The Limiting Poisson Law of Massive MIMO Detection with Box Relaxation

Hong Hu and Yue M. Lu

Abstract

Estimating a binary vector from noisy linear measurements is a prototypical problem for MIMO systems. A popular algorithm, called the box-relaxation decoder, estimates the target signal by solving a least squares problem with convex constraints. This paper shows that the performance of the algorithm, measured by the number of incorrectly-decoded bits, has a limiting Poisson law. This occurs when the sampling ratio and noise variance, two key parameters of the problem, follow certain scalings as the system dimension grows. Moreover, at a well-defined threshold, the probability of perfect recovery is shown to undergo a phase transition that can be characterized by the Gumbel distribution. Numerical simulations corroborate these theoretical predictions, showing that they match the actual performance of the algorithm even in moderate system dimensions.

I. Introduction

A. Motivations

Consider the problem of estimating a binary vector $\boldsymbol{\beta} \in \{-1, 1\}^p$ from noisy linear measurements in the form of

$$\boldsymbol{y} = \boldsymbol{A}\boldsymbol{\beta} + \boldsymbol{w}. \tag{1}$$

Here, $\boldsymbol{A} \in \mathbb{R}^{n\times p}$ is a known sensing matrix and $\boldsymbol{w} \sim \mathcal{N}(\boldsymbol{0}, \sigma_p^2 \boldsymbol{I}_n)$ denotes an unknown noise vector. This is a prototypical model for multi-user detections in MIMO communication systems [1], [2]. It also arises in other applications such as compressed sensing [3], source separation [4], and image processing [5].

H. Hu and Y. M. Lu are with the John A. Paulson School of Engineering and Applied Sciences, Harvard University, Cambridge, MA 02138, USA (e-mails: honghu@g.harvard.edu and yuelu@seas.harvard.edu). This work was supported by the Harvard FAS Dean's Fund for Promising Scholarship, and by the US National Science Foundation under grants CCF-1718698 and CCF-1910410.

Various algorithms have been proposed to solve (1). Examples include sphere decoding [6], zero-forcing [7], approximate message passing [8], Markov chain Monte Carlo methods [9], and semidefinite programming [10]. Among them, a convex-optimization based method, known as the box-relaxation decoder [11]–[13], is popular in practice due to its simplicity and efficiency. The method consists of merely two steps: (1) solve a box-constrained least squares problem

$$\boldsymbol{x}^* = \underset{\boldsymbol{x}\in[-1,1]^p}{\arg\min}\ \frac{1}{2}\|\boldsymbol{y}-\boldsymbol{A}\boldsymbol{x}\|^2, \tag{2}$$

and (2) obtain an estimate of $\boldsymbol{\beta}$ by taking the sign of $\boldsymbol{x}^*$, *i.e.*, $\widehat{\boldsymbol{\beta}} = \mathrm{sign}(\boldsymbol{x}^*)$.

The performance of this algorithm can be measured by the bit error rate (BER):

$$\mathrm{BER} = \frac{1}{p}\sum_{i=1}^{p} \mathbb{1}_{\{\widehat{\beta}_i \neq \beta_i\}}, \tag{3}$$

where $\mathbb{1}_{\{\cdot\}}$ denotes the indicator function. The achievable BER depends on two key parameters: the noise variance σ_p^2, and the sampling ratio $\delta_p \overset{\text{def}}{=} n/p$.

Under the assumption that the sensing matrix $\boldsymbol{A}$ has i.i.d. normal entries, the authors of [12], [13] analyzed the asymptotic BER achieved by the box-relaxation decoder. They show that, as $n, p \to \infty$ with $\delta_p \to \delta \in (\frac{1}{2}, \infty)$ and $\sigma_p^2 \equiv \sigma^2 > 0$, the BER converges in probability to a deterministic limit, *i.e.*,

$$\mathrm{BER} \xrightarrow{\mathcal{P}} \mathcal{E}(\delta, \sigma^2) \in \left(0, \tfrac{1}{2}\right). \tag{4}$$

This means that for any $\sigma^2 > 0$ and $\delta > \frac{1}{2}$, the algorithm can asymptotically achieve a *weak recovery* of $\boldsymbol{\beta}$: it is better than random guess, but $\widehat{\boldsymbol{\beta}}$ always contains a nonzero fraction of errors. Moreover, one can show that

$$\lim_{\delta\to\infty} \mathcal{E}(\delta, \sigma^2) = \lim_{\sigma^2\to 0} \mathcal{E}(\delta, \sigma^2) = 0. \tag{5}$$

The expressions in (5), together with (4), suggest that the asymptotic BER can be made arbitrarily small if we increase the number of measurements or reduce the noise variance. This then raises a tantalizing question: is there a regime of (δ_p, σ_p^2) such that the box-relaxation decoder can *perfectly* recover the target signal? Existing results in [12], [13] cannot answer this question, for two reasons. First, $\mathrm{BER} \xrightarrow{p\to\infty} 0$ only guarantees that the *number of error bits*

$$N_e \overset{\text{def}}{=} \sum_{i=1}^{p} \mathbb{1}_{\{\widehat{\beta}_i \neq \beta_i\}}, \tag{6}$$

is sublinear in p, but it contains no information about the actual distribution of N_e, including whether $N_e = 0$. The second issue is subtle but important. It has to do with the specific order

with which the limits are taken in (4) and (5). There, we first send the dimension $p \to \infty$ *before* letting $\delta_p \to \infty$ or $\sigma_p^2 \to 0$. In practice, p is large but always finite, and thus the speed with which $\delta_p \to \infty$ and $\sigma_p^2 \to 0$ [*e.g.*, $\sigma_p^2 = \mathcal{O}(1/p)$ vs. $\sigma_p^2 = \mathcal{O}(1/\log p)$] makes all the difference.

The goal of this paper is to present a precise asymptotic characterization of the probability distribution of N_e. We show that, in certain scaling regimes of (δ_p, σ_p^2), the distribution of N_e converges to a Poisson law. Moreover, we derive conditions under which the exact recovery of $\boldsymbol{\beta}$ is possible and provide an asymptotic formula for $\mathbb{P}(N_e = 0)$ in the form of a Gumbel distribution.

B. Main Results

We make the following assumptions throughout the paper.

(A.1) The elements of $\boldsymbol{A}$ are drawn from the i.i.d. Gaussian distribution: $A_{ij} \overset{i.i.d.}{\sim} \mathcal{N}(0, \frac{1}{p})$.

(A.2) $\boldsymbol{\beta} = -\mathbf{1}_p$, where $\mathbf{1}_p$ denotes the all-ones vector.

(A.3) The noise is Gaussian: $\boldsymbol{w} \sim \mathcal{N}(\mathbf{0}, \sigma_p^2 \boldsymbol{I}_n)$.

(A.4) $\liminf_{p\to\infty} \delta_p > 1/2$ and $\limsup_{p\to\infty} \delta_p / \log p < \infty$.

(A.5) $\liminf_{p\to\infty} \sigma_p^2 \log^2 p > 0$ and $\limsup_{p\to\infty} \sigma_p^2 < \infty$.

In (A.2), we assume that each coordinate of true signal is -1 to simplify our derivations. All the results still hold for arbitrary $\boldsymbol{\beta}$, due to the rotational symmetry of $\boldsymbol{A}$. In (A.4), the requirement that $\liminf_{p\to\infty} \delta_p > 1/2$ is related to the fundamental limits of convex relaxation for structural signal reconstruction. In [14], it is shown that, if $\limsup_{p\to\infty} \delta_p \le \frac{1}{2}$, the box-relaxation decoder cannot successfully recover $\boldsymbol{\beta}$ even in the noiseless case. In (A.5), we essentially require $\sigma_p^2 > c/\log^2 p$ for some $c > 0$. This restriction is due to the limitations of our current proof techniques. We expect that many of our results still hold without this restriction.

To state our main results, we first need to introduce the following potential function:

$$F_p(\tau; \sigma_p^2, \delta_p) = \frac{\tau}{2}\left(\delta_p - \frac{1}{2}\right) + \frac{\sigma_p^2}{2\tau} + \frac{\tau}{2}\int_{\frac{2}{\tau}}^{\infty} \left(x - \frac{2}{\tau}\right)^2 \Phi(dx), \tag{7}$$

where Φ is the CDF of the standard normal distribution. One can verify that F_p is a strictly convex function of $\tau \in (0, \infty)$. (See Appendix B for details.) Thus, one can uniquely define

$$f_p \overset{\text{def}}{=} \min_{\tau>0} F_p(\tau; \sigma_p^2, \delta_p) \quad \text{and} \quad \tau_p \overset{\text{def}}{=} \arg\min_{\tau>0} F_p(\tau; \sigma_p^2, \delta_p). \tag{8}$$

Another quantity that will be crucial in our analysis is

$$\lambda_p \overset{\text{def}}{=} p\Phi(-\tfrac{1}{\tau_p}). \tag{9}$$

Theorem 1: Under (A.1)-(A.5), and if $\limsup_{p\to\infty} \frac{\lambda_p}{\sqrt{\log p}} < \infty$, then

$$d_{\text{TV}}(N_e, \mathscr{P}(\lambda_p)) \leq \frac{\text{polylog}\, p}{p^{1/5}}, \tag{10}$$

where d_{TV} is the total variation (TV) distance and $\mathscr{P}(\lambda)$ denotes a Poisson distribution with parameter λ.

Remark 1: The theorem, whose proof can be found in Section II-D, characterizes the asymptotic distribution of N_e under certain scaling regimes of (δ_p, σ_p^2). It shows that the law of N_e converges to that of a Poisson random variable with parameter λ_p, if λ_p grows no faster than $\sqrt{\log p}$. This requirement on λ_p is not satisfied in the setting studied in [12] where both δ_p and σ_p^2 are kept as fixed constants and consequently $\lambda_p = \mathcal{O}(p)$. In that case, one can expect that $\sqrt{p}[\frac{N_e}{p} - \Phi(-\frac{1}{\tau_p})]$ converges to a Gaussian distribution.

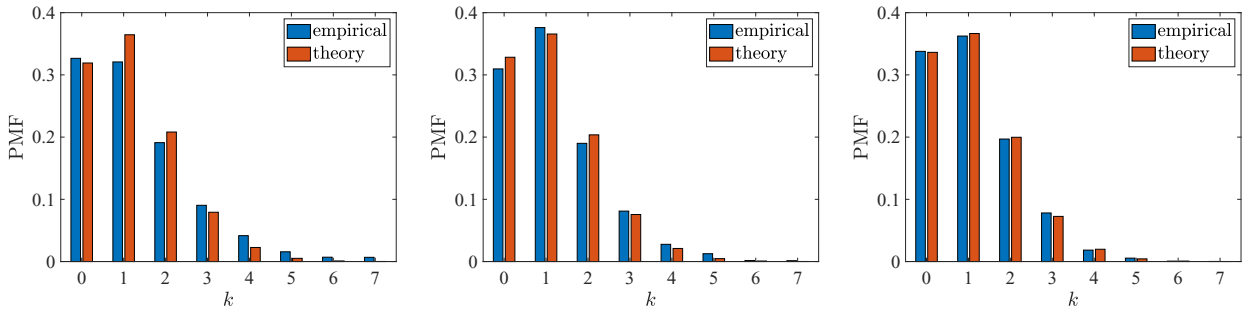
The fact that N_e can have a limiting Poisson law is not surprising. Recall from its definition in (6) that N_e is a sum of p Bernoulli random variables $\{\mathbb{1}_{\{\widehat{\beta}_i \neq \beta_i\}}\}$. Moreover, one can show that $\mathbb{P}(\widehat{\beta}_i \neq \beta_i) \approx \Phi(-\frac{1}{\tau_p})$ and that these Bernoulli random variables are close to being independent. Consequently, the law of N_e is approximately a Binomial distribution $\mathcal{B}(p, \Phi(-\frac{1}{\tau_p}))$ with an expected value equal to λ_p. As $p \to \infty$ with $\lambda_p = \mathcal{O}(\sqrt{\log p})$, it is well-known that the Binomial distribution converges to a Poisson distribution (*i.e.*, the "law of small numbers"). The technical contribution of this paper is to make the above arguments precise and rigorous. The main tool we use is the leave-one-out approach (see, *e.g.*, [15]), also known as the cavity method in statistical physics [16], [17]. It allows us to carry out a detailed probabilistic analysis of the random optimization problem in (2).

In our proof of Theorem 1, we did not attempt to optimize the rate of convergence shown on the right-hand side of (10). The actual rate is likely to be faster. In Figure 1, we compare the empirical distribution of N_e, obtained after averaging over 10^4 independent trials, against the limiting Poisson distribution for three different problem dimensions. We can see that, even at a moderate dimension of $p = 200$, the Poisson approximation is already accurate.

The characterization given in Theorem 1 allows us to study the conditions under which the box-relaxation decoder can perfectly recover the target signal. Let $P_{\text{correct}} \overset{\text{def}}{=} \mathbb{P}(N_e = 0)$ denotes the probability of perfect recovery. We can show that a phase transition of P_{correct} emerges when the following quantity

$$\alpha_p \overset{\text{def}}{=} \frac{\delta_p - 1/2}{2\sigma_p^2 \log p} \tag{11}$$

is near 1.



(a) $p = 200$ (b) $p = 1000$ (c) $p = 10000$

Figure 1: Comparison of the empirical distribution of N_e and the limiting Poisson distribution, over three different problem dimensions. In the experiments, we set $\sigma_p^2 = 1$ and choose δ_p so that $\lambda_p \approx 1.1$ for all three values of p.

Proposition 1: Under (A.1)-(A.5), and if $\lim_{p\to\infty} \alpha_p = \alpha^*$, then

$$\lim_{p\to\infty} P_{\text{correct}} = \begin{cases} 1, & \text{if } \alpha^* > 1, \\ 0, & \text{if } \alpha^* < 1. \end{cases} \tag{12}$$

If $\alpha^* = 1$, a more refined characterization is available. Specifically, assume that

$$\alpha_p(x) = 1 - \frac{\log\log p}{2\log p} + \frac{x - \log\sqrt{4\pi}}{\log p}, \tag{13}$$

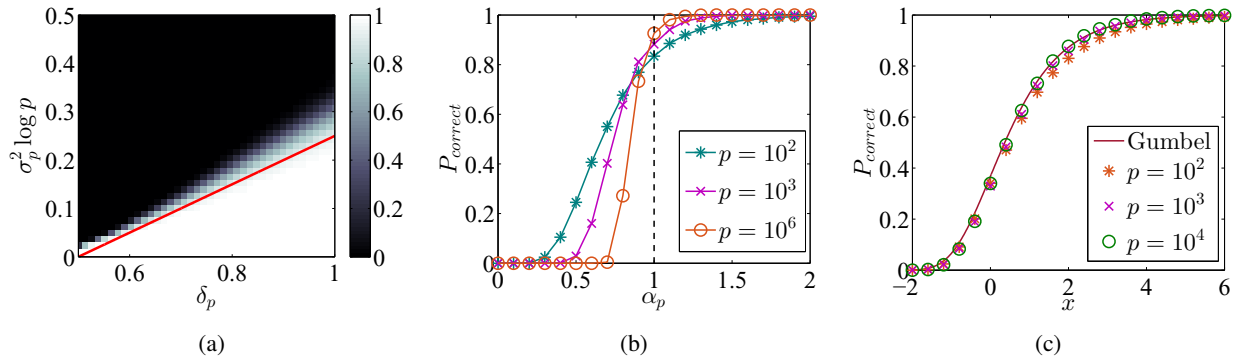
for some constant $x \in \mathbb{R}$ (and thus $\alpha_p(x) \overset{p\to\infty}{\longrightarrow} 1$), then

$$\lim_{p\to\infty} P_{\text{correct}} = e^{-e^{-x}}, \tag{14}$$

where the right-hand side is the CDF of the Gumbel distribution.

Remark 2: The above proposition, proved in Section II-E, characterizes the scaling regimes of (δ_p, σ_p^2) over which perfect recovery is achievable. The possible scalings are also flexible. For example, if we keep the sampling ratio δ_p at a fixed value $\delta > 1/2$, it then follows from (11) and (12) that $\sigma_p^2 = \frac{\delta - 1/2}{2\log p}$ is the critical noise variance threshold for perfect recovery to happen. Alternatively, if we fix the noise variance $\sigma_p^2 \equiv \sigma^2$, then the critical threshold for the sampling ratio is $\delta_p = 1/2 + 2\sigma^2 \log p$.

To illustrate Proposition 1, we show some results from numerical experiments. In Figure 2a, we plot the phase diagram of the empirical values of P_{correct} under different choices of (δ_p, σ_p^2), as well as the theoretical phase transition boundary separating the regimes of perfect/nonperfect recovery. In Figure 2b, we plot P_{correct} as a function of α_p (by fixing $\delta_p = 1$ and varying σ_p^2).



(a) (b) (c)

Figure 2: (a) Phase diagram of the box-relaxation decoder. Each pixel represents the value of P_{correct} under a specific (δ_p, σ_p^2). The red curve is the theoretical transition boundary: $\sigma_p^2 \log p = \frac{\delta_p - 1/2}{2}$. (b) Phase transition of P_{correct} with respect to α_p. The dashed line represents the theoretical threshold. (c) Near the phase transition boundary, P_{correct} is well-approximated by the Gumbel distribution. In all three experiments, P_{correct} is estimated by averaging over 10^4 independent trials. In (b) and (c), we fix $\delta_p = 1$ and vary α_p and x by changing σ_p^2.

A transition indeed takes place near $\alpha_p = 1$, and the transition becomes sharper as we increase the problem dimension p. When p is not very large, a more accurate approximation of P_{correct} is given by the Gumbel distribution. This is illustrated in Figure 2c, where we zoom in the region near the phase transition and compare the empirical success probability against the theoretical prediction given in (14).

C. Related Work

The precise analysis of high-dimensional signal estimation has already been the subject of a vast literature. Underpinning these rich results are several powerful techniques developed over the years, including the nonrigorous replica method from statistical physics [18]–[20], approximate message passing (AMP) [21]–[23], the cavity method [16], [17] and leave-one-out analysis [15], Gaussian min-max theorem (GMT) [24], [25], as well as the geometric framework based on Gaussian width [14] and statistical dimensions [26].

