Probabilistic Best Subset Selection via Gradient-Based Optimization

Mingzhang Yin[⋆], Nhat Ho[†], Bowei Yan[∗],

Xiaoning Qian[‡], Mingyuan Zhou[⋆]

UT Austin[⋆], UC Berkeley[†], Texas A&M University[‡]

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 transform the optimization problem from a discrete space to a continuous one via reparameterization. The new objective function is a reformulation of the exact L_0-regularized regression problem (a.k.a. best subset selection). In the framework of stochastic gradient descent, we propose a family of unbiased and efficient gradient estimators that are used to optimize the best subset selection objective and its variational lower bound. Under this family, we identify the estimator with 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 real data sets, the proposed method outperforms existing ones based on penalized regression or best subset selection, 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]

1 Introduction

Variable selection by penalized regression is widely applied to uncover sparse structures in high dimensional data. Solving L_0-regularized regression, also known as the best-subset

---

[1]Code examples are at https://github.com/mingzhang-yin/Probabilistic-Best-Subset

1

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 parameter. We consider the high-dimensional regime, where the number of covariates p exceeds the sample size n, and can potentially grows 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 theory based model selection methods. 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 goodness of fit and model simplicity, pertaining to the out-of-sample predictive accuracy. In practice, such information criteria are often used in the comparison of alternative models, but rarely as objective functions to directly select a model by optimization. This limits the number of candidate models and the selected model can be sub-optimal. The main challenge is the non-convexity and discontinuity of the L_0 penalty, which makes optimizing L_0-regularized objectives NP-hard (Natarajan, 1995).

In order 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 most. Conversely, the backward stepwise selection starts from all covariates and deletes one at a time that impacts the fitting 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 mostly correlate with the prediction residuals (Joseph, 2013; Donoho et al., 2012). The greedy algorithms yield a sequence of subsets with increasing (or decreasing) size, but none of these selected sets are generally 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 $(q > 0)$, which are $\sum_{j=1}^{p} \beta_j^q$ with β_j as the estimated coefficients. When $q \geq 1$, the L_q penalty is convex; when $q \leq 1$, the regularization encourages sparse estimation and hence achieves variable selection (Fan and Lv, 2010). The widely used least absolute shrinkage and selection operator (LASSO), with $q = 1$, encourages the estimation to be sparse while enjoying the 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 parameter 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. For example, smoothly clipped absolute deviation (SCAD) of Fan and Li (2001) and minimax concave penalty (MCP) of Zhang (2010) approximate the hard-thresholding property of L_0 penalty by piecewise-defined non-convex penalties. The coefficient estimators are known to be consistent and unbiased when the estimation is sufficiently large so that there is high non-vanishing probability. Another line of work directly designs functions, which are called peudo-L_0 penalties, to closely resemble the L_0 pseudo-norm. Liu and Wu (2007) and Shen et al. (2012) approximate the L_0 penalty with a convex function $\min\{|\beta_j|/\tau, 1\}$. The former work combines this approximation with the L_1 penalty and solves the optimization problem with mixed integer optimization (MIO). The latter work develops a difference convex method to solve a primal-dual problem in an iterative manner. Dicker et al. (2013) approximate 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 additional tuning parameter τ to control the approximation accuracy, which induces bias and sensitivity to its value.

Comparing to continuous approximations, the solution of the exact best subset selection enjoys superior statistical properties, such as the unbiased estimation for regression coefficients (Greenshtein, 2006; Zhang and Zhang, 2012; Belloni and Chernozhukov, 2013), known as the oracle property, and low in-sample risk (Foster and George, 1994). For orthogonal design matrix, Johnson et al. (2015) show that the predictive risk of 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

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

with a two-stage algorithm. In the first stage, an iterative hard thresholding method is used to provide 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 a leaps and bounds method (Furnival and Wilson, 1974), to the setting where p is in 1000s. However, the core MIO step relies upon a non-convex optimization tool, which is hard to generalize beyond linear regression problems. Though the speed has been improved by following works (Hazimeh and Mazumder, 2018), MIO nevertheless induces substantial computation load.

It is worth noticing that 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, due to the discontinuity, the two problems are not exactly equivalent in that there may not be a surjection between penalty parameters λ and S (Polson and Sun, 2019). Since there is 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 approach to solve the exact L_0-regularized regression, where we cast the discrete optimization problem to an equivalent continuous one. Second, 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 one with minimal variance and non-vanishing signal-to-noise ratio (SNR) in the univarate case and generalize it to the multivariate case. Third, we theoretically analyze the conditions that guarantee the convergence of the updates from SGD to the ground truth in expectation. Last but not the least, we empirically show our probabilistic approach 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 remainder of our paper is organized as follows. In Section 2, we provide a continuous reformulation of the L_0-regularized linear regression problem. In Section 3, we define a family of efficient and unbiased gradient estimators, and analyze their variance properties. In Section 4, we analyze the conditions that guarantee the convergence to the ground truth in expectation. In Section 5, we construct a tightened variational lower bound and show how to optimize it with the proposed gradient estimators. In Section 6, we verify the effectiveness of our method on a variety of synthetic and real data sets.

Notation: We use n as the sample size, p as the number of covariates, and S as the size of the true active set. We use $\boldsymbol{x}_i$ to denote the ith row of design matrix $\mathbf{X}$ and X_j as its jth column. We use $\mathbf{X}_{\mathcal{A}}$ to denote a submatrix of $\mathbf{X}$ as $\mathbf{X}_{\mathcal{A}} = \{X_j\}_{j \in \mathcal{A}}$, $\mathcal{A} \subseteq \{1, \ldots, p\}$.

2 Continuous 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 $\mathbf{X}_{\mathcal{A}}$ and $\mathcal{A} \subseteq \{1, \ldots, p\}$ is called the active set. The size of the true active set 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 (Mitchell and Beauchamp, 1988; George and McCulloch, 1993; Ishwaran et al., 2005; Zhou et al., 2009; Polson and Sun, 2019), where $\odot$ denotes an element-wise product. The binary vector $\boldsymbol{z} \in \{0, 1\}^p$ indicates the inclusion of covariates in the active set and $\boldsymbol{\alpha} \in \mathbb{R}^p$ encodes the scale of coefficients. With these augmented latent variables, the optimization problem in Eq. (1) can be equivalently expressed as

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

Similar to optimizing $\boldsymbol{\beta}$ itself, however, optimizating $\boldsymbol{z}$ is also an NP-hard combinatorial problem.

In this paper, we reformulate the discrete optimization problem in Eq. (3) to an optimization problem in the continuous space. 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. We denote the active set inferred by $\boldsymbol{z}$ as $\mathcal{Z} := \{j\}_{j:z_j \neq 0}$. Then we transform Eq. (3) 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 (Continuous Reformulation). *The L_0-regularized best subset selection problem in Eq. (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 of problem (3) is a feasible solution of problem (4) that achieves the same object value, and vice versa. The proof is conceptually straightforward and we defer it to Appendix A.1. 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. Notice in the objective function (4) after the reformulation, given $\boldsymbol{z}$, the inner optimization is an ordinary least square (OLS) problem on the design matrix $\mathbf{X}_{\mathcal{Z}}$ which has a closed-form solution. If the sparse pattern reveals, solving the OLS problem is effcient with a small size of active set $\mathcal{Z}$. Though 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 $\boldsymbol{\phi}$ are sufficiently large.

A naive approach that is guaranteed to select the best subset is to exhaust all possible subsets. However, to find the optimal subset from p covariates requires evaluating the objective function 2^p times, which is often computationally infeasible unless p is small. Even if we know the cardinality of the true active set, the computation of exhaustion nevertheless can be overwhelming. For example, to exhaust all subsets with cardinality as 10 from $p = 1000$ covariates, we need to evaluate objective function $\binom{1000}{10}$ times, which is in the order of 10^{23}. In the following sections, we propose a gradient based method that can significantly reduce the number of function evaluations that is required to recover the true active set.

