to these papers for further details. The following result originally conjectured in [15, 16] and finally proved in [4], helps us characterize different discrepancy measures for the AMP estimates.

THEOREM 2.4. *Consider a standard converging sequence* $\{\beta_o(p),X(p),w(p)\}$. *Suppose that* $\beta^t(p)$ *is the estimate of AMP at iteration* $t$. *Then for any pseudo-Lipschitz function* $\psi:\mathbb{R}^2\to\mathbb{R}$ *we have*

$$\lim_{p\to\infty}\frac{1}{p}\sum_{i}\psi\big(\beta_i^t(p),\beta_{o,i}(p)\big)=E_{B,W}\big[\psi\big(\eta\big(B+\sigma^tW;\tau^t\big),B\big)\big],$$

*almost surely where* $B$ *and* $W$ *are two independent random variables with distributions* $p_{\beta}$ *and* $N(0,1)$, *respectively. Furthermore, starting with* $(\sigma^0)^2=\mathbb{E}[B^2]/\delta$, $\sigma^t$ *satisfies*

$$\big(\sigma^{t+1}\big)^2=\sigma_w^2+\frac{1}{\delta}\mathbb{E}_{B,W}\big[\big(\eta\big(B+\sigma^tW;\tau^t\big)-B\big)^2\big]. \tag{6}$$

Equation (6) is known as the *state evolution (SE)* for AMP, and $\sigma^t$ is called the state of AMP at iteration $t$. Similar to the discussion for LASSO in Section 2.2, we can establish that almost surely

$$\lim_{p\to\infty} \frac{\|\beta^t(p)\|_0}{p} = \mathbb{P}\big(|B + \sigma^t W| \geq \tau^t\big), \tag{7}$$

even though $\psi(u, v) = I(v \neq 0)$ is not a pseudo-Lipschitz function [4].

One major feature of AMP that will be employed in this paper is that if we set $\tau^t$ "appropriately," then the fixed point of AMP corresponds to the solution of LASSO in the asymptotic regime. One such choice of parameters is the fixed false alarm threshold given by $\tau^t = \chi\sigma^t$, where $\sigma^t$ satisfies (6) and $\chi$ is a fixed number. The following result, conjectured in [15, 16] and later proved in [3] formalizes this statement.

THEOREM 2.5 ([3]). *Consider a standard converging sequence* $\{\beta_o(p), X(p), w(p)\}$. *Let* $\beta^t(p)$ *be the estimate of the AMP algorithm with parameter* $\tau^t = \chi\sigma^t$, *where* $\sigma^t$ *satisfies* (6). *Assume that* $\lim_{t\to\infty}(\sigma^t)^2 = \hat{\sigma}^2$. *Finally, let* $\hat{\beta}^\lambda$ *denote the solution of the LASSO with parameter* $\lambda$ *that satisfies* $\lambda = \chi\hat{\sigma}(1 - \mathbb{P}(|B + \hat{\sigma}W| \geq \chi\hat{\sigma}))$. *Then, almost surely*

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

**3. Main contributions.** We start this section with our contributions regarding AMP. We will then use these results to explain our approach for consistently tuning the regularization parameter of LASSO.

3.1. *Solution path and optimal tuning of AMP.* This section summarizes our contributions on the approximate message passing algorithm.

3.1.1. *Solution path of AMP.* The parameters $\tau^1, \tau^2, \ldots$ have a major impact on both the final reconstruction error, $\lim_{t\to\infty}\|\beta^t - \beta_o\|_2^2/p$ and the convergence rate of AMP to its final solution. Ideally, one would like to select the parameters in a way that the final reconstruction error is the smallest, and at the same time the algorithm converges to this solution at the fastest achievable rate, that is, if we stop the algorithm after $T$ iterations, the estimate it returns is the best possible estimate for $T$ iterations of AMP. There are two main challenges here: (i) It is not clear if these two criteria can be satisfied simultaneously. (ii) To obtain the minimum of $\lim_{p\to\infty}\|\beta^t - \beta_o\|_2^2/p$, we have to solve a computationally demanding optimization problem on $\tau^1, \tau^2, \ldots, \tau^t$.

To address these two challenges, we study the solution path of AMP in terms of the parameters $\tau^1, \ldots, \tau^t$. We will show that under the asymptotic settings described in the previous section, achieving the fastest convergence rate at every iteration is equivalent to achieving the minimum of $\lim_{t\to\infty}\lim_{p\to\infty}\|\beta^t - \beta_o\|_2^2/p$.

Furthermore, we will prove that the optimization of $\tau^1, \tau^2, \ldots, \tau^t$ does not need to be done jointly. Below we formalize these statements. We start with the definition of the optimal threshold parameters of AMP. We overload the notation $\sigma^t(\tau^1, \tau^2, \ldots, \tau^{t-1})$ to emphasize on the fact that the state of AMP at iteration $t$ depends on all the parameters $\tau^1, \tau^2, \ldots, \tau^{t-1}$.

DEFINITION 3.1. A sequence of threshold parameters $\tau^{*,1}, \tau^{*,2}, \ldots, \tau^{*,T-1}$ is called asymptotically optimal for iteration $T$, if and only if

$$\sigma^T\big(\tau^{*,1}, \ldots, \tau^{*,T-1}\big) \le \sigma^T\big(\tau^1, \tau^2, \ldots, \tau^{T-1}\big), \qquad \forall \tau^1, \tau^2, \ldots, \tau^{T-1} \in [0, \infty)^{T-1}.$$

Note that in the above definition we have assumed that the optimal value of $\sigma^T$ is achieved by $(\tau^{*,1}, \ldots, \tau^{*,T-1})$. This assumption is violated for the case $\beta_o = 0$. While we can generalize the definition to include this case, for notational simplicity we skip this special case.

REMARK 3.2. According to Theorem 2.4, we have $\lim_{p\to\infty} \frac{1}{p}\|\beta^t - \beta_o\|_2^2 = \mathbb{E}_{B,W}[\eta(B + \sigma^t W; \tau^t) - B]^2 = \delta((\sigma^{t+1})^2 - \sigma_w^2)$, almost surely. Hence, the optimal parameters, introduced in Definition 3.1, minimize the asymptotic MSE, also.

According to Definition 3.1, it seems that in order to tune AMP optimally, we need to know the number of iterations $T$ we plan to run it (i.e., usually not known in practice) and then perform a joint optimization over the parameters $\tau^1, \tau^2, \ldots, \tau^{T-1}$ (i.e., computationally infeasible). The following theorem resolves both issues.

THEOREM 3.3. *Let $\tau^{*,1}, \tau^{*,2}, \ldots, \tau^{*,T-1}$ be asymptotically optimal for iteration $T$. Then, $\tau^{*,1}, \tau^{*,2}, \ldots, \tau^{*,t-1}$ are asymptotically optimal for any iteration $t < T$.*

See the proof of this Theorem in Section 4.1. An intriguing implication of this result is that the sequence $\tau^{*,1}, \tau^{*,2}, \ldots$ achieves not only the minimum MSE as $t \to \infty$, but also the fastest convergence rate toward the final solution. Note that Theorem 3.3 provides the first simplification for the tuning of the threshold parameters; if $\tau^{*,1}, \tau^{*,2}, \ldots, \tau^{*,t-1}$ are optimally tuned for iteration $t-1$, then $\tau^{*,t}$ minimizes

$$R_B\big(\sigma^t, \tau^t; p_\beta\big) \triangleq \mathbb{E}\big(\eta\big(B_o + \sigma^t W; \tau^t\big) - B_o\big)^2. \tag{8}$$

In other words, the greedy method that finds the threshold parameter that minimizes the asymptotic mean square error of the next iteration only (assuming that AMP will stop after that iteration) leads to the optimal threshold parameters, defined in Definition 3.1. Before we proceed further, we discuss an important property of $R_B(\sigma^t, \tau^t; p_\beta)$ that will be used later.

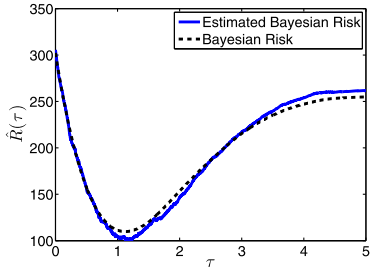


FIG. 1. *The dashed black curve denotes the risk function corresponding to noiseless measurements and the solid blue curve indicates its estimate. For the simulation details, refer to the supplementary file* [29].

DEFINITION 3.4. A quasi-convex function $f : \mathbb{R} \to \mathbb{R}$ is called bowl-shaped if and only if there exists a unique and finite $x_0 \in \mathbb{R}$ at which $f$ achieves its minimum, that is, $f(x_0) \leq f(x),\ \forall x \in \mathbb{R}$.

LEMMA 3.5. *If* $\mathbb{P}(B = 0) < 1$, *then* $R_B(\sigma^t, \tau^t; p_\beta)$ *is a bowl-shaped function of* $\tau^t$. *Furthermore, the derivative of* $R_B(\sigma^t, \tau^t; p_\beta)$ *with respect to* $\tau^t$ *is only zero at the optimal value of* $\tau^t$

This result is similar to Lemma A.5 that is proved in the supplementary material. You may see an example of $R_B(\sigma^t, \tau^t; p_\beta)$ in Figure 1. This lemma confirms that the optimal value of $\tau^t$ exists and is unique. Furthermore, we can use fast convex optimization algorithms such as bisection or gradient descent to find $\tau^{*,t}$. We will clarify this point in the next section.

Despite the success of Theorem 3.3 and Lemma 3.5 in reducing the computational complexity of the optimal tuning of $\tau^1, \ldots, \tau^t$, one major challenge has still remained; The distribution of $B_o$ is not known, and hence $\mathbb{E}(\eta(B_o + \sigma^t W; \tau^t) - B_o)^2$ must be estimated from data. In the next section, we present an asymptotically consistent estimate of this quantity (for a standard converging sequence discussed in Definition 2.1).

