WHICH BRIDGE ESTIMATOR IS OPTIMAL FOR VARIABLE SELECTION?

BY SHUAIWEN WANG∗  HAOLEI WENG† AND ARIAN MALEKI‡

Columbia University ∗†‡

We study the problem of variable selection for linear models under the high-dimensional asymptotic setting, where the number of observations $n$ grows at the same rate as the number of predictors $p$. We consider two-stage variable selection techniques (TVS) in which the first stage uses bridge estimators to obtain an estimate of the regression coefficients, and the second stage simply thresholds the regression coefficients estimate to select the “important” predictors. The asymptotic false discovery proportion (AFDP) and true positive proportion (ATTP) of these TVS are evaluated. We prove that for a fixed ATTP, in order to obtain the smallest AFDP one should pick an estimator that minimizes the asymptotic mean square error in the first stage of TVS. This simple observation enables us to evaluate and compare the performances of different TVS with each other and with some standard variable selection techniques, such as LASSO and Sure Independence Screening. For instance, we prove that a TVS with LASSO in its first stage can outperform LASSO (only one stage) in a large range of ATTP. Furthermore, we will show that for large values of noise, a TVS with ridge in its first stage outperforms TVS with other bridge estimators including the one that has LASSO in its first stage.

1. Introduction.

1.1. Motivation and problem statement. Although linear models can be traced back to two hundred years ago, they keep shining in the modern statistical research. A problem of major interest in this literature is variable selection. Consider the linear regression model

$$y = X\beta + w,$$

with $y \in \mathbb{R}^n$, $X \in \mathbb{R}^{n \times p}$, $\beta \in \mathbb{R}^p$ and $w \in \mathbb{R}^n$, and suppose that many elements of $\beta$ are zero. The problem of variable selection is concerned with finding the locations of the non-zero elements of $\beta$. Motivated by the concerns regarding the instability and high computational cost of classical variable selection techniques, such as best subset selection and forward or backward stepwise, Tibshirani proposed LASSO [Tib96] to perform parameter estimation and variable selection simultaneously. The LASSO solution is given
by

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

where $\lambda \in (0, \infty)$ is the tuning parameter, and $\| \cdot \|_1$ is the $\ell_1$ norm. The regularization term $\|\beta\|_1$ stabilizes the variable selection process and the convex optimization formulation (1.1) reduces the computational cost.

Tibshirani’s seminal work raises an important question. Compared to LASSO, other convex regularizers such as $\|\beta\|_2^2$ can impose larger penalty for large values of $\beta$. Hence, their estimates might be even more stable than LASSO. Even though the solutions of many of these regularizers are not sparse (and thus not automatically performing variable selection), we may threshold their estimates to produce sparse solutions and select variables. This observation leads us to the following fundamental question: can such two-stage variable selection techniques with other regularizers outperform LASSO in variable selection? The goal of this paper is to address this question. In particular, we study the performance of the two-stage variable selection technique discussed above, with the first stage selected from the class of bridge estimators [FF93]:

$$\hat{\beta}(q, \lambda) = \arg\min_{\beta} \frac{1}{2} \|y - X\beta\|_2^2 + \lambda \|\beta\|_q^q,$$

where $\|\beta\|_q^q = \sum_i |\beta_i|^q$ with $q \geq 1$. Our variable selection technique takes $\hat{\beta}(q, \lambda)$ and returns the sparse estimate $\bar{\beta}(q, \lambda, s)$ defined as follows:

$$\bar{\beta}(q, \lambda, s) = \eta_0(\hat{\beta}(q, \lambda); s^2/2),$$

where $\eta_0(u; \chi) = u\mathbb{1}_{\{|u| \geq \sqrt{2\chi}\}}$ denotes the hard threshold function and it operates component-wise for vectors. We can then use this estimator for variable selection. In this paper, we give a thorough investigation of such two-stage variable selection techniques under the asymptotic setting $X_{ij} \sim i.i.d. N(0, 1/n)$, $n/p \to \delta \in (0, \infty)$, and address the following questions:

1. Can we characterize the false discovery proportion and true positive proportion for such two-stage schemes?
2. What is the “optimal” way to tune regularization parameter $\lambda$ and the threshold parameter $s$?
3. Which value of $q$ presents the best variable selection performance?
4. Does LASSO outperform the two-stage method based on other bridge estimators?

A detailed description of our contributions is presented in Section 3.
1.2. Related Work. Traditional methods of variable selection include stepwise procedures and best subset selection. Best subset selection suffers from high computational complexity and high variance. The greedy nature of stagewise procedures reduces the computational complexity, but limits the number of models that are checked by such procedures. To resolve these limitations, [Tib96] proposed the LASSO that aims to perform variable selection and parameter estimation simultaneously.

Both the variable selection and estimation performance of LASSO have been studied extensively in the past decade. For instance, it has been justified in the work of [MB06],[ZY06], [ZH08] that a type of “irrepresentable condition” is almost sufficient and necessary to guarantee sign consistency for the LASSO. Later [Wai09] established sharp conditions under which LASSO can perform a consistent variable selection. One implication of [Wai09] that is relevant to our paper is that, consistent variable selection is impossible under the linear asymptotic regime that we consider in this paper. This result is consistent with that of [SBC15] and our paper. Hence, we should expect that both the true positive proportion (TPP) and false discovery proportion (FDP) play a major role in our analyses and comparisons.

Since LASSO requires strong conditions for variable selection consistency, several authors have considered a few variants, such as adaptive LASSO [Zou06] and thresholded LASSO [MY09]. Thresholded LASSO is an instance of two-stage variable selection schemes we study in this paper. In their paper, Meinshausen and Yu proved that thresholded LASSO can offer a variable selection consistency under weaker conditions than the irrepresentable condition required by LASSO. As we will see later, even the thresholded LASSO does not offer variable selection consistency under the asymptotic framework of this paper. However, we will show that it outperforms the LASSO in variable selection. Other authors have also studied two-step or even multi-step thresholding schemes for variable selection in the hope of weakening the required conditions [Zho09, Z+09]. Note that none of these methods can provide consistent variable selection under the linear asymptotic setting we consider in this paper. Study and comparison of these other schemes under our high-dimensional setting is an interesting open problem for future research.

A more delicate study of the LASSO estimator, or the bridge estimators, is necessary for an accurate analysis of two-stage methods under linear asymptotic regime of this paper. Our analysis relies on the recent results in the study of bridge estimators [DMM09, DMM11, BM11, BM12, WMZ16, MAYB13, SBC15]. These papers use the platform offered by approximate message passing (AMP) to characterize sharp asymptotic properties. In par-
ticular, the most relevant work to our paper is [SBC15] which studies the solution path of LASSO through the trade-off diagram of the asymptotic FDP and TPP. One result of this paper has been adopted in our analysis.

Finally, another line of work that is similar to the two-stage methods discussed in this paper, is the idea of screening [FL08, WR09]. For instance, in [FL08] a preliminary estimate of the $j$th regression coefficient is obtained by regressing $y$ on only the $j$th predictor. Then a hard threshold function is applied to all the estimates to infer the location of the non-zero coefficients. As we will discuss in Section 4.2, this approach is a special form of our two-stage variable selection with a debiasing performed in the first stage, and hence our variable selection technique under appropriate tuning outperforms Sure Independence Screening of [FL08]. Compared to Sure Independence Screening, the work of [WR09] uses more complicated estimators in the first stage, which is more aligned to our approach. However, [WR09] requires data splitting. While we may make data splitting to achieve certain theoretical improvement, in practice (especially in high-dimensions) data splitting may degrade the performance of a variable selection technique. In this paper we avoid data splitting.

2. Our Asymptotic Framework and Some Preliminaries.

2.1. Asymptotic framework. In this section, we review the asymptotic framework under which our studies are performed. We start with the definition of a converging sequence adapted from [BM12].

**Definition 2.1.** The sequence of instances $\{\beta(p), w(p), X(p)\}_{p \in \mathbb{N}}$, indexed by $p$, is said to be a standard converging sequence if

(a) $n = n(p)$ such that $\frac{n}{p} \to \delta \in (0, \infty)$.

(b) The empirical distribution of the entries of $\beta(p)$ converges weakly to a distribution $p_B$ with finite second moment. Further, $\frac{1}{n} \sum_{i=1}^{n} \beta_i(p)^2$ converges to the second moment of $p_B$; and $\frac{1}{p} \sum_{i=1}^{p} I(\beta_i(p) = 0) \to p_B(\{0\})$.

(c) The empirical distribution of the entries of $w(p)$ converges weakly to a zero mean distribution with variance $\sigma^2$. And $\frac{1}{n} \sum_{i=1}^{n} w_i(p)^2 \to \sigma^2$.

(d) $X_{ij}(p) \overset{i.i.d.}{\sim} N(0, \frac{1}{n})$.

We should emphasize that the condition $\frac{1}{p} \sum_{i=1}^{p} I(\beta_i(p) = 0) \to p_B(\{0\})$ in (b) is merely used for Lemma 3.1 and Lemma 3.2. In the rest of the paper, we assume the vector of regression coefficients $\beta$ is sparse. In particular, with the Definition 2.1 we will assume $p_B = (1 - \epsilon)\delta_0 + \epsilon p_G$, where $\delta_0$ denotes
which regularizer is optimal for variable selection?

A point mass at 0 and \( p_G \) is a distribution without any point mass at 0. Accordingly, the mixture proportion \( \epsilon \) represents the sparsity level of \( \beta(p) \) in the converging sequence. In Section 3, we will specify the conditions on \( p_G \) for studying the two-stage variable selection method. We next define the pseudo-Lipschitz function.

**Definition 2.2 (pseudo-Lipschitz function).** A function \( \psi : \mathbb{R}^2 \to \mathbb{R} \) is said to be pseudo-Lipschitz, if \( \exists L > 0 \) s.t., \( \forall x, y \in \mathbb{R}^2, |\psi(x) - \psi(y)| \leq L(1 + \|x\|_2 + \|y\|_2)\|x - y\|_2 \).

The following theorem proved by [BM11] and [WMZ16] is one of the main results that we will use in this paper.

**Theorem 2.1.** ([BM11], [WMZ16]) For a given \( q \in [1, \infty) \), let \( \hat{\beta}(q, \lambda) \) be the bridge estimator defined in (1.2). Consider a converging sequence \( \{\beta(p), X(p), w(p)\} \). Then, for any pseudo-Lipschitz function \( \psi : \mathbb{R}^2 \to \mathbb{R} \), almost surely

\[
\lim_{p \to \infty} \frac{1}{p} \sum_{i=1}^{p} \psi \left( \hat{\beta}_i(q, \lambda), \beta_i(p) \right) = \mathbb{E}(\eta_q(B + \tau Z; \alpha \tau^{2-q}), B),
\]

where \( B \sim p_B \) and \( Z \sim N(0, 1) \) are two independent random variables; \( \alpha \) and \( \tau \) are two positive numbers satisfying the following equations:

\[
\begin{align*}
\tau^2 &= \sigma^2 + \frac{1}{\delta} \mathbb{E}(\eta_q(B + \tau Z; \alpha \tau^{2-q}) - B)^2, \\
\lambda &= \alpha \tau^{2-q} \left( 1 - \frac{1}{\delta} \mathbb{E}\eta'_q(B + \tau Z; \alpha \tau^{2-q}) \right),
\end{align*}
\]

with \( \eta_q(\cdot; \cdot) \) being the proximal operator defined as

\[
\eta_q(u; \chi) = \arg\min_z \frac{1}{2}(u - z)^2 + \chi|z|^q,
\]

and \( \eta'_q(\cdot; \cdot) \) is the derivative of \( \eta_q \) with respect to its first argument.

[BM11] proves this theorem for LASSO. [WMZ16] extends the result to \( q \in (1, 2] \) by using the similar proof ideas proposed in [BM11]. The proof for the case \( q > 2 \) is similar to that in [WMZ16]. Hence, we do not repeat it in this paper.
2.2. Notations and Preliminaries. Before discussing our results, we would like to clarify and summarize our notations. As mentioned above, we will always assume the following sparse structure

\[ p_B = (1 - \epsilon)\delta_0 + \epsilon p_G. \]

\( B \) and \( G \) will be used as random variables with distribution \( p_B \) and \( p_G \) respectively. We use \( Z \) to represent a standard normal random variable. Subscripts like \( i \) attached to a vector are used to denote its \( i \)th component.

We define the asymptotic mean square error (AMSE) of \( \hat{\beta}(q, \lambda) \) as the almost sure limit

\[ (2.4) \quad \text{AMSE}(q, \lambda) \triangleq \lim_{p \to \infty} \frac{1}{p} \| \hat{\beta}(q, \lambda) - \beta \|^2_2. \]

According to Theorem 2.1, \( \text{AMSE}(q, \lambda) \) is well defined for \( q \in [1, \infty) \) and \( \lambda > 0 \). In this paper, one of our focuses will be on bridge estimators with optimal tuning \( \lambda^*_q \) defined as

\[ \lambda^*_q \triangleq \arg \min_{\lambda} \text{AMSE}(q, \lambda). \]

Further, we denote the thresholded estimators as

\[ \bar{\beta}(q, \lambda, s) = \eta_0(\hat{\beta}(q, \lambda); s^2/2) = \hat{\beta}(q, \lambda) \mathbf{1}_{\{ |\hat{\beta}(q, \lambda)| \geq s \}}. \]

For an estimator \( \hat{\beta} \) returned by a given variable selection technique, following [SBC15], we measure its variable selection performance by the false discovery proportion (FDP) and true positive proportion (TPP), defined in the following way:

\[ \text{FDP}(\hat{\beta}) = \frac{\# \{ \hat{\beta}_i \neq 0, \beta_i = 0 \}}{\# \{ \hat{\beta}_i \neq 0 \}}, \quad \text{TPP}(\hat{\beta}) = \frac{\# \{ \hat{\beta}_i \neq 0, \beta_i \neq 0 \}}{\# \{ \hat{\beta}_i \neq 0 \}}. \]

In particular, under current asymptotic setting, our study will be focused on the asymptotic version of FDP and TPP for LASSO \( \hat{\beta}(1, \lambda) \) and thresholded estimators \( \bar{\beta}(q, \lambda, s) \). Hence, we define

\[ \text{AFDP}(1, \lambda) = \lim_{p \to \infty} \text{FDP}(\hat{\beta}(1, \lambda)), \quad \text{AFDP}(q, \lambda, s) = \lim_{p \to \infty} \text{FDP}(\bar{\beta}(q, \lambda, s)). \]

Similar definitions are used for \( \text{ATPP}(1, \lambda) \) and \( \text{ATPP}(q, \lambda, s) \). Finally, for each tuning parameter \( \lambda > 0 \), [BM12] and [WMZ16] have proved that the solution pair \( (\alpha, \tau) \) to the nonlinear Equations (2.1) and (2.2) is unique. We will denote such unique solution pair for the optimal tuning value \( \lambda = \lambda^*_q \) by \( (\alpha_q, \tau_q) \). Notice that we omit the dependency of these two quantities on \( q \), since when they appear in this paper, \( q \) is clear from the context.
3. Our Contribution. The main objective of this paper is to study and compare the performance of the two-stage variable selection techniques described in Section 1.1, under the asymptotic setting formally introduced in Section 2.

3.1. Characterization of AFDP and ATPP and its implications. First note that since under our asymptotic setting the exact recovery of the non-zero locations of $\beta$ is impossible with LASSO [Wai09, RG13], we expect to observe both false positives and false negatives in variable selection. The following result with a minor modification from [BvdBSC13] characterizes the asymptotic FDP and TPP for LASSO.

**Lemma 3.1.** Recall the definition of $\text{AFDP}(1, \lambda)$ and $\text{ATPP}(1, \lambda)$ in Section 2.2. Then, for any given $\lambda > 0$, almost surely

$$\text{AFDP}(1, \lambda) = \frac{(1 - \epsilon)\mathbb{P}(|Z| > \alpha)}{(1 - \epsilon)\mathbb{P}(|Z| > \alpha) + \epsilon\mathbb{P}(|G + \tau Z| > \alpha\tau)},$$

$$\text{ATPP}(1, \lambda) = \mathbb{P}(|G + \tau Z| > \alpha\tau),$$

where $(\alpha, \tau)$ is the unique solution to Equations (2.1) and (2.2) with $q = 1$.

The formulas in this lemma have been derived in terms of converge in probability in [BvdBSC13]. We show they hold almost surely by a straightforward adaption of the proof in [BvdBSC13] in Appendix C.1. Note that one of the main goals of this paper is to compare the performance of two-stage variable selection techniques with LASSO. Hence, in our next lemma we derive the asymptotic FDP and TPP of the thresholded estimate $\hat{\beta}(q, \lambda, s)$. Recall the definition of $\text{AFDP}(q, \lambda, s)$ and $\text{ATPP}(q, \lambda, s)$ in Section 2.2.

**Lemma 3.2.** For any given $q \in [1, \infty), \lambda > 0, s > 0$, almost surely

$$\text{AFDP}(q, \lambda, s) = \frac{(1 - \epsilon)\mathbb{P}(\eta_q(|Z|; \alpha) > \frac{s}{\tau})}{(1 - \epsilon)\mathbb{P}(\eta_q(|Z|; \alpha) > \frac{s}{\tau}) + \epsilon\mathbb{P}(\eta_q(G + \tau Z; \alpha\tau^{2-q}) > s)},$$

$$\text{ATPP}(q, \lambda, s) = \mathbb{P}(\eta_q(G + \tau Z; \alpha\tau^{2-q}) > s),$$

where $(\alpha, \tau)$ is the unique solution to Equations (2.1) and (2.2).

The proof of this lemma is presented in Appendix C.2. We remind the reader that our goal is to compare different variable selection techniques. Toward this goal, we set ATPP to a fixed value $\zeta \in [0, 1]$ for different variable selection schemes and then compare their AFDPs. The two-stage procedures may have many different ways for setting ATPP to $\zeta$. If $q > 1$, Lemma 3.2
shows that for every given value of the regularization parameter $\lambda$, we can set $s$ (the threshold parameter) in a way that it returns the right level of ATPP. Therefore, in order to have a fair comparison among all pairs $(\lambda, s)$ that achieve the same ATPP level $\zeta$, we search for the one that minimizes AFDP. Next theorem explains how this optimal pair can be found.

**Theorem 3.1.** Consider $q \in (1, \infty)$. Given a ATPP level $\zeta \in [0, 1]$, for every value of $\lambda > 0$ there exists $s = s(\lambda, \zeta)$ such that ATPP($q, \lambda, s$) = $\zeta$. Furthermore, the value of $\lambda$ that minimizes AFDP($q, \lambda, s(\lambda, \zeta)$) also minimizes AMSE($q, \lambda$).

The proof of this theorem can be found in Appendix D.1. Note that the important case $q = 1$ is not covered in Theorem 3.1. Hence, before discussing the implications of this theorem, we state the impact of the second stage thresholding on the variable selection performance of LASSO.

**Theorem 3.2.** For any $\zeta \in [0, \text{ATPP}(1, \lambda^*_1)]$, there exists at least one $\lambda$ such that ATPP($1, \lambda$) = $\zeta$. Furthermore, there exists a unique $s = s(\lambda, \zeta)$ such that ATPP($1, \lambda^*_1, s$) = $\zeta$. Among all these estimators, the one that offers the minimal AFDP is $\bar{\beta}(1, \lambda^*_1, s(\lambda, \zeta))$, i.e., the two-stage LASSO with the optimal tuning value $\lambda = \lambda^*_1$.

The proof of this theorem can be found in Appendix D.2. There are a couple of points we would like to emphasize here:

(i) Consider a two-stage variable selection technique. According to Theorems 3.1 and 3.2, for $q \in (1, \infty)$, the optimal choice of $\lambda$ does not depend on the ATPP level $\zeta$ we are interested in. It minimizes the AMSE of the bridge estimate from the first stage. Even for $q = 1$, the optimal choice of $\lambda$ is independent of $\zeta$ in large range of ATPPs.

(ii) Given that the design matrix has iid elements, minimizing AMSE is equivalent to minimizing out-of-sample prediction error. Hence, intuitively speaking, one can use model selection methods that have been developed to optimize the out-of-sample prediction error, such as cross validation, to tune $\lambda$. It has also been recently shown that $\lambda^*_1$ can be estimated accurately via approximate message passing framework [MMB15]. Following the idea of [MMB15], as will be seen in Section 5, we propose a consistent estimate (up to some constants) for AMSE with $q \in [1, \infty)$. Then a grid search of $\lambda$ will be used to optimize that estimate for optimal tuning.

(iii) A simple implication of Theorem 3.2 is, for a wide range of $\zeta$, having a second threshold can help the variable selection of LASSO. Figure 3.1
WHICH REGULARIZER IS OPTIMAL FOR VARIABLE SELECTION?

0.0 0.2 0.4 0.6 0.8 1.0
ATPP 0.0 0.2 0.4 0.6 0.8 1.0
AFDP \( \sigma = 0.5 \)
Lasso
2S-Lasso

Fig 3.1. Comparison of AFDP-ATPP curve between LASSO and two-stage LASSO. Here we pick the setting \( \delta = 0.8, \epsilon = 0.3, \sigma \in \{0.5, 0.22, 0.15\}, p_C = \delta_1 \). For two-stage LASSO, we use optimal tuning \( \lambda_1^* \) in the first stage. All the curves are calculated based on Equations (3.1) and (3.2). The gray dotted line is the upper bound of ATPP that the two-stage LASSO can reach. Notice that even for LASSO, there is an upper bound which it cannot exceed.

compares the AFDP-ATPP curve of LASSO with that of the two-stage LASSO. As is clear in this figure, when the signal-to-noise ratio (SNR) is high, the gap between the performance of two-stage LASSO and LASSO is very large. A similar conclusion has been also mentioned in [MY09]. We should emphasize that the two-stage LASSO (with optimal tuning) can not achieve the level of ATPP beyond that of \( \hat{\beta}(1, \lambda_1^*) \). We discuss debiasing to resolve this issue in Section 4.