The box-constrained least square problem in (2) has been previously analyzed in [12], [13] using GMT techniques. Analysis of similar problems can also be carried out by AMP [8]. However, these existing studies consider the setting where both the sampling ratio δ_p and the noise variance σ_p^2 are kept as constants as $p \to \infty$. Under such scalings, one can establish that

the empirical measure of $\boldsymbol{x}^*$, defined as $\widehat{\mu}(\boldsymbol{x}^*) \stackrel{\text{def}}{=} \frac{1}{p}\sum_{i=1}^{p}\delta_{x_i^*}$, converges to some deterministic limiting measure. However, the convergence of the empirical measure is insufficient for our purpose: flipping the signs of $o(p)$ entries of $\boldsymbol{x}^*$ will completely change the number of error bits N_e, but it has no effect on the limiting empirical measure. In view of this, we choose to use the leave-one-out approach, which allows us to construct a surrogate of $\boldsymbol{x}^*$, denoted by $\widetilde{\boldsymbol{x}}$, in our analysis. We show that $\|\boldsymbol{x}^* - \widetilde{\boldsymbol{x}}\|_\infty \to 0$ but the statistical properties of $\widetilde{\boldsymbol{x}}$ are much easier to obtain. We will elaborate on this point in Sec. II.

Our work considers settings where (δ_p, σ_p^2) can scale with the problem dimension p. Similar settings with flexible scalings have been explored in other contexts, including, *e.g.*, sparse linear regression [27]–[29], spiked matrix estimation [30], and low-rank matrix recovery [31]. These studies established the precise conditions under which perfect recovery in these problems is achievable. In our work, we go one step further by establishing the asymptotic distribution of the number of error bits N_e.

II. Roadmap of Analysis

This section provides a general roadmap to our proof of Theorem 1, which is given in Section II-D. To emphasize readability, we only highlight the main ideas and key intermediate results here, leaving heavier technical details to the subsequent sections and to the appendix.

A. An Equivalent Scalar Problem

To analyze N_e, we need to understand the statistical properties of $\boldsymbol{x}^*$, *i.e.*, the optimal solution of (2). A basic challenge lies in the fact $\boldsymbol{x}^*$ is a high-dimensional vector with no closed-form expressions. The key idea behind the cavity approach [16], [17] or the leave-one-out analysis [15] is to circumvent this issue by focusing instead on a single coordinate of $\boldsymbol{x}^*$. Specifically, to study the ith coordinate x_i, we can first rewrite the original problem (2) as

$$\begin{aligned}
&\arg\min_{x_i\in[-1,1]} \quad \min_{\boldsymbol{x}_{\setminus i}\in[-1,1]^{p-1}} \frac{1}{2}\|\boldsymbol{A}_{\setminus i}\boldsymbol{x}_{\setminus i} + \boldsymbol{a}_i(x_i-\beta_i) - \boldsymbol{y}_{\setminus i}\|^2 \\
=&\arg\min_{x_i\in[-1,1]} \quad \min_{\boldsymbol{x}_{\setminus i}\in[-1,1]^{p-1}} \quad \max_{\boldsymbol{u}} \ \boldsymbol{u}^{\mathsf{T}}[\boldsymbol{A}_{\setminus i}\boldsymbol{x}_{\setminus i} + \boldsymbol{a}_i(x_i-\beta_i) - \boldsymbol{y}_{\setminus i}] - \frac{1}{2}\|\boldsymbol{u}\|^2 \qquad (15)\\
=&\arg\min_{x_i\in[-1,1]} \ \max_{\boldsymbol{u}} \ \boldsymbol{a}_i^{\mathsf{T}}\boldsymbol{u}(x_i-\beta_i) - L_i(\boldsymbol{u}), \qquad (16)
\end{aligned}$$

where $\boldsymbol{x}_{\setminus i}$ is the vector formed by removing x_i (and $\boldsymbol{\beta}_{\setminus i}$ is defined in the same way), $\boldsymbol{a}_i$ is the ith column of $\boldsymbol{A}$, $\boldsymbol{A}_{\setminus i}$ denotes the matrix formed by removing $\boldsymbol{a}_i$ from $\boldsymbol{A}$, $\boldsymbol{y}_{\setminus i} = \boldsymbol{A}_{\setminus i}\boldsymbol{\beta}_{\setminus i} + \boldsymbol{w}$, and

$$L_i(\boldsymbol{u}) = \|\boldsymbol{A}_{\setminus i}^{\mathsf{T}}\boldsymbol{u}\|_1 + \boldsymbol{u}^{\mathsf{T}}\boldsymbol{y}_{\setminus i} + \frac{1}{2}\|\boldsymbol{u}\|^2. \tag{17}$$

In reaching (16), we have also used Sion's minimax theorem [32] to swap the inner minimization and maximization in (15).

Let $\boldsymbol{u}^*_{\setminus i} = \arg\min_{\boldsymbol{u}} L_i(\boldsymbol{u})$ and define a function

$$g_{p,i}(v) \stackrel{\text{def}}{=} \max_{\boldsymbol{u}} \ (\boldsymbol{u} - \boldsymbol{u}^*_{\setminus i})^{\mathsf{T}}\boldsymbol{a}_i v - [L_i(\boldsymbol{u}) - L_i(\boldsymbol{u}^*_{\setminus i})]. \tag{18}$$

We can then check that the optimization problem (16) has the same solution as

$$\arg\min_{x_i \in [-1,1]} \ g_{p,i}(x_i - \beta_i) + \boldsymbol{a}_i^{\mathsf{T}}\boldsymbol{u}^*_{\setminus i}(x_i - \beta_i). \tag{19}$$

Thus, starting from the original problem (2) and after optimizing over all the "nuisance" variables $\boldsymbol{x}_{\setminus i}$, we have reached in (19), an equivalent scalar optimization problem over x_i.

To nonspecialists, the reformulations leading to (19) might look slightly mysterious, but there are several good reasons for doing so. First, note that (19) is obtained by subtracting $-L_i(\boldsymbol{u}^*_{\setminus i})$ from (16). This manipulation does not change the minimizer of (16), but it sets the magnitude of (19) to be $\mathcal{O}(1)$, which facilitates our later analysis. Second, we explicitly pull out $\boldsymbol{a}_i^{\mathsf{T}}\boldsymbol{u}^*_{\setminus i}$ in (19), since its distribution is much easier to characterize than $\boldsymbol{a}_i^{\mathsf{T}}\boldsymbol{u}^*$ in (16), due to the independence between $\boldsymbol{a}_i$ and $\boldsymbol{u}^*_{\setminus i}$. This is in fact a major benefit of the leave-one-out analysis. Third, as we will show next $g_{p,i}(x_i - \beta_i)$, which is a random one-dimensional function $g_{p,i}(v)$ evaluated at $v = x_i - \beta_i$, has a particularly simple limiting form as $p \to \infty$.

B. A Limiting Quadratic Function

The following proposition, whose proof is given in Section III-A, shows that $g_i(v)$ uniformly converges to a simple quadratic function.

Proposition 2: Under (A.1)-(A.5), there exists $c > 0$ such that for any $i \in [p]$ and $\varepsilon > 0$,

$$\mathbb{P}\Big\{ \sup_{v \in [-2,2]} \Big| g_{p,i}(v) - \frac{1}{2}A_p v^2 \Big| > \varepsilon \Big\} \le \frac{c\delta_p}{\varepsilon} e^{-c^{-1}p \min\left\{\frac{\varepsilon^2}{\delta_p}, \varepsilon\right\}}, \tag{20}$$

where

$$A_p = \frac{\mathbb{E}\boldsymbol{w}^{\mathsf{T}}(\boldsymbol{y} - \boldsymbol{A}\boldsymbol{x}^*)}{\sigma_p^2 p}. \tag{21}$$

Moreover, for $\gamma > 2$ and all large enough p, $|A_p - A_p^*| < cp^{-1/(2\gamma)}$, where

$$A_p^* \stackrel{\text{def}}{=} f_p/\tau_p, \tag{22}$$

and f_p and τ_p are the quantities defined in (8).

There is a simple intuitive explanation for why $g_{p,i}(v)$ is approximately a quadratic function. Recall that $\boldsymbol{u}_{\setminus i}^*$ is the minimizer of $L_i(\boldsymbol{u})$. Thus, in a local neighborhood near $\boldsymbol{u}_{\setminus i}^*$, we can approximate $L_i(\boldsymbol{u})$ by a second-order Taylor expansion: $L_i(\boldsymbol{u}) \approx L_i(\boldsymbol{u}_{\setminus i}^*) + \frac{\boldsymbol{\delta}^\intercal \boldsymbol{H}_{\setminus i} \boldsymbol{\delta}}{2}$, where $\boldsymbol{\delta} = \boldsymbol{u} - \boldsymbol{u}_{\setminus i}^*$ and $\boldsymbol{H}_i$ corresponds to the Hessian of $L_i(\boldsymbol{u})$ at $\boldsymbol{u}_{\setminus i}^*$. Substituting this approximation into (18), we can immediately obtain that $g_{p,i}(v) \approx \frac{\boldsymbol{a}_i^\intercal \boldsymbol{H}_i^{-1} \boldsymbol{a}_i}{2} v^2$. Since $\boldsymbol{a}_i \sim \mathcal{N}(\boldsymbol{0}, \frac{\boldsymbol{I}_n}{p})$ and it is independent of $\boldsymbol{H}_i$ due to the leave-one-out construction, we can expect $\boldsymbol{a}_i^\intercal \boldsymbol{H}_i^{-1} \boldsymbol{a}_i$ to concentrate near a constant as $p \to \infty$. Of course, the above explanation is not rigorous in that $L_i(\boldsymbol{u})$ is not smooth and $\boldsymbol{H}_i$ may not exist. This is one technical challenge we address in the proof.

Since $\frac{1}{2} A_p^* v^2$ is a good approximation of $g_{p,i}(v)$, we can now approximate the optimization problem in (19) by

$$\begin{aligned} \widetilde{x}_i &= \arg\min_{x_i \in [-1,1]} \frac{A_p^*(x_i - \beta_i)^2}{2} + \boldsymbol{a}_i^\intercal \boldsymbol{u}_{\setminus i}^* (x_i - \beta_i) \\ &= \text{Prox}_{[-1,1]}\left(\beta_i - \tfrac{\boldsymbol{a}_i^\intercal \boldsymbol{u}_{\setminus i}^*}{A_p^*}\right), \end{aligned} \tag{23}$$

where $\text{Prox}_{[-1,1]}$ denotes the proximal operator of the indicator function on $[-1, 1]$. Its solution, denoted by $\widetilde{x}_i$, provides a good surrogate of x_i^*, as shown in the following proposition.

Proposition 3: Under (A.1)-(A.5), for any $\gamma > 2$, there exists $c > 0$, such that, for any $i \in [p]$ and $\varepsilon \in (0, 1)$,

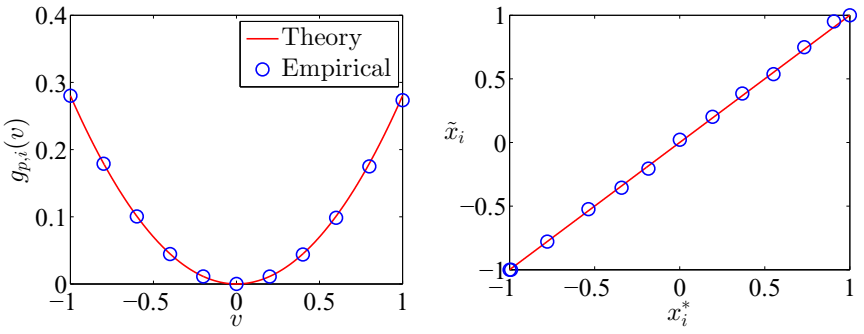
$$\mathbb{P}\left(|x_i^* - \widetilde{x}_i| > \varepsilon\right) < \frac{c}{\varepsilon^2} e^{-p^{\frac{1}{\gamma}} \varepsilon^2 / c}, \tag{24}$$

We prove this result in Section III-B. Here, we demonstrate the accuracy of the approximations stated in (20) and (24) via numerical results shown in Figure 3.

Thanks to the independence between $\boldsymbol{a}_i$ and $\boldsymbol{u}_{\setminus i}^*$, the surrogate solution $\widetilde{x}_i$ is much easier to analyze than x_i^*. Accordingly, we can consider the following approximations of $\widehat{\boldsymbol{\beta}}$ and N_e:

$$\widetilde{\boldsymbol{\beta}} \stackrel{\text{def}}{=} \text{sign}(\widetilde{\boldsymbol{x}}) \quad \text{and} \quad \widetilde{N}_e \stackrel{\text{def}}{=} \sum_{i=1}^{p} \mathbb{1}_{\widetilde{\beta}_i \neq \beta_i}. \tag{25}$$

Applying a union bound to (24) gives us $\max_i |x_i^* - \widetilde{x}_i| \xrightarrow{\mathcal{P}} 0$, *i.e.*, the surrogate vector $\widetilde{\boldsymbol{x}}$ is close to $\boldsymbol{x}^*$ in ℓ_∞ distance. This then allows us to show that $\mathbb{P}(\widehat{\boldsymbol{\beta}} \neq \widetilde{\boldsymbol{\beta}}) \to 0$, which also implies $d_{\text{TV}}(N_e, \widetilde{N}_e) \to 0$.



(a) (b)

Figure 3: Accuracy of the leave-one-out approximation. (a) Comparison of $g_{p,i}(v)$ with its limiting prediction $\frac{1}{2}A_p^* v^2$, (b) Comparison between x_i^* and its leave-one-out approximation $\widetilde{x}_i$. In our experiments, $\sigma_p^2 = 1$, $\delta_p = 1$ and $p = 1000$.

Proposition 4: Under (A.1)-(A.5), it holds that

$$\mathbb{P}(\widetilde{\boldsymbol{\beta}} \neq \widehat{\boldsymbol{\beta}}) \leq \lambda_p p^{-1/5} \operatorname{polylog} p, \tag{26}$$

and accordingly,

$$d_{\mathrm{TV}}(N_e, \widetilde{N}_e) \leq \lambda_p p^{-1/5} \operatorname{polylog} p. \tag{27}$$

The proof of Proposition 4 can be found in Section III-C. It shows that the distribution of N_e is well captured by that of $\widetilde{N}_e$. Therefore, to obtain the limiting distribution of N_e, we just need to analyze $\widetilde{N}_e$, which is what we are going to do next.

C. Approximate independence of $\{\widetilde{\beta}_i\}_{i\in[p]}$

To derive the distribution of $\widetilde{N}_e$, we need to know the joint distribution of $\{\widetilde{x}_i\}_{i\in[p]}$. From (23), we know $\{\widetilde{x}_i\}_{i\in[p]}$ is determined by $\{\boldsymbol{a}_i^\intercal \boldsymbol{u}_{\setminus i}^*\}_{i\in[p]}$. Since for $i \neq j$, $\boldsymbol{u}_{\setminus i}^* \approx \boldsymbol{u}_{\setminus j}^*$, the set of variables $\{\widetilde{x}_i\}_{i\in[p]}$ are correlated, but the correlations are weak. In fact, we can prove something stronger. The following result, proved in Section IV-A, shows that any size-k subset of $\{\boldsymbol{a}_i^\intercal \boldsymbol{u}_{\setminus i}^*\}_{i\in[p]}$ are approximately independent, provided that k is not too large.

Proposition 5: If $k \leq \sqrt{p}$, then there exists $c > 0$ such that, for any $b_i \in \mathbb{R}, i = 1, 2, \ldots, k$ and $\varepsilon > 0$,

$$\mathbb{P}\left(\bigcap_{i=1}^k \{\boldsymbol{a}_i^\intercal \boldsymbol{u}_{\setminus i}^* \leq b_i\}\right) \in \left[\prod_{i=1}^k \Phi\left(\frac{b_i - \sqrt{\delta_p}\varepsilon}{f_p}\right) - \Delta_{p,k}, \prod_{i=1}^k \Phi\left(\frac{b_i + \sqrt{\delta_p}\varepsilon}{f_p}\right) + \Delta_{p,k}\right], \tag{28}$$

where $\Phi(\cdot)$ is the CDF of the standard Gaussian and $\Delta_{p,k} \overset{\text{def}}{=} ckp^{\frac{1}{2}} e^{-c^{-1}p\min\left\{\frac{\varepsilon^2}{k^2}, \frac{\varepsilon}{\sqrt{p}}\right\}}$.

It follows from (23) and (25) that $\left\{\widetilde{\beta}_i \neq \beta_i\right\} = \left\{\boldsymbol{a}_i^{\mathsf{T}} \boldsymbol{u}^*_{\setminus i} \leq -A_p^*\right\}$. (Recall that we have assumed that $\beta_i = -1$ for all i.) By taking $b_i = -A_p^*$ in (28), we can conclude that the k events $\{\widetilde{\beta}_i \neq \beta_i\}_{i\in[k]}$ (or equivalently $\{\mathbb{1}_{\widetilde{\beta}_i \neq \beta_i}\}_{i\in[k]}$) are also approximately independent. This is made precise by the following proposition, whose proof can be found in Appendix E.

Proposition 6: If $k \leq p^{\frac{1}{8}}$, there exists $c > 0$, such that

$$\mathbb{P}\left(\bigcap_{i=1}^{k}\left\{\widetilde{\beta}_i \neq \beta_i\right\}\right) \in \left[\Phi^k\left(-\tfrac{1+cp^{-1/4}}{\tau_p}\right) - ce^{-p^{1/4}/c}, \Phi^k\left(-\tfrac{1-cp^{-1/4}}{\tau_p}\right) + ce^{-p^{1/4}/c}\right]. \tag{29}$$

Moreover, if $\sigma_p^2 \geq \frac{c'}{\log^2 p}$ for some $c' > 0$, then for all large enough p,

$$\left|\mathbb{P}\Big(\bigcap_{i=1}^{k}\left\{\widetilde{\beta}_i \neq \beta_i\right\}\Big) - \Phi^k\left(-\tfrac{1}{\tau_p}\right)\right| \leq \Phi^k\left(-\tfrac{1}{\tau_p}\right) kp^{-1/4}\,\mathrm{polylog}\, p, \tag{30}$$

D. Proof of the Main Theorem

We are now ready to prove Theorem 1 by showing that the limiting distribution of $\widetilde{N}_e$ converges to Poisson. Recall that $\widetilde{N}_e = \sum_{i=1}^{p} \mathbb{1}_{\widetilde{\beta}_i \neq \beta_i}$. The approximate independence of $\{\mathbb{1}_{\widetilde{\beta}_i \neq \beta_i}\}$ makes the analysis tractable. Classical results on Poisson approximation of rare events deal with the sum of p i.i.d. Bernoulli random variables with success probability λ/p. As $p \to \infty$, the sum converges in distribution to a Poisson random variable with rate λ. Things are slightly different in our case, since $\widetilde{N}_e$ is a summation of p weakly correlated Bernoulli random variables. The following proposition, proved in Section IV-B, shows that the Poisson convergence still holds under the weaker condition of approximate independence.