3.1.2. *Stein unbiased risk estimate and optimal tuning of AMP.* A major obstacle in using the results of the last section for tuning the threshold parameters of AMP is the lack of knowledge of $p_\beta$ in most applications. Hence, we consider the following estimate of $R_B(\sigma^t, \tau^t; p_\beta)$:

$$\begin{aligned}\hat{R}^t_{h,p}(\tau^t, \tau^{t-1}, \ldots, \tau^1) \triangleq{}& \frac{1}{p}\|\tilde{\eta}_h(\beta^t + X^* z^t; \tau^t) - (\beta^t + X^* z^t)\|_2^2 + (\sigma^t)^2 \\ &+ \frac{2}{p}(\sigma^t)^2[\mathbf{1}^*(\tilde{\eta}'_h(\beta^t + X^* z^t; \tau^t) - \mathbf{1})],\end{aligned} \tag{9}$$

where $\tilde{\eta}_h(u;\tau) = \eta(u;\tau) * \frac{1}{\sqrt{2\pi}h} e^{-\frac{u^2}{2h^2}}$, with $*$ and $h$ denoting the convolution operator and a small number, respectively. The role of this convolution is to smooth out the soft thresholding function. We would like to emphasize three aspects of this risk estimate:

1. The dependence of $\hat{R}^t_{h,p}$ on $\tau^1, \tau^2, \ldots, \tau^{t-1}$ might not be clear from the expression we have written in (9). However, $\beta^t, z^t$ and $\sigma^t$ depend on $\tau^1, \tau^2, \ldots, \tau^{t-1}$.
2. $\hat{R}^t_{h,p}$ is inspired by the Stein unbiased risk estimate (SURE). Since SURE can be applied to any weakly differentiable function, the introduction of the smoothing kernel $\frac{1}{\sqrt{2\pi}h} e^{-\frac{u^2}{2h^2}}$ seems to be unnecessary. Our simulation results agree with this observation, also. However, we require this modification for proving $\mathbb{P}(\sup_{\tau^t} |\hat{R}^t_{h,p}(\tau^t, \ldots, \tau^1) - \mathbb{R}_B(\sigma^t, \tau^t; p_\beta)| > \varepsilon) \to 0$. This uniform convergence is the base of the tuning approach we propose in this section and is proved in Section 4.2.2. Hence, the introduction of $h$ might be unnecessary and an artifact of our proof technique.
3. Note that $(\sigma^t)^2$ is employed in $\hat{R}^t_{h,p}(\tau^t, \tau^{t-1}, \ldots, \tau^1)$ despite the fact that it is not known in practice. It is straightforward to use Lemma 1 of [4] to show that $\frac{1}{n}(z^t)^* z^t \to (\sigma^t)^2$, almost surely.[3] Hence, we can replace $(\sigma^t)^2$ in (9) with $\frac{1}{n}(z^t)^* z^t$ and all the discussions of this section will be still valid. However, for notational simplicity we assume that $\sigma^t$ is given.

Let $\mathcal{T}^1, \mathcal{T}^2, \ldots$ denote some known compact intervals in $\mathbb{R}$ such that $\tau^{*,i} \in \mathcal{T}^i$. Combining Theorem 3.3 and the risk estimate, $\hat{R}^t_{h,p}(\tau^t, \tau^{t-1}, \ldots, \tau^1)$, we obtain the following algorithm for tuning the parameters of AMP:

(i) Let $\hat{\tau}^1_{p,h} = \arg\min_{\tau^1 \in \mathcal{T}^1} \hat{R}^1_{h,p}(\tau^1)$.

(ii) Fix, $\tau^1, \tau^2, \ldots, \tau^{t-1}$ to $\hat{\tau}^1_{p,h}, \hat{\tau}^2_{p,h}, \ldots, \hat{\tau}^{t-1}_{p,h}$, and calculate $\beta^t, z^t$, $\hat{R}^t_{h,p}(\tau^t, \hat{\tau}^{t-1}_{p,h}, \ldots, \hat{\tau}^1_{p,h})$, and

$$\hat{\tau}^t_{p,h} \triangleq \arg\min_{\tau^t \in \mathcal{T}^t} \hat{R}^t_{h,p}(\tau^t, \hat{\tau}^{t-1}_{p,h}, \ldots, \hat{\tau}^1_{p,h}). \tag{10}$$

Compared with the original AMP algorithm, the only extra calculation that needs to be done for the optimal tuning is the univariate optimization (10) at each iteration. For the moment, we suppose that we can solve this univariate optimization problem efficiently (grid search can be applied at every iteration, however we will describe more efficient algorithms later). Under this assumption, the following Theorem proves the consistency of $\hat{\tau}^t_{p,h}$.

THEOREM 3.6. *Consider a standard converging sequence* $\{\beta_o(p), X(p), w(p)\}$. *Let* $\tau^{*,t}$ *denote the optimal threshold according to Definition* 3.1. *Then, for any fixed iteration* $t$, $\lim_{h\to 0} \lim_{p\to\infty} \hat{\tau}^t_{p,h} = \tau^{*,t}$ *in probability.*

[3] This estimate has been introduced elsewhere [25].

The proof of this theorem is discussed in Section 4.2. The tuning algorithm we described above can be implemented in practice, but requires an exhaustive search over each $\mathcal{T}^t$. Since $R_B(\sigma^t, \tau^t; p_\beta)$ is a quasi-convex function of $\tau^t$, we can employ a bisection method or gradient descent to reduce the computations further. However, the algorithm has to work with the risk estimate $\hat{R}^t_{h,p}(\tau^t, \hat{\tau}^{t-1}_{p,h}, \dots, \hat{\tau}^1_{p,h})$ that is not necessarily quasi-convex. Hence, the last challenge is to modify these algorithms in a way that they can work on $\hat{R}^t_{h,p}(\tau^t, \hat{\tau}^{t-1}_{p,h}, \dots, \hat{\tau}^1_{p,h})$. Here, we present an approximate bisection algorithm, but the interested reader may also see the performance of an approximate gradient descent algorithm in our unpublished report [28].

We assume that $\mathcal{T}^t = [\underline{\tau}^t, \overline{\tau}^t]$. We select two small numbers $\varepsilon$ and $\Delta$, set $\tau = (\underline{\tau} + \overline{\tau})/2$, and do the following: If $(\hat{R}^t_{p,h}(\tau + \Delta) - \hat{R}^t_{p,h}(\tau))/\Delta < -\varepsilon$, then set $\underline{\tau}^t = \tau$ and repeat the process. If $(\hat{R}^t_{p,h}(\tau + \Delta) - \hat{R}^t_{p,h}(\tau))/\Delta > \varepsilon$, then set $\overline{\tau}^t = \tau$ and repeat the process. Otherwise, stop the process and return $\tau$. This is a slight modification of the bisection method that is popular in optimization. We can analyze the performance of this algorithm under the asymptotic settings.

THEOREM 3.7. *Consider a standard converging sequence* $\{\beta_o(p), X(p), w(p)\}$. *Let* $\hat{\tau}^t_{B,p}$ *denote the value of* $\tau$ *at which our bisection algorithm stops. Then there exists* $\bar{\tau} \in [\hat{\tau}^t_{B,p}, \hat{\tau}^t_{B,p} + \Delta]$ *such that with probability one* $\lim_{h\to 0} \lim_{p\to\infty} |\frac{\partial R_B(\sigma, \bar{\tau}; p_\beta)}{\partial \tau}| < \varepsilon$.

This result is a straightforward application of Theorem 4.4 and is skipped here. We have shown in the supplementary material [29] that (i) The performance of the bisection method is not sensitive to the exact value of $\varepsilon$ and $\Delta$, and (ii) These two parameters are easy to tune. For a discussion on the choice of these parameters and the problem size at which these algorithms work, refer to Section G of the supplementary material.

3.2. *Connection to optimal tuning of* $\lambda$ *in LASSO.* In the last section, we showed how the threshold parameters of AMP can be optimized. In this section, we study a connection between the estimates of the optimally-tuned AMP and the solution of LASSO for the optimal value of $\lambda$. Suppose that we run AMP with the optimal parameters $\tau^{*,1}, \tau^{*,2}, \dots$ defined in Definition 3.1, and obtain $\beta^1_*, \beta^2_*, \dots$.

PROPOSITION 3.8. *Consider a standard converging sequence* $\{\beta_o(p), X(p), w(p)\}$. *Let* $\hat{\beta}^\lambda(p)$ *denote the solution of LASSO with regularization parameter* $\lambda$. *Then* $\lim_{t\to\infty} \lim_{p\to\infty} \frac{1}{p}\|\beta_o(p) - \beta^t_*(p)\|_2^2 = \inf_\lambda \lim_{p\to\infty} \frac{1}{p}\|\hat{\beta}^\lambda(p) - \beta_o(p)\|_2^2$.

The proof of this result can be found in Section B of the supplementary material. This theorem implies that the final solution the optimal AMP converges to, has the

same MSE as the solution of LASSO with the optimal value of the regularization parameter $\lambda$.

In Theorem 3.8, the threshold parameters of AMP are set to $\tau^{*,1}, \tau^{*,2}, \ldots$. In the last section, we showed how a consistent estimates of these parameters can be obtained. Our next theorem proves that the estimates of AMP with such data-dependent thresholds are close to $\beta_*^t(p)$, and hence have similar MSE as the solution of LASSO with the optimal regularization parameter.

PROPOSITION 3.9. *Consider a standard converging sequence* $\{\beta_o(p), X(p), w(p)\}$. *Let* $\hat{\tau}^1, \hat{\tau}^2, \ldots, \hat{\tau}^t$ *denote data-driven threshold parameters that satisfy* $\hat{\tau}^i \to \tau^{*,i}$ *in probability for every* $i \in \{0, 1, 2, \ldots, t\}$. *Let* $\tilde{\beta}^t$ *denote the estimate of AMP with thresholds* $\hat{\tau}^1, \hat{\tau}^2, \ldots, \hat{\tau}^t$. *Then, in probability*

$$\lim_{p\to\infty} \frac{1}{p} \|\tilde{\beta}^t(p) - \beta_*^t(p)\|_2^2 = 0.$$

The proof of this result can be found in Section C of the supplementary material.

3.3. *Solution path and optimal tuning of LASSO.* As discussed in Section 3.2, one may use the optimally-tuned AMP to reach the solution of LASSO with the optimal value of $\lambda$. In this section, we propose a direct method to find the optimal value of $\lambda$ in LASSO. The approach we develop in this section can be used for a wide range of regularizers.

Similar to the previous section, we first review some of the properties of LASSO's solution path that will be used for tuning $\lambda$. The two main problems that we address are: (Q1) How does $\frac{1}{p}\|\hat{\beta}^\lambda\|_0$ change as $\lambda$ varies? (Q2) How does $\frac{1}{p}\|\hat{\beta}^\lambda - \beta_o\|_2^2$ change as $\lambda$ varies? The first question is about the number of active (nonzero) elements in the solution of the LASSO, and the second one is about the mean squared error (MSE). Intuitively speaking, one would expect the size of the active set to shrink as $\lambda$ increases and the mean squared error to be a bowl-shaped function of $\lambda$. Unfortunately, the peculiar behavior of LASSO breaks this intuition. See Figure 2 for a counter-example. This figure exhibits the number of active elements in the solution as we increase the value of $\lambda$. It is clear that the size of the active set is not monotonically decreasing. The details of this simulation are described in the supplementary material [29].

Such pathological examples have discouraged further investigation of these problems in the literature. One of the main objectives of this paper is to show that such examples are quite rare, and if we consider the asymptotic setting (that was described in Section 2.2), then we can provide quite intuitive answers to the two questions raised above. Let us summarize our results here in a nonrigorous way: considering the asymptotic setting with $X_{ij} \sim N(0, 1/n)$ and $w_i \sim N(0, \sigma_w^2)$, (A1) $\frac{1}{p}\|\hat{\beta}^\lambda\|_0$ is a decreasing function of $\lambda$ and (A2) $\frac{1}{p}\|\hat{\beta}^\lambda - \beta_o\|_2^2$ is a quasi-convex function of $\lambda$. These results are formally stated below.

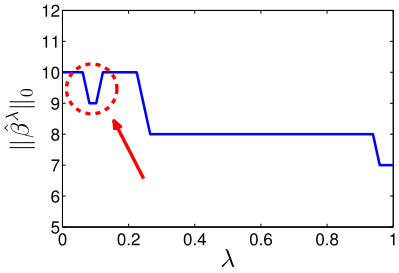


FIG. 2. *The number of active elements in the LASSO's solution as a function of* $\lambda$. *The size of the active set at one location grows as we increase* $\lambda$, *and hence this function does not match the intuition. For the details of this experiment, see the supplementary material* [29].