Theorems 3.1 and 3.2 prove that the optimal way to use two-stage variable selection technique is to set \( \lambda = \lambda_q^* \) for the regularization parameter in the first stage. Note that \( \lambda_q^* \) minimizes \( \text{AMSE}(q, \lambda) \), i.e. is the optimal tuning for parameter estimation. We thus see an interesting connection between variable selection and parameter estimation. In the rest of the paper we will use the notation \( s_q^*(\zeta) \) for the value of threshold that satisfies

\[
\text{ATPP}(q, \lambda_q^*, s_q^*(\zeta)) = \zeta.
\]

So far we have addressed the first two questions raised in Section 1.1. Our next goal is to compare the performance of two-stage variable selection techniques for different values of \( q \).

3.2. Optimality of bridge estimator for variable selection. We now discuss our answer to the third question and compare the performance of two-stage variable selection methods built upon bridge estimators. Consider \( q_1, q_2 \in [1, \infty) \). We would like to compare \( \text{AFDP}(q_1, \lambda_q^*, s_{q_1}^*(\zeta)) \) and \( \text{AFDP}(q_2, \lambda_q^*, s_{q_2}^*(\zeta)) \). Our first result shows the equivalence of the variable selection performance and the estimation performance of bridge estimators.
Theorem 3.3. Let $q_1, q_2 \geq 1$. If $\text{AMSE}(q_1, \lambda_{q_1}^*) < \text{AMSE}(q_2, \lambda_{q_2}^*)$, then for every $\zeta \in [0, 1]$

$$\text{AFDP}(q_1, \lambda_{q_1}^*, s_{q_1}^*(\zeta)) \leq \text{AFDP}(q_2, \lambda_{q_2}^*, s_{q_2}^*(\zeta)),$$

where the equality only holds for $\zeta \in \{0, 1\}$.

The proof of this result is presented in Appendix D.3. Note that $s_{q}^*(\zeta)$ may not exist for certain range of $\zeta$. As will be seen in Section 4, a debiasing can overcome the limitation and hence we can interpret the result in terms of the debiased version whenever $s_{q}^*(\zeta)$ is not well defined. According to Theorem 3.3, in order to see which two-stage method is better, we can compare their AMSE with optimal tuning $\lambda_q^*$. From Theorem 2.1 we know such AMSE is given by

$$\text{(3.3) } \text{AMSE}(q, \lambda_q^*) = E \left( \eta_q (B + \tau^* Z; \alpha^* \tau_2^{-q}) - B \right)^2,$$

where $\tau^*$ and $\alpha^*$ satisfy Equations (2.1) and (2.2) with $\lambda = \lambda_q^*$.

Unfortunately, the complicated form of the AMSE and its dependence on many factors like $\delta, \sigma, p_B$ make it difficult to discover the behavior of AMSE and compare it for different values of $q$. To address this issue, in the rest of the paper we consider a few limiting cases. As will be discussed in this section and Section 5, such analysis sheds light on the behavior of two-stage methods under different scenarios and discovers the performance pattern even when the settings are not very similar to the limiting cases. Below we summarize these limiting cases:

(i) High signal-to-noise ratio (SNR): We assume that the noise level $\sigma$ is small (compared to the variance of the response). Such analysis has been performed in [WMZ16].

(ii) Low SNR: We assume that the noise level $\sigma$ is very large. This regime is in sharp contrast to the previous case and helps us understand the behavior of two-stage variable selection techniques for very noisy datasets.

(iii) Large sample regime: We consider the case when the sample size $\delta$ (relative to the covariates dimension $p$) is large. This regime, as will be seen, is closely related to the classical asymptotic regime where we assume $n \to \infty$ and $p$ fixed.

We now present the theorems characterizing AMSE for each of the three scenarios respectively. In Section 5 we explore and confirm the implications of these results for more practical problem settings through simulations.
3.2.1. **Analysis of AMSE in small noise scenario.** In this section we study the AMSE for a small noise. Next theorem summarizes the main result of this section.

**Theorem 3.4.** Assume $\epsilon \in (0,1)$. As $\sigma \to 0$, we have the following expansions of $\text{AMSE}(q, \lambda^*_q)$ in terms of $\sigma$.

(i) For $q = 1$, if $\mathbb{P}(|G| \geq \mu) = 1$ for some $\mu > 0$, $\delta > M_1(\epsilon)$, and $\mathbb{E}|G|^2 < \infty$, then

$$\text{AMSE}(1, \lambda^*_1) = \frac{\delta M_1(\epsilon)}{\delta - M_1(\epsilon)} \sigma^2 + o\left(\frac{(M_1(\epsilon) - \delta)\bar{\mu}^2}{2\delta^2}\right),$$

where $\bar{\mu}$ can be any positive number smaller than $\mu$.

(ii) For $1 < q < 2$, if $\mathbb{P}(|G| \leq x) = O(x)$ (as $x \to 0$), $\delta > 1$, and $\mathbb{E}|G|^2 < \infty$ then

$$\text{AMSE}(q, \lambda^*_q) = \frac{\sigma^2}{1 - 1/\delta} - \sigma^2 \frac{\delta^{q+1}(1-\epsilon)^2}{(\delta - 1)^q + 1} e \mathbb{E}|G|^{2q-2} + o(\sigma^2).$$

(iii) For $q = 2$, if $\delta > 1$ and $\mathbb{E}|G|^2 < \infty$, we have

$$\text{AMSE}(2, \lambda^*_2) = \frac{\sigma^2}{1 - 1/\delta} - \sigma^4 \frac{\delta^3}{(\delta - 1)^3 e \mathbb{E}|G|^2} + o(\sigma^4).$$

(iv) For $q > 2$, if $\delta > 1$ and $\mathbb{E}|G|^{2q-2} < \infty$, then

$$\text{AMSE}(q, \lambda^*_q) = \frac{\sigma^2}{1 - 1/\delta} - \sigma^4 \frac{\delta^3 e(q-1)^2(\mathbb{E}|G|^{q-2})^2}{(\delta - 1)^3 \mathbb{E}|G|^{2q-2}} + o(\sigma^4).$$

The results for $q \in [1,2]$ are taken from [WMZ16]. We give the proof for the case $q > 2$ in Appendix F. As shown in [WMZ16], $M_1(\epsilon)$ is an increasing function of $\epsilon \in [0,1]$ and $M_1(1) = 1$. This implies that $\text{AMSE}(1, \lambda^*_1)$ is the smallest among all $\text{AMSE}(q, \lambda^*_q)$ with $q \in [1, \infty)$. As is clear, the first order terms in the expansion of $\text{AMSE}(q, \lambda^*_q)$ are the same for all $q \in (1, \infty)$. However, the second dominant term shows that the smaller values of $q$ are preferable (note the strict monotonicity only occurs in the range $(1,2)$).

Combining the above results with Theorem 3.3 implies that in the high SNR setting, two-stage LASSO offers the best variable selection performance. We should also emphasize that as depicted in Figure 3.1, in this regime two-stage LASSO offers a much better variable selection performance than LASSO.

\[ M_1(\epsilon) = \min_{\chi} (1-\epsilon) \mathbb{E} q_1^2(Z \cdot \chi) + \epsilon (1 + \chi^2) \text{ with } Z \sim N(0,1). \]
3.2.2. Analysis of AMSE in large noise scenario. This section is devoted to the analysis of the optimal AMSE of bridge estimators in the low SNR regime. The following theorem is the main result of this section.

**Theorem 3.5.** As $\sigma \to \infty$, we have the following expansions of $\text{AMSE}(q, \lambda_q^*)$ in terms of $\sigma$.

(i) For $q = 1$, when $G$ has sub-Gaussian tail, we have

\[ \text{AMSE}(1, \lambda_1^*) = \epsilon \mathbb{E}|G|^2 + o(e^{-\frac{C^2 \sigma^2}{2}}), \]

where $C$ is any positive number smaller than $C_0$, with $C_0$ a constant only depending on $\epsilon$ and $G$.

(ii) For $1 < q \leq 2$, if all the moments of $G$ are finite, then

\[ \text{AMSE}(q, \lambda_q^*) = \epsilon \mathbb{E}|G|^2 - \frac{c_q^2 \mathbb{E}|G|^2}{\sigma^2} + o(\sigma^{-2}), \]

with $c_q = \left(\frac{\mathbb{E}|Z|^{2q}}{(q-1)^{2} \mathbb{E}|Z|^{q-1}}\right)^{2}$. 

(iii) For $q > 2$, if $G$ has sub-Gaussian tail, then (3.9) holds.

The proofs are different for $q = 1$, $1 < q \leq 2$ and $q > 2$. Hence, we present our proofs for these three cases in Appendix E.2, E.3, and E.4 respectively. Figure 3.2 compares the accuracy of the first order approximation and second order approximation for moderate values of $\sigma$. As is clear, for $q \in (1, \infty)$ the second order approximation provides a more accurate approximation for a wide range of $\sigma$. Moreover, the first order approximation of LASSO is already very accurate as can be justified by its exponentially small second order term in (3.8).

According to this theorem, we can conclude that for sufficiently large $\sigma$, two-stage method with any $q > 1$ can outperform the two-stage LASSO. This is because while the first dominant term is the same for all the bridge estimators with $q \in [1, \infty)$, the second order term for LASSO is exponentially smaller (in magnitude) than that of the other values of $q$. More interestingly, the following lemma shows that in fact $q = 2$ leads to the best AMSE (and consequently variable selection performance) in the large noise regime.

**Lemma 3.3.** The maximum of $c_q$, defined in Theorem 3.5, is achieved at $q = 2$.

---

2 Refer to the proof of this theorem for the exact characterization of $C_0$. 
WHICH REGULARIZER IS OPTIMAL FOR VARIABLE SELECTION?

Fig 3.2. Absolute relative error of 1st order and 2nd order approximations of AMSE under large noise scenario. The orange curves and purple curves display the relative errors $\frac{|E|C^2 - AMSE(q, \lambda_q^*)|}{AMSE(q, \lambda_q^*)}$ and $\frac{|E|C^2 - \sqrt{E|C|^2} - AMSE(q, \lambda_q^*)|}{AMSE(q, \lambda_q^*)}$ respectively, for different values of $\sigma$. In these four figures, $p_B = (1 - \epsilon)\delta_0 + \epsilon\delta_1$, $\delta = 0.4$, $\epsilon = 0.2$.

**Proof.** A simple integration by part yields:

$$E[Z]^{\frac{2-q}{q-1}} = 2 \int_0^\infty z^{\frac{2-q}{q-1}} \phi(z) dz = 2(q-1) \int_0^\infty \phi(z)dz \frac{1}{z^{\frac{1}{q-1}}}$$

$$= 2(q-1) \int_0^\infty z^{\frac{q}{q-1}} \phi(z)dz = (q-1)E[Z]^{\frac{q}{q-1}}$$

We can then apply Hölder’s inequality to obtain

$$c_q = \left(\frac{E[Z]^{\frac{q}{q-1}}}{E[Z]^{\frac{2}{q-1}}}\right)^2 \leq \frac{E[Z]^{\frac{2}{q-1}} E Z^2}{E[Z]^{\frac{2}{q-1}}} = 1 = c_2.$$
Therefore while the AMSE of all bridge estimators share the same first dominant term, ridge offers the largest second dominant term (in magnitude), and hence the lowest AMSE. If we combine this result with Theorem 3.3, we conclude that in low SNR regime, ridge obtains the best variable selection performance among two-stage variable selection schemes with their first stage picked from the class of bridge estimators.

3.2.3. Analysis of AMSE in large sample scenario. Our final limiting analysis is concerned with the large $\delta$ regime. Since $n/p \rightarrow \delta$ in our asymptotic setting, large $\delta$ means large sample size (relative to the dimension $p$). Intuitively speaking, this is similar to the classical asymptotic setting where $n \rightarrow \infty$ and $p$ is fixed (specially if we assume the fixed number $p$ is large). We will later connect the results we derive in the large $\delta$ regime to those obtained in classical asymptotic regime, and provide some new insights.

In our original set-up the elements of the design matrix are $X_{ij} \sim_{i.i.d.} N(0, \frac{1}{n})$. This means the SNR \( \frac{\text{var}(\sum_j X_{ij} \beta_j)}{\text{var}(w_i)} \rightarrow \frac{E|B|^2}{\delta \sigma^2} \) as $n \rightarrow \infty$. Therefore, if we let $\delta \rightarrow \infty$, the SNR will decrease to zero, which is not consistent with the classical asymptotic setting in which the SNR is assumed to be fixed. To resolve this discrepancy we scale the noise term by $\sqrt{\delta}$ and use the model:

\begin{equation}
y = X \beta + \frac{1}{\sqrt{\delta}} w,
\end{equation}

where \( \{\beta, w, X\} \) is the converging sequence in Definition 2.1. Under this model we compare the AMSE of different bridge estimators. The next theorem summarizes the main result of this section.
Theorem 3.6. Consider the model in (3.10) and \( \epsilon \in (0, 1) \). As \( \delta \to \infty \), we have the following expansions.

(i) For \( q = 1 \), if \( \mathbb{P}(|G| \geq \mu) = 1 \) for some \( \mu > 0 \) and \( \mathbb{E}|G|^2 < \infty \), then

\[
\text{AMSE}(1, \lambda_1^*) = \frac{M_1(\epsilon)\sigma^2}{\delta} + o(\delta^{-1}),
\]

where \( M_1(\epsilon) \) is the same as in Theorem 3.4.

(ii) For \( 1 < q < 2 \), if \( \mathbb{P}(|G| \leq x) = O(x) \) (as \( x \to 0 \)) and \( \mathbb{E}|G|^2 < \infty \), then

\[
\text{AMSE}(q, \lambda_q^*) = \frac{\sigma^2}{\delta} - \frac{\sigma^2 q (1 - \epsilon)^2 (\mathbb{E}|Z|^q)^2}{\epsilon \mathbb{E}|G|^{2q-2}} + o(\delta^{-q})
\]

(iii) For \( q = 2 \), if \( \mathbb{E}|G|^2 < \infty \), then we have

\[
\text{AMSE}(2, \lambda_2^*) = \frac{\sigma^2}{\delta} + \frac{\sigma^2}{\delta^2} [1 - \frac{\sigma^2}{\epsilon \mathbb{E}|G|^2}] + o(\delta^{-2})
\]

(iv) For \( q > 2 \), if \( \mathbb{E}|G|^{2q-2} < \infty \), then

\[
\text{AMSE}(q, \lambda_q^*) = \frac{\sigma^2}{\delta} + \frac{\sigma^2}{\delta^2} [1 - \frac{\epsilon(q-1)^2 \sigma^2 (\mathbb{E}|G|^{q-2})^2}{\mathbb{E}|G|^{2q-2}}] + o(\delta^{-2}).
\]

The proof of Theorem 3.6 can be found in Appendix F. Figure 3.4 compares the accuracy of the first and second order expansions in large range of \( \delta \). As is clear from this figure, the second order term often offers a more accurate approximation over a wide range of \( \delta \).

Remark 3.1. As shown in [WMZ16], \( M_1(\epsilon) \) is an increasing function of \( \epsilon \in [0, 1] \) and \( M_1(1) = 1 \). This implies that \( \text{AMSE}(1, \lambda_1^*) \) is the smallest among all \( \text{AMSE}(q, \lambda_q^*) \) with \( q \in [1, \infty) \). Hence, in this regime LASSO gives the smallest estimation error and thus two-stage LASSO offers the best variable selection performance.

Remark 3.2. The \( \text{AMSE}(q, \lambda_q^*) \) with \( q > 1 \) share the same first dominant term, but have different second order terms. Furthermore, for \( q \in (1, 2) \) the smaller \( q \) is, the better its performance will be. Such monotonicity does not hold beyond \( q = 2 \).

We now connect our results in this large \( \delta \) regime to those obtained in classical asymptotic setting. The classical asymptotics (\( p \) fixed) of bridge
estimators for all the values of $q \in [0, \infty)$ is studied in [KF00]. We explain LASSO first. According to [KF00], if $\frac{\lambda}{\sqrt{n}} \to \lambda_0 \geq 0$ and $\frac{1}{n}X^TX \to C$, then

$$\sqrt{n}(\hat{\beta} - \beta) \overset{d}{\to} \arg\min_u V(u),$$

where $V(u) = -2u^TW + u^TCu + \lambda_0 \sum_{j=1}^p [u_j \text{sgn}(\beta_j)1_{\{\beta_j \neq 0\}} + |u_j|1_{\{\beta_j = 0\}}]$ with $W \sim \mathcal{N}(0, \sigma^2C)$. We will do the following calculations to explore the connections. Since $X_{ij} \sim \mathcal{N}(0, 1/n)$ in our paper, we first make the following
changes to LASSO to make our set-up consistent with that of [KF00]:

\[
\frac{1}{2} \|y - X\beta\|_2^2 + \lambda \|\beta\|_1 = \frac{1}{2} \left( \|y - \sqrt{n}X \frac{\beta}{\sqrt{n}}\|_2^2 + 2\sqrt{n}\lambda \|\beta\|_1 \right).
\]

We thus have \[ C = \frac{1}{n}(\sqrt{n}X)^T(\sqrt{n}X) \rightarrow I \] and \( \lambda_0 = 2\lambda \). Now suppose the result (3.15) works for \( \hat{\beta}(1, \lambda) \). Then we have

\[
(3.16) \quad \hat{\beta}(1, \lambda) - \beta \overset{d}{\rightarrow} \arg\min_u V(u),
\]

where \( V(u) = -2u^TW + u^Tu + 2\lambda \sum_{j=1}^p [u_j \text{sgn}(\beta_j) \mathbb{1}_{\{\beta_j \neq 0\}} + |u_j| \mathbb{1}_{\{\beta_j = 0\}}] \) with \( W \sim \mathcal{N}(0, \frac{\sigma^2}{\delta} I) \). It is straightforward to see that the optimal choice of \( u \) in (3.16) has the following form:

\[
\hat{u}_j = \begin{cases} 
W_j - \lambda \text{sgn}(\beta_j) & \text{when } \beta_j \neq 0 \\
W_j - \lambda s(\hat{u}_j) & \text{when } \beta_j = 0
\end{cases}
\]

where \( s(u_j) = \text{sgn}(u_j) \) when \( u_j \neq 0 \) and \( |s(u_j)| \leq 1 \) when \( u_j = 0 \). Furthermore, for the case of \( \beta_j = 0, \hat{u}_j = 0 \) is equivalent to \( |W_j| \leq \lambda \) and \( \text{sgn}(W_j) = \text{sgn}(\hat{u}_j) \) when \( \hat{u}_j \neq 0 \). Based on this result one can do the following heuristic calculation to connect our results with those of [KF00] (one can add some extra conditions to make these heuristic calculations more accurate. But, that is out of the scope of our paper):

\[
\frac{1}{p} \|\hat{\beta}(1, \lambda) - \beta\|_2^2 \approx \frac{1}{p} \mathbb{E} \left[ \sum_{j: \beta_j \neq 0} [W_j^2 - 2\lambda \text{sgn}(\beta_j)W_j + \lambda^2] \right]
\]

\[
+ \sum_{j: \beta_j = 0, \hat{u}_j \neq 0} [W_j^2 - 2\lambda W_j \text{sgn}(\hat{u}_j) + \lambda^2]
\]

\[
\approx \frac{1}{p} \sum_{j: \beta_j \neq 0} \left( \frac{\sigma^2}{\delta} + \lambda^2 \right) + \sum_{j: \beta_j = 0} \mathbb{E} \eta_1^2(W_j; \lambda)
\]

\[
= \frac{k}{p} \left( \frac{\sigma^2}{\delta} + \lambda^2 \right) + \frac{p - k}{p} \mathbb{E} \eta_1^2(W_j; \lambda)
\]

\[
= \frac{\sigma^2}{\delta} \left[ \frac{p - k}{p} \mathbb{E} \eta_1^2(Z; \sqrt{\delta}\lambda/\sigma) + \frac{k}{p}(1 + (\sqrt{\delta}\lambda/\sigma)^2) \right],
\]

where \( k \) is the number of non-zero elements of \( \beta \) and \( Z \sim N(0, 1) \). Note that in our asymptotic setting \( k/p \rightarrow \epsilon \) and we consider the optimal tuning \( \lambda_1^* \).

Therefore following the above calculations we obtain

\[
\min_{\lambda} \frac{1}{p} \|\hat{\beta}(1, \lambda) - \beta\|_2^2 \approx \frac{\sigma^2}{\delta} \min \left( 1 - \epsilon \mathbb{E} \eta_1^2(Z; \chi) + \epsilon(1 + \chi^2) = \frac{M_1(\epsilon)\sigma^2}{\delta} \right).
\]
This is consistent with (3.11) in our asymptotic analysis. We can do similar calculations to show that the asymptotic analysis of [KF00] leads to the first order expansion of AMSE in Theorem 3.6 for the case \( q > 1 \).

We may conclude that the information provided by the classical asymptotic analysis is reflected in the first order term of AMSE\((q, \lambda^*_q)\). Moreover, our large sample analysis is able to derive the second dominant term for \( q > 1 \). This term enables us to compare the performance of different values of \( q > 1 \) more accurately (note they all have the same first order term). Such comparisons cannot be performed in [KF00].

4. Debiasing.

4.1. Implications of debiasing for LASSO. As is clear from Theorem 3.2, since LASSO produces a sparse solution, it is not possible for a LASSO based two-stage method to achieve ATPP values beyond what is already reached by the first stage. This problem can be resolved by debiasing. In this approach, instead of thresholding the LASSO estimate (or in general a bridge estimate), we threshold its debiased version. Below we will add a dagger \( \dagger \) to aforementioned notations to denote their corresponding debiased version.

Recall \( \hat{\beta}(q, \lambda) \) denotes the solution of bridge regression for any \( q \geq 1 \). Define the debiased estimates as

- For \( q = 1 \),
  \[
  \hat{\beta}^{\dagger}(1, \lambda) \triangleq \hat{\beta}(1, \lambda) + X^T \frac{y - X\hat{\beta}(1, \lambda)}{1 - \|\hat{\beta}(1, \lambda)\|_0/n},
  \]
  where \( \| \cdot \|_0 \) counts the number of non-zero elements in a vector.

