Probabilistic Best Subset Selection via Gradient-Based Optimization

Journal:	*Journal of the American Statistical Association*
Manuscript ID	JASA-T&M-2020-0620
Manuscript Type:	Article – Theory & Methods
Keywords:	L_0-regularized regression, Stochastic gradient descent, Variance reduction, Discrete optimization, Strong sparsity

Probabilistic Best Subset Selection via Gradient-Based Optimization

Abstract

In high-dimensional statistics, variable selection is an optimization problem aiming to recover the latent sparse pattern from all possible covariate combinations. In this paper, we propose a novel optimization method to solve the exact L_0-regularized regression problem (a.k.a. best subset selection). We reformulate the optimization problem from a discrete space to a continuous one via probabilistic reparameterization. Within the framework of stochastic gradient descent, we propose a family of unbiased gradient estimators to optimize the L_0-regularized objective and its variational lower bound. Within this family, we identify the estimator with a non-vanishing signal-to-noise ratio and uniformly minimum variance. Theoretically, we study the general conditions under which the method is guaranteed to converge to the ground truth in expectation. In a wide variety of synthetic and semi-synthetic data sets, the proposed method outperforms existing variable selection methods that are based on penalized regression and mixed integer optimization, in both sparse pattern recovery and out-of-sample prediction. Our method can find the true regression model from thousands of covariates in a couple of seconds. [1]

Keywords: L_0-regularized regression, stochastic gradient descent, variance reduction, discrete optimization, strong sparsity

[1] Code is available in the supplementary materials.

1 Introduction

Variable selection by penalized models is widely applied to uncover sparse structures in high dimensional data. Solving L_0-regularized regression, also known as the best-subset selection problem (Friedman et al., 2001; Fan and Lv, 2010), is a natural approach, as it directly regularizes the number of variables included in the regression model. In this paper, we study L_0-regularized regression to optimize the following objective function

$$\min_{\boldsymbol{\beta}\in\mathbb{R}^p} \left\{ \frac{1}{n} \|\boldsymbol{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2 + \lambda \|\boldsymbol{\beta}\|_0 \right\}, \tag{1}$$

where $\boldsymbol{y} = (y_1, \ldots, y_n)^\top \in \mathbb{R}^n$ represents the vector of response variables, $\mathbf{X} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)^\top \in \mathbb{R}^{n\times p}$ is the design matrix, $\boldsymbol{\beta} \in \mathbb{R}^p$ are the regression coefficients, and $\lambda > 0$ is the penalty hyperparameter. We consider the high-dimensional regime, where the number of covariates p exceeds the sample size n, and can potentially grow with n. The L_0 penalty is defined as $\|\boldsymbol{\beta}\|_0 := \sum_{j=1}^p \mathbf{1}_{[\beta_j \neq 0]}$, which counts the number of nonzero elements in $\boldsymbol{\beta}$. Here, $\mathbf{1}_{[\cdot]}$ is an indicator function that equals to one if the condition is true and zero otherwise. The L_0-regularized subset selection is closely related to the standard information theoretic methods for model selection. In particular, when the data is Gaussian distributed, the objective function in Eq. (1) is equivalent to the Akaike information criterion (AIC) (Akaike, 1974, 1998) when $\lambda = 1/n$, and the Bayesian information criterion (BIC) (Schwarz, 1978) when $\lambda = \log(n)/(2n)$. Both AIC and BIC balance the model fit with parsimony, pertaining to the out-of-sample predictive accuracy. In practice, such information criteria are often used in comparison of alternative models, but rarely as optimization objectives for variable selection (Chickering, 2002). This limits the number of candidate models and the selected model is often sub-optimal. The main challenge of directly optimizing (1) is the discontinuity of the L_0 penalty, making the problem NP-hard (Natarajan, 1995).

To improve computational efficiency, the best subset selection problems often resort to some approximate solutions. For example, instead of searching over all possible subsets, greedy schemes sequentially select or prune covariates until a prespecified number of covariates is attained. Forward stepwise selection, starting from a null set, adds one covariate at each step that improves the fitting the most. Conversely, the backward stepwise selection starts from all covariates and deletes one at a time that impacts the fitting the least (Beale et al., 1967; Mallat and Zhang, 1993). Orthogonal matching pursuit (Chen et al., 1989; Pati et al., 1993) is a greedy forward-search algorithm that selects covariates that correlate the most with the prediction residuals (Joseph, 2013; Donoho et al., 2012). The greedy algorithms yield a sequence of subsets with increasing (or decreasing) size, but in general none of these selected sets are the global optimum.

For decades, a major paradigm to approximate the L_0-regularized regression is to use continuous approximations. The Bridge regression (Frank and Friedman, 1993; Fu, 1998) uses the L_q penalties ($q > 0$), defined as $\sum_{j=1}^{p} \beta_j^q$ with estimated coefficient β_j. When $q \geq 1$, the L_q penalty is convex, while when $q \leq 1$, the regularization encourages sparse estimation and hence performs variable selection (Fan and Lv, 2010). In the intersection of these two domains lies the widely used least absolute shrinkage and selection operator (Lasso). It encourages sparsity, and in the meanwhile enjoys computational advantage of convex optimization. Asymptotically, Lasso is accurate for both variable selection and coefficient estimation (Zhao and Yu, 2006; Candès and Plan, 2009; Wainwright, 2009). However, in the finite sample setting, it suffers from downward bias due to the shrinkage effect of the L_1 norm. In a high dimensional setting ($p \gg n$), when the penalty hyperparameter is chosen by cross-validation, Lasso often chooses extra spurious variables which can result in a high false discovery rate (FDR) (Barber and Candès, 2015). To mitigate these problems, a variety of non-convex penalties have been proposed. Smoothly clipped absolute deviation (SCAD) (Fan and Li, 2001) and minimax concave penalty (MCP) (Zhang, 2010),

for example, approximate the hard-thresholding property of L_0 penalty by the piece-wise non-convex penalties. The coefficient estimators are consistent and unbiased when the true coefficients are sufficiently large. Another line of work constructs functions to directly resemble the L_0 pseudo-norm, which are called the pseudo-L_0 penalties. For example, Liu and Wu (2007) and Shen et al. (2012) approximate the L_0 penalty with a convex function $\min\{|\beta_j|/\tau, 1\}$. Dicker et al. (2013) approximates the L_0 penalty with a non-convex function $\log(|\beta_j|/(|\beta_j| + \tau) + 1)$, where the objective is optimized by coordinate descent. Though achieving improved sparsity recovery, the pseudo-L_0 penalties introduce an extra tuning parameter τ to control the approximation accuracy, which induces sensitivity to its value and extra bias.

Comparing to continuous approximations, the solution of exact best subset selection enjoys superior statistical properties, such as the unbiasedness of regression coefficients, also known as the oracle property (Greenshtein, 2006; Zhang and Zhang, 2012; Belloni and Chernozhukov, 2013), and the low in-sample risk (Foster and George, 1994). For the orthogonal design matrix, Johnson et al. (2015) shows that the predictive risk of the L_1-regularized linear regression cannot outperform L_0-regularized regression by more than a constant factor, and in some cases is infinitely worse. Due to the benefits of L_0 penalty and rapid improvements in computational tools, recently there is renewed interest in solving the exact best subset selection problem. With modern optimization tools, Bertsimas et al. (2016) study the constrained best subset selection problem

$$\min_{\boldsymbol{\beta}\in\mathbb{R}^p} \left\{ \frac{1}{n} \|\boldsymbol{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2 \right\} \quad \text{subject to} \quad \|\boldsymbol{\beta}\|_0 \leq \hat{S} \tag{2}$$

with a two-stage algorithm. In the first stage, an iterative hard thresholding method computes a warm-up initialization, and in the second stage, an industrial standard mixed integer optimization (MIO) tool is applied as the solver. The proposed method scales the best subset selection from the setting where the number of covariates $p \leq 30$, as studied

by the leaps and bounds method (Furnival and Wilson, 1974), to the setting where p is in 1000s. However, the core MIO steps rely upon a non-convex optimization tool, which is hard to generalize beyond linear regression problems. Although the speed has been improved by following works (Hazimeh and Mazumder, 2018), MIO nevertheless induces a substantial computation load.

The L_0-regularized best subset selection in Eq. (1) can be considered as the Lagrangian form of the L_0-constrained best subset selection in Eq. (2). However, it is worth noting that due to the discontinuity, the two problems are not equivalent in that there may not exist a surjection between penalty parameters λ and $\hat{S}$ (Polson and Sun, 2019). Since there is a rich literature on variable selection, we refer the reader to several representative publications (Friedman et al., 2001; Fan and Lv, 2010; Bertsimas et al., 2016; Hastie et al., 2017) and the references therein for comprehensive reviews.

Our contributions: In this paper, we first propose a probabilistic objective for the exact L_0-regularized regression, where we cast the discrete optimization problem to an equivalent one in the continuous space. Second, capitalizing on the new objective we design a modern stochastic gradient descent (SGD) based method as an end-to-end solver, with a general framework to construct unbiased gradient estimators. We identify the estimator with minimal variance and non-vanishing signal-to-noise ratio (SNR) in the univariate case and generalize it to the multivariate case. Third, we provide theoretical insights on the conditions that guarantee the convergence of the gradient descent updates to the ground truth in expectation. Finally, we empirically show our gradient-based method can solve the L_0-regularized regression with accurate active set recovery, coefficient estimation, and testing prediction. In terms of efficiency, it can solve the problems with n in 100s, p in 1000s in seconds, significantly faster than previously proposed best subset selection methods.

Organization: The rest of this paper is organized as follows. In Section 2, we provide a probabilistic reformulation of the L_0-regularized linear regression problem. In Section 3, we propose a family of unbiased gradient estimators to solve the reformulated optimization problem. In Section 4, we analyze the conditions that guarantee the convergence to the ground truth in expectation. In Section 5, we generalize the proposed gradient methods to optimize a novel variational lower bound for the best subset selection. Experiments and results are described in Section 6, and discussions follow in Section 7.

Notation: Throughout this work, we use n as the sample size, p as the number of covariates, and S as the number of non-zero true coefficients. We use $\boldsymbol{x}_i$ to denote the i^{th} row of matrix $\mathbf{X}$ and X_j as its j^{th} column. We use $\mathbf{X}_{\mathcal{A}}$ to denote a submatrix of $\mathbf{X}$ as $\mathbf{X}_{\mathcal{A}} = \{X_j\}_{j\in\mathcal{A}}$, $\mathcal{A} \subseteq [p]$, where $[p] := \{1, \ldots, p\}$.

2 Reformulation of L_0-Penalized Regression

The underlying assumption of best subset selection is that the response variables only depend on a subset of covariates $\mathbf{X}_{\mathcal{A}}$, where $\mathcal{A} \subseteq \{1, \ldots, p\}$ is called the active set. The size of the true active set S is assumed to be much smaller than p. We decompose the regression coefficients as $\boldsymbol{\beta} = \boldsymbol{\alpha} \odot \boldsymbol{z}$, using a spike-and-slab construction (George and McCulloch, 1993; Zhou et al., 2009), where $\odot$ denotes an element-wise product. The binary vector $\boldsymbol{z} \in \{0, 1\}^p$ represents the inclusion of covariates in the active set, and $\boldsymbol{\alpha} \in \mathbb{R}^p$ encodes the scale. With these augmented latent variables, the optimization problem (1) can be equivalently expressed as

$$\min_{\boldsymbol{\alpha},\boldsymbol{z}} \ \frac{1}{n} \left\| \boldsymbol{y} - \mathbf{X}(\boldsymbol{\alpha} \odot \boldsymbol{z}) \right\|^2 + \lambda \left\| \boldsymbol{z} \right\|_0 \tag{3}$$

Similar to optimizing $\boldsymbol{\beta}$ itself, optimizing $\boldsymbol{z}$ is a combinatorial problem and remains NP-hard.

A major bottleneck is that the discrete nature of variable $\boldsymbol{z}$ precludes the derivative with respect to it, which could have provided the direction of steepest descent. This motivates us to reformulate the discrete optimization problem (3) to an optimization problem in the continuous space. Such continuity greenlights the gradient-based methods that act as the horsepower in modern machine learning. Instead of directly optimizing $\boldsymbol{z}$, we consider $\boldsymbol{z}$ as a random variable with distribution $p(\boldsymbol{z};\boldsymbol{\pi}) = \prod_{j=1}^{p} \text{Bern}(z_j;\pi_j)$, $\pi_j \in [0,1]$, where $\text{Bern}(z_j;\pi_j)$ stands for Bernoulli distribution with parameter π_j. Then the problem (3) can be transformed to a form of expectation, which allows us to construct a stochastic gradient with Monte Carlo estimation. We first have the following theorem:

Theorem 1 (Probabilistic Reformulation)**.** *The L_0-regularized best subset selection problem* (1) *is equivalent to the following problem*

$$\min_{\boldsymbol{\pi}} \ \mathbb{E}_{\boldsymbol{z}\sim p(\boldsymbol{z};\boldsymbol{\pi})}\left[\min_{\boldsymbol{\alpha}} \frac{1}{n} \ \|\boldsymbol{y} - \mathbf{X}(\boldsymbol{\alpha}\odot\boldsymbol{z})\|_2^2 + \lambda\,\|\boldsymbol{z}\|_0\right], \tag{4}$$

where $\boldsymbol{\pi} = (\pi_1, \pi_2, \ldots, \pi_p) \in [0,1]^p$, $p(z_j = 1) = \pi_j$, $j \in \{1,\ldots,p\}$.

The equivalence can be proved by the fact that the optimal solution to Eq. (3) is a feasible solution of Eq. (4) that achieves the same objective value, and vice versa. The proof is conceptually straightforward and we defer it to Appendix B.1. The objective (4) is a bi-level optimization problem, where the inner optimization for given $\boldsymbol{z}$ is an ordinary least square (OLS) problem on the design matrix $\mathbf{X}_{\mathcal{Z}}$, which has a closed-form solution. Here, we denote the active set inferred by $\boldsymbol{z}$ as $\mathcal{Z} := \{j\}_{j:z_j\neq 0}$. The computational speed of the inner optimization increases when the size of $\mathcal{Z}$ decreases, i.e. the sparser the faster. For computational convenience, we reparameterize $\boldsymbol{\pi} = (\pi_1,\ldots,\pi_p)$ with the sigmoid function as $\pi_j = \sigma(\phi_j) = 1/(1+\exp(-\phi_j))$, $j \in [p]$ and relax the optimization space to an unconstrained continuous space. Although under the sigmoid reparameterization, probability π_j can reach 0 or 1 only when logits ϕ_j goes to the infinity, it can be accurately approximated in practice when the absolute values of logits ϕ_j are sufficiently large.

A naive approach that is guaranteed to select the best subset is to exhaust all possible subsets. However, the exhaustion method requires evaluating the objective function 2^p times with p covariates, which is often computationally infeasible. Even if the cardinality of the true active set is known, for instance, to find the best subset with cardinality as 10 from $p = 1000$ covariates, brute-force search needs to evaluate the objective function in an order of 10^{23} times. To improve the efficiency, our key idea is to guide the function evaluation by the first-order information, which is estimated by the stochastic gradient. In the following sections, we first build a family of unbiased gradient estimators in the univariate case, then the estimators are generalized to high dimension.

3 A Family of Unbiased Gradient Estimators

In this section we consider gradient methods with a general probabilistic objective

$$\min_{\boldsymbol{\phi}} \ \mathcal{E}(\boldsymbol{\phi}) = \mathbb{E}_{\boldsymbol{z}\sim p_{\boldsymbol{\phi}}(\boldsymbol{z})}[f(\boldsymbol{z})], \tag{5}$$

where $p_{\boldsymbol{\phi}}(\boldsymbol{z}) := \prod_{j=1}^{p} \text{Bern}(\sigma(\phi_j))$. The L_0-regularized objective (4) can be considered as a special case when setting $f(\boldsymbol{z}) = \min_{\boldsymbol{\alpha}} \|\boldsymbol{y} - \mathbf{X}(\boldsymbol{\alpha} \odot \boldsymbol{z})\|_2^2 / n + \lambda \|\boldsymbol{z}\|_0$. Taking the gradient of expectation $\mathcal{E}$ in (5) with respect to $\boldsymbol{\phi}$, the derivative and integral can be interchanged as

$$\nabla_{\boldsymbol{\phi}} \mathcal{E}(\boldsymbol{\phi}) = \nabla_{\boldsymbol{\phi}} \mathbb{E}_{\boldsymbol{z}\sim p_{\boldsymbol{\phi}}(\boldsymbol{z})}[f(\boldsymbol{z})] = \mathbb{E}_{\boldsymbol{z}\sim p_{\boldsymbol{\phi}}(\boldsymbol{z})}[f(\boldsymbol{z}) \nabla_{\boldsymbol{\phi}} \log p_{\boldsymbol{\phi}}(\boldsymbol{z})], \tag{6}$$

where the last equality is called the *score method* in statistics and *REINFORCE* in reinforcement learning (Williams, 1992). Though the expectation (6) cannot be computed analytically, it can be estimated by an unbiased Monte Carlo estimation $\frac{1}{K} \sum_{k=1}^{K} f(\boldsymbol{z}_k) \nabla_{\boldsymbol{\phi}} \log p_{\boldsymbol{\phi}}(\boldsymbol{z}_k)$, with $\boldsymbol{z}_1, \ldots, \boldsymbol{z}_K \overset{iid}{\sim} p_{\boldsymbol{\phi}}(\boldsymbol{z})$. One advantage of computing REINFORCE estimator is that the number of function evaluations does not grow with the dimension of variables, which makes it generalizable to high dimensional $\boldsymbol{z}$. Another advantage is that this estimator only needs

the value of $f(\boldsymbol{z})$, which makes it applicable to the situations when $f(\boldsymbol{z})$ is discontinuous or even has no explicit expression (*e.g.*, in reinforcement learning, $f(\boldsymbol{z})$ could be the unknown reward function for action $\boldsymbol{z}$). However, the score function gradient is known for high variance. Though the Monte Carlo estimation with K samples reduces the variance in the order $\mathcal{O}(1/K)$, it nevertheless needs a large number of function evaluations at each gradient step to get an accurate gradient estimation, which can be computationally demanding.

Aiming at variance reduction, we propose a general framework to construct unbiased gradient estimators for the objective (5). Maintaining the unbiased property, our goal is to find a gradient estimator with minimal variance within the proposed estimator family. To cope with the best subset selection problem, we require that the estimators do not rely on the function continuity, and the number of required function evaluations does not increase with the number of covariates.

3.1 Insight from univariate gradient setting

When involving only a single dimension, the gradient can be computed analytically as

$$\nabla_\phi \mathbb{E}_{z\sim \mathrm{Bern}(\sigma(\phi))}[f(z)] = \sigma(\phi)(1-\sigma(\phi))[f(1)-f(0)]. \tag{7}$$

Based on the closed form of expectation, we require the following properties for gradient estimation in the univariate case.