THEOREM 3.10. *Let* $\{\beta_o(p), X(p), w(p)\}$ *denote a standard converging sequence of problem instances as defined in Definition* 2.1. *If* $\hat{\beta}^{\lambda}(p)$ *is the solution of LASSO with regularization parameter* $\lambda$, *then*

$$\frac{d}{d\lambda}\left(\lim_{p\to\infty}\frac{1}{p}\sum_{i=1}^{p}\mathbb{I}(\hat{\beta}_i^{\lambda}(p)\neq 0)\right)<0.$$

*Furthermore,* $\lim_{p\to\infty}\frac{1}{p}\sum_{i=1}^{p}\mathbb{I}(\beta_i^{\lambda}(p)\neq 0)\leq\delta$ *no matter how we select* $\lambda>0$.

We summarize the proof of this theorem in Section D of the supplementary material. Intuitively, Theorem 3.10 claims that, as we increase the regularization parameter $\lambda$, the number of elements in the active set, that is, $\|\hat{\beta}^{\lambda}\|_0$ is decreasing. Also, according to the condition $\lim_{p\to\infty}\frac{1}{p}\sum_{i=1}^{p}\mathbb{I}(\beta_i^{\lambda}(p)\neq 0)\leq\delta$, the largest it can get is $\delta=n/p$. Note that the fact that $\|\hat{\beta}^{\lambda}\|_0\leq n$ is true even under the nonasymptotic settings. For more information, refer to [20]. Since the number of active elements is a decreasing function of $\lambda$, $\delta$ appears only in the limit $\lambda\to 0$. Figure 3 plots the number of active elements as a function of $\lambda$ for a setting described in Section G.8.2 of supplementary material [29].

Our next result is regarding the behavior of the MSE in terms of the regularization parameter $\lambda$. Figure 4 exhibits the behavior of MSE as a function of $\lambda$. The detailed description of this problem instance can be found in Section G.8.3 of the supplementary material [29].

THEOREM 3.11. *Let* $\{\beta_o(p), X(p), w(p)\}$ *denote a standard converging sequence of problem instances as defined in Definition* 2.1. *If* $\hat{\beta}^{\lambda}(p)$ *is the solution of LASSO with regularization parameter* $\lambda$, *then* $\lim_{p\to\infty}\frac{1}{p}\|\hat{\beta}^{\lambda}(p)-\beta_o(p)\|_2^2$ *is a quasi-convex function of* $\lambda$. *Furthermore, if* $p_{\beta}(B=0)\neq 1$, *then the function is bowl-shaped.*

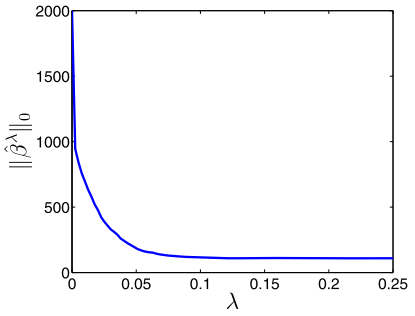


FIG. 3. *The number of active elements in the solution of LASSO as a function of* $\lambda$. *The size of the active set decreases monotonically as we increase* $\lambda$. *See the supplementary material* [29] *for the details.*

For the proof, see Section E of the supplementary material. Our next goal is to show how one can estimate the optimal value of $\lambda$ for LASSO. We start with the definition of the optimal value of $\lambda$.

DEFINITION 3.12. Let $\{\beta_o(p), X(p), w(p)\}$ denote a standard converging sequence of problem instances as defined in Definition 2.1. Also, let $\hat{\beta}^{\lambda}(p)$ be the solution of LASSO with regularization parameter $\lambda$. A regularization parameter $\lambda^*$ is called asymptotically optimal for LASSO if and only if $\lambda^*$ achieves the minimum of the almost sure limit of

$$\lim_{p\to\infty} \frac{\|\hat{\beta}^{\lambda}(p) - \beta_o(p)\|_2^2}{p}.$$

According to Theorem 3.11 if $\beta_o \neq 0$, then $\lambda^*$ exists and is unique.

REMARK 3.13. Since $X_{ij} \sim N(0, \frac{1}{n})$ $\lambda^*$, minimizes both the asymptotic out-of-sample prediction error and the asymptotic mean square error.

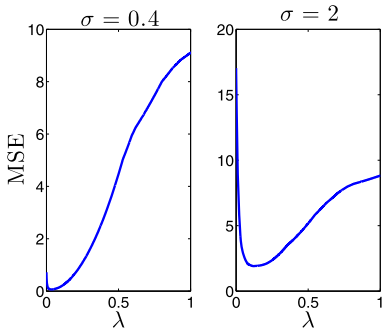


FIG. 4. *Behavior of the MSE as a function of* $\lambda$ *of LASSO for two different noise variances. See the supplementary material* [29] *for details.*

The main obstacle in finding $\lambda^*$ is the estimation of $\lim_{p\to\infty} \frac{1}{p}\|\hat{\beta}^\lambda - \beta_o\|_2^2$. The following theorem plays a pivotal role in estimating this quantity.

THEOREM 3.14. *Consider a standard converging sequence* $\{\beta_o(p), X(p), w(p)\}$. *Let* $\beta^t$ *and* $z^t$ *denote the estimates of AMP with parameter* $\tau^t = \chi\sigma^t$, *where* $\sigma^t$ *satisfies* (6). *Assume that* $\lim_{t\to\infty}(\sigma^t)^2 = \hat{\sigma}^2$, *where* $\hat{\sigma}$ *is a fixed point of* (4). *Let* $\hat{\beta}^\lambda$ *denote the solution of the LASSO with parameter* $\lambda$ *that satisfies* $\lambda = \chi\hat{\sigma}(1 - \mathbb{P}(|B + \hat{\sigma}W| \geq \chi\hat{\sigma}))$. *Then, almost surely*

$$\lim_{t\to\infty}\lim_{p\to\infty}\frac{1}{p}\left\|\beta^t + X^*z^t - \hat{\beta}^\lambda - \frac{X^*(y - X\hat{\beta}^\lambda)}{1 - \frac{\|\hat{\beta}^\lambda\|_0}{n}}\right\|_2^2 = 0.$$

The proof of this theorem is presented in Section F of the supplementary material. It is important to note that the term $\beta^t + X^*z^t$ appears in the estimate of the risk in (9). Hence, from Theorem 3.14 we expect the quantity $\hat{\beta}^\lambda + \frac{X^*(y-X\hat{\beta}^\lambda)}{1-\frac{\|\hat{\beta}^\lambda\|_0}{n}}$ to appear in the estimate of $\lim_{p\to\infty}\frac{\|\hat{\beta}^\lambda - \beta_o\|_2^2}{p}$. The following remarks enable us to construct an estimate of $\lim_{p\to\infty}\frac{1}{p}\|\hat{\beta}^\lambda - \beta_o\|_2^2$:

(i) $\hat{R}_{h,p}(\tau^t, \tau^{t-1}, \ldots, \tau^1)$ converges almost surely to $\mathbb{E}(\eta(B_o + \sigma^t W; \tau^t) - B_o)^2$, which is in turn the almost sure limit of $\|\beta^t - \beta_o\|_2^2/p$.

(ii) According to Theorem 2.5 if $\tau^t = \chi\sigma^t$, then the almost sure limit of $\lim_{t\to\infty}\lim_{p\to\infty}\|\hat{\beta}^\lambda - \beta^t\|_2^2/p$ is zero. This implies that the almost sure limit of $\lim_{p\to\infty}\|\hat{\beta}^\lambda - \beta_o\|_2^2/p$ is equal to the almost sure limit of $\lim_{t\to\infty}\lim_{p\to\infty}\|\beta^t - \beta_o\|_2^2/p$.

If we combine these two facts and Theorem 3.14 with (9), we obtain the following expression as an estimate of $\lim_{p\to\infty}\|\hat{\beta}^\lambda - \beta_o\|_2^2/p$:

$$\begin{aligned}\tilde{r}_{h,p}(\lambda) \triangleq \frac{1}{p}\left\|\tilde{\eta}_h\left(\hat{\beta}^\lambda + \frac{X^*(y - X\hat{\beta}^\lambda)}{1 - \frac{\|\hat{\beta}^\lambda\|_0}{n}}; \chi\hat{\sigma}\right) - \left(\hat{\beta}^\lambda + \frac{X^*(y - X\hat{\beta}^\lambda)}{1 - \frac{\|\hat{\beta}^\lambda\|_0}{n}}\right)\right\|_2^2 \\ + \hat{\sigma}^2 + \frac{2}{p}\hat{\sigma}^2\left[\mathbf{1}^*\left(\tilde{\eta}_h'\left(\hat{\beta}^\lambda + \frac{X^*(y - X\hat{\beta}^\lambda)}{1 - \frac{\|\hat{\beta}^\lambda\|_0}{n}}; \chi\hat{\sigma}\right) - \mathbf{1}\right)\right].\end{aligned} \tag{11}$$

Note that this expression is still not a proper estimator for $\lim_{p\to\infty}\|\hat{\beta}^\lambda - \beta_o\|_2^2/p$ for the following reasons: (i) $\hat{\sigma}$ is not known. (ii) The value of $\chi$ corresponding to $\lambda$ is not known. We address both issues below:

1. Estimating $\hat{\sigma}$: Similar to the proof of Theorem 3.14, we can show that almost surely $\|z^t - \frac{y-X\hat{\beta}^\lambda}{1-\|\hat{\beta}^\lambda\|_0/n}\|_2/\sqrt{p} \to 0$. This is in fact part of the proof of Theorem 3.14. It is straightforward to use Lemma 1 of [4] to prove that

$(z^t)^* z^t/n \to (\sigma^t)^2$. Since $\chi$ is picked such that $\sigma^t \to \hat{\sigma}$, we can use an estimate $\|\frac{y - X\hat{\beta}^\lambda}{1 - \|\hat{\beta}^\lambda\|_0/n}\|_2^2/n$ for $\hat{\sigma}^2$.

2. According to Theorem 3.14, $\chi\hat{\sigma}(1 - \mathbb{P}(|B + \hat{\sigma}W| \geq \chi\hat{\sigma})) = \lambda$. Furthermore, according to Remark 2.3, $\frac{\|\hat{\beta}^\lambda\|_0}{p} \to \mathbb{P}(|B + \hat{\sigma}W| \geq \chi\hat{\sigma})$ almost surely. Hence, we can estimate $\chi\hat{\sigma}$ with $\frac{\lambda}{1 - \frac{\|\hat{\beta}^\lambda\|_0}{p}}$.