- For \( q > 1 \),
  \[
  \hat{\beta}^{\dagger}(q, \lambda) \triangleq \hat{\beta}(q, \lambda) + X^T \frac{y - X\hat{\beta}(q, \lambda)}{1 - f(\hat{\beta}(q, \lambda), \hat{\gamma}_\lambda)/n},
  \]
  where
  \[
  f(v, w) = \sum_{i=1}^{p} \frac{1}{1 + w q (q-1)|v_i|^q |z_i|^q - 1}.
  \]
  and \( \gamma = \hat{\gamma}_\lambda \) is the unique solution of the following equation:
  \[
  \frac{\lambda}{\gamma} = 1 - \frac{1}{n} f(\hat{\beta}(q, \lambda), \gamma).
  \]

We have the following theorem to confirm the validity of the debiasing estimator \( \hat{\beta}^{\dagger}(q, \lambda) \).
Theorem 4.1. For any given \( q \in [1, \infty) \), with probability one, the empirical distribution of the components of \( \hat{\beta}^\dagger(q, \lambda) - \beta \) converges weakly to \( N(0, \tau^2) \), where \( \tau \) is the solution of (2.1) and (2.2).

See Appendix G for the proof. In order to perform variable selection, one may apply the hard thresholding function to these debiased estimates and obtain:

\[
\hat{\beta}^\dagger(q, \lambda, s) = \eta_0(\hat{\beta}^\dagger(q, \lambda); s^2/2) = \hat{\beta}^\dagger(q, \lambda) 1_{\{|\hat{\beta}^\dagger(q, \lambda)| \geq s\}}.
\]

We use the notations \( \text{ATTP}^\dagger(q, \lambda, s) \) and \( \text{AFDP}^\dagger(q, \lambda, s) \) to denote the ATTP and AFDP of \( \hat{\beta}^\dagger(q, \lambda, s) \) respectively. In the case of LASSO, note that unlike \( \hat{\beta}(1, \lambda) \) the debiased estimator \( \hat{\beta}^\dagger(1, \lambda) \) is dense, and hence as the first byproduct, we expect the two-stage variable selection estimate \( \hat{\beta}^\dagger(1, \lambda, s) \) to be able to reach any value of ATTP between \([0, 1]\). The following theorem confirms this claim.

Theorem 4.2. Given the ATTP level \( \zeta \in [0, 1] \), for every value of \( \lambda > 0 \), there exists \( s(\lambda, \zeta) \) such that \( \text{ATTP}^\dagger(1, \lambda, s(\lambda, \zeta)) = \zeta \). Furthermore, whenever \( \hat{\beta}^\dagger(1, \lambda, s) \) and \( \beta(1, \lambda, \tilde{s}) \) reach the same level of ATTP, they have the same AFDP. The value of \( \lambda \) that minimizes \( \text{AFDP}^\dagger(1, \lambda, s(\lambda, \zeta)) \) also minimizes \( \text{AMSE}(1, \lambda) \).

As expected since the solution of bridge regression for \( q > 1 \) is dense, the debiasing step does not help variable selection for \( q > 1 \). Our next theorem confirms this claim.

Theorem 4.3. Consider \( q > 1 \). Given the ATTP level \( \zeta \in [0, 1] \), for every value of \( \lambda > 0 \), there exists \( s(\lambda, \zeta) \) such that \( \text{ATTP}^\dagger(q, \lambda, s(\lambda, \zeta)) = \zeta \). Furthermore, whenever \( \hat{\beta}^\dagger(q, \lambda, s) \) and \( \beta(q, \lambda, \tilde{s}) \) reach the same level of ATTP, they have the same AFDP. Also, the value of \( \lambda \) that minimizes \( \text{AFDP}^\dagger(q, \lambda, s(\lambda, \zeta)) \) also minimizes \( \text{AMSE}(q, \lambda) \). As a result, the optimal value of \( \text{AFDP}^\dagger(q, \lambda, s(\lambda, \zeta)) \) is the same as \( \text{AFDP}(q, \lambda^s_q, s_q^s(\zeta)) \).

For the proof of Theorems 4.2 and 4.3, please refer to Appendix G.

Remark 4.1. Comparing Theorem 4.2 with Theorem 3.2, we see that replacing LASSO in the first stage with the debiased version enables to achieve wider range of ATTP level. On the other hand, given the value of \( \lambda \), if \( \hat{\beta}^\dagger(1, \lambda, s) \) and \( \beta(1, \lambda, \tilde{s}) \) reach the same level of ATTP, their AFDP are equal as well. Therefore, the debiasing for LASSO expands the range
Fig 4.1. Comparison of AFDP-ATPP curve between LASSO and two-stage debiased LASSO. Here we pick the setting $\delta = 0.8$, $\epsilon = 0.3$, $\sigma \in \{0.5, 0.22, 0.15\}$, $p_C = \delta_1$. For the two-stage debiased LASSO, we use optimal tuning $\lambda_1^*$ in the first stage. The gray dotted line is the upper bound for the two-stage LASSO without debiasing can reach.

Remark 4.2. The debiasing does not present any extra gain to the two-stage variable selection technique based on bridge estimators with $q > 1$. In other words, debiasing does not change the AFDP-ATPP curve for $q > 1$.

4.2. Debiasing and Sure Independence Screening. Sure Independence Screening (SIS) is a variable selection scheme proposed for ultra-high dimensional settings. Our asymptotic setting is not considered an ultra-high dimensional asymptotic. We are also aware that SIS is typically used for screening out irrelevant variables and other variable selection methods, such as LASSO, will be applied afterwards. Nevertheless, we present a connection and comparison between our two-stage methods and SIS in the linear asymptotic regime. Such comparisons shed more light on the performance of SIS. It is straightforward to confirm that Sure Independence Screening is equivalent to

$$\tilde{\beta}^\dagger(q, \infty, s) = \eta_0(\hat{\beta}^\dagger(q, \infty); s^2/2) = \eta_0(X^Ty; s^2/2).$$

Therefore, the main difference between the approach we propose in this paper and SIS, is that SIS sets $\lambda$ to $\infty$, while, we select the value of $\lambda$ that
WHICH REGULARIZER IS OPTIMAL FOR VARIABLE SELECTION? 21

Fig 4.2. Comparison of AFDP-ATPP curve between SIS and the two-stage debiased LASSO. Here we pick the setting \( \delta = 0.8, \epsilon = 0.3, \sigma \in \{0.5, 0.22, 0.15\}, p_G = \delta_1 \). For the two-stage debiased LASSO, we use optimal tuning \( \lambda_1^* \) in the first stage. The gray dotted line is the upper bound that the two-stage LASSO without debiasing can reach.

minimizes AMSE. This simple difference may give a major boost to the variable selection performance. The following lemma confirms this statement.

**Lemma 4.1.** Consider \( q \geq 1 \). Given any ATPP level \( \zeta \in [0, 1] \), let \( \text{AFDP}_{\text{sis}}(\zeta), \text{AFDP}^\dagger(q, \lambda, s(\lambda, \zeta)) \) denote the asymptotic FDP of SIS and debiased two-stage bridge estimator respectively, when their ATPP is equal to \( \zeta \). Then, \( \text{AFDP}^\dagger(q, \lambda_1^*, s(\lambda_1^*, \zeta)) \leq \text{AFDP}_{\text{sis}}(\zeta) \).

Refer to Appendix G for the proof. Note that when the variance of the noise \( \sigma^2 \) is large, we expect the optimally tuned \( \lambda \) to be large, and hence the performance of SIS is closer to our two-stage variable selection technique. However, as \( \sigma^2 \) decreases, the gain that we obtain from using a better estimator in the first stage improves. Figure 4.2 compares the performance of SIS and two-stage variable selection based on debiased LASSO under different noise settings.

5. Numerical experiments.

5.1. **Objective and Simulation Set-up.** This section aims to investigate the finite sample variable selection performances of various two-stage variable selection (TVS) estimators under the three different regimes that our analysis in Section 3.2 has been focused on. We will also study to what extent our theory works for more realistic situations in which \( \sigma \) is not “too
small” or “too large”, and δ is not “too large” either. The performances of different methods will be compared via the AFDP-ATPP curves. Note that since we are in the finite sample situations, the curve we calculate is actually FDP-TPP instead of the asymptotic version. With a little abuse of notation, we will call it AFDP-ATPP curve throughout the section.

The following general settings are adopted in this section.

1. Number of variables is fixed at $p = 5000$. Sample size $n = p\delta$ is then decided by $\delta$.
2. Given the values of $\delta$, $\epsilon$, $\sigma$ and the distribution $pG$, we sample $X \in \mathbb{R}^{n \times p}$ with $X_{ij} \sim \mathcal{N}(0, \frac{1}{n})$, $\beta \in \mathbb{R}^p$ with $\beta_i \sim pB = (1 - \epsilon)\delta_0 + \epsilon pG$, $w \in \mathbb{R}^n$ with $w_i \sim \mathcal{N}(0, \sigma^2)$ or $\mathcal{N}(0, \frac{\sigma^2}{\delta})$ depending on large/small noise or large sample scenarios. Construct $y = X\beta + w$.
3. For each $(y, X)$ pair, ATPP-AFDP curves will be generated for different variable selection methods. In each setting, 80 samples are drawn and the average ATPP-AFDP curves are calculated. Furthermore, the average differences $\text{AFDP}_q - \text{AFDP}_2$ (the subscript $q$ indexes the specific bridge estimator used in the two-stage variable selection) in large/small noise, the average difference $\text{AFDP}_q - \text{AFDP}_1$ in large sample scenarios, and their associated 1-standard deviation confidence interval will be plotted.

We compute bridge regression through coordinate descent algorithm, with the proximal operator $\eta_q(x; \tau)$ calculated through a properly implemented Newton’s method.

5.2. Estimating the optimal tuning $\lambda^*_q$. For two-stage variable selection procedures, it is critical to have a good estimator in the first step. One minor challenge here is to search for the optimal tuning that minimizes AMSE of $\hat{\beta}(q, \lambda)$. By comparing the result of Theorem 2.1 with the definition of AMSE in (2.4), it is straightforward to see that $\tau^2 = \sigma^2 + \frac{1}{\delta} \text{AMSE}$. Hence, one can minimize $\tau^2$ to achieve the same optimal tuning. By extending the result of [MMB15], we can obtain a consistent estimator of $\tau^2$:

$$
q = 1 : \quad \hat{\tau}^2 = \frac{1}{n} \left\| \frac{y - X\hat{\beta}(1, \lambda)}{1 - \|\hat{\beta}(1, \lambda)\|_0/n} \right\|_2^2,
$$

(5.1)

$$
q > 1 : \quad \hat{\tau}^2 = \frac{1}{n} \left\| \frac{y - X\hat{\beta}(1, \lambda)}{1 - f(\hat{\beta}(q, \lambda), \hat{\gamma}_\lambda)/n} \right\|_2^2,
$$

where $f(\cdot, \cdot)$, $\hat{\gamma}_\lambda$ are the same as the ones in (4.2) and (4.3). The consistency $\hat{\tau} \xrightarrow{a.s.} \tau$ can be easily seen from the proof of Theorem 4.1. We thus do not re-
peat it. As a result, we approximate $\lambda^*_q$ by searching for the $\lambda$ that minimizes $\tilde{\tau}^2$. Notice that this problem has been studied for LASSO in [MMB15] and a generalization is straightforward for other bridge estimators. For simplicity, here we just use a grid search strategy that proceeds as follows:

- **Initialization:** An initial search region $[a, b]$, a window size $\Delta$ and a grid size $m$;
- **Searching:** A grid with size $m$ is built over $[a, b]$, upon which we search in descending order for $\lambda$ that minimizes $\tilde{\tau}^2$ with warm initialization.
  - If the minimal point $\hat{\lambda} \in (a, b)$, stop searching and return $\hat{\lambda}$.
  - If $\hat{\lambda} = a$ or $b$, update the search region with $[\frac{a}{10}, a]$ or $[b, b + \Delta]$ and do the next round of searching;
- **Stability:** If the optimal $\hat{\lambda}$ obtained from two consecutive search regions are smaller than a threshold $\epsilon_0$, we stop and return the previous optimal $\lambda$; if the number of non-zero locations of a LASSO estimator is larger than $n$ (which may happen numerically for very small tuning), we set its $\hat{\tau}^2$ to $\infty$.

For our experiments, we pick the initial $[a, b] = [0.1, \frac{1}{2}\|X^T y\|_\infty]$, $\Delta = \frac{1}{2}\|X^T y\|_\infty$ and $m = 15$.

5.3. **From large noise to small noise.** Two of our main results, i.e. Theorem 3.5 and Theorem 3.4 showed that in low SNR and high SNR situations, ridge and LASSO offer the best performances. These results are obtained for limiting cases $\sigma \to \infty$ and $\sigma \to 0$. In this section, we run a few simulations to clarify the scope of applicability of our analysis. Toward this goal we fix the distribution $p_G = \delta_8$ and run TVS for $q \in \{1, 1.2, 2, 4\}$ and debiased LASSO under four parameter settings:

1. $\delta = 0.8$, $\epsilon = 0.2$: The results are shown in Figure 5.1. Here we pick $\sigma \in \{1.5, 3, 5\}$. Note that as expected from our theoretical results, for small values of noise LASSO offers the best performance. As we increase the noise, eventually ridge outperforms LASSO. But under this setting, this outperformance happens at a high noise level at which all estimators perform similarly poorly (since the optimal choice of $\lambda$ is very large). Note that in this example $1 > \delta > M_1(\epsilon)$. Refer to Theorem 3.4 for the importance of this condition.

2. $\delta = 2$, $\epsilon = 0.4$: The results are shown in Figure 5.1. Here we pick $\sigma \in \{2, 4, 8\}$. Similar phenomena are observed here. However for all choices of $\sigma$, the ATPP-AFDP curves of different methods are quite close to each other.
Fig 5.1. Top row: under the setting $\delta = 0.8$, $\epsilon = 0.2$, $\sigma \in \{1.5, 3, 5\}$, the averaged ATPP-AFDP curve. Second row: averaged $AFDP_q - AFDP_2$ with 1 standard deviation of this differences. Third row: under the setting $\delta = 2$, $\epsilon = 0.4$, $\sigma \in \{2, 4, 8\}$, the averaged ATPP-AFDP curve. Fourth row: averaged $AFDP_q - AFDP_2$ with 1 standard deviation of this differences.
Which regularizer is optimal for variable selection?

Fig 5.2. Top row: under the setting $\delta = 0.6$, $\epsilon = 0.4$, $\sigma \in \{0.25, 0.75, 2\}$, the averaged ATPP-AFDP curve. Second row: averaged $\text{AFDP}_q - \text{AFDP}_2$ with 1 standard deviation of this differences. Third row: under the setting $\delta = 0.9$, $\epsilon = 0.4$, $\sigma \in \{1.2, 1.5, 1.9\}$, the averaged ATPP-AFDP curve. Bottom row: averaged $\text{AFDP}_q - \text{AFDP}_2$ with 1 standard deviation of this differences.
3. $\delta = 0.6$, $\epsilon = 0.4$: Figure 5.2 contains the results for this part. Here we have $\sigma \in \{0.25, 0.75, 2\}$. An interesting feature of this simulation is that $0.6 < M_1(0.4)$, which does not satisfy the conditions of Theorem 3.4. From another angle the setting is above the phase transition for all values of $q$. Refer to [WMZ16] for more information on the phase transition diagrams and their importance. It is interesting to see that in this case, ridge outperforms LASSO even for small values of the noise. The message is, when the problem setting is above phase transition, LASSO cannot achieve exact recovery for $\sigma = 0$ and will have non-zero AMSE. If this AMSE is larger than that of other estimators at $\sigma = 0$, following the continuity of AMSE in the noise level, LASSO may not be able to outperform other bridge estimators for small $\sigma$. Actually for our setting here, the optimal AMSE for $q = 1, 1.2, 2, 4$ at $\sigma = 0$ are $14.9, 12.2, 10.2, 11.6$, which is perfectly consistent with our theory.

4. $\delta = 0.9$, $\epsilon = 0.4$: Figure 5.2 contains the results for this part. Here we have $\sigma \in \{1.2, 1.5, 1.9\}$. This group of figures provides us with examples where ridge based two-stage variable selection outperforms LASSO based version, and at the same time reaches a quite satisfactory AFDP-ATPP trade-off. For instance, when $\sigma = 1.5$ and AFDP $\approx 0.2$, for ridge we have $\text{ATPP} \approx 0.85$ while that for LASSO is around 0.75. Note that here $M_1(\epsilon) < \delta < 1$.

5.4. Large $\delta$ regime. For this part we will validate the results in Theorem 3.6, which are obtained under limiting case $\delta \to \infty$. Here, we fix the distribution $p_G = \delta_1$ and consider the following settings for $q \in \{1, 1.5, 2, 4\}$ and debiased $\ell_1$:

1. $\epsilon = 0.1$, $\sigma = 0.4$: The results for this setting is shown in Figure 5.3. Here we vary $\delta \in \{2, 3, 4\}$. As can be seen from the figure, LASSO starts to outperforms the others even when $\delta = 2$. Actually the extra gain from LASSO is most obvious in this setting. As we gradually increase $\delta$, although LASSO still leads the performance, all methods are becoming better and the ATPP-AFDP curve starts to get closer.

2. $\epsilon = 0.3$, $\sigma = 0.4$: The results for this setting is shown in Figure 5.3. Again $\delta \in \{2, 3, 4\}$. Similar phenomena are observed. Compare to the previous setting, increase of $\epsilon$ provides stronger signal to the learning problem. As a result, all methods get better compared to the 1st setting.

3. $\epsilon = 0.4$, $\sigma = 0.22$: The results are shown in Figure 5.4. $\delta \in \{0.7, 0.8, 1.2\}$. It can be seen that, when $\delta$ is 0.7, 0.8, $\ell_2$ significantly outperforms other methods. As we increase $\delta$ to 1.2, $\ell_1$ starts to outperform the others. $\ell_1$
Fig 5.3. Top row: under the setting $\delta \in \{2, 3, 4\}$, $\epsilon = 0.1$ and $\sigma = 0.4$, the averaged ATPP-AFDP curve. Second row: averaged $\text{AFDP}_q - \text{AFDP}_1$ with 1 standard deviation of this differences. Third row: under the setting $\delta \in \{2, 3, 4\}$, $\epsilon = 0.3$ and $\sigma = 0.4$, the averaged ATPP-AFDP curve. Fourth row: averaged $\text{AFDP}_q - \text{AFDP}_1$ with 1 standard deviation of this differences.
penalty receives more power from increased sample size and behaves better when $\delta$ is mildly large, while $\ell_2$ degrades slower with respect to decreasing sample size.

6. Conclusion and future work. We studied two-stage variable selection (TVS) schemes for linear models under the high-dimensional asymptotic setting, where the number of observations $n$ grows at the same rate as the number of predictors $p$. Our TVS have a bridge estimator in their first stage and a simple threshold function in their second stage. For such schemes, we proved that for a fixed asymptotic true positive proportions (ATTP), in order to obtain the smallest AFDP one should pick an estimator that minimizes the asymptotic mean square error in the first stage of TVS. This connection between parameter estimation and variable selection further led us to a thorough investigation of the AMSE under small noise, large noise and large sample regimes. Our analyses revealed several interesting phenomena and provided new insights into variable selection. For instance, the variable selection of LASSO can be improved by debiasing and thresholding; a TVS with ridge in its first stage outperforms TVS with other bridge
estimators including LASSO for large values of noise.

Several important directions are left open for future research. For instance, it is important to check the correctness and robustness of our conclusions for dependent design matrices, or to study TVS with non-convex regularization like $\ell_q (q < 1)$ and SCAD [FL01] in the first stage.

REFERENCES

[BM11] Mohsen Bayati and Andrea Montanari. The dynamics of message passing on dense graphs, with applications to compressed sensing. *IEEE Transactions on Information Theory*, 57(2):764–785, 2011.

[BM12] Mohsen Bayati and Andrea Montanari. The lasso risk for gaussian matrices. *IEEE Transactions on Information Theory*, 58(4):1997–2017, 2012.

[BvdBSC13] Małgorzata Bogdan, Ewout van den Berg, Weijie Su, and Emmanuel J Candès. Supplementary materials for statistical estimation and testing via the sorted 1 norm. *Annals of Statistics*, 2013.

[BY93] ZD Bai and YQ Yin. Limit of the smallest eigenvalue of a large dimensional sample covariance matrix. *The annals of Probability*, pages 1275–1294, 1993.

[DMM09] David L Donoho, Arian Maleki, and Andrea Montanari. Message-passing algorithms for compressed sensing. *Proceedings of the National Academy of Sciences*, 106(45):18914–18919, 2009.

[DMM11] David L Donoho, Arian Maleki, and Andrea Montanari. The noise-sensitivity phase transition in compressed sensing. *IEEE Transactions on Information Theory*, 57(10):6920–6941, 2011.

[FF93] LLdiko E Frank and Jerome H Friedman. A statistical view of some chemometrics regression tools. *Technometrics*, 35(2):109–135, 1993.

[FL01] Jianqing Fan and Runze Li. Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American statistical Association*, 96(456):1348–1360, 2001.

[FL08] Jianqing Fan and Jinchu Lv. Sure independence screening for ultrahigh dimensional feature space. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 70(5):849–911, 2008.

[KF00] Keith Knight and Wenjiang Fu. Asymptotics for lasso-type estimators. *Annals of statistics*, pages 1356–1378, 2000.

[MAYB13] Arian Maleki, Laura Anitori, Zai Yang, and Richard G Baraniuk. Asymptotic analysis of complex lasso via complex approximate message passing (cAMP). *IEEE Transactions on Information Theory*, 59(7):4290–4308, 2013.