Definition 1. *For an objective $\mathbb{E}_{z\sim \mathrm{Bern}(\sigma(\phi))}[f(z)]$, assume an estimator of the gradient with respect to ϕ is $g(u;\sigma(\phi))$ where $u\sim \mathrm{Unif}(0,1)$ is a uniform random variable. For function $f:\{0,1\}\to\mathbb{R}$, we assume that the estimator satisfies the following properties:*

- *Unbiasedness:* $\mathbb{E}_{u\sim\mathrm{Unif}(0,1)}[g(u;\sigma(\phi))] = \sigma(\phi)(1-\sigma(\phi))[f(1)-f(0)]$,
- *Functional form:*

$$g(u;\sigma(\phi)) = a(u;\sigma(\phi))f(\mathbf{1}_{[u<\sigma(\phi)]}) + b(u;\sigma(\phi))f(\mathbf{1}_{[u>1-\sigma(\phi)]}) \tag{8}$$

where $a(u;\sigma(\phi))$, $b(u;\sigma(\phi))$ *are independent of function* $f(\cdot)$.

The requirement of the specific functional form ensures efficiency when generalizing to the high dimensional settings, as discussed in Section 3.2. The estimator family in Definition 1 incorporates many popular unbiased gradient estimators. Here we list several representative ones. The REINFORCE estimator can be considered as a special case, with

$$g_R(u;\sigma(\phi)) = f(\mathbf{1}_{[u<\sigma(\phi)]})(\mathbf{1}_{[u<\sigma(\phi)]} - \sigma(\phi)). \tag{9}$$

Recently, a newly proposed augment-REINFORCE-merge (ARM) gradient (Yin and Zhou, 2019) achieves success in deep learning problems with binary latent variables. Its derivation is related to data augmentation and antithetic sampling. In the univariate case, the ARM gradient shares a similar form as

$$g_{\mathrm{ARM}}(u;\sigma(\phi)) = [f(\mathbf{1}_{[u>1-\sigma(\phi)]}) - f(\mathbf{1}_{[u<\sigma(\phi)]})](u-\tfrac{1}{2}). \tag{10}$$

We add an indicator mask to the ARM gradient without changing its univariate distribution, which we call it ARM_0 estimator, with the expression as

$$g_{\mathrm{ARM}_0}(u;\sigma(\phi)) = [f(\mathbf{1}_{[u>\sigma(-\phi)]}) - f(\mathbf{1}_{[u<\sigma(\phi)]})](u-\tfrac{1}{2})\left|\mathbf{1}_{[u>\sigma(-\phi)]} - \mathbf{1}_{[u<\sigma(\phi)]}\right|. \tag{11}$$

In the univariate case, the ARM_0 estimator is identical to the ARM estimator, but when it comes to the multivariate case, ARM_0 can produce sparse gradients where many dimensions may become exact zeros. This straightforward sparsification, though not used in the original ARM algorithm of Yin and Zhou (2019), has been adopted as the default setting by several recent works when utilizing ARM (Boluki et al., 2020; Yue et al., 2020; Dadaneh et al., 2020).

In this paper, we propose a new gradient estimator that can further reduce the gradient variance, which is given by

$$g_{\mathrm{U2G}}(u;\sigma(\phi)) = \frac{\sigma(|\phi|)}{2}[f(\mathbf{1}_{[u>1-\sigma(\phi)]}) - f(\mathbf{1}_{[u<\sigma(\phi)]})](\mathbf{1}_{[u>\sigma(-\phi)]} - \mathbf{1}_{[u<\sigma(\phi)]}). \tag{12}$$

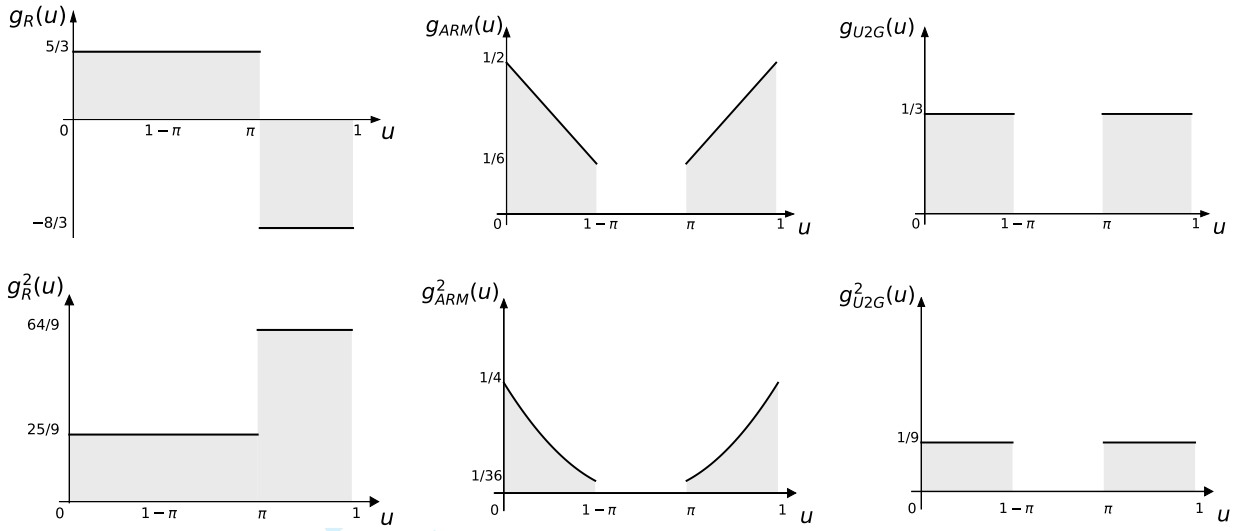
	REINFORCE	ARM	ARM_0	U2G
$a(u;\sigma(\phi))$	$\mathbf{1}_{[u<\sigma(\phi)]}-\sigma(\phi)$	$\frac{1}{2}-u$	$(\frac{1}{2}-u)\lvert\mathbf{1}_{[u>\sigma(-\phi)]}-\mathbf{1}_{[u<\sigma(\phi)]}\rvert$	$\sigma(\lvert\phi\rvert)(\mathbf{1}_{[u<\sigma(\phi)]}-\mathbf{1}_{[u>\sigma(-\phi)]})/2$
$b(u;\sigma(\phi))$	0	$u-\frac{1}{2}$	$(u-\frac{1}{2})\lvert\mathbf{1}_{[u>\sigma(-\phi)]}-\mathbf{1}_{[u<\sigma(\phi)]}\rvert$	$\sigma(\lvert\phi\rvert)(\mathbf{1}_{[u>\sigma(-\phi)]}-\mathbf{1}_{[u<\sigma(\phi)]})/2$

Table 1: Parameterization of unbiased gradient estimators.

Since it takes constant value at the non-zero region, we call it unbiased uniform gradient (U2G) estimator. The estimator can be derived by finding the minimum-variance unbiased estimator (MVUE) from the family in Definition 1, subject to an additional assumption that the gradient has a non-vanishing SNR for arbitrary $f(\cdot)$, defined as $\text{SNR} \coloneqq \mathbb{E}[g(u)]/\sqrt{\text{var}[g(u)]}$. We discuss this property in detail in Section 3.1.1. The U2G estimator is concurrently discovered by Dong et al. (2020) via Rao-Blackwellization over the ARM estimator, where it is called the DisARM estimator and exhibits promising performance in optimizing a generative model with deep neural network.

We summarize the above-mentioned estimators in Table 1 in a form compatible with Definition 1. The unbiased estimators share the same expectation, while the gradient variance differs depending on how the stochastic gradient is expressed as a function of $u \sim \text{Uniform}(0,1)$. As an illustrative example, we plot the functions $g(u)$ and $g^2(u)$ for the aforementioned estimators in Figure 1 with $f(1)=5$, $f(0)=4$, and $\pi=\sigma(\phi)=2/3$. Since the estimators are unbiased, the net signed areas under the curve of the first row in Figure 1 are the same. The variance of each gradient estimator, up to the same additive constant, is represented by the area under the curve in each subplot of the second row.

The intrinsic nature of variable selection is the model comparison, which takes the *relative* difference as a measure of goodness. Comparing the first column of Figure 1 to the other two columns, and comparing Eq. (9) to Eqs. (10) and (12), intuitively we find, if the function $f(\cdot)$ appears in the estimation as a relative difference $f(z)-f(z')$, the scale of



(d) REINFORCE (e) ARM (f) U2G

Figure 1: The characteristic curves of gradient estimators. In this illustrative example, $f(1) = 5$, $f(0) = 4$, and $\pi = \sigma(\phi) = 2/3$. The top row is the function $g(u)$ with respect to u; the second row is the function $g^2(u)$. The unbiased estimators have the same integration in the first row, but different gradient variance as shown in the second row (up to a constant).

the gradient does not increase with the scale of $f(\cdot)$; thus it controls the magnitude of the variance. The variance can be further reduced if the direction and magnitude of gradient estimator do not change with u. Formally, we have the following proposition, whose proof is deferred to Appendix B.2.

Proposition 1. *For positive or negative function $f(z)$, we have*

$$var[g_{\text{U2G}}] \leq var[g_{\text{ARM}}] \leq var[g_R],$$

where the second inequality requires $|f(1) - f(0)| \leq \min\{|f(1)|, |f(0)|\}$.

3.1.1 Optimality of U2G estimator

In this section, we discuss the condition under which U2G is the optimal within the estimator family in Definition 1. To simplify the notation, let $f_1 = f(1)$, $f_0 = f(0)$, $\pi = \sigma(\phi)$, and $\Delta = |f_0 - f_1|$. Without loss of generality, we first assume $\pi > 1/2$. An ideal gradient for variable selection should be capable of distinguishing potential models, even when their difference is small. However, as shown in Eq. (7), the scale of the true gradient diminishes when $\Delta \to 0$. Therefore, to ensure the stochastic gradient estimator has non-diminishing SNR at the non-stationary points, the variance has to decease to zero as the difference between potential models shrinks, that is

$$\lim_{\Delta \to 0} \text{var}[g(u;\pi)] = 0, \quad \text{for all } \pi. \tag{13}$$

This condition ensures the estimated gradient can distinguish the optimal model from the others, even when the objective values are close. Under condition (13), the following proposition shows that U2G is the uniformly minimum-variance unbiased estimator (UMVUE) within the proposed family, which has the optimal efficiency. The proof is deferred to Appendix Section B.3.

Proposition 2. *Among the unbiased gradient estimators defined in Definition 1 and assume* $\forall \pi,\ \lim_{|f(1)-f(0)|\to 0} \ var[g(u;\pi)] = 0$, *U2G has the uniformly minimum variance for all* π.

Specifically, the variance of U2G estimator is

$$\begin{aligned} \text{var}[g_{\text{U2G}}(u;\pi)] = & \ \pi|\pi - \frac{1}{2}|(1-\pi)\max\{\pi, 1-\pi\}[f(1)-f(0)]^2 \\ & \leq C[f(1)-f(0)]^2 \end{aligned}$$

with $C \approx 0.0388$. The SNR for U2G estimator is

$$\text{SNR}(\pi) = \sqrt{\pi(1-\pi)/(|\pi - \frac{1}{2}|\max\{\pi, 1-\pi\})},$$

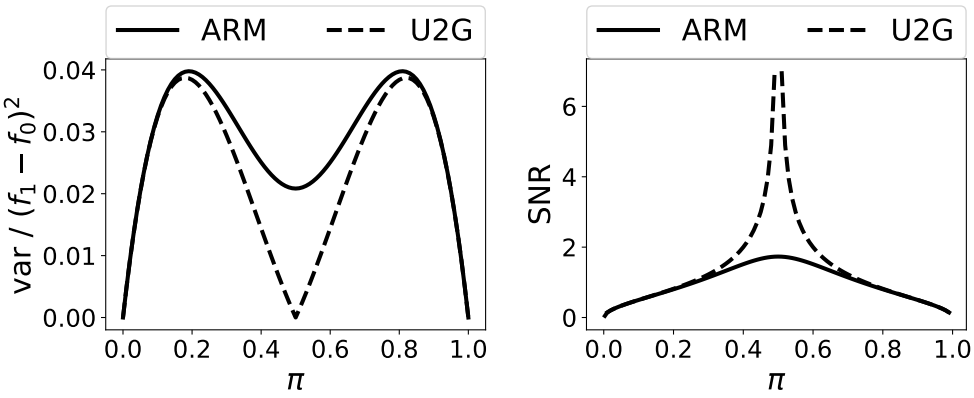


Figure 2: Variance and SNR of univariate ARM and U2G estimators.

which is the same for arbitrary function $f(\cdot)$ in the objective, and only vanishes when the algorithm converges. Similar properties hold for the ARM estimator. The variance and SNR distributions of the ARM and U2G estimators are shown in Figure 2.

3.2 Multivariate Generalization

Coming back to the objective (5), we consider the case when the latent variable $\boldsymbol{z}$ is high dimensional. When z is univariate, gradient estimation is unnecessary since the true gradient is analytic. However, in the multivariate case, the element-wise true gradient is

$$\begin{aligned}\frac{\partial}{\partial \phi_v}\mathcal{E}(\boldsymbol{\phi}) =& \frac{\partial}{\partial \phi_v}\mathbb{E}_{\boldsymbol{z}\sim\prod_{j=1}^{p} p(z_j;\sigma(\phi_j))}[f(\boldsymbol{z})] \\ =& \mathbb{E}_{\boldsymbol{z}_{-v}}[\sigma(\phi_v)(1-\sigma(\phi_v))(f(\boldsymbol{z}_{-v}, z_v = 1) - f(\boldsymbol{z}_{-v}, z_v = 0))]. \end{aligned} \tag{14}$$

To compute the v-th element of the gradient estimation by Eq. (14), we need to set the v-th element z_v as 0 and 1 while keeping other elements $\boldsymbol{z}_{-v}$ the same. This has to be done for each element of $\boldsymbol{z}$ separately. Therefore, it requires at least $2p$ function evaluations to get an unbiased gradient estimation at each step, which is computationally intractable in the large-p setting. Luckily, the estimators in Definition 1 circumvent this problem. Applying

the univariate gradient estimators in Definition 1, for multivariate $\boldsymbol{z}$ we have

$$\begin{aligned}
&\frac{\partial}{\partial\phi_v}\mathbb{E}_{\boldsymbol{z}\sim\prod_{j=1}^p p(z_j;\sigma(\phi_j))}[f(\boldsymbol{z})] = \mathbb{E}_{\boldsymbol{z}_{-v}}\frac{\partial}{\partial\phi_v}\mathbb{E}_{z_v\sim p(z_v;\sigma(\phi_v))}[f(z_v,\boldsymbol{z}_{-v})]\\
=&\mathbb{E}_{\boldsymbol{z}_{-v}}\mathbb{E}_{u_v\sim\text{Unif}(0,1)}[a(u_v;\sigma(\phi_v))f(\mathbf{1}_{[u_v<\sigma(\phi_v)]},\boldsymbol{z}_{-v}) + b(u_v;\sigma(\phi_v))f(\mathbf{1}_{[u_v>1-\sigma(\phi_v)]},\boldsymbol{z}_{-v})]\\
=&\mathbb{E}_{\boldsymbol{u}\sim\prod_{j=1}^p \text{Unif}(0,1)}[a(u_v;\sigma(\phi_v))f(\mathbf{1}_{[\boldsymbol{u}<\sigma(\boldsymbol{\phi})]}) + b(u_v;\sigma(\phi_v))f(\mathbf{1}_{[\boldsymbol{u}>1-\sigma(\boldsymbol{\phi})]})], \qquad (15)
\end{aligned}$$

where the first equality is by the factorization of $p(\boldsymbol{z})$, the second equality is by the unbiasedness of the univariate gradient estimator, and the last equality is by the law of the unconscious statistician (LOTUS) (Ross, 2014). This derivation ensures that the multivariate gradient is unbiased. It induces a key benefit that, to compute the gradient estimator, we can evaluate $f(\mathbf{1}_{[\boldsymbol{u}<\sigma(\boldsymbol{\phi})]})$ and $f(\mathbf{1}_{[\boldsymbol{u}>1-\sigma(\boldsymbol{\phi})]})$ as few as a single time with $\boldsymbol{u}\sim\prod_{j=1}^p\text{Unif}(0,1)$, and share it across all elements of the gradient vector. This greatly reduces computation time at each gradient step. Written in a vector form, the estimators in Section 3.1 have their multivariate form as

$$\begin{aligned}
\boldsymbol{g}_R(\boldsymbol{u};\sigma(\boldsymbol{\phi})) &= f(\mathbf{1}_{[\boldsymbol{u}<\sigma(\boldsymbol{\phi})]})(\mathbf{1}_{[\boldsymbol{u}<\sigma(\boldsymbol{\phi})]}-\sigma(\boldsymbol{\phi})),\\
\boldsymbol{g}_{\text{ARM}_0}(\boldsymbol{u};\sigma(\boldsymbol{\phi})) &= [f(\mathbf{1}_{[\boldsymbol{u}>1-\sigma(\boldsymbol{\phi})]}) - f(\mathbf{1}_{[\boldsymbol{u}<\sigma(\boldsymbol{\phi})]})](\boldsymbol{u}-\tfrac{1}{2})\odot|\mathbf{1}_{[\boldsymbol{u}>1-\sigma(\boldsymbol{\phi})]}-\mathbf{1}_{[\boldsymbol{u}<\sigma(\boldsymbol{\phi})]}|,\\
\boldsymbol{g}_{\text{U2G}}(\boldsymbol{u};\sigma(\boldsymbol{\phi})) &= \tfrac{1}{2}[f(\mathbf{1}_{[\boldsymbol{u}>1-\sigma(\boldsymbol{\phi})]}) - f(\mathbf{1}_{[\boldsymbol{u}<\sigma(\boldsymbol{\phi})]})]\sigma(|\boldsymbol{\phi}|)\odot(\mathbf{1}_{[\boldsymbol{u}>1-\sigma(\boldsymbol{\phi})]}-\mathbf{1}_{[\boldsymbol{u}<\sigma(\boldsymbol{\phi})]}), \qquad (16)
\end{aligned}$$

where $\boldsymbol{u}\sim\prod_{j=1}^p\text{Unif}(0,1)$, and all the operations are element-wise. Due to the indicator mask, the gradient vectors of ARM_0 and U2G are sparse when the probability close to the extremes, such as when it approaches convergence. We observe in practice that the sparsity in gradient estimation, while not required to ensure unbiasedness, can improve the stability of the convergence process.