In summary, we obtain the following estimate for $\lim_{p\to\infty} \frac{\|\hat{\beta}^\lambda - \beta_o\|_2^2}{p}$:

$$\begin{aligned}\tilde{R}_{h,p}(\lambda) \triangleq \frac{1}{p}\left\|\tilde{\eta}_h\left(\hat{\beta}^\lambda + \frac{X^*(y - X\hat{\beta}^\lambda)}{1 - \frac{\|\hat{\beta}^\lambda\|_0}{n}}; \frac{\lambda}{1 - \frac{\|\hat{\beta}^\lambda\|_0}{p}}\right) - \left(\hat{\beta}^\lambda + \frac{X^*(y - X\hat{\beta}^\lambda)}{1 - \frac{\|\hat{\beta}^\lambda\|_0}{n}}\right)\right\|_2^2 \\ + \hat{\sigma}^2 + \frac{2}{p}\hat{\sigma}^2\left[\mathbf{1}^*\left(\tilde{\eta}_h'\left(\hat{\beta}^\lambda + \frac{X^*(y - X\hat{\beta}^\lambda)}{1 - \frac{\|\hat{\beta}^\lambda\|_0}{n}}; \frac{\lambda}{1 - \frac{\|\hat{\beta}^\lambda\|_0}{p}}\right) - \mathbf{1}\right)\right],\end{aligned} \tag{12}$$

where $\hat{\sigma}^2 = \|\frac{y - X\hat{\beta}^\lambda}{1 - \|\hat{\beta}^\lambda\|_0/n}\|^2/n$. Based on this estimate, we propose the following approach for evaluating $\lambda^*$. Suppose that $\Lambda$ is a compact subset of $\mathbb{R}$ with $\lambda^* \in \Lambda$. Define $\hat{\lambda} = \arg\min_{\lambda\in\Lambda} \tilde{R}_{h,p}(\lambda)$. The following result proves the consistency of $\hat{\lambda}$.

THEOREM 3.15. *Consider a standard converging sequence* $\{\beta_o(p), X(p), w(p)\}$. *Let* $\lambda^*$ *denote the optimal regularization parameter according to Definition* 3.12. *Then* $\lim_{h\to 0}\lim_{p\to\infty}\hat{\lambda} = \lambda^*$ *in probability.*

Since the proof is similar to the proof of Theorem 3.6, we skip it. Note that we do not have to solve the LASSO for many different values of $\lambda$. According to Theorem 3.11, the risk is quasi-convex, and hence other methods such as bisection can help as well. Since the approach is similar to what we discussed for AMP, we do not repeat it here.

## 4. Proofs of the main results.

4.1. *Proof of Theorem* 3.3. We start with the following lemma that will be used in our proofs.

LEMMA 4.1. [32] *If* $g : \mathbb{R} \to \mathbb{R}$ *is a weakly differentiable function and* $W \sim N(0, 1)$, *then*

$$\begin{aligned}\mathbb{E}\big(g(B_o + \sigma W; \tau) - B_o\big)^2 = \mathbb{E}\big(g(B_o + \sigma W; \tau) - B_o - \sigma W\big)^2 + \sigma^2 \\ + 2\sigma^2\mathbb{E}\big(g'(B_o + \sigma W; \tau) - 1\big).\end{aligned}$$

We call this result Stein's lemma in this paper. We first prove one of the main features of the risk function defined in (8).

LEMMA 4.2. *If* $\mathbb{P}(B \neq 0) \neq 0$, *then* $\inf_\tau R_B(\sigma, \tau; p_\beta)$ *is an increasing function of* $\sigma$.

PROOF. First, we prove this lemma for the modified risk function defined as $\bar{R}_B(\sigma, \chi; p_\beta) \triangleq \mathbb{E}_{W,B_o}((\eta(B_o + \sigma W; \sigma\chi) - B_o)^2)$. Then we will switch to the original risk function $R_B(\sigma, \tau; p_\beta)$.

According to Lemma A.3 (in the supplementary material), the risk function $\bar{R}_B(\sigma, \chi; p_\beta)$ is a concave function of $\sigma^2$. Therefore, the derivative of the risk function is a decreasing function of $\sigma^2$. Hence, if we prove that the derivative is positive when $\sigma \to \infty$, then for any fixed $\chi$ the risk function is an increasing function of $\sigma$. Therefore, we first prove that $\lim_{\sigma\to\infty} \frac{\partial \bar{R}_B(\sigma,\chi;p_\beta)}{\partial\sigma^2} \geq 0$:

$$
\begin{aligned}
&\frac{\partial \bar{R}_B(\sigma, \chi; p_\beta)}{\partial \sigma^2} \\
&\quad= \frac{1}{2\sigma}\frac{\partial}{\partial\sigma}\mathbb{E}_{W,B_o}\big(\big(\eta(B_o + \sigma W; \sigma\chi) - B_o\big)^2\big) \\
&\quad= \frac{1}{2\sigma}\mathbb{E}_{W,B_o}\big(2\big(\eta(B_o + \sigma W; \sigma\chi) - B_o\big)\big((W - \chi)\mathbb{I}(B_o + \sigma W > \sigma\chi) \\
&\qquad + (W + \chi)\mathbb{I}(B_o + \sigma W < -\sigma\chi)\big)\big) \\
&\quad\overset{(a)}{=} \mathbb{E}_{W,B_o}\big(\sigma W\delta(B_o + \sigma W - \sigma\chi) + \mathbb{I}(B_o + \sigma W > \sigma\chi) \\
&\qquad - 2\sigma\chi\delta(B_o + \sigma W - \sigma\chi) + \tau^2\mathbb{I}(B_o + \sigma W > \sigma\chi) \\
&\qquad - \sigma W\delta(B_o + \sigma W + \sigma\chi) + \mathbb{I}(B_o + \sigma W < -\sigma\chi) \\
&\qquad + \chi^2\mathbb{I}(B_o + \sigma W < -\sigma\chi) - 2\sigma\chi\delta(B_o + \sigma W + \sigma\chi)\big) \\
&\quad= \mathbb{E}_{B_o}\left(\left(\frac{\sigma\chi - B_o}{\sigma}\right)\phi\left(\frac{\sigma\chi - B_o}{\sigma}\right) + \left(\frac{\sigma\chi + B_o}{\sigma}\right)\phi\left(\frac{\sigma\chi + B_o}{\sigma}\right)\right) \\
&\qquad + \mathbb{E}_{B_o,W}\big((1 + \chi^2)\big(\mathbb{I}(B_o + \sigma W > \sigma\chi) + \mathbb{I}(B_o + \sigma W < -\sigma\chi)\big)\big) \\
&\qquad - \mathbb{E}_{B_o}\left(2\chi\left(\phi\left(\frac{\sigma\chi - B_o}{\sigma}\right) + \phi\left(\frac{\sigma\chi + B_o}{\sigma}\right)\right)\right) \\
&\quad= (1 + \chi^2)\mathbb{E}_{B_o}\left(\Phi\left(\frac{B_o}{\sigma} - \chi\right) + \Phi\left(\frac{-B_o}{\sigma} - \chi\right)\right) \\
&\qquad - \chi\mathbb{E}_{B_o}\left(\phi\left(\frac{B_o}{\sigma} - \chi\right) + \phi\left(\frac{B_o}{\sigma} + \chi\right)\right) \\
&\qquad - \frac{1}{\sigma}\mathbb{E}_{B_o}\left(B_o\left(\phi\left(\frac{B_o}{\sigma} - \chi\right) - \phi\left(\frac{B_o}{\sigma} + \chi\right)\right)\right),
\end{aligned}
\tag{13}
$$

where (a) holds because of the Stein's lemma (Lemma 4.1). As a result of (13), we can write

$$
\begin{aligned}
&\lim_{\sigma\to\infty}\frac{\partial R_B(\sigma,\chi;p_\beta)}{\partial\sigma^2}\\
&\quad=\lim_{\sigma\to\infty}\Bigg((1+\chi^2)\mathbb{E}_{B_o}\bigg(\Phi\bigg(\frac{B_o}{\sigma}-\chi\bigg)+\Phi\bigg(\frac{-B_o}{\sigma}-\chi\bigg)\bigg)\\
&\qquad-\chi\mathbb{E}_{B_o}\bigg(\phi\bigg(\frac{B_o}{\sigma}-\chi\bigg)+\phi\bigg(\frac{B_o}{\sigma}+\chi\bigg)\bigg)\\
&\qquad-\frac{1}{\sigma}\mathbb{E}_{B_o}\bigg(B_o\bigg(\phi\bigg(\frac{B_o}{\sigma}-\chi\bigg)-\phi\bigg(\frac{B_o}{\sigma}+\chi\bigg)\bigg)\bigg)\Bigg)\\
&\quad=2\big(1+\chi^2\big)\Phi(-\chi)-2\chi\phi(\chi)\overset{(b)}{=}2\big(1+\chi^2\big)Q(\chi)-2\chi\phi(\chi)\\
&\quad\overset{(c)}{>}2\chi\phi(\chi)-2\chi\phi(\chi)=0,
\end{aligned}
\tag{14}
$$

where in (b), $Q(\chi)\triangleq\int_\chi^\infty\phi(w)\,dw$ and in (c) we have used the well-known lower-bound for the Q-function $(\frac{\chi}{1+\chi^2})\phi(\chi)<Q(\chi)$. Now, since $\bar{R}_B(\sigma,\chi;p_\beta)=\mathbb{E}_{W,B_o}((\eta(B_o+\sigma W;\sigma\chi)-B_o)^2)$ is an increasing function of $\sigma$ for any fixed $\chi$, if $\sigma_1<\sigma_2$, then we can write

$$
\mathbb{E}_{W,B_o}\big(\eta(B_o+\sigma_1W;\sigma_1\chi)-B_o\big)^2<\mathbb{E}_{W,B_o}\big(\eta(B_o+\sigma_2W;\sigma_2\chi)-B_o\big)^2.
\tag{15}
$$

We can take the infimum from both sides of the inequality in (15) and obtain

$$
\begin{aligned}
&\inf_\chi\mathbb{E}_{W,B_o}\big(\eta(B_o+\sigma_1W;\sigma_1\chi)-B_o\big)^2\\
&\qquad<\inf_\chi\mathbb{E}_{W,B_o}\big(\eta(B_o+\sigma_2W;\sigma_2\chi)-B_o\big)^2.
\end{aligned}
\tag{16}
$$

Since $\mathbb{P}(B\neq0)\neq0$, according to Lemma A.5 (in the supplementary material), $\mathbb{E}_{W,B_o}(\eta(B_o+\sigma_2W;\sigma_2\chi)-B_o)^2$ is a bowl shaped function of $\chi$, and hence, $\inf_\chi\mathbb{E}_{W,B_o}(\eta(B_o+\sigma_2W;\sigma_2\chi)-B_o)^2$ is achieved at a finite value of $\bar{\chi}_2$. According to (15), $\mathbb{E}_{W,B_o}(\eta(B_o+\sigma_1W;\sigma_1\bar{\chi}_2)-B_o)^2$ is strictly smaller than $\inf_\chi\mathbb{E}_{W,B_o}(\eta(B_o+\sigma_2W;\sigma_2\chi)-B_o)^2$, and hence (16) is also correct with strict inequality. Let $\tau=\sigma\chi$ and suppose that $\tau^*=\sigma_2\chi^*$ is the threshold by which the infimum of the right-hand side (RHS) of (16) is achieved. Then we have

$$
\begin{aligned}
&\inf_\tau\mathbb{E}_{W,B_o}\big(\big(\eta(B_o+\sigma_1W;\tau)-B_o\big)^2\big)\\
&\quad=\inf_\chi\mathbb{E}_{W,B_o}\big(\big(\eta(B_o+\sigma_1W;\chi\sigma_1)-B_o\big)^2\big)\\
&\quad\leq\mathbb{E}_{W,B_o}\bigg(\bigg(\eta\bigg(B_o+\sigma_1W;\bigg(\frac{\tau^*}{\sigma_2}\bigg)\sigma_1\bigg)-B_o\bigg)^2\bigg)\\
&\quad<\mathbb{E}_{W,B_o}\big(\big(\eta\big(B_o+\sigma_2W;\tau^*\big)-B_o\big)^2\big)\\
&\quad=\inf_\tau\mathbb{E}_{W,B_o}\big(\big(\eta(B_o+\sigma_2W;\tau)-B_o\big)^2\big),
\end{aligned}
\tag{17}
$$

which means $\inf_\tau R_B(\sigma, \tau; p_\beta) = \mathbb{E}_{W,B_o}((\eta(B_o + \sigma W; \tau) - B_o)^2)$ is an increasing function of $\sigma$. $\square$