[MB06] Nicolai Meinshausen and Peter Bühlmann. High-dimensional graphs and variable selection with the lasso. *The annals of statistics*, pages 1436–1462, 2006.

[MB15] Ali Mousavi, Arian Maleki, and Richard G Baraniuk. Consistent parameter estimation for lasso and approximate message passing. *arXiv preprint arXiv:1511.01017*, 2015.

[MY09] Nicolai Meinshausen and Bin Yu. Lasso-type recovery of sparse representations for high-dimensional data. *The Annals of Statistics*, pages 246–270, 2009.
[RG13] Galen Reeves and Michael C Gastpar. Approximate sparsity pattern recovery: Information-theoretic lower bounds. *IEEE Transactions on Information Theory*, 59(6):3451–3465, 2013.

[SBC15] Weijie Su, Malgorzata Bogdan, and Emmanuel Candes. False discoveries occur early on the lasso path. *arXiv preprint arXiv:1511.01957*, 2015.

[Tib96] Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 267–288, 1996.

[Wai09] Martin J Wainwright. Sharp thresholds for high-dimensional and noisy sparsity recovery using-constrained quadratic programming (lasso). *IEEE transactions on information theory*, 55(5):2183–2202, 2009.

[WMZ16] Haolei Weng, Arian Maleki, and Le Zheng. Overcoming the limitations of phase transition by higher order analysis of regularization techniques. *arXiv preprint arXiv:1603.07377*, 2016.

[WR09] Larry Wasserman and Kathryn Roeder. High dimensional variable selection. *Annals of statistics*, 37(5A):2178, 2009.

[Z^+09] Tong Zhang et al. Some sharp performance bounds for least squares regression with l1 regularization. *The Annals of Statistics*, 37(5A):2109–2144, 2009.

[ZH08] Cun-Hui Zhang and Jian Huang. The sparsity and bias of the lasso selection in high-dimensional linear regression. *The Annals of Statistics*, pages 1567–1594, 2008.

[Zho09] Shuheng Zhou. Thresholding procedures for high dimensional variable selection and statistical estimation. In *Advances in Neural Information Processing Systems*, pages 2304–2312, 2009.

[Zou06] Hui Zou. The adaptive lasso and its oracle properties. *Journal of the American statistical association*, 101(476):1418–1429, 2006.

[ZY06] Peng Zhao and Bin Yu. On model selection consistency of lasso. *Journal of Machine Learning Research*, 7(Nov):2541–2563, 2006.
Supplementary material

APPENDIX A: ORGANIZATION

This supplement contains the proofs of all the main results. Below we mention the organization of this supplement to help the readers.

1. Appendix B includes some preliminaries that will be extensively used in the latter proofs.

2. Appendix C proves Lemmas 3.1 and 3.2. These two lemmas characterize the AFDP and ATPP of LASSO and the two-stage bridge estimators.

3. Appendix D contains the proof of Theorems 3.1, 3.2 and 3.3. These theorems compare the variable selection performance of LASSO and different two-stage estimators, and characterize the optimality.

4. Appendix E proves Theorems 3.5. This theorem analyzes the asymptotic mean square error of the bridge estimators in the large noise regime.

5. Appendix F proves Theorems 3.4 and 3.6. The two theorems analyze the asymptotic mean square error of bridge estimators in the small noise and the large sample size regimes, respectively.

6. Appendix G includes the proof of Theorems 4.1, 4.2, 4.3 and Lemma 4.1. Theorem 4.1 justifies the “unbiasedness” of our proposed debiasing estimators. Theorems 4.2 and 4.3 further characterize the variable selection performance of TVS that used unbiased estimates. Lemma 4.1 shows the superiority of the debiased two-stage methods over Sure Independent Screening.

APPENDIX B: PRELIMINARIES

Define

\begin{align}
R_q(\alpha, \tau) &\triangleq \mathbb{E}(\eta_q(B/\tau + Z; \alpha) - B/\tau)^2, \\
\alpha_q(\tau) &\triangleq \arg\min_{\alpha \geq 0} R_q(\alpha, \tau).
\end{align}

For the definition (B.2), if the minimizer is not unique, we choose the one with the smallest value. We also clarify the following notations:

(i) In the subsequent sections, we will use \( \partial_i f \) to denote the partial derivative of \( f(x, y, \ldots) \) to its \( i \)th argument. Also for the ease of organizing
the proof, we may use $\partial_y f$ to be the partial derivative of $f$ with respect to its argument $y$, which is equivalent to $\partial_2 f$.

(ii) We will use DCT as a short name for Dominated Convergence Theorem.

(iii) Recall we have $p_B = (1 - \epsilon)\delta_0 + \epsilon p_G$. By symmetry, it can be easily verified that $B$ and $G$ appearing in all the subsequent proofs can be equivalently replaced by $|B|$ and $|G|$. Therefore, without loss of generality, we assume that $B$ and $G$ are nonnegative random variables.

(iv) Let $\Phi$ and $\phi$ denote the cumulative distribution function and probability density function of a standard normal random variable respectively. Standard result on the expansion of Gaussian tails through integration by parts gives: for $k \in \mathbb{N}^+$, $s > 0$

\[
\Phi(s) = s^{-1/2} + \sum_{i=0}^{k-1} \frac{(-1)^i(2i-1)!!}{s^{2i+1}} + (-1)^k(2k-1)!! \int_s^\infty \frac{\phi(t)}{t^{2k}} dt,
\]

where $(2i-1)!! \triangleq 1 \times 3 \times 5 \times \ldots \times (2i-1)$. Recall the proximal operator function $\eta_q(u; \chi)$ defined in (2.3). Note that $\eta_q(u; \chi)$ does not have an explicit form except for $q = 0, 1, 2$. The next lemma shows a few properties of $\eta_q(u; \chi)$ that will be extensively used in the proofs.

**Lemma B.1.** For any $q \in (1, \infty)$, we have

(i) $\eta_q(u; \chi) = -\eta_q(-u; \chi)$.

(ii) $|u| = |\eta_q(u; \chi)| + \chi q |\eta_q(u; \chi)|^{q-1}$.

(iii) $\eta_q(\alpha u; \alpha^{2-q} \chi) = \alpha \eta_q(u; \chi)$, for $\alpha > 0$.

(iv) $\partial_1 \eta_q(u; \chi) = \frac{1}{1 + \chi q(1-\alpha) |\eta_q(u; \chi)|^{q-1}}$.

(v) $\partial_2 \eta_q(u; \chi) = \frac{-q |\eta_q(u; \chi)|^{q-1} \text{sgn}(u)}{1 + \chi q(1-\alpha) |\eta_q(u; \chi)|^{q-1}}$.

**Proof.** Please refer to Lemmas 7 and 10 in [WMZ16] for the proof of $q \in (1, 2)$. The proof for $q > 2$ is the same, hence we do not repeat it. \hfill \Box

**Lemma B.2.** Consider a nonnegative random variable $X$ with probability distribution $\mu$ and $\mathbb{P}(X > 0) = 1$. Let $\xi > \zeta > 0$ be the points such that $\mathbb{P}(X \leq \zeta) \leq \frac{1}{4}$ and $\mathbb{P}(\zeta < X \leq \xi) \geq \frac{1}{4}$. Let $a, b, c : \mathbb{R}_+ \to \mathbb{R}_+$ be three deterministic positive functions such that $a(s), c(s) \to \infty$ as $s \to \infty$. Then there exists a positive constant $s_0$ depending on $a, c, X$, such that when $s > s_0$,

\[
\int_0^a e^{b(s)x - \frac{x^2}{c(s)}} d\mu(x) \leq 3 \int_\zeta^\xi e^{b(s)x - \frac{x^2}{c(s)}} d\mu(x).
\]
Proof. For large enough $s$ such that $a(s) > \xi$, \[
\int_\zeta^{a(s)} e^{b(s)x - \frac{x^2}{a(s)}} d\mu(x) \geq \int_\zeta^{\xi} e^{b(s)x - \frac{x^2}{a(s)}} d\mu(x) \geq e^{b(s)\zeta - \frac{\xi^2}{a(s)}} \mathbb{P}(\zeta < X \leq \xi) \\
\geq e^{b(s)\zeta - \frac{\xi^2}{a(s)}} \mathbb{P}(X \leq \zeta) \geq e^{-\frac{\xi^2}{a(s)}} \int_0^\zeta e^{b(s)x - \frac{x^2}{a(s)}} d\mu(x).
\]

As a result we have the following inequality, \[
\int_0^{a(s)} e^{b(s)x - \frac{x^2}{a(s)}} d\mu(x) \leq (1 + e^{\frac{\xi^2}{a(s)}}) \int_\zeta^{a(s)} e^{b(s)x - \frac{x^2}{a(s)}} d\mu(x).
\]

For sufficiently large $s$ such that $e^{\frac{\xi^2}{a(s)}} < 2$, the conclusion follows.

APPENDIX C: PROOF OF LEMMAS 3.1 AND 3.2

The proof of Lemmas 3.1 and 3.2 can be found in Sections C.1 and C.2, respectively.

C.1. Proof of Lemma 3.1. The formulas in Lemma 3.1 have been derived in terms of convergence in probability in [BvdBSC13]. We show they hold almost surely by a straightforward adaption of the proof in [BvdBSC13].

Adopting the notations from [BvdBSC13]:
\[
\varphi_v(x, y) = \mathbb{I}(x \neq 0)\mathbb{I}(y = 0), \quad \varphi_v,h(x, y) = (1 - Q(x/h))Q(y/h)
\]
\[
Q(x) = \max(1 - |x|, 0), \quad V = \sum_{i=1}^p \varphi_v(\hat{\beta}_i(1, \lambda), \beta_i).
\]

We consider a positive sequence $h_k \to 0$ as $k \to 0$. Since $\varphi_v,h_k(x, y)$ is a bounded Lipschitz function, Theorem 2.1 implies almost surely
\[
\lim_{k \to \infty} \lim_{p \to \infty} \frac{1}{p} \sum_{i} \varphi_v,h_k(\hat{\beta}_i(1, \lambda), \beta_i) = \lim_{k \to \infty} \mathbb{E}\varphi_v,h_k(\eta_1(B + \tau Z; \alpha \tau), B) = \mathbb{E} \lim_{k \to \infty} \varphi_v,h_k(\eta_1(B + \tau Z; \alpha \tau), B) = \mathbb{P}(\eta_1(B + \tau Z; \alpha \tau) \neq 0, B = 0) = (1 - \epsilon)\mathbb{P}(|Z| > \tau).
\]

Moreover, according to Definition 2.1 (b), we have
\[
\lim_{k \to \infty} \lim_{p \to \infty} \frac{1}{p} \sum_{i} \mathbb{I}(0 < \beta_i < h_k) = \lim_{k \to \infty} \epsilon \mathbb{P}(0 < |G| < h_k) = 0, \ a.s.
\]
Note that
\[
\left| \sum_i \varphi_{v,h_k}(\hat{\beta}_i(1,\lambda),\beta_i) - V \right| \leq \sum_i \mathbb{I}(0 < \hat{\beta}_i(1,\lambda) < h_k) + \sum_i \mathbb{I}(0 < \beta_i < h_k).
\]

Hence, if we can show
\[
\lim_{k \to \infty} \limsup_{p \to \infty} \frac{1}{p} \sum_i \mathbb{I}(0 < \hat{\beta}_i(1,\lambda) < h_k) = 0, \quad \text{a.s.} \tag{C.1}
\]
we will be able to conclude \( V \overset{a.s.}{\to} (1 - \epsilon) \mathbb{P}(|G + \tau Z| > \alpha \tau) \). Then we can use similar arguments to obtain \( \frac{1}{p} \sum_i \mathbb{I}(\hat{\beta}_i(1,\lambda) \neq 0) \overset{a.s.}{\to} \mathbb{P}(|B + \tau Z| > \alpha \tau) \). As a result, we will get the almost sure formula of AFDP. ATPP can be derived in a similar way. To prove (C.1), it is sufficient to prove
\[
\limsup_{p \to \infty} \frac{1}{p} \sum_i \mathbb{I}(0 < \hat{\beta}_i(1,\lambda) < h_k) \leq -\frac{C}{\log h_k}, \quad \text{a.s.}, \tag{C.2}
\]
where \( C \) is a positive constant we specify later. Let \( \mathcal{A} \subset \{1, \ldots, p\} \) be the active set of \( \hat{\beta}(1,\lambda) \). Since almost surely the empirical distribution of \( \hat{\beta}(1,\lambda) \) converges weakly to the distribution of \( \eta_1(B + \tau Z; \alpha \tau) \) (from Theorem 2.1), we know from portmanteau lemma almost surely,
\[
\liminf_{p \to \infty} \frac{|\mathcal{A}|}{p} \geq \mathbb{P}(|B + \tau Z| \geq \alpha \tau) > \rho, \tag{C.3}
\]
with \( \rho > 0 \) being a positive constant. Also according to [BY93] we have
\[
\lim_{p \to \infty} \sigma_{\max}(X) = \frac{1}{\sqrt{\delta}} + 1, \quad \text{a.s.}, \tag{C.4}
\]
where \( \sigma_{\max}(X) \) is the largest singular value of \( X \). We now use some key results in [BvdBSC13] to show (C.2). From the proof of Theorem 1 in [BvdBSC13] we already know that
\[
\sum_p \mathbb{P}\left(\frac{1}{p} \sum_i \mathbb{I}(0 < \hat{\beta}_i(1,\lambda) < h_k) > \frac{-C}{\log h_k}, |\mathcal{A}| > \rho p, \sigma_{\max}(X) < \delta^{-1/2} + 1 + \epsilon\right) \leq \sum_p \left[2^{p+2}(\sqrt{2/\pi h(\delta')^{-1}})^{-\rho C/\log h_k}\right]^p < \infty.
\]
where \( \delta' = (\delta^{-1/2} + 1 + \epsilon')^{-1} \); and the last step is proved by choosing \( C = 2(1 + 3p^{-1}) \). Hence the Borel-Cantelli lemma together with (C.3) and (C.4) implies (C.2).
C.2. Proof of Lemma 3.2. Denote

\[ FP = \sum_{i=1}^{p} \mathbb{I}(\hat{\beta}_i(q, \lambda) \neq 0, \beta_i = 0), \quad TP = \sum_{i=1}^{p} \mathbb{I}(\hat{\beta}_i(q, \lambda) \neq 0, \beta_i \neq 0). \]

First note that according to Theorem 2.1, almost surely the empirical distribution of \((\hat{\beta}(q, \lambda), \beta)\) converges weakly to the distribution of \((\eta_q(B + \tau Z; \alpha \tau^{-2-q}), B)\). We now choose a sequence \(t_m \to 0\) as \(m \to 0\) such that \(G\) does not have any point mass on that sequence. Then by portmanteau lemma we have almost surely

\[ \lim_{p \to \infty} \frac{1}{p} \sum_{i=1}^{p} \mathbb{I}(\hat{\beta}_i(q, \lambda) \neq 0, |\beta_i| \leq t_m) = \lim_{p \to \infty} \frac{1}{p} \sum_{i=1}^{p} \mathbb{I}(|\hat{\beta}_i(q, \lambda)| > s, |\beta_i| \leq t_m) \]

\[ = \mathbb{P}(\eta_q(B + \tau Z; \alpha \tau^{-2-q}) > s, |B| \leq t_m) = (1 - \epsilon)\mathbb{P}(\eta_q(\tau Z; \alpha \tau^{-2-q}) > s) \]

\[ + \epsilon \mathbb{P}(\eta_q(G + \tau Z; \alpha \tau^{-2-q}) > s, |G| \leq t_m), \]

which leads to

\[ \lim_{m \to \infty} \lim_{p \to \infty} \frac{1}{p} \sum_{i=1}^{p} \mathbb{I}(\hat{\beta}_i(q, \lambda, s) \neq 0, |\beta_i| \leq t_m) = (1 - \epsilon)\mathbb{P}(|\eta_q(\tau Z; \alpha \tau^{-2-q})| > s). \]

Moreover, it is clear that

\[ \frac{1}{p} \sum_{i=1}^{p} \mathbb{I}(\hat{\beta}_i(q, \lambda, s) \neq 0, |\beta_i| \leq t_m) - FP \leq \frac{1}{p} \sum_{i=1}^{p} \mathbb{I}(|\hat{\beta}_i(q, \lambda)| > s) \cdot \mathbb{I}(0 < |\beta_i| \leq t_m) \]

\[ \leq \sqrt{\frac{1}{p} \sum_{i=1}^{p} \mathbb{I}(|\hat{\beta}_i(q, \lambda)| > s)} \cdot \sqrt{\frac{1}{p} \sum_{i=1}^{p} \mathbb{I}(0 < |\beta_i| \leq t_m)} \]

\[ \xrightarrow{a.s.} \mathbb{P}(\eta_q(B + \tau Z; \alpha \tau^{-2-q}) > s)^{1/2} \cdot \epsilon^{1/2} \mathbb{P}(0 < |G| \leq t_m)^{1/2} \text{ as } p \to \infty. \]

Hence we obtain almost surely

\[ \lim_{m \to \infty} \lim_{p \to \infty} \frac{1}{p} \sum_{i=1}^{p} \mathbb{I}(\hat{\beta}_i(q, \lambda, s) \neq 0, |\beta_i| \leq t_m) - FP \xrightarrow{a.s.} 0. \]

This combined with (C.5) implies that as \(p \to \infty\)

\[ \frac{FP}{p} \xrightarrow{a.s.} (1 - \epsilon)\mathbb{P}(|\eta_q(\tau Z; \alpha \tau^{-2-q})| > s). \]
We can now conclude that

\[
\text{AFDP}(q, \lambda, s) = \lim_{p \to \infty} \frac{FP}{p} \frac{\lim_{p \to \infty} \sum_{i=1}^{p} I(\hat{\beta}_i(q, \lambda) > s)}{p} = (1 - \epsilon) \mathbb{P}(\eta_q(\tau Z; \alpha \tau^2 - q) > s), \quad \text{a.s.}
\]

The formula of \(\text{AFDP}(q, \lambda, s)\) in Lemma 3.2 can then be obtained by Lemma B.1 part (iii). Regarding \(\text{ATPP}(q, \lambda, s)\) we have

\[
\text{ATPP}(q, \lambda, s) = \frac{\lim_{p \to \infty} \sum_{i=1}^{p} I(\hat{\beta}_i(q, \lambda) > s)}{\lim_{p \to \infty} \sum_{i=1}^{p} I(\beta_i \neq 0)} - \lim_{p \to \infty} \frac{FP}{p} \frac{\lim_{p \to \infty} \sum_{i=1}^{p} I(\beta_i \neq 0)}{p} = \mathbb{P}(|\eta_q(G + \tau Z; \alpha \tau^2 - q)| > s), \quad \text{a.s.}
\]

**APPENDIX D: PROOF OF THEOREMS 3.1, 3.2 AND 3.3**

We present the proof of Theorems 3.1, 3.2 and 3.3 in Sections D.1, D.2 and D.3, respectively.

**D.1. Proof of Theorem 3.1.**

**Proof.** According to Lemma 3.2, we know

\[
\text{ATPP}(q, \lambda, s) = \mathbb{P}(|\eta_q(G + \tau Z; \alpha \tau^2 - q)| > s)
\]

where \((\alpha, \tau)\) is the unique solution to Equations (2.1) and (2.2). From Lemma B.1 part (iv), the proximal function \(\eta_q(u; \chi) = 0\) if and only if \(u = 0\) for \(q > 1\). Since \(G + \tau Z \neq 0\) a.s., we have \(\text{ATPP}(q, \lambda, 0) = 1\). Moreover, it is clear that \(\text{ATPP}(q, \lambda, +\infty) = 0\), and \(\text{ATPP}(q, \lambda, s)\) is a continuous and strictly decreasing function of \(s\) over \([0, \infty]\). Hence there exists a unique \(s\) for which \(\text{ATPP}(q, \lambda, s) = \zeta \in [0, 1]\).

Now consider all possible pairs \((\lambda, s)\) such that \(\text{ATPP}(q, \lambda, s) = \zeta\). Let \((\alpha_{s}, \tau_{s}, s_{s})\) be the triplet corresponding to the optimal tuning \(\lambda^*_q\) (it minimizes \(\text{AMSE}(q, \lambda)\)), and \((\alpha, \tau, s)\) be the one that corresponds to any other \(\lambda\). According to Theorem 2.1, we know

\[
\text{AMSE}(q, \lambda) = \delta(\tau^2 - \sigma^2).
\]

So \(\tau_{s} < \tau\). By the strict monotonicity and symmetry of \(\eta_q\) with respect to its first argument (see Lemma B.1 parts (i)(iv)), \(\text{ATPP}(q, \lambda^*_q, s_{s}) = \text{ATPP}(q, \lambda, s)\) implies that

\[
\mathbb{P}(|G/\tau_{s} + Z| > \eta_q^{-1}(s_{s}/\tau_{s}; \alpha_{s})) = \mathbb{P}(|G/\tau + Z| > \eta_q^{-1}(s/\tau; \alpha)), \quad \text{(D.1)}
\]
where $\eta_q^{-1}$ is the inverse function of $\eta_q$. Now we claim $\text{AFDP}(q, \lambda^*_s, s_*) < \text{AFDP}(q, \lambda, s)$. Otherwise, from the formula of AFDP in (3.2), we will have

$$\mathbb{P}(\eta_q(|Z|; \alpha_s) > s_*/\tau_s) \geq \mathbb{P}(\eta_q(|Z|; \alpha) > s/\tau),$$

which is equivalent to

$$\mathbb{P}(|Z| > \eta_q^{-1}(s_*/\tau_s; \alpha_s)) \geq \mathbb{P}(|Z| > \eta_q^{-1}(s/\tau; \alpha)).$$

This implies $\eta_q^{-1}(s_*/\tau_s; \alpha_s) \leq \eta_q^{-1}(s/\tau; \alpha)$. However, combining this result with $\tau_s < \tau$ and the fact that $\mathbb{P}(\mu + Z > t)$ is an strictly increasing function of $\mu$ over $[0, \infty)$, we must have

$$\mathbb{P}(|G/\tau_s + Z| > \eta_q^{-1}(s_*/\tau_s; \alpha_s)) \geq \mathbb{P}(|G/\tau_s + Z| > \eta_q^{-1}(s/\tau; \alpha)) > \mathbb{P}(|G/\tau + Z| > \eta_q^{-1}(s/\tau; \alpha)).$$

This is in contradiction with (D.1). The conclusion follows.

