DIFFERENTIALLY PRIVATE BEST SUBSET SELECTION VIA INTEGER PROGRAMMING

Kayhan Behdin & Peter Prastakos
Operations Research Center
Massachusetts Institute of Technology
Cambridge, MA 02139, USA
{behdink,pprastak}@mit.edu

Rahul Mazumder
Operations Research Center
Sloan School of Management
Massachusetts Institute of Technology
Cambridge, MA 02139, USA
rahulmaz@mit.edu

ABSTRACT

We study the Best Subset Selection (BSS) problem under Differential Privacy (DP), where one seeks to find a subset of features and a linear estimator supported on the said subset that minimize the least squares error. Due to the combinatorial structure, solving BSS has been traditionally challenging. However, thanks to recent advancements in Mixed Integer Programming (MIP), solving large-scale (non-private) BSS with millions of variables has become feasible in minutes. Despite this, the application of such algorithmic developments to BSS under DP has remained limited. In this paper, we propose a new DP estimator for model selection in BSS. Particularly, inspired by the exponential mechanism, we design a new sampling procedure which makes use of MIP to explore the (non-convex) objective landscape of BSS in a structured fashion. This circumvents the need for combinatorial exhaustive search as required by the exponential mechanism. This allows us to solve BSS with hundreds of features and $(\varepsilon, 0)$-DP guarantees in minutes, whereas standard (exhaustive search based) exponential mechanism would require tens of days of runtime and tens of terabytes of storage. To our knowledge, our work is a first to employ techniques from integer programming to design a differentially private algorithm for BSS, specially under pure differential privacy.

1 INTRODUCTION

With the advancement of data acquisition, processing and storage systems, there has been a growing interest in learning from high-dimensional datasets. However, extracting meaningful models from high-dimensional data can be challenging, for example, due to overfitting or lack of model interpretability. Statistical regularizations that encourage model simplicity from certain perspectives have been successful in addressing such challenges, becoming a staple of high-dimensional statistics and machine learning. One such common regularization is sparsity (Hastie et al., 2015; 2009), where one seeks to choose a small subset of features in the data to form the statistical model.

In this paper, we focus on the sparse linear regression problem. Given the data matrix $\boldsymbol{X} \in \mathbb{R}^{n\times p}$ and the observations $\boldsymbol{y} \in \mathbb{R}^n$, in the sparse linear regression one seeks to obtain a linear estimator such as $\boldsymbol{\beta} \in \mathbb{R}^p$ that describes the data well (i.e., $\boldsymbol{y} \approx \boldsymbol{X}\boldsymbol{\beta}$) with only a few coordinates of $\boldsymbol{\beta}$ being nonzero. A natural first formulation for this problem is the Best Subset Selection (BSS, Miller (2002)),

$$\min_{\boldsymbol{\beta}\in\mathbb{R}^p} \|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\|_2^2 \text{ s.t. } \|\boldsymbol{\beta}\|_0 \le s,\ \|\boldsymbol{\beta}\|_2^2 \le r^2 \tag{1}$$

where $\|\cdot\|_0$ counts the number of nonzero coordinates of a vector. In (1), the objective is the least squares loss, ensuring the resulting estimator describes the data as well as possible and $s > 0$ is the sparsity budget, enforcing the sparsity of $\boldsymbol{\beta}$ through the constraint $\|\boldsymbol{\beta}\|_0 \leq s$. Finally, $\|\boldsymbol{\beta}\|_2^2 \leq r^2$ for some $r > 0$ serves as an additional (ridge) regularization. For a linear estimator such as $\boldsymbol{\beta} \in \mathbb{R}^p$, we call $\{i : \beta_i \neq 0\}$ the support of $\boldsymbol{\beta}$. A sparse linear estimator can be more interpretable and have better statistical performance (Hastie et al., 2009; 2015; Wainwright, 2019).

Real-world datasets often contain confidential and personal information, that should be protected. Hence, recent years have seen a surge in private learning algorithms, hoping to preserve sensitive information while extracting useful statistical knowledge. In particular, Differential Privacy (DP, Dwork (2006)) has garnered significant interest in the machine learning and statistics literature. On a high level, DP aims to ensure one cannot obtain too much information from the private dataset, via querying the statistical model in an adversarial way. A significant body of work is dedicated to designing DP algorithms for general machine learning tasks (McSherry & Talwar, 2007; Dwork et al., 2006; Hardt et al., 2012; Dwork et al., 2010; 2014), as well as specialized algorithms for specific statistical problems. Particularly, there is a long line of work studying the sparse linear regression problem.

In this paper, we develop a new differentially private algorithm for model and variable selection in the BSS problem where one releases the optimal support in (1) (i.e., the location of nonzero coordinates in the optimal $\boldsymbol{\beta}$). In what follows, we review the existing work on both non-private and private sparse linear regression, and then discuss our approach and contributions.

1.1 Related Work

Best Subset Selection (BSS) The BSS problem in (1) serves as a natural formulation for the sparse linear regression, derived from the first principles. Hence, several papers have studied statistical properties of BSS, showing that BSS enjoys strong prediction and variable selection properties (Rad, 2009; Wainwright, 2009a; Guo et al., 2020). Despite strong theoretical and practical strengths of BSS, solving the BSS problem can be computationally challenging due to the combinatorial structure (Natarajan, 1995). Therefore, several relaxations and approximations to BSS have been developed such as Lasso (Tibshirani, 1996), Elastic Net (Zou & Hastie, 2005) and non-concave penalties (Zhang, 2010; Fan & Li, 2001). Theoretical properties of such estimators is well-studied in the literature (Zhao & Yu, 2006; Zhang & Huang, 2008; Wainwright, 2009b).

Thanks to the recent advancements in discrete optimization and Mixed Integer Programming (MIP, Wolsey & Nemhauser (1999)), there has been a renewed interest in studying sparse learning problems that admit a combinatorial structure. Particularly, starting with the work of Bertsimas et al. (2016), a growing body of work has been dedicated to developing efficient and scalable algorithms for the BSS problem. Examples include cutting plane algorithms (Bertsimas & Van Parys, 2020), Branch-and-Bound methods (Hazimeh et al., 2022), as well as approximate algorithms (Hazimeh & Mazumder, 2020). With new optimization algorithms for BSS, obtaining global solutions to BSS problems with millions of variables has become possible in minutes (Hazimeh et al., 2022).

The feasibility of solving large-scale BSS has triggered a new wave of theoretical and empirical studies, investigating the differences between BSS and its convex relaxations such as Lasso. In particular, recent work suggests that BSS can perform better than convex estimators in certain Signal to Noise Ratio (SNR) setups (Mazumder et al., 2023; Hastie et al., 2020; Guo et al., 2023), can have better false discovery control (Zhu & Wu, 2021) and can be robust to design dependence and colinearity (Guo et al., 2020; Zhu & Wu, 2021). This motivates us to directly study (non-convex) BSS, rather than a convex relaxation such as Lasso.

Private Linear Regression Due to practical and methodological importance of (sparse) linear regression, numerous papers have developed DP algorithms for this problem. Most of the existing literature focuses on non-sparse linear regression, or ℓ_2 risk excess in sparse regression (Talwar et al., 2015; Varshney et al., 2022; Jain & Thakurta, 2014; Cai et al., 2021; Kasiviswanathan & Jin, 2016). However, as we mentioned, our focus in this paper is on variable selection, i.e., what features in the data should have nonzero coefficients in the sparse linear estimator. We also note that some papers have studied the problem of removing correlated features from the data (Dick et al., 2024), which is different from BSS.

