Majority Vote for Distributed Differentially Private Sign Selection

Weidong Liu*, Jiyuan Tu†, Xiaojun Mao‡, and Xi Chen §¶

Abstract

Privacy-preserving data analysis has become more prevalent in recent years. In this study, we propose a distributed group differentially private Majority Vote mechanism, for the sign selection problem in a distributed setup. To achieve this, we apply the iterative peeling to the stability function and use the exponential mechanism to recover the signs. For enhanced applicability, we study the private sign selection for mean estimation and linear regression problems, in distributed systems. Our method recovers the support and signs with the optimal signal-to-noise ratio as in the non-private scenario, which is better than contemporary works of private variable selections. Moreover, the sign selection consistency is justified by theoretical guarantees. Simulation studies are conducted to demonstrate the effectiveness of the proposed method.

Keywords: Majority voting; divide-and-conquer; privacy; sign selection

1 Introduction

Recently, large amounts of sensitive data have been collected in a distributed manner. While one wants to extract more accurate statistical information from distributed datasets, critical awareness is required toward the probability of possible sensitive personal information leakages, during the learning process. This calls for the study of distributed learning under privacy constraints (Pathak et al., 2010; Hamm et al., 2016; Jayaraman et al., 2018).

In this study, we consider the following distributed sign recovery problem. Suppose that all $N$ data points $X = \{X_i\}_{i=1}^N$ are stored in a distributed system, that consists of $m$ local machines $\mathcal{H}_j$ for

---

*School of Mathematical Sciences and MoE Key Lab of Artificial Intelligence, Shanghai Jiao Tong University, Shanghai 200240, China.
†School of Mathematics, Shanghai University of Finance and Economics, Shanghai 200433, China.
‡School of Mathematical Sciences and Ministry of Education Key Laboratory of Scientific and Engineering Computing, Shanghai Jiao Tong University, Shanghai 200240, China.
§Stern School of Business, New York University, NY 10012, USA.
¶Xiaojun Mao and Xi Chen are the co-corresponding authors.
1 ≤ j ≤ m, and each local machine has n = N/m observations. We focus on estimating the nonzero positions and signs of a p-dimensional population parameter θ∗, with distributed group differential privacy (DGDP, see Definition 1). Let sgn(θ∗) = (sgn(θ∗₁), . . . , sgn(θ∗ₚ)), where each of its elements takes a value in {1, −1, 0}. In non-distributed and non-privacy settings, estimating sgn(θ∗), which is usually referred to as a sign/variable selection problem, is an important topic in high-dimensional statistics. Many sign/variable selection problems, including estimating the location of the nonzero elements in mean vectors or covariance matrices, support recovery of regression coefficients, and structural estimation of graphical models, have been well studied in the statistical community (see, e.g., Butucea et al. (2018, 2023); Tibshirani (1996); Miller (2002); Meinshausen and Bühlmann (2006); Karouli (2008); Bickel and Levina (2008); Cai and Liu (2011); Cai et al. (2011)). The sign selection problem also finds applications in many fields, including anomaly detection, medical imaging, and genomics (Fan and Li, 2001; Cai et al., 2014). There also exist several well-developed statistical tools such as the thresholding technique, ℓ₁-regularization, and multiple testing to tackle the sign/variable selection problems. However, these problems become more challenging under differential privacy, without their in-depth exploration in the existing literature.

First, we define DGDP for distributed learning. In the privacy literature, differential privacy (DP), first proposed by Dwork et al. (2006), has been the most widely adopted definition of privacy tailored to statistical data analysis. Note that the (ϵ, δ)-differential privacy ((ϵ, δ)-DP) in Dwork et al. (2006), is designed to protect the privacy between adjacent datasets with a Hamming distance of one, indicating two datasets differing only by one sample (see, e.g., Lei (2011)). In distributed learning, a local machine needs to protect the privacy of all its data points. It is natural to extend the standard DP to distributed group differential privacy as follows.

**Definition 1.** (Distributed Group Differential Privacy) In a distributed system, assume that the entire dataset is stored in machines H₁, . . . , Hₘ. A randomized algorithm A : Xᴺ → Θ is (ϵ, δ)-distributed group differentially private ((ϵ, δ)-DGDP) if for any pair of datasets X ∈ Xᴺ and X′ ∈ Xᴺ, whose elements are the same except for one local machine, the following holds

\[ \Pr\{A(X) \in U\} \leq e^ϵ \cdot \Pr\{A(X') \in U\} + δ, \]

for every measurable subset U ⊆ Θ.

Note that, when the sample size of each local machine is n = 1, the DGDP reduces to the classical (ϵ, δ)-DP. Although there is in-depth literature on estimating the parameter θ∗ under the DP constraint, an efficient procedural design for DGDP sign selection in distributed systems is highly nontrivial. However, there are three major challenges in this regard. First, naively adding noise (a common DP technique) to the output of a non-private sign/variable selection method, makes it difficult to maintain the desired sign selection consistency. Second, most of the existing
methods for parameter estimation with DP, are usually inoperative for sign selection. The essential
difference between parameter estimation and sign selection is well known. The latter is closely
related to multiple testing and top-\(k\) selection. To the best of our knowledge, there is insufficient
literature on DGDP multiple testing in distributed systems. On the other hand, the top-\(k\) selection
aims to select the largest \(k\) elements among the \(p\) given values, under the DP constraint (see, e.g.,
Bafna and Ullman (2017); Steinke and Ullman (2017); Dwork et al. (2021); Qiao et al. (2021)).
However, top-\(k\) selection does not imply sign consistency, since obtaining the prior information
about the number of non-zero elements \(s\) is impractical, and it is impossible to set \(k = s\). Third,
sign selection is relatively easy with a large signal-to-noise ratio (SNR) of the nonzero signals.
However, when the SNR is close to the minimax lower bound (typically of the order \(O(\sqrt{\log p/N})\)),
it complicates the theoretical analysis for sign selection consistency. As far as we know, for many
important statistical problems including support recovery of linear regression under DP, no method
attains the minimax lower bound even in a non-distributed setting.

In this paper, we address the above challenges by developing a general Majority Vote mechanism
for sign selection problems, which is particularly adaptive to the distributed system. Assume that
\(m\) parties vote to make a decision. Each party can vote for positive (1), negative (-1), and null
(0). The positive (negative) decision will be made only when there are more than half of the
votes for it; otherwise, it returns null. We leverage this Majority Vote mechanism to sign recovery
problems in a distributed setup. More specifically, let each local machine obtain an initial sign
vector based on its local samples and transmits the initial sign vector to a trustworthy server. By
carefully constructing the stability function using the Majority Vote mechanism and combining it
with the peeling technique and exponential mechanism in McSherry and Talwar (2007), we develop
a Majority Vote mechanism that protects distributed group differential privacy (DGDP). The server
then generates the result based on this DGDP algorithm. Our method has the following advantages.
First, every worker machine can conveniently apply classical techniques such as thresholding and/or
Lasso to obtain an initial estimate, and every component only takes value in \(\{1, -1, 0\}\). Since every
worker machine only needs to transmit the sign vector, the privacy of data on each machine is
protected to a certain degree. Second, the proposed method only incurs very low communication
costs as it only transmits the sign vector. More contributions to this work are summarized below.

- To the best of our knowledge, we proposed the first distributed group differential privacy-
  aware sign selection algorithm in a distributed system. Moreover, the proposed method
  guarantees sign selection consistency.

- For both the sparse mean estimation problem and the sparse regression problem, our
  proposed method allows the minimal signal to have the order \(O(\sqrt{\log p/N})\), which meets the
  optimal “beta-min” condition (Wainwright, 2009) in the non-private case. In addition, the
  DP literature’s often assumed the boundedness condition, is not required by our theory for
data samples.

- Our Majority Vote mechanism and DGDP guarantee are substantially general, and do not assume any special underlying statistical model. Thus, in addition to the mean model and linear regression, they can be further applied to many other sign selection problems, such as sparse Gaussian graphical model estimation.

1.1 Related Works

Variable/sign selection and sparse parameter estimation have been fundamental problems in statistics. In recent years, significant efforts have been made to extend sparse learning/variable selection methods to a distributed setup. Several distributed algorithms, such as the distributed version of the proximal gradient method (Shi et al., 2015), alternating direction method of multipliers (Boyd et al., 2011), approximate Newton method (Wang et al., 2017), and the Frank-Wolfe Algorithm (Bellet et al., 2015) for sparse learning, have been proposed one after another. Moreover, in Zhu et al. (2021), the authors introduced the distributed Bayesian information criterion for variable selection. Furthermore, in Battey et al. (2018) and Lee et al. (2017), the authors proposed aggregating local debiased Lasso estimators in one shot to achieve near-optimal statistical accuracy. On the contrary, our method directly aggregates local Lasso estimators (instead of debiased Lasso estimators) and achieves the near-optimal minimal signal strength condition, which is a novel contribution to the literature, despite privacy concerns.

In the past decade, differential privacy has gained significant attention in computer science and statistics. Many algorithms have been developed to deal with various problems, such as empirical risk minimization (Bassily et al., 2014), principal component analysis (Dwork et al., 2014), demand learning (Chen et al., 2022; Chen et al., 2021) and deep learning (Bu et al., 2020). In addition to the classical differential privacy framework, several proposed privacy notions help adapt to different situations. For example, group DP (Dwork and Roth, 2014), local DP (Wasserman and Zhou, 2010; Duchi et al., 2018), Gaussian DP (Dong et al., 2022), concentrated DP (Bun and Steinke, 2016), Rényi DP (Mironov, 2017), and on-average KL-privacy (Wang et al., 2016). This study proposes the concept of distributed group DP, which is designed for distributed setup. The group structure in the distributed system is predetermined. This differs from the group DP in Dwork and Roth (2014), where the DP is relative to arbitrary subgroups of a given size.

There are many works that study differentially private sparse regression problems, such as Jain and Thakurta (2014); Talwar et al. (2015); Kasiviswanathan and Jin (2016); Cai et al. (2021); Wang and Xu (2019). However, most of these studies considered minimizing the loss function or proving $\ell_2$-consistency. It is not direct to obtain the model selection results from these works. To the best of our knowledge, only three works consider DP variable selection for linear regression in a non-distributed setting: Kifer et al. (2012), Thakurta and Smith (2013), and Lei et al. (2018).
We perform a more detailed comparison between these studies and ours. In Kifer et al. (2012), two algorithms were proposed. One is based on an exponential mechanism (McSherry and Talwar, 2007). Their algorithm requires solving many optimizations constrained in any $s$-sparse subspace. As pointed out in Kifer et al. (2012), this algorithm is computationally inefficient. The other used a resample-and-aggregate method (Nissim et al., 2007; Smith, 2011), which directly counts the number of nonzero elements that are initially estimated from $m$ block of subsamples and applies Peeling based on this counting. Their algorithm ignores the sign information from the initial estimators, which is useful for identifying the zero positions. Moreover, their method does not result in sign selection consistency and requires that the nonzero elements have magnitudes larger than $O(\sqrt{s \log p}/N^{1/4})$, which is not minimax optimal. In Thakurta and Smith (2013), the authors defined two concepts of stability and proposed a PTR-based mechanism for variable selection. This method has a nontrivial probability of outputting the null (no result), which is undesirable in practice. In theory, it requires the boundness of the true parameter $\theta^*$ and the covariates $X$, which is not needed in our method. Moreover, they require the signals to be larger than $O(\sqrt{s \log p}/N)$, which is not minimax optimal. Furthermore, their algorithm was designed for differential privacy rather than DGDP. In Lei et al. (2018), the authors proposed the use of the Akaike or Bayesian information criterion, in conjunction with the exponential mechanism, to choose the proper model. However, their method traverses all possible models and causes a heavy computational burden.

Variable selection is also related to multiple testing, which can be used to find significant signals while controlling the false discovery rate (FDR). Dwork et al. (2021) proposed a differentially private procedure to control the FDR in multiple hypothesis testing. However, it cannot be applied directly to sign selection in a distributed group differential privacy setting.

Our method also shares a similar spirit with the subsample-and-aggregate (SA) methods in the DP literature (Nissim et al., 2007; Smith, 2011; Kifer et al., 2012). Like SA methods, we compute noiseless results for each block and then aggregate them in a DP manner. However, our method differs from existing SA methods in several aspects. In our model, the data are already distributed, and our method ensures the privacy of each entire block, as opposed to existing SA methods that only protect the privacy of each element; SA was originally proposed in Nissim et al. (2007) to compute the smooth sensitivity of statistics, which has a different motivation from our method; In Smith (2011), SA was advocated to achieve the same efficiency as non-private statistics. However, their method is limited to low-dimensional statistics and does not extend well to high-dimensional problems like model selection. In Kifer et al. (2012), SA was employed for model selection with a voting mechanism as the aggregator. In contrast, our method uses a different voting rule, and our theoretical results are much stronger than theirs. For a comprehensive comparison of theoretical results, please refer to the preceding two paragraphs.
1.2 Paper Organization and Notations

The remainder of this paper is organized as follows. In Section 2, we introduce a general Majority Vote procedure and develop the DGDP algorithm. In Section 3, we apply this to the private sign recovery problem for sparse mean estimation, in a distributed system. The theoretical results on sign selection consistency are provided. In Section 4, we study the sign recovery problem for sparse linear regression and simultaneously state the sign selection consistency results. The outcomes of simulation studies are presented in Section 5, to demonstrate the proposed method’s effectiveness. Finally, concluding remarks are provided in Section 6. All proofs of the theory are relegated to the Appendix.