D.2. Proof of Theorem 3.2. According to Lemma 3.1,

$$\text{ATPP}(1, \lambda) = \mathbb{P}(|G + \tau Z| > \alpha \tau) = \mathbb{E}[\Phi(G/\tau - \alpha) + \Phi(-G/\tau - \alpha)].$$

It has been shown in [BM12] that, $\alpha$ is an increasing and continuous function of $\lambda$, and $\alpha \to \infty$ as $\lambda \to \infty$. Hence, $\text{ATPP}(1, \lambda)$ is a continuous function of $\lambda$ and $\lim_{\lambda \to \infty} \text{ATPP}(1, \lambda) = \lim_{\alpha \to \infty} \mathbb{P}(|G + \tau Z| > \alpha \tau) = 0$. Now let $(\alpha_*, \tau_*)$ be the solution to (2.1) and (2.2) when $\lambda = \lambda^*_1$. As we decrease $\lambda$ from $\infty$ to $\lambda^*_1$, $\text{ATPP}(1, \lambda)$ continuously changes from 0 to $\text{ATPP}(1, \lambda^*_1)$. Therefore, for any $\text{ATPP}$ level $\zeta \in [0, \text{ATPP}(1, \lambda^*_1)]$, there always exists at least a value of $\lambda \in [\lambda^*_1, \infty]$ such that $\text{ATPP}(1, \lambda) = \zeta$. Regarding the thresholded LASSO $\tilde{\beta}(1, \lambda^*_1, s)$, Lemma 3.2 shows that

$$\text{ATPP}(1, \lambda^*_1, s) = \mathbb{P}(|\eta_1(G + \tau_s Z; \alpha_s \tau_s)| > s).$$

Note that when $s = 0$ the thresholded LASSO is equal to LASSO and thus $\text{ATPP}(1, \lambda^*_1, 0) = \text{ATPP}(1, \lambda^*_1)$. It is also clear that $\text{ATPP}(1, \lambda^*_1, s)$ is a continuous and strictly decreasing function of $s$ on $[0, \infty]$. As a result, a unique threshold $s_\zeta$ exists s.t. $\text{ATPP}(1, \lambda^*_1, s_\zeta)$ reaches a given level $\zeta \in [0, \text{ATPP}(1, \lambda^*_1)]$.

We now compare the AFDP of different estimators that have the same ATPP. Suppose $\tilde{\beta}(1, \lambda)$ and $\tilde{\beta}(1, \lambda^*_1, s)$ reach the same level of ATPP. We have

$$\mathbb{P}(|\eta_1(G + \tau Z; \alpha \tau)| > 0) = \mathbb{P}(|\eta_1(G + \tau_s Z; \alpha_s \tau_s)| > s),$$
which is equivalent to

\[ P(|G/\tau + Z| > \alpha) = P(|G/\tau_s + Z| > \alpha_s + s/\tau_s). \]

Similar to the argument in the proof of Theorem 3.1, we have \( \alpha < \alpha_s + s/\tau_s \), since otherwise the left hand side in (D.2) will be smaller than the right hand side. Hence, we obtain

\[ P(|Z| > \alpha) > P(|Z| > \alpha_s + s/\tau_s) = P(|\eta_1(Z; \alpha_s)| > s/\tau_s). \]

This implies \( \text{AFDP}(1, \lambda) > \text{AFDP}(1, \lambda^*_1, s) \) based on Lemmas 3.1 and 3.2. By the same argument, we can show that \( \tilde{\beta}(1, \lambda^*_1, s) \) also has smaller AFDP than \( \tilde{\beta}(1, \lambda, s) \) if \( \lambda \neq \lambda^*_1 \).

D.3. Proof of Theorem 3.3. This theorem compares the two-stage estimators \( \tilde{\beta}(q, \lambda^*_i, s) \) for \( q \in [1, \infty) \). Consider \( q_1, q_2 \geq 1 \), and \( \text{AMSE}(q_1, \lambda^*_1) < \text{AMSE}(q_2, \lambda^*_2) \). Let \( (\alpha_{q_i}, \tau_{q_i}) \) be the solution to (2.1) and (2.2) when \( \lambda = \lambda^*_i \), for \( i = 1, 2 \). Then, according to Theorem 2.1, \( \tau_{q_1} < \tau_{q_2} \). Suppose \( \text{ATPP}(q_1, \lambda_{q_1}, s_1) = \text{ATPP}(q_2, \lambda_{q_2}, s_2) \), i.e.,

\[ P(\eta_{q_1}(G + \tau_{q_1}Z; \alpha_{q_1}^\tau_{q_1} - 2q_1) > s_1) = P(\eta_{q_2}(G + \tau_{q_2}Z; \alpha_{q_2}^\tau_{q_2} - 2q_2) > s_2). \]

When the ATPP level is 0 or 1, we see \( s_1 \) and \( s_2 \) are either both \( \infty \) or 0. The corresponding AFDP will be the same. We now consider the level of ATPP belong to \((0, 1)\). Using arguments similar to the ones presented in the proof of Theorem 3.1, we can conclude \( \eta_{q_1}^{-1}(s_1/\tau_{q_1}^*; \alpha_{q_1}^*) > \eta_{q_2}^{-1}(s_2/\tau_{q_2}^*; \alpha_{q_2}^*) \).

This gives us

\[ P(\eta_{q_1}(Z; \alpha_{q_1}^*) > s_1/\tau_{q_1}^*) = P(Z > \eta_{q_1}^{-1}(s_1/\tau_{q_1}^*; \alpha_{q_1}^*)) \]
\[ < P(Z > \eta_{q_2}^{-1}(s_2/\tau_{q_2}^*; \alpha_{q_2}^*)) = P(\eta_{q_2}(Z; \alpha_{q_2}^*) > s_2/\tau_{q_2}^*), \]

implying \( \text{AFDP}(q_1, \lambda_{q_1}, s_1) < \text{AFDP}(q_2, \lambda_{q_2}, s_2) \).

APPENDIX E: PROOF OF THEOREM 3.5

E.1. Roadmap. Since the proof of this theorem is long, we lay out the roadmap of the proof here to help readers navigate through the details. According to Theorem 2.1, in order to evaluate \( \text{AMSE}(q, \lambda^*_i) \), the crucial step is to characterize the solution \((\alpha_s, \tau_s)\) to Equations (2.1) and (2.2) with \( \lambda = \lambda^*_i \). From Corollary 1 in [WMZ16], we know that \( \tau_s \) can be obtained by solving the following single equation:

\[ \tau_s^2 = \sigma^2 + \frac{1}{\delta} \min_{\alpha \geq 0} \mathbb{E}(\eta_q(B + \tau_sZ; \alpha \tau_s^{2-q}) - B)^2. \]

\[ \text{Note that } \eta_1^{-1}(u; \chi) \text{ is not well defined for } u = 0 \text{ and we define it as } \eta_1^{-1}(0; \chi) = \chi. \]
Accordingly AMSE\((q, \lambda^*_q)\) can be computed as 
\[(E.2) \quad \text{AMSE}(q, \lambda^*_q) = \delta (\tau^2_* - \sigma^2). \]

It is clear from \((E.1)\) that \(\tau_* \to \infty\) as \(\sigma \to \infty\). However, to derive the second order expansion of \(\text{AMSE}(q, \lambda^*_q)\) as \(\sigma \to \infty\), we need to obtain the convergence rate of \(\tau_*\). We will achieve this goal by first characterizing the convergence rate of the term \(\min_{\alpha \geq 0} \mathbb{E}(\eta_q(B+\tau_* Z; \alpha \tau^2_* - q) - B)^2\) as \(\tau_* \to \infty\). We then use that result to derive the convergence rate of \(\tau_*\) based on \((E.1)\) and finally calculate \(\text{AMSE}(q, \lambda^*_q)\) through \((E.2)\). Since the proof techniques look different for \(q = 1, 1 < q \leq 2, q > 2\), we prove the theorem for these three cases in Sections E.2, E.3 and E.4 respectively.

**E.2. Proof of Theorem 3.5 for \(q = 1\).** According to Definitions (B.1), (B.2) and Lemma B.1 part (iii), it is clear that Equation \((E.1)\) can be rewritten:
\[(E.3) \quad \tau^2_* = \sigma^2 + \frac{1}{\delta \tau^2_*} R_q(\alpha_q(\tau_*), \tau_*). \]

As explained in the roadmap of the proof, the key step is to characterize the convergence rate of \(\alpha_q(\tau)\) as \(\tau \to \infty\) for \(q = 1\). We first prove \(\alpha_q(\tau) \to \infty\) as \(\tau \to \infty\) in the next lemma.

**Lemma E.1.** Recall the definition of \(\alpha_q(\tau)\) in (B.2). Assume \(\mathbb{E}|G|^2 < \infty\). Then, \(\alpha_q(\tau) \to \infty\) as \(\tau \to \infty\).

**Proof.** Suppose this is not true, then there exists a sequence \(\{\tau_n\}\) such that \(\alpha_q(\tau_n) \to \alpha_0 < \infty\) and \(\tau_n \to \infty\), as \(n \to \infty\). Notice that 
\[|\eta_q(B/\tau_n + Z; \alpha_q(\tau_n))| \leq |B|/\tau_n + Z \leq |B| + Z, \]
for sufficiently large \(n\). We can apply DCT to obtain 
\[\lim_{n \to \infty} R_q(\alpha_q(\tau_n), \tau_n) = \mathbb{E} \eta_q^2(Z; \alpha_0) > 0. \]

On the other hand, since \(\alpha = \alpha_q(\tau_n)\) minimizes \(R_q(\alpha, \tau_n)\)
\[\lim_{n \to \infty} R_q(\alpha_q(\tau_n), \tau_n) \leq \lim_{n \to \infty} \lim_{\alpha \to \infty} R_q(\alpha, \tau_n) = 0. \]

A contradiction arises. \(\square\)
Based on Lemma E.1, we can further derive the convergence rate of $\alpha_q(\tau)$.

**Lemma E.2.** If $G$ has a sub-Gaussian tail, then

$$\lim_{\tau \to \infty} \frac{\alpha_q(\tau)}{\tau} = C_0,$$

where $C = C_0$ is the unique solution of the following equation:

$$E\left(e^{CG}(CG - 1) + e^{-CG}(-CG - 1)\right) = \frac{2(1 - \epsilon)}{\epsilon}.$$  

**Proof.** Since $\alpha = \alpha_q(\tau)$ minimizes $R_q(\alpha, \tau)$, we know

$$\partial_1 R_q(\alpha_q(\tau), \tau) = 0.$$  

(E.4)

To simplify the notation, we will simply write $\alpha$ for $\alpha_q(\tau)$ in the rest of this proof. Rearranging the terms in (E.4) gives us

$$\frac{2(1 - \epsilon)}{\epsilon} = E_{\alpha, \tau} \left[ \frac{a^2}{\phi'(\alpha)} \left( \alpha \Phi\left( \frac{|G|}{\tau} - \alpha \right) \right. \right.$$

$$\left. + \alpha \Phi\left( \frac{|G|}{\tau} - \alpha \right) - \phi\left( \frac{|G|}{\tau} - \alpha \right) - \phi\left( \frac{|G|}{\tau} + \alpha \right) \right].$$

Fixing $t \in (0, 1)$, we reformulate the above equation in the following way:

$$\frac{2(1 - \epsilon)}{\epsilon} = E[T(G, \alpha, \tau)I(|G| \leq t\tau \alpha)] + E[T(G, \alpha, \tau)I(|G| > t\tau \alpha)].$$

(E.5)

We now analyze the two terms on the right hand side of the above equation. Since $G$ has a sub-Gaussian tail, there exists a constant $\gamma > 0$ such that $P(|G| > x) \leq e^{-\gamma x^2}$ for $x$ large. We can then have the following bound,

$$|E[T(G, \alpha, \tau)I(|G| > t\tau \alpha)]| \leq \frac{a^2}{\phi'(\alpha)} (2\alpha + \sqrt{2/\pi}) P(|G| > t\tau \alpha)$$

$$\leq \alpha^2 (2\sqrt{2/\pi} + 2) e^{-\gamma t^2 \tau^2 - \frac{1}{2}a^2} \to 0, \quad \text{as} \quad \tau \to \infty,$$

where we have used the fact that $\alpha \to \infty$ as $\tau \to \infty$ from Lemma E.1. This result combined with (E.5) implies that as $\tau \to \infty$

$$\frac{2(1 - \epsilon)}{\epsilon} = E[T(G, \alpha, \tau)I(|G| \leq t\tau \alpha)].$$

(E.6)
Moreover, using the tail approximation of normal distribution in (B.3) with \( k = 3 \), we have for sufficiently large \( \tau \),

\[
\mathbb{E}[T(G, \alpha, \tau) \mathbb{I}(|G| \leq t\tau \alpha)] \\
\leq \mathbb{E}\left[ \frac{\alpha}{\alpha - |G|/\tau} e^{-\frac{\alpha|G|}{2\tau^2}} \left( \frac{\alpha|G|}{\tau} - \frac{\alpha^2}{(\alpha - |G|/\tau)^2} + \frac{3\alpha^2}{(\alpha - |G|/\tau)^4} \right) + \frac{\alpha}{\alpha + |G|/\tau} e^{-\frac{\alpha|G|}{2\tau^2}} \left( -\frac{\alpha|G|}{\tau} - \frac{\alpha^2}{(\alpha + |G|/\tau)^2} + \frac{3\alpha^2}{(\alpha + |G|/\tau)^4} \right) \right] \cdot \mathbb{I}(|G| \leq t\tau \alpha).
\]

Similarly applying (B.3) with \( k = 2 \) gives us for large \( \tau \)

\[
\mathbb{E}[T(G, \alpha, \tau) \mathbb{I}(|G| \leq t\tau \alpha)] \geq \mathbb{E}\left[ \frac{\alpha}{\alpha - |G|/\tau} e^{-\frac{\alpha|G|}{2\tau^2}} \left( \frac{\alpha|G|}{\tau} - \frac{\alpha^2}{(\alpha - |G|/\tau)^2} \right) + \frac{\alpha}{\alpha + |G|/\tau} e^{-\frac{\alpha|G|}{2\tau^2}} \left( -\frac{\alpha|G|}{\tau} - \frac{\alpha^2}{(\alpha + |G|/\tau)^2} \right) \right] \cdot \mathbb{I}(|G| \leq t\tau \alpha).
\]

We can conclude based on the two bounds that \( \lim_{\tau \to \infty} \frac{\alpha}{\tau} = C_1 \) with \( 0 < C_1 < \infty \). Otherwise

- If \( C_1 = \infty \), there exists a sequence \( \alpha_n/\tau_n \to \infty \) and \( \tau_n \to \infty \), as \( n \to \infty \). Since \( |L_2(G, \alpha_n, \tau_n)| \leq e^{-\frac{\alpha_n|G|}{\tau_n}} \left( \frac{\alpha_n|G|}{\tau_n} + 1 \right) \leq 2 \), we can apply DCT to obtain

\[
\lim_{n \to \infty} \mathbb{E}(L_2(G, \alpha_n, \tau_n) \mathbb{I}(|G| \leq t\tau_n \alpha_n)) = 0.
\]

Furthermore, we choose a positive constant \( \zeta > 0 \) satisfying the condition in Lemma B.2 for the nonnegative random variable \( |G| \). Then

\[
\mathbb{E}(L_1(G, \alpha_n, \tau_n) \mathbb{I}(|G| \leq t\tau_n \alpha_n)) \\
\geq \mathbb{E}\left[ e^{-\frac{\alpha_n|G|}{\tau_n}} \left( \frac{\alpha_n|G|}{\tau_n} - \frac{1}{(1 - t)^3} \right) \mathbb{I}(|G| \leq t\tau_n \alpha_n) \right] \\
\geq \int_{\zeta g \leq t\tau_n \alpha_n} e^{-\frac{\alpha_n g}{\tau_n}} \left( \frac{\alpha_n g}{\tau_n} - \frac{1}{(1 - t)^3} \right) dF(g) - \int_{g \leq t\tau_n \alpha_n} \frac{1}{(1 - t)^3} e^{-\frac{\alpha_n g}{\tau_n}} \frac{g^2}{2\tau_n^2} dF(g) \\
\geq (a) \left( \frac{\zeta \alpha_n}{\tau_n} - \frac{2}{(1 - t)^3} \right) \int_{\zeta g \leq t\tau_n \alpha_n} e^{-\frac{\alpha_n g}{\tau_n}} \frac{g^2}{2\tau_n^2} dF(g) \\
\geq \left( \frac{\zeta \alpha_n}{\tau_n} - \frac{2}{(1 - t)^3} \right) e^{-\frac{\alpha_n \zeta}{\tau_n}} \int_{\zeta g \leq t\tau_n \alpha_n} e^{-\frac{g^2}{2\tau_n^2}} dF(g) \to \infty,
\]

WHICH REGULARIZER IS OPTIMAL FOR VARIABLE SELECTION? 41
where we have used Lemma B.2 in (a). This forms a contradiction.

• If \( C_1 = 0 \), for large enough \( \tau \) we have \( \frac{\alpha}{\tau} < 1 \) and then on \( |G| \leq t\tau\alpha \),

\[
|U_1(G, \alpha, \tau) + U_2(G, \alpha, \tau)| \leq \frac{2}{1-t}e^{G\left[G + \frac{1}{(1-t)^2} + \frac{3}{\alpha^2(1-t)^4}\right]},
\]

which is integrable since \( G \) has sub-Gaussian tail. Hence we apply DCT to obtain as \( \tau \to \infty \)

\[
\mathbb{E} [(U_1(G, \alpha, \tau) + U_2(G, \alpha, \tau)) \mathbb{I}(|G| \leq t\tau\alpha)] \to -2
\]

This forms another contradiction.

Similar to the above arguments, we can conclude that \( \lim_{\tau \to \infty} \frac{\alpha}{\tau} = C_2 \in (0, \infty) \). Now that \( \frac{\alpha}{\tau} = O(1) \), we can use DCT to obtain

\[
\lim_{\tau \to \infty} \mathbb{E} \left[ \frac{\alpha}{\alpha \pm |G|/\tau} e^{\frac{\alpha|G|}{\tau}} \frac{3\alpha^2}{(\alpha \pm |G|/\tau)^2} \mathbb{I}(|G| \leq t\tau\alpha) \right] = 0.
\]

This result combined together with (E.6) and the upper and lower bounds on \( \mathbb{E}[T(G, \alpha, \tau) \mathbb{I}(|G| \leq t\tau\alpha)] \) enables us to show

\[
\lim_{\tau \to \infty} \mathbb{E} [(L_1(G, \alpha, \tau) + L_2(G, \alpha, \tau)) \mathbb{I}(|G| \leq t\tau\alpha)] = \frac{2(1 - \epsilon)}{\epsilon}.
\]

Now consider a convergent sequence \( \frac{\alpha_n}{\tau_n} \to C_1 \in (0, \infty) \) and \( \tau_n \to \infty \) as \( n \to \infty \). On \( |G| \leq t\tau_n\alpha_n \) we can bound for large \( n \)

\[
|L_1(G, \alpha_n, \tau_n) + L_2(G, \alpha_n, \tau_n)| \leq \frac{2}{1-t}e^{2C_1G\left(2C_1G + \frac{1}{(1-t)^2}\right)},
\]

which is again integrable. Thus DCT gives us

\[
\frac{2(1 - \epsilon)}{\epsilon} = \lim_{n \to \infty} \mathbb{E} [(L_1(G, \alpha_n, \tau_n) + L_2(G, \alpha_n, \tau_n)) \mathbb{I}(|G| \leq t\tau_n\alpha_n)]
\]

\[
= \mathbb{E} [e^{C_1|G|}(C_1|G| - 1) + e^{-C_1|G|}(-C_1|G| - 1)].
\]

For \( C_2 \) the same equation holds. By calculating the derivative we can easily verify \( h(c) = e^{c|G|}(c|G| - 1) + e^{-c|G|}(-c|G| - 1) \), as a function of \( c \) over \( (0, \infty) \), is strictly increasing. This determines \( C_1 = C_2 \). Above all we have shown

\[
\frac{\alpha_q(\tau)}{\tau} \to C_0, \quad \text{as} \ \tau \to \infty,
\]

where \( \mathbb{E} [e^{C_0G}(C_0G - 1) + e^{-C_0G}(-C_0G - 1)] = \frac{2(1-t)}{\epsilon} \).
WHICH REGULARIZER IS OPTIMAL FOR VARIABLE SELECTION?

43

E.2.2. Bounding the convergence rate of \( R_q(\alpha_q(\tau), \tau) \) as \( \tau \to \infty \) for \( q = 1 \).

We state the main result in the next lemma.