Proposition 7: If $\limsup_{p\to\infty} \frac{\lambda_p}{\sqrt{\log p}} < \infty$, then

$$d_{\mathrm{TV}}(\widetilde{N}_e, \mathscr{P}(\lambda_p)) \leq p^{-1/5}\,\mathrm{polylog}\, p, \tag{31}$$

where $\mathscr{P}(\lambda)$ denotes a Poisson distribution with parameter λ.

Finally, since the TV distance is a metric, the statement of Theorem 1 immediately follows from (27), (31) and the triangle inequality.

E. Proof of Proposition 7

Using the Gaussian tail bounds (133) and (134) given in Appendix F, we can get

$$\lim_{p\to\infty} \lambda_p = \begin{cases} 0, & \alpha^* > 1, \\ \infty, & \alpha^* < 1. \end{cases} \tag{32}$$

Therefore, if $\alpha^* > 1$, it directly follows from Theorem 1 that $\mathbb{P}(N_e = 0) = 1$.

The case that $\alpha^* < 1$ is more complicated. One can show that $\lambda_p \geq p^{c(\alpha^*)}$, where $c(\alpha^*)$ is some constant, so it is possible $\lim_{p\to\infty} d_{\text{TV}}(\widetilde{N}_e, N_e) \not\to 0$. Instead, we can look at a subset $\mathcal{K} \subset [p]$. Define $N_{e,\mathcal{K}}$ as the number of error bits in $\mathcal{K}$. and $\lambda_{p,\mathcal{K}} \stackrel{\text{def}}{=} |\mathcal{K}|\Phi(-\tau_p^{-1})$. We can find $\mathcal{K}$ satisfying $\lambda_{p,\mathcal{K}} \asymp \sqrt{\log p}$. Then following same steps of proving Proposition 4 and Proposition 10 in Appendix G, we can show $\lim_{p\to\infty} \mathbb{P}(N_{e,\mathcal{K}} = 0) = 0$, which indicates that $\lim_{p\to\infty} \mathbb{P}(N_e = 0) = 0$, since $N_{e,\mathcal{K}} \leq N_e$.

Finally, we prove (14). If α_p satisfies (13), then for large p, $\sigma_p^2 \asymp (\log p)^{-1}$. Letting $t = \sigma_p^2$ in (80), it follows that if $\alpha_p \to \alpha^*$, then $2\alpha_p\tau_p^2 \log p \to 1$. On the other hand, from the auxiliary bounds (131) given in Appendix E, we can get $\frac{m(-\tau_p^{-1})}{\tau_p} \to 1$. Applying (9) and (10) gives us

$$\begin{aligned}
\lim_{p\to\infty} \mathbb{P}(N_e = 0) &= \lim_{p\to\infty} \exp\left\{-p\Phi(-1/\tau_p)\right\} \\
&\stackrel{\text{(a)}}{=} \lim_{p\to\infty} \exp\left\{-p\cdot\tau_p\varphi(-1/\tau_p)\right\} \\
&\stackrel{\text{(b)}}{=} \lim_{p\to\infty} \exp\left\{-p(2\alpha_p\log p)^{-1/2}\frac{e^{-\alpha_p\log p}}{\sqrt{2\pi}}\right\} \\
&= \lim_{p\to\infty} \exp\left\{-\exp\left\{-\log p\left(\alpha_p - 1 + \frac{\log(\alpha_p)}{2\log p} + \frac{\log(4\pi)+\log\log p}{2\log p}\right)\right\}\right\} \\
&\stackrel{\text{(c)}}{=} e^{-e^{-x}},
\end{aligned}$$

where step (a) follows from $\frac{m(-\tau_p^{-1})}{\tau_p} \to 1$, step (b) follows from $2\alpha_p\tau_p^2\log p \to 1$ and we use (13) in step (c).

III. The Limiting Quadratic Function

The goal of this technical section is to make the approximations shown in Figure 3 rigorous.

A. *Proof of Proposition 2*

To lighten notation, we will sometimes omit the leave-one-out subscript as used in Sec. II-A. For example, $\boldsymbol{A}_{\setminus i}$ will be replaced by $\boldsymbol{A}$, and $\boldsymbol{a}_i$ by $\boldsymbol{a}$, as long as doing so causes no confusion.

Let us first introduce the following function:

$$\mathcal{G}_p(\boldsymbol{s}) \stackrel{\text{def}}{=} \max_{\boldsymbol{u}}[\boldsymbol{s}^\mathsf{T}\boldsymbol{u} - L(\boldsymbol{u})] - [\boldsymbol{s}^\mathsf{T}\boldsymbol{u}^* - L(\boldsymbol{u}^*)],, \tag{33}$$

where $L(\boldsymbol{u}) = \|\boldsymbol{A}^\mathsf{T}\boldsymbol{u}\|_1 + \boldsymbol{u}^\mathsf{T}\boldsymbol{y} + \frac{1}{2}\|\boldsymbol{u}\|^2$ and $\boldsymbol{u}^* = \arg\min_{\boldsymbol{u}} L(\boldsymbol{u})$. Using $\mathcal{G}_p(\boldsymbol{s})$ and omitting subscript i, scalar function $g_{p,i}(v)$ defined in (18) can be also expressed as:

$$g_p(v) = \mathcal{G}_p(\boldsymbol{a}v)$$

and correspondingly, we re-write (19) as:

$$\min_{-1\le x\le 1} g_p(x-\beta)+\boldsymbol{a}^\intercal \boldsymbol{u}^*(x-\beta). \tag{34}$$

It can be seen that $\mathcal{G}_p(\boldsymbol{s})$ is related with the conjugate function of $L(\boldsymbol{u})$, which is a strongly convex function. Therefore, $\mathcal{G}_p(\boldsymbol{s})$ and $g_p(v)$ possess some nice properties that will be useful in our proof. We gather them together in Appendix A.

We first show that $g_p(v)$ concentrates around its expectation, which is the following proposition. Its proof will be given in Appendix C.

Proposition 8: There exists $c>0$, s.t. for any $\varepsilon>0$,

$$\mathbb{P}\left(\sup_{v\in[-2,2]}|g_p(v)-\mathbb{E}g_p(v)|>\varepsilon\right)\le \frac{c\delta_p}{\varepsilon}e^{-c^{-1}p\min\left\{\frac{\varepsilon^2}{\delta_p},\varepsilon\right\}}. \tag{35}$$

The next result shows that $\mathbb{E}g_p(v)$ is essentially a quadratic function in the large p limit.

Proposition 9: For any $v\in[-2,2]$,

$$\left|\mathbb{E}g_p(v)-\frac{1}{2}A_p v^2\right|\le \frac{16\delta_p}{\sigma_p^2 p}, \tag{36}$$

where A_p is defined in (21).

Proof: First we introduce the following auxiliary functions:

$$Q_p(\theta)\overset{\text{def}}{=}\min_{\boldsymbol{x}\in[-1,1]^p}\frac{\|\boldsymbol{A}\boldsymbol{x}-\boldsymbol{y}+\sqrt{\theta}\widetilde{\boldsymbol{a}}\|^2}{2p},\ \theta\ge 0, \tag{37}$$

where $\widetilde{\boldsymbol{a}}\sim\mathcal{N}(0,\boldsymbol{I}_n)$, independent of $\boldsymbol{A},\boldsymbol{w}$. Clearly, the original problem (2) is the special case when $\theta=0$. For notational convenience, we also define the expectation of $Q_p(\theta)$ as:

$$\begin{aligned}\overline{Q}_p(\theta)&\overset{\text{def}}{=}\mathbb{E}Q_p(\theta)\\&=\frac{1}{p}\mathbb{E}\max_{\boldsymbol{u}}\boldsymbol{u}^\intercal\sqrt{\theta}\widetilde{\boldsymbol{a}}-L(\boldsymbol{u}),\end{aligned} \tag{38}$$

where $L(\boldsymbol{u})$ is given in (33). Note that the connection between $\overline{Q}_p(\theta)$ and $\mathbb{E}g_p(v)$ is:

$$\mathbb{E}g_p(v)=\frac{\overline{Q}_p(v^2/p)-\overline{Q}_p(0)}{v^2/p}v^2, \tag{39}$$

i.e., $\mathbb{E}g_p(v)$ can be approximated by the derivative of $\overline{Q}_p(\theta)$ at $\theta=0$. To make this intuition rigorous, we need to study the analytical properties of $\overline{Q}_p(\theta)$.

First, we show that $\overline{Q}_p(\theta)$ is differentiable on $[0,\infty)$ and $\overline{Q}_p'(\theta)$ is Lipschitz continuous. Indeed, from (33) and (38),

$$\begin{aligned}\overline{Q}_p'(\theta) &= \frac{1}{p}\frac{\partial}{\partial\theta}\left[\mathbb{E}\max_{\boldsymbol{u}} \boldsymbol{u}^\mathsf{T}\sqrt{\theta}\widetilde{\boldsymbol{a}} - L(\boldsymbol{u})\right] \\ &= \frac{1}{p}\frac{\partial}{\partial\theta}\mathbb{E}\max_{\boldsymbol{u}} -\left(\|\boldsymbol{A}^\mathsf{T}\boldsymbol{u}\|_1 + \boldsymbol{\beta}^\mathsf{T}\boldsymbol{A}^\mathsf{T}\boldsymbol{u} + \frac{1}{2}\|\boldsymbol{u}\|^2 + \sqrt{\theta+\sigma_p^2}\boldsymbol{u}^\mathsf{T}\widetilde{\boldsymbol{w}}\right) \qquad (40)\\ &\overset{\text{(a)}}{=} -\frac{\mathbb{E}\widetilde{\boldsymbol{w}}^\mathsf{T}\hat{\boldsymbol{u}}_\theta}{2p\sqrt{\theta+\sigma_p^2}}, \qquad (41)\end{aligned}$$

where $\widetilde{\boldsymbol{w}} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_n)$ and $\hat{\boldsymbol{u}}_\theta$ corresponds to the optimal solution of (40). In step (a), we use dominated convergence theorem (DCT) to interchange derivative and expectation. By the same argument of (77) in Appendix A, we have for any $b, c \geq 0$,

$$\|\hat{\boldsymbol{u}}_b - \hat{\boldsymbol{u}}_c\| \leq \left|\sqrt{b+\sigma_p^2} - \sqrt{c+\sigma_p^2}\right| \|\widetilde{\boldsymbol{w}}\|. \qquad (42)$$

On the other hand, for any $\theta \geq 0$,

$$\|\hat{\boldsymbol{u}}_\theta\| = \min_{\boldsymbol{x}\in[-1,1]^p} \|\boldsymbol{A}\boldsymbol{x} - (\boldsymbol{A}\boldsymbol{\beta} + \sqrt{\theta+\sigma_p^2}\widetilde{\boldsymbol{w}})\| \leq \sqrt{\theta+\sigma_p^2}\|\widetilde{\boldsymbol{w}}\|. \qquad (43)$$

Combining (41), (42) and (43), for any $b > c \geq 0$, we can get

$$\left|\overline{Q}_p'(b) - \overline{Q}_p'(c)\right| \leq \frac{\delta_p|b-c|}{\sigma_p^2}. \qquad (44)$$

Therefore, $\overline{Q}_p'(h)$ is $\frac{\delta_p}{\sigma_p^2}$-Lipschitz.

Now we are ready to analyze $\mathbb{E}g_p(v)$. By the mean value theorem, we get from (39) that

$$\mathbb{E}g_p(v) = \overline{Q}_p'\left(\tfrac{\kappa_p v^2}{p}\right) v^2, \qquad (45)$$

where $\kappa_p \in [0,1]$. From (44) and (45), we deduce that

$$\left|\mathbb{E}g_p(v) - \overline{Q}_p'(0)v^2\right| \leq \frac{v^4\delta_p}{\sigma_p^2 p} \leq \frac{16\delta_p}{\sigma_p^2 p}. \qquad (46)$$

On the other hand, from (41),

$$\overline{Q}_p'(0) = -\frac{\mathbb{E}\widetilde{\boldsymbol{w}}^\mathsf{T}\hat{\boldsymbol{u}}_0}{2\sigma_p^2 p} = -\frac{\mathbb{E}\boldsymbol{w}^\mathsf{T}\boldsymbol{u}^*}{2\sigma_p^2 p}, \qquad (47)$$

It can be checked from (15) that $\boldsymbol{u}^* = \boldsymbol{A}\boldsymbol{x}^* - \boldsymbol{y}$. Combining (46) and (47), we get (36). ■

Remark 3: It will be shown later [c.f. (59)] that $A_p \geq C\delta_p$, for some constant $C > 0$. Therefore, we know from (36) that the quadratic approximation of $\mathbb{E}g_p(v)$ is accurate for large p, if $\sigma_p \gg p^{-1/2}$. We will prove that, when $\sigma_p < \frac{c}{\sqrt{\log p}}$ for some constant c, perfect recovery

is achieved with high probability. This means that $\sigma_p \gg p^{-1/2}$ already covers the regime where we are most interested in. In the following, we will take $\sigma_p \geq \frac{1}{\log p}$.

Proposition 8 and 9 immediately implies the first part of Proposition 2, i.e., (20). Next we show A_p converges to A_p^* in the high-dimensional limit. From (47),

$$A_p = 2\overline{Q}_p'(0). \tag{48}$$

Hence, it boils down to analyzing $\overline{Q}_p'(\theta)$ and its limit, which can be done as follows.

1) Convergence of $Q_p(\theta)$: The CGMT framework in [13], [33] can be readily applied to computing the limit of $Q_p(\theta)$ in high dimensions.

Lemma 1: There exists $c > 0$, s.t., for any $\varepsilon > 0$ and $\theta \in [0, 1]$,

$$\mathbb{P}\left(|Q_p(\theta) - Q_p^*(\theta)| > \varepsilon\right) \leq \frac{ce^{-p\min\left\{\frac{\varepsilon^2}{\delta_p}, \varepsilon\right\}/c}}{\min\{\frac{\varepsilon}{\delta_p}, \sqrt{\frac{\varepsilon}{\delta_p}}\}}, \tag{49}$$

where

$$Q_p^*(\theta) = \frac{1}{2}\left[\min_{\tau>0} F_p\big(\tau; \theta + \sigma_p^2, \delta_p\big)\right]^2, \tag{50}$$

with F_p defined in (7). Also for any $\gamma > 2$, there exists $c > 0$ such that

$$\sup_{\theta\in[0,1]} |\overline{Q}_p(\theta) - Q_p^*(\theta)| < cp^{-1/\gamma}. \tag{51}$$

Remark 4: The proof of Lemma 1 will be given in Appendix D. We can find $Q_p^*(0) = \frac{f_p^2}{2}$, where f_p is defined in (8). This can be understood from (37) and (49), since $Q_p^*(\theta)$ is the limiting value of the squared fitting error when the noise variance is $\theta + \sigma_p^2$.

2) Smoothness of $Q_p^*(\theta)$:

Lemma 2: $Q_p^*(\theta)$ is twice differentiable over $\theta \geq 0$, with

$$Q_p^{*\,\prime}(0) = \frac{f_p}{2\tau_p} \tag{52}$$

and $Q_p^{*\,\prime\prime}(\theta) \leq C$, for all $\theta \geq 0$, where C is some constant.

Proof: Note that $Q_p^*(\theta)$ is a composition of $R_p(t)$ and $t(\theta) = \theta + \sigma_p^2$, where $R_p(t)$ is defined in Appendix B. By chain rule, $Q_p^*(\theta)$ is twice differentiable, with $Q_p^{*\,\prime}(0) = \frac{f_p}{2\tau_p}$ and

$$\begin{aligned} Q_p^{*\,\prime\prime}(\theta) &= R_p''(t)t'(\theta) + R_p'(t)t''(\theta) \\ &= R_p''(\theta + \sigma_p^2). \end{aligned} \tag{53}$$

Then together with bound (84) shown in Appendix B, we know there exists $C > 0$, s.t., $Q_p^{*\,\prime\prime}(\theta) \leq C$, for all $\theta \geq 0$. ■

3) Convergence of A_p to A_p^*: Now we can show the convergence of the curvature A_p, which also implies the simple limiting form of $g_p(v)$.

Lemma 3: There exists $c > 0$ such that

$$|A_p - A_p^*| < cp^{-1/(2\gamma)}. \tag{54}$$

Proof: For $\gamma > 2$, there exists $C > 0$, s.t. for $\theta \in (0, 1]$,

$$\begin{aligned} |A_p - A_p^*| &\overset{\text{(a)}}{\leq} 2\left|\overline{Q}_p'(0) - \tfrac{\overline{Q}_p(\theta)-\overline{Q}_p(0)}{\theta}\right| + 2\left|\tfrac{\overline{Q}_p(\theta)-\overline{Q}_p(0)}{\theta} - \tfrac{Q_p^*(\theta)-Q_p^*(0)}{\theta}\right| \\ &\quad + 2\left|\tfrac{Q_p^*(\theta)-Q_p^*(0)}{\theta} - {Q_p^*}'(0)\right| \\ &\overset{\text{(b)}}{\leq} C\left(\tfrac{\theta\delta_p}{\sigma_p^2} + \tfrac{\sqrt{\delta_p}p^{-\frac{1}{\gamma}}}{\theta} + \theta\right), \end{aligned} \tag{55}$$

where in step (a), we use (22), (48) and (52) and in step (b), we use (44), (51) and Lemma 2. Therefore, taking $\theta = p^{-\frac{1}{2\gamma}}$ and using Assumptions (A.4) and (A.5), we can get (54). ■

B. Proof of Proposition 3

Proposition 2 indicates that the original scalar problem (34) can be well approximated by

$$\min_{x\in[-1,1]} \frac{1}{2}A_p(x-\beta)^2 + \boldsymbol{a}^\intercal \boldsymbol{u}^*(x-\beta), \tag{56}$$

which has an explicit optimal solution:

$$\breve{x} = \text{Prox}_{[-1,1]}\left(\beta - \frac{\boldsymbol{a}^\intercal \boldsymbol{u}^*}{A_p}\right). \tag{57}$$

Note that the difference between $\breve{x}$ and $\widetilde{x}$ should be small, as implied by (23), (57) and (54). In fact, we can directly prove $\widetilde{x} \to x^*$ without considering $\breve{x}$. The reason for us to introduce this intermediate variable is to achieve a better convergence rate in our proof.