We consider the variance of all elements in the gradient estimator, which is the diagonal of the covariance matrix. By the law of total variance, the variance of element v in the gradient vector can be decomposed as

$$\text{var}_{\boldsymbol{u}}[\boldsymbol{g}_v(\boldsymbol{u};\sigma(\boldsymbol{\phi}))] = \text{var}\{\mathbb{E}[\boldsymbol{g}_v(\boldsymbol{u};\sigma(\boldsymbol{\phi}))|\boldsymbol{u}_{-v}]\} + \mathbb{E}\{\text{var}[\boldsymbol{g}_v(\boldsymbol{u};\sigma(\boldsymbol{\phi}))|\boldsymbol{u}_{-v}]\}. \qquad (17)$$

Algorithm 1 Best subset selection with probabilistic reformulation

input : Bernoulli distribution $\{q_{\phi_j}(z_j)\}_{j\in[p]}$ with probability $\{\sigma(\phi_j)\}_{j\in[p]}$, target $\mathcal{E}(\boldsymbol{\phi}) = \mathbb{E}_{\boldsymbol{z}\sim p_{\boldsymbol{\phi}}(\boldsymbol{z})}[f(\boldsymbol{z})]$, $\boldsymbol{z} = (z_1, \cdots, z_p)$, $\boldsymbol{\phi} = (\phi_1, \cdots, \phi_p)$, $p_{\boldsymbol{\phi}}(\boldsymbol{z}) = \prod_{j=1}^{p} p_{\phi_j}(z_j)$

output : Maximum likelihood estimator of $p_{\boldsymbol{\phi}}(\boldsymbol{z})$ as $\hat{\boldsymbol{z}} = \mathbf{1}_{[\sigma(\boldsymbol{\phi})>1/2]}$

Initialize $\boldsymbol{\phi}$ randomly

while *not converged* **do**

Sample $\boldsymbol{u}_k \overset{i.i.d.}{\sim} \prod_{j=1}^{p} \mathrm{Unif}(0,1)$ for $k = 1, \cdots, K$

Evaluate $f(\mathbf{1}_{[\boldsymbol{u}_k > 1-\sigma(\boldsymbol{\phi})]})$ and $f(\mathbf{1}_{[\boldsymbol{u}_k < \sigma(\boldsymbol{\phi})]})$

Compute $\boldsymbol{g}_k = g(\boldsymbol{u}_k; \sigma(\boldsymbol{\phi}), f)$ by an estimator in Eq. (16)

Update $\boldsymbol{\phi} = \boldsymbol{\phi} - \frac{1}{K}\rho \sum_{k=1}^{K} \boldsymbol{g}_k$ with stepsizes ρ

end

The first term on the right-hand side (RHS) of Eq. (17) is the irreducible variance, shared by all unbiased gradient estimators, which can be further computed as

$$\mathrm{var}_{\boldsymbol{u}_{-v}}\{\mathbb{E}_{u_v}[\boldsymbol{g}_v(\boldsymbol{u}; \sigma(\boldsymbol{\phi}))|\boldsymbol{u}_{-v}]\} = (\pi_v)^2(1-\pi_v)^2 \mathrm{var}_{\boldsymbol{u}}[\Delta_{\boldsymbol{z},v} f],$$

with $\boldsymbol{z} = \mathbf{1}_{[\boldsymbol{u}<\sigma(\boldsymbol{\phi})]}$, $\tilde{\boldsymbol{z}} \in \{0,1\}^p$ which differs from $\boldsymbol{z}$ only at the v-th dimension, and $\Delta_{\boldsymbol{z},v} f := f(\tilde{\boldsymbol{z}}) - f(\boldsymbol{z})$. It measures the variance of the effect on $f(\cdot)$ by knocking out the v-th covariate. The second term measures the average variance in a single dimension. Given a fixed $\boldsymbol{u}_{-v}$, as shown in the univariate case, U2G estimator has the minimal variance for all estimators in Definition 1 with non-vanishing SNR. Therefore by averaging over all $\boldsymbol{u}_{-v}$, the second term of U2G estimator is small; hence the total variance is well controlled in the multivariate case.

4 Convergence in Expectation

In this section, we provide theoretical insights to the convergence properties of the gradient method under the expectation of data generation and gradient estimation. We assume

that the observations $(\mathbf{X}, \boldsymbol{y})$ are generated from the following model with the active set $\mathcal{A} \subset \{1, 2, \cdots, p\}$

$$\boldsymbol{y} = \mathbf{X}\boldsymbol{\beta}^* + \boldsymbol{\epsilon}, \quad \boldsymbol{\epsilon} \sim \mathcal{N}(0, \sigma^2 \mathbf{I}), \tag{18}$$

where $\beta_j^* = 0$ for $j \notin \mathcal{A}$. Let $\boldsymbol{z}^* \in \{0,1\}^p$ indicate the true active set where z_j^* equals 1 if $j \in \mathcal{A}$ and 0 otherwise. We assume a random design matrix $\mathbf{X} = (\boldsymbol{x}_1, \cdots, \boldsymbol{x}_n)^\top$ in which $\boldsymbol{x}_i \sim \mathcal{N}(0, \mathbf{I}_p)$ for $i \in [n]$. In order to ease the presentation, we denote

$$f_{\mathbf{X},\boldsymbol{y}}(\boldsymbol{z}) = \min_{\boldsymbol{\alpha}} \frac{1}{n} \ \|\boldsymbol{y} - \mathbf{X}(\boldsymbol{\alpha} \odot \boldsymbol{z})\|_2^2 + \lambda \|\boldsymbol{z}\|_0 . \tag{19}$$

Here we use subscripts to make the dependency of f on $(\mathbf{X}, \boldsymbol{y})$ explicit. Denote $\mathbf{X}_{\boldsymbol{z}} \in \mathbb{R}^{n \times \|\boldsymbol{z}\|_0}$ as the matrix consisting of $\{X_j : z_j \neq 0\}$, and $\mathbf{X}_{-\boldsymbol{z}}$ the complement in the design matrix. Furthermore, for any $\boldsymbol{z}, \tilde{\boldsymbol{z}} \in \{0,1\}^p$ such that $z_k = 0$, $\tilde{z}_k = 1$, $z_j = \tilde{z}_j$ for all $j \neq k$, we denote $\Delta_{\boldsymbol{z},k} f := f_{\mathbf{X},\boldsymbol{y}}(\tilde{\boldsymbol{z}}) - f_{\mathbf{X},\boldsymbol{y}}(\boldsymbol{z})$. All the proof details in this section are given in the appendix.

First, we have the following lemma for the expectation of the gradient over the randomness of $\boldsymbol{u}$, given any fixed training data $(\mathbf{X}, \boldsymbol{y})$.

Lemma 1. *Consider $g_{\mathrm{U2G}}(\boldsymbol{u}; \sigma(\boldsymbol{\phi}))$ as in Eq. (16). For each $(\mathbf{X}, \boldsymbol{y})$, we have*

$$\mathbb{E}_{\boldsymbol{u} \sim \prod_{j=1}^p \mathrm{Unif}(u_j; 0, 1)}[\boldsymbol{g}_{\mathrm{U2G}}(\boldsymbol{u}; \sigma(\boldsymbol{\phi}))] = \boldsymbol{\pi}(1 - \boldsymbol{\pi}) \odot \mathbb{E}_{\boldsymbol{u}}[\Delta_{\boldsymbol{z}} f],$$

where $\boldsymbol{\pi} = (\sigma(\phi_1), \cdots, \sigma(\phi_p))$, $\boldsymbol{z} = \mathbf{1}_{[\boldsymbol{u} > 1 - \sigma(\boldsymbol{\phi})]}$, and $\Delta_{\boldsymbol{z}} f = (\Delta_{\boldsymbol{z},1} f, \cdots, \Delta_{\boldsymbol{z},p} f)$.

Lemma 1 shows that the gradient is closely related to $\Delta_{\boldsymbol{z}} f$, whose randomness comes from latent variable $\boldsymbol{u}$ and data $(\mathbf{X}, \boldsymbol{y})$. By analyzing the expectation of $\Delta_{\boldsymbol{z}} f$, the following result establishes the expectation of stochastic gradients

Lemma 2. *Given $g_{\mathrm{U2G}}(\boldsymbol{u}; \sigma(\boldsymbol{\phi}))$ in Eq. (16), the expected gradient is*

$$\mathbb{E}_{\mathbf{X},\boldsymbol{y},\boldsymbol{u}}[g_{\mathrm{U2G}}(\boldsymbol{u}; \sigma(\boldsymbol{\phi}))_k] = \Bigg[\lambda - \Bigg(\frac{(n - \mathbb{E}[\|\boldsymbol{z}\|_0 \,\big|\, \|\boldsymbol{z}\|_0 < n] - 1)(\beta_k^*)^2}{n}$$

$$+\frac{\sigma^2+\mathbb{E}_{\boldsymbol{u}}[\left\|\boldsymbol{\beta}^*_{-\boldsymbol{z}}\right\|_2^2\Big|\|\boldsymbol{z}\|_0<n]}{n}\Big)p(\|\boldsymbol{z}\|_0<n)\Big]\times\pi_k(1-\pi_k), \tag{20}$$

for any $k\in\{1,2,\cdots,p\}$ *where* $\boldsymbol{\pi}=(\sigma(\phi_1),\cdots,\sigma(\phi_p))$ *and* $\boldsymbol{z}=\mathbf{1}_{[\boldsymbol{u}>1-\sigma(\boldsymbol{\phi})]}$.

In the following proposition, based on the results of Lemma 2, we show that if the sample size and true coefficient magnitude are not too small, then with proper hyper-parameter λ controlling the penalty strength, each element of the expected gradient points to the direction that can recover the true active set.

Proposition 3. *Assume* $\sum_{j=1}^p\sigma(\boldsymbol{\phi}_j)/(n-1)\le 1-\eta$, *for certain* $\eta\in(\frac{1}{n},1)$. *If* n *is sufficiently large such that* $\sqrt{p\log(n)/2(n-1)^2}\le\eta$ *and* $(\|\boldsymbol{\beta}^*\|_2^2+\sigma^2)/(n-1)\min_{k\in\mathcal{A}}(\beta_k^*)^2\le\eta$, *then there exists* $\lambda>0$ *such that*

$$\mathbb{E}_{\mathbf{X},\boldsymbol{y},\boldsymbol{u}}[g_{\mathrm{U2G}}(\boldsymbol{u};\sigma(\boldsymbol{\phi}))_j]<0,\ \forall j\in\mathcal{A};\quad \mathbb{E}_{\mathbf{X},\boldsymbol{y},\boldsymbol{u}}[g_{U2G}(\boldsymbol{u};\sigma(\boldsymbol{\phi}))_j]>0,\ \forall j\notin\mathcal{A}.$$

Remark 1. *If the gradient points to the right direction element-wisely, then for each gradient step, in expectation,* π_j *increases if and only if* $j\in\mathcal{A}$. *Therefore, if*

$$\varpi=\left(1-\min\left\{\sqrt{\frac{p\log(n)}{2(n-1)^2}},\frac{\|\boldsymbol{\beta}^*\|_2^2+\sigma^2}{(n-1)\min_{k\in\mathcal{A}}(\beta_k^*)^2}\right\}\right)n>S, \tag{21}$$

with initialization

$$\sum_{j=1}^p\pi_j^{(0)}\le\varpi-S$$

in expectation, $\boldsymbol{\pi}$ *in Algorithm 1 converges to the indicator of the true active set.*

The proof of Proposition 3 provides a guidance in choosing hyperparameter λ as

$$\lambda\in\left(\frac{\|\boldsymbol{\beta}^*\|_2^2+\sigma^2}{n},\frac{n-1}{n}(\eta-\frac{1}{n})\min_{j\in\mathcal{A}}(\beta_j^*)^2\right). \tag{22}$$

Though in practice the true coefficient $\boldsymbol{\beta}^*$ is unknown a priori, choosing $\lambda = \log(n)/(2n)$ as BIC falls in the region (22) asymptotically, and serves as a good initial point for the cross validation in the finite sample case. Now, we study the convergence rate of the updates of Algorithm 1 in expectation, namely, with precise gradient each step. We show that these updates converge to the ground truth after $\mathcal{O}(1/\epsilon)$ steps where $\epsilon > 0$ is the desired accuracy.

Theorem 2. *Let the update be* $\boldsymbol{\phi}^{(t+1)} = \boldsymbol{\phi}^{(t)} - \rho \mathbb{E}_{\mathbf{X},\boldsymbol{y},\boldsymbol{u}}[g_{\mathrm{U2G}}(\boldsymbol{u}; \sigma(\boldsymbol{\phi}^{(t)}))]$ *where* $\rho > 0$ *is the given step size. We assume that* $\lambda \in \mathcal{I}$ *where* $\mathcal{I}$ *is defined as in Eq. (22) Furthermore, the initialization* $\boldsymbol{\phi}^{(0)}$ *satisfies that* $\sum_{j=1}^{p} \sigma(\phi_j^{(0)}) \leq \varpi - S$ *where* ϖ *is defined in Eq. (21). Then, the following holds:*

(a) For any $j \in \mathcal{A}$*, as long as* $t \geq t_j^1$

$$\left(1 - \sigma(\phi_j^{(t)})\right)\left[1 - c_1\sigma^2(\phi_j^{(t)})\left(1 - \sigma(\phi_j^{(t)})\right)\right] \leq 1 - \sigma(\phi_j^{(t+1)}) \leq \left(1 - \sigma(\phi_j^{(t)})\right) \times \left[1 - C_1(\sigma(\phi_j^{(t)}))^2\left(1 - \sigma(\phi_j^{(t)})\right)\right].$$

(b) For any $j \notin \mathcal{A}$ *and* $t \geq t_j^2$

$$\sigma(\phi_j^{(t)})\left[1 - c_2\left(1 - \sigma(\phi_j^{(t)})\right)^2\sigma(\phi_j^{(t)})\right] \leq \sigma(\phi_j^{(t+1)}) \leq \sigma(\phi_j^{(t)}) \times \left[1 - C_2\left(1 - \sigma(\phi_j^{(t)})\right)^2\sigma(\phi_j^{(t)})\right].$$

Here, with model parameters $\tau = \{n, p, \sigma, \boldsymbol{\beta}^*\}$*,* c_1, c_2, C_1, C_2 *are some positive constants depending only on* τ *and* ρ*.* t_j^1 *and* t_j^2 *are constants depending on* τ*,* ρ *and initial* $\phi_j^{(0)}$*.*

Remark 2. *(i) The upper bounds of Theorem 2 demonstrate that when* $j \in \mathcal{A}$*,* $1 - \sigma(\phi_j^{(t)}) \leq \epsilon$ *after* $t = \mathcal{O}(\epsilon^{-1})$ *steps, which is* sub-linear*. Similarly, when* $j \notin \mathcal{A}$*, it takes* $t = \mathcal{O}(\epsilon^{-1})$ *number of iterations for* $\sigma(\phi_j^{(t)})$ *to be within* ϵ *radius from 0. The lower bounds in Theorem 2 indicate that these sub-linear complexities are tight. As a consequence, in expectation, the updates of Algorithm 1 converge to the global optima at the sub-linear rate* $\mathcal{O}(\epsilon^{-1})$*.*

(ii) The results of Theorem 2 also yield an insight into the choice of step size ρ. Based on the specific forms of c_1 and c_2 in the proof, we need the step size ρ to satisfy

$$\rho < \min\left\{\frac{2}{\lambda}, \frac{2}{\max_j\{(\beta_j^*)^2\} - \lambda + (\sigma^2 + \|\boldsymbol{\beta}^*\|_2^2)/n}\right\}. \tag{23}$$

The convergence properties we present in this section are under the expectation. The empirical performance of a low variance gradient estimator such as U2G can be close to the theoretical results, as shown in Section 6. Before that, we extend the proposed gradient methods from solving the frequentist objective (4) to solving the L_0-regularized regression in the Bayesian paradigm.

5 Bayesian L_0-Regularized Regression

The objective function studied in Section 3.2 is a general objective function with the penalized linear regression function (4) as a special case. In this section, we cast the best subset selection as a posterior inference problem, and use the proposed gradient estimators to solve the new objective function. The Bayesian linear regression model can be expressed hierarchically, where the latent variables $\boldsymbol{\alpha} \in \mathbb{R}^p$ and $\boldsymbol{z} \in \{0,1\}^p$ have a Gaussian and Bernoulli prior, respectively. In particular, the model is given by:

$$\begin{aligned} y_i &\sim \mathcal{N}(\boldsymbol{x}_i^\top(\boldsymbol{\alpha} \odot \boldsymbol{z}), \sigma^2), \ i \in [n] \\ \boldsymbol{\alpha} &\sim \mathcal{N}(\boldsymbol{\alpha}; \mathbf{0}, \boldsymbol{\Sigma}_{\boldsymbol{\alpha}}), \quad z_j \sim \text{Bern}(\sigma(-\lambda_0)), \ j \in [p]. \end{aligned} \tag{24}$$

The hyper-parameter $\boldsymbol{\Sigma}_{\boldsymbol{\alpha}}$ is set as $\sigma_\alpha^2\mathbf{I}$ and λ_0 controls *a priori* degree of shrinkage. The priors can be jointly written as

$$p(\boldsymbol{\alpha}, \boldsymbol{z}; \lambda_0, \boldsymbol{\Sigma}_{\boldsymbol{\alpha}})\sigma(-\lambda_0)^{\|\boldsymbol{z}\|_0}(1-\sigma(-\lambda_0))^{p-\|\boldsymbol{z}\|_0}(2\pi)^{-p/2}|\boldsymbol{\Sigma}_{\boldsymbol{\alpha}}|^{-1/2}\exp(-\tfrac{1}{2}\boldsymbol{\alpha}^\top\boldsymbol{\Sigma}_{\boldsymbol{\alpha}}^{-1}\boldsymbol{\alpha}). \tag{25}$$

In Bayesian statistics, a standard approach to the variable selection problem is to utilize the spike-and-slab prior. Setting the regression parameter $\boldsymbol{\beta} = \boldsymbol{\alpha} \odot \boldsymbol{z}$, the prior for $\boldsymbol{\beta}$ is a

spike-and-slab prior which has a slab Gaussian component and a spike component at 0:

$$p(\boldsymbol{\beta}) = \prod_{j=1}^{p} \left[\sigma(\lambda_0)\delta_0 + (1-\sigma(\lambda_0))\mathcal{N}(0, \sigma_\alpha^2)\right]. \tag{26}$$

With the likelihood and prior, the posterior distribution is

$$p(\boldsymbol{\alpha}, \boldsymbol{z} \,|\, \mathbf{X}, \boldsymbol{y}; \lambda_0, \boldsymbol{\Sigma}_{\boldsymbol{\alpha}}) \propto \exp\Big(-(2\sigma^2)^{-1} \left\|\boldsymbol{y} - \mathbf{X}(\boldsymbol{\alpha} \odot \boldsymbol{z})\right\|_2^2 - \boldsymbol{\alpha}^\top \boldsymbol{\Sigma}_{\boldsymbol{\alpha}}^{-1} \boldsymbol{\alpha}/2 - \lambda_0 \left\|\boldsymbol{z}\right\|_0 \Big). \tag{27}$$

To find the *maximum a posterior* (MAP) estimator, we can minimize the negative logarithm of the posterior as:

$$\min_{\boldsymbol{\alpha}, \boldsymbol{z}} \frac{1}{2} \left\|\boldsymbol{y} - \mathbf{X}(\boldsymbol{\alpha} \odot \boldsymbol{z})\right\|_2^2 + \frac{\sigma^2}{2\sigma_\alpha^2} \left\|\boldsymbol{\alpha}\right\|_2^2 + \sigma^2 \lambda_0 \left\|\boldsymbol{z}\right\|_0 . \tag{28}$$

Hence the MAP solution is equivalent to the linear regression solution with combined L_2 and L_0 penalties. It has been observed that adding additional L_2 penalty to the L_0 penalty can in practice improve the computational efficiency (Liu and Wu, 2007; Hazimeh and Mazumder, 2018). When the variance σ_α^2 of the slab component in the $\boldsymbol{\beta}$ prior is large, the MAP solution is close to the best subset solution.