3 A Family of Unbiased Gradient Estimators

In this section we consider a general optimization objective

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

where $p_\phi(z) := \prod_{j=1}^{p} \mathrm{Bern}(\sigma(\phi_j))$. The objective function in Eq. (4) can be considered as a special case of the objective function in Eq. (5) when $f(z) = \min_{\alpha} \|y - \mathbf{X}(\alpha \odot z)\|_2^2 / n + \lambda \|z\|_0$. Taking gradient of the function $\mathcal{E}$ in (5) with respect to ϕ, we have

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

which is called the *score method* in statistical literature and *REINFORCE* in reinforcement learning (Williams, 1992). Even though the expectation in Eq. (6) cannot be computed analytically, the key is that an unbiased Monte Carlo estimation of that expectation can be obtained by $\frac{1}{K} \sum_{k=1}^{K} f(z_k) \nabla_\phi \log p_\phi(z_k)$ with $z_1, \ldots, z_K \overset{iid}{\sim} p_\phi(z)$. One advantage of the REINFORCE estimator is that the number of evaluations for function f does not grow with the dimension of the covariate space, which makes it generalizable to high dimensional z. Another advantage is that this estimator only needs the value of $f(z)$, which makes it widely applicable to the situations when $f(z)$ is discontinuous or even has no explicit expression (*e.g.*, in reinforcement learning, $f(z)$ could be the reward returned by the environment when the agent takes action 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 evaluations of function f at each gradient step to get a gradient estimation with sufficiently low variance, which can be computationally demanding.

We propose a general framework to construct unbiased gradient estimators for the objective in Eq. (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 further require that the gradient estimators do not require function continuity and the number of function evaluations do 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:

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

We require the following properties for gradient estimation in the univariate case.

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

- *Unbiasedness:*

$$\mathbb{E}_{u \sim \text{Unif}(0,1)}[g(u; \sigma(\phi))] = \sigma(\phi)(1 - \sigma(\phi))[f(1) - f(0)], \tag{8}$$

- *Functional form:*

$$g(u; \sigma(\phi)) = a(u; \sigma(\phi))f(\mathbf{1}_{[u<\sigma(\phi)]}) + b(u; \sigma(\phi))f(\mathbf{1}_{[u>1-\sigma(\phi)]}) \tag{9}$$

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

The estimator family in Definition 1 incorporates many popular unbiased gradient estimators. For example, the univariate REINFORCE estimator can be considered as a special case, with gradient

$$g_R(u; \sigma(\phi)) = f(\mathbf{1}_{[u<\sigma(\phi)]})(\mathbf{1}_{[u<\sigma(\phi)]} - \sigma(\phi)). \tag{10}$$

Recently, a newly proposed augment-REINFORCE-merge (ARM) gradient (Yin and Zhou, 2019) has achieved 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

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

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

$$g_{\text{ARM}_0}(u; \sigma(\phi)) = [f(\mathbf{1}_{[u>\sigma(-\phi)]}) - f(\mathbf{1}_{[u<\sigma(\phi)]})](u - \tfrac{1}{2})\big|\mathbf{1}_{[u>\sigma(-\phi)]} - \mathbf{1}_{[u<\sigma(\phi)]}\big|. \tag{12}$$

	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.

In the univariate case, the ARM_0 estimator is identical to the ARM estimator, but when it comes to the multivariate case, as discussed in Section 3.2, it can produce sparse gradient where many different dimensions may become exactly zeros, which in practice improves the gradient estimation. Note this straightforward sparsification, though not used in the original ARM algorithm of Yin and Zhou (2019), has already been adopted as the default setting by several recent works when utilizing ARM or its categorical generalization in their respective applications (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:

$$g_{\mathrm{U2G}}(u;\sigma(\phi)) = \frac{\sigma(\lvert\phi\rvert)}{2}[f(\mathbf{1}_{[u>1-\sigma(\phi)]}) - f(\mathbf{1}_{[u<\sigma(\phi)]})](\mathbf{1}_{[u>\sigma(-\phi)]} - \mathbf{1}_{[u<\sigma(\phi)]}). \qquad (13)$$

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 defined in Definition 1, subject to an additional assumption that the gradient has a non-vanishing SNR, defined as $\mathrm{SNR}_g := \mathbb{E}[g(u)]/\sqrt{\mathrm{var}[g(u)]}$. We discuss this property in detail in Section 3.1.1. We summarize the above-mentioned estimators in Table 1 in a form compatible with Definition 1.

The unbiased estimators have the same expectation, which equals to the true gradient in Eq. (7). What differs between the estimators is how the stochastic gradient is expressed as a function of $u \sim \mathrm{Uniform}(0,1)$, hence different gradient variance. As an illustrative example, we plot the functions $g(u)$ and $g^2(u)$ for the above-mentioned 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 of Figure 1.

The intrinsic nature of variable selection is the comparison between potential models, which takes the *relative* difference as a measure of goodness. Comparing the first column
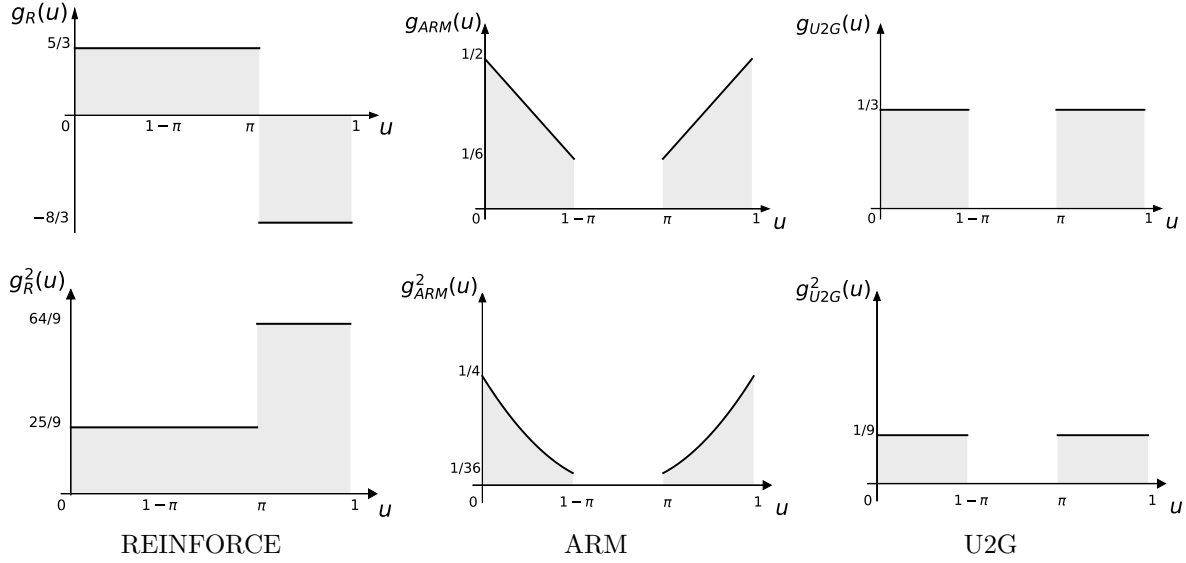
Figure 1: The characteristic curves of gradient estimators. In this illustrative example, $f(1) = 5$, $f(0) = 4$, and $\pi = \sigma(\phi) = 2/3$. The top row is the function $g(u)$ with respect to u; the second row is the function $g^2(u)$ As shown in the first row, the unbiased estimators have the same integration. But as shown in the second row, they have different gradient variance, represented by the area under a curve (up to an additive constant).

of Figure 1 to the other two columns, and comparing Eq. (10) to Eqs. (11) and (13), intuitively we find, if the function $f(z)$ appears in the estimation as a relative difference $f(z) - f(z')$, the scale of the gradient does not increase with the scale of $f(z)$; thus it controls the magnitude of the second moment. The gradient variance can be further reduced if the direction and magnitude of gradient estimator do not change with u. Formally, we have the following proposition with the proof deferred to Appendix A.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 aim to answer the question that under what condition, U2G is the optimal within the estimator family defined 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 $\pi > 1/2$. As shown in Eq. (7), the scale of the true gradient that the estimators concentrate on diminishes when $\Delta \to 0$. Therefore, to ensure non-diminishing in the gradient estimator, a necessary condition that the estimator has to satisfy is that when the difference between potential models gets small the variance deceases to zero, that is

$$\lim_{\Delta \to 0} \text{var}[g(u; \pi)] = 0, \quad \text{for all } \pi. \tag{14}$$

This condition ensures the estimated gradient can distinguish the optimal model from the others, even when the objective function values are close.

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. (9) 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

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

11

which contradicts the condition (14). Therefore if the estimator has its variance upper bounded by the model difference, 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

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

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. Therefore within the proposed family, U2G is the uniformly minimum-variance unbiased estimator (UMVUE), which has the optimal efficiency. We summarize the above analysis in the following proposition.