Other papers have studied variable selection under differential privacy. As Lasso tends to promote sparsity, an interesting line of work is based on releasing the variables selected by Lasso in a private fashion (Thakurta & Smith, 2013; Liu et al., 2022; Kifer et al., 2012), using mechanisms such as private majority voting or resample-and-aggregate (Nissim et al., 2007). Lei et al. (2018) propose an algorithm based on the exponential mechanism, requiring to enumerate all feasible supports in (1), limiting the scalability of their method. Recently, Roy & Tewari (2023) have proposed a new method based on the notion of Markov chain mixing to obtain DP solutions for BSS, resulting in a statistically strong estimator.

1.2 Our contributions

Despite significant work done in private variable selection, there has been limited exploration regarding how to use recently developed discrete optimization algorithms and techniques for BSS in a private setting, leaving a gap between private and non-private BSS algorithms. In this work, we bridge this gap by introducing a new differentially private algorithm for BSS, building on recent algorithmic advancements pertaining to BSS. Our algorithm makes extensive use of discrete optimization techniques for BSS, an approach with limited exploration in DP literature.

On a high level, our algorithm builds on the following observation: In general, when applied to a problem with finite output space such as support selection in BSS, the exponential mechanism (McSherry & Talwar, 2007) has a non-zero probability of outputting each solution in the outcome space, which makes applying this mechanism infeasible for problems with a large outcome space. However, by design, this mechanism is most likely to output solutions that are close to the optimal solution in an "objective" sense. Therefore, if one can enumerate such near optimal solutions, it is possible to closely approximate the exponential mechanism without the need to sample from a probability distribution with exponentially large outcome space. As it turns out, in the BSS problem, MIP enables us to quickly enumerate such solutions that are likely to be returned by the exponential mechanism.

Our contributions can be summarized as follows: **(1)** We modify the exponential mechanism to design a DP variable selection method for the (non-convex) BSS problem with the cardinality constraint. **(2)** Our selection procedure does not require enumerating all $\binom{p}{s}$ feasible s-sparse supports in (1). Instead, our method "approximates" the exponential mechanism by enumerating feasible solutions that are likely to be returned by the mechanism. To obtain such likely supports, our procedure requires solving a series of convex MIPs. **(3)** We show numerically, our sampling procedure can obtain BSS solutions with hundreds of features in minutes, and with $(\varepsilon, 0)$-DP guarantees. An exhaustive enumeration method would require days for such problem sizes. We do not know of any existing scalable $(\varepsilon, 0)$-DP method for BSS.

Notation. We let $[p] = \{1, \cdots, p\}$. Throughout this paper, we assume the data points follow $\boldsymbol{x}_i \in \mathcal{X} \subseteq \mathbb{R}^p, y_i \in \mathcal{Y} \subseteq \mathbb{R}$ for $i \in [n]$. We also let $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$. The rows of matrix $\boldsymbol{X} \in \mathbb{R}^{n \times p}$ contain $\boldsymbol{x}_1, \cdots, \boldsymbol{x}_n$, while $\boldsymbol{y} \in \mathbb{R}^n$ contains the values of $y_1, \cdots, y_n$. We also use the notation $\mathcal{D} = (\boldsymbol{X}, \boldsymbol{y}) \in \mathcal{Z}^n$ for a dataset containing n observations.

2 Method

Background on Differential Privacy Before continuing with our selection procedure, let us formalize the notion of differential privacy.

Definition 1 (Dwork (2006))**.** Given the privacy parameters $(\varepsilon, \delta) \in \mathbb{R}^+ \times \mathbb{R}^+$, a randomized algorithm $\mathcal{A}(\cdot)$ is said to satisfy the (ε, δ)-DP property if

$$\mathbb{P}(\mathcal{A}(\mathcal{D}) \in K) \leq e^{\varepsilon} \mathbb{P}(\mathcal{A}(\mathcal{D}') \in K) + \delta$$

for any measurable event $K \in \text{range}(\mathcal{A})$ and for any pair of neighboring datasets $\mathcal{D}$ and $\mathcal{D}'$.

We note that in Definition 1, the probability is taken over the randomness of the algorithm $\mathcal{A}$. When $\delta = 0$, the special case of $(\varepsilon, 0)$-DP is commonly referred to as pure differential privacy.

Next, let us briefly review the exponential mechanism (McSherry & Talwar, 2007), a general mechanism to achieve pure DP. Consider a general task where the dataset $\mathcal{D} \in \mathcal{Z}^n$ is given, and we seek to design a procedure such as $\mathcal{A} : \mathcal{Z}^n \to \mathcal{O}$ to choose the outcome of the task, where $\mathcal{O}$ is the set of

possible outcomes. We also assume we are given an objective function such as $\mathcal{R} : \mathcal{O} \times \mathcal{Z}^n \to \mathbb{R}$, where a smaller objective indicates a more desirable outcome. The global sensitivity of the objective is then defined as

$$\Delta = \max_{o \in \mathcal{O}} \max_{\substack{\mathcal{D}, \mathcal{D}' \in \mathcal{Z}^n \\ \mathcal{D}, \mathcal{D}' \text{ are neighbors}}} \mathcal{R}(o, \mathcal{D}) - \mathcal{R}(o, \mathcal{D}'). \tag{2}$$

Lemma 1 (Exponential Mechanism, McSherry & Talwar (2007))**.** *The exponential mechanism* $\mathcal{A}_E(\cdot)$ *that follows*

$$\mathbb{P}(\mathcal{A}_E(\mathcal{D}) = o) \propto \exp\left(-\frac{\varepsilon \mathcal{R}(o, \mathcal{D})}{2\Delta}\right), \ \ \forall o \in \mathcal{O} \tag{3}$$

ensures $(\varepsilon, 0)$*-DP.*

2.1 Selection Procedure

A First Attempt The main inspiration for our selection procedure is the exponential mechanism, defined in Lemma 1. In particular, in the BSS problem we seek to select a subset of features with size s that are a good linear predictor of n observations $\boldsymbol{y}$. Therefore, a natural choice for the outcome set in BSS is the set of all subsets of $[p]$ with size s, $\mathcal{O} = \{S \subseteq [p] : |S| = s\}$. Next, a natural choice for the objective in the BSS problem for each S is the least squares loss, when the regression coefficients can only be nonzero for features in S. Formally,

$$\mathcal{R}(S, \mathcal{D}) = \min_{\boldsymbol{\beta} \in \mathbb{R}^{|S|}} \|\boldsymbol{y} - \boldsymbol{X}_S \boldsymbol{\beta}\|_2^2 \ \text{ s.t. } \ \|\boldsymbol{\beta}\|_2^2 \leq r^2 \tag{4}$$

where $\boldsymbol{X}_S$ is the submatrix of $\boldsymbol{X}$ with columns indexed by S. As formalized in Lemma 2 below, we can bound the global sensitivity of $\mathcal{R}(\mathcal{D}, S)$.