For every vector $v = (v_1, ..., v_p)^T$, define $|v|_2 = \sqrt{\sum_{l=1}^{p} v_l^2}$, $|v|_1 = \sum_{l=1}^{p} |v_l|$, and $|v|_{\infty} = \max_{1 \leq l \leq p} |v_l|$. Moreover, we use $\text{supp}(v) = \{l \mid v_l \neq 0, 1 \leq l \leq p\}$ to denote the support of the vector $v$, and define $v_{-l} = (v_1, ..., v_{l-1}, v_{l+1}, ..., v_p)^T$. For each matrix $A \in \mathbb{R}^{p_1 \times p_2}$, $\|A\| = \sup_{\|v\|_2 = 1} |Av|_2$, $\|A\|_{\infty} = \max_{1 \leq l_1 \leq p_1, 1 \leq l_2 \leq p_2} |A_{l_1, l_2}|$ and $\|A\|_{L_{\infty}} = \sup_{\|v\|_{\infty} = 1} |Av|_{\infty}$. Furthermore, let $\Lambda_{\text{max}}(A)$ and $\Lambda_{\text{min}}(A)$ denote the largest and smallest eigenvalues of $A$, respectively. We use $I(\cdot)$ to denote the indicator function, and $\text{sgn}(\cdot)$ to denote the sign function. For two sequences $\{a_n\}, \{b_n\}$, we say $a_n \asymp b_n$ if $a_n = O(b_n)$ and $a_n = \Theta(b_n)$ hold simultaneously. For simplicity, we let $S^{p-1}$ and $\mathbb{B}^p$ denote the unit sphere and the unit ball in $\mathbb{R}^p$ centered on $0$. For a sequence of vectors $\{v_i\}_{i=1}^n \subseteq \mathbb{R}^p$, we define $\text{med}(\cdot)$ as the coordinate-wise median. Finally, the generic constants were assumed to be independent of $m, n,$ and $p$. We define $[p]$ to be the set $\{1, \ldots, p\}$.

2 A General Majority Vote Procedure for Privacy Preserving Sign Recovery

In this section, we introduce the Majority Vote procedure in a distributed system. Suppose that local machines estimate $\text{sgn}(\theta^*)$ using the local dataset $\{X_i, i \in \mathcal{H}_j\}$ separately and send the results to a trustworthy server. The server then aggregates the initial estimators received and provides a final result to the user. In this process, two pivotal procedures require careful design. First, the aggregation algorithm on the server integrates the sign information from initial estimators as much as possible. Moreover, a specific noise-addition technique was developed for DGDP. This requires an innovative way to integrate existing DP methods. Second, local machines require a feasible approach to generate good initial estimators. A natural strategy to address this problem is the regularization technique. However, in the distributed setting, we argue in Section 3 that the classical choice of regularization parameters can not attain the minimax optimal rate at the signal level. Therefore, careful design of the regularization parameter is important.
2.1 A Majority Vote Procedure

First, we formally introduce the Majority Vote procedure. Let $Q_j^c$ be the sign vector obtained by the local machine $H_j$ based on $\{X_i, i \in H_j\}$ (where $1 \leq j \leq m$). The server then has a $p \times m$ sign matrix $Q = (Q_{l,j}) = (Q_1^c, \ldots, Q_m^c)$ where each $Q_{l,j}$ takes a value of $\{1, -1, 0\}$, representing positive, negative and null, respectively. Here $Q_j^c = (Q_{1,j}, \ldots, Q_{p,j})^T$ denotes each column vector received from the $j$-th local machine (where $1 \leq j \leq m$). For each coordinate $1 \leq l \leq p$, we consider the Majority Vote with the row vector $Q_l^r = (Q_{l,1}, \ldots, Q_{l,m})$. The mechanism is as follows. When more than half of the elements in $Q_l^r$ have values of 1 (or $-1$), the output is this value. Otherwise, it returns null (0). Formally, for each row $Q_l^r$, we compute the number of positives, negatives, and nulls as follows:

$$
N_l^+ = \sum_{j=1}^{m} I(Q_{l,j} = 1), \quad N_l^- = \sum_{j=1}^{m} I(Q_{l,j} = -1), \quad N_l^0 = \sum_{j=1}^{m} I(Q_{l,j} = 0).
$$

(1)

Then $N_l^+ + N_l^- + N_l^0 = m$ always holds, and the Majority Vote of $Q_l^r$ can be equivalently computed as

$$
\text{MajVote}\{Q_l^r\} = \begin{cases} 
1 & \text{if } N_l^+ \geq N_l^0 + N_l^- + 1, \\
-1 & \text{if } N_l^- \geq N_l^0 + N_l^+ + 1, \\
0 & \text{otherwise.}
\end{cases}
$$

Let $\overline{Q}$ be the resulting vector of $Q$ by the Majority Vote, that is,

$$
\overline{Q} = (\text{MajVote}\{Q_l^r\}, 1 \leq l \leq p)^T.
$$

(2)

The merit of the Majority Vote will be illustrated in Section 3.1. For the mean model and linear regression model, we show that the vector $\overline{Q}$ has sign selection consistency under some weak conditions (see Propositions 1 and 2 in the Appendix). Although $\overline{Q}$ has already protected the privacy of the local dataset in some sense (note that local machines only send $\{1, -1, 0\}$ to the server), it does not satisfy DGDP in Definition 1. To solve this problem, we integrate the Majority Vote with the Peeling algorithm and exponential mechanism and develop a distributed group differentially private algorithm. Furthermore, we prove that the new algorithm can still exhibit sign selection consistency.

Before introducing our DGDP algorithm, we first review some classical definitions of the sensitivity function, exponential mechanism, and composition theorem for differential privacy. For simplicity, we define $\overline{S}$ as the index set of nonzero elements in $\overline{Q}$ and let $s = |\overline{S}|$. 
2.2 Privacy Preliminaries

Popularly, a key step to developing a differentially private algorithm is to impose additional randomness on the selection procedure. The scale of such randomness is typically determined by the global sensitivity of the algorithm.

**Definition 2 (Global Sensitivity).** Given an algorithm \( A : \mathcal{X}^N \rightarrow \Theta \), the global sensitivity of \( A \) is defined as
\[
\text{GS}_A = \sup \{ |A(X) - A(X')| : X, X' \in \mathcal{X}^N, H(X, X') = 1 \},
\]
where \( H(\cdot, \cdot) \) denotes the Hamming distance.

For the global sensitivity for DGDP in the distributed setting, we let \( H(X, X') = 1 \) denote that all elements are the same except for one local machine. Global sensitivity measures the magnitude of the change in the output of \( A \), resulting from replacing one element (or one group of elements) in the input dataset. Intuitively, when the noise is scaled proportionally to the global sensitivity, the output of the differentially private version of \( A \) is relatively stable, regardless of the presence or absence of any individual data in the dataset. Indeed, we have the following result.

**Lemma 1** (The Laplace mechanism, Theorem 3.6 of Dwork and Roth (2014)). For any algorithm \( A \) satisfying \( \text{GS}_A < \infty \), \( A_1 = A + g \), where \( g \) is sampled from \( \text{Lap}(\text{GS}_A/\epsilon) \), achieves \((\epsilon, 0)\)-differential privacy.

Here, \( \text{Lap}(b) \) denotes the Laplace distribution with the density function \( \frac{1}{2b} \exp(-|x|/b) \) for \( x \in (-\infty, \infty) \), and the scale parameter \( b > 0 \). Our algorithm acts on every coordinate as the composition of multiple actions, aiding us to introduce two composition theorems for the convenient construction of our algorithm.

**Lemma 2** (Composition theorem, Theorem B.1 of Dwork and Roth (2014)). Let \( A_1 : \mathcal{X}^n \rightarrow \Theta_1 \) be an \((\epsilon_1, \delta_1)\)-differentially private mechanism, and \( A_2 : (\mathcal{X}^n, \Theta_1) \rightarrow \Theta_2 \) be an \((\epsilon_2, \delta_2)\)-differentially private mechanism for every fixed element \( \theta_1 \in \Theta_1 \). Then, the composition of the mappings, \( (A_1, A_2) : \mathcal{X}^n \rightarrow (\Theta_1, \Theta_2) \) by mapping \( X \in \mathcal{X}^n \) to \( (A_1(X), A_2(X, A_1(X))) \), is an \((\epsilon_1 + \epsilon_2, \delta_1 + \delta_2)\)-differentially private mechanism.

**Lemma 3** (Advanced composition theorem, Corollary 3.21 in Dwork and Roth (2014)). For all \( \epsilon \in (0, 1), \delta, \delta' \in [0, 1] \), the \( k \)-fold adaptive composition of the class of \((\epsilon, \delta)\)-differentially private mechanisms preserves \((2\sqrt{2k \log(1/\delta')\epsilon}, k\delta + \delta')\)-differential privacy.

We refer to Section 3.5.2 of Dwork and Roth (2014) for a more detailed introduction to the definition of differential privacy, under the \( k \)-fold adaptive composition. Invariably, the two aforementioned composition theorems still hold for the DGDP.
Note that our goal is to recover the sign vector, whereas the direct addition of Laplace noise to \( \mathbf{Q} \) can completely destroy its value. It is well known that when the algorithm takes values with some special structure \( \Theta \) (e.g. \( \Theta = \{1, -1, 0\} \) in sign recovery), we can apply the exponential mechanism invented by McSherry and Talwar (2007). The exponential mechanism maintains the same structure as \( \Theta \) for the final output. In more detail, this mechanism assigns each pair of values in \( \Theta \) and the dataset \( \mathbf{X} \), with a utility function \( f : \mathcal{X}^n \times \Theta \rightarrow \mathbb{R} \), and generates a noisy output according to the global sensitivity of the utility function. In particular, we define

\[
\Delta f = \max_{\theta \in \Theta} \max_{H(\mathbf{X}, \mathbf{X}') = 1} |f(\mathbf{X}, \theta) - f(\mathbf{X}', \theta)|.
\]  

The exponential mechanism outputs a result \( \theta \in \Theta \) with probability proportional to \( \exp(\epsilon f(\mathbf{X}, \theta)/(2\Delta f)) \), that is,

\[
P(\text{OUTPUT} = \theta) = c \exp(\epsilon f(\mathbf{X}, \theta)/(2\Delta f)),
\]

where \( c \) is the scale constant of the probability distribution. We refer the reader to Section 3.4 in Dwork and Roth (2014) for more details about the exponential mechanism. The following lemma provides the privacy guarantee for the exponential mechanism.

**Lemma 4.** (Theorem 3.10 in Dwork and Roth (2014)) The exponential mechanism preserves \((\epsilon, 0)\)-differential privacy.

The exponential mechanism is also \((\epsilon, 0)\)-DGDP when we modify \( H(\mathbf{X}, \mathbf{X}') = 1 \) in (3) to denote that all elements are the same, except for one local machine. In fact, we can view the samples in one local machine as one higher-dimensional vector. Thus, all the conclusions on DP hold naturally for DGDP. The above lemma provides us with a method to modify the Majority Vote to a differentially private counterpart. We must leverage the knowledge of the sparseness of the true parameter. If we ignore the sparseness and apply the DP algorithm to every coordinate, by Lemma 3, we know that each coordinate should preserve \((\tilde{\epsilon}, \tilde{\delta})\) privacy, where \( \tilde{\epsilon} \asymp \frac{\epsilon}{\sqrt{p \log(1/\delta)}} \). The noise level imposed on each coordinate should then be scaled as \( \frac{\sqrt{p \log(1/\delta)}}{\epsilon} \), which may significantly deteriorate the performance when \( p \) is large. Therefore, it is important to use sparseness to reduce the level of noise and improve performance. To this end, we use the Peeling algorithm, which is typically used in the problem of differentially private top-k selection and was also used in Dwork and Roth (2014); Dwork et al. (2021); Cai et al. (2021). Using the Peeling algorithm, we can reduce the dimension and select a subset of \( \{1, 2, ..., p\} \) that asymptotically covers the true support. The challenge is to construct an efficient utility function \( f \) that is required in both the Peeling algorithm and exponential mechanism. That is, we must construct a measure of the closeness of the dataset \( \mathbf{Q} \) to the discrete set \( \{-1, 0, 1\} \). In the next section, we solve these problems by introducing the
Algorithm 1 Peeling ($\text{Peeling}(Q, \tilde{s}, \epsilon, \delta)$)

**Input:** The set of signs $Q = \{Q_1, \ldots, Q_m\}$; the number of selections $\tilde{s}$; privacy level $(\epsilon, \delta)$; initial $\tilde{S} = \emptyset$.

1. for $1 \leq l \leq p$ do
2. Compute the stability level $f^S(Q^l)$ based on (4).
3. end for
4. for $1 \leq t \leq \tilde{s}$ do
5. Generate $g_{t,1}, \ldots, g_{t,p} \sim \text{Lap}(4\sqrt{2\tilde{s} \log(1/\delta)}/\epsilon)$;
6. Add $l^* = \arg\max_{l \in [p]} \tilde{S} f^S(Q^l) + g_{t,l}$ to $\tilde{S}$.
7. end for

**Output:** Return the sets $\tilde{S}$.

concept of a stability function.

2.3 Private Sign Selection According to Majority Vote

Recall the definitions of $\{N_l^+, N_l^-, N_l^0\}$ given in (1). To solve the questions addressed at the end of the previous section, for each row $Q^l$ where $1 \leq l \leq p$, we define the stability level of dataset $Q^l$ with respect to the MajVote($\cdot$) function as the minimal number of elements in $Q^l$ which need to be flipped to change the value of MajVote{$Q^l$}. This can also be explicitly computed as:

$$f^S(Q^l) = \begin{cases} 
N_l^+ - N_l^0 - N_l^- & \text{if MajVote}\{Q^l\} = 1, \\
N_l^- - N_l^0 - N_l^+ & \text{if MajVote}\{Q^l\} = -1, \\
-\min \{N_l^+ + N_l^0 - N_l^- , N_l^- + N_l^0 - N_l^+\} & \text{if MajVote}\{Q^l\} = 0.
\end{cases} \quad (4)$$

Note that we assign the stability $f^S(Q^l)$ with a minus sign for MajVote{$Q^l$} = 0 because we only want to select the nonzero elements.

To select a subset that covers the support of $\bar{Q}$ in a private way, we leverage the Peeling algorithm with the stability function $f^S(Q^l)$. This procedure is presented in Algorithm 1. Note that in this algorithm we repeatedly select the maximal element in a private manner. Recently, Qiao et al. (2021) proposed a one-shot Top-$\tilde{s}$ selection algorithm that reduces computational burden. However, it uses a larger scale of noise, and theoretically only accepts small $\epsilon$ and $\delta$ (namely, $\epsilon \leq 0.2$ and $\delta \leq 0.05$). Thus, it is not adopted in our study.