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

Specifically, the variance of U2G estimator is

$$
\mathrm{var}[g_{\mathrm{U2G}}(u; \pi)] = \pi|\pi - \frac{1}{2}|(1 - \pi)\max\{\pi, 1 - \pi\}[f(1) - f(0)]^2 \tag{15}
$$

$$
\leq C[f(1) - f(0)]^2 \tag{16}
$$

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

$$
\mathrm{SNR}_g(\pi) = \sqrt{\frac{\pi(1 - \pi)}{|\pi - \frac{1}{2}|\max\{\pi, 1 - \pi\}}}, \tag{17}
$$

which is the same with 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_g distributions of the ARM and U2G estimators are shown in Figure 2.
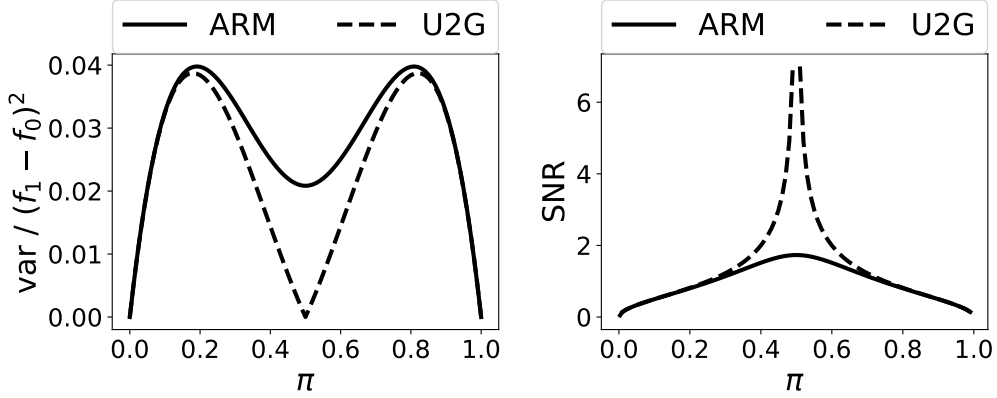
Figure 2: Variance and SNR of univariate ARM and U2G estimators.

Remark 1. *Comparing to the REINFORCE estimator with leave-one-out baseline (LOO) ([Kool et al., 2019](#)), which uses two independent samples to construct unbiased estimator as*

$$g_{LOO}(u_1, u_2) = \frac{1}{2}[f(\mathbf{1}_{[u_1 < \sigma(\phi)]}) - f(\mathbf{1}_{[u_2 < \sigma(\phi)]})](\mathbf{1}_{[u_1 < \sigma(\phi)]} - \mathbf{1}_{[u_2 < \sigma(\phi)]}), \tag{18}$$

the U2G estimator has lower variance uniformly, as $\forall \phi$, $\mathbb{E}[g_{U2G}^2(u)]/\mathbb{E}[g_{LOO}^2(u_1, u_2)] = \sigma(|\phi|)/2 \leq 1$, and they have the same expectation.

3.2   Multivariate Generalization

We consider the problem in Eq. (5) when latent variable $\boldsymbol{z}$ is high dimensional. When z is univarate, the true gradient can be calculated analytically as in Eq. (7), and hence it is unnecessary to estimate the gradient. In the multivariate case, the true gradient can be computed element-wisely as

$$\frac{\partial}{\partial \phi_v} \mathcal{E}(\boldsymbol{\phi}) = \frac{\partial}{\partial \phi_v} \mathbb{E}_{\boldsymbol{z} \sim \prod_{j=1}^p p(z_j; \sigma(\phi_j))}[f(\boldsymbol{z})]$$
$$= \mathbb{E}_{\boldsymbol{z}_{\backslash v}}[\sigma(\phi_v)(1 - \sigma(\phi_v))(f(\boldsymbol{z}_{\backslash v}, z_v = 1) - f(\boldsymbol{z}_{\backslash v}, z_v = 0))]. \tag{19}$$

However, to compute the v-th element of the gradient estimation, we need to set the v-th element z_v as 0 and 1 while keeping other elements $\boldsymbol{z}_{\backslash v}$ the same for the two function evaluations in Eq. (19), which can be considered as measuring the effect of a randomized controlled trial. This has to be done for each element of $\boldsymbol{z}$ separately. Therefore, it requires at least $2p$ function evaluations to get an estimated gradient at each step, which

13

can be hard to generalize to a large-p setting. For multivariate $\boldsymbol{z}$, applying the univariate gradient estimators defined in Definition 1 and following the related derivation in Yin and Zhou (2019), we have

$$
\frac{\partial}{\partial \phi_v} \mathbb{E}_{\boldsymbol{z} \sim \prod_{j=1}^{p} p(z_j; \sigma(\phi_j))}[f(\boldsymbol{z})]
$$

$$
= \mathbb{E}_{\boldsymbol{z}_{\backslash v}} \frac{\partial}{\partial \phi_v} \mathbb{E}_{z_v \sim p(z_v; \sigma(\phi_v))}[f(z_v, \boldsymbol{z}_{\backslash v})]
$$

$$
= \mathbb{E}_{\boldsymbol{z}_{\backslash v}} \mathbb{E}_{u_v \sim \mathrm{Unif}(0,1)}[a(u_v; \sigma(\phi_v))f(\mathbf{1}_{[u_v < \sigma(\phi_v)]}, \boldsymbol{z}_{\backslash v}) + b(u_v; \sigma(\phi_v))f(\mathbf{1}_{[u_v > 1 - \sigma(\phi_v)]}, \boldsymbol{z}_{\backslash v})]
$$

$$
= \mathbb{E}_{\boldsymbol{u} \sim \prod_{j=1}^{p} \mathrm{Unif}(0,1)}[a(u_v; \sigma(\phi_v))f(\mathbf{1}_{[\boldsymbol{u} < \sigma(\boldsymbol{\phi})]}) + b(u_v; \sigma(\phi_v))f(\mathbf{1}_{[\boldsymbol{u} > 1 - \sigma(\boldsymbol{\phi})]})] \tag{20}
$$

where the first equality is by 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). The key is that to compute an 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 $\boldsymbol{u} \sim \prod_{j=1}^{p} \mathrm{Unif}(0, 1)$, and share it across all elements of the gradient vector. This greatly reduces computation time for each gradient step. Eq. (20) ensures the multivariate gradient is unbiased. Written in a vector form, the estimators in Section 3.1 have the multivariate generalization as

$$
\boldsymbol{g}_R(\boldsymbol{u}; \sigma(\boldsymbol{\phi})) = f(\mathbf{1}_{[\boldsymbol{u} < \sigma(\boldsymbol{\phi})]})(\mathbf{1}_{[\boldsymbol{u} < \sigma(\boldsymbol{\phi})]} - \sigma(\boldsymbol{\phi})) \tag{21}
$$

$$
\boldsymbol{g}_{\mathrm{ARM}_0}(\boldsymbol{u}; \sigma(\boldsymbol{\phi})) = [f(\mathbf{1}_{[\boldsymbol{u} > 1 - \sigma(\boldsymbol{\phi})]}) - f(\mathbf{1}_{[\boldsymbol{u} < \sigma(\boldsymbol{\phi})]})](\boldsymbol{u} - \tfrac{1}{2}) \odot |\mathbf{1}_{[\boldsymbol{u} > 1 - \sigma(\boldsymbol{\phi})]} - \mathbf{1}_{[\boldsymbol{u} < \sigma(\boldsymbol{\phi})]}| \tag{22}
$$

$$
\boldsymbol{g}_{\mathrm{U2G}}(\boldsymbol{u}; \sigma(\boldsymbol{\phi})) = \tfrac{1}{2}[f(\mathbf{1}_{[\boldsymbol{u} > 1 - \sigma(\boldsymbol{\phi})]}) - f(\mathbf{1}_{[\boldsymbol{u} < \sigma(\boldsymbol{\phi})]})]\sigma(|\boldsymbol{\phi}|) \odot (\mathbf{1}_{[\boldsymbol{u} > 1 - \sigma(\boldsymbol{\phi})]} - \mathbf{1}_{[\boldsymbol{u} < \sigma(\boldsymbol{\phi})]}), \tag{23}
$$

where $\boldsymbol{u} \sim \prod_{j=1}^{p} \mathrm{Unif}(0, 1)$, 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, for example 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 covariance matrix. By the law of total variance, the variance of element v in the gradient vector can be decomposed as

$$
\mathrm{var}_{\boldsymbol{u}}[\boldsymbol{g}_v(\boldsymbol{u}; \sigma(\boldsymbol{\phi}))] = \mathrm{var}\{\mathbb{E}[\boldsymbol{g}_v(\boldsymbol{u}; \sigma(\boldsymbol{\phi}))|\boldsymbol{u}_{\backslash v}]\} + \mathbb{E}\{\mathrm{var}[\boldsymbol{g}_v(\boldsymbol{u}; \sigma(\boldsymbol{\phi}))|\boldsymbol{u}_{\backslash v}]\} \tag{24}
$$

The first term on the right-hand side (RHS) of Eq. (24) is the irreducible variance,

shared by all unbiased gradient estimators, which can be further computed as

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

with $\boldsymbol{z} = \mathbf{1}_{[\boldsymbol{u}<\sigma(\boldsymbol{\phi})]}$. It measures the variance of knocking out one covariate. The second term measures the average variance in a single dimension. Given a fixed $\boldsymbol{u}_{\backslash v}$, 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 $\boldsymbol{u}_{\backslash v}$, the second term of U2G estimator is small; hence the total variance is low in the multivariate case.