Having completed the proof of Lemma 4.2, the proof of Theorem 3.3 is performed using an induction argument. Suppose that $\tau^{*,1}, \tau^{*,2}, \ldots, \tau^{*,T-1}$ is optimal for iteration $T$. Our goal is to show that $\tau^{*,1}, \tau^{*,2}, \ldots, \tau^{*,T-2}$ is optimal for iteration $T-1$ as well. Now we use a contradiction argument. Suppose that $\tau^{*,1}, \tau^{*,2}, \ldots, \tau^{*,T-2}$ is not optimal for iteration $T-1$; then there exists $\tau^1, \tau^2, \ldots, \tau^{T-2}$ such that

$$\sigma^{T-1}(\tau^1, \ldots, \tau^{T-2}) < \sigma^{T-1}(\tau^{*,1}, \tau^{*,2}, \ldots, \tau^{*,T-2}).$$

Define $\tau^{**,T-1}$ as

$$\arg\min_\tau \mathbb{E}_{B_o,W}\big[\big(\eta\big(B_o + \sigma^{T-1}(\tau^1, \ldots, \tau^{T-2})W; \tau\big) - B_o\big)^2\big].$$

We can now prove that

$$\sigma^T(\tau^1, \ldots, \tau^{T-2}, \tau^{**,T-1}) < \sigma^T(\tau^{*,1}, \tau^{*,2}, \ldots, \tau^{*,T-1}).$$

From Theorem 2.4, we have

$$(\sigma^{t+1})^2 = \sigma_w^2 + \frac{1}{\delta}\mathbb{E}_{B_o,W}\big[\big(\eta\big(B_o + \sigma^t W; \tau^t\big) - B_o\big)^2\big]. \tag{18}$$

Since $\sigma^{T-1}(\tau^1, \ldots, \tau^{T-2}) < \sigma^{T-1}(\tau^{*,1}, \tau^{*,2}, \ldots, \tau^{*,T-2})$, Lemma 4.2 combined with (18) prove that $\sigma^T(\tau^1, \ldots, \tau^{T-2}, \tau^{**,T-1}) < \sigma^T(\tau^{*,1}, \tau^{*,2}, \ldots, \tau^{*,T-1})$, that contradicts the optimality of $\tau^{*,1}, \tau^{*,2}, \ldots, \tau^{*,T-1}$. Therefore, we conclude that if $\tau^{*,1}, \tau^{*,2}, \ldots, \tau^{*,T-1}$ is optimal for iteration $T$, then it is optimal for every iteration $t < T$. The rest of the induction argument is similar, and hence for the sake of brevity we skip it.

4.2. *Proof of Theorem 3.6.*

4.2.1. *Roadmap of the proof.* We break the rather long proof of this theorem into two steps:

1. First, we prove that the risk estimate presented in (9) provides a consistent estimate of the risk $R_B(\sigma^t, \tau^t; p_\beta)$. Since we would like to optimize the risk estimate over the parameter $\tau^t$, we require a uniform notion of consistency, that is,
$$\lim_{h\to 0}\lim_{p\to\infty}\sup_{\tau^t\in\mathcal{T}^t}|\hat{R}^t_{h,p}(\tau^t) - R_B(\sigma^t, \tau^t; p_\beta)| = 0,$$
in probability. Note that the convergence is uniform on $\mathcal{T}^t$. After discussing several useful lemmas, we prove this claim in Theorem 4.4.
2. Once we prove this claim, we use the properties of the solution path of AMP, in particular Theorem 3.3, to show the consistency of our parameter tuning scheme across $t$ iterations.

4.2.2. *Uniform convergence of the risk estimate.* We start with a few lemmas that will be later used to prove

$$\lim_{h\to 0}\lim_{p\to\infty}\sup_{\tau^t\in\mathcal{T}^t}\left|\hat{R}^t_{h,p}(\tau^t)-R_B(\sigma^t,\tau^t;p_\beta)\right|=0,$$

in probability. Our first lemma is concerned with the pointwise (with respect to $\tau^t$) convergence of the risk estimate to $R_B(\sigma^t,\tau^t;p_\beta)$.

LEMMA 4.3. *Let $\{\beta_o(p),X(p),w(p)\}$ be a standard converging sequence. Furthermore, let $\hat{R}^t_{h,p}(\tau^t)$ denote the estimate of the risk at iteration $t$ of AMP as defined in (9). Then*

$$\lim_{p\to\infty}\hat{R}^t_{h,p}(\tau^t,\tau^{t-1},\ldots,\tau^1)=\mathbb{E}_{B_o,W}\big[\big(\tilde{\eta}_h(B_o+\sigma^tW;\tau^t)-B_o\big)^2\big], \tag{19}$$

*almost surely, where $B_o$ and $W$ are two independent random variables with distributions $p_\beta$ and $N(0,1)$, respectively, and $\sigma^t$ satisfies (6).*

PROOF. By applying Lemma 4.1 to the right-hand side of (19), we can rewrite it as

$$\begin{aligned}&\mathbb{E}_{B_o,W}\big[\big(\tilde{\eta}_h(B_o+\sigma^tW;\tau^t)-B_o\big)^2\big]\\&\quad=\mathbb{E}_{B_o,W}\big[\big(\tilde{\eta}_h(B_o+\sigma^tW;\tau^t)-(B_o+\sigma^tW)\big)^2\big]+(\sigma^t)^2\\&\qquad+2(\sigma^t)^2\mathbb{E}_{B_o,W}\big[\big(\tilde{\eta}'_h(B_o+\sigma^tW;\tau^t)-1\big)\big].\end{aligned} \tag{20}$$

Similarly, we can decompose the left-hand side (LHS) of (19) to

$$\begin{aligned}\hat{R}^t{}_{h,p}(\tau^t,\tau^{t-1},\ldots,\tau^1)&=\frac{1}{p}\big\|\tilde{\eta}_h(\beta^t+X^*z^t;\tau^t)-(\beta^t+X^*z^t)\big\|_2^2+(\sigma^t)^2\\&\quad+\frac{2}{p}(\sigma^t)^2\big[\mathbf{1}^*\big(\tilde{\eta}'_h(\beta^t+X^*z^t;\tau^t)-\mathbf{1}\big)\big].\end{aligned} \tag{21}$$

Let $X_{(:,i)}$ denote the $i$th column of $X$. Define

$$b^t\triangleq\beta^t+X^*z^t-\beta_o. \tag{22}$$

Considering the following function:

$$\begin{aligned}\psi_1(b_i^t,\beta_{o,i})&\triangleq\big(\tilde{\eta}_h(b_i^t+\beta_{o,i};\tau^t)-(b_i^t+\beta_{o,i})\big)^2\\&=\big(\tilde{\eta}_h(\beta_i^t+X^*_{(:,i)}z^t;\tau^t)-(\beta_i^t+X^*_{(:,i)}z^t)\big)^2.\end{aligned} \tag{23}$$

It is straightforward to use Lemma 1 of [4] to prove

$$\lim_{p\to\infty}\frac{1}{p}\sum_{i=1}^{p}\psi_1(b_i^t,\beta_{o,i})=\mathbb{E}_{B_o,W}\big[\big(\tilde{\eta}_h(B_o+\sigma^tW;\tau^t)-(B_o+\sigma^tW)\big)^2\big], \tag{24}$$

almost surely. Furthermore, it is straightforward to note that the derivative of $\tilde{\eta}_h$ is bounded, and hence by Lemma 1 of [4] almost surely

$$\lim_{p\to\infty} \frac{\mathbf{1}^*(\tilde{\eta}_h'(\beta^t + X^* z^t; \tau^t) - \mathbf{1})}{p} = \mathbb{E}(\tilde{\eta}_h'(B_o + \sigma^t W; \tau^t) - 1). \tag{25}$$

Combining (20), (24) and (25) completes the proof. □

Lemma 4.3 is only concerned with the pointwise convergence of the risk. The next theorem proves the uniform convergence. Define

$$R_A^t(\tau^t, \tau^{t-1}, \ldots, \tau^1) \triangleq R_B(\tau^t, \sigma^t; p_\beta), \tag{26}$$

where $\sigma^t$ is derived from the iterations of (6). This new notation will be useful in our proofs.

THEOREM 4.4. *Let $\{\beta_o(p), X(p), w(p)\}$ be a standard converging sequence. Furthermore, let $\hat{R}_{h,p}^t(\tau^t, \tau^{t-1}, \ldots, \tau^1)$ denote the estimate of the Bayes risk at iteration $t$ of AMP as defined in (9). Let $\mathcal{T}^t \subset \mathbb{R}$ denote a compact set. Then*

$$\lim_{h\to 0} \lim_{p\to\infty} \sup_{\tau^t \in \mathcal{T}^t} |\hat{R}_{h,p}^t(\tau^t, \tau^{t-1}, \ldots, \tau^1) - R_A^t(\tau^t, \tau^{t-1}, \ldots, \tau^1)| = 0 \tag{27}$$

*in probability, for every $\tau^1, \ldots, \tau^{t-1} \in \mathcal{T}^1 \times \cdots \times \mathcal{T}^{t-1}$.*

PROOF. We first define the the following function:

$$R_{A,h}^t(\tau^t, \tau^{t-1}, \ldots, \tau^1) \triangleq \mathbb{E}(\tilde{\eta}_h(B + \sigma^t W; \tau^t) - B)^2,$$

where $B$ and $W$ are two independent random variables with distributions $p_\beta$ and $N(0, 1)$, respectively, and $\sigma^t$ satisfies (6). Note that this is the asymptotic risk of AMP for the smoothed version of the soft thresholding function. By the triangle inequality, we have

$$\begin{aligned} &|\hat{R}_{h,p}^t(\tau^t, \tau^{t-1}, \ldots, \tau^1) - R_A^t(\tau^t, \tau^{t-1}, \ldots, \tau^1)| \\ &\quad \le |\hat{R}_{h,p}^t(\tau^t, \tau^{t-1}, \ldots, \tau^1) - R_{A,h}^t(\tau^t, \tau^{t-1}, \ldots, \tau^1)| \\ &\qquad + |R_{A,h}^t(\tau^t, \tau^{t-1}, \ldots, \tau^1) - R_A^t(\tau^t, \tau^{t-1}, \ldots, \tau^1)|. \end{aligned} \tag{28}$$