The Peeling algorithm finds a subset $\tilde{S}$ that asymptotically covers the support of $\bar{Q}$. Notably, this does not generate the signs for each coordinate. Therefore, after generating the set $\tilde{S}$ for each coordinate $l$ (where $1 \leq l \leq p$), we apply the exponential mechanism by defining the following
Algorithm 2 Differentially Private Majority Vote for Sign Recovery (DPVote($Q, \tilde{s}, \epsilon, \delta$))

**Input:** The set of signs $Q = \{Q_1, \ldots, Q_m\}$; the number of selections $\tilde{s}$; privacy level $(\epsilon, \delta)$.

1. Apply Peeling($Q, \tilde{s}, \epsilon/2, \delta/2$) to select the index set $\tilde{S} = \{l_1, l_2, \ldots, l_{\tilde{s}}\}$.
2. for $l \in \tilde{S}$ do
3.   Compute the quantities $f_l(Q^+_l, 1)$, $f_l(Q^-_l, -1)$, $f_l(Q^0_l, 0)$ based on (5).
4.   Generate the random sign $Q_l$ according to the following distribution
   \[
   \begin{aligned}
   P(Q_l = 1) &= \frac{P^+_l}{P^+_l + P^-_l + P^0_l}, \\
   P(Q_l = 0) &= \frac{P^0_l}{P^+_l + P^-_l + P^0_l}, \\
   P(Q_l = -1) &= \frac{P^-_l}{P^+_l + P^-_l + P^0_l},
   \end{aligned}
   \]
   where
   \[
   \begin{aligned}
   P^+_l &= \exp\{\epsilon' f_l(Q^+_l, 1)/4\}, \\
   P^0_l &= \exp\{\epsilon' f_l(Q^0_l, 0)/4\}, \\
   P^-_l &= \exp\{\epsilon' f_l(Q^-_l, -1)/4\}.
   \end{aligned}
   \]
5. end for

**Output:** Return the sets $\tilde{S} = \{l | l \in \tilde{S}, Q_l \neq 0\}$ and $\hat{Q} = \{Q_l | l \in \tilde{S}\}$.

utility function:
\[
\begin{aligned}
f_l(Q^+_l, 1) &= N^+_l - N^0_l - N^-_l, \\
f_l(Q^-_l, -1) &= N^-_l - N^0_l - N^+_l, \\
f_l(Q^0_l, 0) &= \min\{N^+_l + N^0_l - N^-_l, N^-_l + N^0_l - N^+_l\}.
\end{aligned}
\] (5)

Using this utility function, we can easily verify that the global sensitivity in (3) is $\Delta f = 2$. That is, changing the entire dataset in one local machine can at most change the value of $f$ by $\pm 2$. This results in the realization of our private Majority Vote procedure as presented in Algorithm 2.

The concept of stability in differential privacy is related to the classical “Propose-Test-Release” mechanism. The “Propose-Test-Release” mechanism, or the PTR mechanism, was first proposed in Dwork and Lei (2009). We also refer the reader to Thakurta and Smith (2013); Dwork and Roth (2014); Vadhan (2017) for further exploration of this topic. The PTR mechanism uses stability to measure the distance of the database from the sensitive dataset and decides whether to halt the algorithm. We leverage this idea by using the stability function as the utility function in the Peeling algorithm and the exponential mechanism. Unlike the PTR mechanism, our algorithm does not halt and always outputs a non-empty set of signs.

### 2.4 Theories of DPVote Method

We show that the Peeling algorithm and the coordinate-wise exponential mechanism both attain $(\frac{\epsilon}{2}, \frac{\delta}{2})$-DGDP. Then, from the composition theorem in Lemma 2, we know that the entire procedure is $(\epsilon, \delta)$-DGDP. Namely, we have the following theorem:
Theorem 1 (Differential privacy of DPVote). The proposed DPVote mechanism in Algorithm 2 is $(\epsilon, \delta)$-distributed group differentially private.

Note that the $(\epsilon, \delta)$-DGDP conclusion for our algorithm holds for any sign matrix, $Q$. Next, we show that the algorithm outputs the same signs as $\overline{Q}$ in (2) asymptotically under weak conditions. Let $\overline{S}$ be the support for $\overline{Q}$ and $\overline{s} = |\overline{S}|$.

Theorem 2 (Sign consistency of DPVote). Given the privacy level $(\epsilon, \delta)$, suppose $\overline{s} \geq \tilde{s}$, and for $1 \leq l \leq p$, the stability function satisfies

$$f_l(\overline{Q}_r^l, \overline{Q}_l) \geq 8(\gamma + 1)\sqrt{2\overline{s} \log(2/\delta)} \log p/\epsilon,$$

where $\gamma > 2$. Then, the proposed DPVote algorithm produces $(\hat{S}, \hat{Q})$ that has the same signs as $\overline{Q}$ with a probability (with respect to the randomness in the algorithm) of no less than $1 - O(p^{-\gamma+2})$.

In this paper, we focus on the high-dimensional scenario where the dimension $p$ tends to infinity. Therefore, Theorem 2 implies consistency under this assumption. Condition (6) indicates that, if a sign (c.f. +1) obtains sufficient votes (exceeding the other two $8(\gamma + 1)\sqrt{2\overline{s} \log(2/\delta)} \log p/\epsilon$ votes), then the algorithm outputs the true sign of $\overline{Q}$ with high probability. The term $8(\gamma + 1)\sqrt{2\overline{s} \log(2/\delta)} \log p/\epsilon$ is determined by the noise level in the Peeling algorithm. Recall that $\overline{Q}$ is the estimator (without noise addition) for the sign vector of the population parameter. The proof shows that $\overline{Q}$ has sign consistency for the mean vector and linear regression models, which indicates that DPVote can have sign consistency for these two models. In these settings, we can deduce the corresponding minimal signal strength conditions (see (12) and (19)) from (6), which are regular in variable selection.

Notably, Theorem 2 does not assume any underlying model for the dataset in the local machines. That is, DPVote works for many statistical models once we can construct statistics in local machines to ensure sign consistency for $\overline{Q}$. Therefore, except for the mean model and linear regression problem given in Sections 3 and 4, DPVote can also be applied to other variable selection problems such as Gaussian graphical model estimation.

In the following sections, we apply the proposed DPVote algorithm to the sign selection for the mean vector and linear regression.

3 Private Sign Selection of Mean Vector

Private mean estimation is a fundamental problem in private statistical analysis and has been intensively studied (Dwork et al., 2006; Lei, 2011; Bassily et al., 2014; Cai et al., 2021). The standard approach is to project the data onto a known bounded domain and then apply the noises according to the diameter of the feasible domain and the privacy level. This requires that the input
**Algorithm 3** Differentially private Majority Vote (DPVote) for sparse mean.

**Input:** Dataset $X = \{X_1, \ldots, X_N\}$ stored in $m$ local machines (where $j = 1, \ldots, m$), the universal thresholding parameter $\lambda_N$, number of selections $\tilde{s}$, privacy level $(\epsilon, \delta)$.

1. for $j = 1, \ldots, m$ do
2. Compute the local sample mean $\overline{X}_j = n_j^{-1} \sum_{i \in H_j} X_i$ on the $j$-th machine and obtain the sign vector $Q_j = Q_{\lambda N}(\overline{X}_j)$. Send $Q_j$ to the server.
3. end for
4. Apply DPVote $\hat{Q}(X) = \text{DPVote}(\{Q_j\}_{1 \leq j \leq m}, \tilde{s}, \epsilon, \delta)$. (8)

**Output:** The sign vector $\hat{Q}(X)$.

Data or true parameter lie in a known bounded domain, which seems unsatisfactory in practice. Furthermore, few studies have examined the sign selection for the mean vector under DGDP. As the first application of the DPVote algorithm, in this section, we investigate the private sign recovery problem for a sparse mean vector in a distributed system. The results of sign selection consistency and privacy guarantee are provided.

### 3.1 DPVote for Mean Estimation

Let $\theta^* = (\theta_1^*, \ldots, \theta_p^*)^T$ be the true population parameter of interest. We denote $S$ as the support of $\theta^*$ and $s = |S|$. We assume that the vector is sparse in the sense that many entries $\theta_i^*$ are zero. There are $N$ i.i.d. observations $X_i$ that satisfy $E[X_i] = \theta^*$ and are stored in $m$ machines $H_j$ (where $1 \leq j \leq m$), each holds $n_j$ (where $1 \leq j \leq m$) samples. We assume that $n_j$’s have the same order as $n$, namely $n_1 \asymp \ldots \asymp n_m \asymp n$. We define $X = \{X_1, \ldots, X_N\}$ as the full dataset. Our task is to identify $\text{sgn}(\theta^*)$ under DGDP.

To present our method more clearly, we define the quantization function $Q_\lambda(\cdot)$ as follows:

$$Q_\lambda(x) = \text{sgn}\{\text{sgn}(x) \cdot (|x| - \lambda)_+\} = \begin{cases} \text{sgn}(x) & \text{if } |x| > \lambda, \\ 0 & \text{if } |x| \leq \lambda. \end{cases}$$ (7)

Here $\lambda$ is the threshold parameter. When $x$ is a vector, $Q_\lambda(x)$ performs the aforementioned operation in a coordinated manner. In particular, when $\lambda = 0$, the function $Q_0(\cdot)$ is the same as sign function $\text{sgn}(\cdot)$. We then present our method in Algorithm 3. The choice of thresholding parameter is discussed after Theorem 3 in Section 3.2.

**Majority Vote vs direct thresholding.** Before presenting the theoretical results, we first briefly discuss the advantages of the Majority Vote in sign/variable selection. Using the thresholding
Figure 1: This figure visualizes the mechanism of the majority vote approach. Denote $S^+$, $S^-$ and $S^c$ as the sets of positives, negatives, and zeros of the true parameter $\theta^*$, respectively. The black dots and white dots in each column represent the estimated positive and negative locations on each local machine. By aggregating these local sign vectors with the proposed method, we can recover the true signs with high probability.

estimator as the initial statistic, we let

$$\bar{Q} = \bar{Q}(\bar{X}) = \text{MajVote}(Q_{\lambda_N}(\bar{X}_j) | 1 \leq j \leq m),$$

(9)

where $\bar{X}_j$’s are the local sample means. To estimate support $S$, we claim that the Majority Vote procedure is more efficient than direct thresholding on local sample means. Suppose that the sample $X_i \sim \theta^* + \mathcal{N}(0, I)$. The local sample size is $n$. It is well known that a common threshold level for the local sample mean is $\lambda_n = c\sqrt{\log p/n}$ with $c = \sqrt{2}$, such that the local machine can identify all zero positions in $\theta^*$. It is easy to see that any smaller thresholding constant $c < \sqrt{2}$ results in false positives. With the thresholding level $\lambda_n$, the minimal signal of nonzero position should be no less than the order $O(\sqrt{\log p/n})$, otherwise, selection consistency becomes impossible. In contrast, the Majority Vote procedure allows for a thresholding level smaller than $\sqrt{2\log p/n}$. In fact, by using the Majority Vote procedure, we let $\lambda_N = c\sqrt{\log p/N}$ for some $c > 0$. Recall that $N$ denotes the total number of samples. This thresholding level is much smaller than $\sqrt{2\log p/n}$, and hence, many false nonzero positions are misidentified by local machines. However, due to the symmetry of the normal distribution, the number of false 1’s and −1’s are likely equal, and both are less than half. Thus, after applying the Majority Vote rule, the output is still 0. Now with a smaller $\lambda_N$,
we allow a weaker minimal signal strength assumption for nonzero positions. The mechanism of MajVote is illustrated in Figure 1.

3.2 Theory of Mean Vector Sign Selection

To discuss the theoretical properties of our method, we introduce the distribution space $P$ for the population $X$:

$$\mathcal{P}(\theta^*, C) = \left\{ P \left| \mathbb{E}_P[X] = \theta^*, \max_{1 \leq l \leq p} \mathbb{E}_P[|X_l - \theta^*_l|^3] \leq C \right. \right\},$$

where $C > 0$ is a constant. This is a relatively weak condition for the distribution of $X$. We obtain the following sign consistency result.

Theorem 3. (Sign consistency of DPVote for the mean vector) Let $\{X_1, \ldots, X_N\}$ be $N$ i.i.d. random vectors sampled from $P(\theta^*, C)$ distributed in $m$ machines $H_1, \ldots, H_m$, with each machine holds $n_j$ ($1 \leq j \leq m$) samples. Here we assume $n_1 \asymp \ldots \asymp n_m \asymp n$. Moreover, assume that there exist sufficiently large constants $C_1, C_2, \gamma_0 > 0$ such that

(a) The dimension $p$ satisfies $p = O(n^{\gamma_0})$, and the number of machines $m$ satisfies $\sqrt{s \log(1/\delta)} \log p / \epsilon = o(m)$. We take

$$\lambda_N \geq C_1 \left( \sqrt{\frac{\log p}{N}} + \frac{\sqrt{s \log(1/\delta)} \log p}{m \sqrt{n \epsilon}} + \frac{1}{n} \right);$$

(b) Define $S = \text{supp}(\theta^*)$, and assume that

$$\min_{l \in S} |\theta^*_l| \geq C_2 \left( \sqrt{\frac{\log p}{N}} + \frac{\sqrt{s \log(1/\delta)} \log p}{m \sqrt{n \epsilon}} + \frac{1}{n} \right).$$

Then for $\tilde{s} \geq s, \gamma_1 > 0$ and $\hat{Q}(X)$ defined in Algorithm 3, we have

$$\lim_{N \to \infty} \mathbb{P}(\hat{Q}(X) = \text{sgn}(\theta^*)) = 1.$$

As can be seen from assumptions (a) and (b), there are three terms in (11) and (12). The first term corresponds to the optimal signal-to-noise ratio, the second term is incurred by the privacy constraint, and the last term is caused by the error in the Berry-Esseen bound.