The first lemma below shows that the objective function of (56), i.e.,

$$\widehat{\ell}_p(x) = \frac{1}{2}A_p(x-\beta)^2 + \boldsymbol{a}^\intercal \boldsymbol{u}^*(x-\beta) \tag{58}$$

is strongly convex.

Lemma 4: There exists $K > 0$, s.t., $A_p \geq K\delta_p$ for all p large enough. Therefore, $\widehat{\ell}_p(x)$ is $K\delta_p$-strongly convex.

Proof: By (8) and the definition of A_p^*, we have

$$A_p^* = \frac{1}{2}\left(\delta_p - \frac{1}{2}\right) + \frac{\sigma_p^2}{2\tau_p^2} + \frac{1}{2}\int_{\frac{2}{\tau_p}}^{\infty}\left(x - \frac{2}{\tau_p}\right)^2 \Phi(dx) \geq \frac{1}{2}\left(\delta_p - \frac{1}{2}\right). \tag{59}$$

Then from assumption (A.5) and (54), we know there exists $K > 0$ s.t. $A_p \geq K\delta_p > 0$ and $\widehat{\ell}_p(x)$ is $K\delta_p$-strongly convex. ■

Then together with uniform convergence proved in Proposition 2, we can show $x^* \to \breve{x}$.

Lemma 5: There exists $c > 0$ s.t., for $\varepsilon \in (0,1)$,

$$\mathbb{P}\left(|x^* - \breve{x}| > \varepsilon\right) < \frac{c}{\varepsilon^2} e^{-p\varepsilon^4/c}. \tag{60}$$

Proof: Since $\widehat{\ell}_p(x)$ is $K\delta_p$-strongly convex,

$$\widehat{\ell}_p(x^*) - \widehat{\ell}_p(\breve{x}) \geq \frac{1}{2} K\delta_p (x^* - \breve{x})^2. \tag{61}$$

Let $\ell_p(x)$ be the objective function in (19). From (20) we know there exists $c > 0$, s.t., for $\varepsilon \in (0,1)$, $|\widehat{\ell}_p(x^*) - \ell_p(x^*)| \leq \delta_p\varepsilon$ and $|\widehat{\ell}_p(\breve{x}) - \ell_p(\breve{x})| \leq \delta_p\varepsilon$ with probability greater than $1 - \frac{c}{\varepsilon} e^{-p\varepsilon^2/c}$. This indicates

$$\widehat{\ell}_p(x^*) - \hat{\ell}_p(\breve{x}) \leq [\ell_p(x^*) + \sqrt{\delta_p}\varepsilon] - [\ell_p(\breve{x}) - \sqrt{\delta_p}\varepsilon] \leq 2\delta_p\varepsilon. \tag{62}$$

From (61) and (62), we can get there exists $c > 0$ s.t. for all $\varepsilon \in (0,1)$, $\mathbb{P}\left(|x^* - \breve{x}| > \sqrt{\varepsilon}\right) < \frac{c}{\varepsilon} e^{-p\varepsilon^2/c}$. Then changing $\sqrt{\varepsilon}$ to ε in the above, we get (60). ■

Furthermore, using (54) we can also show $\breve{x} \to \widetilde{x}$.

Lemma 6: For $\gamma > 2$, there exists $c > 0$, s.t., for $\varepsilon \in (0,1)$,

$$\mathbb{P}\left(|\breve{x} - \widetilde{x}| > \varepsilon\right) < \frac{c}{\varepsilon} e^{-p^{\frac{1}{\gamma}}\varepsilon^2/c}.$$

Proof: By the non-expansiveness of proximal operator $\mathrm{Prox}_{[-1,1]}(\cdot)$, from (23) and (57) we know there exists $C > 0$, s.t.,

$$|\breve{x} - \widetilde{x}| \leq \left|\frac{1}{A_p} - \frac{1}{A_p^*}\right| |\boldsymbol{a}^\intercal \boldsymbol{u}^*| \leq \frac{C}{\delta_p^2} |\boldsymbol{a}^\intercal \boldsymbol{u}^*| p^{-\frac{1}{2\gamma}}, \tag{63}$$

where we have used (54) and (59). Recall that $\boldsymbol{u}^* = \boldsymbol{A}\boldsymbol{x}^* - \boldsymbol{y}$, so similar to (104) and (105), we obtain that there exists $c > 0$, s.t., for all $\varepsilon > 0$, $\mathbb{P}\left(\left|\frac{\|\boldsymbol{u}^*\|}{\sqrt{p}} - f_p\right| > \varepsilon\right) \leq \frac{c\sqrt{\delta_p}}{\varepsilon} e^{-p\varepsilon^2/c}$. Since $\boldsymbol{a}$ and $\boldsymbol{u}^*$ are independent, then from (63) it is not hard to show there exists $c > 0$, s.t., for all $\varepsilon \in (0,1)$, $\mathbb{P}\left(|\breve{x} - \widetilde{x}| > \varepsilon\right) \leq \frac{c}{\varepsilon} e^{-p^{1/\gamma}\varepsilon^2/c}$. ■

Lemma 5 and 6 imply Proposition 3, based on which we can now prove Proposition 4.

C. Proof of Proposition 4

Our strategy is to show that $\mathbb{P}(\widetilde{\boldsymbol{\beta}} \neq \hat{\boldsymbol{\beta}})$ is small, which implies $\mathbb{P}(\widetilde{N}_e \neq N_e)$ is small and so is $d_{\mathrm{TV}}(\widetilde{N}_e, N_e)$. Recall that $\widetilde{N}_e$ and N_e are in the same probability space, and we have assumed $\beta_i = -1$, for any $i \in [p]$. Then the following simple relation holds:

$$\{\widetilde{\beta}_i \neq \hat{\beta}_i\} \subset \{|\tilde{x}_i - x_i^*| > p^{-\frac{1}{5}}\} \bigcup \{\tilde{x}_i \in [-p^{-\frac{1}{5}}, p^{-\frac{1}{5}}]\}. \tag{64}$$

Since $\tilde{x}_i = \mathrm{Prox}_{[-1,1]}\left(\beta_i - \frac{\boldsymbol{a}_i^\intercal \boldsymbol{u}_{\setminus i}^*}{A_p^*}\right)$, for $U \in (-1,1)$, $|\tilde{x}_i| \leq U \Leftrightarrow \left|\frac{\boldsymbol{a}_i^\intercal \boldsymbol{u}_{\setminus i}^*}{A_p^*} + 1\right| \leq U$. Then letting $b_i = A_p^*(-1 + p^{-\frac{1}{5}})$ and $A_p^*(-1 - p^{-\frac{1}{5}})$, $k = 1$ and $\varepsilon = p^{-\frac{1}{5}}$ in (28), we can show $\mathbb{P}(|\tilde{x}_i| \leq p^{-\frac{1}{5}}) \leq \Phi(-1/\tau_p) p^{-\frac{1}{5}} \operatorname{polylog} p$, similar to (126) shown in Appendix E. On the other hand, letting $\varepsilon = p^{-\frac{1}{5}}$ in (24), $\mathbb{P}(|x_i^* - \tilde{x}_i| > p^{-\frac{1}{5}}) < p^{\frac{2}{5}} e^{-p^{1/12}/c}$. These together with (64) indicate

$$\mathbb{P}(\widetilde{\beta}_i \neq \hat{\beta}_i) \leq \Phi(-1/\tau_p) p^{-\frac{1}{5}} \operatorname{polylog} p. \tag{65}$$

By union bound,

$$\mathbb{P}(\widetilde{\boldsymbol{\beta}} \neq \hat{\boldsymbol{\beta}}) \leq \sum_{i=1}^{p} \mathbb{P}(\widetilde{\beta}_i \neq \hat{\beta}_i) \leq \lambda_p p^{-\frac{1}{5}} \operatorname{polylog} p.$$

Since $d_{\mathrm{TV}}(\widetilde{N}_e, N_e) \leq \mathbb{P}(\widetilde{N}_e \neq N_e) \leq \mathbb{P}(\widetilde{\boldsymbol{\beta}} \neq \hat{\boldsymbol{\beta}})$, we obtain (27).

IV. Asymptotic Distributions

This is another technical section. Our main goal here is to derive the asymptotic distribution of $\{\widetilde{x}_i\}$ and that of $\widetilde{N}_e$.

A. Proof of Proposition 5

By the exchangeability of $\left\{\boldsymbol{a}_i^\intercal \boldsymbol{u}_{\setminus i}^*\right\}_{i\in[p]}$, we just need to consider the joint distribution of $\left\{\boldsymbol{a}_i^\intercal \boldsymbol{u}_{\setminus i}^*\right\}_{i\in[k]}$, i.e., the first k coordinates. A key result we are going to establish is that $\left\{\boldsymbol{a}_i^\intercal \boldsymbol{u}_{\setminus i}^*\right\}_{i\in[k]}$ are approximately independent, provided that k is not too large.

Let $\boldsymbol{u}_{\setminus[k]}^*$ be the optimal solution of

$$\min_{\boldsymbol{u}} \|\boldsymbol{A}_{\setminus[k]}^\intercal \boldsymbol{u}\|_1 + \boldsymbol{u}^\intercal \boldsymbol{A}_{\setminus[k]} \boldsymbol{\beta}_{\setminus[k]} + \frac{1}{2}\|\boldsymbol{u}\|^2 + \boldsymbol{u}^\intercal \boldsymbol{w}, \tag{66}$$

where $\boldsymbol{A}_{\setminus[k]}$ is the matrix formed by removing the first k columns of $\boldsymbol{A}$ and $\boldsymbol{\beta}_{\setminus[k]}$ is defined in the same way. In other words, $\boldsymbol{u}_{\setminus[k]}^*$ is the leave-k-out solution of $\min_{\boldsymbol{u}} L(\boldsymbol{u})$. Also define

$$\widetilde{\boldsymbol{u}}_{\setminus[k]} \stackrel{\text{def}}{=} \frac{\sqrt{p} f_p \boldsymbol{u}_{\setminus[k]}^*}{\|\boldsymbol{u}_{\setminus[k]}^*\|}. \tag{67}$$

Since $\boldsymbol{a}_i \overset{i.i.d.}{\sim} \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_p/p)$, $i = 1, 2, \ldots, k$ and $\widetilde{\boldsymbol{u}}_{\setminus[k]}$ is independent of $\{\boldsymbol{a}_i\}_{i\in[k]}$, with fixed norm $\sqrt{p}f_p$, the joint distribution of $\left\{\boldsymbol{a}_i^\intercal \widetilde{\boldsymbol{u}}_{\setminus[k]}\right\}_{i\in[k]}$ is:

$$\left(\boldsymbol{a}_1^\intercal \widetilde{\boldsymbol{u}}_{\setminus[k]} \quad \boldsymbol{a}_2^\intercal \widetilde{\boldsymbol{u}}_{\setminus[k]} \quad \ldots \quad \boldsymbol{a}_k^\intercal \widetilde{\boldsymbol{u}}_{\setminus[k]}\right)^\intercal \sim \mathcal{N}(\boldsymbol{0}, f_p^2 \boldsymbol{I}_p). \tag{68}$$

Our proof of approximate independence of $\left\{\boldsymbol{a}_i^\intercal \boldsymbol{u}^*_{\setminus i}\right\}_{i\in[k]}$ consists of two steps:

1) Show the joint distribution of $\left\{\boldsymbol{a}_i^\intercal \boldsymbol{u}^*_{\setminus i}\right\}_{i\in[k]}$ is closed to that of $\left\{\boldsymbol{a}_i^\intercal \boldsymbol{u}^*_{\setminus[k]}\right\}_{i\in[k]}$. This is proved in Lemma 9.
2) Show the joint distribution of $\left\{\boldsymbol{a}_i^\intercal \boldsymbol{u}^*_{\setminus[k]}\right\}_{i\in[k]}$ is closed to that of $\left\{\boldsymbol{a}_i^\intercal \widetilde{\boldsymbol{u}}_{\setminus[k]}\right\}_{i\in[k]}$, which are mutually independent. This is proved in Lemma 12.

Details of the proof can be found in Appendix E.

B. The Limiting Poisson Law of $\widetilde{N}_e$

Before presenting the actual proof, it would help to first show some heuristic derivations. We employ the following general inclusion-exclusion principle [34, p.106]: for any $k \in [p]$, the probability P_k that exactly k among p events $A_1, \ldots, A_p$ occur is

$$P_k = \sum_{m=k}^{p} \binom{m}{k} (-1)^{m-k} S_m, \tag{69}$$

where

$$S_m = \begin{cases} 1 & m = 0, \\ \sum_{1\le i_1<\cdots<i_m\le p} \mathbb{P}\left(\bigcap_{j=1}^m A_{i_j}\right) & 1 \le m \le p. \end{cases} \tag{70}$$

In our setting, $A_i = \{\tilde{\beta}_i \neq \beta_i\}, i = 1, 2, \ldots, p$ and $P_k = \mathbb{P}(\widetilde{N}_e = k)$.

By the exchangeability of $\{\tilde{\beta}_i\}_{i\in[p]}$, we have $S_m = \binom{p}{m} S_{[m]}, ,$ with $S_{[m]} = \mathbb{P}\left(\tilde{\beta}_i \neq \beta_i, i \in [m]\right)$. From Proposition 6, for large enough p and "reasonably large" m, $S_{[m]} \approx \Phi^m\left(-\frac{1}{\tau_p}\right)$, so

$$S_m = \frac{p! S_{[m]}}{m!(p-m)!} \approx \frac{\lambda_p^m}{m!}, \tag{71}$$

where λ_p is defined in (9). Then combining (69) and (71), we have

$$\begin{aligned} \mathbb{P}\left(\widetilde{N}_e = k\right) &= \sum_{m=0}^{p-k} \binom{k+m}{k} (-1)^m S_{k+m} \\ &\approx \sum_{m=0}^{p-k} \frac{(k+m)!}{m!k!} (-1)^m \frac{\lambda_p^{k+m}}{(k+m)!} \\ &\approx \frac{\lambda_p^k}{k!} e^{-\lambda_p}, \end{aligned} \tag{72}$$

which implies that the PMF of $\widetilde{N}_e$ is approximately Poisson with rate λ_p.

We now quantitatively analyze the error of approximation in (72). First, we approximate the right-hand side of (69) by a truncated sum: $\sum_{m=k}^{L}\binom{m}{k}(-1)^{m-k}S_m$, with $L \leq p$. The reason for this operation is that $S_{[m]} \approx \Phi^m\left(-\frac{1}{\tau_p}\right)$ may not be accurate for large m, since we only have approximate finite event independence. We then need to control the error caused by the truncation. Accordingly, we can apply Bonferroni's inequality [34, p.110], stated as follows. Under the same setting as (69), for $k+1 \leq L \leq p$, we have

1) If $L-k$ is odd,

$$\sum_{m=k}^{L}\binom{m}{k}(-1)^{m-k}S_m \leq P_k \leq \sum_{m=k}^{L-1}\binom{m}{k}(-1)^{m-k}S_m. \tag{73}$$

2) If $L-k$ is even,

$$\sum_{m=k}^{L-1}\binom{m}{k}(-1)^{m-k}S_m \leq P_k \leq \sum_{m=k}^{L}\binom{m}{k}(-1)^{m-k}S_m. \tag{74}$$

Therefore, we need to choose a reasonably large L to attain a good trade-off between the approximation error of (71) and the truncation error of (73) and (74), such that they are both properly bounded. Our proof of Proposition 7 follows this idea. The details can be found in Appendix H.

V. Conclusion

In this paper, we have presented an exact performance characterization of the box-relaxation decoder in high dimensions. We show that, under certain scalings of the sampling ratio and the noise variance, the number of incorrectly-decoded bits has a limiting Poisson distribution. In addition, a phase transition from nonperfect to perfect recovery takes place at a well-defined critical threshold. Numerical simulations show that the actual performance of the algorithm is well captured by our theoretical predictions. Finally, it is worth mentioning that, although we have assumed that the sensing matrix has i.i.d. normal entries, the results on the limiting Poisson law should hold under more general matrix ensembles. We leave this as an interesting line of work for future investigation.

Appendix

A. Properties of $\mathcal{G}_p(\boldsymbol{s})$ and $g_p(v)$

Lemma 7: For any $\boldsymbol{A}$ and $\boldsymbol{y}$, it holds that:

1) $\mathcal{G}_p(\boldsymbol{s})$ is convex and differentiable in $\mathbb{R}^n$, with

$$\nabla \mathcal{G}_p(\boldsymbol{s}) = \boldsymbol{u}_{\boldsymbol{s}}^* - \boldsymbol{u}^*, \tag{75}$$

where $\boldsymbol{u}_{\boldsymbol{s}}^* \stackrel{\text{def}}{=} \arg\max_{\boldsymbol{u}} \boldsymbol{s}^\intercal \boldsymbol{u} - L(\boldsymbol{u})$.

2) $\nabla \mathcal{G}_p(\boldsymbol{s})$ is 1-Lipschitz continuous, i.e., $\forall \boldsymbol{r}, \boldsymbol{s} \in \mathbb{R}^n$

$$\|\nabla \mathcal{G}_p(\boldsymbol{r}) - \nabla \mathcal{G}_p(\boldsymbol{s})\| \le \|\boldsymbol{r} - \boldsymbol{s}\| \tag{76}$$

or equivalently,

$$\|\boldsymbol{u}_{\boldsymbol{r}}^* - \boldsymbol{u}_{\boldsymbol{s}}^*\| \le \|\boldsymbol{r} - \boldsymbol{s}\|. \tag{77}$$

3) $g_p(v)$ is convex and differentiable with

$$\left|g_p'(v)\right| \le 2\|\boldsymbol{a}\|^2, \tag{78}$$

Proof: Let $L^*(\boldsymbol{s}) \stackrel{\text{def}}{=} \max_{\boldsymbol{u}} \boldsymbol{s}^\intercal \boldsymbol{u} - L(\boldsymbol{u})$, which is the conjugate function of $L(\boldsymbol{u})$. We know $\nabla \mathcal{G}_p(\boldsymbol{s}) = \nabla L^*(\boldsymbol{s}) - \boldsymbol{u}^*$. Since $L(\boldsymbol{u})$ is closed and 1-strongly convex, $L^*(\boldsymbol{s})$ is convex and differentiable with $\nabla L^*(\boldsymbol{s}) = \boldsymbol{u}_{\boldsymbol{s}}^*$ and $\nabla L^*(\boldsymbol{s})$ is 1-Lipschitz continuous [35, Chapter X]. Therefore, from (33) we know $\mathcal{G}_p(\boldsymbol{s})$ is convex. Since $\nabla \mathcal{G}_p(\boldsymbol{r}) - \nabla \mathcal{G}_p(\boldsymbol{s}) = \nabla L^*(\boldsymbol{r}) - \nabla L^*(\boldsymbol{s})$, we get (75) and (76).