Directly solving Eq. (28) is a combinatorial problem. To overcome the computational challenge, we resort to variational inference (VI) to approximate the posterior distribution and MAP estimation. To be consistent with VI nomenclature, here we deviate from the notation in Eq. (5), and use $p(\boldsymbol{z})$ as the prior and $q_{\boldsymbol{\phi}}(\boldsymbol{\alpha}, \boldsymbol{z})$ as the variational distribution with parameter $\boldsymbol{\phi}$. The VI methods find an approximated posterior by minimizing the Kullback–Leibler (KL) divergence from $p(\boldsymbol{\alpha}, \boldsymbol{z} \,|\, \mathbf{X}, \boldsymbol{y}))$ to $q_{\boldsymbol{\phi}}(\boldsymbol{\alpha}, \boldsymbol{z})$, denoted as $D_{\mathrm{KL}}(q_{\boldsymbol{\phi}}(\boldsymbol{\alpha}, \boldsymbol{z}) || p(\boldsymbol{\alpha}, \boldsymbol{z} \,|\, \mathbf{X}, \boldsymbol{y}))$. Since the true posterior is often unknown, equivalently we can maximize the evidence lower bound (ELBO) (Blei et al., 2017) as a tractable objective, defined as

$$\begin{aligned} \mathcal{L}(\boldsymbol{\phi}) = & \log p(\boldsymbol{y}|\mathbf{X}) - D_{\mathrm{KL}}(q_{\boldsymbol{\phi}}(\boldsymbol{\alpha}, \boldsymbol{z}) || p(\boldsymbol{\alpha}, \boldsymbol{z}|\mathbf{X}, \boldsymbol{y})) \\ = & \mathbb{E}_{q_{\boldsymbol{\phi}}(\boldsymbol{\alpha}, \boldsymbol{z})} \log \left[p(\boldsymbol{y}|\mathbf{X}, \boldsymbol{z}, \boldsymbol{\alpha}) p(\boldsymbol{\alpha}, \boldsymbol{z}; \lambda_0, \sigma_\alpha^2) / q_{\boldsymbol{\phi}}(\boldsymbol{\alpha}, \boldsymbol{z})\right]. \end{aligned} \tag{29}$$

Due to the limited expressiveness of the variational distribution and the zero-forcing property of the KL divergence, variational method often underestimates the posterior uncertainty. Recent analysis, however, provides theoretical guarantees to its accuracy of point estimation. For specific models such as Latent Dirichlet Allocation (LDA) and Stochastic Blockmodel (SBM), Bickel et al. (2013); Pati et al. (2018); Zhang and Zhou (2017) and Yin et al. (2020) have proved the consistency and asymptotic normality of the VI point estimation. For more general cases, Wang and Blei (2018) have proved a variational Bernstein-von Mises theorem, which states that the variational posterior converges to the KL minimizer of a normal distribution, centered at the truth. Since we are most interested in the MAP solution, the point estimation of VI is accurate, as validated in simulation in Section 6.

To further improve the accuracy, we propose a tightened ELBO. The gap between ELBO and marginal likelihood is equivalent to the KL divergence from the posterior to variational distribution. With the chain rule of KL divergence, this gap can be decomposed as

$$D_{\mathrm{KL}}(q_{\boldsymbol{\phi}}(\boldsymbol{\alpha},\boldsymbol{z})||p(\boldsymbol{\alpha},\boldsymbol{z}|\mathbf{X},\boldsymbol{y})) = D_{\mathrm{KL}}(q_{\boldsymbol{\phi}}(\boldsymbol{z})||p(\boldsymbol{z}|\mathbf{X},\boldsymbol{y})) + \mathbb{E}_{q(\boldsymbol{z})} D_{\mathrm{KL}}(q(\boldsymbol{\alpha}|\boldsymbol{z})||p(\boldsymbol{\alpha}|\mathbf{X},\boldsymbol{z},\boldsymbol{y})). \quad (30)$$

Therefore, we can choose $q(\boldsymbol{\alpha}|\boldsymbol{z}) = p(\boldsymbol{\alpha}|\mathbf{X},\boldsymbol{z},\boldsymbol{y})$ and set the second term on the RHS of Eq. (30) as 0. Marginalizing out the latent variable $\boldsymbol{\alpha}$, we get a tightened ELBO as

$$\mathcal{L}(\boldsymbol{\phi}) = \mathbb{E}_{q_{\boldsymbol{\phi}}(\boldsymbol{z})} \log \Big[p(\boldsymbol{y}|\mathbf{X},\boldsymbol{z};\sigma_\alpha^2) p(\boldsymbol{z};\lambda_0)/q_{\boldsymbol{\phi}}(\boldsymbol{z}) \Big]. \quad (31)$$

The variational distribution is chosen as $q_{\boldsymbol{\phi}}(\boldsymbol{z}) = \prod_{j=1}^{p} \mathrm{Bern}(z_j;\sigma(\phi_j))$ and the likelihood, after marginalizing out $\boldsymbol{\alpha}$, is

$$p(\boldsymbol{y}|\mathbf{X},\boldsymbol{z};\sigma_\alpha^2) = \mathcal{N}(\boldsymbol{y};\mathbf{0},\mathbf{X}(\sigma_\alpha^2\mathbf{I}\odot(\boldsymbol{z}\boldsymbol{z}^\top))\mathbf{X}^\top + \sigma^2\mathbf{I}_n). \quad (32)$$

Setting $f(\boldsymbol{z}) = \log[p(\boldsymbol{y}|\mathbf{X},\boldsymbol{z};\sigma_\alpha^2)p(\boldsymbol{z};\lambda_0)/q_{\boldsymbol{\phi}}(\boldsymbol{z})]$, the objective (31) can be considered as a special case of the general optimization objective (5). Therefore, the unbiased gradient estimators can be directly applied to maximizing the ELBO. The variational objective,

comparing to the frequentist objective in Eq. (4), does not require computing an OLS solution when evaluating $f(\boldsymbol{z})$, hence improves efficiency, especially when n is large.

6 Experimental Results

In this section, we examine the performance of our approach on a variety of synthetic and semi-synthetic datasets whose underlying ground truth sparsity patterns are accessible. Let $\widehat{\boldsymbol{\beta}}$ denote the estimated coefficients, $\boldsymbol{\beta}^*$ the true coefficients, and $(\boldsymbol{x}, y)$ a sample. We assume the covariates are centered, i.e. $\boldsymbol{x} \sim \mathcal{N}(0, \boldsymbol{\Sigma})$. To measure the level of information in data, we use the data SNR, defined as

$$\mathrm{SNR}_d := \mathrm{var}(\boldsymbol{x}^\top \boldsymbol{\beta}^*)/\mathrm{var}(\epsilon) = \boldsymbol{\beta}^{*\top} \boldsymbol{\Sigma} \boldsymbol{\beta}^* / \sigma^2.$$

The evaluation metrics throughout can be categorized as two groups: one group of metrics measures the out-of-sample predictive performance and the other group measures the recovery quality of the sparsity pattern (Bertsimas et al., 2016; Hastie et al., 2017). The metrics for the predictive performance that we use are

- *Relative risk (RR)* that measures how model prediction deviates from the oracle prediction, the perfect score being 0:

$$\mathrm{RR}(\widehat{\boldsymbol{\beta}}) = \frac{\mathbb{E}(\boldsymbol{x}^\top \widehat{\boldsymbol{\beta}} - \boldsymbol{x}^\top \boldsymbol{\beta}^*)^2}{\mathbb{E}(\boldsymbol{x}^\top \boldsymbol{\beta}^*)^2} = \frac{(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*)^\top \Sigma (\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*)}{\boldsymbol{\beta}^{*\top} \Sigma \boldsymbol{\beta}^*}.$$

- *Relative test error (RTE)* that measures the relative test MSE compared with the oracle predictor, the perfect score being 1:

$$\mathrm{RTE}(\widehat{\boldsymbol{\beta}}) = \frac{\mathbb{E}(y - \boldsymbol{x}^\top \widehat{\boldsymbol{\beta}})^2}{\mathbb{E}(y - \boldsymbol{x}^\top \boldsymbol{\beta}^*)^2} = \frac{(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*)^\top \Sigma (\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*) + \sigma^2}{\sigma^2}.$$

- *Proportion of variance explained (PVE)* that measures the proportion of variance in the response variable explained by the model, the perfect score being $\mathrm{SNR}_d/(1 +$

SNR_d):

$$\mathrm{PVE}(\widehat{\boldsymbol{\beta}}) = 1 - \frac{\mathbb{E}(y - \boldsymbol{x}^T\widehat{\boldsymbol{\beta}})^2}{\mathrm{var}(y)} = 1 - \frac{(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*)^T\Sigma(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*) + \sigma^2}{\boldsymbol{\beta}^{*T}\Sigma\boldsymbol{\beta}^* + \sigma^2}.$$

To evaluate sparse pattern recovery, we consider the *size of estimated active set* (Linero, 2018) as well as the *precision*, *recall*, and *F1* scores, given by prec = TP/(TP + FP), rec = TP/(TP +FN), and F1 = $2 \cdot$ prec $\cdot$ rec/(prec + rec), where TP denotes the number of predictors correctly flagged as influential, FP denotes the number of predictors incorrectly flagged as influential, and FN denotes the number of predictors incorrectly flagged as noninfluential. The F1 score is an overall summary that balances precision and recall.

We compare the proposed methods with the representatives from a set of sparse variable selection methods. We choose Lasso (Tibshirani, 1996) as a convex penalty regularized method, SCAD (Fan and Li, 2001) as a non-convex penalty regularized method, and MIO (Bertsimas et al., 2016) as a best subset selection method. Throughout the experiments, if not specified, we set the number of Monte Carlo samples for the gradient-based method as $K = 20$, which takes seconds to converge when the number of covariates is in thousands, running on a cluster node with two Intel E5-2690 v3 12-core (Haswell) processors. To determine the convergence, we compute the entropy for the j-th covariate as $H_j = -p_j \log(p_j)$, and stop the training when the average of the 1% largest entropy is below 0.1. According to the theoretical analysis in Section 4, for the gradient-based methods we choose hyperparameter λ via cross validation with starting point $\log(n)/(2n)$, and use a constant step-size in SGD sufficiently smaller than $2/\lambda$. For methods in comparison, Lasso is implemented by R package *"glmnet"* (Friedman et al., 2010) and SCAD is implemented by R package *"picasso"* (Ge et al., 2019). The R package *"bestsubset"* (Hastie et al., 2018) is used to obtain the results for the best subset selection with MIO (Bertsimas et al., 2016). For MIO, $\hat{S}$ in Eq. (2) is set to the oracle sparsity level S. If not specified, we use the default configurations in the R packages.

Experiment 1: Synthetic data with correlated covariates

We consider the example in Fan and Li (2001) with increased dimension. The true coefficient is set as $\boldsymbol{\beta}^* = (3, 1.5, 0, 0, 2, \underbrace{0, \cdots, 0}_{195}) \in \mathbb{R}^{200}$. The design matrix $\mathbf{X}$ is a collection of n i.i.d. samples generated from $\mathcal{N}(\mathbf{0}, \Sigma)$ where $\Sigma_{ij} = \rho^{|i-j|}$ and $\boldsymbol{y} \sim \mathcal{N}(\mathbf{X}\boldsymbol{\beta}^*, \sigma^2\mathbf{I})$. In the experiment, we set $\rho = 0.5$. We test the comparative algorithms in both high and low SNR_d regimes, by setting the standard deviation of noise as $\sigma = 1$ and $\sigma = 3$.

Table 2: Results of the variable selection simulation study with $n = 60, p = 200, S = 3$. Reported results are the mean of 100 independent trials.

	Precision	Recall	F1	Nonzero	RR	RTE	PVE
$n = 60, p = 200, \sigma = 1, \text{SNR}_d = 21.3$							
LASSO	0.780	1.000	0.852	4.65	0.039	1.830	0.918
SCAD	0.983	1.000	0.990	3.07	0.013	1.271	0.943
MIO	1.000	1.000	1.000	3.00	0.003	1.056	0.952
REINFORCE	0.089	0.657	0.153	32.4	1.601	35.01	-
ARM_0	0.992	1.000	0.996	3.03	0.003	1.067	0.952
U2G	0.990	1.0000	0.994	3.04	0.003	1.069	0.952
ARM_0(VI)	0.950	1.000	0.971	3.21	0.005	1.107	0.950
U2G(VI)	0.950	1.000	0.971	3.21	0.005	1.107	0.95
$n = 60, p = 200, \sigma = 3, \text{SNR}_d = 2.4$							
LASSO	0.747	0.850	0.745	4.32	0.284	1.671	0.503
SCAD	0.722	0.777	0.721	3.51	0.214	1.506	0.552
MIO	0.780	0.780	0.780	3.00	0.125	1.294	0.615
REINFORCE	0.092	0.503	0.151	20.2	1.092	3.579	-
ARM_0	0.856	0.863	0.844	3.15	0.107	1.251	0.628

U2G	0.860	0.853	0.857	3.14	0.111	1.263	0.624
ARM_0(VI)	0.921	0.890	0.889	2.96	0.081	1.190	0.646
U2G(VI)	0.920	0.887	0.885	2.95	0.081	1.190	0.646

In general, we find that the non-L_0-based methods tend to select larger active sets than the L_0-regularized regression. Furthermore, the L_0 penalty only forces the coefficients of irrelevant covariates to 0 and does not shrink the coefficients of the relevant covariates. In the high SNR_d setting, due to this oracle property, when the performances of active set recovery are similar, the best subset methods achieve better predictive accuracy. When SNR_d is low, among the best subset methods, the gradient-based methods have better out-of-sample prediction and active set recovery than MIO. In both SNR settings and for both objectives (4) and (31), U2G performs on par with or better than ARM_0, while REINFORCE estimator fails to select variables due to high gradient variance.

We show the regularization path of L_0 regression in Figure 3, with $n = 60, p = 200, \sigma = 1$, and independent covariates. When λ decreases, the number of selected variable increases. The test error first decreases when the correct covariates join the selection, and then increases as additional incorrect covariates are selected. As the top panel shows, for a wide range of λ values, the L_0-regularized regression recovers the true active set and keeps the estimation of the relevant covariates to their true value without shrinkage.

Experiment 2: Synthetic data with independent covariates

We consider the experiment in Bertsimas et al. (2016) and Hastie et al. (2017). The true coefficients have the first 10 elements equal to 1 as $\boldsymbol{\beta}^* = (\underbrace{1, \ldots, 1}_{10}, \underbrace{0, \ldots, 0}_{990})$. The covariates $\boldsymbol{x}_i \in \mathbb{R}^{1000}, i \in [n]$ are sampled i.i.d. from a zero mean isotropic Gaussian distribution. The error variance σ^2 is adjusted such that the SNR_d equals 5 or 7. In this example, we set $(n, p, S) = (100, 1000, 10)$. We set the maximum time for Gurobi running in MIO as

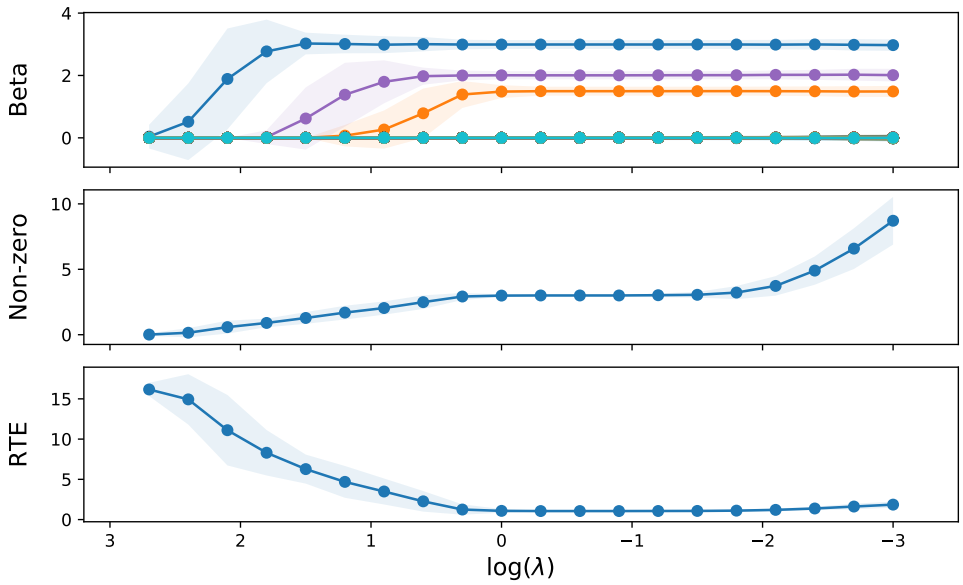


Figure 3: Regularized path for L_0-regularized regression estimated by U2G gradient, with $n = 60, p = 200, \sigma = 1$. The dotted curves are the mean of 100 independent trials and the shaded areas represent the standard deviation.

200 seconds, which doubles the default setting and takes about 5-10 times longer than the gradient-based methods to converge. From Table 3, the gradient-based method outperforms MIO on both active set recovery and predictive accuracy, on both SNR_d settings. Due to the L_0 constraint and the hard thresholding, MIO gets the number of estimated nonzero elements the same as what the $\hat{S}$ is set to be, but there are many false-positive and false-negative estimations, as shown by the precision and recall metrics. In contrast, our methods can get the sparsity level close to the ground truth, achieving accurate sparsity recovery, and hence low testing error. In the following experiments, we will stick to the U2G estimator, since it performs no worse than other gradient estimators both theoretically and empirically.

Table 3: Results of the variable selection simulation study, with $n = 100, p = 1000, S = 10$. Reported results are the mean of 100 independent trials.