Lemma 2. *Suppose that* $|y| \leq b_y$ *for* $y \in \mathcal{Y}$*, and* $\|\boldsymbol{x}\|_\infty \leq b_x$ *for* $\boldsymbol{x} \in \mathcal{X}$*. Then,*

$$\Delta \leq 2b_y^2 + 2b_x^2 r^2 s.$$

Remark 1. In practice, one might not know the exact values of b_x, b_y, or such values might not exist. In such cases, one can clip the values of $\boldsymbol{X}, \boldsymbol{y}$ to satisfy the boundedness requirements of Lemma 2.

Remark 2. We also note that the bound in Lemma 2 improves upon the global sensitivity of least squares as proposed by Roy & Tewari (2023). In particular, our bound holds for all $\mathcal{D}$, while the bound of Roy & Tewari (2023) holds with high probability leading to DP guarantees that only hold with high probability. Moreover, our sensitivity bound only needs coordinate-wise bounds on the data, while Roy & Tewari (2023)'s requires additional assumptions on the data generation mechanism, as well as bounds on eigenvalues of numerous submatrices of $\boldsymbol{X}$ which can be hard to certify.

Lemma 2 shows that one can directly apply the exponential mechanism on $\mathcal{R}(S, \mathcal{D})$ and achieve a $(\varepsilon, 0)$-DP procedure for the BSS problem. The issue, however, is that forming the probability distribution in (3) requires calculating $\mathcal{R}(S, \mathcal{D})$ for all $S \in \mathcal{O}$, which can be computationally infeasible even for small problems (Roy & Tewari, 2023) as $|\mathcal{O}| = \binom{p}{s}$ which grows exponentially with s.

Our Proposal As discussed above, the difficulty in solving BSS under DP constraints arises from the need to enumerate all feasible solutions in $\mathcal{O}$. However, one can argue that if a support S is far from the optimal one, the least-squares objective $\mathcal{R}(S, \mathcal{D})$ is likely to be large, therefore, the probability mass of S in (3) should be small. Therefore, one might ask:

Is it necessary to have access to $\mathcal{R}(S, \mathcal{D})$ ***for all*** $S \in \mathcal{O}$ ***in the exponential mechanism?***

Specifically, for the moment, suppose we have access to an oracle that for a fixed $R > 1$, can return R feasible supports from $\mathcal{O}$ that have the smallest objectives. Formally, assume we can access $\hat{S}_1(\mathcal{D}), \cdots, \hat{S}_R(\mathcal{D})$ where

$$\hat{S}_k(\mathcal{D}) \in \arg\min_S \mathcal{R}(S, \mathcal{D}) \ \text{ s.t. } \ S \subseteq [p], \ |S| = s, \ S \neq \hat{S}_i(\mathcal{D}), i = 1, \cdots, k-1. \tag{5}$$

In particular, $\hat{S}_1(\mathcal{D})$ is the optimal support for BSS in (1). Then, based on our discussion above, if R is sufficiently large, the values $\mathcal{R}(\hat{S}_k(\mathcal{D}), \mathcal{D})$ for $k \geq R$ are expected to be significantly larger

than $\mathcal{R}(\hat{S}_k(\mathcal{D}), \mathcal{D})$ for $k \ll R$. Therefore, most of the probability mass of the distribution in (3) is concentrated around $\hat{S}_k(\mathcal{D})$ for $k \ll R$. Hence, we might not need to have access to the exact values of $\mathcal{R}(\hat{S}_k(\mathcal{D}), \mathcal{D})$ for $k \geq R$, as long as we can replace them with a suitable lower bound. This lower bound can be taken as $\mathcal{R}(\hat{S}_R(\mathcal{D}), \mathcal{D})$. To this end, we propose the sampling procedure $\mathcal{M}$, shown as Algorithm 1 below, where $\mathbb{P}_0$ is the probability distribution following

$$\mathbb{P}_0(k) \propto \begin{cases} \exp\left(-\varepsilon \mathcal{R}(\hat{S}_k(\mathcal{D}), \mathcal{D})/(2\Delta)\right) & \text{if } k \leq R \\ \left(\binom{p}{s} - R\right) \exp\left(-\varepsilon \mathcal{R}(\hat{S}_R(\mathcal{D}), \mathcal{D})/(2\Delta)\right) & \text{if } k = R+1. \end{cases} \tag{6}$$

Algorithm 1 BSS with DP guarantees

procedure $\mathcal{M}(\mathcal{D}, b_x, b_y, r, R, T)$
  Clip $\boldsymbol{X}, \boldsymbol{y}$ to b_x, b_y, respectively, as in Lemma 2. Take Δ as in Lemma 2. Form $\mathbb{P}_0$ in (6).
  Draw $a(\mathcal{D}) \sim \mathbb{P}_0$
  if $a(\mathcal{D}) \leq R$ **then**
    return $\hat{S}_{a(\mathcal{D})}(\mathcal{D})$
  else
    return $\mathcal{M}_0(\mathcal{D}, R, T)$
procedure $\mathcal{M}_0(\mathcal{D}, R, T)$
  for $t \leq T$ **do**
    Draw $S \in \mathcal{O}$ uniformly at random, independent of $\mathbb{P}_0$.
    if $S \in \{\hat{S}_k(\mathcal{D}), k > R\}$ **then**
      Break
  return S

Intuitively speaking, $\mathcal{M}$ replaces $\mathcal{R}(\hat{S}_k(\mathcal{D}), \mathcal{D})$ for $k \geq R$ with $\mathcal{R}(\hat{S}_R(\mathcal{D}), \mathcal{D})$ and then "approximately" samples from the exponential mechanism. To this end, let

$$\hat{\mathcal{R}}(S, \mathcal{D}) = \begin{cases} \mathcal{R}(S, \mathcal{D}) & \text{if } S \in \{\hat{S}_1(\mathcal{D}), \cdots, \hat{S}_R(\mathcal{D})\} \\ \mathcal{R}(\hat{S}_R(\mathcal{D}), \mathcal{D}) & \text{otherwise} \end{cases} \tag{7}$$

where we substitute $\mathcal{R}(\hat{S}_k(\mathcal{D}), \mathcal{D})$ for $k \geq R$ with $\mathcal{R}(\hat{S}_R(\mathcal{D}), \mathcal{D})$. Suppose $\hat{\mathcal{A}}_E$ is the exponential mechanism that uses the objective $\hat{\mathcal{R}}$. If $a(\mathcal{D}) \leq R$ in Algorithm 1, we return $\hat{S}_{a(\mathcal{D})}(\mathcal{D})$. Note that $\mathbb{P}(\hat{\mathcal{A}}_E(\mathcal{D}) = \hat{S}_{a(\mathcal{D})}(\mathcal{D})) = \mathbb{P}_0(a(\mathcal{D}))$ in this case, showing $\mathcal{M}$ mimics the exponential mechanism $\hat{\mathcal{A}}_E$. If $a(\mathcal{D}) = R+1$, to mimic $\hat{\mathcal{A}}_E$, we have to sample uniformly from the set $\mathbb{S} = \{\hat{S}_k(\mathcal{D}), k > R\}$ as $\mathbb{P}(\hat{\mathcal{A}}_E)$ is uniform on $\mathbb{S}$, by the definition of $\hat{\mathcal{R}}$ in (7). However, $\mathbb{S}$ is exponentially large in general. Therefore, we invoke $\mathcal{M}_0$ that in the limit of $T \to \infty$, samples uniformly from $\mathbb{S}$. Below, we show this procedure is indeed DP.