Since $\Delta_{\boldsymbol{u},v}f$ is bounded by the maximal residual reduction by adding the v-th covariate to the active set and $\sigma(\phi_v)(1-\sigma(\phi_v)) \to 0$ when the optimization convergences, the first term on the RHS of Eq. (24) is well-controlled. When estimated with Monte Carlo integration, it can be further reduced by multiple random samples. This term is the same for all unbiased estimators, so the differences of estimators are reflected on the second term. As shown in Section 3.1, ARM and U2G control the scale of the second term; therefore they reduce the total variance of the gradient vector element-wisely.

Algorithm 1 Best subset selection with continuous 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 $\boldsymbol{u}_k \overset{i.i.d.}{\sim} \prod_{j=1}^p \text{Unif}(0,1)$ 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. (22)
  Update $\boldsymbol{\phi} = \boldsymbol{\phi} - \frac{1}{K}\rho_t \sum_{k=1}^K \boldsymbol{g}_k$ with stepsizes ρ_t
end

---

4 Convergence Analysis

In this section, we assume that the observations $(\mathbf{X}, \boldsymbol{y})$ are generated from the following model with active set $\mathcal{A}$:

$$\boldsymbol{y} = \mathbf{X}\boldsymbol{\beta}^* + \boldsymbol{\epsilon}, \quad \boldsymbol{\epsilon} \sim \mathcal{N}(0, \sigma^2\mathbf{I}) \tag{25}$$

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 model $\mathbf{X} = (\boldsymbol{x}_1, \cdots, \boldsymbol{x}_n)^T$ in which $\boldsymbol{x}_i \sim \mathcal{N}(0, \mathbf{I}_p)$ for $i \in [n]$. In order to ease the presentation, we denote

$$f_{\mathbf{X},\boldsymbol{y}}(\boldsymbol{z}) = \min_{\boldsymbol{\alpha}} \frac{1}{n} \ \|\boldsymbol{y} - \mathbf{X}(\boldsymbol{\alpha} \odot \boldsymbol{z})\|_2^2 + \lambda \|\boldsymbol{z}\|_0. \tag{26}$$

Here we use subscripts to make the dependency of f on (X, y) explicit. Denote $\mathbf{X}_{\boldsymbol{z}} \in \mathbb{R}^{n \times \|z\|_0}$ as the matrix which consists of $\{X_j : z_j \neq 0\}$, and $\mathbf{X}_{-\boldsymbol{z}}$ is the complement in the design matrix. The following lemma calculates the expectation of the gradient over the randomness of $\boldsymbol{u}$, for any fixed data $(\mathbf{X}, \boldsymbol{y})$.

Lemma 1. *Consider* $g_{\mathrm{U2G}}(\boldsymbol{u}; \sigma(\boldsymbol{\phi}))$ *as in Eq. (22). For fixed data* $(\mathbf{X}, \boldsymbol{y})$*, 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* $\Delta_{\boldsymbol{z}} f = (\Delta_{\boldsymbol{z},1} f, \cdots, \Delta_{\boldsymbol{z},p} f)$.

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 $\boldsymbol{g}_{\mathrm{U2G}}(\boldsymbol{u}; \sigma(\boldsymbol{\phi})) = \frac{f(\boldsymbol{z}) - f(\tilde{\boldsymbol{z}})}{2} \sigma(|\boldsymbol{\phi}|) \odot (\mathbf{1}_{[\boldsymbol{u} > 1 - \sigma(\boldsymbol{\phi})]} - \mathbf{1}_{[\boldsymbol{u} < \sigma(\boldsymbol{\phi})]})$. 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 jth dimension of the gradient vector

$$\mathbb{E}_{\boldsymbol{u}}[g(\boldsymbol{u})_j] = \frac{\sigma(|\phi_j|)}{2} \mathbb{E}_{\boldsymbol{u}} \sum_{i=1}^p (f_{\mathbf{X},\boldsymbol{y}}(\boldsymbol{z}^{i-1}) - f_{\mathbf{X},\boldsymbol{y}}(\boldsymbol{z}^i))(\mathbf{1}_{[u_j > \sigma(-\phi_j)]} - \mathbf{1}_{[u_j < \sigma(\phi_j)]}) \tag{27}$$

Note that $\boldsymbol{z}^{i-1}$ and $\boldsymbol{z}^i$ only differ on the ith dimension, and different dimensions of $\boldsymbol{u}$ are independent. Consider the ith element of the summation in Eq.(27) 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

$$\mathbb{E}_{\boldsymbol{u}} \frac{\sigma(|\phi_j|)}{2} (f_{\mathbf{X},\boldsymbol{y}}(\boldsymbol{z}^{i-1}) - f_{\mathbf{X},\boldsymbol{y}}(\boldsymbol{z}^i))(\mathbf{1}_{[u_j > \sigma(-\phi_j)]} - \mathbf{1}_{[u_j < \sigma(\phi_j)]})$$

$$= \frac{\sigma(|\phi_j|)}{2} \mathbb{E}_{\boldsymbol{u}_{-i}} \left[\mathbb{E}_{u_i} [f_{\mathbf{X},\boldsymbol{y}}(\boldsymbol{z}^{i-1}) - f_{\mathbf{X},\boldsymbol{y}}(\boldsymbol{z}^i)](\mathbf{1}_{[u_j > \sigma(-\phi_j)]} - \mathbf{1}_{[u_j < \sigma(\phi_j)]}) | \boldsymbol{u}_{-i}] \right]$$

$$= \frac{\sigma(|\phi_j|)}{2} \mathbb{E}_{\boldsymbol{u}_{-i}} \left[(\mathbf{1}_{[u_j > \sigma(-\phi_j)]} - \mathbf{1}_{[u_j < \sigma(\phi_j)]}) \mathbb{E}_{u_i} [f_{\mathbf{X},\boldsymbol{y}}(\boldsymbol{z}^{i-1}) - f_{\mathbf{X},\boldsymbol{y}}(\boldsymbol{z}^i)] | \boldsymbol{u}_{-i}] \right]$$

$$= \mathbb{E}_{\boldsymbol{u}_{-i}} \left[(\mathbf{1}_{[u_j > \sigma(-\phi_j)]} - \mathbf{1}_{[u_j < \sigma(\phi_j)]}) \left(\int_0^{\sigma(-\phi_i)} (f_{\mathbf{X},\boldsymbol{y}}(\boldsymbol{z}^{i-1} | z_i = 0) - f_{\mathbf{X},\boldsymbol{y}}(\boldsymbol{z}^i | z_i = 1)) du_i \right. \right.$$

16

$$+ \int_{\sigma(\phi_i)}^{1} \left(f_{\mathbf{X},\boldsymbol{y}}(\boldsymbol{z}^{i-1}|z_i=1) - f_{\mathbf{X},\boldsymbol{y}}(\boldsymbol{z}^{i}|z_i=0)\right)du_i\Big)|\boldsymbol{u}_{-i}\Big]\frac{\sigma(|\phi_j|)}{2}$$