**Lemma E.3.** If \( G \) has sub-Gaussian tail, then as \( \tau \to \infty \)

\[
R_q(\alpha_q(\tau), \tau) = \frac{eE[G^2]}{\tau^2} + o\left( \frac{\phi(\alpha_q(\tau))}{\alpha_q^3(\tau)} \right).
\]

**Proof.** For notational simplicity, we will use \( \alpha \) to denote \( \alpha_q(\tau) \) in the rest of the proof. Since \( \eta_1(u; \chi) = \text{sgn}(u)(|u| - \chi)_+ \), we can write \( R_q(\alpha, \tau) \) in the following form:

\[
R_q(\alpha, \tau) = 2(1 - \epsilon)[(1 + \alpha^2)\Phi(-\alpha) - \alpha\phi(\alpha)]
\]

\[
+ \epsilon E\left[ (1 + \alpha^2 - G^2/\tau^2)\left( \Phi(G/\tau - \alpha) + \Phi(-G/\tau - \alpha) \right) S_1(G, \alpha, \tau) \right]
\]

\[
-(G/\tau + \alpha)\phi(\alpha - G/\tau) + (G/\tau - \alpha)\phi(\alpha + G/\tau) + G^2/\tau^2 S_2(G, \alpha, \tau) \right].
\]

Hence, we have

\[
\lim_{\tau \to \infty} \frac{\alpha^3}{\phi(\alpha)} \left( R_q(\alpha, \tau) - \frac{eE[G^2]}{\tau^2} \right) = 2(1 - \epsilon) \lim_{\tau \to \infty} \frac{\alpha^3}{\phi(\alpha)} [(1 + \alpha^2)\Phi(-\alpha) - \alpha\phi(\alpha)]
\]

\[
+ \epsilon \lim_{\tau \to \infty} \frac{\alpha^3}{\phi(\alpha)} E[S_1(G, \alpha, \tau) + S_2(G, \alpha, \tau)]
\]

\[
\text{(E.7)} \quad \equiv 4(1 - \epsilon) + \epsilon \lim_{\tau \to \infty} \frac{\alpha^3}{\phi(\alpha)} E[S_1(G, \alpha, \tau) + S_2(G, \alpha, \tau)].
\]

We have used the tail expansion (B.3) with \( k = 3,4 \) to obtain (a). Note that since \( |x\phi(x)| \leq \frac{e^{-x^2/2\pi}}{\sqrt{2\pi}} \), we have

\[
|S_1(G, \alpha, \tau) + S_2(G, \alpha, \tau)| \leq \frac{2e^{-1/2}}{\sqrt{2\pi}} + \frac{4\alpha}{\sqrt{2\pi}} + 2\left( 1 + \alpha^2 + \frac{G^2}{\tau^2} \right).
\]

Moreover, it is not hard to use the sub-Gaussian condition \( \mathbb{P}(|G| > x) \leq e^{-\gamma x^2} \) to obtain

\[
\mathbb{E}(G^2 I(|G| > t\tau\alpha)) = \int_0^{t\tau\alpha} 2x\mathbb{P}(G > t\tau\alpha) dx + \int_{t\tau\alpha}^{\infty} 2x\mathbb{P}(G > x) dx
\]

\[
\leq (t\tau\alpha)^2 e^{-\gamma t^2\tau^2\alpha^2} + \frac{1}{\gamma} e^{-\gamma t^2\tau^2\alpha^2},
\]
where \( t \in (0, 1) \) is a constant. Combining the last two bounds we can derive

\[
\frac{\alpha^3}{\phi(\alpha)} E[(S_1(G, \alpha, \tau) + S_2(G, \alpha, \tau)) \mathbb{I}(\{|G| > t\tau\alpha\})] \\
\leq \alpha^3 \left(2e^{-1/2} + 4\alpha + 2\sqrt{2\pi(1 + \alpha^2)}e^{-(\gamma t^2\tau^2 - \frac{1}{2})\alpha^2} + \frac{2\sqrt{2\pi\alpha^3}}{\tau^2}(t^2\tau^2\alpha^2 + 1/\gamma)e^{-(\gamma t^2\tau^2 - \frac{1}{2})\alpha^2} \right) \to 0, \quad \text{as } \tau \to \infty.
\]

On the other hand, we can build an upper bound and lower bound for \( |S_1(G, \alpha, \tau) + S_2(G, \alpha, \tau)| \) on \( \{|G| \leq t\tau\alpha\} \) with the tail expansion (B.3) as we did in the proof of Lemma E.2, For both bounds we can argue they converge to the same limit as \( \tau \to \infty \) by using DCT and Lemma E.2. Here we give the details of using DCT for the upper bound. Using (B.3) with \( k = 3 \) we can obtain the upper bound,

\[
\frac{\alpha^3}{\phi(\alpha)} (S_1(G, \alpha, \tau) + S_2(G, \alpha, \tau)) \\
\leq \frac{\alpha^3}{\phi(\alpha)} \left( \frac{2G^2/\tau^2 - 2\alpha G/\tau - 1}{(\alpha - G/\tau)^3} + \frac{3(1 + \alpha^2 - G^2/\tau^2)}{(\alpha - G/\tau)^5} \right) + \frac{\alpha^3}{\phi(\alpha)} \left( \frac{2G^2/\tau^2 + 2\alpha G/\tau - 1}{(\alpha + G/\tau)^3} + \frac{3(1 + \alpha^2 - G^2/\tau^2)}{(\alpha + G/\tau)^5} \right).
\]

It is straightforward to see that on \( \{|G| \leq t\tau\alpha\} \) for sufficiently large \( \alpha \), there exist three positive constants \( C_1, C_2, C_3 \) such that the upper bound can be further bounded by \( \left[ \frac{C_1|G|+1}{(1+1)^2} + \frac{C_2|G|+1}{(1+1)^2} + \frac{C_3|G|}{(1+1)^2} \right] e^{C_3|G|} \), which is integrable by the condition that \( G \) has sub-Gaussian tail. Hence we can apply DCT to derive the limit of the upper bound. Similar arguments enable us to calculate the limit of the lower bound. By calculating the limits of the upper and lower bounds we can obtain the following result:

\[
\frac{\alpha^3}{\phi(\alpha)} E[(S_1(G, \alpha, \tau) + S_2(G, \alpha, \tau)) \mathbb{I}(\{|G| \leq t\tau\alpha\})] \\
\to -2E \left( e^{C_0G}(C_0G - 1) + e^{-C_0G}(-C_0G - 1) \right) = -\frac{4(1 - e)}{e}.
\]

This completes the proof. \( \Box \)

E.2.3. Deriving the expansion of AMSE\((q, \lambda_q^*)\) for \( q = 1 \). We are now in the position to derive the result (3.8) in Theorem 3.5. As we explained in the roadmap, we know

\[
\text{AMSE}(q, \lambda_q^*) = \tau^2_q R_q(\alpha_q(\tau_q), \tau_q) = \delta(\tau^2_q - \sigma^2).
\]
First note that \( \tau_\star \rightarrow \infty \) as \( \sigma \rightarrow \infty \) since \( \tau_\star \geq \sigma \). Then according to Lemma E.3 and (E.8), we have

\[
\lim_{\sigma \rightarrow \infty} \frac{\sigma^2}{\tau_\star^2} = \lim_{\tau_\star \rightarrow \infty} \frac{\sigma^2}{\tau_\star^2} = \lim_{\tau_\star \rightarrow \infty} \left( 1 - \frac{R_q(\alpha_q(\tau_\star), \tau_\star)}{\delta} \right) = 1.
\]

Furthermore, Lemma E.2 shows that

\[
\lim_{\sigma \rightarrow \infty} \alpha_q(\tau_\star) = \lim_{\tau_\star \rightarrow \infty} \alpha_q(\tau_\star) = C_0.
\]

Combining Lemma E.3 with (E.8), (E.9), and (E.10) we obtain as \( \sigma \rightarrow \infty \),

\[
e^{-C_2^2 \sigma^2} (\text{AMSE}(q, \lambda^*_q) - \epsilon \mathbb{E}[G]^2) = e^{-C_2^2 \sigma^2 \tau_\star^2 (R_q(\alpha_q(\tau_\star), \tau_\star) - \epsilon \mathbb{E}[G]^2/\tau_\star^2)}
= e^{-C_2^2 \sigma^2 \tau_\star^2} e^{-\frac{\alpha^2_{q(\tau_\star)}}{2}} (\alpha_q(\tau_\star))^{-3} o(1)
= e^{-\frac{C_2^2 \sigma^2 \tau_\star^2}{q^2} \tau_\star^2 - C^2} (\alpha_q(\tau_\star))^{-3} o(1) = o(1).
\]

We have used the fact \( 0 < C < C_0 \) to get the last equality.

E.3. Proof of Theorem 3.5 for \( q \in (1, 2] \). The basic idea of the proof for \( q \in (1, 2] \) is the same as that for \( q = 1 \). We characterize the convergence rate of \( R_q(\alpha_q(\tau), \tau) \) in Section E.3.1. We can derive the expansion of \( \text{AMSE}(q, \lambda^*_q) \) in Section E.3.2.

E.3.1. Characterizing the convergence rate of \( R_q(\alpha_q(\tau), \tau) \) as \( \tau \rightarrow \infty \) for \( q \in (1, 2] \). We first derive the convergence rate of \( \alpha_q(\tau) \) as \( \tau \rightarrow \infty \).

**Lemma E.4.** For \( q \in (1, 2] \), assume \( G \) has finite moments of all order. We have as \( \tau \rightarrow \infty \),

\[
\frac{\alpha_q(\tau)}{\tau^{2(q-1)}} \rightarrow \left( \frac{q - 1}{q^{\frac{1}{q-1}}} \frac{\mathbb{E}[Z]^{\frac{2}{q-1}}}{\mathbb{E}[B^2 \mathbb{E}[Z]^{\frac{2}{q-1}}} \right)^{q-1}.
\]

**Proof.** First note that Lemma E.1 holds for \( q \in (1, 2] \) as well. Hence \( \alpha_q(\tau) \rightarrow \infty \) as \( \tau \rightarrow \infty \). We aim to characterize its convergence rate. Since \( \eta_2(u; \chi) = \frac{u}{1+2u} \), the result can be easily verified for \( q = 2 \). We will focus on the case \( q \in (1, 2) \). For notational simplicity, we will use \( \alpha \) to represent \( \alpha_q(\tau) \) in the rest of the proof. By the first order condition of the optimality,
we have $\partial_1 R_q(\alpha, \tau) = 0$, which can be further written out:

$$0 = \mathbb{E}[\eta_q(\beta/\tau + Z; \alpha) - \partial_2 \eta_q(\beta/\tau + Z; \alpha)]$$

$$= \mathbb{E} \left[ -q|\eta_q(\beta/\tau + Z; \alpha)|^q \left( \frac{1}{1 + \alpha q(q - 1)|\eta_q(\beta/\tau + Z; \alpha)|^q} \right) \right]$$

$$H_1 = \frac{1}{\tau^2} \mathbb{E} \left[ \eta_q(\beta/\tau + Z; \alpha) \right]$$

$$= \frac{1}{\tau^2} \mathbb{E} \left[ \eta_q(\beta/\tau + Z; \alpha) \right]$$

$$H_2 = \frac{1}{\tau^2} \mathbb{E} \left[ \eta_q(\beta/\tau + Z; \alpha) \right]$$

$$= \frac{1}{\tau^2} \mathbb{E} \left[ \eta_q(\beta/\tau + Z; \alpha) \right]$$

where we have used Lemma B.1 part (v). We now analyze the two terms $H_1$ and $H_2$ respectively. Regarding $H_1$ from Lemma B.1 part (ii) we have

$$\alpha \frac{q + 1}{q - 1} |\eta_q(\beta/\tau + Z; \alpha)|^q \leq \frac{1}{q - 1} \frac{1}{\tau^{q-1}} \frac{\eta_q(\beta/\tau + Z; \alpha)}{q - 1}$$

$$\leq \frac{1}{q^{q-1}} (q - 1), \text{ for } \tau \geq 1.$$
By a similar argument and using DCT, it is not hard to see that,

\[
\lim_{\tau \to \infty} \tau^{\frac{q}{q-1}} I_1 = \frac{E B^2 E|Z|^{\frac{2-q}{q-1}}}{q^{\frac{1}{q-1}} (q-1)},
\]

\[
\lim_{\tau \to \infty} \tau^{\frac{q+1}{q-1}} I_3 = \frac{E B E(\frac{3-q}{q-1} \text{sgn}(Z))}{q^{\frac{2}{q-1}} (q-1)} = 0.
\]

Regarding the term \(I_2\), by using Stein’s lemma and Taylor expansion, we can obtain a sequel of equalities:

\[
I_2 = E \left[ B(Z^2 - 1)\eta_\gamma(B/\tau + Z; \alpha) \right] \alpha^\tau = E \left[ B(Z^2 - 1)\partial_1 \eta_\gamma(\gamma B/\tau + Z; \alpha)B/\tau \right] \alpha^\tau = \frac{1}{\alpha^2} E \left[ \frac{B^2(Z^2 - 1)}{1 + \alpha q(q-1)|\eta_\gamma(\gamma B/\tau + Z; \alpha)|^{q-2}} \right],
\]

where the second step is simply due to Lemma B.1 part (i); \(\gamma \in (0,1)\) is a random variable depending on \(B\) and \(Z\). Again with a similar argument to verify the conditions of DCT we obtain

\[
\lim_{\tau \to \infty} \alpha^\frac{q}{q-1} \tau^2 I_2 = \frac{(2-q)EB^2E|Z|^{\frac{2-q}{q-1}}}{q^{\frac{1}{q-1}} (q-1)^2}.
\]

Finally, based on (E.11) and collecting the results from (E.12), (E.13) and (E.14) enables us to have as \(\tau \to \infty\),

\[
\frac{\alpha}{\tau^{2(q-1)}} = \left[ \frac{\alpha^{q+1} (I_3 - H_1)}{\tau^{2(q-1)} (I_1 + I_2)} \right]^{q-1} \to \left( q-1 \frac{E|Z|^{\frac{2}{q-1}}}{q^{\frac{1}{q-1}} E B^2 E|Z|^{\frac{2-q}{q-1}}} \right)^{q-1}.
\]

We now characterize the convergence rate of \(R_q(\alpha_q(\tau), \tau)\) in the lemma below.

**Lemma E.5.** Suppose \(1 < q \leq 2\) and \(G\) has finite moments of all orders, then as \(\tau \to \infty\),

\[
R_q(\alpha_q(\tau), \tau) = \frac{e^2E|G|^2}{\tau^2} - \frac{e^2(E|G|^2E|Z|^{\frac{2-q}{q-1}})^2}{(q-1)^2E|Z|^{\frac{2}{q-1}}} \frac{1}{\tau^4} + o(1/\tau^4).
\]
Proof. It is straightforward to prove the result for $q = 2$. From now on we only consider $1 < q < 2$. We write $\alpha$ for $\alpha_q(\tau)$ in the rest of the proof to simplify the notation. First we have

\[
R_q(\alpha, \tau) - \frac{\epsilon E|G|^2}{\tau^2} = E\eta_q^2(B/\tau + Z; \alpha) - 2E[\eta_q(B/\tau + Z; \alpha)B/\tau]
\]

\[
= E\eta_q^2(B/\tau + Z; \alpha) - 2E[(\eta_q(Z; \alpha) + \partial_1\eta_q(\gamma B/\tau + Z; \alpha)B/\tau)]B/\tau
\]

(E.15) \[
= E\eta_q^2(B/\tau + Z; \alpha) - 2E[\partial_1\eta_q(\gamma B/\tau + Z; \alpha)B^2/\tau^2],
\]

where we have used Taylor expansion in the second step and $\gamma \in (0,1)$ is a random variable depending on $B, Z$. According to Lemma B.1 part (ii),

\[
\alpha^{2-1} \eta_q^2(B/\tau + Z; \alpha) = q^{\frac{2}{1-q}}(||B/\tau + Z| - |\eta_q(B/\tau + Z; \alpha)||)^2
\]

\[
\leq q^{\frac{2}{1-q}}(||B| + |Z||)^2, \quad \text{for } \tau \geq 1.
\]

The upper bound is integrable since $G$ has finite moments of all orders. Hence we can apply DCT to obtain

(E.16) \[
\lim_{\tau \to \infty} \alpha^{2-1} E\eta_q^2(B/\tau + Z; \alpha) = q^{\frac{2}{1-q}} E|Z|^{\frac{2}{1-q}}.
\]

We can follow a similar argument to use DCT to have

(E.17) \[
\lim_{\tau \to \infty} \frac{\alpha^{2-1}}{\tau^2} E[\partial_1\eta_q(\gamma B/\tau + Z; \alpha)B^2/\tau^2]
\]

\[
(\alpha) \lim_{\tau \to \infty} \frac{1}{\tau^2} \lim_{\tau \to \infty} \frac{B^2}{\alpha^{2-1} + q(q - 1)|\alpha^{1-1} \eta_q(\gamma B/\tau + Z; \alpha)|^{q-2}}
\]

\[
(\beta) \frac{q - 1}{q^{\frac{1}{q-1}}} \frac{E|Z|^{\frac{2}{q-1}}}{EB^2E|Z|^\frac{2-q}{q-1}} \frac{EB^2E|Z|^\frac{2-q}{q-1}}{q^{\frac{1}{q-1}}(q - 1)} = q^{\frac{2}{1-q}} E|Z|^{\frac{2}{1-q}},
\]

where (a) holds due to Lemma B.1 part (iv); we have used Lemma E.4 and DCT to obtain (b). Finally, we put the results (E.15), (E.16), (E.17) and
Lemma E.4 together to derive

\[
\lim_{\tau \to \infty} \tau^4 (R_q(\alpha, \tau) - \epsilon \mathbb{E}[G]^2/\tau^2) = \lim_{\tau \to \infty} \frac{\tau^4}{\alpha_q^2} \cdot \left[ \lim_{\tau \to \infty} \alpha^{2m+1} \mathbb{E}\eta_q^2(B/\tau + Z; \alpha) - 2 \lim_{\tau \to \infty} \alpha^{2m+1} \mathbb{E}(\partial_1 \eta_q(\gamma B/\tau + Z; \alpha)B^2/\tau^2) \right]
\]

\[
= \left( \frac{q-1}{q^{2-4}} \mathbb{E}[G]\mathbb{E}[Z]^{2-4} \right)^{-2} \cdot (q^{2-4} \mathbb{E}[G]\mathbb{E}[Z]^{2-4} - 2q^{2-4} \mathbb{E}[G]\mathbb{E}[Z]^{2-4})
\]

\[
= \frac{\epsilon^2 (\mathbb{E}[G]^2 \mathbb{E}[Z]^{2-4})^2}{(q-1)^2 \mathbb{E}[G] \mathbb{E}[Z]^{2-4}}.
\]

This finishes the proof. \(\square\)

E.3.2. Deriving the expansion of \(\text{AMSE}(q, \lambda_q^*)\) for \(q \in (1, 2]\). The way we derive the result (3.9) of Theorem 3.5 is similar to that in Section E.2.3. We hence do not repeat all the details. The key step is applying Lemma E.5 to obtain

\[
\lim_{\sigma \to \infty} \sigma^2 (\text{AMSE}(q, \lambda_q^*) - \epsilon \mathbb{E}[G]^2) = \lim_{\tau \to \infty} \sigma^2 (\text{AMSE}(q, \lambda_q^*) - \epsilon \mathbb{E}[G]^2)
\]

\[
= \lim_{\tau \to \infty} \tau^4 (R_q(\alpha_q^*(\tau_q^*), \tau_q^*) - \epsilon \mathbb{E}[G]^2/\tau_q^*)
\]

\[
= -\epsilon^2 (\mathbb{E}[G]^2)^2 c_q.
\]

E.4. Proof of Theorem 3.5 for \(q > 2\). We aim to prove the same results as presented in Lemmas E.4 and E.5. However, many of the limits we took when proving for the case \(1 < q \leq 2\) become invalid for \(q > 2\) because DCT may not be applicable. Therefore, here we assume a slightly stronger condition that \(G\) has a sub-Gaussian tail and use a different reasoning to validate the results in Lemmas E.4 and E.5. Throughout this section, we use \(\alpha\) to denote \(\alpha_q(\tau)\) for simplicity. First note that Lemma E.1 holds for \(q > 2\) as well. Hence we already know \(\alpha \to \infty\) as \(\tau \to \infty\). The following key lemma paves our way for the proof.