Hence, we first prove that

$$\lim_{p\to\infty} \sup_{\tau^t \in \mathcal{T}^t} |\hat{R}_{h,p}^t(\tau^t, \tau^{t-1}, \ldots, \tau^1) - R_{A,h}^t(\tau^t, \tau^{t-1}, \ldots, \tau^1)| = 0, \tag{29}$$

in probability for every $h > 0$ and every $\tau^1, \ldots, \tau^{t-1}$. Second, we prove that

$$\lim_{h\to 0} \sup_{\tau \in \mathcal{T}^t} |R_{A,h}^t(\tau^t, \tau^{t-1}, \ldots, \tau^1) - R_A^t(\tau^t, \tau^{t-1}, \ldots, \tau^1)| = 0. \tag{30}$$

Combining these two results will establish the theorem. To establish (29), we start with the following definitions whose importance will become clear later:

$$
\begin{aligned}
U_{h,p}(b_i, \beta_{o,i}, \tau, \sigma) &\triangleq \big(\tilde{\eta}_h(b_i + \beta_{o,i}; \tau) - (b_i + \beta_{o,i})\big)^2 + \sigma^2 \\
&\quad + 2\sigma^2\big[\big(\tilde{\eta}_h'(b_i + \beta_{o,i}; \tau) - 1\big)\big], \qquad (31)\\
\bar{U}(b_i, \beta_{o,i}, \rho, \tau, \sigma) &\triangleq \sup_{\tilde{\tau}:|\tilde{\tau}-\tau|\le\rho} U_{h,p}(b_i, \beta_{o,i}, \tilde{\tau}, \sigma) - \mathbb{E}U_{h,p}(\sigma W, B, \tilde{\tau}, \sigma),
\end{aligned}
$$

where $B$ and $W$ are two independent random variables with distributions $p_\beta$ and $N(0, 1)$, respectively. The following remarks clarify some of the main features and connections of these definitions:

1. It is straightforward to verify that
$$\hat{R}^t_{h,p}(\tau^t, \tau^{t-1}, \ldots, \tau^1) = \frac{1}{p}\sum_{i=1}^{p} U_{h,p}(b_i^t, \beta_{o,i}, \tau^t, \sigma^t),$$
where $b^t = \beta^t + X^* z^t - \beta_o$ and $\beta^t$ is the estimate of AMP with threshold parameters $\tau^i$ at the $i$th iteration.
2. According to Lemma 4.3, $\frac{1}{p}\sum_{i=1}^{p} U_{h,p}(b_i^t, \beta_{o,i}, \tau^t, \sigma^t) \overset{\text{a.s.}}{\to} R^t_{A,h}(\tau^t, \tau^{t-1}, \ldots, \tau^1)$.
3. According to Lemma 4.1, $R_{A,h}(\tau^t, \tau^{t-1}, \ldots, \tau^1) = \mathbb{E}U_{h,p}(\sigma^t W, B, \tau, \sigma^t)$, where the expectation is with respect to two independent random variables $W \sim N(0, 1)$ and $B \sim p_\beta$.

The next four lemmas prove several basic properties of $R_{A,h}$, $U_{h,p}$ and $\bar{U}_{h,p}$ that will be useful later in our proof.

LEMMA 4.5. *$R^t_{A,h}(\tau^t, \tau^{t-1}, \ldots, \tau^1)$ is a continuous function of $\tau^t$, for every $\tau^1, \tau^2, \ldots, \tau^{t-1} \in \mathcal{T}^1 \times \mathcal{T}^2, \ldots, \mathcal{T}^{t-1}$.*

PROOF. The proof is a straightforward application of the dominated convergence theorem:

$$
\begin{aligned}
&\lim_{\tilde{\tau}^t \to \tau^t} R^t_{A,h}(\tilde{\tau}^t, \tau^{t-1}, \ldots, \tau^1) \\
&\quad = \lim_{\tilde{\tau}^t \to \tau^t} \mathbb{E}U_{h,p}(\sigma^t W, B, \tilde{\tau}^t, \sigma^t) \\
&\quad \overset{(a)}{=} \mathbb{E} \lim_{\tilde{\tau}^t \to \tau^t} U_{h,p}(\sigma^t W, B, \tilde{\tau}^t, \sigma^t) \overset{(b)}{=} \mathbb{E}U_{h,p}(\sigma^t W, B, \tau^t, \sigma^t).
\end{aligned}
$$

Equality (a) is due to the fact that $U_{h,p}$ is a bounded function of both $\sigma^t W$ and $B$, and hence dominated convergence theorem can be applied. Equality (b) uses the continuity of $U_{h,p}$ with respect to $\tau^t$. □

LEMMA 4.6. *Let $\mathcal{T}$ denote a compact subset of $\mathbb{R}$. $U_{h,p}(b_i, \beta_{o,i}, \tau, \sigma)$ is a Lipschitz function of $(b_i, \beta_{o,i})$ with Lipschitz constant*:

$$L_U \triangleq \sqrt{2} \max_{\tau \in \mathcal{T}} 2\tau \Big( \sup_{\zeta} |\tilde{\eta}_h'(\zeta; \tau)| + 1 \Big) + 2\sigma^2 \sup_{\tilde{\zeta}} |\tilde{\eta}_h''(\tilde{\zeta}; \tau)|.$$

It is important to note that both $|\tilde{\eta}_h'(\zeta; \tau)|$ and $|\tilde{\eta}_h''(\tilde{\zeta}; \tau)|$ are bounded functions of $\zeta$ and $\tilde{\zeta}$, respectively, for a fixed $\tau$. Since $\mathcal{T}$ is a compact set, $L_U$ is bounded as well.

PROOF. Define $s_i \triangleq b_i + \beta_{o,i}$ and $\tilde{s}_i \triangleq \tilde{b}_i + \tilde{\beta}_i$:

$$\begin{aligned}
&\big|U_{h,p}(b_i, \beta_{o,i}, \tau, \sigma) - U_{h,p}(\tilde{b}_i, \tilde{\beta}_{o,i}, \tau, \sigma)\big| \\
&\quad = \big|\big(\tilde{\eta}_h(s_i; \tau) - s_i\big)^2 + 2\sigma^2 \tilde{\eta}_h'(s_i; \tau) - \big(\tilde{\eta}_h(\tilde{s}_i; \tau) - \tilde{s}_i\big)^2 + 2\sigma^2 \tilde{\eta}_h'(\tilde{s}_i; \tau)\big| \\
&\quad \overset{(a)}{=} \big|\big(\tilde{\eta}_h'(\zeta; \tau) + 1\big)(s_i - \tilde{s}_i)\big(\tilde{\eta}_h(s_i; \tau) - s_i + \tilde{\eta}_h(\tilde{s}_i; \tau) - \tilde{s}_i\big)\big| \\
&\qquad + 2\sigma^2 |\tilde{\eta}_h''(\tilde{\zeta}; \tau)||(s_i - \tilde{s}_i)| \\
&\quad \overset{(b)}{\leq} 2\tau \Big( \sup_{\zeta} |\tilde{\eta}_h'(\zeta; \tau)| + 1 \Big) + 2\sigma^2 \Big( \sup_{\tilde{\zeta}} |\tilde{\eta}_h''(\tilde{\zeta}; \tau)| \Big) |(s_i - \tilde{s}_i)|.
\end{aligned} \tag{32}$$

Note that equality (a) is derived from the mean value theorem. To obtain inequality (b) we used the fact that $|\eta_h(s, \tau) - s| \leq \tau$. Finally, to show that the function is Lipschitz we employ the inequality $|s_i - \tilde{s}_i| \leq \sqrt{2}\sqrt{(b_i - \tilde{b}_i)^2 + (\beta_i^t - \tilde{\beta}_i)^2}$. □

LEMMA 4.7. *$\bar{U}(b_i, \beta_{o,i}, \rho, \tau, \sigma)$ is also a Lipschitz function of $(b_i, \beta_{o,i})$ with Lipschitz constant $L_U$ defined in Lemma 4.6.*

PROOF. From Lemma 4.6, we have

$$U_{h,p}(b_i, \beta_{o,i}, \tilde{\tau}, \sigma) \leq U_{h,p}(\tilde{b}_i, \tilde{\beta}_{o,i}, \tilde{\tau}, \sigma) + L_U \sqrt{(b_i - \tilde{b}_i)^2 + \big(\beta_i^t - \tilde{\beta}_i\big)^2}.$$

By subtracting the constant (in terms of $b_i$ and $\beta_{o,i}$) $\mathbb{E}U_{h,p}(\sigma^t W, B, \tilde{\tau}, \sigma)$ and taking the supremum with respect to $\tilde{\tau}$, we obtain

$$\bar{U}(b_i, \beta_{o,i}, \rho, \tau, \sigma) \leq \bar{U}(\tilde{b}_i, \tilde{\beta}_{o,i}, \rho, \tau, \sigma) + L_U \sqrt{(b_i - \tilde{b}_i)^2 + \big(\beta_i^t - \tilde{\beta}_i\big)^2}.$$

The proof of the other direction is similar. □

LEMMA 4.8. *Let $W$ and $B$ denote two independent random variables with distributions $N(0, 1)$ and $p_\beta$, respectively. Then* $\lim_{\rho \to 0} \mathbb{E}\bar{U}(\sigma W, B, \rho, \tau, \sigma) = 0$.

PROOF. Since $\bar{U}$ is a bounded function, we can exchange the order of $\lim_{\rho\to 0}$ and $\mathbb{E}$. Hence,

$$\begin{aligned}\lim_{\rho\to 0}\mathbb{E}\bar{U}(\sigma W, B, \rho, \tau, \sigma) &= \mathbb{E}\lim_{\rho\to 0}\bar{U}(\sigma W, B, \rho, \tau, \sigma)\\ &\overset{(a)}{=} \mathbb{E}U_{h,p}\big(\sigma^t W, B, \tau, \sigma\big) - \mathbb{E}U_{h,p}\big(\sigma^t \tilde{W}, \tilde{B}, \tau, \sigma\big)\\ &= 0.\end{aligned} \tag{33}$$

Note that to obtain equality (a) we use the continuity of $U_{h,p}$ in $\tau$. □

Lemma 4.8 implies that for any $\varepsilon > 0$, there exists $\rho_{\tau_0}$ such that if $|\tau - \tau_0| < \rho_{\tau_0}$, then $\mathbb{E}\bar{U}(\sigma^t W, B, \rho, \tau) < \varepsilon$. Note that we have a subscript $\tau_0$ for $\rho_{\tau_0}$ to emphasize on the fact that $\rho$ is dependent on the choice of $\tau_0$. Define $\mathcal{B}(c, \rho) = \{\tau \in \mathcal{T} || \tau - c| \le \rho\}$. Consider the set of all the balls $\mathcal{B}(\tau, \rho_\tau)$ for every $\tau \in \mathcal{T}$. This set forms a covering of $\mathcal{T}$. Since $\mathcal{T}$ is compact, it has a finite subcover. Let $\mathcal{B}(\tau_1^*, \rho_1^*)$, $\mathcal{B}(\tau_2^*, \rho_2^*), \ldots, \mathcal{B}(\tau_M^*, \rho_M^*)$ denote this finite subcover. We have

$$\begin{aligned}&\mathbb{P}\left(\sup_\tau \frac{1}{p}\sum_{i=1}^p U_{h,p}\big(b_i^t, \beta_{o,i}, \tau, \sigma^t\big) - R_{A,h}\big(\tau, \tau^{t-1}, \ldots, \tau^1\big) > 2\varepsilon\right)\\ &\quad\le \mathbb{P}\left(\max_i \frac{1}{p}\sum_{i=1}^p \bar{U}_{h,p}\big(b_i^t, \beta_{o,i}, \rho_i^*, \tau_i^*, \sigma^t\big) > 2\varepsilon\right)\\ &\quad< M\max_i \mathbb{P}\left(\frac{1}{p}\sum_{i=1}^p \bar{U}_{h,p}\big(b_i^t, \beta_{o,i}, \rho_i^*, \tau_i^*, \sigma^t\big) > 2\varepsilon\right).\end{aligned} \tag{34}$$