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

$$+ \left(f_{\mathbf{X},\boldsymbol{y}}(\boldsymbol{z}^{i-1}|z_i=1) - f_{\mathbf{X},\boldsymbol{y}}(\boldsymbol{z}^{i}|z_i=0)\right)(1-\sigma(\phi_i))\Big)|\boldsymbol{u}_{-i}\Big]\frac{\sigma(|\phi_j|)}{2}$$

$$= 0.$$

Whereas for $i = j$, we have

$$\frac{\sigma(|\phi_j|)}{2}\mathbb{E}_{\boldsymbol{u}}(f_{\mathbf{X},\boldsymbol{y}}(\boldsymbol{z}^{j-1}) - f_{\mathbf{X},\boldsymbol{y}}(\boldsymbol{z}^{j}))(\mathbf{1}_{[u_j>\sigma(-\phi_j)]} - \mathbf{1}_{[u_j<\sigma(\phi_j)]})$$

$$= \frac{\sigma(\phi_j)}{2}\mathbb{E}_{\boldsymbol{u}_{-j}}\left[\mathbb{E}_{u_j}[(f_{\mathbf{X},\boldsymbol{y}}(\boldsymbol{z}^{j-1}) - f_{\mathbf{X},\boldsymbol{y}}(\boldsymbol{z}^{j}))(\mathbf{1}_{[u_j>\sigma(-\phi_j)]} - \mathbf{1}_{[u_j<\sigma(\phi_j)]})|\boldsymbol{u}_{-j}]\right]$$

$$= \frac{\sigma(\phi_j)}{2}\mathbb{E}_{\boldsymbol{u}_{-j}}\left[\left(\int_{0}^{\sigma(-\phi_j)}(f_{\mathbf{X},\boldsymbol{y}}(\boldsymbol{z}^{j-1}|z_j=0) - f_{\mathbf{X},\boldsymbol{y}}(\boldsymbol{z}^{j}|z_j=1))(-1)du_j\right.\right.$$

$$\left.\left.+ \int_{\sigma(\phi_j)}^{1}(f_{\mathbf{X},\boldsymbol{y}}(\boldsymbol{z}^{j-1}|z_j=1) - f_{\mathbf{X},\boldsymbol{y}}(\boldsymbol{z}^{j}|z_j=0))du_j\right)\Big|\boldsymbol{u}_{-j}\right]$$

$$= \mathbb{E}_{\boldsymbol{u}_{-j}}\left[\sigma(\phi_j)(1-\sigma(\phi_j))[f_{\mathbf{X},\boldsymbol{y}}(\boldsymbol{z}^{j-1}|z_j=1) - f_{\mathbf{X},\boldsymbol{y}}(\boldsymbol{z}^{j}|z_j=0)]\right]$$

$$= \pi_j(1-\pi_j)\mathbb{E}_{\boldsymbol{u}}[\Delta_{\boldsymbol{z},j}f].$$

The same derivation holds true when the logit $\phi_i \leq 0$. Hence for each dimension there is only one non-zero element in the summation of Eq.(27). Rewriting the result in vector form proves the lemma. $\qquad\square$

Lemma 1 shows that the gradient is closely related to $\Delta_{\boldsymbol{z}}f$, whose randomness comes from latent variable $\boldsymbol{u}$ and data $(\mathbf{X}, \boldsymbol{y})$. We first look at the effect of changing one element of $\boldsymbol{z}$. Then, we consider the gradient behavior in expectation. Now, the following result establishes concentration behaviors of $\Delta_{\boldsymbol{z}}f$ based on the value of $\|z\|_0$.

Lemma 2. *Let* $\boldsymbol{z}, \tilde{\boldsymbol{z}} \in \{0,1\}^p$ *be such that they only differ on one dimension, i.e.,* $\boldsymbol{z}_k = 0, \tilde{\boldsymbol{z}}_k = 1, \boldsymbol{z}_j = \tilde{\boldsymbol{z}}_j, \forall j \neq k$ *for some* k. *We denote* $S = \|\boldsymbol{\beta}^*\|_0$ *and*

$$\Delta_{\boldsymbol{z},k}f = f_{\mathbf{X},\boldsymbol{y}}(\tilde{\boldsymbol{z}}) - f_{\mathbf{X},\boldsymbol{y}}(\boldsymbol{z}) \tag{28}$$

Then, there exist positive universal constants C_1 *and* C_2 *such that the following holds*

(a) When $\|\boldsymbol{z}\|_0 \leq n-1$, we have

$$\left| \Delta_{\boldsymbol{z},k} f - \left(\lambda - \frac{(n - \|\boldsymbol{z}\|_0 - 1)(\beta_k^*)^2 + \|\boldsymbol{\beta}_{-\boldsymbol{z}}^*\|_2^2 + \sigma^2}{n} \right) \right| \leq C\sqrt{\frac{\log(1/\delta)}{n}},$$

with probability at least $1 - 4\delta$.

(b) When $\|\boldsymbol{z}\|_0 \geq n$, then

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

with probability one.

The proof of Lemma 2 is in Appendix A.3. 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 $\sum_{j=1}^p \sigma(\boldsymbol{\phi}_j) \leq (1-t)n$, for certain $t \in (0,1)$. If n is sufficiently large such that $\sqrt{\frac{p \log(n)}{2n^2}} \leq t$, $\frac{\|\boldsymbol{\beta}^*\|_2^2 + \sigma^2}{(n-1)\min_k (\beta_k^*)^2} \leq t$, then there exists $\lambda > 0$ such that*

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

Proof. By Lemma 1, we find that

$$\mathbb{E}_{\boldsymbol{u} \sim \prod_{j=1}^p \mathrm{Unif}(u_j;0,1)}[g_{\mathrm{U2G}}(\boldsymbol{u};\sigma(\boldsymbol{\phi}))_j]/(\pi_j(1-\pi_j))$$
$$= \mathbb{E}_{\boldsymbol{u}}[\Delta_{\boldsymbol{z},j} f]$$
$$= \mathbb{E}_{\boldsymbol{u}}[\Delta_{\boldsymbol{z},j} f | \|\boldsymbol{z}\|_0 < n]p(\|\boldsymbol{z}\|_0 < n) + \mathbb{E}_{\boldsymbol{u}}[\Delta_{\boldsymbol{z},j} f | \|\boldsymbol{z}\|_0 \geq n]p(\|\boldsymbol{z}\|_0 \geq n)$$
$$= \mathbb{E}_{\boldsymbol{u}}[\Delta_{\boldsymbol{z},j} f | \|\boldsymbol{z}\|_0 < n]p(\|\boldsymbol{z}\|_0 < n) + \lambda p(\|\boldsymbol{z}\|_0 \geq n),$$

which can be viewed as a mixture of two components. Here, $\boldsymbol{z} = \mathbf{1}_{[\boldsymbol{u} > 1 - \sigma(\boldsymbol{\phi})]}$, and $\Delta_{\boldsymbol{z}} f = (\Delta_{\boldsymbol{z},1} f, \cdots, \Delta_{\boldsymbol{z},p} f)$. Furthermore, the final equation is due to Lemma 2. Now, we consider cases whether the covariates are in the true active set separately.

1) $\{j : \beta_j^* = 0\}$: when $\|\boldsymbol{z}\|_0 < n$, by Lemma 2 if $\lambda \geq \frac{\|\boldsymbol{\beta}^*\|_2^2 + \sigma^2}{n}$, we have

$$\mathbb{E}_{\mathbf{X},\boldsymbol{y},\boldsymbol{u}}[\Delta_{\boldsymbol{z},j} f] = \lambda - \frac{\sigma^2}{n} - \mathbb{E}_{\boldsymbol{u}}[\frac{\|\boldsymbol{\beta}_{-\boldsymbol{z}}^*\|_2^2}{n}] > 0.$$

18

This holds for all $\boldsymbol{z}$, so

$$\mathbb{E}_{\mathbf{X},\boldsymbol{y},\boldsymbol{u}}[\Delta_{\boldsymbol{z},j}f] > \lambda p(\|\boldsymbol{z}\|_0 \geq n) > 0,$$

Therefore, $\mathbb{E}_{\mathbf{X},\boldsymbol{y},\boldsymbol{u}}[g_{\text{U2G}}(\boldsymbol{u};\sigma(\boldsymbol{\phi}))_j] > 0$ for all $j \notin \mathcal{A}$.

ii) $\{j : \beta_j^* \neq 0\}$**:** by Hoeffding's inequality of sub-Gaussian random variables, and by assumption $p \leq 2t^2n^2/\log(n)$, we have

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

Conditionally on $\|\boldsymbol{z}\|_0 < n$, we have

$$\mathbb{E}\left[\|\boldsymbol{z}\|_0 \Big| \|\boldsymbol{z}\|_0 < n\right] \leq \mathbb{E}[\|\boldsymbol{z}\|_0] = \sum_{j=1}^p \pi_j.$$

By Lemma 2 if $\lambda < \frac{n-1}{n}t\min_k (\beta_k^*)^2$, we have

$$\begin{aligned}
\mathbb{E}_{\mathbf{X},\boldsymbol{y},\boldsymbol{u}}[\Delta_{\boldsymbol{z},j}f] \leq & \lambda - \frac{(n - \sum_{j=1}^p \pi_j - 1)(\beta_j^*)^2}{n} - \frac{\left\|\boldsymbol{\beta}_{-\boldsymbol{z}}^*\right\|_2^2 + \sigma^2}{n} \\
\leq & \lambda - t(\beta_j^*)^2 - \frac{\sigma^2}{n} \\
\leq & -\frac{\lambda}{n-1} - \frac{\sigma^2}{n}.
\end{aligned}$$

Collecting the above results, we find that

$$\mathbb{E}_{\mathbf{X},\boldsymbol{y},\boldsymbol{u}}[\Delta_{\boldsymbol{z},j}f] \leq -\frac{\lambda}{n-1}\frac{n-1}{n} + \frac{\lambda}{n} - \frac{\sigma^2}{n+2} < 0.$$

Hence, $\mathbb{E}_{\mathbf{X},\boldsymbol{y},\boldsymbol{u}}[g_{\text{U2G}}(\boldsymbol{u};\sigma(\boldsymbol{\phi}))_j] < 0$ for all $j \in \mathcal{A}$.