From Theorem 1, DPVote is $(\epsilon, \delta)$-DGDP. We now comment on the lower bound condition in (12). First, we fix the privacy level $(\epsilon, \delta)$. Assume that the number of local machines $m = O(n)$ and take $\tilde{s}$ in the Peeling algorithm satisfying $\sqrt{s \log(1/\delta)} / \epsilon = O(\sqrt{m})$. Thus, (12) becomes $\min_{l \in S} |\theta^*_l| \geq C \sqrt{\log p / N}$. This is the classical minimax rate-optimal lower bound for variable selection. That is, if we replace $C \sqrt{\log p / N}$ with $o(\sqrt{\log p / N})$, then for any statistics $T_N$ constructed from $\{X_1, \ldots, X_N\}$, there is a sufficiently small constant $c > 0$ and a parameter.
\( \theta^* \) satisfying \( \min_{l \in S} |\theta^*_l| = c\sqrt{\log p / N} \) such that the support of \( \theta^* \) can not be estimated by \( T_N \) consistently. See Theorem 3 in Cai et al. (2014) for further details.

At the end of this section, we remark that the private sign selection of the mean vector has a wide range of applications. Because we allow each coordinate of the vector \( X \) to be correlated, it is not difficult to extend our method to differentially private support recovery for the sparse covariance matrix (Karoui, 2008; Bickel and Levina, 2008; Cai and Liu, 2011) and the sparse precision matrix (Yuan, 2010; Cai et al., 2011).

4 Private Sign Selection of Linear Regression

In this section, we leverage the idea of DPVote for the sign selection problem for sparse linear regression in distributed systems. With some commonly adopted assumptions on the covariate vector and the noise, we prove the effectiveness and privacy of our method.

4.1 DPVote for Regression Parameter

Let \((X_i, Y_i)\) (where \(i = 1, \ldots, N\)) be the i.i.d. observations from the model

\[
Y = X^T \theta^* + z, \tag{13}
\]

where \( \theta^* = (\theta^*_1, \ldots, \theta^*_p)^T \) is the true sparse regression parameter, and \( z \) is the noise independent of the covariate \( X \). The full dataset is denoted by \( \mathcal{X} = \{(X_1, Y_1), \ldots, (X_N, Y_N)\} \), and it is assumed that \( \mathcal{X} \) is divided into \( m \) local machines \( \mathcal{H}_j \) \( (1 \leq j \leq m) \), each holds \( n_j \) samples. Similarly, we attempt to recover the sign vector \( \text{sgn}(\theta^*) = (\text{sgn}(\theta^*_1), \ldots, \text{sgn}(\theta^*_p))^T \).

By leveraging the idea of DPVote, we present Algorithm 4 for the sign recovery of sparse regression. Let

\[
b_{\theta_j}(\lambda) = \text{argmin}_{\theta \in \mathbb{R}^p} \frac{1}{2n_j} \sum_{i \in \mathcal{H}_j} (Y_i - X_i^T \theta)^2 + \lambda|\theta|_1. \tag{14}
\]

We use \( \text{sgn}(\hat{\theta}_j(\lambda)) \) for Majority Vote. We need to choose a problem-specific regularization parameter \( \lambda_j \) for each local machine. In particular, we let \( \lambda_j \) be the smallest number such that the local estimator \( \hat{\theta}_j(\lambda_j) \) having at most \( \tilde{s} \) nonzero elements and no less than a universal constant \( \lambda_N \) in (18). More precisely, we let

\[
\lambda_j = \min \left\{ \lambda \mid |\hat{\theta}_j(\lambda)|_0 \leq \tilde{s}, \lambda \geq \lambda_N \right\}. \tag{15}
\]

It is worth noting that \( \lambda_j \)'s could be different for every local machine. It should also be noted that the \( \lambda_j \)'s always exist. To demonstrate this, we observe that the resulting estimator \( \hat{\theta}_j(\lambda) \) is 0 if \( \lambda \) is sufficiently large. This implies that the set on the right-hand side of (15) is always nonempty. For
Algorithm 4 Differentially private Majority Vote for sparse linear regression (DPVote Lasso)

Input: Data on local machines \( \{(X_i, Y_i) \mid i \in \mathcal{H}_j\} \) for \( j = 1, \ldots, m \), the universal regularization parameter \( \lambda_N \) in (18) and \( \lambda_j \) in (15), privacy level \((\epsilon, \delta)\).

1: for \( j = 1, \ldots, m \) do
2: Let \( b_j = \theta_j(\lambda_j) \) in (14), and obtain the sign vector \( Q_j = \text{sgn}(b_j) \). Send \( Q_j \) to the server.
3: end for
4: Apply DPVote

\[
\hat{Q}(X) = \text{DPVote}(\{Q_j\}_{1 \leq j \leq m}, \tilde{s}, \epsilon, \delta).
\] (16)

Output: The sign vector \( \hat{Q}(X) \).

the implementation of (14), we can use the LARS algorithm (Efron et al., 2004) because it obtains all the solution paths efficiently. Hence the choice of \( \lambda_j \) can be quickly implemented.

4.2 Theory of Regression Parameter Sign Recovery

For linear regression, we consider the following distribution space

\[
\mathcal{P}_{X,Y}(\theta^*, \rho, \eta_1, C_1, \eta_2, C_2) = \left\{ P \mid \rho \leq \Lambda_{\min}(\Sigma) \leq \Lambda_{\max}(\Sigma) \leq \rho^{-1}, \right. \\
\sup_{|v|_2=1} \mathbb{E}_P\left\{ \exp(\eta_1|v^T X|^2) \right\} \leq C_1, \\
\left. z = Y - X^T \theta^*, z \perp X, \mathbb{E}_P\left\{ \exp(\eta_2|z|^2) \right\} \leq C_2 \right\}, 
\] (17)

where \( \rho, C_1, C_2, \eta_1, \eta_2 \) are positive constants. This implies that both the covariate vector \( X \) and the noise \( z \) are sub-Gaussian. Then, we obtain the following sign consistency result.

**Theorem 4.** (Sign consistency of DPVote Lasso) Let \( X = \{(X_1, Y_1), \ldots, (X_N, Y_N)\} \) be \( N \) i.i.d. random vectors sampled from \( \mathcal{P}_{X,Y}(\theta^*, \rho, \eta_1, C_1, \eta_2, C_2) \) distributed in \( m \) machines \( \mathcal{H}_1, \ldots, \mathcal{H}_m \), with each machine holds \( n_j \) \( (1 \leq j \leq m) \) samples. Here we assume \( n_1 \asymp \ldots \asymp n_m \asymp n \). Moreover, assume that sufficiently large constants \( C_3, C_4, \gamma_0 \) exist such that

(a) The dimension, the number of machines and the sparsity level satisfy \( p = O(n^{\gamma_0}) \),

\[
\sqrt{s} \log(1/\delta) \log p/\epsilon = o(m), \tilde{s} = o(\sqrt{n/\log p}), \text{ and we take} \\
\lambda_N = C_3 \left( \frac{\log p}{N} + \frac{\sqrt{s} \log(1/\delta) \log p}{m \sqrt{n \epsilon}} + \frac{1}{n} \right); 
\] (18)

(b) For \( S = \text{supp}(\theta^*) \), the minimal signal satisfies

\[
\min_{l \in S} |\theta^*_l| \geq C_4 \left( \frac{\log p}{N} + \frac{\sqrt{s} \log(1/\delta) \log p}{m \sqrt{n \epsilon}} + \frac{1}{n} + \max_{1 \leq j \leq m} \lambda_j \right), 
\] (19)

with probability tending to 1.
(c) The covariance matrix $\Sigma = \mathbb{E}XX^T$ is positive definite. Let $\Sigma^{-1} = (\omega_1, \ldots, \omega_p)$, assume that
\[
\max_{l \in S^c} \frac{|\omega_{-l}|}{\omega_{l,l}} \leq 1 - \Delta_0. \tag{20}
\]

Then for $\tilde{s} \geq s$, the estimator $\hat{Q}(X)$ defined in Algorithm 4 satisfies
\[
\lim_{N \to \infty} \mathbb{P}\left( \hat{Q}(X) = \text{sgn}(\theta^*) \right) = 1.
\]

From Theorem 1 we know DPVote regression is $(\epsilon, \delta)$-DGDP. We now make some comments regarding these assumptions. As can be seen from assumption (a), the lower bound of the regularization parameter $\lambda_N$ has three terms, which is the same as in (11). In assumption (b), we assume (19) holds with a probability tending to 1. Here, randomness comes from the selection of $\lambda_j$’s. Note that each $\lambda_j$ is chosen according to (15) which depends on the data. Assumption (c) implies that, for $l \in S^c$, the $l$-th row of the precision machine $\Sigma^{-1}$ is dominated by the diagonal entry $\omega_{l,l}$.

**Remark 1.** In the lower bound condition (19) on the signals, there is an additional term $\max_{1 \leq j \leq m} \lambda_j$, which depends on the magnitude of each $\lambda_j$. From the classical sparsity result for Lasso, under some regular conditions, $\text{supp}(\hat{\theta}_j) \subseteq S$ with probability tending to one, when $\lambda_j \geq \tilde{C} \sqrt{\log p/n}$ for a large $\tilde{C} > 0$ (see Wainwright (2009)). Therefore, it can be seen that $\lambda_j$ in (15) lies in the interval $(\lambda_N, \tilde{C} \sqrt{\log p/n})$ with a probability tending to one. Furthermore, if the dimension $p$ satisfies $p = o(\sqrt{n}/\log n)$ and $\sqrt{p \log p \log(1/\delta)}/\epsilon = O(\sqrt{m}) = O(\sqrt{n})$, we can simply set $\tilde{s} = p$, then, we have $\lambda_j = \lambda_N$. In this case, $\lambda_N \approx C \sqrt{\log p/N}$ and condition (19) becomes
\[
\min_{l \in S} |\theta^*_l| \geq C \sqrt{\log p/N}
\]
for some $C > 0$. This lower bound is optimal for variable selection consistency.

## 5 Simulation Study

In this section, we conduct two experiments to examine the performance of our Majority Vote procedure and the relevant differentially private versions. Since our study is the first to examine DGDP in a distributed system, it is difficult to compare it with other methods. However, we still choose the methods of Cai et al. (2021) as the object for comparison. There are two main reasons for this choice. First, the approaches suggested by Cai et al. (2021) are based on iterative algorithms and can be implemented in a distributed system. Second, from comparisons with Cai et al. (2021), we can see that the methods designed for parameter estimation perform poorly for sign/variable selection. Therefore, it is important to investigate the problem of sign/variable selection separately, as we claimed in the introduction. As we pointed out in the section on related works, there are also some special works on the variable selection with DP. However, some of them are computationally inefficient, and others cannot be applied directly to the distributed setting.
5.1 Results for Sparse Mean Estimation

In the first experiment, we consider the sparse mean estimation problem. The observations \( \{X_1, \ldots, X_N\} \) are sampled from the model

\[
X_i = \theta^* + z_i,
\]

where the noises \( z_i \)'s are drawn from a multivariate normal distribution \( \mathcal{N}(0, \Sigma) \). Here, the covariance matrix \( \Sigma \) is a \( p \times p \) Toeplitz matrix with its \((i, j)\)-th entry \( \Sigma_{ij} = 0.5^{|i-j|} \), where \( 1 \leq i, j \leq p \).

We set the dimension \( p \) to 500. The parameter of interest is defined as:

\[
\theta^* = (1, 0.8, \cdots, 0.2, -0.2, \cdots, -0.8, -1, 0_{p-10}^T)^T,
\]

which implies that the sparsity level \( s \) is fixed at 10. For the regularization parameter \( \lambda_N \) on each local machine, we take \( \lambda_N = 0.1 \) for simplicity. For the privacy level \((\epsilon, \delta)\), we take \( \epsilon = 0.5 \) and \( \delta = 0.05 \). Next, we check the performance of the following three methods: Majority Vote without adding noise, DPvote in Algorithm 3, and the method by Cai et al. (2021).

(a) Vote: Majority Vote without adding noise;

(b) DPvote: Majority Vote with DGDP;

(c) CWZ: The method proposed in Cai et al. (2021) for all samples on a single machine. Here we take the truncation level \( R \) in Cai et al. (2021) to be 2.

Note that Vote is a non-privacy algorithm. For DPvote and CWZ, we set the selection sparsity level \( \tilde{s} \) to 15. The original CWZ is designed to protect the privacy of a single element. To protect the privacy of the whole local samples, we multiply the noise level with the local sample size \( n \) by the composition theorem in Lemma 2.

The performance of sign recovery is measured using the following two criteria:

- **False Discovery Rate (FDR).**

\[
\text{FDR} = \frac{\sum_{l \notin S} \mathbb{I}\{\hat{Q}_l(X) \neq 0\} + \sum_{l \in S^-} \mathbb{I}\{\hat{Q}_l(X) = 1\} + \sum_{l \in S^+} \mathbb{I}\{\hat{Q}_l(X) = -1\}}{\max[p \sum_{l=1}^p \mathbb{I}\{\hat{Q}_l(X) \neq 0\}, 1]}.
\]

- **Power.**

\[
\text{Power} = \frac{\sum_{l \in S} \mathbb{I}\{\hat{Q}_l(X) = \text{sgn}(\theta^*_l)\}}{\max[|S|, 1]}.
\]

**Effect of the number of machines** To evaluate the effect of the number of machines \( m \), we fix the local sample size \( n \) to be 500 and vary the number of machines \( m \) from 500 to 1500 by 100. The total sample size \( N = nm \) varies accordingly. Then we report the FDR and Power of the three methods: DPvote, Vote, and CWZ. The results are shown in Figure 2.
As we can see, Vote always performs the best on both the power and FDR. This is not surprising, as we do not add any noise to Vote. The Power of DPVote is also close to one and is higher than that of CWZ significantly. On the other hand, the FDRs of Vote and DPVote are close to zero, whereas the FDRs of CWZ are quite large. The behavior of CWZ indicates that procedures work well in parameter estimation may be not suitable for sign selection.

Effect of the privacy budget To study the effect of the privacy budget $\epsilon$, we fix the number of machines $m$ to be 800, the local sample size $n$ to be 500, and vary the privacy level $\epsilon$ from 0.1 to 1 by 0.1. Because Vote is a non-private algorithm, we focus on the other two private algorithms in this part: DPvote and CWZ. The FDR and Power of these methods are presented in Figure 3.