Theorem 1. *Suppose $T > 1$, $1 < R < \binom{p}{s}$. The procedure $\mathcal{M}$ in Algorithm 1 is $(\varepsilon', 0)$-DP where*

$$\varepsilon' = \log\left(e^{\varepsilon} + \gamma\right) - \log\left(1 - q^T\right), \quad \gamma = \frac{R^T \exp(n\varepsilon b_y^2/(2\Delta))}{\binom{p}{s}^{T-1}}, \quad q = \frac{R}{\binom{p}{s}}.$$

In particular, if $T = \infty$, the procedure $\mathcal{M}$ is $(\varepsilon, 0)$-DP. Moreover, $\mathbb{P}(\mathcal{M}(\mathcal{D}) = \hat{S}_1(\mathcal{D})) \geq \mathbb{P}_0(1)$.

The sampling procedure in Algorithm 1 only requires sampling from $\mathbb{P}_0$ (which is supported on $R+1$ different values), and sampling sparse supports from a uniform distribution (in procedure $\mathcal{M}_0$), which can be done efficiently. Therefore, this procedure circumvents the need to sample from a non-uniform distribution with exponentially large support. Importantly, Algorithm 1 satisfies pure $(\varepsilon, 0)$-DP. To our knowledge, no such algorithm exists for BSS that can scale to problems with hundreds of variables. Moreover, the final part of Theorem 1 shows the probability that $\mathcal{M}$ outputs the optimal BSS support $\hat{S}_1$ is at least $\mathbb{P}_0(1)$, showing $\mathcal{M}$ does not perform worse than the exponential mechanism that uses $\hat{\mathcal{R}}$, $\hat{\mathcal{A}}_E$, in terms of utility.

MIP Formulations Next, we discuss how the top R supports in terms of least squares loss, $\hat{S}_1(\mathcal{D}), \cdots, \hat{S}_R(\mathcal{D})$, can be obtained by solving a series of MIPs. In particular, for $k \in [R]$ consider:

$$
\begin{aligned}
(\boldsymbol{z}^{(k)}, \boldsymbol{\beta}^{(k)}, \boldsymbol{\theta}^{(k)}) \in \underset{\boldsymbol{z},\boldsymbol{\beta},\boldsymbol{\theta}}{\arg\min} \quad & \|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\|_2^2 \qquad (8)\\
\text{s.t.} \quad & \boldsymbol{\beta}, \boldsymbol{\theta} \in \mathbb{R}^p, \boldsymbol{z} \in \{0,1\}^p,\ \boldsymbol{\theta} \geq 0,\ \sum_{i=1}^p z_i = s,\ \sum_{i=1}^p \theta_i \leq r^2,\\
& \beta_i^2 \leq \theta_i z_i \ \ \forall i \in [p]\\
& \sum_{i \in \hat{S}_j(\mathcal{D})} z_i \leq s - \frac{1}{2},\ \ j = 1, \cdots, k-1.
\end{aligned}
$$

Proposition 1. *For $k \geq 1$, $\{i : z_i^{(k)} \neq 0\} = \hat{S}_k(\mathcal{D})$.*

Problem (8) involves a convex quadratic objective function. Moreover, if one relaxes the binary variables $\boldsymbol{z}$ to take values in $[0,1]^p$, the feasible set of (8) becomes a second order cone which is convex. In other words, the interval relaxation of (8) is convex. Therefore, Problem (8) is typically known as a Mixed-Integer Second Order Cone Program (MISOCP), a standard class of MIPs. Most modern off-the-shelf solvers such as Gurobi and Mosek are able to obtain globally optimal solutions to Problem (8) for moderately-sized datasets. As we demonstrate in Section 3, this allows us to scale our DP algorithm to problems with $p \approx 250$ in a few minutes, benefiting from the enormous progress made in the discrete optimization literature in the recent years.

Finally, let us also discuss the role of R in Algorithm 1. For $k > R$, we underestimate $\mathcal{R}(\hat{S}_k, \mathcal{D})$ with $\mathcal{R}(\hat{S}_R, \mathcal{D})$, consequently increasing the probability mass given to supports $\hat{S}_k$ in the procedure $\hat{\mathcal{A}}_E$. This reduces the probability mass for the best support $\hat{S}_1$. Therefore, in practice, we like to choose a larger R to explore the objective landscape better, however, a very large R can make the computation slower.

3 Numerical Experiments

In our experiments, we draw the data points as $y_i = \boldsymbol{x}_i^T \boldsymbol{\beta}^* + \epsilon_i$ for $i \in [n]$, where $\boldsymbol{x}_1, \cdots, \boldsymbol{x}_n \overset{\text{iid}}{\sim} \mathcal{N}(\boldsymbol{0}, \boldsymbol{\Sigma}) \in \mathbb{R}^p$ and the independent noise follows $\boldsymbol{\epsilon} \sim \mathcal{N}(\boldsymbol{0}, \sigma^2 \boldsymbol{I}_n)$ where $\boldsymbol{I}_n$ is the identity matrix of size n. Moreover, for $i, j \in [p]$, we set $\Sigma_{i,j} = 0.1^{|i-j|}$ and set nonzero coordinates of $\boldsymbol{\beta}^*$ to take value $1/\sqrt{s}$ at indices $\{1, 3, \cdots, 2s-1\}$. We define the Signal to Noise Ratio as SNR $= \|\boldsymbol{X}\boldsymbol{\beta}^*\|_2^2 / \|\boldsymbol{\epsilon}\|_2^2$. We consider various values of n, p, s, SNR and ε.

In Algorithm 1, we set $R = 100, b_x = b_y = 1, r = 1.1$ and $T = \infty$. We use Gurobi to obtain values of $\hat{S}_1, \cdots, \hat{S}_R$ to form $\mathbb{P}_0$ in (6), using the `Solution Pool` option. Our experiments are run on a standard MacBook Pro, and Gurobi returned the top-R solutions in less than 5 minutes for most cases. In Figure 1, we plot the probability of exact correct support recovery by Algorithm 1, averaged over 10 independent repetitions. For comparison, we also tried DP Lasso of Talwar et al. (2015), however, their method failed to recover the exact correct support in all cases. Therefore, we do not plot their results here.

As we see, the probability of recovery increases as we increase n, as one expects. Moreover, we observe that larger values of p make the recovery harder. Interestingly, the effect of increasing p is more announced when s is larger as in the case of $s = 7$. Overall, we also observe decreasing ε reduces the recovery probability, as this requires stronger privacy guarantees.

To demonstrate the power of our MIP-based estimator, we note that for $p = 250, s = 7$, we have $\binom{250}{7} \approx 10^{13}$. To put this in perspective, to use the standard exponential mechanism and enumerate all feasible supports, assuming computing each feasible support takes 10^{-6} seconds and 16 bits of storage, one would need **128 days and 22 terabytes of storage**. This shows the usefulness of our MIP-based estimator in practice, enabling us to solve BSS with DP for problem sizes that otherwise would be prohibitive.