Combining the two cases, by setting

$$\lambda \in \left(\frac{\|\boldsymbol{\beta}^*\|_2^2 + \sigma^2}{n}, \frac{n-1}{n}t\min_k (\beta_k^*)^2\right), \tag{29}$$

the proposition is proved. $\qquad\square$

Remark 2. *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

$$\varpi = \left(1 - \min\left\{\sqrt{\frac{p\log(n)}{2n^2}}, \frac{\|\boldsymbol{\beta}^*\|_2^2 + \sigma^2}{(n-1)\min_k (\beta_k^*)^2}\right\}\right) n > S,$$

with initialization

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

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

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 in Eq. (4) as a special case. In this section, we show the proposed gradient estimators can solve the approximate L_0 regularized regression by optimizing non-linear objective functions. In the Bayesian paradigm, the linear regression model can be expressed hierarchically with latent variables $\boldsymbol{\alpha} \in \mathbb{R}^p$ and $\boldsymbol{z} \in \{0,1\}^p$, which have a Gaussian and Bernoulli prior respectively. In particular, the model is given by:

$$y_i \sim \mathcal{N}(\boldsymbol{x}_i^\top(\boldsymbol{\alpha} \odot \boldsymbol{z}), \sigma^2), \;\; i \in [n]$$
$$\boldsymbol{\alpha} \sim \mathcal{N}(\boldsymbol{\alpha}; \mathbf{0}, \boldsymbol{\Sigma}_{\boldsymbol{\alpha}}), \quad z_j \sim \text{Bern}(\sigma(-\lambda_0)), \;\; j \in [p]. \tag{30}$$

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

$$p(\boldsymbol{\alpha}, \boldsymbol{z}; \lambda_0, \boldsymbol{\Sigma}_{\boldsymbol{\alpha}})\sigma(-\lambda_0)^{\|\boldsymbol{z}\|_0}(1 - \sigma(-\lambda_0))^{p-\|\boldsymbol{z}\|_0}(2\pi)^{-p/2}|\boldsymbol{\Sigma}_{\boldsymbol{\alpha}}|^{-1/2}\exp(-\tfrac{1}{2}\boldsymbol{\alpha}^\top\boldsymbol{\Sigma}_{\boldsymbol{\alpha}}^{-1}\boldsymbol{\alpha}). \tag{31}$$

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:

$$p(\boldsymbol{\beta}) = \prod_{j=1}^{p} \left[\sigma(\lambda_0)\delta_0 + (1 - \sigma(\lambda_0))\mathcal{N}(0, \sigma_{\alpha}^2)\right]. \tag{32}$$

With the likelihood and prior, the posterior distribution is

$$p(\boldsymbol{\alpha}, \boldsymbol{z} \,|\, \mathbf{X}, \boldsymbol{y}; \lambda_0, \boldsymbol{\Sigma}_{\boldsymbol{\alpha}}) \propto \exp\Big(- \tfrac{1}{2\sigma^2} \|\boldsymbol{y} - \mathbf{X}(\boldsymbol{\alpha} \odot \boldsymbol{z})\|_2^2 - \tfrac{1}{2}\boldsymbol{\alpha}^\top \boldsymbol{\Sigma}_{\boldsymbol{\alpha}}^{-1} \boldsymbol{\alpha} - \lambda_0 \|\boldsymbol{z}\|_0 \Big). \quad (33)$$

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

$$\min_{\boldsymbol{\alpha}, \boldsymbol{z}} \frac{1}{2} \|\boldsymbol{y} - \mathbf{X}(\boldsymbol{\alpha} \odot \boldsymbol{z})\|_2^2 + \frac{\sigma^2}{2\sigma_\alpha^2} \|\boldsymbol{\alpha}\|_2^2 + \sigma^2 \lambda_0 \|\boldsymbol{z}\|_0 . \quad (34)$$

Hence the MAP solution is equivalent to the linear regression solution with combined L_2 and L_0 penalties (Polson and Sun, 2019). It has been observed that adding additional L_2 penalty can improve the computational efficiency in practice (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. (34) is a combinatorial problem. To overcome the computational challenge, we resort to the variational inference (VI) to approximate the posterior distribution and MAP estimation. To be consistent with VI nomenclature, here we deviate from the notation in (5), and use $p(\boldsymbol{z})$ as the prior, $q_{\boldsymbol{\phi}}(\boldsymbol{\alpha}, \boldsymbol{z})$ as the variational distribution parameterized by $\boldsymbol{\phi}$. The VI methods find an approximated posterior by minimizing the Kullback-Leibler (KL) divergence $D_{\mathrm{KL}}(q_{\boldsymbol{\phi}}(\boldsymbol{\alpha}, \boldsymbol{z})||p(\boldsymbol{\alpha}, \boldsymbol{z} \,|\, \mathbf{X}, \boldsymbol{y}))$. 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

$$\begin{aligned}
\mathcal{L}(\boldsymbol{\phi}) &= \log p(\boldsymbol{y}|\mathbf{X}) - D_{\mathrm{KL}}(q_{\boldsymbol{\phi}}(\boldsymbol{\alpha}, \boldsymbol{z})||p(\boldsymbol{\alpha}, \boldsymbol{z}|\mathbf{X}, \boldsymbol{y})) \\
&= \mathbb{E}_{q_{\boldsymbol{\phi}}(\boldsymbol{\alpha}, \boldsymbol{z})} \log \big[p(\boldsymbol{y}|\mathbf{X}, \boldsymbol{z}, \boldsymbol{\alpha}) p(\boldsymbol{\alpha}, \boldsymbol{z}; \lambda_0, \sigma_\alpha^2) / q_{\boldsymbol{\phi}}(\boldsymbol{\alpha}, \boldsymbol{z}) \big].
\end{aligned} \quad (35)$$

Due to the limited expressiveness of the variational distribution and the zero-forcing property of the KL divergence, variational methods often underestimate the posterior uncertainty. Recent analysis, however, provides theoretical guarantees to the accuracy of point estimation by variational methods. For specific models such as the 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 highly accurate, as shown in Section 6.

To further improve the accuracy, we tighten the ELBO by marginalizing out the latent variable $\boldsymbol{\alpha}$. The gap between ELBO and marginal likelihood is equivalent to the KL divergence from the variational distribution to the posterior. With the chain rule of KL divergence, this gap can be decomposed as

$$D_{\mathrm{KL}}(q_{\boldsymbol{\phi}}(\boldsymbol{\alpha}, \boldsymbol{z})||p(\boldsymbol{\alpha}, \boldsymbol{z}|\mathbf{X}, \boldsymbol{y})) = D_{\mathrm{KL}}(q_{\boldsymbol{\phi}}(\boldsymbol{z})||p(\boldsymbol{z}|\mathbf{X}, \boldsymbol{y})) + \mathbb{E}_{q(\boldsymbol{z})} D_{\mathrm{KL}}(q(\boldsymbol{\alpha}|\boldsymbol{z})||p(\boldsymbol{\alpha}|\mathbf{X}, \boldsymbol{z}, \boldsymbol{y})). \quad (36)$$

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. (36) as 0. In this way, we get a tight ELBO as

$$\mathcal{L}(\boldsymbol{\phi}) = \mathbb{E}_{q_{\boldsymbol{\phi}}(\boldsymbol{z})} \log \left[p(\boldsymbol{y}|\mathbf{X}, \boldsymbol{z}; \sigma_{\alpha}^2) p(\boldsymbol{z}; \lambda_0) / q_{\boldsymbol{\phi}}(\boldsymbol{z}) \right]. \quad (37)$$

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

$$p(\boldsymbol{y}|\mathbf{X}, \boldsymbol{z}; \sigma_{\alpha}^2) = \mathcal{N}(\boldsymbol{y}; \boldsymbol{0}, \mathbf{X}(\sigma_{\alpha}^2 \mathbf{I} \odot (\boldsymbol{z}\boldsymbol{z}^\top))\mathbf{X}^\top + \sigma^2 \mathbf{I}_n). \quad (38)$$