	Precision	Recall	F1	Nonzero	RR	RTE	PVE
$n = 100, p = 1000, \text{SNR}_d = 7$							
Lasso	0.173	0.992	0.292	60.96	0.239	2.675	0.666
SCAD	0.550	0.908	0.675	17.65	0.439	4.072	0.491
MIO	0.772	0.772	0.772	10.00	0.293	3.050	0.619
REINFORCE	0.012	0.488	0.02	470.1	1.071	8.502	-
ARM_0	0.949	0.921	0.934	9.695	0.114	1.798	0.775
U2G	0.947	0.924	0.934	9.763	0.112	1.782	0.777
ARM_0(VI)	0.981	0.997	0.988	10.20	0.025	1.178	0.853
U2G(VI)	0.975	0.992	0.983	10.20	0.034	1.24	0.845
$n = 100, p = 1000, \text{SNR}_d = 5$							
Lasso	0.179	0.971	0.299	59.24	0.33	2.649	0.559
SCAD	0.465	0.886	0.603	20.13	0.472	3.361	0.44
MIO	0.674	0.674	0.674	10.00	0.444	3.220	0.463
REINFORCE	0.011	0.497	0.021	471.6	1.085	6.426	-
‘ ARM_0	0.906	0.906	0.904	10.01	0.154	1.771	0.705
U2G	0.905	0.909	0.905	10.07	0.151	1.755	0.708
ARM_0(VI)	0.954	0.967	0.96	10.13	0.075	1.374	0.771
U2G(VI)	0.955	0.964	0.959	10.07	0.077	1.385	0.769

Experiment 3: Semi-synthetic data

We further benchmark our methods on the Prostate cancer dataset, a real-world microar-

ray dataset (Singh et al., 2002). The regularized models are widely used for gene selection as biomarkers in analyzing microarray data, which is often high-dimensional with a large number of genes and a small number of samples. The original Prostate dataset contains the expression profiles of 12,600 genes for 50 normal tissues and 52 prostate tumor tissues. Similar to Bertsimas et al. (2016), we reduce the number of covariates by choosing 1000 genes which maximally correlate (in absolute value) with the tumor type. For the active set, we first choose five gene biomarkers correlated the most with the tumor type, which induces high multi-collinearity in the chosen genes. We also choose five gene biomarkers with pairwise correlation in $(-0.7, 0.7)$ so that the multi-collinearity is moderate. The pairwise correlations are shown in Appendix Figure A.1. For each type of active set, we separately create a semi-synthetic data set $\boldsymbol{y} \sim \mathcal{N}(\mathbf{X}\boldsymbol{\beta}^*, \sigma^2\mathbf{I})$, $\mathbf{X} \in \mathbb{R}^{102\times 1000}$, where the coefficients are one for the chosen covariates and zero for the rest. The σ^2 is set to let $\text{SNR}_d = 5$.

We compare the gradient-based methods with LASSO, SCAD and MIO, as shown in Table 4 and Appendix Table 1. We find that if multi-collinearity is high in the true active set, the L_0 penalty often leads to selecting a subset of true covariates, reflecting in a higher precision than recall. When the multi-collinearity is moderate, the gradient-based method can recover the true active set with high probability. In both cases, LASSO and SCAD select many spurious covariates, and MIO has high prediction error at new data.

Experiment 4: Compressive sensing

We further test whether the L_0-based method improves the sparse signal recovery in compressive sensing. The traditional compressive sensing combines the random projection method with L_1-relaxation (Wainwright, 2019). It finds the sparse pattern of the observed signal under a set of orthonormal bases while maintaining the exact reconstruction under random projection by a measurement matrix. Following Ji et al. (2008), we consider $\boldsymbol{\theta}$ as the coordinates of observations in transformed space with length $p = 1024$, where 10

Table 4: Results of the Prostate cancer dataset, with $n = 102, p = 1000, S = 5$, *moderate* multi-collinearity. Reported results are the average of 100 independent trials.

	Precision	Recall	F1	Nonzero	RR	RTE	PVE
Lasso	0.161	0.966	0.273	33.3	0.071	1.351	0.775
SCAD	0.350	0.736	0.461	11.5	0.341	2.704	0.549
MIO	0.458	0.458	0.458	5.00	0.268	2.332	0.609
U2G	0.924	0.916	0.919	4.97	0.038	1.191	0.801
U2G(VI)	0.951	0.946	0.947	4.99	0.031	1.156	0.807

elements are randomly picked as the signal with magnitude ± 1. In this example, most of the entries in the true signal are identically zero, which is called *strong sparsity* (Carvalho et al., 2010). Construing $\mathbf{A} \in \mathbb{R}^{n \times p}$ as a multiplication of the random projection matrix and orthonormal transformation matrix, each row of $\mathbf{A}$ is generated from isotropic Gaussian distribution $\mathcal{N}(0, \mathbf{I}_p)$ and normalized to have the unit norm. We add the Gaussian white noise with a standard deviation σ to the measurements $\boldsymbol{y}$. Aligned with our probabilistic objective, we solve the Lagrangian form of

$$\min_{\boldsymbol{\theta} \in \mathbb{R}^p} \|\boldsymbol{\theta}\|_0, \qquad \text{such that } \mathbf{A}\boldsymbol{\theta} = \boldsymbol{y}.$$

We compare U2G with basis pursuit (BP) (Chen et al., 2001) and Bayesian compressive sensing (BCS) (Ji et al., 2008) in different SNR_d settings by changing the magnitude of σ. For U2G, we use $K = 5$ Monte Carlo samples in the gradient estimation. The numerical results are summarized in Table 5 and Appendix Table 2. As shown in Appendix Figure A.2, in high SNR_d regime, all three methods can reconstruct the sparse signal reasonably well, but the probabilistic best subset method can identify the locations of true signals, while BP and BCS identify excessively large active sets. Consequently, the L_0-regularized

Table 5: Results of the signal reconstruction when SNR_d is low, with $n = 500$, $p = 1000$, $S = 10$, $\sigma = 0.1$. L_0-based method has the best performance in recovering strong sparsity in the signal.

	Precision	Recall	F1	Nonzero	RR	RTE	PVE
BP	0.009	1.000	0.019	1024	0.298	298	0.702
BCS	0.029	1.000	0.057	336.1	2.380	2381	-
U2G	1.000	1.000	1.000	10.00	0.014	15.1	0.985
U2G(VI)	1.000	1.000	1.000	10.00	0.018	18.6	0.981

method has higher predictive precision. This phenomenon is amplified when SNR_d drops. When SNR_d is low, as shown in Figure 4, BP and BCS only recover *weak sparsity* where the signals are dense yet most of the entries are small compared to several large ones. In both high and low SNR_d regimes, the gradient-based methods accurately recover the strong sparsity in signal, improve the predictive accuracy of BP and BCS by several orders of magnitude.

7 Discussion

We propose a probabilistic reformulation to solve the exact best subset selection problem by gradient-based optimization. In order to efficiently solve the L_0-regularized regression in high dimensional settings, a family of unbiased gradient estimators is proposed to approximate the gradient of the objective function. We identify the estimator with non-vanishing signal-to-noise ratio and minimal variance, which theoretically can recover the true sparse pattern in expectation, as also demonstrated in empirical studies.

There are future directions arising naturally from our work. First, our probabilistic reformulation is currently only for binary variables. In applications of latent variable models,

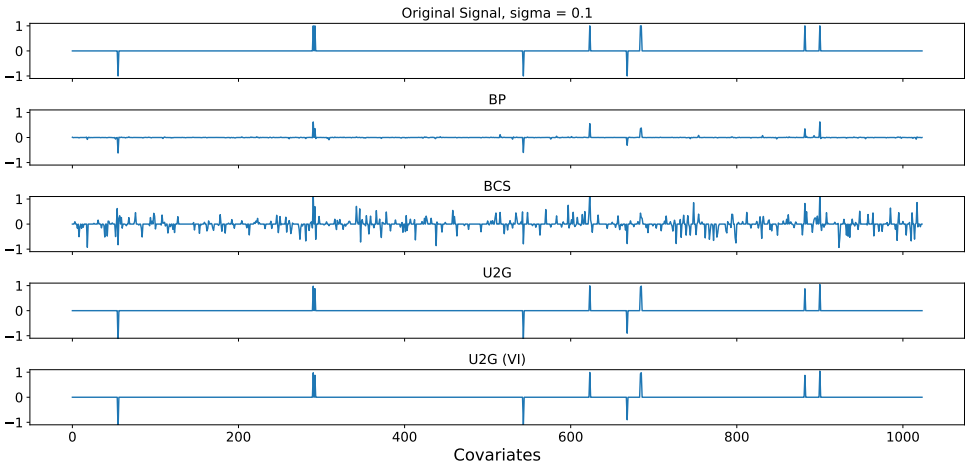


Figure 4: The reconstruction of signals when SNR_d is low, with $n = 500$, $p = 1024$, $\sigma = 0.1$.

the discrete latent variables often consist of more than two values (Jang et al., 2017; Tucker et al., 2017; Yin et al., 2019). Finding the optimal unbiased gradient estimators for these models is an important direction. Second, our theoretical analysis in this paper focuses on the random design matrix with independent covariates, and the convergence under expectation. We leave the study of convergence property of our method when the design matrix is in finite sample setting or has multi-collinearity as future work. It is intriguing to combine L_0 penalty with Tikhonov regularization such that the covariates are selected in groups. Finally, in this paper, the unbiased gradient estimator is applied in the stochastic gradient descent framework. One future direction is to incorporate accelerated gradient and momentum to the update rule. Since the optimization landscape of (4) is locally weakly concave around $\boldsymbol{\pi}$, a fundamental trade-off between instability of SGD and its accelerated versions is important to study (Ho et al., 2020).

SUPPLEMENTARY MATERIAL

Supplementary materials: Detailed proofs of theoretical results in the main paper and additional experimental results. (PDF document)

Code: Python, R and Matlab code to perform the numerical experiments described in the article. The package also contains the Prostate data set used in the article. (ZIP archive)

References

Akaike, H. (1974). A new look at the statistical model identification. In *Selected Papers of Hirotugu Akaike*, pp. 215–222. Springer.

Akaike, H. (1998). Information theory and an extension of the maximum likelihood principle. In *Selected papers of Hirotugu Akaike*, pp. 199–213. Springer.

Barber, R. F. and E. J. Candès (2015). Controlling the false discovery rate via knockoffs. *The Annals of Statistics 43*(5), 2055–2085.

Beale, E., M. Kendall, and D. Mann (1967). The discarding of variables in multivariate analysis. *Biometrika 54*(3-4), 357–366.

Belloni, A. and V. Chernozhukov (2013). Least squares after model selection in high-dimensional sparse models. *Bernoulli 19*(2), 521–547.

Bertsimas, D., A. King, and R. Mazumder (2016). Best subset selection via a modern optimization lens. *The Annals of Statistics 44*(2), 813–852.

Bickel, P., D. Choi, X. Chang, and H. Zhang (2013). Asymptotic normality of maximum likelihood and its variational approximation for stochastic blockmodels. *The Annals of Statistics 41*(4), 1922–1943.

Blei, D. M., A. Kucukelbir, and J. D. McAuliffe (2017). Variational inference: A review for statisticians. *Journal of the American Statistical Association 112*(518), 859–877.

Boluki, S., R. Ardywibowo, S. Z. Dadaneh, M. Zhou, and X. Qian (2020). Learnable Bernoulli dropout for Bayesian deep learning. In *Artificial Intelligence and Statistics*.

Candès, E. J. and Y. Plan (2009). Near-ideal model selection by ℓ_1 minimization. *The Annals of Statistics 37*(5A), 2145–2177.

Carvalho, C. M., N. G. Polson, and J. G. Scott (2010). The horseshoe estimator for sparse signals. *Biometrika 97*(2), 465–480.

Chen, S., S. A. Billings, and W. Luo (1989). Orthogonal least squares methods and their application to non-linear system identification. *International Journal of control 50*(5), 1873–1896.

Chen, S. S., D. L. Donoho, and M. A. Saunders (2001). Atomic decomposition by basis pursuit. *SIAM review 43*(1), 129–159.

Chickering, D. M. (2002). Optimal structure identification with greedy search. *Journal of machine learning research 3*(Nov), 507–554.

Dadaneh, S. Z., S. Boluki, M. Yin, M. Zhou, and X. Qian (2020). Pairwise supervised hashing with Bernoulli variational auto-encoder and self-control gradient estimator. In *Uncertainty in Artificial Intelligence*.

Dicker, L., B. Huang, and X. Lin (2013). Variable selection and estimation with the seamless-ℓ_0 penalty. *Statistica Sinica 23*, 929–962.

Dong, Z., A. Mnih, and G. Tucker (2020, June). DisARM: an antithetic gradient estimator for binary latent variables.

Donoho, D. L., Y. Tsaig, I. Drori, and J.-L. Starck (2012). Sparse solution of underdetermined systems of linear equations by stagewise orthogonal matching pursuit. *IEEE transactions on Information Theory 58*(2), 1094–1121.

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

Fan, J. and J. Lv (2010). A selective overview of variable selection in high dimensional feature space. *Statistica Sinica 20*(1), 101.

Foster, D. P. and E. I. George (1994). The risk inflation criterion for multiple regression. *The Annals of Statistics 22*(4), 1947–1975.

Frank, L. E. and J. H. Friedman (1993). A statistical view of some chemometrics regression tools. *Technometrics 35*(2), 109–135.

Friedman, J., T. Hastie, and R. Tibshirani (2001). *The Elements of Statistical Learning*, Volume 1. Springer series in statistics New York.

Friedman, J., T. Hastie, and R. Tibshirani (2010). Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software 33*(1), 1.

Fu, W. J. (1998). Penalized regressions: the bridge versus the lasso. *Journal of Computational and Graphical Statistics 7*(3), 397–416.

Furnival, G. M. and R. W. Wilson (1974). Regressions by leaps and bounds. *Technometrics 16*(4), 499–511.

Ge, J., X. Li, H. Jiang, M. Wang, T. Zhang, H. Liu, and T. Zhao (2019). *picasso: Pathwise Calibrated Sparse Shooting Algorithm*. R Foundation for Statistical Computing. R package version 1.3.1.

George, E. I. and R. E. McCulloch (1993). Variable selection via Gibbs sampling. *Journal of the American Statistical Association 88*(423), 881–889.

Greenshtein, E. (2006). Best subset selection, persistence in high-dimensional statistical learning and optimization under ℓ_1 constraint. *The Annals of Statistics 34*(5), 2367–2386.

Hastie, T., R. Tibshirani, and R. Tibshirani (2018). *Bestsubset: Tools for best subset selection in regression*. R Foundation for Statistical Computing. R package version 1.0.10.

Hastie, T., R. Tibshirani, and R. J. Tibshirani (2017). Extended comparisons of best subset selection, forward stepwise selection, and the lasso.

Hazimeh, H. and R. Mazumder (2018). Fast best subset selection: Coordinate descent and local combinatorial optimization algorithms.

Ho, N., K. Khamaru, R. Dwivedi, M. J. Wainwright, M. I. Jordan, and B. Yu (2020). Instability, computational efficiency and statistical accuracy.

Jang, E., S. Gu, and B. Poole (2017). Categorical reparameterization with Gumbel-softmax. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net.

Ji, S., Y. Xue, and L. Carin (2008). Bayesian compressive sensing. *IEEE Transactions on signal processing 56*(6), 2346.

Johnson, K. D., D. Lin, L. H. Ungar, D. P. Foster, and R. A. Stine (2015). A risk ratio comparison of ℓ_0 and ℓ_1 penalized regression.

Joseph, A. (2013). Variable selection in high-dimension with random designs and orthogonal matching pursuit. *The Journal of Machine Learning Research 14*(1), 1771–1800.

Linero, A. R. (2018). Bayesian regression trees for high-dimensional prediction and variable selection. *Journal of the American Statistical Association 113*(522), 626–636.

Liu, Y. and Y. Wu (2007). Variable selection via a combination of the ℓ_0 and ℓ_1 penalties. *Journal of Computational and Graphical Statistics 16*(4), 782–798.

Mallat, S. G. and Z. Zhang (1993). Matching pursuits with time-frequency dictionaries. *IEEE Transactions on Signal Processing 41*(12), 3397–3415.

Natarajan, B. K. (1995). Sparse approximate solutions to linear systems. *SIAM Journal on Computing 24*(2), 227–234.

Pati, D., A. Bhattacharya, and Y. Yang (2018). On statistical optimality of variational Bayes. In *International Conference on Artificial Intelligence and Statistics*, pp. 1579–1588.

Pati, Y. C., R. Rezaiifar, and P. S. Krishnaprasad (1993). Orthogonal matching pursuit: Recursive function approximation with applications to wavelet decomposition. In *Proceedings of 27th Asilomar conference on signals, systems and computers*, pp. 40–44. IEEE.

Polson, N. G. and L. Sun (2019). Bayesian l 0-regularized least squares. *Applied Stochastic Models in Business and Industry 35*(3), 717–731.

Ross, S. M. (2014). *Introduction to probability models*. Academic press.

Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics 6*(2), 461–464.

Shen, X., W. Pan, and Y. Zhu (2012). Likelihood-based selection and sharp parameter estimation. *Journal of the American Statistical Association 107*(497), 223–232.

Singh, D., P. G. Febbo, K. Ross, D. G. Jackson, J. Manola, C. Ladd, P. Tamayo, A. A. Renshaw, A. V. D'Amico, J. P. Richie, et al. (2002). Gene expression correlates of clinical prostate cancer behavior. *Cancer cell 1*(2), 203–209.

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

Tucker, G., A. Mnih, C. J. Maddison, J. Lawson, and J. Sohl-Dickstein (2017). Rebar: Low-variance, unbiased gradient estimates for discrete latent variable models. In *Advances in Neural Information Processing Systems*, pp. 2627–2636.

Wainwright, M. J. (2009). Sharp thresholds for high-dimensional and noisy sparsity recovery using ℓ_1-constrained quadratic programming (lasso). *IEEE Transactions on Information Theory 55*(5), 2183–2202.

Wainwright, M. J. (2019). *High-dimensional statistics: A non-asymptotic viewpoint*, Volume 48. Cambridge University Press.

Wang, Y. and D. M. Blei (2018). Frequentist consistency of variational Bayes. *Journal of the American Statistical Association 114*(527), 1147–1161.

Williams, R. J. (1992). Simple statistical gradient-following algorithms for connectionist reinforcement learning. In *Reinforcement Learning*, pp. 5–32. Springer.

Yin, M., Y. Wang, and P. Sarkar (2020). A theoretical case study of structured variational

inference for community detection. In *International Conference on Artificial Intelligence and Statistics*.

Yin, M., Y. Yue, and M. Zhou (2019). ARSM: Augment-reinforce-swap-merge estimator for gradient backpropagation through categorical variables. In *International Conference on Machine Learning*, pp. 7095–7104.

Yin, M. and M. Zhou (2019). ARM: Augment-REINFORCE-merge gradient for stochastic binary networks. In *International Conference on Learning Representations*.

Yue, Y., Y. Tang, M. Yin, and M. Zhou (2020). Discrete action on-policy learning with action-value critic. In *Artificial Intelligence and Statistics*.