With an increase in the privacy level $\epsilon$, the performance of DPvote improves. This is because a larger privacy level $\epsilon$ induces smaller random noise in private algorithms. Note that DPvote outperforms CWZ uniformly at all privacy levels. Although CWZ behaves better as $\epsilon$ increases, it still has a large FDR even when $\epsilon = 1$.

![Figure 2: The FDR and Power over the number of machines for sparse mean estimation. The number of machines varies from 500 to 1500, the local sample size is 500, and the dimension $p$ is 500.](image)

5.2 Results for Sparse Linear Regression

In this section, we consider the linear model defined by (13). The i.i.d. noises $z_i$’s are drawn from $\mathcal{N}(0, 1)$ and the i.i.d. covariate vectors $\mathbf{X}_i = (X_{i,1}, \ldots, X_{i,p})^T$ ($i = 1, \ldots, N$) are drawn from a multivariate normal distribution $\mathcal{N}(0, \Sigma)$. The covariance matrix $\Sigma$ is a $p \times p$ Toeplitz matrix with the $(i, j)$-th entry $\Sigma_{ij} = 0.5^{|i-j|}$, where $1 \leq i, j \leq p$. We fix the dimension $p = 200$. Moreover, we set the true coefficient $\theta^*$ as the same as $\theta^*$ in (21). In the CWZ method proposed in Cai et al.
Figure 3: The FDR and Power over the privacy level $\epsilon$ for sparse mean estimation. The local sample size is 500, the number of machines is 800, and the dimension $p$ is 500. (2021), we take the truncation level $R$ to be 2, the number of steps $T$ to be 20, and the step size $\eta$ to be 0.1.

We adopt the same settings as the counterpart in the experiment for the sparse mean estimation. The FDR and Power are shown in Figures 4 and 5. It is not surprising that all these methods perform similarly to mean estimation. That is, with an increase in the number of machines $m$ or the privacy level $\epsilon$, the FDR of DPVote approaches zero and Power approaches one. Again, DPVote outperforms CWZ in all settings.

Figure 4: The FDR and Power over the number of machines for sparse linear regression. The number of machines varies from 500 to 1500, the local sample size is 500, and the dimension $p$ is 500.
Figure 5: The FDR and Power over the privacy level $\epsilon$ for sparse linear regression. The local sample size is 500, the number of machines is 800, and the dimension $p$ is 500.

6 Conclusion

In summary, this study considers the sign selection problem under privacy constraints and applies it to private variable selection for sparse mean vectors and sparse linear regression in a distributed setup. Using the differentially private Majority Vote algorithm, we can consistently recover the signs. The admissible signal-to-noise ratio can have the same order as that of the non-private algorithm. In the following, we will discuss several additional topics and intriguing avenues for future research.

In the following section, we will explore several additional topics and intriguing avenues for future research.

- **More sparse recovery problems.** In this paper, we apply the DPvote algorithm to variable selection for sparse mean vectors and sparse linear regression. It is natural to extend our method to differentially private support recovery for sparse covariance matrix (Karoui, 2008; Bickel and Levina, 2008; Cai and Liu, 2011) and sparse precision matrix (Yuan, 2010; Cai et al., 2011). It is also interesting to consider applying our method to the debias-Lasso problem (Javanmard and Montanari, 2014; Lee et al., 2017; Javanmard and Montanari, 2018);

- **Untrusted curator.** In the DGDP definition, we make the assumption that a trusted curator is present to collect information from each machine and execute the algorithm. In scenarios where the curator cannot be trusted, secure multi-party computation (SMC) techniques are typically employed to prevent any party from learning intermediate information (Goldreich, 1998; Du and Atallah, 2001; Evans et al., 2018). While this setting is beyond the scope of our current paper, it would be interesting to explore combining our method with SMC
techniques to address this issue. Moreover, this scenario bears a resemblance to the local DP setup, where additional noise is added when transmitting data to an untrusted curator (Duchi et al., 2018). It would be possible to explore combining techniques from local DP with our method to tackle this privacy concern;

- **Optimal minimal signal condition.** In condition (b) of Theorem 3, we can find that the minimal signal condition is lower bounded by three terms. The second term $\sqrt{s}\log(1/\delta) \log p/(m\sqrt{\epsilon \delta})$ is new and arises from the privacy constraint. In Butucea et al. (2023), the author provided a $d/(\epsilon\sqrt{n})$ lower bound of the minimal signal under local DP setup (see Corollary 2.2). We acknowledge that the proof techniques for DP algorithms and local DP algorithms are different (see Cai et al. (2021) and Duchi et al. (2018)). Additionally, in Bafna and Ullman (2017) and Steinke and Ullman (2017), the authors demonstrated sample complexity lower bounds for DP selection, independent of the privacy levels $\epsilon$ and $\delta$. It would be interesting to develop a minimax theory for minimal signal under DP, as well as our proposed DGDP setups.

References

Arratia, R., Goldstein, L., and Gordon, L. (1989), “Two moments suffice for Poisson approximations: The Chen-Stein method,” *Ann. Probab.*, 17, 9–25.

Bafna, M. and Ullman, J. (2017), “The price of selection in differential privacy,” in *Proceedings of the 2017 Conference on Learning Theory*, PMLR, vol. 65 of *Proceedings of Machine Learning Research*, pp. 151–168.

Bassily, R., Smith, A., and Thakurta, A. (2014), “Private empirical risk minimization: Efficient algorithms and tight error bounds,” in *Proceedings of the 2014 IEEE 55th Annual Symposium on Foundations of Computer Science*, IEEE Computer Society, FOCS ’14, p. 464–473.

Battey, H., Fan, J., Liu, H., Lu, J., and Zhu, Z. (2018), “Distributed estimation and inference with statistical guarantees,” *Ann. Statist.*, 46, 1352–1382.

Bellet, A., Liang, Y., Garakani, A. B., Balcan, M.-F., and Sha, F. (2015), *A Distributed Frank-Wolfe Algorithm for Communication-Efficient Sparse Learning*, pp. 478–486.

Bickel, P. J. and Levina, E. (2008), “Covariance regularization by thresholding,” *Ann. Statist.*, 36, 2577 – 2604.

Boyd, S., Parikh, N., Chu, E., Peleato, B., and Eckstein, J. (2011), “Distributed optimization and statistical learning via the alternating direction method of multipliers,” *Found. Trends Mach. Learn.*, 3, 1–122.
Bu, Z., Dong, J., Long, Q., and Su, W. J. (2020), “Deep learning with gaussian differential privacy,” *Harvard data science review*, 2020.

Bun, M. and Steinke, T. (2016), “Concentrated differential privacy: Simplifications, extensions, and lower bounds,” in *TCC (B1)*, Springer, pp. 635–658.

Butucea, C., Dubois, A., and Saumard, A. (2023), “Phase transitions for support recovery under local differential privacy,” *Math. Stat. Learn.*, 6, 1–50.

Butucea, C., Ndaoud, M., Stepanova, N. A., and Tsybakov, A. B. (2018), “Variable selection with Hamming loss,” *Ann. Statist.*, 46, 1837–1875.

Cai, T. and Liu, W. (2011), “Adaptive thresholding for sparse covariance matrix estimation,” *J. Amer. Statist. Assoc.*, 106, 672–684.

Cai, T., Liu, W., and Luo, X. (2011), “A constrained ℓ1 minimization approach to sparse precision matrix estimation,” *J. Amer. Statist. Assoc.*, 106, 594–607.

Cai, T. T., Liu, W., and Xia, Y. (2014), “Two-sample test of high dimensional means under dependence,” *J. R. Stat. Soc. Series B Stat. Methodol.*, 76, 349–372.

Cai, T. T., Wang, Y., and Zhang, L. (2021), “The cost of privacy: Optimal rates of convergence for parameter estimation with differential privacy,” *Ann. Statist.*, 49, 2825–2850.

Chen, X., Miao, S., and Wang, Y. (2021), “Differential privacy in personalized pricing with non-parametric demand models,” *Oper. Res.*, To appear.

Chen, X., Simchi-Levi, D., and Wang, Y. (2022), “Privacy-preserving dynamic personalized pricing with demand learning,” *Manag. Sci.*, 68, 4878–4898.

Chow, Y. and Teicher, H. (2012), *Probability Theory: Independence, Interchangeability, Martingales*, Springer Texts in Statistics, Springer New York.

Dong, J., Roth, A., and Su, W. J. (2022), “Gaussian differential privacy,” *J. R. Stat. Soc. Series B Stat. Methodol.*, 84, 3–37.

Du, W. and Atallah, M. J. (2001), “Secure Multi-Party Computation Problems and Their Applications: A Review and Open Problems,” in *Proceedings of the 2001 Workshop on New Security Paradigms*, Association for Computing Machinery, NSPW ’01, p. 13–22.

Duchi, J. C., Jordan, M. I., and Wainwright, M. J. (2018), “Minimax optimal procedures for locally private estimation,” *J. Amer. Statist. Assoc.*, 113, 182–201.
Dwork, C. and Lei, J. (2009), “Differential privacy and robust statistics,” in *Proceedings of the Forty-First Annual ACM Symposium on Theory of Computing*, Association for Computing Machinery, STOC ’09, p. 371–380.

Dwork, C., McSherry, F., Nissim, K., and Smith, A. (2006), “Calibrating noise to sensitivity in private data analysis,” in *Theory of Cryptography*, Springer Berlin Heidelberg, pp. 265–284.

Dwork, C. and Roth, A. (2014), “The algorithmic foundations of differential privacy,” *Found. Trends Theor. Comput. Sci.*, 9, 211–407.

Dwork, C., Su, W., and Zhang, L. (2021), “Differentially private false discovery rate control,” *J. priv. confid.*, 11.

Dwork, C., Talwar, K., Thakurta, A., and Zhang, L. (2014), “Analyze Gauss: Optimal bounds for privacy-preserving principal component analysis,” in *Proceedings of the Forty-Sixth Annual ACM Symposium on Theory of Computing*, Association for Computing Machinery, STOC ’14, p. 11–20.

Efron, B., Hastie, T., Johnstone, I., and Tibshirani, R. (2004), “Least angle regression,” *Ann. Statist.*, 32, 407–499.

Evans, D., Kolesnikov, V., and Rosulek, M. (2018), “A Pragmatic Introduction to Secure Multi-Party Computation,” *Foundations and Trends® in Privacy and Security*, 2, 70–246.

Fan, J. and Li, R. (2001), “Variable selection via nonconcave penalized likelihood and its oracle properties,” *J. Amer. Statist. Assoc.*, 96, 1348–1360.

Goldreich, O. (1998), “Secure multi-party computation,” *Manuscript. Preliminary version*, 78.

Hamm, J., Cao, Y., and Belkin, M. (2016), “Learning privately from multiparty data,” in *Proceedings of The 33rd International Conference on Machine Learning*, PMLR, vol. 48 of *Proceedings of Machine Learning Research*, pp. 555–563.

Jain, P. and Thakurta, A. G. (2014), “(Near) Dimension independent risk bounds for differentially private learning,” in *Proceedings of the 31st International Conference on Machine Learning*, vol. 32, pp. 476–484.

Javanmard, A. and Montanari, A. (2014), “Confidence intervals and hypothesis testing for high-dimensional regression,” *J. Mach. Learn. Res.*, 15, 2869–2909.

— (2018), “Debiasing the lasso: Optimal sample size for Gaussian designs,” *Ann. Statist.*, 46, 2593–2622.
Jayaraman, B., Wang, L., Evans, D., and Gu, Q. (2018), “Distributed learning without distress: privacy-preserving empirical risk minimization,” in Advances in Neural Information Processing Systems 31, Curran Associates, Inc., pp. 6343–6354.

Karoui, N. E. (2008), “Operator norm consistent estimation of large-dimensional sparse covariance matrices,” Ann. Statist., 36, 2717 – 2756.

Kasiviswanathan, S. P. and Jin, H. (2016), “Efficient private empirical risk minimization for high-dimensional learning,” in Proceedings of The 33rd International Conference on Machine Learning, vol. 48, pp. 488–497.

Kifer, D., Smith, A., and Thakurta, A. (2012), “Private convex empirical risk minimization and high-dimensional regression,” in Proceedings of the 25th Annual Conference on Learning Theory, PMLR, vol. 23 of Proceedings of Machine Learning Research, pp. 25.1–25.40.

Lee, J. D., Liu, Q., Sun, Y., and Taylor, J. E. (2017), “Communication-efficient sparse regression,” J. Mach. Learn. Res., 18, 115–144.

Lei, J. (2011), “Differentially private M-estimators,” in Advances in Neural Information Processing Systems 24, Curran Associates, Inc., pp. 361–369.

Lei, J., Charest, A., Slavkovic, A., Smith, A., and Fienberg, S. (2018), “Differentially private model selection with penalized and constrained likelihood,” J. R. Stat. Soc. Ser. A Stat. Soc., 181, 609–633.

McSherry, F. and Talwar, K. (2007), “Mechanism design via differential privacy,” in 48th Annual IEEE Symposium on Foundations of Computer Science (FOCS’07), pp. 94–103.

Meinshausen, N. and Bühlmann, P. (2006), “High-dimensional graphs and variable selection with the Lasso,” Ann. Statist., 34, 1436–1462.

Miller, A. (2002), Subset Selection in Regression, New York: Chapman and Hall/CRC.

Mironov, I. (2017), “Rényi differential privacy,” in 2017 IEEE 30th Computer Security Foundations Symposium (CSF), pp. 263–275.

Nissim, K., Raskhodnikova, S., and Smith, A. (2007), “Smooth sensitivity and sampling in private data analysis,” in Proceedings of the Thirty-Ninth Annual ACM Symposium on Theory of Computing, Association for Computing Machinery, STOC ’07, p. 75–84.

Pathak, M., Rane, S., and Raj, B. (2010), “Multiparty differential privacy via aggregation of locally trained classifiers,” in Advances in Neural Information Processing Systems 23, Curran Associates, Inc., pp. 1876–1884.
Qiao, G., Su, W., and Zhang, L. (2021), “Oneshot differentially private Top-k selection,” in *Proceedings of the 38th International Conference on Machine Learning*, PMLR, vol. 139 of *Proceedings of Machine Learning Research*, pp. 8672–8681.