Since $g_p(v) = \mathcal{G}_p(\boldsymbol{a}v)$, $g_p(v)$ is also convex and differentiable with $g_p'(v) = \boldsymbol{a}^\intercal \nabla \mathcal{G}_p(\boldsymbol{a}v)$. From (75) and (76), we know $\|\nabla \mathcal{G}_p(\boldsymbol{a}v)\| \le \|\boldsymbol{a}\|v$. Therefore, (78) follows from Cauchy-Schwartz inequality and the fact that $|v| \le 2$. ∎

B. Properties of the Optimization Problem (8)

In this section, we collect some useful properties of the one-dimensional optimization (8), which was first studied in [13]. For our purpose, we consider a slightly more general setting:

$$\begin{aligned} f_p(t) &= \min_{\tau>0} F_p(\tau; t, \delta_p) \\ &= \min_{\tau>0} \frac{\tau}{2}\left(\delta_p - \frac{1}{2}\right) + \frac{t}{2\tau} + \frac{\tau}{2}\int_{\frac{2}{\tau}}^{\infty} \left(x - \frac{2}{\tau}\right)^2 \Phi(dx), \end{aligned} \tag{79}$$

where $t > 0$ is a parameter. Note that (8) and the inline optimization of (50) are the cases where $t = \sigma_p^2$ and $t = (1+\theta)^2\sigma_p^2$, respectively. Also we define the squared loss function: $R_p(t) \stackrel{\text{def}}{=} \frac{f_p^2(t)}{2}$ and evidently, $R_p\left[(1+\theta)^2\sigma_p^2\right] = Q_p^*(\theta)$, where $Q_p^*(\theta)$ is defined in (50).

1) Uniqueness of Optimal Solution: Let $\tau(t)$ be the minimizer of (79), which is the solution of stationary equation:

$$h(\tau) \overset{\text{def}}{=} \delta_p - \frac{1}{2} + \int_{\frac{2}{\tau}}^{\infty} \left(x^2 - \frac{4}{\tau^2}\right) \Phi(dx) - \frac{t}{\tau^2} = 0. \tag{80}$$

By direct differentiation of $h(\tau)$ above, we can show $h'(\tau) = \int_{\frac{2}{\tau}}^{\infty} \frac{8}{\tau^3}\Phi(dx) + \frac{2t}{\tau^3} > 0$, so it is a strictly increasing function. Also $\lim_{\tau\to 0} h(\tau) = -\infty$ and $\lim_{\tau\to\infty} h(\tau) = \delta_p > 0$. This also establishes that the strict convexity of $f_p(t)$. Therefore, $\tau(t)$ is unique for any $t > 0$. Besides, we can directly check that $\tau(t)$ is differentiable with

$$\tau'(t) = \frac{\tau(t)}{8\int_{2/\tau(t)}^{\infty} \Phi(dx) + 2t} > 0, \tag{81}$$

so $\tau(t)$ is strictly increasing.

2) Upper and Lower Bounds of $\tau(t)$: Since $h\left(\sqrt{\frac{t}{\delta_p}}\right) < -\frac{1}{2} + \int_{\frac{2}{\tau}}^{\infty} x^2\Phi(dx) < 0$, by $h(0^+) < 0$, $h(\infty) > 0$ and uniqueness of $\tau(t)$, we have $\tau(t) \geq \sqrt{\frac{t}{\delta_p}}$. Similarly, we can get $\tau(t) \leq \min\{\sqrt{\frac{t}{\delta_p - 1/2}}, \sqrt{\frac{4+t}{\delta_p}}\}$ and $\tau(t) \geq \sqrt{\frac{t}{\delta_p - 1/2 + v_p}}$, where $v_p = \int_{b_p}^{\infty} x^2\Phi(dx)$, with $b_p = 2\sqrt{\frac{\delta_p - 1/2}{t}}$ and evidently, $v_p < 1/2$. Therefore, $\tau(t)$ can be bounded as:

$$\sqrt{\frac{t}{\delta_p - 1/2 + v_p}} \leq \tau(t) \leq \min\left\{\sqrt{\frac{t}{\delta_p - 1/2}}, \sqrt{\frac{4+t}{\delta_p}}\right\}. \tag{82}$$

3) Properties of $f_p(t)$: From (79) we get $f_p(t) \geq 0$, $f_p'(t) = \frac{1}{2\tau(t)} > 0$ and $f_p''(t) = -\frac{\tau'(t)}{2\tau^2(t)} < 0$, so $f_p(t)$ is nonnegative, strictly increasing and concave. On the other hand, letting $\tau = \sqrt{\frac{t}{\delta_p}}$ in (79) we can get $f_p(t) \leq C\left(\frac{\sqrt{t\delta_p}}{2} + 1\right)$, where C is some constant.

4) Properties of $R_p(t)$: By the chain rule, $R_p'(t) = \frac{f_p(t)}{2\tau(t)}$ and $R_p''(t) = \frac{\int_{2/\tau(t)}^{\infty} x\Phi(dx)}{\tau(t)(8\int_{2/\tau(t)}^{\infty}\Phi(dx) + 2t)}$. Therefore, $R_p(t)$ is strictly increasing and convex. From (80), we can show $R_p'(t)$ is bounded:

$$R_p'(t) = \frac{1}{2}\left[\delta_p - \tfrac{1}{2} + \int_{\frac{2}{\tau(t)}}^{\infty} x^2 - \tfrac{2x}{\tau(t)}\Phi(dx)\right] \leq \frac{\delta_p}{2}. \tag{83}$$

On the other hand, $R_p''(t)$ satisfies: $R_p''(t) \leq \frac{\varphi(-2/\tau(t))}{2\tau(t)t}$, where $\varphi(x)$ is the PDF of standard Gaussian. Then using (82) and Assumption (A.4), we know there exists $C > 0$, s.t., for $t > 0$,

$$R_p''(t) \leq \sqrt{\frac{\delta_p}{8\pi}} e^{-\frac{2(\delta_p - 1/2)}{t}} t^{-\frac{3}{2}} \leq C. \tag{84}$$

C. Proof of Proposition 8

We first prove the pointwise convergence of $g_p(v)$ to $\mathbb{E}g_p(v)$: there exists $c > 0$, s.t., for any $v \in [0, 2]$ and $\varepsilon > 0$,

$$\mathbb{P}\left(|g_p(v) - \mathbb{E}g_p(v)| > \varepsilon\right) \leq ce^{-c^{-1}p\min\left\{\frac{\varepsilon^2}{\delta_p}, \varepsilon\right\}}. \tag{85}$$

Recall that $g_p(v) = \mathcal{G}_p(\boldsymbol{a}v)$, so it is equivalent to prove $|\mathcal{G}_p(\boldsymbol{a}v) - \mathbb{E}\mathcal{G}_p(\boldsymbol{a}v)| \to 0$. We first control the moment generating function of $\mathcal{G}_p(\boldsymbol{a}v) - \mathbb{E}\mathcal{G}_p(\boldsymbol{a}v)$. Let $\boldsymbol{b}$ be an i.i.d. copy of $\boldsymbol{a}$. For all $|\lambda| \le \frac{p}{2\sqrt{2}\pi}$, we can apply Theorem 2.2 of [36, p.176] to get

$$\begin{aligned}\mathbb{E}[\exp\lambda(\mathcal{G}_p(\boldsymbol{a}v) - \mathbb{E}\mathcal{G}_p(\boldsymbol{a}v))] &\le \mathbb{E}e^{\frac{\pi\lambda}{2}(\boldsymbol{b}v)^\intercal\nabla\mathcal{G}_p(\boldsymbol{a}v)} \\ &\overset{\text{(a)}}{\le} \mathbb{E}_{\boldsymbol{A},\boldsymbol{w}}\mathbb{E}_{\boldsymbol{a}}e^{\frac{2\lambda^2\pi^2}{p}\|\boldsymbol{a}\|^2} \\ &= \exp\left[-\frac{n}{2}\log\left(1 - \frac{4\lambda^2\pi^2}{p^2}\right)\right] \\ &\overset{\text{(b)}}{\le} \exp\left(\frac{4\delta_p\lambda^2\pi^2}{p}\right).\end{aligned}$$

In step (a), we take expectation over $\boldsymbol{b}$ and use $|v| \le 2$ and $\|\nabla\mathcal{G}_p(\boldsymbol{a}v)\| \le 2\|\boldsymbol{a}\|$, as implied by (76); In step (b), we use the inequality $\log(1+x) \ge \frac{x}{1+x}$, for $x > -1$ and the condition that $|\lambda| \le \frac{p}{2\sqrt{2}\pi}$. As a result, for any $\varepsilon \ge 0$ and $\lambda \in \left[0, \frac{p}{2\sqrt{2}\pi}\right]$,

$$\mathbb{P}\left(g_p(v) - \mathbb{E}g_p(v) > \varepsilon\right) \le e^{-\lambda\varepsilon + \frac{4\delta_p\lambda^2\pi^2}{p}}. \tag{86}$$

After minimizing the exponent on the RHS of (86) over $\lambda \in \left[0, \frac{p}{2\sqrt{2}\pi}\right]$, we can get for any $\varepsilon \in [0, \sqrt{8}\pi\delta_p]$, $\mathbb{P}\left(g_p(v) - \mathbb{E}g_p(v) > \varepsilon\right) \le e^{-\frac{p\varepsilon^2}{16\delta_p\pi^2}}$; for any $\varepsilon > \sqrt{8}\pi\delta_p$, $\mathbb{P}\left(g_p(v) - \mathbb{E}g_p(v) > \varepsilon\right) \le e^{-\frac{p\varepsilon}{4\sqrt{2}\pi}}$. The other direction also holds by the same reasoning. Thus,

$$\mathbb{P}\left(|g_p(v) - \mathbb{E}g_p(v)| > \varepsilon\right) \le 2e^{-\frac{p}{16\pi^2}\min\{\frac{\varepsilon^2}{\delta_p}, \varepsilon\}}. \tag{87}$$

To show uniform convergence (35), it suffices to prove the Lipschitz continuity of $g_p(v)$ and $\mathbb{E}g_p(v)$. From Lemma 1 of [37], we have for all $x > 0$, $\mathbb{P}\left(\|\boldsymbol{a}\|^2 \ge \delta_p + \frac{2\sqrt{\delta_p x}}{\sqrt{p}} + \frac{2x}{p}\right) \le \exp(-x)$. Let $x = n(\sqrt{y+1} - 1)^2$, we have for $y \ge 2$, $\mathbb{P}\left(\frac{\|\boldsymbol{a}\|^2}{\delta_p} - 1 \ge y\right) \le \exp\left(-\frac{ny}{4}\right)$. Therefore, by taking $y = K/\delta_p$, we get for any $K \ge 2\delta_p$,

$$\mathbb{P}(\|\boldsymbol{a}\|^2 > K) \le \mathbb{P}(\|\boldsymbol{a}\|^2 - \delta_p > K/2) \le \exp(-\tfrac{pK}{4}). \tag{88}$$

Combining it with (78), we know for $K \ge 2\delta_p$, $g_p(v)$ is $2K$-Lipschitz with probability greater than $1 - \exp(-\frac{pK}{4})$. From (78), we can also get $\left|\frac{d\mathbb{E}g_p(v)}{dv}\right| \le 2\delta_p$, so $\mathbb{E}g_p(v)$ is $2\delta_p$-Lipschitz continuous over $v \in [0, 2]$. Combining the Lipschitz continuity of $g_p(v)$ and $\mathbb{E}g_p(v)$ with (85), we can obtain (35) by a standard epsilon-net argument as follows. We need to consider different values of ε:

1) If $\varepsilon \geq \delta_p$, we construct an epsilon-net of $[0,2]$ formed by the following points: $v_k = \frac{k}{4}$, $k = 1, 2, \ldots, 8$. For any $v \in [0,2]$, denote v^* as the closest point to v in the above epsilon-net. By construction, $|v - v^*| \leq \frac{1}{8}$. If $g_p(v)$ is $2K$-Lipschitz, then for any $v \in [0,2]$,

$$\begin{aligned} |g_p(v) - \mathbb{E}g_p(v)| &\leq |g_p(v) - g_p(v^*)| + |g_p(v^*) - \mathbb{E}g_p(v^*)| + |\mathbb{E}g_p(v^*) - \mathbb{E}g_p(v)| \\ &\leq \frac{K}{4} + |g_p(v^*) - \mathbb{E}g_p(v^*)| + \frac{\varepsilon}{2}, \end{aligned} \tag{89}$$

where we have used the Lipschitz continuity of $g_p(v)$ and $\mathbb{E}g_p(v)$, as well as $\varepsilon \geq \delta_p$. Then $\sup_{v\in[0,2]} |g_p(v) - \mathbb{E}g_p(v)| \geq 2\varepsilon$, only if at least one of following holds: (i) $K \geq 2\varepsilon \geq 2\delta_p$, (ii) there exists a $k \in \{1, 2, \ldots, 8\}$, s.t., $|g_p(v_k) - \mathbb{E}g_p(v_k)| \geq \varepsilon$. Combining (87) and (88) and applying the union bound, we get for $\varepsilon \geq \delta_p$,

$$\mathbb{P}(\sup_{v\in[0,2]} |g_p(v) - \mathbb{E}g_p(v)| \geq 2\varepsilon) \leq 18e^{-\frac{p\varepsilon}{16\pi^2}}. \tag{90}$$

2) If $\varepsilon < \delta_p$, we construct an epsilon-net of $[0,2]$ formed by the following points: $v_k = 2k/\lceil\frac{8\delta_p}{\varepsilon}\rceil$, $k = 1, 2, \ldots, \lceil\frac{8\delta_p}{\varepsilon}\rceil$. In this case, for any $v \in [0,2]$, we have $|v - v^*| \leq \frac{\varepsilon}{8\delta_p}$. Then similar as previous argument, we have $\sup_{v\in[0,2]} |g_p(v) - \mathbb{E}g_p(v)| \geq 2\varepsilon$, only if at least one of following holds: (i) $g_p(v)$ is not $4\delta_p$-Lipschitz, (ii) there exists a $k \leq \lceil\frac{8\delta_p}{\varepsilon}\rceil$, s.t., $|g_p(v_k) - \mathbb{E}g_p(v_k)| \geq \varepsilon$. Combining (87) and (88) and applying the union bound, we get:

$$\mathbb{P}(\sup_{v\in[0,2]} |g_p(v) - \mathbb{E}g_p(v)| \geq 2\varepsilon) \leq \frac{16\delta_p}{\varepsilon} e^{-\frac{p\varepsilon^2}{16\pi^2\delta_p}}. \tag{91}$$

Combining (90) and (91), together with symmetry and the union bound, we directly get (35).

D. Proof of Lemma 1

The proof follows the CGMT framework [12], [13]. The optimization in (37) is equivalent to

$$Q_p(\theta) = p^{-\frac{3}{2}} \min_{\boldsymbol{x}\in[-1,1]^p} \max_{\boldsymbol{u}} \boldsymbol{u}^\mathsf{T} \begin{bmatrix} \sqrt{p}\boldsymbol{A} & -\widetilde{\boldsymbol{w}} \end{bmatrix} \begin{bmatrix} \boldsymbol{x} - \boldsymbol{\beta} \\ \sqrt{p(\theta + \sigma_p^2)} \end{bmatrix} - \frac{\sqrt{p}\|\boldsymbol{u}\|^2}{2}, \tag{92}$$

where $\widetilde{\boldsymbol{w}} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_n)$. The corresponding auxiliary problem (AO) of (92) is

$$\begin{aligned} Q_{\text{AO},p}(\theta) &= \min_{\boldsymbol{x}\in[-1,1]^p} \max_{\boldsymbol{u}} -\sqrt{\tfrac{\|\boldsymbol{x}-\boldsymbol{\beta}\|^2}{p} + \theta + \sigma_p^2}\, \tfrac{\boldsymbol{g}^\mathsf{T}\boldsymbol{u}}{p} + \tfrac{\|\boldsymbol{u}\|}{\sqrt{p}} \left[\tfrac{\boldsymbol{h}^\mathsf{T}(\boldsymbol{x}-\boldsymbol{\beta})}{p} + \tfrac{h_0\sqrt{\theta+\sigma_p^2}}{\sqrt{p}} \right] - \tfrac{\|\boldsymbol{u}\|^2}{2p} \\ &= \frac{1}{2} \left(\min_{\boldsymbol{x}\in[-1,1]^p} \sqrt{\tfrac{\|\boldsymbol{x}-\boldsymbol{\beta}\|^2}{p} + \theta + \sigma_p^2}\, \tfrac{\|\boldsymbol{g}\|}{\sqrt{p}} + \tfrac{\boldsymbol{h}^\mathsf{T}(\boldsymbol{x}-\boldsymbol{\beta})}{p} + \tfrac{h_0\sqrt{\theta+\sigma_p^2}}{\sqrt{p}} \right)_+^2. \end{aligned} \tag{93}$$

where $(x)_+ \overset{\text{def}}{=} \max\{x, 0\}$, $\boldsymbol{g} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_n), \boldsymbol{h} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_p)$, $h_0 \sim \mathcal{N}(0,1)$ and they are mutually independent.