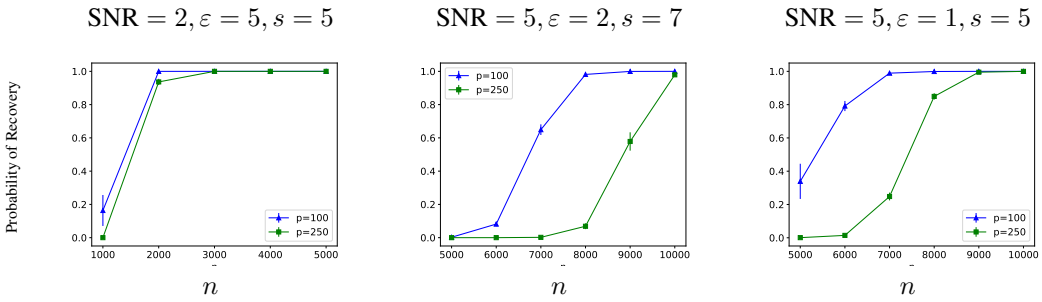

Figure 1: Numerical experiments for three different settings. On the x-axis, we vary the value of n and plot the probability of correct recovery by Algorithm 1, for two different values of p.

REFERENCES

Dimitris Bertsimas and Bart Van Parys. Sparse high-dimensional regression: Exact scalable algorithms and phase transitions. 2020.

Dimitris Bertsimas, Angela King, and Rahul Mazumder. Best subset selection via a modern optimization lens. 2016.

T Tony Cai, Yichen Wang, and Linjun Zhang. The cost of privacy: Optimal rates of convergence for parameter estimation with differential privacy. *The Annals of Statistics*, 49(5):2825–2850, 2021.

Travis Dick, Jennifer Gillenwater, and Matthew Joseph. Better private linear regression through better private feature selection. *Advances in Neural Information Processing Systems*, 36, 2024.

Cynthia Dwork. Differential privacy. In *International colloquium on automata, languages, and programming*, pp. 1–12. Springer, 2006.

Cynthia Dwork, Frank McSherry, Kobbi Nissim, and Adam Smith. Calibrating noise to sensitivity in private data analysis. In *Theory of Cryptography: Third Theory of Cryptography Conference, TCC 2006, New York, NY, USA, March 4-7, 2006. Proceedings 3*, pp. 265–284. Springer, 2006.

Cynthia Dwork, Guy N Rothblum, and Salil Vadhan. Boosting and differential privacy. In *2010 IEEE 51st Annual Symposium on Foundations of Computer Science*, pp. 51–60. IEEE, 2010.

Cynthia Dwork, Aaron Roth, et al. The algorithmic foundations of differential privacy. *Foundations and Trends® in Theoretical Computer Science*, 9(3–4):211–407, 2014.

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

Yilin Guo, Haolei Weng, and Arian Maleki. Signal-to-noise ratio aware minimaxity and higher-order asymptotics. *IEEE Transactions on Information Theory*, 2023.

Yongyi Guo, Ziwei Zhu, and Jianqing Fan. Best subset selection is robust against design dependence. *arXiv preprint arXiv:2007.01478*, 2020.

Moritz Hardt, Katrina Ligett, and Frank McSherry. A simple and practical algorithm for differentially private data release. *Advances in neural information processing systems*, 25, 2012.

Trevor Hastie, Robert Tibshirani, Jerome H Friedman, and Jerome H Friedman. *The elements of statistical learning: data mining, inference, and prediction*, volume 2. Springer, 2009.

Trevor Hastie, Robert Tibshirani, and Martin Wainwright. *Statistical learning with sparsity: the lasso and generalizations*. CRC press, 2015.

Trevor Hastie, Robert Tibshirani, and Ryan Tibshirani. Best subset, forward stepwise or lasso? analysis and recommendations based on extensive comparisons. 2020.

Hussein Hazimeh and Rahul Mazumder. Fast best subset selection: Coordinate descent and local combinatorial optimization algorithms. *Operations Research*, 68(5):1517–1537, 2020.

Hussein Hazimeh, Rahul Mazumder, and Ali Saab. Sparse regression at scale: Branch-and-bound rooted in first-order optimization. *Mathematical Programming*, 196(1-2):347–388, 2022.

Prateek Jain and Abhradeep Guha Thakurta. (near) dimension independent risk bounds for differentially private learning. In Eric P. Xing and Tony Jebara (eds.), *Proceedings of the 31st International Conference on Machine Learning*, volume 32 of *Proceedings of Machine Learning Research*, pp. 476–484, Bejing, China, 22–24 Jun 2014. PMLR.

Shiva Prasad Kasiviswanathan and Hongxia Jin. Efficient private empirical risk minimization for high-dimensional learning. In Maria Florina Balcan and Kilian Q. Weinberger (eds.), *Proceedings of The 33rd International Conference on Machine Learning*, volume 48 of *Proceedings of Machine Learning Research*, pp. 488–497, New York, New York, USA, 20–22 Jun 2016. PMLR.

Daniel Kifer, Adam Smith, and Abhradeep Thakurta. Private convex empirical risk minimization and high-dimensional regression. In *Conference on Learning Theory*, pp. 25–1. JMLR Workshop and Conference Proceedings, 2012.

Jing Lei, Anne-Sophie Charest, Aleksandra Slavkovic, Adam Smith, and Stephen Fienberg. Differentially private model selection with penalized and constrained likelihood. *Journal of the Royal Statistical Society Series A: Statistics in Society*, 181(3):609–633, 2018.

Weidong Liu, Jiyuan Tu, Xiaojun Mao, and Xi Chen. Majority vote for distributed differentially private sign selection. *arXiv preprint arXiv:2209.04419*, 2022.

Rahul Mazumder, Peter Radchenko, and Antoine Dedieu. Subset selection with shrinkage: Sparse linear modeling when the snr is low. *Operations Research*, 71(1):129–147, 2023.

Frank McSherry and Kunal Talwar. Mechanism design via differential privacy. In *48th Annual IEEE Symposium on Foundations of Computer Science (FOCS'07)*, pp. 94–103. IEEE, 2007.

Alan Miller. *Subset selection in regression*. CRC Press, 2002.

Balas Kausik Natarajan. Sparse approximate solutions to linear systems. *SIAM journal on computing*, 24(2):227–234, 1995.

Kobbi Nissim, Sofya Raskhodnikova, and Adam Smith. Smooth sensitivity and sampling in private data analysis. In *Proceedings of the thirty-ninth annual ACM symposium on Theory of computing*, pp. 75–84, 2007.

Kamiar Rahnama Rad. Sharp sufficient conditions on exact sparsity pattern recovery. *arXiv preprint arXiv:0910.0456*, 2009.

Saptarshi Roy and Ambuj Tewari. On the computational complexity of private high-dimensional model selection via the exponential mechanism. *arXiv preprint arXiv:2310.07852*, 2023.

Kunal Talwar, Abhradeep Guha Thakurta, and Li Zhang. Nearly optimal private lasso. *Advances in Neural Information Processing Systems*, 28, 2015.

Abhradeep Guha Thakurta and Adam Smith. Differentially private feature selection via stability arguments, and the robustness of the lasso. In *Conference on Learning Theory*, pp. 819–850. PMLR, 2013.

Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 58(1):267–288, 1996.