Shi, W., Ling, Q., Wu, G., and Yin, W. (2015), “A Proximal Gradient Algorithm for Decentralized Composite Optimization,” *IEEE Trans. Signal Process.*, 63, 6013–6023.

Smith, A. (2011), “Privacy-preserving statistical estimation with optimal convergence rates,” in *Proceedings of the Forty-Third Annual ACM Symposium on Theory of Computing*, Association for Computing Machinery, STOC ’11, p. 813–822.

Steinke, T. and Ullman, J. (2017), “Tight lower bounds for differentially private selection,” in 2017 *IEEE 58th Annual Symposium on Foundations of Computer Science (FOCS)*, pp. 552–563.

Talwar, K., Guha Thakurta, A., and Zhang, L. (2015), “Nearly optimal private LASSO,” in *Advances in Neural Information Processing Systems*, Curran Associates, Inc., vol. 28.

Thakurta, A. G. and Smith, A. (2013), “Differentially private feature selection via stability arguments, and the robustness of the Lasso,” in *Proceedings of the 26th Annual Conference on Learning Theory*, PMLR, vol. 30 of *Proceedings of Machine Learning Research*, pp. 819–850.

Tibshirani, R. (1996), “Regression shrinkage and selection via the lasso,” *J. R. Stat. Soc. Series B Stat. Methodol.*, 58, 267–288.

Vadhan, S. (2017), “The complexity of differential privacy,” in *Tutorials on the Foundations of Cryptography*, Springer, pp. 347–450.

Wainwright, M. J. (2009), “Sharp thresholds for high-dimensional and noisy sparsity recovery using $\ell_1$-constrained quadratic programming (Lasso),” *IEEE Trans. Inform. Theory*, 55, 2183–2202.

Wang, D. and Xu, J. (2019), “On sparse linear regression in the local differential privacy model,” in *Proceedings of the 36th International Conference on Machine Learning*, vol. 97, pp. 6628–6637.

Wang, J., Kolar, M., Srebro, N., and Zhang, T. (2017), “Efficient distributed learning with sparsity,” in *Proceedings of the 34th International Conference on Machine Learning*, vol. 70, pp. 3636–3645.

Wang, Y.-X., Lei, J., and Fienberg, S. E. (2016), “On-average KL-privacy and its equivalence to generalization for max-entropy mechanisms,” in *Privacy in Statistical Databases*, Cham: Springer International Publishing, pp. 121–134.

Wasserman, L. and Zhou, S. (2010), “A Statistical Framework for Differential Privacy,” *J. Amer. Statist. Assoc.*, 105, 375–389.
Yuan, M. (2010), “High dimensional inverse covariance matrix estimation via linear programming,”
*J. Mach. Learn. Res.*, 11, 2261–2286.

Zhu, X., Li, F., and Wang, H. (2021), “Least-Square Approximation for a Distributed System,” *J. Comput. Graph. Statist.*, 30, 1004–1018.
7 Appendix

7.1 Proof of Results in Section 2.4

Proof of Theorem 1. Notice that the stability function $f^S(\cdot)$ has sensitive 2. By Lemma 2.4 of Dwork et al. (2021), we know each round of selection is $(\epsilon', 0)$ private, where

$$
\epsilon' = \frac{\epsilon}{4\sqrt{2s \log(2/\delta)}}.
$$

Then by advanced composition theorem (Lemma 3), we know that the selection of the set $\tilde{S}$ with $\tilde{s}$ indices is $(\frac{\epsilon}{2}, \frac{\delta}{2})$-differentially private. Notice that the functions $f_l(\cdot, 1), f_l(\cdot, -1), f_l(\cdot, 0)$ all have global sensitivity 2. By Theorem 3.10 of Dwork and Roth (2014), for every $l \in \tilde{S}$, the generation of $\tilde{Q}_l$ is $(\epsilon', 0)$-differentially private. Then by the advanced composition theorem (Lemma 3), we know that the generation of the $\tilde{s}$-tuple $\{\tilde{Q}_l | l \in \tilde{S}\}$ is $(\frac{\epsilon}{2}, \frac{\delta}{2})$-differentially private. Next, we apply the composition theorem (Lemma 2) and conclude that the composition of these two mechanisms is $(\epsilon, \delta)$-differentially private. \qed

Proof of Theorem 2. To prove the selection consistency of DPVote, we first prove the consistency of Peeling.

**Consistency of Peeling:** Let us directly consider the extreme case, where $f^S(Q^t_l)$

$$
[f_{\tilde{S}}(Q^t_l) + g_l \geq 0] \text{ for } l \in \tilde{S}, \text{ and } f^S(Q^t_l) = 0 \text{ for } l \notin \tilde{S}.
$$

Clearly the probability of DPVote algorithm produces the correct signs for any dataset $Q$ is not less than this case. For simplicity denote it as $P(\text{extreme}).$

On the other hand, for the $t$-th selection step (where $1 \leq t \leq \tilde{s}$), suppose there left $\tilde{s} - t + 1$ elements in $\tilde{S}$ and $p - \tilde{s}$ elements in $\tilde{S}^c$, the probability that the peeling algorithm select element in $\tilde{S}$ is greater than the case when there left only one element in $\tilde{S}$, let’s denote this probability as $P(1 \text{ vs } p - \tilde{s})$. Since each round of selection is independent, we know

$$
P(\text{extreme}) \geq P(1 \text{ vs } p - \tilde{s}). \quad (22)
$$

Therefore, it left to bound the probability $P(1 \text{ vs } p - \tilde{s})$.

By Lemma 6, we know that

$$
P(\max_{l \in \tilde{S}} f^S(Q^t_l) + g_l \geq 0) \leq P(\max_{l \in \tilde{S}} g_l \geq 8\sqrt{2\tilde{s}\log(2/\delta)} \log p) \leq p^{-\gamma + 1},
$$

$$
P(\max_{l \in \tilde{S}} f^S(Q^t_l) + g_l \leq 0) \leq P(\max_{l \in \tilde{S}} g_l \leq -8\sqrt{2\tilde{s}\log(2/\delta)} \log p) \leq p^{-\gamma + 1}.
$$

29
Therefore,
\[
\mathbb{P}(1 \text{ vs } p - s) \geq 1 - \mathbb{P}(\max_{l \in \check{S}} f^S(Q_l^0) + g_l \geq 0) - \mathbb{P}(\max_{l \in \check{S}} f^S(Q_l^0) + g_l \leq 0) \geq 1 - 2p^{-\gamma + 1}.
\]

Plugging it into (22) we have that
\[
\mathbb{P} \text{(extreme)} \geq \mathbb{P}^S(1 \text{ vs } p - s) \geq (1 - 2p^{-\gamma + 1})^\tilde{S} \geq 1 - 2\delta p^{-\gamma + 1} \geq 1 - O(p^{-\gamma + 2}),
\]
(23)
which proves the first part.

**Consistency of DPVote:** Suppose that the Peeling algorithm has selected the support covering \(\check{S}\), i.e. \(\check{S} \subseteq \tilde{S}\). For each \(l \in \check{S}\), without loss of generality assume \(Q_l^0 = 1\). Then from assumption we know
\[
f^S(Q_l^0) = N_l^+ - N_l^0 - N_l^- \geq 8(\gamma + 1)\sqrt{2\delta \log(2/\delta)} \log p/\epsilon.
\]

Then we have that
\[
\mathbb{P}(\check{Q}_l = 1) = \frac{P_l^0}{P_l^+ + P_l^- + P_l^0} = \frac{1}{1 + \exp\{e'(N_l^0 - N_l^+ + N_l^-)/2\} + \exp\{e'(N_l^- - N_l^+)/2\}}
\]
\[
\geq \frac{1}{1 + \exp\{-e'f_l(Q_l^0, 1)/2\} + \exp\{e'f_l(Q_l^0, 1)/2\}}
\]
\[
\geq 1 + 2p^{-\gamma - 1} \geq 1 - 2p^{-\gamma - 1}.
\]

For \(l \in \tilde{S}\setminus \check{S}\), we have that
\[
\mathbb{P}(\check{Q}_l = 0) = \frac{P_l^0}{P_l^+ + P_l^- + P_l^0} = \frac{1}{1 + \exp\{-e' \min\{N_l^0 - N_l^+ + N_l^-\}/2\} + \exp\{e' \min\{N_l^0 - N_l^- + N_l^+\}/2\}}
\]
\[
\geq 1 + 2 \exp\{-e' \min\{N_l^0 - N_l^+ + N_l^-\}/2\} + \exp\{e' \min\{N_l^0 - N_l^- + N_l^+\}/2\}
\]
\[
\geq 1 + 2p^{-\gamma - 1} \geq 1 - 2p^{-\gamma - 1}.
\]

Then we have that
\[
\mathbb{P}(\check{Q}_l = Q_l, l \in \check{S}) \geq \prod_{l \in \check{S}} \mathbb{P}(\check{Q}_l = Q_l) \geq (1 - 2p^{-\gamma - 1})^\check{S} \geq 1 - 2\delta p^{-\gamma - 1} \geq 1 - p^{-\gamma}.
\]
Therefore,
\[ P(\hat{\mathcal{Q}} \cap \{ P(\hat{\mathcal{Q}}) \geq \hat{Q}_l, l \in \hat{S} \}) - (1 - P(\text{extreme})) \geq 1 - p^{-\gamma} - p^{-2\gamma + 2} = 1 - O(p^{-\gamma + 2}), \]
which proves the Theorem 2.

\[ \square \]

**Lemma 5.** (Theorem 1 of Arratia et al. (1989)) Let \( \{ g_i \mid i \in I \} \) be the collection of random variables in an index set \( I \). For each \( i \in I \), let \( I_i \) be a subset of \( I \) with \( i \in I_i \). For a given \( x \in \mathbb{R} \), set \( y = \sum_{i \in I} P(g_i > x) \). Then
\[
\left| P\left( \max_{i \in I} g_i \leq x \right) - e^{-y} \right| \leq (1 \wedge y^{-1})(T_1 + T_2 + T_3),
\]
where
\[
T_1 = \sum_{i \in I} \sum_{j \in I_i} P(g_i > x)P(g_j > x),
\]
\[
T_2 = \sum_{i \in I} \sum_{j \in I_i \backslash \{ \}} P(g_i > x, g_j > x),
\]
\[
T_3 = \sum_{i \in I} \mathbb{E}[\left| P(g_i > x \mid g_j, j \notin I_i) - P(g_i > x) \right|].
\]

**Lemma 6.** Let \( g_1, \ldots, g_p \) be \( p \) i.i.d. random variable sampled from \( \text{Lap}(\lambda) \), then for each \( \gamma \geq 2 \) we have that
\[ P\left( \max_{i \in I} g_i \geq \gamma \lambda \log p \right) \leq p^{-\gamma + 1}. \]

**Proof.** To apply Lemma 5, we let the index \( I = \{1, \ldots, p\} \), for every \( i \), let \( I_i = \{i\} \). Then by taking \( x = \gamma \lambda \log p \), we know that
\[
y = \sum_{i \in I} P(g_i > x) = \frac{p}{2} e^{-x/\lambda} = \frac{1}{2} p^{-\gamma + 1},
\]
\[
T_2 = T_3 = 0,
\]
\[
T_1 = p \sum_{i \in I} P^2(g_i > x) = \frac{p}{4} e^{-2x/\lambda} = \frac{1}{4} p^{-2\gamma + 1}.
\]
By plugging them into (24) we have that
\[
\left| P\left( \max_{i \in I} g_i \leq \gamma \lambda \log p \right) - e^{-y/2p^{-\gamma + 1}} \right| \leq \frac{1}{4} p^{-2\gamma + 1},
\]
\[ \Rightarrow P\left( \max_{i \in I} g_i > \gamma \lambda \log p \right) \leq 1 - e^{-y/2p^{-\gamma + 1}} + \frac{1}{4} p^{-2\gamma + 1} \leq p^{-\gamma + 1}, \]
which proves the lemma. \[ \square \]
7.2 Proof of Results in Section 3.2

**Lemma 7.** (Berry-Esseen Theorem, Theorem 9.1.3 in Chow and Teicher (2012)) If \( \{X_i, i \geq 1\} \) are i.i.d. mean-zero random variables with \( \mathbb{E}[X_i^2] = \sigma^2, \mathbb{E}|X_i|^3 < \infty \). Then there exists a constant \( C_B > 0 \) such that

\[
\sup_{-\infty < x < \infty} \left| \mathbb{P} \left( \sum_{i=1}^{n} X_i < \sqrt{n} \sigma x \right) - \Phi(x) \right| \leq \frac{C_B}{\sqrt{n}}.
\]

**Lemma 8.** (Exponential Inequality, Lemma 1 in Cai and Liu (2011)) Let \( X_1, \ldots, X_n \) be i.i.d. random variables with zero mean. Suppose that there exist some \( \eta > 0 \) and \( C > 0 \) such that \( \mathbb{E}[X_i^2 e^{\eta |X_i|}] \leq C \). Then uniformly for \( 0 < x \leq C \) and \( n \geq 1 \), there is

\[
\mathbb{P} \left\{ \frac{1}{n} \sum_{i=1}^{n} X_i \geq (\eta + \eta^{-1})x \right\} \leq \exp \left( -\frac{nx^2}{C} \right).
\]

**Lemma 9.** Let \( N \) i.i.d. random variables \( X_1, \ldots, X_N \) distributed in \( m \) subsets \( \mathcal{H}_1, \ldots, \mathcal{H}_m \), with each subset holds \( n_j \) (\( 1 \leq j \leq m \)) samples. Here we assume \( n_1 \approx \ldots \approx n_m \approx n \). Suppose \( \mathbb{E}[X_i] = 0, \mathbb{E}[X_i^2] = \sigma^2, \mathbb{E}|X_i|^3 < \infty \). Denote \( \bar{X}_j = \sum_{i \in \mathcal{H}_j} X_i/n_j \) as the local sample mean on \( \mathcal{H}_j \). For every \( \gamma > 1 \) and non-negative constant \( k \), denote

\[
c_{\gamma,k} = \tilde{C} \left( \frac{1}{n} + \frac{k}{m \sqrt{n}} + \sqrt{\frac{\gamma \log p}{mn}} \right),
\]

where \( \tilde{C} > 0 \) is sufficiently large enough. Then there is

\[
\mathbb{P} \left\{ \sum_{j=1}^{m} \mathbb{I} (\bar{X}_j < c_{\gamma,k}) \leq \frac{m}{2} + k \right\} + \mathbb{P} \left\{ \sum_{j=1}^{m} \mathbb{I} (\bar{X}_j > -c_{\gamma,k}) \leq \frac{m}{2} + k \right\} = O(p^{-\gamma}). \tag{25}
\]