Now we analyze the inline optimization problem of (93), which can be simplified as:

$$\phi(\theta,\boldsymbol{g},\boldsymbol{h}) = \min_{\boldsymbol{x}\in[-1,1]^p}\inf_{\tau>0}\sqrt{\delta_p}\left[\frac{\tau}{2}+\frac{\|\boldsymbol{x}-\boldsymbol{\beta}\|^2}{2\tau p}+\frac{\theta+\sigma_p^2}{2\tau}\right]\frac{\|\boldsymbol{g}\|}{\sqrt{n}}+\frac{\boldsymbol{h}^\intercal(\boldsymbol{x}-\boldsymbol{\beta})}{p}+\frac{h_0\sqrt{\theta+\sigma_p^2}}{\sqrt{p}} \tag{94}$$

$$= \underbrace{\inf_{\tau>0}\left[\frac{\tau\delta_p}{2}+\frac{\theta+\sigma_p^2}{2\tau}\right]\frac{\|\boldsymbol{g}\|}{\sqrt{n}}+\frac{1}{p}\sum_{i=1}^{p}v\left(h_i;\tau,\boldsymbol{g}\right)+\frac{h_0\sqrt{\theta+\sigma_p^2}}{\sqrt{p}}}_{F(\tau;\theta,\boldsymbol{g},\boldsymbol{h})}, \tag{95}$$

where in (95) we make a change of variable: $\frac{\tau}{\sqrt{\delta_p}}\to\tau$ and the parametric function $v\left(h;\tau,\boldsymbol{g}\right)$ is defined as:

$$v\left(h;\tau,\boldsymbol{g}\right)\overset{\text{def}}{=}\begin{cases}0 & h\geq 0,\\ -\frac{\tau\sqrt{n}}{2\|\boldsymbol{g}\|}h^2 & h\in[\frac{-2\|\boldsymbol{g}\|}{\tau\sqrt{n}},0),\\ 2\left(\frac{\|\boldsymbol{g}\|}{\tau\sqrt{n}}+h\right) & h<\frac{-2\|\boldsymbol{g}\|}{\tau\sqrt{n}}.\end{cases} \tag{96}$$

Denote $\tau^*_{\text{AO}}(\theta)$ as the optimal solution in (95). From (94) and the fact that we did a change of variable in (95), it can be seen $\tau^*_{\text{AO}}(\theta)=\sqrt{\frac{\|\boldsymbol{x}^*-\boldsymbol{\beta}\|^2}{p\delta_p}+\frac{\theta+\sigma_p^2}{\delta_p}}$. Therefore, for $\theta\in[0,1]$, $\tau^*_{\text{AO}}(\theta)\in\Omega(\sigma_p,\delta_p)$, where $\Omega(\sigma_p,\delta_p)\overset{\text{def}}{=}\left[\frac{\sigma_p}{\sqrt{\delta_p}},\frac{\sqrt{5+\sigma_p^2}}{\sqrt{\delta_p}}\right]$. Note that this is consistent with (82).

We now show objective function $F(\tau;\theta,\boldsymbol{g},\boldsymbol{h})$ in (95) converges to $F(\tau;\theta)\overset{\text{def}}{=}F_p(\tau;\theta+\sigma_p^2,\delta_p)$ with high probability over $\tau\in\Omega(\sigma_p,\delta_p)$. The first and third term in RHS of (95) is relatively easy to deal with. By the concentration of $\frac{\|\boldsymbol{g}\|}{\sqrt{n}}$ (e.g. [38, p.44]) and $\frac{h_0}{\sqrt{p}}$, there exists $c>0$, s.t., for any $\varepsilon>0$ and $\tau\in\Omega(\sigma_p,\delta_p)$,

$$\mathbb{P}\left(\left(\frac{\tau\delta_p}{2}+\frac{\theta+\sigma_p^2}{2\tau}\right)\left|\frac{\|\boldsymbol{g}\|}{\sqrt{n}}-1\right|>\sqrt{\delta_p}\varepsilon\right)\leq c\exp(-n\varepsilon^2/c) \tag{97}$$

and

$$\mathbb{P}\left(\left|\frac{h_0\sqrt{\theta+\sigma_p^2}}{\sqrt{p}}\right|>\varepsilon\right)\leq c\exp(-p\varepsilon^2/c). \tag{98}$$

Here in (97), we have used the fact that for $\tau\in\Omega(\sigma_p,\delta_p)$, $\frac{\tau\delta_p}{2}+\frac{\theta+\sigma_p^2}{2\tau}\leq C\sqrt{\delta_p}$, where C is some constant. For the second term, define the following function: $V(\boldsymbol{h};\tau,\boldsymbol{g})\overset{\text{def}}{=}\frac{\sum_{i=1}^p v(h_i;\tau,\boldsymbol{g})}{p}$, where $v\left(h;\tau,\boldsymbol{g}\right)$ is given in (96). We now show there exists $c>0$, s.t., for any $\varepsilon\geq 0$,

$$\mathbb{P}(|V(\boldsymbol{h};\tau,\boldsymbol{g})-f(\tau)|>\varepsilon)\leq c\exp(-p\varepsilon^2/c), \tag{99}$$

where

$$f(t)\overset{\text{def}}{=}-\frac{t}{4}+\frac{t}{2}\int_{2/t}^{\infty}\left(x-\frac{2}{t}\right)^2\Phi(dx). \tag{100}$$

First, note that for any fixed $\boldsymbol{g}$, $v\left(h;\tau,\boldsymbol{g}\right)$ is 2-Lipschitz continuous, so $V(\boldsymbol{h};\tau,\boldsymbol{g})$ is $\frac{2}{\sqrt{p}}$-Lipschitz continuous w.r.t. $\boldsymbol{h}$. Also we can verify that $\mathbb{E}_h v(h;\tau,\boldsymbol{g}) = f(\tau_{\boldsymbol{g}})$, with $\tau_{\boldsymbol{g}} \overset{\text{def}}{=} \frac{\tau\sqrt{n}}{\|\boldsymbol{g}\|}$. Then using Theorem 2.1 in [36, p.176], we have for any $\boldsymbol{g}$ and $\varepsilon > 0$,

$$\mathbb{P}(|V(\boldsymbol{h};\tau,\boldsymbol{g}) - f(\tau_{\boldsymbol{g}})| > \varepsilon) \leq 2\exp(-\tfrac{p\varepsilon^2}{2\pi^2}). \tag{101}$$

It can be checked that $f(t)$ in (100) satisfies $f(t) \in [-1, 0]$ for any $t > 0$. Combining this with (100) and (101), we know (99) holds for $\varepsilon > \frac{1}{2}$. On the other hand, by a direct differentiation, we have $f'(t) = -\frac{1}{4} + \frac{1}{2}\int_{2/t}^{\infty}\left(x^2 - \frac{4}{t^2}\right)\Phi(dx)$. It is not hard to verify $|f'(t)| \leq 1/4$, for all $t > 0$. Therefore, for any $\varepsilon \in (0, 1/2)$, on the event $E_\varepsilon = \left\{\left|\frac{\|\boldsymbol{g}\|}{\sqrt{n}} - 1\right| < \varepsilon\right\}$, which happens with probability $\mathbb{P}(E_\varepsilon) \geq 1 - ce^{-n\varepsilon^2/c}$, there exists $c > 0$, s.t., $|\tau_{\boldsymbol{g}} - \tau| \leq c\varepsilon$. As a result, there exists $c > 0$, s.t., for $\varepsilon \in (0, 1/2)$, $\mathbb{P}(|f(\tau_{\boldsymbol{g}}) - f(\tau)| > \varepsilon) \leq ce^{-n\varepsilon^2/c}$. This together with (101) implies there exists $c > 0$, s.t., for $\varepsilon \in (0, 1/2)$, inequality (99) still holds.

Combining (97) and (99), we get that there exists $c > 0$, s.t., for any $\varepsilon > 0$, $\tau \in \Omega(\sigma_p, \delta_p)$ and $\theta \in [0, 1]$,

$$\mathbb{P}\left(|F(\tau;\theta,\boldsymbol{g},\boldsymbol{h}) - F(\tau;\theta)| > \varepsilon\right) \leq ce^{-p\varepsilon^2/c}. \tag{102}$$

On the other hand, it can be verified from definition that there exists $C > 0$, s.t., $F(\tau;\theta,\boldsymbol{g},\boldsymbol{h})$ and $F(\tau;\theta)$ are both $C\delta_p$-Lipschitz over $\tau \in \Omega(\sigma_p, \delta_p)$. Then by a similar epsilon-net argument as in the proof of Proposition 8, we can get:

$$\mathbb{P}\Big(\sup_{\tau\in\Omega(\sigma_p,\delta_p)} |F(\tau;\theta,\boldsymbol{g},\boldsymbol{h}) - F(\tau;\theta)| > \varepsilon\Big) \leq \tfrac{c\sqrt{\delta_p}}{\varepsilon}e^{-p\varepsilon^2/c}. \tag{103}$$

Since $\phi(\theta,\boldsymbol{g},\boldsymbol{h}) = \min_{\tau\in\Omega(\sigma_p,\alpha_p)} F(\tau;\theta,\boldsymbol{g},\boldsymbol{h})$ and $\sqrt{2Q_p^*(\theta)} = \min_{\tau\in\Omega(\sigma_p,\alpha_p)} F(\tau;\theta)$, from (103) we know there exists $c > 0$, s.t., for any $\varepsilon > 0$,

$$\mathbb{P}\left(|\phi(\theta,\boldsymbol{g},\boldsymbol{h}) - \sqrt{2Q_p^*(\theta)}| > \varepsilon\right) \leq \tfrac{c\sqrt{\delta_p}}{\varepsilon}e^{-p\varepsilon^2/c}. \tag{104}$$

Since $\sqrt{2Q_{\text{AO},p}(\theta)} = \max\{\phi(\theta,\boldsymbol{g},\boldsymbol{h}), 0\}$, from (104) we have

$$\mathbb{P}\left(|\sqrt{2Q_{\text{AO},p}(\theta)} - \sqrt{2Q_p^*(\theta)}| > \varepsilon\right) \leq \tfrac{c\sqrt{\delta_p}}{\varepsilon}e^{-p\varepsilon^2/c}. \tag{105}$$

Taking into account the fact $Q_p^*(\theta) \leq C\delta_p$ (as shown in Appendix B), we can further obtain the following Bernstein's type inequality: there exists $c > 0$, s.t., for any $\varepsilon > 0$ and $\theta \in [0, 1]$,

$$\mathbb{P}\left(|Q_{\text{AO},p}(\theta) - Q_p^*(\theta)| > \varepsilon\right) \leq \frac{ce^{-p\min\{\frac{\varepsilon^2}{\delta_p},\varepsilon\}/c}}{\min\{\frac{\varepsilon}{\delta_p}, \sqrt{\frac{\varepsilon}{\delta_p}}\}}. \tag{106}$$

Then by CGMT (e.g., [33, Corollary 5.1]), (106) implies that there exists $c>0$, s.t.,

$$\mathbb{P}\left(|Q_p(\theta)-Q_p^*(\theta)|>\varepsilon\right)\leq \frac{ce^{-p\min\{\frac{\varepsilon^2}{\delta_p},\varepsilon\}/c}}{\min\{\frac{\varepsilon}{\delta_p},\sqrt{\frac{\varepsilon}{\delta_p}}\}}. \tag{107}$$

Finally, from (107) we know there exists $c>0$, s.t., for any $\eta>0$ and $\theta\in[0,1]$,

$$\begin{aligned}\mathbb{E}|Q_p(\theta)-Q_p^*(\theta)| &= \int_0^\infty \mathbb{P}\left(|Q_p(\theta)-Q_p^*(\theta)|\geq t\right)dt\\ &\leq \sqrt{\delta_p}\eta+\int_{\sqrt{\delta_p}\eta}^\infty \frac{c\delta_p}{t}e^{-\frac{pt^2}{c\delta_p}}dt+\int_{\sqrt{\delta_p}\eta}^\infty c\sqrt{\frac{\delta_p}{t}}e^{-\frac{pt}{c}}dt \qquad (108)\\ &\leq \sqrt{\delta_p}\eta+\frac{c^2\delta_p}{p}\left(\frac{e^{-p\eta^2/c}}{\eta^2}+\frac{e^{-\sqrt{\delta_p}p\eta/c}}{\sqrt{\eta}}\right). \qquad (109)\end{aligned}$$

Then for $\gamma>2$, letting $\eta=p^{-1/\gamma}$ in (109) and taking into account Assumption (A.4), we can get $\mathbb{E}|Q_p(\theta)-Q_p^*(\theta)|\leq cp^{-1/\gamma}$ for some $c>0$ and all the sufficiently large p. As a result,

$$|\overline{Q}_p(\theta)-Q_p^*(\theta)|\leq \mathbb{E}|Q_p(\theta)-Q_p^*(\theta)|\leq cp^{-1/\gamma}.$$

Since the constant c above does not depend on θ, we get (51).

E. Approximate k-wise Independence

1) $\left\{\boldsymbol{a}_i^\mathsf{T}\boldsymbol{u}_{\setminus i}^*\right\}_{i\in[k]}\overset{d}{\approx}\left\{\boldsymbol{a}_i^\mathsf{T}\boldsymbol{u}_{\setminus[k]}^*\right\}_{i\in[k]}$: We first prove that the joint distribution of $\left\{\boldsymbol{a}_i^\mathsf{T}\boldsymbol{u}_{\setminus i}^*\right\}_{i\in[k]}$ is close to $\left\{\boldsymbol{a}_i^\mathsf{T}\boldsymbol{u}_{\setminus[k]}^*\right\}_{i\in[k]}$. To prove this, we can show $\boldsymbol{a}_i^\mathsf{T}\boldsymbol{u}_{\setminus i}^*\approx\boldsymbol{a}_i^\mathsf{T}\boldsymbol{u}_{\setminus[k]}^*$, for any $i\in[k]$ and use the fact that $\boldsymbol{a}_i^\mathsf{T}\boldsymbol{u}_{\setminus i}^*$ and $\boldsymbol{a}_i^\mathsf{T}\boldsymbol{u}_{\setminus[k]}^*$ are in the same probability space.

Lemma 8: There exists $c>0$, s.t., for any $\varepsilon>0$ and $i=1,2,\ldots,p-1$,

$$\mathbb{P}\left(\left|\boldsymbol{a}_i^\mathsf{T}(\boldsymbol{u}_{\setminus[i]}^*-\boldsymbol{u}_{\setminus[i+1]}^*)\right|>\sqrt{\delta_p}\varepsilon\right)\leq ce^{-c^{-1}p\min\left\{\varepsilon^2,\varepsilon\right\}}. \tag{110}$$

Proof: To lighten notation, define $\Delta_{[i]}\overset{\text{def}}{=}\boldsymbol{u}_{\setminus[i]}^*-\boldsymbol{u}_{\setminus[i+1]}^*$. Denote the objective function in (66) as $L_{\setminus[i]}(\boldsymbol{u})$, (with k replaced by i here). By strong convexity of $L_{\setminus[i]}(\boldsymbol{u})$, we have

$$L_{\setminus[i]}(\boldsymbol{u}_{\setminus[i+1]}^*)-L_{\setminus[i]}(\boldsymbol{u}_{\setminus[i]}^*)\geq\tfrac{1}{2}\|\Delta_{[i]}\|^2. \tag{111}$$

and

$$\begin{aligned}L_{\setminus[i]}(\boldsymbol{u}_{\setminus[i+1]}^*)-L_{\setminus[i]}(\boldsymbol{u}_{\setminus[i]}^*) &= |\boldsymbol{a}_{i+1}^\mathsf{T}\boldsymbol{u}_{\setminus[i+1]}^*|-|\boldsymbol{a}_{i+1}^\mathsf{T}\boldsymbol{u}_{\setminus[i]}^*|-\Delta_{[i]}^\mathsf{T}\boldsymbol{a}_{i+1}\beta_{i+1}\\ &\quad+L_{\setminus[i+1]}(\boldsymbol{u}_{\setminus[i+1]}^*)-L_{\setminus[i+1]}(\boldsymbol{u}_{\setminus[i]}^*)\\ &\leq -\tfrac{1}{2}\|\Delta_{[i]}\|^2+2\|\Delta_{[i]}\|\cdot\|\boldsymbol{a}_{i+1}\|, \qquad (112)\end{aligned}$$

where we use the fact $|\beta_i| = 1$ and Cauchy-Schwartz inequality in the last step. From (111) and (112), we can get $\|\Delta_{[i]}\| \le 2\|\boldsymbol{a}_{i+1}\|$. Therefore, there exists $c > 0$, s.t., for any $\varepsilon, D > 0$,

$$\begin{aligned}\mathbb{P}\left(|\boldsymbol{a}_i^\intercal \Delta_{[i]}| > \sqrt{\delta_p}\varepsilon\right) &\le \mathbb{P}\left(|\boldsymbol{a}_i^\intercal \Delta_{[i]}| > \sqrt{\delta_p}\varepsilon \bigcap \|\Delta_{[i]}\| \le D\right) + \mathbb{P}\left(\|\Delta_{[i]}\| > D\right)\\ &\le \mathbb{P}\left(\left|\boldsymbol{a}_i^\intercal \tfrac{D\Delta_{[i]}}{\|\Delta_{[i]}\|}\right| > \sqrt{\delta_p}\varepsilon\right) + \mathbb{P}\left(\|\boldsymbol{a}_{i+1}\| > \tfrac{D}{2}\right)\\ &\le e^{-\frac{p\delta_p\varepsilon^2}{2D^2}} + ce^{-c^{-1}p\left(\frac{D}{2\sqrt{\delta_p}}-1\right)_+^2}, \end{aligned} \tag{113}$$

where $(x)_+ \overset{\text{def}}{=} \max\{x, 0\}$. Then by choosing $D \asymp \sqrt{\delta_p}$ for small ε and $D \asymp \sqrt{\delta_p\varepsilon}$ for large ε, we can get (110). ■

Lemma 9: There exists $c > 0$, s.t., for any $b_i \in \mathbb{R}$, $i = 1, 2, \ldots, k$ and $\varepsilon > 0$,

$$\mathbb{P}\left(\bigcap_{i=1}^k \left\{\boldsymbol{a}_i^\intercal \boldsymbol{u}^*_{\setminus i} \le b_i\right\}\right) \ge \mathbb{P}\left(\bigcap_{i=1}^k \left\{\boldsymbol{a}_i^\intercal \boldsymbol{u}^*_{\setminus [k]} \le b_i - \sqrt{\delta_p}\varepsilon\right\}\right) - ck^2 e^{-c^{-1}p\min\left\{\frac{\varepsilon^2}{k^2}, \frac{\varepsilon}{k}\right\}} \tag{114}$$

and

$$\mathbb{P}\left(\bigcap_{i=1}^k \left\{\boldsymbol{a}_i^\intercal \boldsymbol{u}^*_{\setminus i} \le b_i\right\}\right) \le \mathbb{P}\left(\bigcap_{i=1}^k \left\{\boldsymbol{a}_i^\intercal \boldsymbol{u}^*_{\setminus [k]} \le b_i + \sqrt{\delta_p}\varepsilon\right\}\right) + ck^2 e^{-c^{-1}p\min\left\{\frac{\varepsilon^2}{k^2}, \frac{\varepsilon}{k}\right\}}. \tag{115}$$