Note that the first inequality is due to the definition of $\bar{U}_{h,p}$ and the second inequality is a simple application of the union bound. The last step of the proof is to show that

$$\mathbb{P}\left(\frac{1}{p}\sum_{i=1}^p \bar{U}_{h,p}\big(b_i^t, \beta_{o,i}, \rho_i^*, \tau_i^*, \sigma^t\big) > 2\varepsilon\right) \to 0, \tag{35}$$

as $p \to \infty$. Note that if we combine Lemma 4.7 and Lemma 1 of [4] we obtain

$$\mathbb{P}\left(\frac{1}{p}\sum_{i=1}^p \bar{U}_{h,p}\big(b_i^t, \beta_{o,i}, \rho_i^*, \tau_i^*, \sigma^t\big) - \mathbb{E}\bar{U}_{h,p}\big(\sigma^t W, B, \rho_i^*, \tau_i^*, \sigma^t\big) > \varepsilon\right) \to 0, \tag{36}$$

as $p \to \infty$. Furthermore, from the construction of the covering we have

$$\max_{i=1,\ldots,M} \mathbb{E}\bar{U}_{h,p}\big(\sigma^t W, B, \rho_i^*, \tau_i^*, \sigma^t\big) < \varepsilon. \tag{37}$$

Hence, by combining (36) and (37) we obtain (35). □

At this point, we refer the reader to (28). So far, we have proved (29). Hence, if we prove (30), it will establish (28) and will complete the proof

of Theorem 4.4. Hence, in this step our goal is to prove that the function $\sup_{\tau^t \in \mathcal{T}^t} |R^t_{A,h}(\tau^t, \tau^{t-1}, \ldots, \tau^1) - R^t_A(\tau^t, \tau^{t-1}, \ldots, \tau^1)|$, is a continuous function of $h$. In Lemma 4.5, we showed that $R^t_{A,h}(\tau^t, \tau^{t-1}, \ldots, \tau^1)$ is a continuous function of $\tau^t$. It is straightforward to use the same argument to show that it is a continuous function of $(h, \tau^t)$. Therefore, we can show that the function $|R^t_{A,h}(\tau^t, \tau^{t-1}, \ldots, \tau^1) - R^t_A(\tau^t, \tau^{t-1}, \ldots, \tau^1)|$ is also a continuous function of $(h, \tau)$. We require the following standard result from analysis.

LEMMA 4.9. *Let $f(h, \tau)$ denote a continuous function from $\mathbb{R}^2$ to $\mathbb{R}$. Also, assume that $\mathcal{T}$ is a compact subset of $\mathbb{R}$. Then, $\lim_{h \to h_o} \sup_{\tau \in \mathcal{T}} f(h, \tau) = \sup_{\tau \in \mathcal{T}} f(h_o, \tau)$.*

This is a standard result and its proof can be found elsewhere. For instance, it is equivalent to Lemma 12 in [40]. According to this lemma, $\sup_{\tau \in \mathcal{T}} f(h, \tau)$ is a continuous function of $h$. Applying Lemma 4.9 to $|R^t_{A,h}(\tau^t, \tau^{t-1}, \ldots, \tau^1) - R^t_A(\tau^t, \tau^{t-1}, \ldots, \tau^1)|$ proves (30).

4.2.3. *Consistency of the parameter tuning.* At this point, we remind the reader that as we discussed in Section 4.2.1 we broke the proof of Theorem 3.6 in two steps. The first step was to prove:

$$\lim_{h \to 0} \lim_{p \to \infty} \sup_{\tau^t \in \mathcal{T}} \left| \hat{R}^t_{h,p}(\tau^t, \tau^{t-1}, \ldots, \tau^1) - R^t_A(\tau^t, \tau^{t-1}, \ldots, \tau^1) \right| = 0 \tag{38}$$

in probability, for every $\tau^1, \ldots, \tau^{t-1} \in \mathcal{T}^1 \times \cdots \times \mathcal{T}^{t-1}$, that was established in Theorem 4.4. In this section, we would like to prove the second step, that is, the consistency of $\hat{\tau}^1_{p,h}, \hat{\tau}^2_{p,h}, \ldots, \hat{\tau}^t_{p,h}$. For the proof, we employ an induction. As a base of induction, we first prove that $\hat{\tau}^1_{p,h} \to \tau^{*,1}$ in probability. First, note that from the proof of Lemma A.5 (in the supplementary material), we conclude that for every $\varepsilon > 0$ we have

$$\inf_{\tau : |\tau - \tau^{*,1}| > \varepsilon} R^1_A(\tau) > R^1_A(\tau^{*,1}).$$

In the rest of the proof, we assume that $\inf_{\tau : |\tau - \tau^{*,1}| > \varepsilon} R^1_A(\tau) - R^1_A(\tau^{*,1}) = 2\gamma$, where $\gamma > 0$ is a fixed number. We proved in Theorem 4.4 that

$$\sup_{\tau \in \mathcal{T}} |R^1_{A,h}(\tau) - R^1_A(\tau)| \to 0, \tag{39}$$

as $h \to 0$. Hence, we can find $h_o$ such that for every $h < h_o$, $\sup_{\tau \in \mathcal{T}} |R^1_{A,h}(\tau) - R^1_A(\tau)| < \gamma/2$. This implies that for $h < h_0$

$$R^1_{A,h}(\tau^{*,1}) < R^1_A(\tau^{*,1}) + \gamma/2, \tag{40}$$

$$\inf_{\tau : |\tau - \tau^{*,1}| > \varepsilon} R^1_{A,h}(\tau) > \inf_{\tau : |\tau - \tau^{*,1}| > \varepsilon} R^1_A(\tau) - \gamma/2. \tag{41}$$

In Theorem 4.4, we proved that

$$\mathbb{P}\Big(\sup_{\tau}|\hat{R}^1_{h,p}(\tau) - R^1_{A,h}(\tau)| > \gamma/2\Big) \to 0,$$

as $p \to \infty$. As a result, we conclude that

$$\mathbb{P}\big(\hat{R}^1_{h,p}(\tau^{*,1}) > R^1_{A,h}(\tau^{*,1}) + \gamma/2\big) \to 0, \tag{42}$$

$$\mathbb{P}\Big(\inf_{\tau:|\tau-\tau^{*,1}|>\varepsilon} \hat{R}^1_{h,p}(\tau) < \inf_{\tau:|\tau-\tau^{*,1}|>\varepsilon} R^1_{A,h}(\tau) - \gamma/2\Big) \to 0. \tag{43}$$

Hence, by combining (40) and (42) we conclude that

$$\mathbb{P}\big(\hat{R}^1_{h,p}(\tau^{*,1}) > R^1_{A}(\tau^{*,1}) + \gamma\big) \to 0. \tag{44}$$

It is also straightforward to combine (41) and (43) and obtain

$$\mathbb{P}\Big(\inf_{\tau:|\tau-\tau^{*,1}|>\varepsilon} \hat{R}^1_{h,p}(\tau) < \inf_{\tau:|\tau-\tau^{*,1}|>\varepsilon} R^1_{A,h}(\tau) - \gamma\Big) \to 0. \tag{45}$$

Combining (44) and (45) proves that if $h < h_0$, then

$$\mathbb{P}\big(|\hat{\tau}^1_{p,h} - \tau^{*,1}| > \varepsilon\big) \to 0, \tag{46}$$

as $p \to \infty$. Now we use an induction to show that if $\hat{\tau}^1_{p,h} \xrightarrow{p} \tau^{*,1}, \hat{\tau}^2_{p,h} \xrightarrow{p} \tau^{*,2}, \ldots, \hat{\tau}^t_{p,h} \xrightarrow{p} \tau^{*,t}$, then $\hat{\tau}^{t+1}_{p,h} \xrightarrow{p} \tau^{*,t+1}$. To keep the notation simpler, we only prove this claim for $t = 1$. The proof for an arbitrary $t$ is the same. Our proof uses the following steps:

1. We first prove that $|\hat{R}^2_{h,p}(\tau^2, \hat{\tau}^1_{p,h}) - R^2_A(\tau^2, \tau^{*,1})| \xrightarrow{p} 0$. Note that the main reason this cannot be derived from Theorem 4.4 is that now we have used a data-dependent threshold $\hat{\tau}^1_{p,h}$ in the first iteration. In Theorem 4.4, the threshold does not depend on data.
2. Next, we prove that $\sup_{\tau^2\in\mathcal{T}^2} |\hat{R}_{h,p}(\tau^2, \hat{\tau}^1_{p,h}) - R_B(\tau^2, \tau^{*,1})| \xrightarrow{p} 0$. Using the proof technique in Theorem 4.4 and the fact that we have already proved $|\hat{R}_{h,p}(\tau^2, \hat{\tau}^1_{p,h}) - R_B(\tau^2, \tau^{*,1})| \xrightarrow{p} 0$, the proof of this statement is straightforward and will be skipped.
3. Finally, we use the fact that $\sup_{\tau^2\in\mathcal{T}^2} |\hat{R}_{h,p}(\tau^2, \hat{\tau}^1_{p,h}) - R_B(\tau^2, \tau^{*,1})| \xrightarrow{p} 0$ and the proof technique developed in (46) to show that $\hat{\tau}^2_{p,h} \to \tau^{2,*}$. Note that by Theorem 3.3 we already know that even though $\tau^1$ is set to $\tau^{*,1}$, the optimal choice of $\tau^2$ can still be achieved. Since this is exactly the same as the proof of (46), we will skip the proof.