Setting $f(\boldsymbol{z}) = \log[p(\boldsymbol{y}|\mathbf{X}, \boldsymbol{z}; \sigma_{\alpha}^2) p(\boldsymbol{z}; \lambda)/q_{\boldsymbol{\phi}}(\boldsymbol{z})]$, the objective in Eq. (37) can be considered as a special case of the general optimization objective in Section 3.2. 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

Let $\widehat{\boldsymbol{\beta}}$ denote the estimated coefficients, $\boldsymbol{\beta}^*$ the true coefficients, and $(\boldsymbol{x}, y)$ a test sample. Here, $x \sim \mathcal{N}(0, \Sigma)$. 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; Hazimeh and Mazumder, 2018). 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 $\mathrm{SNR}_d/(1 + \mathrm{SNR}_d)$:

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

For the evaluation of sparse pattern recovery, we consider the *precision, recall, F1* scores and the *size of estimated active set* (Linero, 2018), given by prec = TP/(TP + FP), rec = TP/(TP +FN), and F1 = 2 · prec · rec/(prec + rec), respectively, 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. To measure the level of information in data, the data SNR is defined as

$$\mathrm{SNR}_d := \mathrm{var}(\boldsymbol{x}^{\top}\boldsymbol{\beta})/\mathrm{var}(\epsilon) = \boldsymbol{\beta}^{*\top}\Sigma\boldsymbol{\beta}^*/\sigma^2. \tag{39}$$

We compare the proposed methods with the representatives from a set of sparse variable selection methods. We choose LASSO (Tibshirani, 1996) as convex penalty regularized method, SCAD (Fan and Li, 2001) as a non-convex penalty regularized method, and MIO as a best subset selection method (Bertsimas et al., 2016). Throughout the experiments, if not specified, we set the number of Monte Carlo samples for continuous reformulation methods 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 $H_j = -p_j \log(p_j), j \in [p]$ and stop the training when the average of the 1% largest entropy is below 0.1.

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 $\sigma = 3$. The 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 of best subset selection with MIO (Bertsimas et al., 2016). The K in Eq. (2) for MIO is set as 3 which is the oracle sparsity level. MIO is run for sufficient time to reach the optimal solution. For the continuous reformulation methods, we use a constant step-size in SGD as $0.02/\lambda$.

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 of L_0 penalty, when the active set recovery performances are similar, the best subset methods achieve better predictive accuracy. When SNR_d is low, continuous reformulation methods have better out-of-sample prediction and active set recovery than MIO. In both SNR settings and for both objectives (4) and (37), 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

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
$n = 60, p = 200, \sigma = 1, \mathrm{SNR}_d = 21.3$							
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	0.995	1.000	0.997	3.02	0.003	1.059	0.952
REINFORCE	0.015	1.0	0.03	200.0	0.513	11.894	0.465
ARM_0	0.992	1.000	0.996	3.03	0.003	1.067	0.952
U2G	0.990	1.000	0.994	3.04	0.003	1.069	0.952
$\mathrm{ARM}_0(\mathrm{VI})$	0.954	1.000	0.974	3.19	0.005	1.103	0.950
U2G(VI)	0.950	1.000	0.971	3.21	0.005	1.107	0.950
$n = 60, p = 200, \sigma = 3, \mathrm{SNR}_d = 2.4$							
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	0.015	1.0	0.03	200.0	0.749	2.768	0.176
ARM_0	0.853	0.86	0.841	3.15	0.107	1.252	0.627
U2G	0.85	0.853	0.834	3.14	0.111	1.263	0.624
$\mathrm{ARM}_0(\mathrm{VI})$	0.822	0.803	0.786	3.06	0.116	1.273	0.621
U2G(VI)	0.907	0.883	0.878	2.99	0.082	1.194	0.645

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 $(n, p, S) = (100, 1000, 10)$. The hyper-parameter λ is chosen so that the prediction error on a validation set is minimized. The value of λ falls in the range given by Eq. (29), which validates the theoretical analysis. For MIO, we set the S in Eq. (2) as 10, the oracle sparsity level. We set the maximum running time of MIO as 200 seconds which takes about 5-10 times longer than continuous reformulation methods to converge. From Table 3, the continuous reformulation 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 can get the number of estimated nonzero elements the same as what the S is set to be, but there are many false-positive and false-negative estimations, as shown by the precision and recall metrics. By contrast,
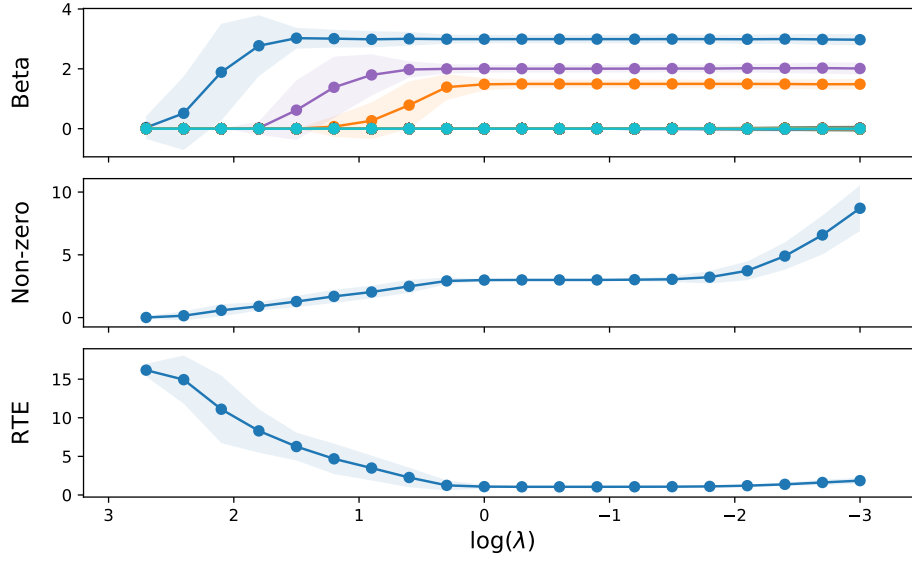
Figure 3: Regularized path for L_0-regularized regression estimated by U2G gradient, with $n = 60, p = 200, \sigma = 1$. The dotted curves are the mean of 100 independent trials and the shaded areas represent the standard deviation.

our methods can get the sparsity level close to the ground truth, achieving accurate sparsity recovery and hence low testing error. Since theoretically and empirically, the U2G estimator performs no worse than other gradient estimators in the proposed family, we will stick to the U2G estimator in the following experiments.

Experiment 3: Semi-synthetic data

We further benchmark our methods on the Prostate cancer dataset, a real-world microarray dataset (Singh et al., 2002). The regularization approaches are widely used for gene selection when analyzing microarray data, which is often high-dimensional with a large number of genes and a small number of samples (Liang et al., 2013). 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 that are maximally correlated (in absolute value) with the tumor type. We greedily choose five gene biomarkers with pairwise correlation in $(-0.7, 0.7)$ and create a semisynthetic data set $\boldsymbol{y} \sim \mathcal{N}(\mathbf{X}\boldsymbol{\beta}^*, \sigma^2 \mathbf{I})$, where the coefficients are one for the chosen covariates and zero for the rest. The σ^2 is set to let $\mathrm{SNR}_d = 5$.

We compare the continuous reformulation methods with LASSO and Elastic Net, as

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
$n = 100, p = 1000, \mathrm{SNR}_d = 7$							
Lasso	0.297	0.992	0.452	35.31	0.239	2.675	0.666
SCAD	0.55	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	0.01	1.0	0.02	1000.0	0.916	7.417	0.073
ARM_0	0.945	0.906	0.923	9.55	0.129	1.905	0.762
U2G	0.949	0.93	0.938	9.8	0.105	1.735	0.783
ARM_0(VI)	0.975	0.992	0.983	10.2	0.034	1.24	0.845
U2G(VI)	0.971	0.992	0.981	10.25	0.035	1.248	0.844
$n = 100, p = 1000, \mathrm{SNR}_d = 5$							
Lasso	0.306	0.971	0.458	34.25	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	0.01	1.0	0.02	1000.0	0.923	5.615	0.064
' ARM_0	0.906	0.906	0.904	10.01	0.154	1.771	0.705
U2G	0.9	0.908	0.903	10.13	0.155	1.778	0.704
ARM_0(VI)	0.954	0.966	0.959	10.11	0.076	1.379	0.77
U2G(VI)	0.952	0.968	0.959	10.16	0.074	1.369	0.772

shown in Table 4. We find that if there are high correlation among certain covariates in the true active set, the L_0-regularized regression tends to select only one of these highly correlated covariates. When the correlation is moderate, the continuous reformulation method can recover the true active set with high probability. In comparison, LASSO and Elastic Net select many spurious covariates. The consequential overfitting negatively impacts the prediction accuracy at new data, and results in high false positive rate.