**Proof.** For every \( x > 0 \), there is

\[
\mathbb{P} \left\{ \sum_{j=1}^{m} \mathbb{I} (\bar{X}_j < x) \leq \frac{m}{2} + k \right\}
\]

\[
= \mathbb{P} \left\{ \frac{1}{m} \sum_{j=1}^{m} \mathbb{I} (\bar{X}_j < x) - \mathbb{P} (\bar{X}_j < x) \right\} \leq \frac{1}{2} + \frac{k}{m} - \frac{1}{m} \sum_{j=1}^{m} \mathbb{P} (\bar{X}_j < x)
\]

\[
\leq \mathbb{P} \left\{ \frac{1}{m} \sum_{j=1}^{m} \mathbb{I} (\bar{X}_j < x) - \mathbb{P} (\bar{X}_j < x) \right\} \leq \frac{1}{2} + \frac{k}{m} + \frac{1}{m} \sum_{j=1}^{m} \left\{ \frac{C_B}{\sqrt{n_j}} - \Phi \left( \frac{\sqrt{n_j}x}{\sigma} \right) \right\}
\]

\[
\leq \mathbb{P} \left\{ \frac{1}{m} \sum_{j=1}^{m} \mathbb{I} (\bar{X}_j < x) - \mathbb{P} (\bar{X}_j < x) \right\} \leq -\sqrt{\frac{C \gamma \log p}{m}},
\]

32
where the last line uses Berry-Esseen Theorem (Lemma 7), and \( x \) is given by

\[
x = \max_{1 \leq j \leq m} \frac{\sigma}{\sqrt{n_j}} \Phi^{-1} \left( \frac{1}{2} + \frac{k}{m} + \frac{C_B \log p}{\sqrt{n}} \right) = \frac{\sigma}{\sqrt{n}} \Phi^{-1} \left( \frac{1}{2} + \frac{k}{m} + \frac{C_B \log p}{\sqrt{n}} \right).
\]

Applying Lemma 8 to the i.i.d. sequence \( I(X_j < x) - P(X_1 < x) \), we have

\[
P \left\{ \frac{1}{m} \sum_{j=1}^{m} I(X_j < x) - P(X_1 < x) \leq -\sqrt{\frac{C_B \log p}{n}} \right\} = O(p^{-\gamma}),
\]

for some \( C \) large enough. Moreover, we have the following elementary facts

\[
\Phi^{-1}(x_0) = \Phi^{-1}(x_0) - \Phi^{-1}(1/2) \leq |x_0 - 1/2| \Phi^{-1}'(x_0) \leq \frac{|x_0 - 1/2|}{\psi\{\Phi^{-1}(c_\phi)\}},
\]

holds for any \( c_\phi \leq x_0 < 1 - c_\phi \), where \( c_\phi \in (0, 1/2) \) is some constant. On the other hand, we know that \( c_\phi \leq \min_j \Phi(\sqrt{n_j} x/\sigma) \leq \max_j \Phi(\sqrt{n_j} x/\sigma) < 1 - c_\phi \) holds for \( m, n \) sufficiently large. Denote \( C_\delta = 1/\psi\{\Phi^{-1}(c_\phi)\} \), then there is

\[
P \left\{ \frac{1}{m} \sum_{j=1}^{m} I(X_j < x) - P(X_1 < x) \leq -\sqrt{\frac{C_B \log p}{n}} \right\} \leq p^{-\gamma}.
\]

Therefore, if we choose \( \tilde{C} \geq \max\{C_\delta C_B \sigma, C_\delta \sqrt{C} \sigma\} \), the bound of the first term in the left hand side of (25) is proved. By repeating the same procedure, we can prove the bound of the second term, which yields the desired result.

Before proving Theorem 3, we shall first prove the following proposition.

**Proposition 1.** (Sign consistency of MajVote) Let \( N \) i.i.d. random vectors \( \{X_1, \ldots, X_N\} \) be sampled from \( P(\theta^*, C) \). Moreover, there are sufficiently large constants \( C_3, C_4, \gamma_0 > 0 \), such that

(a’) The dimension \( p \) satisfies \( p = O(n^{\gamma_0}) \), and take \( \lambda_N = C_3(\sqrt{\log p/N} + 1/n) \);

(b’) Denote \( S = \text{supp}(\theta^*) \), then there is min_{i \in S} |\theta_i^*| \geq C_4 \lambda_N.

Then \( \overline{Q}(X) \) defined in (9) satisfies that, for some large \( \gamma_2 \) depends on \( C_1, C_2, \gamma_0 \), there is

\[
P\left( \overline{Q}(X) = \text{sgn}(\theta^*) \right) \geq 1 - O(p^{-\gamma_2}).
\]
Proof of Proposition 1. If we can show that
\[ \max_{1 \leq l \leq p} \mathbb{P} \left( \overline{Q}_l(X) \neq \text{sgn}(\theta^*_l) \right) = O(p^{-\gamma}), \] (26)
where \( \gamma > 0 \) is large enough. Then this theorem is proved as follows
\[ 1 - \mathbb{P} \left( \overline{Q}(X) = \text{sgn}(\theta^*) \right) \leq p \max_{1 \leq l \leq p} \mathbb{P} \left( \overline{Q}_l(X) \neq \text{sgn}(\theta^*_l) \right) = O(pp^{-\gamma}) = O(p^{-\gamma+\gamma_0}), \]
provided that \( \gamma > \gamma_0 \) and sufficient large \( \gamma_0 \). For the \( l \)-th coordinate, we firstly suppose that \( \theta^*_l = 0 \).
Since \( X_i \)'s are sampled from the distribution (10), by Lemma 9 with \( k = 0 \), if
\[ \lambda_N = C_0 \left( \sqrt{\frac{\log p}{mn}} + \frac{1}{n} \right) \geq c_{\gamma,0}, \]
we have
\[ \mathbb{P} \left( \overline{Q}_l(X) \neq \text{sgn}(\theta^*_l) \right) \leq \mathbb{P} \left\{ \sum_{j=1}^{m} \mathbb{I} (X_{j,l} < \lambda_N) \leq \frac{m}{2} \right\} + \mathbb{P} \left\{ \sum_{j=1}^{m} \mathbb{I} (X_{j,l} > -\lambda_N) \leq \frac{m}{2} \right\} = O(p^{-\gamma}). \]
Next we assume \( \theta^*_l > 0 \). We use Lemma 9 again, if
\[ \theta^*_l \geq C_1 \left( \sqrt{\frac{\log p}{mn}} + \frac{1}{n} \right) \geq c_{\gamma,0} + \lambda_N, \]
then there is
\[ \mathbb{P} \left( \overline{Q}_l(X) \neq \text{sgn}(\theta^*_l) \right) \leq \mathbb{P} \left\{ \sum_{j=1}^{m} \mathbb{I} (X_{j,l} > \lambda_N) \leq \frac{m}{2} \right\} \leq \mathbb{P} \left\{ \sum_{j=1}^{m} \mathbb{I} (X_{j,l} - \theta^*_l > -c_{\gamma,0}) \leq \frac{m}{2} \right\} = O(p^{-\gamma}). \]
Lastly, when \( \theta^*_l < 0 \), the proof is the same as above. Therefore (26) is proved.

Proof of Theorem 3. By Theorem 1, we know that \( \overline{Q}(X) = \text{sgn}(\theta^*) \) holds with probability not less than \( 1 - p^{-\gamma_1} \). To prove the sign consistency of \( \hat{Q}(X) \), from Theorem 2 we know that it is enough to show for each \( l \in S \), there is
\[ f^S(\hat{Q}_l^r) \geq 8(\gamma + 1) \sqrt{2S \log(2/\delta) \log p/\epsilon} \]
hold with probability $1 - O(p^{-\gamma})$. For $l \in S$, without loss of generality assume that $\text{sgn}(\theta^*_l) = \text{sgn}(\overline{Q}_l(X)) = 1$, then

$$f_l(Q^*_l, \overline{Q}_l(X)) = N^+_l - N^-_l - N^0_l = 2N^+_l - m$$

$$= 2 \sum_{j=1}^m \mathbb{I}(\overline{X}_{j,l} > \lambda_N) - m$$

$$= 2 \{ \sum_{j=1}^m \mathbb{I}(\overline{X}_{j,l} - \theta^*_l > \lambda_N - \theta^*_l) - \frac{m}{2} \}.$$

Therefore,

$$\mathbb{P}\left( f^S(Q^*_l) \geq 8(\gamma + 1)\sqrt{2\bar{s}\log(2/\delta) \log p/\epsilon} \right)$$

$$= \mathbb{P}\left( \sum_{j=1}^m \mathbb{I}(\overline{X}_{j,l} - \theta^*_l > \lambda_N - \theta^*_l) - \frac{m}{2} \geq 4(\gamma + 1)\sqrt{2\bar{s}\log(2/\delta) \log p/\epsilon} \right)$$

$$\geq 1 - O(p^{-\gamma}),$$

as long as

$$\theta^*_l - \lambda_N \geq C_\delta \sigma \left( \frac{C_B}{\sqrt{n}} + \frac{4(\gamma + 1)\sqrt{2\bar{s}\log(2/\delta) \log p}}{m\epsilon} + \sqrt{\frac{C\gamma_1 \log p}{m}} \right),$$

which proves the first part of the proposition. Similarly, when $\bar{s} > s$, for $l \notin S$, there is

$$f_l(Q^*_l, 0) = \min\{N^0_l - N^+_l + N^-_l, N^0_l + N^+_l - N^-_l\}$$

$$= \min\left\{ 2 \sum_{j=1}^m \mathbb{I}(\overline{X}_{j,l} \leq \lambda_N), 2 \sum_{j=1}^m \mathbb{I}(\overline{X}_{j,l} \geq -\lambda_N) \right\} - m.$$
as long as
\[
\lambda_N \geq \frac{C_5 \sigma}{\sqrt{n}} \left( \frac{C_B}{\sqrt{n}} + \frac{4(\gamma + 1)\sqrt{2s \log(2/\delta) \log p}}{m\epsilon} + \sqrt{\frac{C\gamma_1 \log p}{m}} \right).
\]

Similarly, we can show that
\[
P\left( \sum_{j=1}^{m} \mathbb{I} \left( X_{j,l} \geq -\lambda_N \right) - m \geq 8(\gamma + 1)\sqrt{2s \log(2/\delta) \log p/\epsilon} \right) \geq 1 - O(p^{-\gamma_1}).
\]

Therefore we have
\[
P\left( f_l(Q^l_1, 0) \geq 8(\gamma + 1)\sqrt{2s \log(2/\delta) \log p/\epsilon} \right) \geq 1 - O(p^{-\gamma_1}),
\]
and Theorem 2 applies.

### 7.3 Proof of Results in Section 4.2

**Lemma 10.** Let \( X_1, \ldots, X_n \) be i.i.d. random vectors sampled from the distribution in (17). Denote its covariance matrix as \( \Sigma \), and the sample covariance matrix as \( \hat{\Sigma} \). Then for every \( \gamma > 1 \), there exists a constant \( \tilde{C} > 0 \) such that
\[
P\left( \| \Sigma^{-1} \hat{\Sigma} - I \|_\infty \geq \tilde{C} \sqrt{\frac{\log p}{n}} \right) = O(p^{-\gamma}).
\]

**Proof.** Recall that the inverse covariance matrix is denoted as \( \Sigma^{-1} = (\omega_1, \ldots, \omega_p) \), and \( e_l \) as the \( l \)-th coordinate vector. Then the \((l_1, l_2)\)-entry of the matrix \( \Sigma^{-1} \hat{\Sigma} - I \) is
\[
(\Sigma^{-1} \hat{\Sigma} - I)_{l_1,l_2} = \frac{1}{n} \sum_{i=1}^{n} \omega_{l_1}^T X_i \cdot e_{l_2}^T X_i - \delta_{l_1,l_2}.
\]

Since the dimension \( p \) is assumed to be bounded, and the covariance matrix is positive definite, there exist a constant \( \rho \in (0,1) \) such that
\[
\rho \leq \Lambda_{\min}(\Sigma) \leq \Lambda_{\max}(\Sigma) \leq \rho^{-1}.
\]

Then we have
\[
\max_{1 \leq |l| \leq p} |\omega_l|^2 \leq \| \Sigma^{-1} \| \leq \rho^{-1}.
\]
Since $X$ is sub-Gaussian in (17), we obtain

$$\max_{1 \leq i_1, i_2 \leq p} \mathbb{E}\{ \exp \eta_1 \rho | \omega_{i_1}^T X_i \cdot e_{i_2}^T X_i - \delta_{i_1, i_2} | \} \leq e^{n \rho} \cdot \sup_{|v| \leq 1} \mathbb{E}\{ |v| |v^T X|^2 \} \leq e^{n \rho} C_1.$$ 

Therefore, we can apply Lemma 8 to each coordinate and yield

$$\mathbb{P}\left( \| \Sigma^{-1} \hat{\Sigma} - I \|_\infty \geq \tilde{C} \sqrt{\log p/n} \right) \leq p^2 \max_{1 \leq i_1, i_2 \leq p} \mathbb{P}\left( \left| \frac{1}{n} \sum_{i=1}^n \omega_{i_1}^T X_i \cdot e_{i_2}^T X_i - \delta_{i_1, i_2} \right| \geq \tilde{C} \sqrt{\log p/n} \right) = O(p^2 p^{-\gamma - 2}) = O(p^{-\gamma}),$$

for some $\tilde{C}$ sufficiently large. Therefore, the lemma is proved.