Zhang, A. Y. and H. H. Zhou (2017). Theoretical and computational guarantees of mean field variational inference for community detection.

Zhang, C.-H. (2010). Nearly unbiased variable selection under minimax concave penalty. *The Annals of Statistics 38*(2), 894–942.

Zhang, C.-H. and T. Zhang (2012). A general theory of concave regularization for high-dimensional sparse estimation problems. *Statistical Science 27*(4), 576–593.

Zhao, P. and B. Yu (2006). On model selection consistency of lasso. *Journal of Machine learning research 7*(Nov), 2541–2563.

Zhou, M., H. Chen, L. Ren, G. Sapiro, L. Carin, and J. W. Paisley (2009). Non-parametric Bayesian dictionary learning for sparse image representations. In *Advances in neural information processing systems*, pp. 2295–2303.

SUPPLEMENTARY MATERIAL

This supplementary document contains detailed proofs and derivation of theoretical results presented in the main paper "Probabilistic Best Subset Selection via Gradient-Based Optimization", and additional experimental results. In particular, Section A contains a necessary lemma for the proof of convergence properties. Section B contains proofs of the theoretical results presented in the main paper. Section C contains additional simulation results on the semi-synthetic data and compressive sensing.

A Auxiliary Lemmas

The following lemma describes the linear independence between random Gaussian vectors, which is useful when the sample size exceeds the number of covariates. Closely following the proof in Tao (2008), which contains a thorough discussion on the singularity of random matrix ensembles, we have the following result:

Lemma A.1. *Let $X_j \in \mathbb{R}^n$ are i.i.d. random Gaussian vectors with distribution $\mathcal{N}(0, \mathbf{I}_n)$, for $j = 1, \cdots, k$, $k \leq n$. Then $\{X_1, \cdots, X_k\}$ are linearly independent with probability one.*

Proof. Let event $\mathcal{E}$ be the event that $\{X_1, \cdots, X_k\}$ are linearly dependent. Then $\mathcal{E}$ is equivalent to that X_j lies in the span of $X_1, \cdots, X_{j-1}$ for some j. Thus

$$p(\mathcal{E}) \leq \sum_{j=2}^{k} p(X_j \in V_j),$$

where $V_j \coloneqq \text{span}(X_1, \cdots, X_{j-1})$. For each $2 \leq j \leq k$, conditional on vectors $X_1, \cdots, X_{j-1}$, the vector space V_j is fixed, has positive codimension, and thus has measure zero. Since the distribution of X_j is absolutely continuous, and is independent of $X_1, \cdots, X_{j-1}$, we have

$$p(X_j \in V_j | X_1, \cdots, X_{j-1}) = 0$$

for all $(X_1, \cdots, X_{j-1})$. Integrating over $(X_1, \cdots, X_{j-1})$, we have $p(X_j \in V_j) = 0$; therefore

$$p(\mathcal{E}) \leq \sum_{j=2}^{k} p(X_j \in V_j) = 0,$$

which proves that $\{X_1, \cdots, X_k\}$ are linearly independent with probability 1. As a consequence, we obtain the conclusion of the lemma. Q.E.D.

B Proofs

In this section, we provide proofs for theoretical results in the paper.

B.1 Proof of Theorem 1

Proof. On the one hand, we show the optimal solution to problem (3) is in the set of feasible solutions to problem (4). Assuming $(\boldsymbol{\alpha}^*, \boldsymbol{z}^*)$ is the optimal solution to problem (3), setting $p_j = z_j^*,\ j \in [p]$ and $\boldsymbol{\alpha} = \boldsymbol{\alpha}^*$ would be a feasible solution to problem (4) which gives the same object value as what $(\boldsymbol{\alpha}^*, \boldsymbol{z}^*)$ achieves in problem (3).

On the other hand, we show the optimal solution to problem (4) is in the set of feasible solutions to problem (3). Let $h(\boldsymbol{z}) = \frac{1}{n} \|\boldsymbol{y} - \mathbf{X}(\boldsymbol{\alpha} \odot \boldsymbol{z})\|^2 + \lambda \|\boldsymbol{z}\|_0$, $f(\boldsymbol{z}) = \min_{\boldsymbol{\alpha}} h(\boldsymbol{z})$, $g(\boldsymbol{z}) = \arg\min_{\boldsymbol{\alpha}} h(\boldsymbol{z})$, and assume $\boldsymbol{\pi}^*$ is the optimal $\boldsymbol{\pi}$ in problem (4). First, if $p(\boldsymbol{z}|\boldsymbol{\pi}^*)$ is a point mass density $\delta_{\boldsymbol{z}^*}$ with $\pi_j^* = z_j^*, j \in [p]$ and $\boldsymbol{\alpha}^* = g(\boldsymbol{z}^*)$, by setting $\boldsymbol{z} = \boldsymbol{\pi}^*$, $\boldsymbol{\alpha} = \boldsymbol{\alpha}^*$, it would give a feasible solution to problem (3) with the same objective value as problem (4). Second, if $p(\boldsymbol{z}|\boldsymbol{\pi}^*)$ is not a point mass density, assume $\text{supp}[p(\boldsymbol{z}|\boldsymbol{\pi}^*)] = \{\boldsymbol{z}_1, \cdots, \boldsymbol{z}_K\}$. We show by contradiction that all the points in $\text{supp}[p(\boldsymbol{z}|\boldsymbol{\pi}^*)]$ would give the same objective value $f(\boldsymbol{z})$. Otherwise there exist $\boldsymbol{z}_s, \boldsymbol{z}_l \in \text{supp}[p(\boldsymbol{z}|\boldsymbol{\pi}^*)]$ with $f(\boldsymbol{z}_s) < f(\boldsymbol{z}_l)$ and $f(\boldsymbol{z}_s) \leq f(\boldsymbol{z}_k)$ for $k \neq s, l$. By setting $\hat{\pi}_j = z_{ij}$, we would have $\mathbb{E}_{\boldsymbol{z} \sim p(\boldsymbol{z}|\hat{\boldsymbol{\pi}})} f(\boldsymbol{z}) < \mathbb{E}_{\boldsymbol{z} \sim p(\boldsymbol{z}|\boldsymbol{\pi}^*)} f(\boldsymbol{z})$ which contradicts with the assumption that $\boldsymbol{\pi}^*$ is optimal. Therefore, we have $f(\boldsymbol{z}_1) = f(\boldsymbol{z}_2) =$

$\cdots = f(\boldsymbol{z}_K)$. Hence all points in $\{\boldsymbol{z}_k, g(\boldsymbol{z}_k)\}_{k\in[K]}$ are feasible solutions to problem (3) which give the same objective value as what the optimal solution gives in problem (4).

In summary, we show problem (3) and (4) have the same global optima and the same objective value at such points, so they are equivalent problems. Q.E.D.

B.2 Proof of Proposition 1

Proof. Without loss of generality, we assume that $f(1), f(0) > 0$, $\pi = \sigma(\phi) \geq 1/2$, and let $\Delta = f(1) - f(0)$.

For the first inequality, direct calculation shows that

$$\begin{aligned}\mathrm{var}[g_{\mathrm{ARM}}] - \mathrm{var}[g_R] =& \mathbb{E}_u[g^2_{\mathrm{ARM}}] - \mathbb{E}_u[g^2_R] \\ =& s_1 f(1)^2 + s_2 f(0)^2 + s_3 f(1)f(0) \\ =& (s_1 + s_2 + s_3) f(0)^2 + s_1\Delta^2 + (2s_1 + s_3) f(0)\Delta \\ \leq& (s_1 + s_2 + s_3) f(0)^2 + (3s_1 + s_3) f(0)^2\end{aligned}$$

where $s_1 = -\frac{5}{3}\pi^3 + 3\pi^2 - \frac{3}{2}\pi + \frac{1}{6}$, $s_2 = \frac{1}{3}\pi^3 - \frac{1}{2}\pi + \frac{1}{6}$, $s_3 = \frac{4}{3}\pi^3 - 2\pi^2 + \pi - \frac{1}{3}$.

Re-organizing the coefficients, we have

$$\mathrm{var}[g_{\mathrm{ARM}}] - \mathrm{var}[g_R] \leq -\big[\pi(1 - \frac{\pi}{6}) + \frac{21\pi + 1}{6}(1 - \pi)\big] f(0)^2 \leq 0.$$

For the second inequality, we find that

$$\mathrm{var}[g_{\mathrm{U2G}}] - \mathrm{var}[g_{\mathrm{ARM}}] = \mathbb{E}_u[g^2_{\mathrm{U2G}}] - \mathbb{E}_u[g^2_{\mathrm{ARM}}] = -\frac{(1-\pi)^3}{6}(f(1) - f(0))^2 \leq 0.$$

As a consequence, we obtain the conclusion of the proposition. Q.E.D.

B.3 Proof of Proposition 2

Proof. We consider a constrained optimization problem

$$\min_g \int_0^1 g^2(u)du, \quad \text{subject to} \quad \mathbb{E}[g(u)] = \mu,$$

where $\mu = \pi(1-\pi)(f_1 - f_0)$. For simplicity, we omit the conditional notation on π if it is clear. The integration can be decomposed into three intervals $[0, 1-\pi], (1-\pi, \pi], (\pi, 1]$, so we can rewrite Eq. (8) into a piece-wise function.

$$g(u) = \begin{cases} g_1(u) := a(u)f_1 + b(u)f_0, & u \in [0, 1-\pi] \\ g_2(u) := a(u)f_1 + b(u)f_1, & u \in (1-\pi, \pi] \\ g_3(u) := a(u)f_0 + b(u)f_1, & u \in (\pi, 1] \end{cases}$$

And we would like to minimize

$$\mathbb{E}[g^2(u)] = \mathbb{E}_{u\in[0,1-\pi]}[g_1^2(u)] + \mathbb{E}_{u\in[1-\pi,\pi]}[g_2^2(u)] + \mathbb{E}_{u\in[\pi,1]}[g_3^2(u)].$$

If there exists π and a positive measure subset $\mathcal{S}_\pi \subset (1-\pi, \pi]$ where $|a(u;\pi) + b(u;\pi)| \geq \epsilon_\pi > 0$ and $f_1 \neq 0$, then

$$\mathbb{E}[g^2(u;\pi)] \geq \mathbb{E}[g_2^2(u;\pi)] \geq \epsilon_\pi^2 |\mathcal{S}_\pi| f_1^2.$$

This means that

$$\begin{aligned} \lim_{\Delta\to 0} \text{var}[g(u;\pi)] &= \lim_{\Delta\to 0} \mathbb{E}[g^2(u;\pi)] - \mu^2 \\ &= \lim_{\Delta\to 0} \mathbb{E}[g^2(u;\pi)] \geq \epsilon_\pi^2 |\mathcal{S}_\pi| f_1^2 > 0, \end{aligned}$$

which contradicts the condition (13). Therefore if the estimator has positive SNR for $\pi \in (0,1)$, it has to satisfy $a(u) + b(u) = 0$ almost surely (a.s.) in $(1-\pi, \pi]$ or has $f_1 = 0$, where in both cases $g(u) = 0$ a.s. for $u \in (1-\pi, \pi]$. Assume $g(u) = 0$ for $u \in (1-\pi, \pi]$ a.s., we have

$$\begin{aligned} \text{var}[g(u;\pi)] =& (1-\pi)[\int_0^{1-\pi} \frac{1}{1-\pi} g^2(u)du + \int_\pi^1 \frac{1}{1-\pi} g^2(u)du] - \mu^2 \\ \geq& \frac{1}{1-\pi}\Big\{[\int_0^{1-\pi} g(u)du]^2 + [\int_\pi^1 g(u)du]^2\Big\} - \mu^2 \end{aligned}$$

$$
\begin{aligned}
&= \frac{1}{1-\pi}(2s^2 - 2\mu s + \mu^2) - \mu^2 \\
&\geq \frac{2\pi - 1}{2(1-\pi)}\mu^2,
\end{aligned}
$$

where $s = \int_0^{1-\pi} g(u)du$; both inequalities are equalities if and only if $g(u) = \frac{\mu}{2(1-\pi)}$ for $u \in [0, 1-\pi] \cup (\pi, 1]$. Together with the premise $g(u) = 0$ for $u \in (1-\pi, \pi]$ a.s., we get the U2G estimator. The same argument holds for $\pi < 0.5$ because of the symmetry. Q.E.D.

B.4 Proof of Lemma 1

Proof. In order to ease the presentation, we denote $\boldsymbol{z} = \mathbf{1}_{[\boldsymbol{u} > 1-\sigma(\boldsymbol{\phi})]}, \tilde{\boldsymbol{z}} = \mathbf{1}_{[\boldsymbol{u} < \sigma(\boldsymbol{\phi})]}$, $\pi_j = \sigma(\phi_j)\ \forall j \in \{1, \cdots, p\}$. Then we have $\boldsymbol{g}_{\text{U2G}}(\boldsymbol{u}; \sigma(\boldsymbol{\phi})) = \frac{f(\boldsymbol{z}) - f(\tilde{\boldsymbol{z}})}{2}\sigma(|\boldsymbol{\phi}|) \odot (\mathbf{1}_{[\boldsymbol{u} > 1-\sigma(\boldsymbol{\phi})]} - \mathbf{1}_{[\boldsymbol{u} < \sigma(\boldsymbol{\phi})]})$. Construct a sequence of binary code $\boldsymbol{z}^0 = \boldsymbol{z}, \boldsymbol{z}^1, \cdots, \boldsymbol{z}^p = \tilde{\boldsymbol{z}}$ by flipping one dimension of $\boldsymbol{z}$ to the value in $\tilde{\boldsymbol{z}}$ at a time, i.e., $\boldsymbol{z}^i = (\tilde{z}_1, \cdots, \tilde{z}_i, z_{i+1}, \cdots, z_p)'$. Hence $f_{\mathbf{X},\boldsymbol{y}}(\boldsymbol{z}) - f_{\mathbf{X},\boldsymbol{y}}(\tilde{\boldsymbol{z}}) = \sum_{i=1}^p (f_{\mathbf{X},\boldsymbol{y}}(\boldsymbol{z}^{i-1}) - f_{\mathbf{X},\boldsymbol{y}}(\boldsymbol{z}^i))$. We prove the statement for the gradient vector element-wisely. Consider the j^{th} dimension of the gradient vector

$$
\mathbb{E}_{\boldsymbol{u}}[g(\boldsymbol{u})_j] = \frac{\sigma(|\phi_j|)}{2}\mathbb{E}_{\boldsymbol{u}} \sum_{i=1}^p (f_{\mathbf{X},\boldsymbol{y}}(\boldsymbol{z}^{i-1}) - f_{\mathbf{X},\boldsymbol{y}}(\boldsymbol{z}^i))(\mathbf{1}_{[u_j > \sigma(-\phi_j)]} - \mathbf{1}_{[u_j < \sigma(\phi_j)]}) \quad \text{(B.1)}
$$

Note that $\boldsymbol{z}^{i-1}$ and $\boldsymbol{z}^i$ only differ on the i^{th} dimension, and different dimensions of $\boldsymbol{u}$ are independent. Consider the i^{th} element of the summation in Eq.(B.1) and W.L.O.G. we first assume the logit $\phi_i \geq 0$. For $i \neq j$, due to the symmetry of the sigmoid function, we have

$$
\begin{aligned}
&\mathbb{E}_{\boldsymbol{u}} \frac{\sigma(|\phi_j|)}{2}(f_{\mathbf{X},\boldsymbol{y}}(\boldsymbol{z}^{i-1}) - f_{\mathbf{X},\boldsymbol{y}}(\boldsymbol{z}^i))(\mathbf{1}_{[u_j > \sigma(-\phi_j)]} - \mathbf{1}_{[u_j < \sigma(\phi_j)]}) \\
&= \frac{\sigma(|\phi_j|)}{2}\mathbb{E}_{\boldsymbol{u}_{-i}}\left[\mathbb{E}_{u_i}[f_{\mathbf{X},\boldsymbol{y}}(\boldsymbol{z}^{i-1}) - f_{\mathbf{X},\boldsymbol{y}}(\boldsymbol{z}^i))(\mathbf{1}_{[u_j > \sigma(-\phi_j)]} - \mathbf{1}_{[u_j < \sigma(\phi_j)]})|\boldsymbol{u}_{-i}]\right] \\
&= \frac{\sigma(|\phi_j|)}{2}\mathbb{E}_{\boldsymbol{u}_{-i}}\left[(\mathbf{1}_{[u_j > \sigma(-\phi_j)]} - \mathbf{1}_{[u_j < \sigma(\phi_j)]})\mathbb{E}_{u_i}[f_{\mathbf{X},\boldsymbol{y}}(\boldsymbol{z}^{i-1}) - f_{\mathbf{X},\boldsymbol{y}}(\boldsymbol{z}^i))|\boldsymbol{u}_{-i}]\right] \\
&= \mathbb{E}_{\boldsymbol{u}_{-i}}\Big[(\mathbf{1}_{[u_j > \sigma(-\phi_j)]} - \mathbf{1}_{[u_j < \sigma(\phi_j)]})\Big(\int_0^{\sigma(-\phi_i)} \big(f_{\mathbf{X},\boldsymbol{y}}(\boldsymbol{z}^{i-1}|z_i = 0) - f_{\mathbf{X},\boldsymbol{y}}(\boldsymbol{z}^i|z_i = 1)\big)du_i \\
&\quad + \int_{\sigma(\phi_i)}^1 \big(f_{\mathbf{X},\boldsymbol{y}}(\boldsymbol{z}^{i-1}|z_i = 1) - f_{\mathbf{X},\boldsymbol{y}}(\boldsymbol{z}^i|z_i = 0)\big)du_i\Big)|\boldsymbol{u}_{-i}\Big]\frac{\sigma(|\phi_j|)}{2}
\end{aligned}
$$

$$= \mathbb{E}_{\boldsymbol{u}_{-i}}\Big[(\mathbf{1}_{[u_j>\sigma(-\phi_j)]} - \mathbf{1}_{[u_j<\sigma(\phi_j)]})\Big(\big(f_{\mathbf{X},\boldsymbol{y}}(\boldsymbol{z}^{i-1}|z_i=0) - f_{\mathbf{X},\boldsymbol{y}}(\boldsymbol{z}^{i}|z_i=1)\big)(1-\sigma(\phi_i))$$
$$+ \big(f_{\mathbf{X},\boldsymbol{y}}(\boldsymbol{z}^{i-1}|z_i=1) - f_{\mathbf{X},\boldsymbol{y}}(\boldsymbol{z}^{i}|z_i=0)\big)(1-\sigma(\phi_i))\Big)|\boldsymbol{u}_{-i}\Big]\frac{\sigma(|\phi_j|)}{2}$$
$$= 0.$$

Whereas for $i = j$, we have

$$\frac{\sigma(|\phi_j|)}{2}\mathbb{E}_{\boldsymbol{u}}(f_{\mathbf{X},\boldsymbol{y}}(\boldsymbol{z}^{j-1}) - f_{\mathbf{X},\boldsymbol{y}}(\boldsymbol{z}^{j}))(\mathbf{1}_{[u_j>\sigma(-\phi_j)]} - \mathbf{1}_{[u_j<\sigma(\phi_j)]})$$
$$= \frac{\sigma(\phi_j)}{2}\mathbb{E}_{\boldsymbol{u}_{-j}}\left[\mathbb{E}_{u_j}[(f_{\mathbf{X},\boldsymbol{y}}(\boldsymbol{z}^{j-1}) - f_{\mathbf{X},\boldsymbol{y}}(\boldsymbol{z}^{j}))(\mathbf{1}_{[u_j>\sigma(-\phi_j)]} - \mathbf{1}_{[u_j<\sigma(\phi_j)]})|\boldsymbol{u}_{-j}]\right]$$
$$= \frac{\sigma(\phi_j)}{2}\mathbb{E}_{\boldsymbol{u}_{-j}}\left[\left(\int_0^{\sigma(-\phi_j)}(f_{\mathbf{X},\boldsymbol{y}}(\boldsymbol{z}^{j-1}|z_j=0) - f_{\mathbf{X},\boldsymbol{y}}(\boldsymbol{z}^{j}|z_j=1))(-1)du_j\right.\right.$$
$$\left.\left. + \int_{\sigma(\phi_j)}^1 (f_{\mathbf{X},\boldsymbol{y}}(\boldsymbol{z}^{j-1}|z_j=1) - f_{\mathbf{X},\boldsymbol{y}}(\boldsymbol{z}^{j}|z_j=0))du_j\right)\Big|\boldsymbol{u}_{-j}\right]$$
$$= \mathbb{E}_{\boldsymbol{u}_{-j}}\Big[\sigma(\phi_j)(1-\sigma(\phi_j))[f_{\mathbf{X},\boldsymbol{y}}(\boldsymbol{z}^{j-1}|z_j=1) - f_{\mathbf{X},\boldsymbol{y}}(\boldsymbol{z}^{j}|z_j=0)]\Big]$$
$$= \pi_j(1-\pi_j)\mathbb{E}_{\boldsymbol{u}}[\Delta_{\boldsymbol{z},j}f].$$

The same derivation holds true when the logit $\phi_i \le 0$. Hence for each dimension there is only one non-zero element in the summation of Eq.(B.1). Rewriting the result in vector form proves the lemma. Q.E.D.