**Lemma E.6.** Suppose function \(h : \mathbb{R}^2 \to \mathbb{R}\) satisfies \(|h(x, y)| \leq C(|x|^{m_1} + |y|^{m_2})\) for some \(C > 0\) and \(0 \leq m_1, m_2 < \infty\). \(B\) has sub-Gaussian tail. Then
the following result holds for any constants \( v \geq 0, \gamma \in [0, 1] \) and \( q > 2 \),

\[
\text{(E.18)} \quad \lim_{\tau \to \infty} \alpha^{\frac{v+1}{q-1}} \mathbb{E} \left[ \frac{h(B, Z)|\eta_q(B/\tau + Z; \alpha)|^v}{1 + \alpha q(q - 1)|\eta_q(\gamma B/\tau + Z; \alpha)|^{q-2}} \right] = \frac{q^{-\frac{v+1}{q-1}}}{q-1} \mathbb{E}[h(B, Z)|Z|^{\frac{v+2-q}{q-1}}], \quad \text{as } \tau \to \infty.
\]

Moreover, there is a finite constant \( K \) such that for sufficiently large \( \tau \),

\[
\text{(E.19)} \quad \max_{0 \leq \gamma \leq 1} \alpha^{\frac{v+1}{q-1}} \mathbb{E} \left[ \frac{|h(B, Z)||\eta_q(B/\tau + Z; \alpha)|^v}{1 + \alpha q(q - 1)|\eta_q(\gamma B/\tau + Z; \alpha)|^{q-2}} \right] \leq K.
\]

**Proof.** Define

\[ A = \{|\eta_q(\gamma B/\tau + Z; \alpha)| \leq \frac{1}{2} |\gamma B/\tau + Z|\}. \]

We evaluate the expectation on the set \( A \) and its complement \( A^c \) respectively. Recall we use \( p_B \) to denote the distribution of \( B \). By a change of variable we then have

\[
\mathbb{E} \left[ \frac{h(B, Z)|\eta_q(B/\tau + Z; \alpha)|^v}{1 + \alpha q(q - 1)|\eta_q(\gamma B/\tau + Z; \alpha)|^{q-2}} \right] = \int h(x, y - \gamma x/\tau)|\eta_q(y + (1 - \gamma)x/\tau; \alpha)|^v \frac{\phi(y - \gamma x/\tau)dy}{1 + \alpha q(q - 1)|\eta_q(y; \alpha)|^{q-2}} dp_B(x).
\]

We have on \( \mathbb{I}_{\{|\eta_q(y; \alpha)| \leq \frac{1}{2} |y|\}} \) when \( \tau \) is large enough,

\[
\alpha^\frac{v+1}{q-1} \frac{h(x, y - \gamma x/\tau)|\eta_q(y + (1 - \gamma)x/\tau; \alpha)|^v}{1 + \alpha q(q - 1)|\eta_q(y; \alpha)|^{q-2}} \phi(y - \gamma x/\tau)
\]

\[
(a) \leq \frac{|h(x, y - \gamma x/\tau)||\eta_q(y + (1 - \gamma)x/\tau; \alpha)|^v}{q(q - 1)(|y| - |\eta_q(y; \alpha)|)^\frac{q}{q-1}} \phi(y/\sqrt{2})e^{-1/4(y - \frac{2\gamma x}{\sqrt{v}})^2 + \frac{\gamma^2 x^2}{2v^2}}
\]

\[
(b) \leq \frac{q^{-\frac{v}{q-1}}|h(x, y - \gamma x/\tau)||\eta_q(y + (1 - \gamma)x/\tau; \alpha)|^v}{q(q - 1)|y|^{\frac{q}{q-1}}} \phi(y/\sqrt{2})e^\frac{\gamma^2 x^2}{2v}
\]

\[
(c) \leq \frac{2^{\frac{v-2}{q}} q^{-\frac{v}{q-1}}|h(x, y - \gamma x/\tau)||\eta_q(y + (1 - \gamma)x/\tau; \alpha)|^v}{q(q - 1)|y|^{\frac{q}{q-1}}} \phi(y/\sqrt{2})e^\frac{\gamma^2 x^2}{2v^2}
\]

\[
(d) \leq \frac{2^{\frac{v-2}{q}} q^{-\frac{v}{q-1}} (|x|^{m_1} + (|y| + |x|)^{m_2}) |\eta_q(y + (1 - \gamma)x/\tau; \alpha)|^v}{q(q - 1)|y|^{\frac{q}{q-1}}} \phi(y/\sqrt{2})e^\frac{\gamma^2 x^2}{2v^2}.
\]

We have used Lemma B.1 part (ii) to obtain (a)(b); (c) is due to the condition \( |\eta_q(y; \alpha)| \leq \frac{1}{2} |y| \); and (d) holds because of the condition on the function
$h(x,y)$. Notice that the numerator of the upper bound is essentially a polynomial in $|x|$ and $|y|$. Since $B$ has sub-Gaussian tail, if we choose $c_0$ small enough (when $τ$ is sufficiently large), the integrability with respect to $x$ can be guaranteed. The integrability w.r.t. $y$ is clear since $(2 - q)/(q - 1) > -1$. Thus we can apply DCT to obtain

$$
\lim_{τ→∞} \frac{α^{v+1}}{q-v} \mathbb{E} \left[ \frac{h(B, Z) |η_q(B/τ + Z; α)^v}{1 + αq(q - 1)|η_q(γB/τ + Z; α)^q - 2 A} \right]
$$

$$
= \lim_{τ→∞} \int \frac{h(x, y - γx/τ)|α^{1/v} |η_q(y + (1 - γ)x/τ; α)^v}{α^{1/v} + q(q - 1)|η_q(y; α)|^{q-2}} φ(y - γx/τ) \{ |η_q(y; α)| ≤ \frac{1}{2} |y| \} dydp_B(x)
$$

$$
= \int \frac{q^{1-v} h(x, y)}{(q - 1)|y|^{q-2}} φ(y) dydp_B(x) = \frac{q^{1-v}}{q - 1} \mathbb{E}[h(B, Z) |Z|^{\frac{q+2-v}{q-1}}].
$$

We now evaluate the expectation on the event $A^c$. Note that $A^c$ implies

$$
|γB/τ + Z| = αq|η_q(γB/τ + Z; α)|^{q-1} + |η_q(γB/τ + Z; α)| > \frac{αq}{2q-1} |γB/τ + Z|^{q-1} + \frac{1}{2} |γB/τ + Z|
$$

$$
⇒ |γB/τ + Z| < 2(αq)^{\frac{1}{q-γ}}.
$$

Hence we have the following bounds,

$$
\frac{α^{v+1}}{q-v} \mathbb{E} \left[ \frac{|h(B, Z)| \cdot |η_q(B/τ + Z; α)^v}{1 + αq(q - 1)|η_q(γB/τ + Z; α)^q - 2 A} \right]
$$

$$
≤ \frac{α^{v+1}}{q-v} \mathbb{E} \left( |h(B, Z)| \cdot |η_q(B/τ + Z; α)|^v \right) A^c
$$

$$
≤ \frac{α^{1/v}}{q-1} \int |y|^{q-2(αq)} \frac{1}{^{q-γ}} |h(x, y - γx/τ)| \cdot |α^{1/v} |η_q(y + (1 - γ)x/τ; α)^v| \cdot \phi(y) e^{γy} dydp_B(x)
$$

$$
(c) \leq q^{1-v} \frac{α^{1/v}}{q-1} \int |y|^{q-2(αq)} \frac{1}{^{q-γ}} \left(|x|^{m_1} + (|y| + |x|)^{m_2}\right) (|y| + |x|)^{q-γ} φ(y) e^{2(αq)^{\frac{1}{q-γ}} x} dydp_B(x)
$$

$$
≤ q^{1-v} \frac{α^{1/v}}{q-1} \int |y|^{q-2(αq)} \frac{1}{^{q-γ}} \left(|x|^{m_1} + (|y| + |x|)^{m_2}\right) P(|x|, |y|) φ(y) e^{2(αq)^{\frac{1}{q-γ}} x} dydp_B(x)
$$

$$
(f) \leq c_1 α^{1/v} \frac{1}{q-1} \int \tilde{P}(|x|) e^{\tau x} dP_B(x) \leq c_2 α^{1/(q-1)} \to \infty \text{ as } τ \to \infty,
$$

where $(c)$ is due to Lemma B.1 part (ii) and condition on $h(x,y); P(·,·), \tilde{P}(·)$ are two polynomials; the extra term $α^{1/v}$ in step $(f)$ is derived from the condition $|y| < 2(αq)^{\frac{1}{q-γ}}$. We thus have finished the proof of (E.18). Finally,
note that the two upper bounds we derived do not depend on $\gamma$, hence (E.19) follows directly.

We are now in position to prove Theorem 3.5 for $q > 2$. We will be proving the results of Lemmas E.4 and E.5 for $q > 2$. After that the exactly same arguments presented in Section E.3.2 will close the proof. Since the basic idea of proving Lemmas E.4 and E.5 for $q > 2$ is the same as for the case $q \in (1, 2]$, we do not detail out the entire proof and instead highlight the differences. The major difference is that we apply Lemma E.6 to make some of the limiting arguments valid in the case $q > 2$. Adopting the same notations in Section E.3.1, we list the settings in the use of Lemma E.6 below.

- **Lemma E.4 $I_1$**: set $h(x, y) = x^2, v = 0, \gamma = 1$.
- **Lemma E.4 $I_2$**: set $h(x, y) = x \text{sgn}(\frac{x}{\tau} + y), v = 1, \gamma = 1$. Note that the dependence of $h(x, y)$ on $\tau$ does not affect the result.
- **Lemma E.4 $I_3$**: Notice we have

$$\alpha^{\frac{q}{q-1}} \tau^2 I_2 = \alpha^{\frac{1}{q-1}} \tau^2 \mathbb{E} \left[ B(Z^2 - 1) \left( \eta_q(Z; \alpha) + \frac{B}{\tau} \int_0^1 \partial_1 \eta_q(sB/\tau + Z; \alpha) ds \right) \right]$$

$$= \alpha^{\frac{1}{q-1}} \tau^2 \mathbb{E} \left[ B^2(Z^2 - 1) \partial_1 \eta_q(sB/\tau + Z; \alpha) \right] ds$$

$$= \int_0^1 \alpha^{\frac{1}{q-1}} \mathbb{E} \left[ \frac{B^2(Z^2 - 1)}{1 + \alpha q(q - 1) \mid \eta_q(sB/\tau + Z; \alpha) \mid^{q-2}} \right] ds.$$

We have switched the integral and expectation in the second step above due to the integrability. Set $h(x, y) = x^2(y^2 - 1), v = 0, \gamma = s$; then by the bound (E.19) in Lemma E.6, we can bring the limit $\tau \to \infty$ inside the above integral to obtain the result of $\lim_{\tau \to \infty} \alpha^{\frac{q}{q-1}} \tau^2 I_2$.

- In Lemma E.5, we need rebound the term $\mathbb{E}[\eta_q(B/\tau + Z; \alpha)B/\tau]$ in (E.15).

$$\alpha^{\frac{1}{q-1}} \tau^2 \mathbb{E}[\eta_q(B/\tau + Z; \alpha)B/\tau]$$

$$= \alpha^{\frac{1}{q-1}} \tau^2 \mathbb{E} \left[ \frac{B}{\tau} \left( \eta_q(Z; \alpha) + \frac{B}{\tau} \int_0^1 \partial_1 \eta_q(sB/\tau + Z; \alpha) ds \right) \right]$$

$$= \int_0^1 \alpha^{\frac{1}{q-1}} \mathbb{E} \left[ \frac{B^2}{1 + \alpha q(q - 1) \mid \eta_q(sB/\tau + Z; \alpha) \mid^{q-2}} \right] ds$$

We set $h(x, y) = x^2, v = 0, \gamma = s$. The rest arguments are similar to the previous one.
APPENDIX F: PROOF OF THEOREMS 3.4 AND 3.6

The proof of Theorems 3.4 and 3.6 can be found in Sections F.1 and F.1 respectively. Since the roadmap of these two proofs are similar to that of Theorem 3.5, we will not repeat it. We suggest the reader study Appendix E before reading this appendix.

F.1. Proof of Theorem 3.4. The proof for $q \in [1, 2]$ has been shown in [WMZ16]. We now prove the case $q > 2$. We first derive the convergence rate of $\alpha_q(\tau)$ as $\tau \to 0$.

**Lemma F.1.** For a given $q \in (2, \infty)$, assume $E|G|^{2q-2} < \infty$, then

$$
\lim_{\tau \to 0} \frac{\alpha_q(\tau)}{\tau^q} = \frac{(q-1)E|G|^{q-2}}{qE|G|^{2q-2}}.
$$

**Proof.** The basic idea of the proof is the same as the one for $q \in (1, 2]$ in [WMZ16]. But due to the range of $q$ in the current setting, several steps can be significantly simplified. In the rest of the proof, we will use $\alpha$ to denote $\alpha_q(\tau)$ for notational simplicity. We first claim $\tau^{2-q}\alpha_n \to 0$ as $\tau \to 0$. Otherwise there exists a sequence such that $\tau_n^{2-q}\alpha_n \to C > 0$ and $\tau_n \to 0$ as $n \to \infty$. Because $\alpha_n$ is the optimal tuning, we have as $n \to \infty$

$$
1 = R_q(0, \tau_n) \geq R_q(\alpha_n, \tau_n) > \epsilon\tau_n^{2-q}E(\eta_q(G + \tau_nZ; \tau_n^{2-q}\alpha_n) - G)^2 \to +\infty,
$$

causing a contradiction.

Since $\alpha$ is the minimizer of $R_q(\alpha, \tau)$, we know $\partial_1 R(\alpha, \tau) = 0$, which can be further written out as follows:

$$
0 = (1 - \epsilon)E\eta_q(Z; \alpha)\partial_2 \eta_q(Z; \alpha) + \epsilon E(\eta_q(G/\tau + Z; \alpha) - G/\tau)\partial_2 \eta_q(G/\tau + Z; \alpha)
\overset{(a)}{=} (1 - \epsilon) E \frac{-q|\eta_q(Z; \alpha)|^q}{1 + \alpha q(q-1)|\eta_q(Z; \alpha)|^{q-2}} H_1 + \epsilon E(Z\partial_2 \eta_q(G/\tau + Z; \alpha)) H_2
\overset{(F.1)}{=} + \epsilon \alpha E \frac{q^2|\eta_q(G/\tau + Z; \alpha)|^{2q-2}}{1 + \alpha q(q-1)|\eta_q(G/\tau + Z; \alpha)|^{q-2}} H_3,
$$

where we have used Lemma B.1 parts (ii)(v) to obtain (a). Since $|\eta_q(Z; \alpha)| \leq |Z|$ and $\tau^{2-q}\alpha \to 0$ as $\tau \to 0$, we can apply DCT to have

$$
\lim_{\tau \to 0} H_1 = -qE|Z|^q.
$$
Finally, based on the Equation (F.1), we are able to obtain
\[
\lim_{\tau \to 0} \tau^{q-2} H_3 = q^2 \mathbb{E}|G|^{2q-2}.
\]

Regarding \( H_2 \), we can use Stein’s lemma to reformulate it as:
\[
H_2 = q(1-q) \mathbb{E} \frac{|\eta_q(G/\tau + Z; \alpha)|^{q-2}}{(1+\alpha q(q-1))|\eta_q(G/\tau + Z; \alpha)|^{q-2}} J_1(G,Z,\tau)
+ q^2(1-q) \mathbb{E} \frac{\alpha|\eta_q(G/\tau + Z; \alpha)|^{2q-4}}{(1+\alpha q(q-1))|\eta_q(G/\tau + Z; \alpha)|^{2q-2}} J_2(G,Z,\tau).
\]

Clearly \(|\tau^{q-2} J_1(G, Z, \tau)| \leq (|G| + |Z|)^{q-2}\) for \( \tau < 1 \). We can thus use DCT to obtain
\[
\lim_{\tau \to 0} \mathbb{E}|\tau^{q-2} J_1(G, Z, \tau)| = \mathbb{E}|G|^{q-2}.
\]

Similar argument shows \( \lim_{\tau \to 0} \mathbb{E}|\tau^{q-2} J_2(G, Z, \tau)| = 0 \). Note that we have used the result \( \tau^{2-q} \alpha = o(1) \) we proved at the beginning. Hence,
\[
\lim_{\tau \to 0} \tau^{q-2} H_2 = q(1-q) \mathbb{E}|G|^{q-2}.
\]

Finally based on the Equation (F.1), we are able to obtain
\[
\lim_{\tau \to 0} \frac{\alpha}{\tau^q} = \frac{(\epsilon - 1) \lim_{\tau \to 0} \tau^{q-2} H_1 - \epsilon \lim_{\tau \to 0} \tau^{q-2} H_2}{\epsilon \lim_{\tau \to 0} \tau^{2q-2} H_3} = \frac{(q - 1) \mathbb{E}|G|^{q-2}}{q \mathbb{E}|G|^{2q-2}}.
\]

We now derive the convergence rate of \( R_q(\alpha_q(\tau), \tau) \) in the next lemma.

**Lemma F.2.** For a given \( q \in (2, \infty) \), assume \( \mathbb{E}|G|^{2q-2} < \infty \), then
\[
R_q(\alpha_q(\tau), \tau) = 1 - \tau^2 \frac{\epsilon(q-1)^2}{\mathbb{E}|G|^{q-2}} + o(\tau^2).
\]

**Proof.** We write \( \alpha \) for \( \alpha_q(\tau) \) for simplicity. Note that,
\[
R_q(\alpha, \tau) - 1 = (1 - \epsilon) \mathbb{E}(\eta_q^2(Z; \alpha) - Z^2) + \epsilon \mathbb{E}(\eta_q(G/\tau + Z; \alpha) - G/\tau - Z)^2
+ 2\epsilon \mathbb{E}[Z(\eta_q(G/\tau + Z; \alpha) - G/\tau - Z)]
\]
\[
\overset{(a)}{=} (1 - \epsilon) \mathbb{E}(\eta_q^2(Z; \alpha) - Z^2) + \epsilon \underbrace{\alpha^2 q^2 \mathbb{E}|\eta_q(G/\tau + Z; \alpha)|^{2q-2}}_{K_1}
- 2\epsilon \alpha q(q-1) \underbrace{\mathbb{E} \frac{|\eta_q(G/\tau + Z; \alpha)|^{q-2}}{1 + \alpha q(q-1)|\eta_q(G/\tau + Z; \alpha)|^{q-2}}}_{K_3}.
\]
where we have used Stein’s lemma and Lemma B.1 parts (ii)(iv) to obtain (a). We now analyze the three terms $K_1, K_2, K_3$ respectively. For $K_1$, it is clear that

$$K_1 = (1 - \epsilon)\mathbb{E}(\eta_q(Z; \alpha) + Z)(\eta_q(Z; \alpha) - Z)$$

$$= (1 - \epsilon)\mathbb{E}(\eta_q(Z; \alpha) + Z)(\alpha q \eta_q(Z; \alpha)|^{q-1}\text{sign}(Z))$$

$$= (1 - \epsilon)\alpha q (\mathbb{E}|\eta_q(Z; \alpha)|^q + \mathbb{E}|Z||\eta_q(Z; \alpha)|^{q-1}).$$

Then according to Lemma F.1, it is straightforward to confirm as $\tau \to 0$

$$\tau^{-2}K_1 = o(1).$$

Regarding $K_2$, we can apply DCT and Lemma F.1 to have

$$\lim_{\tau \to 0} \tau^{-2}K_2 = \epsilon q^2 \lim_{\tau \to 0} \alpha^2 \tau^{-2q} \cdot \lim_{\tau \to 0} \mathbb{E}|\eta_q(G + \tau Z; \tau^{-2q}\alpha)|^{2q-2}$$

$$= \epsilon(q - 1)^2(\mathbb{E}|G|^2q-2)^2(\mathbb{E}|G|^{2q-2})^{-1}.$$ 

A similar argument gives us

$$\lim_{\tau \to 0} \tau^{-2}K_3 = -2\epsilon(q - 1)^2(\mathbb{E}|G|^2q-2)^2(\mathbb{E}|G|^{2q-2})^{-1}.$$ 

Finally, we are able to obtain

$$\lim_{\tau \to 0} \tau^{-2}(R_q(\alpha, \tau) - 1) = \lim_{\tau \to 0} \tau^{-2}K_1 + \lim_{\tau \to 0} \tau^{-2}K_2 + \lim_{\tau \to 0} \tau^{-2}K_3$$

$$= -\epsilon(q - 1)^2(\mathbb{E}|G|^2q-2)^2(\mathbb{E}|G|^{2q-2})^{-1}.$$ 

\[\square\]

We are now in position to derive the expansion of AMSE($q, \lambda_q^*$) in (3.7). From Theorem 2.1 we know

(F.2) \hspace{1cm} AMSE($q, \lambda_q^*$) = $\tau_q^* R_q(\alpha_q(\tau_q^*), \tau_q^*) = \delta(\tau_q^2 - \sigma^2).$

Since $\alpha = \alpha_q(\tau_q)$ minimizes $\alpha_q(\alpha, \tau_q)$, we have

$$\delta(\tau_q^2 - \sigma^2) \leq \tau_q^2 R_q(0, \tau_q) = \tau_q^2,$$

which is equivalent to $(\delta - 1)\tau_q^2 \leq \delta\sigma^2$. By the condition $\delta > 1$, we obtain $\tau_q \to 0$ as $\sigma \to 0$. Based on Equation (F.2), this result combined with Lemma F.2 gives us

(F.3) \hspace{1cm} \lim_{\sigma \to 0} \frac{\tau_q^2}{\sigma^2} = \frac{\delta}{\delta - 1}.$
It is straightforward to confirm the following from (F.2):

\[
\sigma^{-4} \left( \text{AMSE}(q, \lambda^*_q) - \frac{\sigma^2}{1 - 1/\delta} \right) = \sigma^{-4} \left( \tau_*^2 R_q(\alpha_q(\tau_*), \tau_*) - \frac{\delta}{\delta - 1}(\tau_*^2 - \tau_*^2 R_q(\alpha_q(\tau_*), \tau_*)/\delta) \right) = \frac{\delta}{\delta - 1} \cdot [\tau_*^{-2}(R_q(\alpha_q(\tau_*), \tau_*) - 1)] \cdot (\tau_*^4 \sigma^{-4})
\]

Letting \( \sigma \to 0 \) above and using (F.3) and Lemma F.2 finishes the proof.

**F.2. Proof of Theorem 3.6.** We remind the reader that in the large sample regime, we have scaled the noise term. Hence \( \tau_* \) will satisfy

\[
(\text{F.4}) \quad \tau_*^2 = \frac{\sigma^2}{\delta} + \frac{\tau_*^2 R_q(\alpha_q(\tau_*), \tau_*)}{\delta}.
\]

We first derive the convergence rate of \( \tau_* \) as \( \delta \to \infty \).

**Lemma F.3.** For a given \( q \in [1, \infty) \), as \( \delta \to \infty \),

\[
\tau_*^2 = \frac{\sigma^2}{\delta} + o \left( \frac{1}{\delta} \right).
\]

**Proof.** Since \( \alpha = \alpha_q(\tau_*) \) minimizes \( R_q(\alpha, \tau_*) \), from (F.4) we obtain

\[
(\text{F.5}) \quad \delta(\tau_*^2 - \sigma^2/\delta) \leq R_q(0, \tau_*) = \tau_*^2,
\]

which gives us \( \tau_*^2 \leq \sigma^2/(\delta - 1) \to 0 \) as \( \delta \to \infty \). This result combined with (F.5) completes the proof.

Lemma F.3 shows that \( \tau_* \to 0 \) as \( \delta \to \infty \). Hence we need to characterize the convergence rate of \( R_q(\alpha_q(\tau), \tau) \) as \( \tau \to 0 \). Luckily the results have been derived in the small noise regime analysis. We collect the results together in the next lemma.