Before proving Theorem 4, we first provide the sign consistency of the non-private MajVote Lasso estimator. Namely, we define

$$\bar{Q}(X) = \text{MajVote}(Q_0(\hat{\theta}_j)|1 \leq j \leq m).$$

(27)

Then we have the following proposition.

**Proposition 2.** (Sign consistency of MajVote Lasso) Let $N$ i.i.d. random vectors $X = \{(X_1, Y_1), \ldots, (X_N, Y_N)\}$ be sampled from $P_{X,Y}(\theta^*, \rho, \eta_1, C_1, \eta_2, C_2)$. Moreover, we assume the same conditions as in Theorem 4 holds. Then $\bar{Q}(X)$ defined in (27) satisfies that, there is constant $\gamma_2 > 0$ such that

$$\mathbb{P}\left( \bar{Q}(X) = \text{sgn}(\theta^*) \right) \geq 1 - O(p^{-\gamma_2}).$$

**Proof of Proposition 2.** For each $j \in \{1, \ldots, m\}$, taking sub-gradient of (14) at $\hat{\theta}_j$, we have that

$$-\frac{1}{n_j} \sum_{i \in H_j} (Y_i - X_i^T \hat{\theta}_j) X_i + \lambda_j Z_j = 0,$$

where $Z_j$ is the sub-gradient satisfying $|Z_j|_\infty \leq 1$. Rearranging the terms and multiplying $\Sigma^{-1}$ on the both sides, we have

$$\hat{\theta}_j - \theta^* = (I - \Sigma^{-1} \frac{1}{n_j} \sum_{i \in H_j} X_i X_i^T)(\hat{\theta}_j - \theta^*) + \frac{1}{n_j} \sum_{i \in H_j} \Sigma^{-1} X_i z_i - \lambda_j \Sigma^{-1} Z_j.$$  

(28)

Taking $\eta = \min\{\eta_1, \eta_2\}$, since the noise and covariates are assumed to be sub-Gaussian in (17),
for each coordinate \( l \in \{1, \ldots, p \} \), we have

\[
\max_{1 \leq l \leq p} \mathbb{E} \left\{ \exp \eta_1 |\mathbf{z}^T \mathbf{w}_l \cdot \mathbf{z}| \right\}
\leq \max_{1 \leq l \leq p} \mathbb{E} \left\{ \exp \left( \frac{1}{2} \eta_2 |\mathbf{w}_l \cdot \mathbf{z}|^2 + \frac{1}{2} \eta_2 |\mathbf{z}|^2 \right) \right\}
\leq \max_{1 \leq l \leq p} \left[ \mathbb{E} \{ \exp \eta_2 |\mathbf{w}_l \cdot \mathbf{z}|^2 \} \cdot \mathbb{E} \{ \exp \eta_2 |\mathbf{z}|^2 \} \right]^{1/2} \leq \sqrt{C_1 C_2}.
\]

Therefore by Lemma 8, we know that there exists a constant \( C_1 > 0 \) such that

\[
\max_{1 \leq j \leq m} \left| \frac{1}{n_j} \sum_{i \in H_j} \mathbf{z}_i \mathbf{X}_{i,j} \right|_{\infty} \leq C_1 \sqrt{\frac{\log p}{n}}.
\]

(29)

with probability larger than \( 1 - O(p^{-\gamma}) \). On the other hand, by the fact that \( |\mathbf{Z}_j|_{\infty} \leq 1 \), we have

\[
|\lambda_j \mathbf{Z}_j|_{\infty} \leq \lambda_j \| \Sigma^{-1} \|_{L_\infty}.
\]

(30)

Moreover, since \( |\hat{\mathbf{\theta}}_j|_0 \leq \tilde{s} \) by Lemma 10, we know

\[
\max_{1 \leq j \leq m} \left| \left( I - \Sigma^{-1} \frac{1}{n_j} \sum_{i \in H_j} \mathbf{X}_i \mathbf{X}_i^T \right) (\hat{\mathbf{\theta}}_j - \mathbf{\theta}^*) \right|_{\infty} = O_p \left( \tilde{s} \sqrt{\frac{\log p}{n}} \right) \leq \frac{1}{2} |\hat{\mathbf{\theta}}_j - \mathbf{\theta}^*|_{\infty}.
\]

(31)

By substituting equations (29), (30), and (31) into (28), we have

\[
|\hat{\mathbf{\theta}}_j - \mathbf{\theta}^*|_{\infty} \leq 2\lambda_j \| \Sigma^{-1} \|_{L_\infty} + 2\tilde{C}_1 \sqrt{\frac{\log p}{n}}.
\]

(32)

From (28), the \( l \)-th coordinate can be written in the following form

\[
\hat{\theta}_{j,l} - \theta^*_l = -\lambda_j \omega_{l,l} Z_{j,l} - \lambda_j \omega_{l,-l}^T Z_{j,-l} + \frac{1}{n_j} \sum_{i \in H_j} \omega_{l,i}^T \mathbf{X}_i \mathbf{z}_i + o_p \left( \frac{\log p}{n} \right).
\]

(33)

It left to rehash the argument in the proof of Theorem 1. From Lemma 11 below we have that

\[
1 - P \left( \mathcal{Q}(\mathcal{X}) = \text{sgn}(\mathbf{\theta}^*) \right) \leq p \cdot \max_{1 \leq l \leq p} P \left( \mathcal{Q}_l(\mathcal{X}) \neq \text{sgn}(\mathbf{\theta}^*_l) \right) = O(p^{-\gamma + 1}).
\]

Therefore we proved Proposition 2. \( \square \)

Lemma 11. Assume the same conditions in Proposition 2. For every \( 1 \leq l \leq p \), we have

\[
P \left( \mathcal{Q}_l(\mathcal{X}) \neq \text{sgn}(\mathbf{\theta}^*_l) \right) = O(p^{-\gamma}).
\]
for arbitrarily fixed \( \gamma > 1 \).

**Proof.** When \( \theta_l^* = 0 \), we know that

\[
\mathbb{P}\left( Q_l(X) \neq \text{sgn}(\theta_l^*) \right) \leq \mathbb{P}\left\{ \sum_{j=1}^{m} \mathbb{I}(\hat{\theta}_{j,l} < 0) \geq \frac{m}{2} \right\} + \mathbb{P}\left\{ \sum_{j=1}^{m} \mathbb{I}(\hat{\theta}_{j,l} > 0) \geq \frac{m}{2} \right\}.
\]

By symmetry of the formulation, it is enough to prove

\[
\mathbb{P}\left\{ \sum_{j=1}^{m} \mathbb{I}(\hat{\theta}_{j,l} > 0) \geq \frac{m}{2} \right\} = O(p^{-\gamma}).
\] (33)

When \( \hat{\theta}_{j,l} > 0 \), we know that

\[
Z_{j,l} = 1 \ (\text{see (32)}).
\]

Moreover, by condition (c), we have

\[
0 < \hat{\theta}_{j,l} \leq -\lambda_j \omega_{l,l} + \lambda_j |\omega_{l,-l}|_{H_j} + \frac{1}{n_j} \sum_{i \in H_j} \omega_i^T X_i z_i + o_p\left( \frac{\log p}{n} \right)
\]

\[
\leq -\Delta_0 \lambda_j \omega_{l,l} + \frac{1}{n_j} \sum_{i \in H_j} \omega_i^T X_i z_i + o_p\left( \frac{\log p}{n} \right)
\]

\[
\leq -\frac{1}{2} \Delta_0 \lambda_N \omega_{l,l} + \frac{1}{n_j} \sum_{i \in H_j} \omega_i^T X_i z_i.
\]

Therefore from (32) we have that

\[
\mathbb{P}\left\{ \sum_{j=1}^{m} \mathbb{I}(\hat{\theta}_{j,l} > 0) \geq \frac{m}{2} \right\} \leq \mathbb{P}\left\{ \sum_{j=1}^{m} \mathbb{I}\left( \frac{1}{n_j} \sum_{i \in H_j} \omega_i^T X_i z_i \geq \frac{1}{2} \Delta_0 \lambda_N \omega_{l,l} \right) \geq \frac{m}{2} \right\}. \tag{34}
\]

Applying Lemma 9 with \( k = 0 \) to the i.i.d. random variables \( \omega_i^T X_i z_i \), we can prove (33) by taking

\[
\lambda_N \geq \frac{2\tilde{C}_1}{\Delta_0 \omega_{l,l}} \left( \sqrt{\frac{\log p}{mn}} + \frac{1}{n} \right), \tag{35}
\]

with \( \tilde{C} \) sufficiently large. Repeat the argument for the other half, we can prove the case for \( \theta_l^* = 0 \).

When \( \theta_l^* > 0 \), again from equation (32), we have

\[
\hat{\theta}_{j,l} \geq \theta_l^* + \frac{1}{n_j} \sum_{i \in H_j} \omega_i^T X_i z_i - \| \Sigma^{-1} \|_{L_\infty} \lambda_j + o_p\left( \frac{\log p}{n} \right).
\]

Therefore

\[
\mathbb{P}\left( Q_l(X) \neq \text{sgn}(\theta_l^*) \right) \leq \mathbb{P}\left\{ \sum_{j=1}^{m} \mathbb{I}(\hat{\theta}_{j,l} \leq 0) \geq \frac{m}{2} \right\}
\]

\[
\leq \mathbb{P}\left\{ \sum_{j=1}^{m} \mathbb{I}\left( -\frac{1}{n_j} \sum_{i \in H_j} \omega_i^T X_i z_i + \theta_l^* - 2 \| \Sigma^{-1} \|_{L_\infty} \lambda_j \geq \frac{m}{2} \right) \right\}. \tag{36}
\]
Applying Lemma 9 we can show that

\[ P \left\{ \sum_{j=1}^{m} \mathbb{I} \left( -\frac{1}{n} \sum_{i \in \mathcal{H}_j} \omega_i^T X_i z_i \geq \theta^*_j - 2 \| \Sigma^{-1} \|_{L_\infty} \lambda_j \right) \geq \frac{m}{2} \right\} = O(p^{-\gamma}), \]

provided that

\[ \theta^*_j - 2 \| \Sigma^{-1} \|_{L_\infty} \lambda_j \geq \tilde{C}_2 \left( \sqrt{\frac{\log p}{mn}} + \frac{1}{n} \right), \]

for all \( j \in \{1, \ldots, m\} \) with \( \tilde{C}_2 \) sufficiently large. Therefore we have that

\[ \theta^*_j \geq \tilde{C}_3 \left( \sqrt{\frac{\log p}{mn}} + \frac{1}{n} + \max_{1 \leq j \leq m} \lambda_j \right), \]

for some \( \tilde{C}_3 > 0 \). When \( \theta^*_j < 0 \), the prove is essentially the same as above, hence we omit it for brevity. Thus the lemma is proved.

Proof of Theorem 4. Similarly as the proof of Proposition 3, for \( l \in S \) such that \( \text{sgn}(\theta^*_l) = \text{sgn}(Q_l(X)) = 1 \), we wish there holds

\[ f^S(Q^*_l) = 2 \sum_{j=1}^{m} \mathbb{I} \left( \tilde{\theta}_{j,l} > 0 \right) - m \geq 8(\gamma + 1) \sqrt{2\tilde{s} \log(2/\delta)} \log p/\epsilon. \]

By (36) we know

\[ \mathbb{P}(f^S(Q^*_l) \geq 8(\gamma + 1) \sqrt{2\tilde{s} \log(2/\delta)} \log p/\epsilon) \]

\[ = \mathbb{P} \left( \sum_{j=1}^{m} \mathbb{I} \left( -\frac{1}{n} \sum_{i \in \mathcal{H}_j} \omega_i^T X_i z_i \geq \theta^*_j - 2 \| \Sigma^{-1} \|_{L_\infty} \lambda_j \right) - \frac{m}{2} \geq 4(\gamma + 1) \sqrt{2\tilde{s} \log(2/\delta)} \log p/\epsilon \right) \]

\[ \geq 1 - O(p^{-\gamma}), \]

as long as

\[ \theta^*_j - 2 \| \Sigma^{-1} \|_{L_\infty} \max_{1 \leq j \leq m} \lambda_j \geq \tilde{C} \left( \frac{1}{n} + \frac{4(\gamma + 1) \sqrt{2\tilde{s} \log(2/\delta)} \log p}{m \sqrt{mn}} + \frac{\sqrt{\log p}}{mn} \right), \]

which proves the first part of the proposition. Similarly, when \( \tilde{s} > s \), for \( l \notin S \), there is

\[ f_l(Q^*_l, 0) = \min \left\{ 2 \sum_{j=1}^{m} \mathbb{I} \left( \tilde{\theta}_{j,l} \leq 0 \right), 2 \sum_{j=1}^{m} \mathbb{I} \left( \tilde{\theta}_{j,l} \geq 0 \right) \right\} - m. \]
Therefore by (34),

\[
\mathbb{P}\left( f_1(Q^*_l, 0) \geq 8(\gamma + 1)\sqrt{2s \log(2/\delta)} \log p/\epsilon \right)
\]

\[
\geq 1 - \mathbb{P}\left( \sum_{j=1}^{m} \mathbb{I}\left( \frac{1}{n_j} \sum_{i \in \mathcal{H}_j} \omega_i^T x_i z_i \leq \frac{1}{2} \Delta_0 \lambda_N \omega_{l,l} \right) - \frac{m}{2} \leq 4(\gamma + 1)\sqrt{2s \log(2/\delta)} \log p/\epsilon \right)
\]

\[
- \mathbb{P}\left( \sum_{j=1}^{m} \mathbb{I}\left( \frac{1}{n_j} \sum_{i \in \mathcal{H}_j} \omega_i^T x_i z_i \geq -\frac{1}{2} \Delta_0 \lambda_N \omega_{l,l} \right) - \frac{m}{2} \leq 4(\gamma + 1)\sqrt{2s \log(2/\delta)} \log p/\epsilon \right)
\]

\[
\geq 1 - O(p^{-\gamma}),
\]

as long as

\[
\lambda_N \geq \tilde{C} \left( \frac{1}{n} + \frac{4(\gamma + 1)\sqrt{2s \log(2/\delta)} \log p}{mn\sqrt{n\epsilon}} + \sqrt{\frac{\log p}{mn}} \right).
\]