Proof: From Lemma 8, for any $k \in [p]$, there exists $c > 0$, s.t., for any $\varepsilon > 0$,

$$\begin{aligned}\mathbb{P}\left(\left|\boldsymbol{a}_1^\intercal(\boldsymbol{u}^*_{\setminus 1} - \boldsymbol{u}^*_{\setminus [k]})\right| > \sqrt{\delta_p}\varepsilon\right) &\le \sum_{i=1}^{k-1} \mathbb{P}\left(\left|\boldsymbol{a}_1^\intercal(\boldsymbol{u}^*_{\setminus [i]} - \boldsymbol{u}^*_{\setminus [i+1]})\right| > \frac{\sqrt{\delta_p}\varepsilon}{k-1}\right)\\ &\le cke^{-c^{-1}p\min\left\{\frac{\varepsilon^2}{k^2}, \frac{\varepsilon}{k}\right\}}.\end{aligned}$$

By the exchangeability of $\left\{\boldsymbol{a}_i^\intercal \boldsymbol{u}^*_{\setminus i}, \boldsymbol{a}_i^\intercal \boldsymbol{u}^*_{\setminus [k]}\right\}_{i\in[k]}$, we have for any $i \in [k]$, it holds that

$$\mathbb{P}\left(\left|\boldsymbol{a}_i^\intercal(\boldsymbol{u}^*_{\setminus i} - \boldsymbol{u}^*_{\setminus [k]})\right| > \sqrt{\delta_p}\varepsilon\right) \le cke^{-c^{-1}p\min\left\{\frac{\varepsilon^2}{k^2}, \frac{\varepsilon}{k}\right\}}.$$

Therefore, we have for any $\varepsilon > 0$,

$$\begin{aligned}\mathbb{P}\left(\bigcap_{i=1}^k \{\boldsymbol{a}_i^\intercal \boldsymbol{u}^*_{\setminus i} \le b_i\}\right) &= \mathbb{P}\left(\bigcap_{i=1}^k \{\boldsymbol{a}_i^\intercal \boldsymbol{u}^*_{\setminus [k]} \le b_i - \boldsymbol{a}_i^\intercal(\boldsymbol{u}^*_{\setminus i} - \boldsymbol{u}^*_{\setminus [k]})\}\right)\\ &\le \mathbb{P}\left(\bigcap_{i=1}^k \{\boldsymbol{a}_i^\intercal \boldsymbol{u}^*_{\setminus [k]} \le b_i + \sqrt{\delta_p}\varepsilon\}\right) + \mathbb{P}\left(\bigcup_{i=1}^k \{\left|\boldsymbol{a}_i^\intercal(\boldsymbol{u}^*_{\setminus i} - \boldsymbol{u}^*_{\setminus [k]})\right| > \sqrt{\delta_p}\varepsilon\}\right)\\ &\le \mathbb{P}\left(\bigcap_{i=1}^k \{\boldsymbol{a}_i^\intercal \boldsymbol{u}^*_{\setminus [k]} \le b_i + \sqrt{\delta_p}\varepsilon\}\right) + ck^2 e^{-c^{-1}p\min\left\{\frac{\varepsilon^2}{k^2}, \frac{\varepsilon}{k}\right\}},\end{aligned}$$

which is (114). The other direction (115) can be obtained in the same way. ■

2) $\left\{\boldsymbol{a}_i^\intercal \boldsymbol{u}^*_{\setminus[k]}\right\}_{i\in[k]} \overset{d}{\approx} \left\{\boldsymbol{a}_i^\intercal \widetilde{\boldsymbol{u}}_{\setminus[k]}\right\}_{i\in[k]}$*:* Next we show the joint distribution of $\left\{\boldsymbol{a}_i^\intercal \boldsymbol{u}^*_{\setminus[k]}\right\}_{i\in[k]}$ is close to $\left\{\boldsymbol{a}_i^\intercal \widetilde{\boldsymbol{u}}_{\setminus[k]}\right\}_{i\in[k]}$. First we show $\frac{\|\boldsymbol{u}^*_{\setminus[k]}\|}{\sqrt{p}} \approx \frac{\|\widetilde{\boldsymbol{u}}_{\setminus[k]}\|}{\sqrt{p}} = f_p$.

Lemma 10: When $k \le \frac{p}{2}$, there exist $C, c > 0$, s.t. for any $\varepsilon > 0$,

$$\mathbb{P}\left(\left|\frac{\|\boldsymbol{u}^*_{\setminus[k]}\|}{\sqrt{p}} - f_p\right| > \sqrt{\delta_p}\varepsilon\right) \le \frac{c\sqrt{\delta_p} e^{-c^{-1}n\left(\varepsilon - \frac{Ck}{p}\right)_+^2}}{\max\left\{\varepsilon - \frac{Ck}{p}, n^{-\frac{1}{2}}\right\}}. \tag{116}$$

Proof: By the definition of $\boldsymbol{u}^*_{\setminus[k]}$, we can get

$$\begin{aligned}\frac{\|\boldsymbol{u}^*_{\setminus[k]}\|}{\sqrt{p}} &= \frac{1}{\sqrt{p}} \min_{\boldsymbol{x}\in[-1,1]^{p-k}} \|\boldsymbol{A}_{\setminus[k]}\boldsymbol{x} - (\boldsymbol{A}_{\setminus[k]}\boldsymbol{\beta}_{\setminus[k]} + \boldsymbol{w})\| \\ &= \frac{\delta_p}{\delta_{p,\setminus[k]}} \cdot \frac{\min_{\boldsymbol{x}\in[-1,1]^{p-k}} \|\widetilde{\boldsymbol{A}}_{\setminus[k]}\boldsymbol{x} - (\widetilde{\boldsymbol{A}}_{\setminus[k]}\boldsymbol{\beta}_{\setminus[k]} + \widetilde{\boldsymbol{w}})\|}{\sqrt{p-k}},\end{aligned} \tag{117}$$

where $\delta_{p,\setminus[k]} = \frac{n}{p-k}$, i.e., the sampling ratio after removing k predictors, $\widetilde{\boldsymbol{A}}_{\setminus[k]} \overset{i.i.d.}{\sim} \mathcal{N}\left(0, \frac{1}{p-k}\right)$ and $\widetilde{\boldsymbol{w}} \sim \mathcal{N}\left(\boldsymbol{0}, \frac{\delta_{p,\setminus[k]}\sigma_p^2}{\delta_p}\boldsymbol{I}_n\right)$. Define

$$S_p^*(\delta) \overset{\text{def}}{=} \frac{\delta_p}{\delta} \min_{\tau>0} F_p\left(\tau; \frac{\delta\sigma_p^2}{\delta_p}, \delta\right), \tag{118}$$

where F_p is defined in (7). Similar to (104), we can get for $k \le \frac{p}{2}$, $\exists c > 0$, s.t., $\forall \varepsilon > 0$,

$$\mathbb{P}\left(\left|\frac{\|\boldsymbol{u}^*_{\setminus[k]}\|}{\sqrt{p}} - S^*_{p,\setminus[k]}\right| > \varepsilon\right) \le \frac{c\sqrt{\delta_p}}{\varepsilon} e^{-(p-k)\varepsilon^2/c}, \tag{119}$$

where $S^*_{p,\setminus[k]} \overset{\text{def}}{=} S_p^*(\delta_{p,\setminus[k]})$.

On the other hand, $|S^*_{p,\setminus[k]} - f_p|$ can be bounded as follows. From (118), we can show when $k \le \frac{p}{2}$, there exists $C > 0$, s.t., $\left|\frac{dS_p^*(\delta)}{d\delta}\right| \le \frac{C}{\sqrt{\delta_p}}$ for any $\delta \in [\delta_p, \delta_{p,\setminus[k]}]$. Since $f_p = S_p^*(\delta_p)$ and $S^*_{p,\setminus[k]} = S_p^*(\delta_{p,\setminus[k]})$, by the mean value theorem, we can get for $k \le \frac{p}{2}$, there exists $C > 0$, s.t.,

$$\left|S^*_{p,\setminus[k]} - f_p\right| \le \frac{Ck\sqrt{\delta_p}}{p}. \tag{120}$$

Now combining (119) , (120) and the condition $k \le p/2$, we can obtain (116). ∎

Based on Lemma 10, we can now show $\boldsymbol{a}_i^\intercal \boldsymbol{u}^*_{\setminus[k]} \approx \boldsymbol{a}_i^\intercal \widetilde{\boldsymbol{u}}_{\setminus[k]}$, if k is not too large.

Lemma 11: If $k \le \sqrt{p}$, then there exists $c > 0$, s.t., for any $\varepsilon > 0$ and $i \in [k]$,

$$\mathbb{P}\left(\left|\boldsymbol{a}_i^\intercal(\boldsymbol{u}^*_{\setminus[k]} - \widetilde{\boldsymbol{u}}_{\setminus[k]})\right| > \sqrt{\delta_p}\varepsilon\right) \le cp^{\frac{1}{2}} e^{-\sqrt{p}\varepsilon/c}. \tag{121}$$

Proof: Using (116) and following the similar steps as (113), we can get:

$$\mathbb{P}\left(|\boldsymbol{a}_i^\intercal(\boldsymbol{u}^*_{\setminus[k]} - \widetilde{\boldsymbol{u}}_{\setminus[k]})| > \sqrt{\delta_p}\varepsilon\right) \le Ce^{-\frac{p\varepsilon^2}{2D^2}} + \frac{C\sqrt{\delta_p} e^{-C^{-1}n\left(\frac{D}{\sqrt{p}} - \frac{Ck}{p}\right)_+^2}}{\max\left\{\frac{D}{\sqrt{p}} - \frac{Ck}{p}, n^{-\frac{1}{2}}\right\}}, \tag{122}$$

where C is some constant. Setting $D = p^{\frac{1}{4}}\varepsilon^{\frac{1}{2}}$ in (122), we can obtain (121). ■

Using Lemma 11, we can show that the joint distributions of $\{\boldsymbol{a}_i^\intercal \boldsymbol{u}^*_{\setminus[k]}\}_{i\in[k]}$ and $\{\boldsymbol{a}_i^\intercal \widetilde{\boldsymbol{u}}_{\setminus[k]}\}_{i\in[k]}$ are similar.

Lemma 12: If $k \le \sqrt{p}$, there exists $c > 0$, s.t., for any $b_i \in \mathbb{R}$, $i = 1, 2, \ldots, p$ and $\varepsilon > 0$,

$$\mathbb{P}(\boldsymbol{a}_i^\intercal \boldsymbol{u}^*_{\setminus[k]} \le b_i, i \in [k]) \le \mathbb{P}\left(\boldsymbol{a}_i^\intercal \widetilde{\boldsymbol{u}}_{\setminus[k]} \le b_i + \sqrt{\delta_p}\varepsilon, i \in [k]\right) + ckp^{\frac{1}{2}} e^{-\sqrt{p}\varepsilon/c} \tag{123}$$

and

$$\mathbb{P}(\boldsymbol{a}_i^\intercal \boldsymbol{u}^*_{\setminus[k]} \le b_i, i \in [k]) \ge \mathbb{P}\left(\boldsymbol{a}_i^\intercal \widetilde{\boldsymbol{u}}_{\setminus[k]} \le b_i - \sqrt{\delta_p}\varepsilon, i \in [k]\right) - ckp^{\frac{1}{2}} e^{-\sqrt{p}\varepsilon/c}. \tag{124}$$

Proof: The proof is similar to Lemma 9 and is omitted here. ■

3) Proof of Proposition 5: The proof follows directly Lemma 9 and Lemma 12.

4) Proof of Proposition 6: Letting $b_i = -A_p^*$ in (28), we have

$$\mathbb{P}\Big(\bigcap_{i=1}^{k}\Big\{\tilde{\beta}_i \ne \beta_i\Big\}\Big) \ge \Phi^k\left(-\tfrac{1+\sqrt{\delta_p}\varepsilon/A_p^*}{\tau_p}\right) - \Delta_{p,k} \tag{125}$$

$$\ge \Phi^k\left(-\tfrac{1}{\tau_p}\right)\left[1 - \tfrac{h(1/\tau_p)\sqrt{\delta_p}\varepsilon}{\tau_p A_p^*}\right]^k - \Delta_{p,k}, \tag{126}$$

where $h(x) = \frac{\varphi(-x)}{\Phi(-x)}$ is the so-called inverse Mills ratio. By (59), (126) and (132) given in Appendix F, there exists $c > 0$, s.t., for any $k \le \sqrt{p}$ and small enough $\varepsilon > 0$,

$$\mathbb{P}\Big(\bigcap_{i=1}^{k}\Big\{\tilde{\beta}_i \ne \beta_i\Big\}\Big) \ge \Phi^k\left(-\tfrac{1}{\tau_p}\right)\left(1 - \tfrac{ck\varepsilon}{\sigma_p^2}\right) - \Delta_{p,k}. \tag{127}$$

On the other hand, we can also get the similar bounds as (125) and (127) for the other direction.

Now consider the case $k \le p^{\frac{1}{8}}$. Accordingly, we set $\varepsilon = p^{-\frac{1}{4}}$. Then there exists $c, c' > 0$, s.t.,

$$\Delta_{p,k} \le c' p^{\frac{5}{8}} e^{-p^{1/4}/c'} \le c e^{-p^{1/4}/c}. \tag{128}$$

As a result, from (59), (125) and (128), if $k \le p^{\frac{1}{8}}$, there exists $c > 0$, s.t.,

$$\mathbb{P}\Big(\bigcap_{i=1}^{k}\Big\{\tilde{\beta}_i \ne \beta_i\Big\}\Big) \ge \Phi^k\Big(-\tfrac{1+cp^{-\frac{1}{4}}}{\tau_p}\Big) - ce^{-p^{1/4}/c}. \tag{129}$$

Meanwhile, we can also get for $\sigma_p^2 \ge \frac{c'}{\log^2 p}$,

$$\begin{aligned}
\mathbb{P}\Big(\bigcap_{i=1}^{k}\Big\{\tilde{\beta}_i \ne \beta_i\Big\}\Big) - \Phi^k\left(-\tfrac{1}{\tau_p}\right) &\overset{(a)}{\ge} -c\left[\Phi^k\left(-\tfrac{1}{\tau_p}\right) kp^{-\frac{1}{4}}\operatorname{polylog} p + e^{-p^{1/4}/c}\right] \\
&\overset{(b)}{\ge} -c\Phi^k\left(-\tfrac{1}{\tau_p}\right)\left[kp^{-\frac{1}{4}}\operatorname{polylog} p + \tfrac{e^{-p^{1/4}/c}}{\Phi^k\left(-\sqrt{\delta_p}/\sigma_p\right)}\right] \\
&\overset{(c)}{\ge} -\Phi^k\left(-\tfrac{1}{\tau_p}\right) kp^{-\frac{1}{4}}\operatorname{polylog} p,
\end{aligned} \tag{130}$$

where in step (a), we use (127), in step (b), we use (82) and step (c) follows from inequality (131) and conditions $k \leq p^{\frac{1}{8}}$ and $\sigma_p^2 \geq \frac{c'}{\log^2 p}$. The other directions of (129) and (130) can be derived similarly, which lead to (29) and (30).

F. Gaussian Tail Bounds

Here we gather some properties of the Gaussian tail bounds that will be used in our proof. Let $\Phi(x)$ and $\varphi(x)$ be the CDF and PDF of the standard Gaussian distribution, respectively. It is well known that (see [38, p.14] for a proof), for any $x > 0$,

$$\frac{1}{x} - \frac{1}{x^3} \leq m(x) \leq \frac{1}{x}, \tag{131}$$

where $m(x) \overset{\text{def}}{=} \frac{\Phi(-x)}{\varphi(-x)}$ is known as the Mills ratio. Correspondingly, the inverse Mills ratio is defined as $h(x) \overset{\text{def}}{=} 1/m(x)$. This provides us a way to approximate the tail probability $\Phi(-x)$ by $\varphi(x)$, which has an explicit form. In view of (82) and (131), there exists $M > 1$, s.t., for all $\eta \in [-1/2, 1/2]$,

$$\tfrac{1+\eta}{\tau_p} \leq h\Big(-\tfrac{1+\eta}{\tau_p}\Big) \leq \tfrac{M(1+\eta)}{\tau_p}. \tag{132}$$

Meanwhile, from (82) and (132), for all $\eta \in [-1/2, 1/2]$,

$$\Phi\left(-\tfrac{1+\eta}{\tau_p}\right) \leq \tfrac{1}{1+\eta}\sqrt{\tfrac{\sigma_p^2}{\delta_p - 1/2}}\tfrac{1}{\sqrt{2\pi}} e^{-\frac{(1+\eta)^2(\delta_p - 1/2)}{2\sigma_p^2}} \tag{133}$$

and

$$\Phi\left(-\tfrac{1+\eta}{\tau_p}\right) \geq \tfrac{1}{M(1+\eta)}\sqrt{\tfrac{\sigma_p^2}{\delta_p}}\tfrac{1}{\sqrt{2\pi}} e^{-\frac{(1+\eta)^2(\delta_p - 1/2 + v_p)}{2\sigma_p^2}}, \tag{134}$$

where $v_p = \int_{b_p}^{\infty} x^2 \Phi(dx)$, with $b_p = 2\sqrt{\frac{\delta_p - 1/2}{\sigma_p^2}}$.

G. An Auxiliary Result

Proposition 10: As $p \to \infty$, it holds that

$$\lim_{p\to\infty} \mathbb{P}(\widetilde{N}_e = 0) = \begin{cases} 1, & \liminf_{p\to\infty} \alpha_p > 1, \\ 0, & \limsup_{p\to\infty} \alpha_p < 1. \end{cases} \tag{135}$$

Proof: When $\liminf_{p\to\infty} \alpha_p > 1$, $\frac{4\sigma_p^2}{\delta_p - 1/2} \leq \frac{2}{\log p}$ for large enough p. Combining (29) and (133) in Appendix F gives us

$$\begin{aligned} \mathbb{E}\widetilde{N}_e &\leq Cp\Phi\left(-\tfrac{1+\eta}{\tau_p}\right) + cpe^{-\sqrt[4]{p}/c} \\ &\leq \tfrac{C}{1+\eta}\sqrt{\tfrac{2}{\log p}} p^{1-\alpha_p[1+o(\eta)]}, \end{aligned} \tag{136}$$

where $\eta = -cp^{-\frac{1}{4}}$ and C is some constant. Therefore, from (136) and Markov's inequality, $\lim_{p\to\infty}\mathbb{P}(\widetilde{N}_e \geq 1) = 0$.