B.5 Proof of Lemma 2

Proof. Based on Lemma 1, it suffices to compute the expectation of $\mathbb{E}_{\mathbf{X},\boldsymbol{y},\boldsymbol{u}}[\Delta_{\boldsymbol{z}}f]$ to obtain the conclusion of Lemma 2, where $\Delta_{\boldsymbol{z}}f = (\Delta_{\boldsymbol{z},1}f, \cdots, \Delta_{\boldsymbol{z},p}f)$. For any $k \in [p]$, direct application of conditional expectation formulations leads to

$$\mathbb{E}_{\mathbf{X},\boldsymbol{y},\boldsymbol{u}}[\Delta_{\boldsymbol{z},k}f] = \mathbb{E}[\Delta_{\boldsymbol{z},k}f\Big|\|\boldsymbol{z}\|_0 < n-1]p(\|\boldsymbol{z}\|_0 < n-1) \tag{B.2}$$
$$+ \mathbb{E}[\Delta_{\boldsymbol{z},k}f\Big|\|\boldsymbol{z}\|_0 = n-1]p(\|\boldsymbol{z}\|_0 = n-1) + \mathbb{E}[\Delta_{\boldsymbol{z},k}f\Big|\|\boldsymbol{z}\|_0 \ge n]p(\|\boldsymbol{z}\|_0 \ge n).$$

Conditioned on the event $\|\boldsymbol{z}\|_0 < n-1$, we denote projection matrix $P_{\boldsymbol{z}} = \mathbf{X}_{\boldsymbol{z}}(\mathbf{X}_{\boldsymbol{z}}^\top \mathbf{X}_{\boldsymbol{z}})^{-1}\mathbf{X}_{\boldsymbol{z}}^\top$ for any $\boldsymbol{z}$ and the OLS estimator $\hat{\boldsymbol{\alpha}}_{\boldsymbol{z}} = (\mathbf{X}_{\boldsymbol{z}}^\top \mathbf{X}_{\boldsymbol{z}})^{-1}\mathbf{X}_{\boldsymbol{z}}^\top \boldsymbol{y}$. Under that event, given the definition of $\Delta_{\boldsymbol{z},k} f$ we obtain that

$$\begin{aligned}\Delta_{\boldsymbol{z},k} f &= \lambda + \frac{1}{n}\|\boldsymbol{y} - \mathbf{X}_{\tilde{\boldsymbol{z}}}\hat{\boldsymbol{\alpha}}_{\tilde{\boldsymbol{z}}}\|_2^2 - \frac{1}{n}\|\boldsymbol{y} - \mathbf{X}_{\boldsymbol{z}}\hat{\boldsymbol{\alpha}}_{\boldsymbol{z}}\|_2^2 \\ &= \lambda + \underbrace{\frac{1}{n}\|\boldsymbol{y} - \mathbf{X}_{\tilde{\boldsymbol{z}}}\hat{\boldsymbol{\alpha}}_{\tilde{\boldsymbol{z}}}\|_2^2 - \frac{1}{n}\|\boldsymbol{y} - \mathbf{X}_{\tilde{\boldsymbol{z}}}\boldsymbol{\beta}^*_{\tilde{\boldsymbol{z}}}\|_2^2}_{:=T_1} \\ &\quad + \underbrace{\frac{1}{n}\|\boldsymbol{y} - \mathbf{X}_{\tilde{\boldsymbol{z}}}\boldsymbol{\beta}^*_{\tilde{\boldsymbol{z}}}\|_2^2 - \frac{1}{n}\|\boldsymbol{y} - \mathbf{X}_{\boldsymbol{z}}\boldsymbol{\beta}^*_{\boldsymbol{z}}\|_2^2}_{:=T_2} + \underbrace{\frac{1}{n}\|\boldsymbol{y} - \mathbf{X}_{\boldsymbol{z}}\boldsymbol{\beta}^*_{\boldsymbol{z}}\|_2^2 - \frac{1}{n}\|\boldsymbol{y} - \mathbf{X}_{\boldsymbol{z}}\hat{\boldsymbol{\alpha}}_{\boldsymbol{z}}\|_2^2}_{:=T_3}, \end{aligned} \tag{B.3}$$

where $\tilde{\boldsymbol{z}} \in \{0,1\}^p$ is such that $\boldsymbol{z}$ and $\tilde{\boldsymbol{z}}$ only differ on dimension k, i.e., $\boldsymbol{z}_k = 0, \tilde{\boldsymbol{z}}_k = 1, \boldsymbol{z}_j = \tilde{\boldsymbol{z}}_j, \forall j \neq k$. Note that, $\|\tilde{\boldsymbol{z}}\|_0 = \|\boldsymbol{z}\|_0 + 1 \leq n-1$ when $\|\boldsymbol{z}\|_0 < n-1$; therefore, the OLS estimator $\hat{\boldsymbol{\alpha}}_{\tilde{\boldsymbol{z}}} = (\mathbf{X}_{\tilde{\boldsymbol{z}}}^\top \mathbf{X}_{\tilde{\boldsymbol{z}}})^{-1}\mathbf{X}_{\tilde{\boldsymbol{z}}}^\top \boldsymbol{y}$ is valid. Regarding term T_1 in equation (B.3), direct computation shows that

$$T_1 = -\frac{1}{n}(\mathbf{X}_{-\tilde{\boldsymbol{z}}}\boldsymbol{\beta}^*_{-\tilde{\boldsymbol{z}}} + \boldsymbol{\epsilon})^\top P_{\tilde{\boldsymbol{z}}}(\mathbf{X}_{-\tilde{\boldsymbol{z}}}\boldsymbol{\beta}^*_{-\tilde{\boldsymbol{z}}} + \boldsymbol{\epsilon}) = -\frac{1}{n}\left\|P_{\tilde{\boldsymbol{z}}}(\mathbf{X}_{-\tilde{\boldsymbol{z}}}\boldsymbol{\beta}^*_{-\tilde{\boldsymbol{z}}} + \boldsymbol{\epsilon})\right\|_2^2. \tag{B.4}$$

By Lemma A.1, conditioned on $\boldsymbol{u}$, the rank of $X_{\tilde{\boldsymbol{z}}}$ equals $\|\tilde{\boldsymbol{z}}\|_0$ with probability one. Taking the expectation of T_1 with respect to $\mathbf{X}, \boldsymbol{y}$, a key observation is that for any $X_{\tilde{\boldsymbol{z}}}$ with full column rank,

$$\mathbb{E}_{\mathbf{X},\boldsymbol{y}}[T_1 | X_{\tilde{\boldsymbol{z}}}, \|\boldsymbol{z}\|_0 < n-1] = -\frac{(\sigma^2 + \left\|\boldsymbol{\beta}^*_{-\tilde{\boldsymbol{z}}}\right\|_2^2)\|\tilde{\boldsymbol{z}}\|_0}{n}. \tag{B.5}$$

Therefore, we obtain that

$$\mathbb{E}_{\mathbf{X},\boldsymbol{y}}[T_1 \Big| \|\boldsymbol{z}\|_0 < n-1] = \mathbb{E}_{\mathbf{X},\boldsymbol{y}}[\mathbb{E}[T_1 | X_{\tilde{\boldsymbol{z}}}, \|\boldsymbol{z}\|_0 < n-1]] = -\frac{(\sigma^2 + \left\|\boldsymbol{\beta}^*_{-\tilde{\boldsymbol{z}}}\right\|_2^2)\|\tilde{\boldsymbol{z}}\|_0}{n}. \tag{B.6}$$

The above result leads to

$$\mathbb{E}_{\mathbf{X},\boldsymbol{y},\boldsymbol{u}}[T_1 \Big| \|\boldsymbol{z}\|_0 < n-1] = -\mathbb{E}_{\boldsymbol{u}}\left[\frac{(\sigma^2 + \left\|\boldsymbol{\beta}^*_{-\tilde{\boldsymbol{z}}}\right\|_2^2)\|\tilde{\boldsymbol{z}}\|_0}{n} \mid \|\boldsymbol{z}\|_0 < n-1\right].$$

For the term T_3 in equation (B.3), similar argument proves that

$$\mathbb{E}_{\mathbf{X},\boldsymbol{y},\boldsymbol{u}}[T_3 \Big| \|\boldsymbol{z}\|_0 < n-1] = \mathbb{E}_{\boldsymbol{u}}\left[\frac{(\sigma^2 + \left\|\boldsymbol{\beta}^*_{-\boldsymbol{z}}\right\|_2^2)\,\|\boldsymbol{z}\|_0}{n} \mid \|\boldsymbol{z}\|_0 < n-1\right].$$

For the term T_2 in equation (B.3), direct calculation shows that

$$T_2 = -(\beta_k^*)^2 \frac{1}{n}\|X_k\|^2 - \frac{2}{n}\beta_k^* X_k^\top R_{\tilde{\boldsymbol{z}}},$$

where $R_{\tilde{\boldsymbol{z}}} = \mathbf{X}_{-\tilde{\boldsymbol{z}}}\boldsymbol{\beta}^*_{-\tilde{\boldsymbol{z}}} + \boldsymbol{\epsilon}$. Therefore, $\mathbb{E}_{\mathbf{X},\boldsymbol{y},\boldsymbol{u}}[T_2 \Big| \|\boldsymbol{z}\|_0 < n-1] = -(\beta_k^*)^2$.

Putting the above results together, we find that

$$\mathbb{E}_{\mathbf{X},\boldsymbol{y},\boldsymbol{u}}[\Delta_{\boldsymbol{z},k} f \Big| \|\boldsymbol{z}\|_0 < n-1] = \lambda - \frac{1}{n}\Big((\beta_k^*)^2(n - \mathbb{E}_{\boldsymbol{u}}[\|\boldsymbol{z}\|_0 \mid \|\boldsymbol{z}\|_0 < n-1] - 1) + \sigma^2 + \mathbb{E}_{\boldsymbol{u}}[\left\|\boldsymbol{\beta}^*_{-\boldsymbol{z}}\right\|_2^2 \mid \|\boldsymbol{z}\|_0 < n-1]\Big). \tag{B.7}$$

Conditioned on the event $\|\boldsymbol{z}\|_0 = n-1$, recall that

$$\Delta_{\boldsymbol{z},k} f = \lambda - \frac{1}{n}\left\|\boldsymbol{y} - \mathbf{X}_{\boldsymbol{z}}\hat{\boldsymbol{\alpha}}_{\boldsymbol{z}}\right\|_2^2 = \lambda - \frac{1}{n}\left\|(I_n - P_z)W_{\boldsymbol{z}}\right\|_2^2,$$

where $W_z = \mathbf{X}_{-\boldsymbol{z}}\boldsymbol{\beta}^*_{-\boldsymbol{z}} + \boldsymbol{\epsilon}$. Conditioned on $\boldsymbol{u}$, $W_z \sim \mathcal{N}(0, (\sigma^2 + \|\boldsymbol{\beta}^*_{-\boldsymbol{z}}\|_2^2)I_n)$. Therefore, we obtain that

$$\mathbb{E}_{\mathbf{X},\boldsymbol{y}}\left[\frac{1}{n}\left\|(I_n - P_z)W_{\boldsymbol{z}}\right\|_2^2 \mid \|\boldsymbol{z}\|_0 = n-1\right] = \frac{\sigma^2 + \|\boldsymbol{\beta}^*_{-\boldsymbol{z}}\|_2^2}{n}.$$

The above inequality shows that

$$\mathbb{E}_{\mathbf{X},\boldsymbol{y},\boldsymbol{u}}[\Delta_{\boldsymbol{z},k} f \Big| \|\boldsymbol{z}\|_0 = n-1] = \lambda - \frac{\sigma^2 + \mathbb{E}_{\boldsymbol{u}}[\|\boldsymbol{\beta}^*_{-\boldsymbol{z}}\|_2^2 \mid \|\boldsymbol{z}\|_0 = n-1]}{n}. \tag{B.8}$$

Finally, we compute $\mathbb{E}_{\mathbf{X},\boldsymbol{y},\boldsymbol{u}}[\Delta_{\boldsymbol{z},k} f \Big| \|\boldsymbol{z}\|_0 \geq n]$. Under the setting $\|\boldsymbol{z}\|_0 \geq n$, we have $\|\tilde{\boldsymbol{z}}\|_0 \geq n+1$. By Lemma A.1, conditioned on $\boldsymbol{u}$, with probability one, we have $\min_{\boldsymbol{\alpha}} \frac{1}{n}\left\|\boldsymbol{y} - \mathbf{X}_{\boldsymbol{z}}\boldsymbol{\alpha}\right\|_2^2 = 0$ and $\min_{\boldsymbol{\alpha}} \frac{1}{n}\left\|\boldsymbol{y} - \mathbf{X}_{\tilde{\boldsymbol{z}}}\boldsymbol{\alpha}\right\|_2^2 = 0$. Therefore, conditioned on $\boldsymbol{u}$ and $\|\boldsymbol{z}\|_0 \geq n$, with probability one, we obtain that

$$\Delta_{\boldsymbol{z},k} f = \lambda.$$

It directly leads to

$$\mathbb{E}_{\mathbf{X},\boldsymbol{y},\boldsymbol{u}}[\Delta_{\boldsymbol{z},k} f \Big| \|\boldsymbol{z}\|_0 \geq n] = \lambda. \tag{B.9}$$

Plugging the results of equations (B.7), (B.8), and (B.9) into the equation (B.2), we obtain the conclusion of the lemma. Q.E.D.

B.6 Proof of Proposition 3

Proof. We consider cases whether the covariates are in the true active set separately.

Case 1 - $\{j : \beta_j^* = 0\}$: when $\|\boldsymbol{z}\|_0 < n$, by Lemma 2, we have

$$\begin{aligned}\mathbb{E}_{\mathbf{X},\boldsymbol{y},\boldsymbol{u}}[g_{U2G}(\boldsymbol{u}; \sigma(\boldsymbol{\phi}))_j] &= \left(\lambda - \frac{\sigma^2 + \mathbb{E}_{\boldsymbol{u}}[\left\|\boldsymbol{\beta}^*_{-\boldsymbol{z}}\right\|_2^2 \Big| \|\boldsymbol{z}\|_0 < n]}{n}\right)\pi_j(1-\pi_j) \\ &\geq \left(\lambda - \frac{\sigma^2 + \|\boldsymbol{\beta}^*\|_2^2}{n}\right)\pi_j(1-\pi_j),\end{aligned}$$

where the inequality is due to $\|\boldsymbol{\beta}^*_{-\boldsymbol{z}}\|_2^2 \leq \|\boldsymbol{\beta}^*\|_2^2$ for all $\boldsymbol{z}$. Therefore, if $\lambda > (\sigma^2 + \|\boldsymbol{\beta}^*\|_2^2)/n$, we have $\mathbb{E}_{\mathbf{X},\boldsymbol{y},\boldsymbol{u}}[g_{\mathrm{U2G}}(\boldsymbol{u}; \sigma(\boldsymbol{\phi}))_j] > 0$ for all $j \notin \mathcal{A}$.

Case 2 - $\{j : \beta_j^* \neq 0\}$: by Hoeffding's inequality for sub-Gaussian random variables, and by the assumption $p \leq 2\eta^2(n-1)^2/\log(n)$, we have

$$p(\|\boldsymbol{z}\|_0 \geq n-1) \leq \exp\left(-\frac{2(n-1-\sum_{k=1}^p \pi_j)^2}{p}\right) \leq \frac{1}{n}.$$

Furthermore, simple algebra shows that

$$\mathbb{E}\Big[\|\boldsymbol{z}\|_0 \Big| \|\boldsymbol{z}\|_0 < n-1\Big] \leq \mathbb{E}[\|\boldsymbol{z}\|_0] = \sum_{k=1}^p \pi_k.$$

Collecting the previous results and using the result of Lemma 2, we find that

$$\mathbb{E}_{\mathbf{X},\boldsymbol{y},\boldsymbol{u}}[g_{U2G}(\boldsymbol{u}; \sigma(\boldsymbol{\phi}))_j] \leq \left(\lambda - \frac{(\beta_j^*)^2(n-1)}{n}(1-\frac{1}{n}) + \frac{(\beta_j^*)^2(\sum_{k=1}^p \pi_k)}{n}\right)\pi_j(1-\pi_j)$$

$$\leq \left(\lambda - (\beta_j^*)^2\left[\frac{(n-1)^2}{n^2} - \frac{\sum_{k=1}^p \pi_k}{n}\right]\right)\pi_j(1-\pi_j)$$
$$\leq \left(\lambda - (\beta_j^*)^2\frac{n-1}{n}(\eta - \frac{1}{n})\right)\pi_j(1-\pi_j). \quad \text{(B.10)}$$

Therefore, as long as $\lambda < \frac{n-1}{n}(\eta - \frac{1}{n})\min_{k\in\mathcal{A}}(\beta_k^*)^2$, we have $\mathbb{E}_{\mathbf{X},\boldsymbol{y},\boldsymbol{u}}[g_{U2G}(\boldsymbol{u};\sigma(\boldsymbol{\phi}))_j] < 0$ for all $j \in \mathcal{A}$.