Prateek Varshney, Abhradeep Thakurta, and Prateek Jain. (nearly) optimal private linear regression via adaptive clipping. *arXiv preprint arXiv:2207.04686*, 2022.

Martin J Wainwright. Information-theoretic limits on sparsity recovery in the high-dimensional and noisy setting. *IEEE transactions on information theory*, 55(12):5728–5741, 2009a.

Martin J Wainwright. Sharp thresholds for high-dimensional and noisy sparsity recovery using l1-constrained quadratic programming (lasso). *IEEE transactions on information theory*, 55(5): 2183–2202, 2009b.

Martin J Wainwright. *High-dimensional statistics: A non-asymptotic viewpoint*, volume 48. Cambridge university press, 2019.

Laurence A Wolsey and George L Nemhauser. *Integer and combinatorial optimization*, volume 55. John Wiley & Sons, 1999.

Cun-Hui Zhang. Nearly unbiased variable selection under minimax concave penalty. 2010.

Cun-Hui Zhang and Jian Huang. The sparsity and bias of the lasso selection in high-dimensional linear regression. 2008.

Peng Zhao and Bin Yu. On model selection consistency of lasso. *The Journal of Machine Learning Research*, 7:2541–2563, 2006.

Ziwei Zhu and Shihao Wu. On the early solution path of best subset selection. *arXiv e-prints*, pp. arXiv–2107, 2021.

Hui Zou and Trevor Hastie. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 67(2):301–320, 2005.

A PROOFS OF MAIN RESULTS

A.1 PROOF OF LEMMA 2

Proof. Suppose $\mathcal{D}, \mathcal{D}'$ are two neighboring datasets. Fix a support $S \in \mathcal{O}$ and suppose

$$\hat{\boldsymbol{\beta}} \in \arg\min_{\boldsymbol{\beta} \in \mathbb{R}^s} \|\boldsymbol{y}' - \boldsymbol{X}'_S \boldsymbol{\beta}\|_2^2 \text{ s.t. } \|\boldsymbol{\beta}\|_2^2 \leq r^2.$$

Then,

$$\mathcal{R}(S, \mathcal{D}) - \mathcal{R}(S, \mathcal{D}') \leq \|\boldsymbol{y} - \boldsymbol{X}_S \hat{\boldsymbol{\beta}}\|_2^2 - \|\boldsymbol{y}' - \boldsymbol{X}'_S \hat{\boldsymbol{\beta}}\|_2^2.$$

Let us assume without loss of generality that $\mathcal{D}, \mathcal{D}'$ differ in the n-th observations. Hence, we have that

$$\begin{aligned}
\|\boldsymbol{y} - \boldsymbol{X}_S \hat{\boldsymbol{\beta}}\|_2^2 - \|\boldsymbol{y}' - \boldsymbol{X}'_S \hat{\boldsymbol{\beta}}\|_2^2 &= \sum_{i=1}^{n-1} [(y_i - (\boldsymbol{x}_i)_S^T \hat{\boldsymbol{\beta}})^2 - (y_i - (\boldsymbol{x}_i)_S^T \hat{\boldsymbol{\beta}})^2] + (y_n - (\boldsymbol{x}_n)_S^T \hat{\boldsymbol{\beta}})^2 - (y'_n - (\boldsymbol{x}_n)_S^T \hat{\boldsymbol{\beta}})^2 \\
&\leq (y_n - (\boldsymbol{x}_n)_S^T \hat{\boldsymbol{\beta}})^2 \\
&\leq 2y_n^2 + 2((\boldsymbol{x}_n)_S^T \hat{\boldsymbol{\beta}})^2 \\
&\leq 2b_y^2 + 2b_x^2 r^2 s
\end{aligned}$$

where the last step uses Cauchy-Schwartz inequality and the fact that $|S| = s$. □

A.2 PROOF OF THEOREM 1

First, we prove some technical results that will be used in the proof of Theorem 1. Define

$$\hat{\mathcal{R}}(S, \mathcal{D}) = \begin{cases} \mathcal{R}(S, \mathcal{D}) & \text{if } S \in \{\hat{S}_1(\mathcal{D}), \cdots, \hat{S}_R(\mathcal{D})\} \\ \mathcal{R}(\hat{S}_R(\mathcal{D}), \mathcal{D}) & \text{otherwise.} \end{cases} \tag{A.1}$$

Lemma A.1. *Let Δ to be taken as in* (2)*. Then,*

$$\max_{k \geq 1} \max_{\substack{\mathcal{D}, \mathcal{D}' \in \mathcal{Z}^n \\ \mathcal{D}, \mathcal{D}' \text{ are neighbors}}} \mathcal{R}(\hat{S}_k(\mathcal{D}), \mathcal{D}) - \mathcal{R}(\hat{S}_k(\mathcal{D}'), \mathcal{D}') \leq \Delta.$$

Proof. Fix $k \geq 1$ and let us consider the following cases:
Case 1: $\mathcal{R}(\hat{S}_k(\mathcal{D}), \mathcal{D}') \leq \mathcal{R}(\hat{S}_k(\mathcal{D}'), \mathcal{D}')$. Then, by the definition of Δ,

$$\begin{aligned}
\mathcal{R}(\hat{S}_k(\mathcal{D}), \mathcal{D}) - \mathcal{R}(\hat{S}_k(\mathcal{D}'), \mathcal{D}') &\leq \mathcal{R}(\hat{S}_k(\mathcal{D}), \mathcal{D}') - \mathcal{R}(\hat{S}_k(\mathcal{D}'), \mathcal{D}') + \Delta \\
&\leq \Delta.
\end{aligned} \tag{A.2}$$

Case 2: $\mathcal{R}(\hat{S}_k(\mathcal{D}), \mathcal{D}) \leq \mathcal{R}(\hat{S}_k(\mathcal{D}'), \mathcal{D})$. Then,

$$\begin{aligned}
\mathcal{R}(\hat{S}_k(\mathcal{D}), \mathcal{D}) - \mathcal{R}(\hat{S}_k(\mathcal{D}'), \mathcal{D}') &\leq \mathcal{R}(\hat{S}_k(\mathcal{D}), \mathcal{D}) - \mathcal{R}(\hat{S}_k(\mathcal{D}'), \mathcal{D}) + \Delta \\
&\leq \Delta.
\end{aligned} \tag{A.3}$$

Case 3: $\mathcal{R}(\hat{S}_k(\mathcal{D}), \mathcal{D}) > \mathcal{R}(\hat{S}_k(\mathcal{D}'), \mathcal{D})$ and $\mathcal{R}(\hat{S}_k(\mathcal{D}), \mathcal{D}') > \mathcal{R}(\hat{S}_k(\mathcal{D}'), \mathcal{D}')$. Trivially, in this case we must have $\binom{p}{s} > k \geq 2$. Then, there must exist $S_0 \subseteq [p]$, $|S_0| = s$ such that

$$\mathcal{R}(S_0, \mathcal{D}) \geq \mathcal{R}(\hat{S}_k(\mathcal{D}), \mathcal{D}), \mathcal{R}(S_0, \mathcal{D}') \leq \mathcal{R}(\hat{S}_k(\mathcal{D}'), \mathcal{D}').$$