When $\limsup_{p\to\infty}\alpha_p < 1$, then $\sigma_p \geq \frac{1}{\log p}$ for large enough p and we have

$$\begin{aligned}\mathbb{E}\widetilde{N}_e &\overset{\text{(a)}}{\geq} p\Phi\left(-\tfrac{1}{\tau_p}\right)\left(1-p^{-\frac{1}{4}}\operatorname{polylog} p\right)\\ &\overset{\text{(b)}}{\geq} \frac{2\left(1-p^{-\frac{1}{4}}\operatorname{polylog} p\right)}{M\log p}e^{-\frac{v_p\log^2 p}{2}}p^{1-\alpha_p},\end{aligned} \tag{137}$$

where step (a) follows from (30) and step (b) follows from (134) in Appendix F. In addition, it can be checked that v_p as defined in (82) satisfies $v_p \leq \frac{e^{-b_p^2/2}(b_p+2)^2}{2}$, where $b_p = \frac{2\sqrt{\delta_p - 1/2}}{\sigma_p}$. If $\alpha_p \in \left[\frac{1}{2}, 1\right)$, $b_p \geq \frac{2\log p}{\sqrt{\delta_p - 1/2}}$. Hence, there exists $C > 0$, such that for large enough p, $v_p \leq \frac{1}{p^C}$. Then from (137) we can get $\lim_{p\to\infty}\mathbb{E}\widetilde{N}_e = \infty$. If $\alpha_p < \frac{1}{2}$, since τ_p is strictly increasing with respect to σ_p^2 as shown in (81), by step (a) above, it still holds that $\lim_{p\to\infty}\mathbb{E}\widetilde{N}_e = \infty$.

We now prove that $\lim_{p\to\infty}\mathbb{P}(\widetilde{N}_e = 0) = 0$, when $\lim_{p\to\infty}\mathbb{E}\widetilde{N}_e = \infty$. The key lies in the approximate independence established in (30). First,

$$\begin{aligned}\operatorname{Var}\left(\widetilde{N}_e\right) &= \sum_{i=1}^{p}\operatorname{Var}\left(\mathbb{1}_{\tilde{\beta}_i\neq\beta_i}\right) + \sum_{i\neq j}\operatorname{Cov}\left(\mathbb{1}_{\tilde{\beta}_i\neq\beta_i}, \mathbb{1}_{\tilde{\beta}_j\neq\beta_j}\right)\\ &\leq \sum_{i=1}^{p}\mathbb{P}\left(\boldsymbol{a}_i^{\mathsf{T}}\boldsymbol{u}^*_{\setminus i} \leq -A_p^*\right)\\ &\quad + \sum_{i\neq j}\left|\mathbb{P}(\boldsymbol{a}_i^{\mathsf{T}}\boldsymbol{u}^*_{\setminus i} \leq -A_p^*, \boldsymbol{a}_j^{\mathsf{T}}\boldsymbol{u}^*_{\setminus j} \leq -A_p^*) - \mathbb{P}(\boldsymbol{a}_i^{\mathsf{T}}\boldsymbol{u}^*_{\setminus i} \leq -A_p^*)\mathbb{P}(\boldsymbol{a}_j^{\mathsf{T}}\boldsymbol{u}^*_{\setminus j} \leq -A_p^*)\right|\\ &\overset{\text{(a)}}{\leq} \mathbb{E}\widetilde{N}_e\left(1+p^{-\frac{1}{4}}\operatorname{polylog} p\right) + (\mathbb{E}\widetilde{N}_e)^2 p^{-\frac{1}{4}}\operatorname{polylog} p,\end{aligned} \tag{138}$$

where in step (a), we have used (30), with $k = 1, 2$ and also (136). Let $\mathbb{P}(\widetilde{N}_e = 0) = 1 - q_p$, $q_p \in [0, 1]$. For any p, $\mathbb{E}\left(\widetilde{N}_e \mid \widetilde{N}_e > 0\right) = \frac{\mathbb{E}\widetilde{N}_e}{q_p}$ and hence $\mathbb{E}\left(\widetilde{N}_e^2 \mid \widetilde{N}_e > 0\right) \geq \left(\frac{\mathbb{E}\widetilde{N}_e}{q_p}\right)^2$, which indicates that $q_p \geq \frac{(\mathbb{E}\widetilde{N}_e)^2}{(\mathbb{E}\widetilde{N}_e)^2 + \operatorname{Var}(\widetilde{N}_e)}$. This combined with (138) and $\lim_{p\to\infty}\mathbb{E}\widetilde{N}_e = \infty$ leads to: $\lim_{p\to\infty} q_p = 1$. Therefore, we conclude that $\lim_{p\to\infty}\mathbb{P}(\widetilde{N}_e = 0) = 0$. ■

H. Proof of Proposition 7

If $\alpha_p \geq 2$, from (136) and (133), we know there exists $c > 0$, s.t., $\lambda_p \leq cp^{-\frac{1}{2}}$ and $\mathbb{E}\widetilde{N}_e \leq cp^{-\frac{1}{2}}$. Hence, $d_{\text{TV}}(\widetilde{N}_e, \mathscr{P}_{\lambda_p})$ can be bounded as:

$$d_{\text{TV}}(\widetilde{N}_e, \mathscr{P}_{\lambda_p}) \leq \frac{1}{2}\left|\mathbb{P}\left(\widetilde{N}_e = 0\right) - e^{-\lambda_p}\right| + \frac{1}{2}\mathbb{P}\left(\widetilde{N}_e \geq 1\right) + \frac{1}{2}\left(1 - e^{-\lambda_p}\right) \leq 2cp^{-\frac{1}{2}}.$$

On the other hand, if $\alpha_p < 2$, then for large enough p, it holds that $\sigma_p \geq \frac{1}{\log p}$. Choose L in (73) to be $L = \lfloor 5 \log p \rfloor$. Without loss of generality, assume $L-k$ is odd (otherwise we add L by 1). Then from Bonferroni's inequality (73), for $k \leq \lfloor \log p \rfloor$,

$$\begin{aligned}
\mathbb{P}\left(\widetilde{N}_e = k\right) &\leq \sum_{m=0}^{L-1-k} \tfrac{(-1)^m}{k!m!} p^{m+k} S_{[m+k]} \\
&\overset{(a)}{\leq} \sum_{m=0}^{L-1-k} \tfrac{(-1)^m}{k!m!} p^{m+k} \Phi^{m+k}\left(-\tfrac{1}{\tau_p}\right) + \sum_{m=0}^{L-1-k} \frac{p^{m+k}\Phi^{m+k}\left(-\frac{1}{\tau_p}\right)}{k!m!} p^{-1/4} L_p \\
&\overset{(b)}{\leq} \tfrac{\lambda_p^k}{k!} e^{-\lambda_p} \left(1 + \left(\tfrac{\lambda_p e}{L-k}\right)^{L-k} e^{\lambda_p - 1}\right) + \tfrac{\lambda_p^k}{k!} e^{\lambda_p} p^{-1/4} L_p \\
&\overset{(c)}{\leq} \frac{\lambda_p^k}{k!} e^{-\lambda_p} \left[1 + \left(\tfrac{C}{\log^2 p}\right)^{\log p} + p^{-1/5} L_p\right].
\end{aligned} \tag{139}$$

Here, L_p is the shorthand notation for a term of order $\mathcal{O}(\operatorname{polylog} p)$ and C is some constant, step (a) follows from (30), in step (b) we use Taylor approximation and inequality $n! \geq e\left(\frac{n}{e}\right)^n$ and step (c) follows from conditions $L = \lfloor 5 \log p \rfloor$, $k \leq \lfloor \log p \rfloor$ and $\limsup_{p\to\infty} \frac{\lambda_p}{\sqrt{\log p}} < \infty$. In a similar manner, for the other direction, we can also obtain

$$\mathbb{P}\left(\widetilde{N}_e = k\right) \geq \frac{\lambda_p^k}{k!} e^{-\lambda_p} \left[1 - \left(\tfrac{C}{\log^2 p}\right)^{\log p} - p^{-1/5} L_p\right]. \tag{140}$$

By (139) and (140), for $k \leq \lfloor \log p \rfloor$,

$$\left|\mathbb{P}\left(\widetilde{N}_e = k\right) - \tfrac{\lambda_p^k}{k!} e^{-\lambda_p}\right| \leq \tfrac{\lambda_p^k}{k!} e^{-\lambda_p} p^{-1/5} L_p. \tag{141}$$

Then $d_{\mathrm{TV}}(\widetilde{N}_e, \mathscr{P}(\lambda_p))$ can be bounded as:

$$\begin{aligned}
d_{\mathrm{TV}}(\widetilde{N}_e, \mathscr{P}(\lambda_p)) &\leq \tfrac{1}{2} \sum_{k=0}^{\lfloor \log p \rfloor} \left|\mathbb{P}(\widetilde{N}_e = k) - \tfrac{\lambda_p^k e^{-\lambda_p}}{k!}\right| + \tfrac{1}{2} \sum_{k=\lfloor \log p \rfloor + 1}^{\infty} \mathbb{P}(\widetilde{N}_e = k) + \sum_{k=\lfloor \log p \rfloor + 1}^{\infty} \tfrac{\lambda_p^k e^{-\lambda_p}}{2k!} \\
&\overset{(a)}{\leq} \tfrac{p^{-1/5} L_p}{2} \sum_{k=0}^{\lfloor \log p \rfloor} \tfrac{\lambda_p^k e^{-\lambda_p}}{k!} + \tfrac{1}{2}\left[1 - \sum_{k=0}^{\lfloor \log p \rfloor} \tfrac{\lambda_p^k e^{-\lambda_p}}{k!}(1 - p^{-\frac{1}{5}} L_p)\right] + \sum_{k=\lfloor \log p \rfloor + 1}^{\infty} \tfrac{\lambda_p^k e^{-\lambda_p}}{2k!} \\
&\overset{(b)}{\leq} p^{-1/5} L_p,
\end{aligned}$$

where in step (a) we use (141), in step (b) we use Chernoff's bound for the tail probability of Poisson random variables [38, p.20]: for $X \sim \mathscr{P}(\lambda), k > \lambda$, $\mathbb{P}(X > k) \leq e^{-\lambda}\left(\frac{e\lambda}{k}\right)^k$ and the condition that $\limsup_{p\to\infty} \frac{\lambda_p}{\sqrt{\log p}} < \infty$.

REFERENCES

[1] A. J. Viterbi, *CDMA: Principles of Spread Spectrum Communication*. Addison Wesley Longman Publishing Co., Inc., 1995.

[2] F. Rusek, D. Persson, B. K. Lau, E. G. Larsson, T. L. Marzetta, O. Edfors, and F. Tufvesson, "Scaling up MIMO: Opportunities and challenges with very large arrays," *IEEE Signal Processing Magazine*, vol. 30, no. 1, pp. 40–60, 2012.

[3] A. K. Das and S. Vishwanath, "On finite alphabet compressive sensing," in *2013 IEEE Int. Conf. on Acoustics, Speech and Signal Processing*. IEEE, 2013, pp. 5890–5894.

[4] A. Aissa-El-Bey, D. Pastor, S. M. A. Sbai, and Y. Fadlallah, "Sparsity-based recovery of finite alphabet solutions to underdetermined linear systems," *IEEE Trans. Inf. Theory*, vol. 61, no. 4, pp. 2008–2018, 2015.

[5] J.-H. Ahn, "Compressive sensing and recovery for binary images," *IEEE Trans. Image Process.*, vol. 25, no. 10, pp. 4796–4802, 2016.

[6] M. Pohst, "On the computation of lattice vectors of minimal length, successive minima and reduced bases with applications," *ACM Sigsam Bulletin*, vol. 15, no. 1, pp. 37–44, 1981.

[7] M. Grötschel, L. Lovász, and A. Schrijver, *Geometric Algorithms and Combinatorial Optimization*. Springer Science & Business Media, 2012, vol. 2.

[8] C. Jeon, R. Ghods, A. Maleki, and C. Studer, "Optimality of large MIMO detection via approximate message passing," in *2015 IEEE Int. Symp. on Inf. Theory (ISIT)*. IEEE, 2015, pp. 1227–1231.

[9] B. Hassibi, M. Hansen, A. G. Dimakis, H. A. J. Alshamary, and W. Xu, "Optimized Markov chain Monte Carlo for signal detection in MIMO systems: An analysis of the stationary distribution and mixing time," *IEEE Trans. Signal Process.*, vol. 62, no. 17, pp. 4436–4450, 2014.

[10] Z.-Q. Luo, W.-K. Ma, A. M.-C. So, Y. Ye, and S. Zhang, "Semidefinite relaxation of quadratic optimization problems," *IEEE Signal Processing Magazine*, vol. 27, no. 3, pp. 20–34, 2010.

[11] P. H. Tan, L. K. Rasmussen, and T. J. Lim, "Box-constrained maximum-likelihood detection in CDMA," in *2000 Int. Zurich Seminar on Broadband Communications. Accessing, Transmission, Networking. Proceedings (Cat. No. 00TH8475)*. IEEE, 2000, pp. 55–62.

[12] C. Thrampoulidis, E. Abbasi, W. Xu, and B. Hassibi, "BER analysis of the box relaxation for BPSK signal recovery," in *2016 IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2016, pp. 3776–3780.

[13] C. Thrampoulidis, W. Xu, and B. Hassibi, "Symbol error rate performance of box-relaxation decoders in massive MIMO," *IEEE Trans. Signal Process.*, vol. 66, no. 13, pp. 3377–3392, 2018.

[14] V. Chandrasekaran, B. Recht, P. A. Parrilo, and A. S. Willsky, "The convex geometry of linear inverse problems," *Foundations of Computational Mathematics*, vol. 12, no. 6, pp. 805–849, 2012.

[15] N. El Karoui, "On the impact of predictor geometry on the performance on high-dimensional ridge-regularized generalized robust regression estimators," *Probability Theory and Related Fields*, vol. 170, no. 1-2, pp. 95–175, 2018.

[16] P. Luo and K. M. Wong, "Cavity approach to noisy learning in nonlinear perceptrons," *Physical Review E*, vol. 64, no. 6, p. 061912, 2001.

[17] M. Ramezanali, P. P. Mitra, and A. M. Sengupta, "The cavity method for analysis of large-scale penalized regression," *arXiv:1501.03194*, 2015.

[18] T. Tanaka, "A statistical-mechanics approach to large-system analysis of CDMA multiuser detectors," *IEEE Trans. Inf. Theory*, vol. 48, no. 11, pp. 2888–2910, 2002.

[19] Y. Kabashima, T. Wadayama, and T. Tanaka, "A typical reconstruction limit for compressed sensing based on ℓ_p-norm minimization," *Journal of Statistical Mechanics: Theory and Experiment*, vol. 2009, no. 09, p. L09003, 2009.

[20] L. Zdeborová and F. Krzakala, "Statistical physics of inference: Thresholds and algorithms," *Advances in Physics*, vol. 65, no. 5, pp. 453–552, 2016.

[21] M. Bayati and A. Montanari, "The dynamics of message passing on dense graphs, with applications to compressed sensing," *IEEE Trans. Inf. Theory*, vol. 57, no. 2, pp. 764–785, 2011.

[22] D. L. Donoho, A. Maleki, and A. Montanari, "Message-passing algorithms for compressed sensing," *Proceedings of the National Academy of Sciences*, vol. 106, no. 45, pp. 18 914–18 919, 2009.

[23] S. Rangan, P. Schniter, and A. K. Fletcher, "Vector approximate message passing," *IEEE Trans. Inf. Theory*, vol. 65, no. 10, pp. 6664–6684, 2019.

[24] M. Stojnic, "A framework to characterize performance of LASSO algorithms," *arXiv:1303.7291*, 2013.

[25] C. Thrampoulidis, E. Abbasi, and B. Hassibi, "Precise error analysis of regularized M-estimators in high-dimensions," *IEEE Trans. Inf. Theory*, 2018.

[26] D. Amelunxen, M. Lotz, M. B. McCoy, and J. A. Tropp, "Living on the edge: Phase transitions in convex programs with random data," *Information and Inference: A Journal of the IMA*, vol. 3, no. 3, pp. 224–294, 2014.

[27] M. J. Wainwright, "Sharp thresholds for high-dimensional and noisy sparsity recovery using ℓ_1-constrained quadratic programming (Lasso)," *IEEE Trans. Inf. Theory*, vol. 55, no. 5, pp. 2183–2202, 2009.

[28] G. David and Z. Ilias, "High dimensional regression with binary coefficients. Estimating squared error and a phase transtition," in *Conference on Learning Theory*, 2017, pp. 948–953.

[29] G. Reeves, J. Xu, and I. Zadik, "The all-or-nothing phenomenon in sparse linear regression," in *Conference on Learning Theory*, 2019, pp. 2652–2663.

[30] J. Barbier and N. Macris, "0-1 phase transitions in sparse spiked matrix estimation," *arXiv:1911.05030*, 2019.

[31] E. Abbe, A. S. Bandeira, and G. Hall, "Exact recovery in the stochastic block model," *IEEE Trans. Inf. Theory*, vol. 62, no. 1, pp. 471–487, 2015.

[32] M. Sion, "On general minimax theorems," *Pacific Journal of Mathematics*, vol. 8, no. 1, pp. 171–176, 1958.

[33] L. Miolane and A. Montanari, "The distribution of the Lasso: Uniform control over sparse balls and adaptive parameter tuning," *arXiv:1811.01212*, 2018.

[34] W. Feller, *An Introduction to Probability Theory and Its Applications*. John Wiley & Sons, 1968, vol. 1.

[35] J.-B. Hiriart-Urruty and C. Lemaréchal, *Convex Analysis and Minimization Algorithms II: Advanced Theory and Bundle Methods*. Springer Science & Business Media, 2013.

[36] G. Pisier, "Probabilistic methods in the geometry of Banach spaces," in *Probability and Analysis*. Springer, 1986, pp. 167–241.

[37] B. Laurent and P. Massart, "Adaptive estimation of a quadratic functional by model selection," *Annals of Statistics*, pp. 1302–1338, 2000.

[38] R. Vershynin, *High-dimensional Probability: An Introduction with Applications in Data Science*. Cambridge University Press, 2018.