We only prove the first of the above three steps. First, note that

$$\begin{aligned}|\hat{R}^2_{h,p}(\tau^2, \hat{\tau}^1_{p,h}) - R^2_A(\tau^2, \tau^{*,1})| \le{}& |\hat{R}^2_{h,p}(\tau^2, \hat{\tau}^1_{p,h}) - \hat{R}^2_{h,p}(\tau^2, \tau^{*,1})| \\ &+ |\hat{R}^2_{h,p}(\tau^2, \tau^{*,1}) - R^2_{A,h}(\tau^2, \tau^{*,1})| \\ &+ |R^2_{A,h}(\tau^2, \tau^{*,1}) - R^2_A(\tau^2, \tau^{*,1})|.\end{aligned} \tag{47}$$

We consider each of the three terms on the right and prove that they converge to zero in probability. $\hat{R}^2_{h,p}(\tau^2, \tau^1)$ is differentiable in terms of $\tau^1$ and $\tau^2$. Furthermore, the derivative is bounded with probability one. Hence, by the mean value theorem we have

$$|\hat{R}^2_{h,p}(\tau^2, \hat{\tau}^1_{p,h}) - \hat{R}^2_{h,p}(\tau^2, \tau^{*,1})| \le C|\hat{\tau}^1_{p,h} - \tau^{*,1}|,$$

where $C$ is an upper bound on the derivative of $\hat{R}_{h,p}$ in terms of $\tau^1$. Hence, it is straightforward to use the base of the induction and prove that

$$\begin{aligned}&\mathbb{P}\big(|\hat{R}^2_{h,p}(\tau^2, \hat{\tau}^1_{p,h}) - \hat{R}^2_{h,p}(\tau^2, \tau^{*,1})| > \varepsilon\big)\\ &\quad\le \mathbb{P}\Big(\sup_{(\tau_1,\tau_2)\in\mathcal{T}_1\times\mathcal{T}_2} (\hat{R}^2_{h,p})'(\tau_2, \tau_1) > C\Big) + \mathbb{P}\big(C|\hat{\tau}^1_{p,h} - \tau^{*,1}| > \varepsilon\big).\end{aligned}$$

Since both probabilities go to zero as $p \to \infty$, we conclude that

$$\mathbb{P}\big(|\hat{R}^2_{h,p}(\tau^2, \hat{\tau}^1_{p,h}) - \hat{R}^2_{h,p}(\tau^2, \tau^{*,1})| > \varepsilon\big) \to 0. \tag{48}$$

Furthermore, according to Theorem 4.4 we have

$$|\hat{R}^2_{h,p}(\tau^2, \tau^{*,1}) - R^2_{A,h}(\tau^2, \tau^{*,1})| \xrightarrow{p} 0. \tag{49}$$

By combining (48) and (49), we obtain $|\hat{R}^2_{h,p}(\tau^2, \hat{\tau}^1_{p,h}) - R^2_{A,h}(\tau^2, \tau^{*,1})| \xrightarrow{p} 0$.

The proof of $|R^2_{A,h}(\tau^2, \tau^{*,1}) - R^2_A(\tau^2, \tau^{*,1})| \to 0$ as $h \to 0$, is a straightforward application of the continuity of $R^2_{A,h}(\tau^2, \tau^{*,1})$ with respect to $(h, \tau^2)$ and is hence skipped. This completes our proof of the consistency of $\hat{\tau}^2_{p,h}$.

**5. Conclusions.** In this paper, we have characterized the performance of LASSO and AMP for estimating a sparse vector from undersampled, noisy observations. By considering a model in which the design matrix and noise are zero-mean i.i.d. Gaussian, we proposed a computationally efficient, data-driven approach for estimating the free parameters of LASSO and AMP. We have shown that our estimates are consistent in the sense that they converge to their asymptotically optimal values in probability. Finally, we have proved asymptotic properties of the solution path of LASSO and AMP.

## SUPPLEMENTARY MATERIAL

**Supplement to "Consistent parameter estimation for LASSO and approximate message passing"** (DOI: 10.1214/16-AOS1529SUPP; .pdf). This supplementary material includes the proof of theorems and simulation results.

## REFERENCES


[1] AMELUNXEN, D., LOTZ, M., MCCOY, M. B. and TROPP, J. A. (2014). Living on the edge: A geometric theory of phase transitions in convex optimization. Preprint. Available at arXiv:1303.6672.

[2] BAYATI, M., LELARGE, M. and MONTANARI, A. (2012). Universality in polytope phase transitions and message passing algorithms. Preprint. Available at arXiv:1207.7321.

[3] BAYATI, M. and MONTANARI, A. (2012). The LASSO risk for Gaussian matrices. *IEEE Trans. Inform. Theory* **58** 1997–2017. MR2951312

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

[5] BICKEL, P. J., RITOV, Y. and TSYBAKOV, A. B. (2009). Simultaneous analysis of LASSO and Dantzig selector. *Ann. Statist.* 1705–1732. MR2533469

[6] BUNEA, F., TSYBAKOV, A. and WEGKAMP, M. (2007). Sparsity oracle inequalities for the Lasso. *Electron. J. Stat.* **1** 169–194. MR2312149

[7] CAI, T. T., WANG, L. and XU, G. (2010). Shifting inequality and recovery of sparse signals. *IEEE Trans. Signal Process.* **58** 1300–1308.

[8] CAI, T. T., WANG, L. and XU, G. (2010). Stable recovery of sparse signals and an oracle inequality. *IEEE Trans. Inform. Theory* **56** 3516–3522. MR2799010

[9] CAI, T. T., XU, G. and ZHANG, J. (2009). On recovery of sparse signals via $\ell_1$-minimization. *IEEE Trans. Inform. Theory* **55** 3388–3397.

[10] CANDÈS, E. J., ROMBERG, J. K. and TAO, T. (2006). Stable signal recovery from incomplete and inaccurate measurements. *Comm. Pure Appl. Math.* **59** 1207–1223. MR2230846

[11] CHATTERJEE, S. and JAFAROV, J. (2015). Prediction error of cross-validated lasso. Preprint. Available at arXiv:1502.06291.

[12] CHEN, S. S., DONOHO, D. L. and SAUNDERS, M. A. (1998). Atomic decomposition by basis pursuit. *SIAM J. Sci. Comput.* **20** 33–61. MR1639094

[13] DONOHO, D. and TANNER, J. (2009). Observed universality of phase transitions in high-dimensional geometry, with implications for modern data analysis and signal processing. *Philos. Trans. R. Soc. Lond. Ser. A Math. Phys. Eng. Sci.* **367** 4273–4293. MR2546388

[14] DONOHO, D. L. (2006). High-dimensional centrally symmetric polytopes with neighborliness proportional to dimension. *Discrete Comput. Geom.* **35** 617–652. MR2225676

[15] DONOHO, D. L., MALEKI, A. and MONTANARI, A. (2009). Message passing algorithms for compressed sensing. *Proc. Natl. Acad. Sci. USA* **106** 18914–18919.

[16] DONOHO, D. L., MALEKI, A. and MONTANARI, A. (2011). Noise sensitivity phase transition. *IEEE Trans. Inform. Theory* **57** 6920–6941. MR2882271

[17] DONOHO, D. L. and TANNER, J. (2005). Neighborliness of randomly projected simplices in high dimensions. *Proc. Natl. Acad. Sci. USA* **102** 9452–9457. MR2168716

[18] DONOHO, D. L. and TANNER, J. (2009). Counting faces of randomly projected polytopes when the projection radically lowers dimension. *J. Amer. Math. Soc.* **22** 1–53. MR2449053

[19] DONOHO, D. L. and TANNER, J. (2010). Precise undersampling theorems. *Proc. IEEE* **98** 913–924.

[20] EFRON, B., HASTIE, T., JOHNSTONE, I. and TIBSHIRANI, R. (2004). Least angle regression. *Ann. Statist.* **32** 407–499. MR2060166

[21] HASTIE, T., TIBSHIRANI, R., FRIEDMAN, J. and FRANKLIN, J. (2005). The elements of statistical learning: Data mining, inference and prediction. *Math. Intelligencer* **27** 83–85.

[22] HOMRIGHAUSEN, D. and MCDONALD, D. J. (2014). Leave-one-out cross-validation is risk consistent for lasso. *Mach. Learn.* **97** 65–78.

[23] KNIGHT, K. and FU, W. (2000). Asymptotics for LASSO-type estimators. *Ann. Statist.* **28** 1356–1378.

[24] LOCKHART, R., TAYLOR, J., TIBSHIRANI, R. J. and TIBSHIRANI, R. (2012). A significance test for the LASSO. Preprint. Available at arXiv:1301.7161.

[25] MALEKI, A. (2010). Approximate message passing algorithm for compressed sensing. Ph.D. Thesis, Stanford Univ., Stanford, CA.

[26] MEINSHAUSEN, N. and BÜHLMANN, P. (2006). High-dimensional graphs and variable selection with the lasso. *Ann. Statist.* **34** 1436–1462. MR2278363

[27] MEINSHAUSEN, N. and YU, B. (2009). Lasso-type recovery of sparse representations for high-dimensional data. *Ann. Statist.* **37** 246–270. MR2488351

[28] MOUSAVI, A., MALEKI, A. and BARANIUK, R. G. (2013). Parameterless optimal approximate message passing. Preprint. Available at arXiv:1311.0035.

[29] MOUSAVI, A., MALEKI, A. and BARANIUK, R. G. (2017). Supplement to "Consistent parameter estimation for LASSO and approximate message passing." DOI:10.1214/16-AOS1529SUPP.

[30] OYMAK, S. and HASSIBI, B. (2012). On a relation between the minimax risk and the phase transitions of compressed recovery. In *Proc. Allerton Conf. Communication, Control, and Computing* 1018–1025, Urbana-Champaign, IL.

[31] RASKUTTI, G., WAINWRIGHT, M. J. and YU, B. (2011). Minimax rates of estimation for high-dimensional linear regression over $\ell_q$-Balss. *IEEE Trans. Inform. Theory* **57** 6976–6994.

[32] STEIN, C. M. (1981). Estimation of the mean of a multivariate normal distribution. *Ann. Statist.* 1135–1151.

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

[34] TIBSHIRANI, R. J. (2013). The lasso problem and uniqueness. *Electron. J. Stat.* **7** 1456–1490. MR3066375

[35] VAN DE GEER, S. A. and BÜHLMANN, P. (2009). On the conditions used to prove oracle results for the Lasso. *Electron. J. Stat.* **3** 1360–1392. MR2576316

[36] WASSERMAN, L. and ROEDER, K. (2009). High-dimensional variable selection. *Ann. Statist.* **37** 2178–2201. MR2543689

[37] ZHANG, C.-H. and HUANG, J. (2008). The sparsity and bias of the LASSO selection in high-dimensional linear regression. *Ann. Statist.* **36** 1567–1594. MR2435448

[38] ZHANG, T. (2009). Some sharp performance bounds for least squares regression with $L_1$ regularization. *Ann. Statist.* **37** 2109–2144. MR2543687

[39] ZHAO, P. and YU, B. (2006). On model selection consistency of Lasso. *J. Mach. Learn. Res.* **7** 2541–2563. MR2274449

[40] ZHENG, L., MALEKI, A., WANG, X. and LONG, T. (2015). Does $\ell_p$-minimization outperform $\ell_1$-minimization? Preprint. Available at arXiv:1501.03704.

[41] ZOU, H., HASTIE, T. and TIBSHIRANI, R. (2007). On the degrees of freedom of the LASSO. *Ann. Statist.* **35** 2173–2192. MR2363967

ECE DEPARTMENT
RICE UNIVERSITY
HOUSTON, TEXAS 77005
USA
E-MAIL: ali.mousavi@rice.edu
richb@rice.edu

DEPARTMENT OF STATISTICS
COLUMBIA UNIVERSITY
NEW YORK, NEW YORK 10027
USA
E-MAIL: arian@stat.columbia.edu