To this end, define

$$\mathbb{S}_1 = \{\hat{S}_1(\mathcal{D}), \cdots, \hat{S}_{k-1}(\mathcal{D})\},\ \mathbb{S}_2 = \{\hat{S}_{k+1}(\mathcal{D}), \cdots\},\ \mathbb{S}' = \{\hat{S}_1(\mathcal{D}'), \cdots, \hat{S}_{k-1}(\mathcal{D}')\}.$$

As $\mathcal{R}(\hat{S}_k(\mathcal{D}), \mathcal{D}') > \mathcal{R}(\hat{S}_k(\mathcal{D}'), \mathcal{D}')$, we have that $\hat{S}_k(\mathcal{D}) \notin \mathbb{S}'$ so $\mathbb{S}' \subset \mathbb{S}_1 \cup \mathbb{S}_2$. On the other hand, as $\mathcal{R}(\hat{S}_k(\mathcal{D}), \mathcal{D}) > \mathcal{R}(\hat{S}_k(\mathcal{D}'), \mathcal{D})$, $\hat{S}_k(\mathcal{D}') \in \mathbb{S}_1$ and as $\hat{S}_k(\mathcal{D}') \notin \mathbb{S}'$, we have $|\mathbb{S}' \cap \mathbb{S}_1| \leq k - 2$. As $|\mathbb{S}'| = k - 1$, we must have $|\mathbb{S}' \cap \mathbb{S}_2| \geq 1$ which proves the existence of S_0. Next, note that

$$\mathcal{R}(\hat{S}_k(\mathcal{D}), \mathcal{D}) - \mathcal{R}(\hat{S}_k(\mathcal{D}'), \mathcal{D}') \leq \mathcal{R}(S_0, \mathcal{D}) - \mathcal{R}(S_0, \mathcal{D}') \leq \Delta. \tag{A.4}$$

□

Lemma A.2. *Let Δ to be taken as in* (2)*. Then,*

$$\max_{\substack{S\subseteq[p]\\|S|=s}} \max_{\substack{\mathcal{D},\mathcal{D}'\in\mathcal{Z}^n\\ \mathcal{D},\mathcal{D}' \text{ are neighbors}}} \hat{\mathcal{R}}(S,\mathcal{D}) - \hat{\mathcal{R}}(S,\mathcal{D}') \leq \Delta.$$

Proof. Suppose $S = \hat{S}_{k_1}(\mathcal{D}) = \hat{S}_{k_2}(\mathcal{D}')$. Let us consider the following cases:

Case 1: $k_1, k_2 \geq R$: Then, we have $\hat{\mathcal{R}}(S,\mathcal{D}) = \mathcal{R}(\hat{S}_R(\mathcal{D}),\mathcal{D})$ and $\hat{\mathcal{R}}(S,\mathcal{D}') = \mathcal{R}(\hat{S}_R(\mathcal{D}'),\mathcal{D}')$. Therefore,

$$\hat{\mathcal{R}}(S,\mathcal{D}) - \hat{\mathcal{R}}(S,\mathcal{D}') = \mathcal{R}(\hat{S}_R(\mathcal{D}),\mathcal{D}) - \mathcal{R}(\hat{S}_R(\mathcal{D}'),\mathcal{D}') \leq \Delta \tag{A.5}$$

by Lemma A.1.

Case 2: $k_1 < R, k_2 \geq R$: Then, we have $\hat{\mathcal{R}}(S,\mathcal{D}) = \mathcal{R}(S,\mathcal{D}) \leq \mathcal{R}(\hat{S}_R(\mathcal{D}),\mathcal{D})$ and $\hat{\mathcal{R}}(S,\mathcal{D}') = \mathcal{R}(\hat{S}_R(\mathcal{D}'),\mathcal{D}')$. Then,

$$\hat{\mathcal{R}}(S,\mathcal{D}) - \hat{\mathcal{R}}(S,\mathcal{D}') \leq \mathcal{R}(\hat{S}_R(\mathcal{D}),\mathcal{D}) - \mathcal{R}(\hat{S}_R(\mathcal{D}'),\mathcal{D}') \leq \Delta. \tag{A.6}$$

Case 3: $k_1 \geq R, k_2 < R$: Then, we have $\hat{\mathcal{R}}(S,\mathcal{D}) = \mathcal{R}(\hat{S}_R(\mathcal{D}),\mathcal{D}) \leq \mathcal{R}(S,\mathcal{D})$ and $\hat{\mathcal{R}}(S,\mathcal{D}') = \mathcal{R}(S,\mathcal{D}')$. Then,

$$\hat{\mathcal{R}}(S,\mathcal{D}) - \hat{\mathcal{R}}(S,\mathcal{D}') \leq \mathcal{R}(S,\mathcal{D}) - \mathcal{R}(S,\mathcal{D}') \leq \Delta. \tag{A.7}$$

Case 4: $k_1, k_2 < R$: Then, we have $\hat{\mathcal{R}}(S,\mathcal{D}) = \mathcal{R}(S,\mathcal{D})$ and $\hat{\mathcal{R}}(S,\mathcal{D}') = \mathcal{R}(S,\mathcal{D}')$. The result follows. □

Lemma A.3. *Suppose $\mathcal{M}$ is as defined in Algorithm 1, and $\hat{\mathcal{A}}_E$ is an exponential mechanism with the objective $\hat{\mathcal{R}}$,*

$$\mathbb{P}(\hat{\mathcal{A}}_E(\mathcal{D}) = S) \propto \exp\left(-\frac{\varepsilon\hat{\mathcal{R}}(S,\mathcal{D})}{2\Delta}\right), \ \forall S \in \mathcal{O}. \tag{A.8}$$

Then, for $S \in \mathcal{O}$,

$$(1-q^T)\mathbb{P}(\hat{\mathcal{A}}_E(\mathcal{D}) = S) \leq \mathbb{P}(\mathcal{M}(\mathcal{D}) = S) \leq \mathbb{P}(\hat{\mathcal{A}}_E(\mathcal{D}) = S) + q^T \tag{A.9}$$

where

$$q = \frac{R}{\binom{p}{s}}.$$

Proof. Fix $S \in \mathcal{O}$ and suppose $S = \hat{S}_k(\mathcal{D})$. Moreover, let $\mathbb{S}_R = \{\hat{S}_k(\mathcal{D}) : k \leq R\}$. Consider the following cases:
Case 1: $k \leq R$**:** Then, based on Algorithm 1,

$$\mathbb{P}(\mathcal{M}(\mathcal{D}) = S) = \mathbb{P}\left(\{a(\mathcal{D}) = k\} \cup \{a(\mathcal{D}) = R+1, \mathcal{M}_0(\mathcal{D}) = S\}\right).$$