Combining the two cases, by setting

$$\lambda \in \left(\frac{\|\boldsymbol{\beta}^*\|_2^2 + \sigma^2}{n}, \frac{n-1}{n}(\eta - \frac{1}{n})\min_{k\in\mathcal{A}}(\beta_k^*)^2\right) := \mathcal{I}, \quad \text{(B.11)}$$

the proposition is proved. Q.E.D.

B.7 Proof of Theorem 2

Proof. For the simplicity of the presentation, we denote

$$G(\boldsymbol{\phi}) := \mathbb{E}_{\mathbf{X},\boldsymbol{y},\boldsymbol{u}}[g_{U2G}(\boldsymbol{u};\sigma(\boldsymbol{\phi}))]/[\boldsymbol{\pi}(1-\boldsymbol{\pi})],$$

where $\boldsymbol{\pi} = \sigma(\boldsymbol{\phi})$ and the division is element-wise. Furthermore, we denote

$$\eta = \min\left\{\sqrt{\frac{p\log(n)}{2(n-1)^2}}, \frac{\|\boldsymbol{\beta}^*\|_2^2 + \sigma^2}{(n-1)\min_{k\in\mathcal{A}}(\beta_k^*)^2}\right\}.$$

We now state our proof with two parts.

(a) Since $\lambda \in \mathcal{I}$ and $\sum_{j=1}^p \sigma(\phi_j^{(0)}) \leq \varpi - S$, we have $G(\phi_j^{(t)}) < 0$ for all $j \in \mathcal{A}$. It shows that the updates $\phi_j^{(t+1)}$ are monotonically increasing, i.e., $\phi_j^{(t+1)} > \phi_j^{(t)}$ for any $t \geq 0$ and $j \in \mathcal{A}$. For any $t \geq 0$, we obtain that

$$\phi_j^{(t+1)} = \phi_j^{(0)} - \rho\left(\sum_{i=0}^t \mathbb{E}_{\mathbf{X},\boldsymbol{y},\boldsymbol{u}}[g_{U2G}(\boldsymbol{u};\sigma(\boldsymbol{\phi}^{(i)}))_j]\right).$$

We now study the lower bound of t such that $\phi_j^{(t+1)} \geq 0$ for the first time. In order to study that, we assume that $\phi_j^{(i)} < 0$ for all $0 \leq i \leq t$. Based on the proof of Proposition 3, for any

$0 \leq i \leq t$, we have

$$\begin{aligned}\mathbb{E}_{\mathbf{X},\boldsymbol{y},\boldsymbol{u}}[g_{U2G}(\boldsymbol{u};\sigma(\boldsymbol{\phi}^{(i)}))_j] \leq & \left(\lambda - (\beta_j^*)^2\frac{n-1}{n}(\eta - \frac{1}{n})\right)\sigma(\phi_j^{(i)})(1-\sigma(\phi_j^{(i)})) \\ \leq & \left(\lambda - (\beta_j^*)^2\frac{n-1}{n}(\eta - \frac{1}{n})\right)\sigma(\phi_j^{(0)})(1-\sigma(\phi_j^{(0)})),\end{aligned}$$

where the second inequality is due to the fact that $\sigma_j^{(0)} \leq \sigma_j^{(i)} < 0$ for all $0 \leq i \leq t$. Collecting the above results, we find that

$$\phi_j^{(t+1)} \geq \phi_j^{(0)} - \rho t\left(\lambda - (\beta_j^*)^2\frac{n-1}{n}(\eta - \frac{1}{n})\right)\sigma(\phi_j^{(0)})(1-\sigma(\phi_j^{(0)})).$$

Therefore, as long as $t \geq \dfrac{\phi_j^{(0)}}{\rho\left(\lambda-(\beta_j^*)^2\frac{n-1}{n}(\eta-\frac{1}{n})\right)\sigma(\phi_j^{(0)})(1-\sigma(\phi_j^{(0)}))} := T_1$, we have $\phi_j^{(t+1)} \geq 0$.

By using the inequality $\sigma(x+y) \leq \sigma(x) + y\sigma'(x)$ for any $x, y > 0$, for any $j \in \mathcal{A}$ and $t \geq T_1$ we obtain that

$$\begin{aligned}1 - \sigma(\phi_j^{(t+1)}) &\geq 1 - \sigma(\phi_j^{(t)}) + \rho G(\phi_j^{(t)})\sigma(\phi_j^{(t)})(1-\sigma(\phi_j^{(t)}))\sigma'(\phi_j^{(t)}) \\ &= \left(1-\sigma(\phi_j^{(t)})\right)\left[1 + \rho\sigma^2(\phi_j^{(t)})G(\phi_j^{(t)})\left(1-\sigma(\phi_j^{(t)})\right)\right]. \qquad \text{(B.12)}\end{aligned}$$

Based on the result of Lemma 2, we find that

$$G(\phi_j^{(t)}) \geq \lambda - \frac{\sigma^2 + \|\boldsymbol{\beta}^*\|_2^2 + (n-1)(\beta_i^*)^2}{n}, \qquad \text{(B.13)}$$

which is due to $\left\|\boldsymbol{\beta}^*_{-\boldsymbol{z}}\right\|_2^2 \leq \|\boldsymbol{\beta}^*\|$ and $\|\boldsymbol{z}\|_0 \geq 0$ for all $\boldsymbol{z}$. Plugging this inequality into equation (B.12), we achieve the conclusion of the lower bound in part (a) with $c_1 = \rho[(\sigma^2 + \|\boldsymbol{\beta}^*\|_2^2 + (n-1)(\beta_i^*)^2)/n - \lambda] > 0$.

Regarding the upper bound, we first prove that

$$\sigma(x+y) \geq \sigma(x) + \exp(-y)y\sigma'(x) \qquad \text{(B.14)}$$

for all $x, y > 0$. In fact, from the mean-value theorem, there exists $\xi \in (x, x+y)$ such that $\sigma'(\xi) = (\sigma(x+y) - \sigma(x))/y$. Since the function $\sigma'(.)$ is monotonically decreasing in

$(0,\infty)$, we have $\sigma'(\xi) \geq \sigma'(x+y)$. Simple calculation yields that $\sigma'(x+y)/\sigma'(x) \geq \exp(-y)$ for all $x, y > 0$. Therefore, we obtain the conclusion of inequality (B.14). Given the inequality (B.14), for any $t \geq T_1$ we have

$$1 - \sigma(\phi_j^{(t+1)}) \leq 1 - \sigma(\phi_j^{(t)}) + \exp\left(\rho\sigma(\phi_j^{(t)})(1-\sigma(\phi_j^{(t)}))G(\phi_j^{(t)})\right)\rho G(\phi_j^{(t)})\sigma^2(\phi_j^{(t)})(1-\sigma(\phi_j^{(t)}))^2. \quad \text{(B.15)}$$

Since $G(\phi_j^{(t)}) < 0$ and $\phi_j^{(t)} \geq 0$, an application of the bound (B.13) and standard inequality $\sigma(\phi_j^{(t)})(1-\sigma(\phi_j^{(t)})) \leq 1/4$ leads to

$$\sigma(\phi_j^{(t)})(1-\sigma(\phi_j^{(t)}))G(\phi_j^{(t)}) \geq \frac{1}{4}\left(\lambda - \frac{\sigma^2 + \|\boldsymbol{\beta}^*\|_2^2 + (n-1)(\beta_i^*)^2}{n}\right). \quad \text{(B.16)}$$

Furthermore, using the proof argument of Proposition 3, we find that

$$G(\phi_j^{(t)}) \leq \lambda - (\beta_i^*)^2\frac{n-1}{n}(\eta - \frac{1}{n}), \quad \text{(B.17)}$$

where $\eta = \min\left\{\sqrt{p\log(n)/2n^2}, (\|\boldsymbol{\beta}^*\|_2^2 + \sigma^2)/(n-1)\min_{k\in\mathcal{A}}(\beta_k^*)^2\right\}$. Plugging the bounds (B.16) and (B.17) into the equation (B.15), we obtain the conclusion of the upper bound in part (a) with C_1 is given by

$$C_1 = \rho\exp\left(\frac{\rho}{4}\left(\lambda - \frac{\sigma^2 + \|\boldsymbol{\beta}^*\|_2^2 + (n-1)(\beta_i^*)^2}{n}\right)\right)\left((\beta_i^*)^2\frac{n-1}{n}(\eta - \frac{1}{n}) - \lambda\right) > 0.$$

As a consequence, we obtain the conclusion of the upper bound in part (a).

(b) The proof of part (b) follows the same line of proof argument as that in part (a). Indeed, since $\lambda \in \mathcal{I}$ and $\sum_{j=1}^p \sigma(\phi_j^{(0)}) \leq \varpi - S$, we have $G(\phi_j^{(t)}) > 0$ for all $j \notin \mathcal{A}$. It shows that the updates $\phi_j^{(t+1)}$ are monotonically decreasing, i.e., $\phi_j^{(t+1)} < \phi_j^{(t)}$ for any $t \geq 0$ and $j \notin \mathcal{A}$. Similar to part (a), we first find a lower bound on t such that $\phi_j^{(t+1)} \leq 0$. In fact, we assume that $\phi_j^{(i)} > 0$ for all $0 \leq i \leq t$. Then, based on the proof of Proposition 3, for any $0 \leq i \leq t$, we find that

$$\mathbb{E}_{\mathbf{X},\boldsymbol{y},\boldsymbol{u}}[g_{U2G}(\boldsymbol{u}; \sigma(\boldsymbol{\phi}^{(i)}))_j] \geq \left(\lambda - \frac{\sigma^2 + \|\boldsymbol{\beta}^*\|_2^2}{n}\right)\sigma(\phi_j^{(i)})(1-\sigma(\phi_j^{(i)}))$$

$$\geq \left(\lambda - \frac{\sigma^2 + \|\boldsymbol{\beta}^*\|_2^2}{n}\right)\sigma(\phi_j^{(0)})(1-\sigma(\phi_j^{(0)})),$$

where the second inequality is due to the fact that $\sigma_j^{(0)} \geq \sigma_j^{(i)} > 0$ for any $0 \leq i \leq t$. Therefore, we obtain that

$$\phi_j^{(t+1)} \leq \phi_j^{(0)} - \rho t\left(\lambda - \frac{\sigma^2 + \|\boldsymbol{\beta}^*\|_2^2}{n}\right)\sigma(\phi_j^{(0)})(1-\sigma(\phi_j^{(0)})).$$

The above inequality demonstrates that as long as $t \geq \dfrac{\phi_j^{(0)}}{\rho\left(\lambda - \frac{\sigma^2+\|\boldsymbol{\beta}^*\|_2^2}{n}\right)\sigma(\phi_j^{(0)})(1-\sigma(\phi_j^{(0)}))} := T_2$, we have $\phi_j^{(t+1)} \leq 0$.

Now, we can check that $\sigma(x-y) \geq \sigma(x) - y\sigma'(x)$ for all $x < 0$ and $y > 0$. Using this inequality, for any $j \notin \mathcal{A}$ and $t \geq T_2$ we obtain that

$$\sigma(\phi_j^{(t+1)}) \geq \sigma(\phi_j^{(t)}) - \rho G(\phi_j^{(t)})\sigma(\phi_j^{(t)})(1-\sigma(\phi_j^{(t)}))\sigma'(\phi_j^{(t)}).$$

From the result of Lemma 2, it is clear that $G(\phi_j^{(t)}) \leq \lambda$ for all $j \notin \mathcal{A}$. Therefore, we have

$$\sigma(\phi_j^{(t+1)}) \geq \sigma(\phi_j^{(t)})\left[1 - \lambda\rho(1-\sigma(\phi_j^{(t)}))^2\sigma(\phi_j^{(t)})\right].$$

We achieve the conclusion of the lower bound in part (b) with $c_2 = \rho\lambda > 0$.

Moving to the upper bound, with similar argument as that of equation (B.14), we can check that

$$\sigma(x-y) \geq \sigma(x) - \exp(-y)y\sigma'(x),$$

for any $x < 0$ and $y > 0$. Given that inequality, for any $t \geq T_2$ we obtain that

$$\begin{aligned}\sigma(\phi_j^{(t+1)}) &\leq \sigma(\phi_j^{(t)}) \\ &\quad - \exp\left(-\rho\sigma(\phi_j^{(t)})(1-\sigma(\phi_j^{(t)}))G(\phi_j^{(t)})\right)\rho G(\phi_j^{(t)})\sigma(\phi_j^{(t)})(1-\sigma(\phi_j^{(t)}))\sigma'(\phi_j^{(t)}). \quad \text{(B.18)}\end{aligned}$$

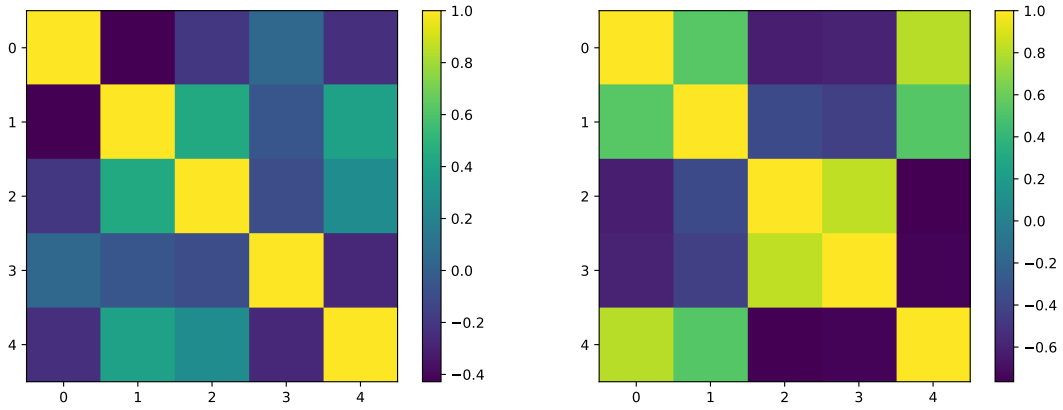
Based on the result of Lemma 2, we can check that $\sigma(\phi_j^{(t)})(1-\sigma(\phi_j^{(t)})G(\phi_j^{(t)}) \leq \lambda/4$ and $G(\phi_j^{(t)}) \geq \lambda - (\sigma^2 + \|\boldsymbol{\beta}^*\|_2^2)/n$. Putting the above results together, we have

$$\sigma(\phi_j^{(t+1)}) \leq \sigma(\phi_j^{(t)}) \left[1 - C_2(1-\sigma(\phi_j^{(t)}))^2 \sigma(\phi_j^{(t)})\right],$$

for any $t \geq T_2$ where $C_2 = \rho \exp(-\rho\lambda/4)\left(\lambda - (\sigma^2 + \|\boldsymbol{\beta}^*\|_2^2)/n\right) > 0$. As a consequence, we reach the conclusion of the upper bound in part (b). Q.E.D.

C Additional Results

In this section, we provide additional experimental results. For the Prostate cancer dataset, Figure A.1 shows the pairwise correlations of the two types of chosen covariates in the true active set. Table 1 summarizes the results for the variable selection methods when the multi-collinearity in the true active set is high.



(a) Moderate multi-collinearity (b) High multi-collinearity

Figure A.1: Two types of correlations between the covariates in the true active sets, Prostate cancer dataset.

Table 1: Results of the Prostate cancer dataset, with $n = 102, p = 1000, S = 5$, *high* multicollinearity. Reported results are the average of 100 independent trials.

	Precision	Recall	F1	Nonzero	RR	RTE	PVE
LASSO	0.091	0.712	0.160	44.9	0.090	1.454	0.759
SCAD	0.270	0.368	0.306	7.04	0.427	3.144	0.478
MIO	0.474	0.474	0.474	5.00	0.197	1.989	0.669
U2G	0.852	0.796	0.818	4.73	0.058	1.292	0.785
U2G(VI)	0.936	0.812	0.863	4.34	0.057	1.286	0.786

For the compressive sensing, as shown in Figure A.2 and Table 2, when SNR_d is high, all methods can achieve low prediction error but BP and BCS cannot set the coefficients of the non-signal covariates as exact zero. As shown in Figure 4, when SNR_d is low, methods with L_1 penalty cannot recover strong sparsity, but our method with L_0 penalty can.

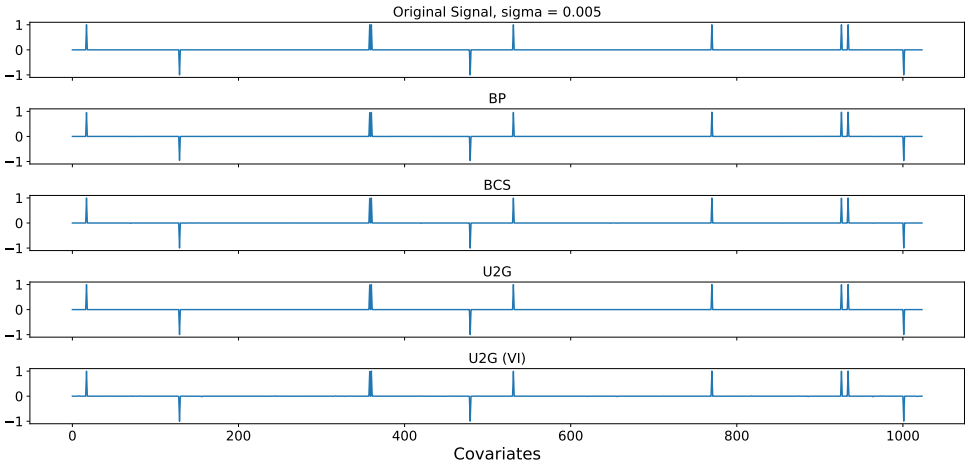


Figure A.2: Results of compressive sensing when SNR_d is high, with $n = 500$, $p = 1024$, $\sigma = 0.005$.

Table 2: Results of compressive sensing when the SNR_d is high, with $n = 500$, $p = 1024$, $\sigma = 0.005$.

	Precision	Recall	F1	Nonzero	RR	RTE	PVE
BP	0.009	1.000	0.019	1024	1e-3	508.0	0.9987
BCS	0.322	1.000	0.488	31	1e-4	55.4	0.9998
U2G	1.000	1.000	1.000	10	2e-5	9.2	1.0000
U2G(VI)	1.000	1.000	1.000	10	3e-5	13.7	1.0000

References

Tao, T. (2008). Singularity and determinant of random matrices. Technical report, Lewis Memorial Lecture.