**Lemma F.4.** As \( \tau \to 0 \) we have

1. For \( q = 1 \), assume \( \mathbb{P}(|G| \geq \mu) = 1 \) with \( \mu \) being a positive constant and \( \mathbb{E}|G|^2 < \infty \), then

\[
R_q(\alpha_q(\tau), \tau) - f(\chi_0) = O(\phi(\mu/\tau - \chi_0)),
\]

where \( \chi = \chi_0 \) is the minimizer of the function \( f(\chi) = (1-\epsilon)\mathbb{E}r_q^2(Z; \chi) + \epsilon(1 + \chi^2) \).
(2) For $1 < q < 2$, assume $P(|G| \leq x) = O(x)$ (as $x \to 0$) and $E|G|^2 < \infty,$

$$R_q(\alpha_q(\tau), \tau) = 1 - \frac{(1 - \epsilon)^2 (E|Z|^q)^2}{\epsilon E|G|^{2q-2}} \tau^{2q-2} + o(\tau^{2q-2}).$$

(3) For $q > 2$, assume $E|G|^{2q-2} < \infty$, then

$$R_q(\alpha_q(\tau), \tau) = 1 - \tau^2 \epsilon (q - 1)^2 (E|G|^{q-2})^2 + o(\tau^2).$$

PROOF. Result (1) is Lemma 5 in [WMZ16]; Result (2) is Lemma 20 in [WMZ16]; Result (3) is Lemma F.2 in Section F.1.

We now use the results in Lemmas F.3 and F.4 to prove Theorem 3.6. We only present the proof for $q \in (1, 2)$. Similar arguments work for other values of $q$. By Theorem 2.1 and (F.4),

$$\delta^q(AMSE(q, \lambda^*) - \sigma^2/\delta) = \delta^q(\tau^2 R_q(\alpha_q(\tau_*), \tau_* - \tau_*^2 + \tau_*^2 R_q(\alpha_q(\tau_*), \tau_*))/\delta)
= \delta^q \tau_s^2 (R_q(\alpha_q(\tau_*), \tau_* - 1) + \delta^q \tau_s^2 R_q(\alpha_q(\tau_*), \tau_*))
\overset{(a)}{\to} -\sigma q (1 - \epsilon)^2 (E|Z|^q)^2/\epsilon E|G|^{2q-2}, \text{ as } \delta \to \infty.$$

The step (a) is due to Lemma F.3 and Lemma F.4 part (2). This finishes the proof.

APPENDIX G: PROOF OF THEOREMS 4.1, 4.2, 4.3 AND LEMMA 4.1

The proof of Theorems 4.1, 4.2 (4.3) and Lemma 4.1 can be found in Sections G.1, G.2 and G.3 respectively.

G.1. Proof of Theorem 4.1. Since some technical details for $q = 1$ and $q > 1$ are different, we prove the two cases separately in Sections G.1.2 and G.1.1 respectively.

G.1.1. Proof of Theorem 4.1 for $q = 1$. In this section, we apply the approximate message passing (AMP) framework to prove the result for LASSO. We first briefly review the approximate message passing algorithm and state some relevant results that will be later used in the proof. We then describe the main proof steps.

I. Approximate message passing algorithms. [BM12] has utilized AMP theory to characterize the sharp asymptotic risk of LASSO. The authors considered a sequence of estimates $\beta^t \in \mathbb{R}^p$ generated from an approximate message passing algorithm with the following iterations (initialized at
\( \beta^0 = 0, z^0 = y \): 

\[
\begin{align*}
\beta^{t+1} &= \eta_q (X^T z^t + \beta^t; \alpha \tau^{2-q}_t), \\
z^t &= y - X \beta^t + \frac{1}{\delta} z^{t-1} \langle \partial_q \eta_q (X^T z^{t-1} + \beta^{t-1}; \alpha \tau^{2-q}_{t-1}) \rangle,
\end{align*}
\]

(G.1)

where \( \langle v \rangle = \frac{1}{p} \sum_{i=1}^{p} v_i \) denotes the average of a vector’s components; \( \alpha \) is the solution to Equations (2.1) and (2.2); and \( \tau_t \) satisfies \( \tau^2_0 = \sigma^2 + E |B|^2 / \delta \):

\[
\tau^2_{t+1} = \sigma^2 w + \frac{1}{\delta} E [\eta_q (B + \tau_t Z; \alpha \tau^{2-q}_t) - B]^2, \quad t \geq 0.
\]

(G.2)

Remarkably, the asymptotics of many quantities in the AMP algorithm can be sharply characterized. We summarize some results of [BM12] that we will use in our proof.

**Theorem G.1 ([BM12])**. Let \( \{\beta(p), X(p), w(p)\} \) be a converging sequence, and \( \psi : \mathbb{R}^2 \to \mathbb{R} \) be a pseudo-Lipschitz function. For \( q = 1 \), almost surely

\[
\begin{align*}
(i) & \quad \lim_{t \to \infty} \lim_{p \to \infty} \frac{1}{p} \| \hat{\beta}(1, \lambda) - \beta^t \|^2_2 = 0, \\
(ii) & \quad \lim_{n \to \infty} \frac{1}{n} \| z^t \|^2_2 = \tau^2_t, \quad \lim_{t \to n \to \infty} \frac{1}{n} \| z^t - z^{t-1} \|^2_2 = 0, \\
(iii) & \quad \lim_{p \to \infty} \frac{1}{p} \| \beta^t \|^0_0 = \mathbb{P}(|B + \tau_t Z| > \alpha \tau_t), \\
(iv) & \quad \lim_{p \to \infty} \frac{1}{p} \sum_{i=1}^{p} \psi(\beta^t_i + (X^T z^t)_i, \beta_i) = \mathbb{E} \psi(B + \tau_t Z, B), \\
(v) & \quad \lim_{p \to \infty} \frac{1}{p} \sum_{i=1}^{p} \psi(\beta^t_i, \beta_i) = \mathbb{E} \psi(\eta_1 (B + \tau_t Z), B),
\end{align*}
\]

where \( \hat{\beta}(1, \lambda) \) is the LASSO solution and \( \tau_t \) is defined in (G.2).
II. Main proof steps. We first have the following bounds:

\[
\frac{1}{p} \| \hat{\beta}^t(1, \lambda) - \beta^t - X^T \frac{y - X\beta^t}{1 - \|\beta^t\|_0/n} \|^2_2 \leq \frac{2}{p} \| \hat{\beta}(1, \lambda) - \beta^t \|^2_2 + Q_1
\]

\[
\frac{8}{p(1 - \mathbb{P}(|B + \tau Z| > \alpha \tau)/\delta)} \| X^T X(\hat{\beta}(1, \lambda) - \beta^t) \|^2_2 + Q_2
\]

\[
\frac{8}{p} \| X^T(y - X\hat{\beta}(1, \lambda)) \|^2_{2} \left( \frac{1}{1 - \|\hat{\beta}(1, \lambda)\|_0/n} - \frac{1}{1 - \mathbb{P}(|B + \tau Z| > \alpha \tau)/\delta} \right)^2 Q_3
\]

\[
\frac{8}{p} \left[ \frac{1}{1 - \|\beta^t\|_0/n} - \frac{1}{1 - \mathbb{P}(|B + \tau Z| > \alpha \tau)/\delta} \right] Q_4
\]

where \((\alpha, \tau)\) is the solution to (2.1) and (2.2). From Theorem G.1 part (i), we know \(\lim_{t \to \infty} \lim_{p \to \infty} Q_1 = 0\), a.s. Since the largest singular value of \(X\) is bounded almost surely [BY93], we can also obtain \(\lim_{t \to \infty} \lim_{p \to \infty} Q_2 = 0\), a.s. Moreover, from Theorem 2.1 we can easily see the term \(\|X^T(y - X\hat{\beta}(1, \lambda))\|^2_2/p \leq 2\|X^T X(\hat{\beta}(1, \lambda) - \beta)\|^2_2/p + 2\|X^T w\|^2_2/p\) is almost surely bounded. Also we know from [BvdBSC13] that \(\frac{1}{p}\|\hat{\beta}(1, \lambda)\|_0 = \mathbb{P}(|B + \tau Z| > \alpha \tau), a.s.\). Therefore, we obtain \(\lim_{p \to \infty} Q_3 = 0\), a.s. Regarding \(Q_4\), it is not hard to see from (G.2) that \(\tau_t \to \tau\) as \(t \to \infty\). Then a similar argument as for \(Q_3\) combined with Theorem G.1 parts (i)(iii) gives us \(\lim_{t \to \infty} \lim_{p \to \infty} Q_4 = 0\), a.s. Above all we are able to derive almost surely

\[
\lim_{t \to \infty} \lim_{p \to \infty} \frac{1}{p} \| \hat{\beta}^t(1, \lambda) - \beta^t - X^T \frac{y - X\beta^t}{1 - \|\beta^t\|_0/n} \|^2_2 = 0.
\]

Next from Equation (G.1) we have the following,

\[
X^T z^t - X^T \frac{y - X\beta^t}{1 - \|\beta^t\|_0/n} = X^T \|\beta^t\|_0/(n(-z^t + z^{t-1}n/(p\delta)) \frac{1}{1 - \|\beta^t\|_0/n}
\]

Using the result of Theorem G.1 part (ii) and \(n/p \to \delta\), we can obtain

\[
\lim_{t \to \infty} \lim_{p \to \infty} \frac{1}{p} \| X^T z^t - X^T \frac{y - X\beta^t}{1 - \|\beta^t\|_0/n} \|^2_2 = 0, \text{ a.s.}
\]

The results (G.3) and (G.4) together imply that

\[
\lim_{t \to \infty} \lim_{p \to \infty} \frac{1}{p} \| \hat{\beta}^t(1, \lambda) - \beta^t - X^T z^t \|^2_2 = 0, \text{ a.s.}
\]
According to Theorem G.1 part (iv), for any bounded Lipschitz function $L(x) : \mathbb{R} \to \mathbb{R}$:

$$\lim_{p \to \infty} \frac{1}{p} \sum_{i=1}^{p} L(\beta^i + (X^T z)^i - \beta_i) = \mathbb{E}L(\tau_i Z).$$

Putting the last two results together, it is not hard to confirm

$$\lim_{p \to \infty} \frac{1}{p} \sum_{i=1}^{p} L(\hat{\beta}^i(1, \lambda) - \beta_i) = \mathbb{E}L(\tau Z).$$

That means the empirical distribution of $\hat{\beta}(1, \lambda) - \beta$ converges to the distribution of $\tau Z$ and finishes the proof.

G.1.2. Proof of Theorem 4.1 for $q > 1$. The basic proof idea for $q > 1$ is the same as for $q = 1$. However, since the debiasing estimator for $q > 1$ takes a different form, we need to take care of some subtle details. Recall the definition of $f(v, w)$ and $\hat{\gamma}_\lambda$ in (4.2). We first obtain the bounds:

$$\frac{1}{p} \left\| \frac{\hat{\beta}(q, \lambda) - \beta^i - X^T (y - X\beta^i)}{1 - f(\beta^i, \alpha \tau^2 - q)/n} \right\|_2^2 \leq \frac{2}{p} \left\| \frac{\hat{\beta}(q, \lambda) - \beta^i}{1 - f(\hat{\beta}(q, \lambda), \hat{\gamma}_\lambda)/n} \right\|_2^2 +$$

$$\frac{8}{p(1 - f(\eta_q(B + \tau Z; \alpha \tau^2 - q), \alpha \tau^2 - q)/\delta)} \left\| X^T (\hat{\beta}(q, \lambda) - \beta^i) \right\|_2^2 +$$

$$\frac{8}{p} \left\| X^T (y - X\hat{\beta}(q, \lambda)) \right\|_2^2 \left( \frac{1}{1 - f(\beta^i, \alpha \tau^2 - q)/n} - \frac{1}{1 - f(\eta_q(B + \tau Z; \alpha \tau^2 - q), \alpha \tau^2 - q)/\delta} \right)^2.$$

As in the proof of $q = 1$, we show that $Q_i(i = 1, 2, 3, 4)$ vanishes asymptotically. For that purpose we first note that Theorem G.1 (except part (iii)) holds for $q > 1$ as well. Hence the same argument for $q = 1$ gives us $Q_1, Q_2 \to 0$. Regarding $Q_3$, by the facts that the empirical distribution of $\hat{\beta}(q, \lambda)$ converges weakly to the distribution of $\eta_q(B + \tau Z; \alpha \tau^2 - q)$ and $\frac{1}{1 + \alpha \tau^2 - q(\tau - 1)|x|^2}$ is a bounded continuous function of $x$, we have

$$\lim_{p \to \infty} \frac{1}{p} f(\hat{\beta}(q, \lambda), \alpha \tau^2 - q) = f(\eta_q(B + \tau Z; \alpha \tau^2 - q), \alpha \tau^2 - q), \text{ a.s.}$$
Moreover, according to Lemma G.1 we obtain as $p \to \infty$,

$$\frac{1}{p} |f(\hat{\beta}(q, \lambda), \alpha \tau^{2-q}) - f(\hat{\beta}(q, \lambda), \hat{\gamma}_\lambda)| \leq \alpha^{-1} \tau^{q-2} |\hat{\gamma}_\lambda - \alpha \tau^{2-q}| \overset{a.s.}{\to} 0.$$ 

The last two results together lead to $Q_3 \overset{a.s.}{\to} 0$. For $Q_4$, it is not hard to apply Theorem G.1 part (v) and the fact $\tau_p \to \tau$ to show

$$\lim_{t \to \infty} \lim_{p \to \infty} \frac{1}{p} f(\beta^t, \alpha \tau^{2-q}) = f(\eta_\theta(B + \tau Z; \alpha \tau^{2-q}), \alpha \tau^{2-q}), \ a.s.$$ 

which implies $Q_4 \overset{a.s.}{\to} 0$. The rest of the proof is almost the same as the one for $q = 1$. We hence do not repeat the arguments.

**Lemma G.1.** For $\hat{\gamma}_\lambda$ defined in (4.3), as $p \to \infty$

$$\hat{\gamma}_\lambda \overset{a.s.}{\to} \alpha \tau^{2-q}.$$ 

**Proof.** Denote $a(\gamma) = \delta(1 - \frac{\lambda}{\gamma})$, $\hat{b}(\gamma) = \text{Ave} \left[ \frac{1}{1 + \gamma q (q - 1)} |\eta_\theta(B + \tau Z; \alpha \tau^{2-q})|^{q-2} \right]$, $b(\gamma) = \mathbb{E} \left[ \frac{1}{1 + \gamma q (q - 1)} |\eta_\theta(B + \tau Z; \alpha \tau^{2-q})|^{q-2} \right]$, and $\gamma_\lambda = \alpha \tau^{2-q}$. Clearly from (4.3) and (2.2), $\hat{\gamma}_\lambda$ is the unique solution of $a(\gamma) = \hat{b}(\gamma)$, and $\gamma_\lambda$ is the unique solution of $a(\gamma) = b(\gamma)$.

As a simple corollary of Theorem 2.1, almost surely the empirical distribution of $\hat{\beta}(q, \lambda)$ converges weakly to the distribution of $\eta_\theta(B + \tau Z; \alpha \tau^{2-q})$. As a result, for $h(x) = \frac{1}{1 + \gamma q (q - 1) |x|^{q-2}}$ which is bounded and continuous on $\mathbb{R}$, we have almost surely

$$\hat{b}(\gamma) \to b(\gamma), \ a.s. \to \infty.$$ 

The above convergence is pointwise in $\gamma$. In fact we can obtain a stronger result. That is, there is a $\Omega_0 \subset \Omega$ with $\mathbb{P}(\Omega_0) = 1$, such that for any $\omega \in \Omega_0$, $\hat{b}(\gamma, \omega) \to b(\gamma)$ for all $\gamma \geq 0$. (we can first construct $\Omega_0$ for $\gamma \in \mathbb{Q}$, then extend to $\mathbb{R}$, by continuity and monotonicity of $a(\gamma)$, $\hat{b}(\gamma)$ and $b(\gamma)$.)

Now for any $\omega \in \Omega_0$, for any $\epsilon > 0$, consider the closed neighborhood $[\gamma_\lambda - \epsilon, \gamma_\lambda + \epsilon]$. Let $\eta_\epsilon = \min \{ b(\gamma_\lambda - \epsilon) - a(\gamma_\lambda - \epsilon), a(\gamma_\lambda + \epsilon) - b(\gamma_\lambda + \epsilon) \}$. Monotonicity of $a(\gamma)$, $b(\gamma)$ and uniqueness of the solution $\gamma_\lambda$ guarantee $\eta_\epsilon > 0$. On the two end points $\gamma_\lambda - \epsilon$ and $\gamma_\lambda + \epsilon$, we know as $p \to \infty$,

$$\hat{b}(\gamma_\lambda - \epsilon, \omega) \to b(\gamma_\lambda - \epsilon), \ \hat{b}(\gamma_\lambda + \epsilon, \omega) \to b(\gamma_\lambda + \epsilon).$$

Thus there exists $N_\epsilon(\omega)$, for any $p > N_\epsilon(\omega)$,

$$|\hat{b}(\gamma_\lambda - \epsilon, \omega) - b(\gamma_\lambda - \epsilon)| < \frac{\eta_\epsilon}{2}, \ |\hat{b}(\gamma_\lambda + \epsilon, \omega) - b(\gamma_\lambda + \epsilon)| < \frac{\eta_\epsilon}{2}.$$
By noticing the distance between \(a(\gamma)\) and \(b(\gamma)\) on the two end-points, we have \(b(\gamma_{\lambda} - \epsilon, \omega) - a(\gamma_{\lambda} - \epsilon) > \frac{\epsilon}{2}\) and \(a(\gamma_{\lambda} + \epsilon) - b(\gamma_{\lambda} + \epsilon, \omega) > \frac{\epsilon}{2}\). The monotonicity of the function \(\hat{b}(\gamma, \omega)\) determines that \(\hat{\gamma}_\lambda(\omega) \in (\gamma_{\lambda} - \epsilon, \gamma_{\lambda} + \epsilon)\), i.e., \(|\hat{\gamma}_\lambda(\omega) - \gamma_{\lambda}| < \epsilon\). As a conclusion, we have \(\hat{\gamma}_\lambda \overset{a.s.}{\longrightarrow} \gamma_{\lambda}\).

G.2. Proof of Theorems 4.2 and 4.3. We only prove the case \(q = 1\). The proof for \(q > 1\) is almost the same. In the proof of Theorem 4.1, we have showed

\[
\lim_{t \to \infty} \lim_{p \to \infty} \| \hat{\beta}^t(1, \lambda) - (\beta^t + X^T z^t) \|_2 = 0, \ a.s.
\]

Combining this result with Theorem G.1 part (iv), we know for any bounded Lipschitz function \(\psi : \mathbb{R}^2 \to \mathbb{R}\)

\[
\lim_{p \to \infty} \frac{1}{p} \sum_{i=1}^{p} \psi(\hat{\beta}^t_i(1, \lambda), \beta_i) = E\psi(B + \tau Z, B).
\]

Hence the empirical distribution of \((\hat{\beta}^1(1, \lambda), \beta)\) converges weakly to the distribution of \((B + \tau Z, B)\). As a result, we can follow the same calculations as we did in the proof of Lemma 3.2 to obtain

\[
\text{AFDP}^\dagger(1, \lambda, s) = \frac{(1 - \epsilon)P(\tau|Z| > s)}{(1 - \epsilon)P(\tau|Z| > s) + \epsilon P(|G + \tau Z| > s)}
\]

\[
\text{ATPP}^\dagger(1, \lambda, s) = P(|G + \tau Z| > s)
\]

(G.5)

Notice that \(\tau > 0\) and \(G + \tau Z\) is continuous. Thus as we vary \(s\), \(\text{ATPP}^\dagger(1, \lambda, s)\) can reach all values in \([0, 1]\). Furthermore, by comparing the formula above with that in Lemma 3.2 (when \(q = 1\)), it is clear that \(\text{ATPP}^\dagger(1, \lambda, s) = \text{ATPP}(1, \lambda, \tilde{s})\) will imply \(\text{AFDP}^\dagger(1, \lambda, s) = \text{AFDP}(1, \lambda, \tilde{s})\). The proof for the second part is similar to that of Theorem 3.1. We do not repeat the argument here.

G.3. Proof of Lemma 4.1. Notice that SIS is thresholding \(X^T y\), which is also the initialization of AMP in (G.1). By setting \(t = 0\) in Theorem G.1 (iv), we obtain

\[
\lim_{p \to \infty} \sum_{i=1}^{p} \psi((X^T y)_i, \beta_i) = E\psi(B + \tau_0 Z, B)
\]

(G.6)

where \(\tau_0^2 = \sigma^2 + \frac{E\beta^2}{\lambda} > 0\). This implies that almost surely the empirical distribution of \(\{((X^T y)_i, \beta_i)\}_{i=1}^{p}\) converges weakly to \((B + \tau_0 Z, B)\). Following
the same argument as in the proof of Lemma 3.2, we have

$$AFDP_{sis}(s) = \frac{(1 - \epsilon)\mathbb{P}(\tau_0 | Z > s)}{(1 - \epsilon)\mathbb{P}(\tau_0 | Z > s) + \epsilon \mathbb{P}(|G + \tau_0 Z| > s)}$$

$$ATPP_{sis}(s) = \mathbb{P}(|G + \tau_0 Z| > s)$$

On the other hand, based on Equation (G.5), we obtain

$$AFDP^{\dagger}(q, \lambda^*_q, s) = \frac{(1 - \epsilon)\mathbb{P}(\tau^*_s | Z > s)}{(1 - \epsilon)\mathbb{P}(\tau^*_s | Z > s) + \epsilon \mathbb{P}(|G + \tau^*_s Z| > s)}$$

$$ATPP^{\dagger}(q, \lambda^*_q, s) = \mathbb{P}(|G + \tau^*_s Z| > s)$$

Note that

$$\tau^2_s = \frac{1}{\delta} \min_{\alpha} \mathbb{E}(\eta_q(B + \tau_s Z; \alpha \tau^2_s - q) - B)^2$$

$$\leq \sigma^2 + \frac{1}{\delta} \lim_{\alpha \to \infty} \mathbb{E}(\eta_q(B + \tau_s Z; \alpha \tau^2_s - q) - B)^2 = \tau^2_0.$$