Therefore,

$$\begin{aligned}\mathbb{P}(a(\mathcal{D}) = k) \leq \mathbb{P}(\mathcal{M}(\mathcal{D}) = S) &= \mathbb{P}\left(\{a(\mathcal{D}) = k\} \cup \{a(\mathcal{D}) = R+1, \mathcal{M}_0(\mathcal{D}) = S\}\right)\\ &\overset{(a)}{\leq} \mathbb{P}(a(\mathcal{D}) = k) + \mathbb{P}(\mathcal{M}_0(\mathcal{D}) \in \mathbb{S}_R)\\ &\overset{(b)}{\leq} \mathbb{P}(a(\mathcal{D}) = k) + q^T\end{aligned} \tag{A.10}$$

where (a) is true as $S \in \mathbb{S}_R$, and (b) is true as $\mathcal{M}_0$ return a support in $\mathbb{S}_R$ if it selects some support from $\mathbb{S}_R$ for all T iterations, showing $\mathbb{P}(\mathcal{M}_0(\mathcal{D}) \in \mathbb{S}_R) = q^T$. Note that for $k \leq R$, $\mathbb{P}(a(\mathcal{D}) = k) = \mathbb{P}(\hat{\mathcal{A}}_E(\mathcal{D}) = S)$, therefore,

$$\mathbb{P}(\hat{\mathcal{A}}_E(\mathcal{D}) = S) \leq \mathbb{P}(\mathcal{M}(\mathcal{D}) = S) \leq \mathbb{P}(\hat{\mathcal{A}}_E(\mathcal{D}) = S) + q^T. \tag{A.11}$$

Case 2: $k > R$**:** Then, from (6),

$$\begin{aligned}\mathbb{P}(a(\mathcal{D}) = R+1) &= \frac{\left(\binom{p}{s} - R\right)\exp\left(-\varepsilon\mathcal{R}(\hat{S}_R(\mathcal{D}),\mathcal{D})/(2\Delta)\right)}{\sum_{k=1}^{R}\exp\left(-\varepsilon\mathcal{R}(\hat{S}_k(\mathcal{D}),\mathcal{D})/(2\Delta)\right) + \left(\binom{p}{s} - R\right)\exp\left(-\varepsilon\mathcal{R}(\hat{S}_R(\mathcal{D}),\mathcal{D})/(2\Delta)\right)} \\ &= \frac{\left(\binom{p}{s} - R\right)\exp\left(-\varepsilon\hat{\mathcal{R}}(S,\mathcal{D})/(2\Delta)\right)}{\sum_{k=1}^{R}\exp\left(-\varepsilon\hat{\mathcal{R}}(\hat{S}_k(\mathcal{D}),\mathcal{D})/(2\Delta)\right) + \sum_{k\geq R+1}\exp\left(-\varepsilon\hat{\mathcal{R}}(\hat{S}_k(\mathcal{D}),\mathcal{D})/(2\Delta)\right)} \\ &= \left(\binom{p}{s} - R\right)\mathbb{P}(\hat{\mathcal{A}}_E(\mathcal{D}) = S). \end{aligned} \tag{A.12}$$

Hence, one can write

$$\begin{aligned}\mathbb{P}(\mathcal{M}(\mathcal{D}) = S) &= \mathbb{P}\left(\{a(\mathcal{D}) = R+1\} \cap \{\mathcal{M}_0(\mathcal{D}) = S\}\right) \\ &\overset{(a)}{=} \mathbb{P}(a(\mathcal{D}) = R+1)\left(\sum_{i=1}^{T}\frac{q^{i-1}}{\binom{p}{s}}\right) \\ &= \frac{1}{\binom{p}{s}}\left[\binom{p}{s} - R\right]\mathbb{P}(\hat{\mathcal{A}}_E(\mathcal{D}) = S)\frac{1-q^T}{1-q} \\ &= (1-q^T)\mathbb{P}(\hat{\mathcal{A}}_E(\mathcal{D}) = S) \end{aligned} \tag{A.13}$$

where (a) is true as

$$\mathbb{P}(\mathcal{M}_0(\mathcal{D}) = S) = \sum_{i=1}^{T}\mathbb{P}(\mathcal{M}_0(\mathcal{D}) = S, \mathcal{M}_0 \text{ stops after } i \text{ iterations }) = \sum_{i=1}^{T}\frac{q^{i-1}}{\binom{p}{s}}.$$

The proof is complete by (A.11) and (A.13). □

Next, let us prove an important intermediate result on Algorithm 1.

Theorem A.1. *Suppose $T > 1$, $1 < R < \binom{p}{s}$. The procedure $\mathcal{M}$ in Algorithm 1 is (ε', δ)-DP where*

$$\varepsilon' = \varepsilon - \log\left(1 - \left[R/\binom{p}{s}\right]^T\right), \quad \delta = \left[R/\binom{p}{s}\right]^T.$$

Proof. From Lemma A.2 and Lemma 1, we know that $\hat{\mathcal{A}}_E$ is an $(\varepsilon, 0)$-DP procedure. Suppose $\mathcal{D}, \mathcal{D}'$ are neighboring datasets. Then, from Lemma A.3,

$$\begin{aligned}\mathbb{P}(\mathcal{M}(\mathcal{D}) = S) &\leq \mathbb{P}(\hat{\mathcal{A}}_E(\mathcal{D}) = S) + q^T \\ &\leq e^{\varepsilon}\mathbb{P}(\hat{\mathcal{A}}_E(\mathcal{D}') = S) + q^T \\ &\leq \frac{1}{1-q^T}e^{\varepsilon}\mathbb{P}(\mathcal{M}(\mathcal{D}') = S) + q^T \end{aligned} \tag{A.14}$$

where the first and last inequality use Lemma A.3. □

Proof of Theorem 1. Note that by definition, for $S \in \mathcal{O}$, we have $0 \leq \mathcal{R}(S,\mathcal{D}) \leq \|\boldsymbol{y}\|_2^2 \leq nb_y^2$. Then,

$$\begin{aligned}\mathbb{P}(\hat{\mathcal{A}}_E(\mathcal{D}) = S) &= \frac{\exp\left(-\varepsilon\hat{\mathcal{R}}(S,\mathcal{D})/(2\Delta)\right)}{\sum_{k=1}^{R}\exp\left(-\varepsilon\hat{\mathcal{R}}(\hat{S}_k(\mathcal{D}),\mathcal{D})/(2\Delta)\right) + \sum_{k\geq R+1}\exp\left(-\varepsilon\hat{\mathcal{R}}(\hat{S}_k(\mathcal{D}),\mathcal{D})/(2\Delta)\right)} \\ &\geq \frac{\exp(-n\varepsilon b_y^2/(2\Delta))}{\binom{p}{s}} := \delta_0. \end{aligned} \tag{A.15}$$

Then, from (A.14),

$$\begin{aligned}
\mathbb{P}(\mathcal{M}(\mathcal{D}) = S) &\leq e^{\varepsilon}\mathbb{P}(\hat{\mathcal{A}}_E(\mathcal{D}') = S) + q^T \\
&= \left(e^{\varepsilon} + \frac{q^T}{\delta_0}\right)\mathbb{P}(\hat{\mathcal{A}}_E(\mathcal{D}') = S) - \frac{q^T}{\delta_0}\mathbb{P}(\hat{\mathcal{A}}_E(\mathcal{D}') = S) + q^T \\
&\overset{(a)}{\leq} \left(e^{\varepsilon} + \frac{q^T}{\delta_0}\right)\mathbb{P}(\hat{\mathcal{A}}_E(\mathcal{D}') = S) \\
&\overset{(b)}{\leq} \frac{1}{1-q^T}\left(e^{\varepsilon} + \frac{q^T}{\delta_0}\right)\mathbb{P}(\mathcal{M}(\mathcal{D}') = S) \qquad \text{(A.16)}
\end{aligned}$$

where (a) is by (A.15) and (b) is by Lemma A.3. □