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

	Precision	Recall	F1	Nonzero	RR	RTE	PVE
Lasso	0.275	0.98	0.423	19.6	0.071	1.354	0.774
SCAD	0.35	0.736	0.461	11.49	0.341	2.704	0.549
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

Experiment 4: Compressive sensing

We further test whether the L_0-based method improves the sparsity recovery in comparison with continuous relaxation methods 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 the transformed space with length $p = 1024$, where 10 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 $\boldsymbol{y}$. For continuous reformulation, 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 the continuous reformulation method, we use $K = 5$ Monte Carlo

Table 5: Results of the signal reconstruction, with $n = 500$, $p = 1000$, $S = 10$. L_0-based continuous reformulation method has the best performance in active set recovery and predictive accuracy, in both high and low SNR_d regimes.

	Precision	Recall	F1	Nonzero	RR	RTE	PVE
$n = 500, p = 1024, \sigma = 0.005$							
BP	0.009	1.000	0.019	1024	1e-3	508	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
$n = 500, p = 1024, \sigma = 0.1$							
BP	0.009	1.000	0.019	1024	0.298	298	0.702
BCS	0.029	1.000	0.057	336	2.380	2381	-
U2G	1.000	1.000	1.000	10	0.014	15.1	0.985
U2G (VI)	1.000	1.000	1.000	10	0.018	18.6	0.981

samples for the gradient estimation, and select hyper-parameter which produces lowest prediction error on a validation set. U2G takes around 20 - 30 seconds to converge when $n = 500$, $p = 1000$, and sparsity level $S = 10$. The numerical results are summarized in Table 5. As shown in Appendix Figure 5, in high SNR_d regime, all three methods can reconstruct the sparse signal reasonably well, but the continuous reformulation method can identify the locations of true signal, while BP and BCS identify excessively large active set. Consequently, the L_0-regularized 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 select fairly dense active set yest most of the entries are small compared to several large signals, which recovers *weak sparsity*. In both high and low SNR_d regimes, the continuous reformulation methods accurately recover the strong sparsity in signal, improve the predictive accuracy of BP and BCS by several order of magnitudes.
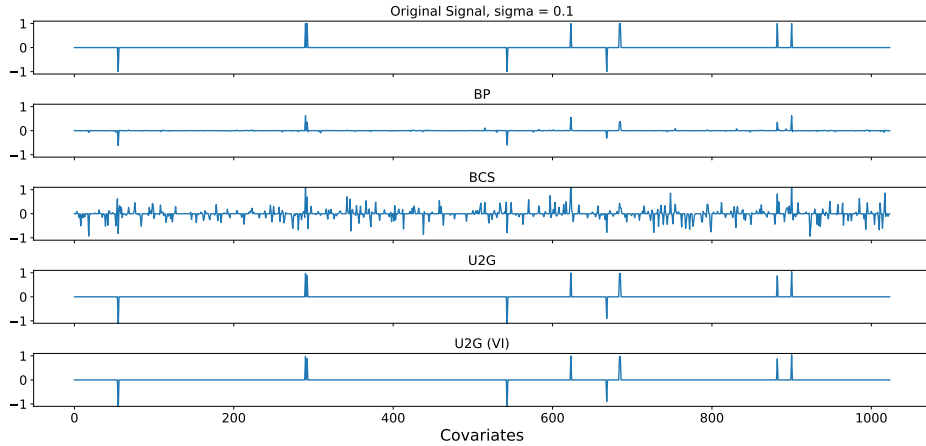
Figure 4: The reconstruction of signals when the SNR_d is low, with $n = 500$, $p = 1024$, $\sigma = 0.1$.

7   Discussion

In the paper, we propose a continuous reformulation to solve the exact best subset selection problem by gradient-based optimization. In order to efficiently solve the continuous reformulation problem 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 SNR and minimal variance, which can efficiently recover the true sparse pattern theoretically and empirically.

We now discuss a few directions arising naturally from our work. First, our continuous reformulation method is currently only for binary variables. In applications of latent variable models, the discrete latent variables often consist of more than two values (Jang et al., 2017; Maddison et al., 2016; Tucker et al., 2017; Grathwohl et al., 2017; Yin et al., 2019; Yue et al., 2020; Fan et al., 2020). 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. We leave the study of convergence property of our method when the design matrix has multi-collinearity as future work. In this case, it is intriguing to combine L_0 penalty with Tikhonov regularization such that the covariates are selected in groups. The proposed method can be further applied to the generalized linear models with L_0 penalty. 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 the continuous reformulation method (4) is locally weakly concave around $\boldsymbol{\pi}$, a fundamental trade-off between instability of SGD and its accelerated versions as well as their computational efficiency and statistical accuracy is important to study (Ho et al., 2020).

A Proof

In this appendix, we provide proofs for key remaining results in the paper.

A.1 Proof of Theorem 1

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

Secondly, we show the optimal solution of problem (4) is in the set of feasible solutions of 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). Notice all the points $\boldsymbol{z}$ in the active set of $p(\boldsymbol{z} \,|\, \boldsymbol{\pi}^*)$ would give the same objective value $f(\boldsymbol{z})$. Otherwise there exist $\boldsymbol{z}_i, \boldsymbol{z}_j \in \mathrm{supp}[p(\boldsymbol{z}|\boldsymbol{\pi}^*)]$ with $f(\boldsymbol{z}_i) < f(\boldsymbol{z}_j)$. By setting $\hat{\pi}_i = \pi_i^* + \pi_j^*$, $\hat{\pi}_j = 0$, $\hat{\pi}_k = \pi_k^*, k \neq i, j$, 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, when $p(\boldsymbol{z}|\boldsymbol{\pi}^*)$ is a point mass density

$\delta_{\boldsymbol{z}^*}$ with $\pi_j^* = z_j^*, j \in [p]$ and $\boldsymbol{\alpha}^* = g(\boldsymbol{z}^*)$, setting $\boldsymbol{z} = \boldsymbol{\pi}^*$, $\boldsymbol{\alpha} = \boldsymbol{\alpha}^*$ would give a feasible solution to problem (3) with the same objective value as problem (4). When $p(\boldsymbol{z}|\boldsymbol{\pi}^*)$ is not a point mass density, assuming $\text{supp}[p(\boldsymbol{z}|\boldsymbol{\pi}^*)] = \{\boldsymbol{z}_1, \cdots, \boldsymbol{z}_K\}$, we have $f(\boldsymbol{z}_1) = f(\boldsymbol{z}_2) = \cdots = f(\boldsymbol{z}_K)$. Then all $\{\boldsymbol{z}_k, g(\boldsymbol{z}_k)\}_{k \in [K]}$ are feasible solutions to problem (3) which give the same objective value as what the optimal solution gives in problem (4).

A.2   Proof of Proposition 1

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

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

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

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

For the second inequality, we find that

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

As a consequence, we obtain the conclusion of the proposition.

A.3   Proof of Lemma 2

The proof of the lemma is divided into three regimes of $\|\boldsymbol{z}\|_0$: $\|\boldsymbol{z}\|_0 \leq n-2$, $\|\boldsymbol{z}\|_0 = n-1$, and $\|\boldsymbol{z}\|_0 \geq n$.

When $\|\boldsymbol{z}\|_0 \leq n-2$: Under this setting, $\|\tilde{\boldsymbol{z}}\|_0 \leq n-1$. In order to simplify the presentation, 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 $\hat{\boldsymbol{\alpha}}_{\boldsymbol{z}} = (\mathbf{X}_{\boldsymbol{z}}^\top \mathbf{X}_{\boldsymbol{z}})^{-1} \mathbf{X}_{\boldsymbol{z}}^\top \boldsymbol{y}$. Given the definition of $\Delta_{\boldsymbol{z},k} f$, we obtain that

$$
\begin{aligned}
\Delta_{\boldsymbol{z},k} f &= \lambda + \frac{1}{n} \|\boldsymbol{y} - \mathbf{X}_{\tilde{\boldsymbol{z}}} \hat{\boldsymbol{\alpha}}_{\tilde{\boldsymbol{z}}}\|_2^2 - \frac{1}{n} \|\boldsymbol{y} - \mathbf{X}_{\boldsymbol{z}} \hat{\boldsymbol{\alpha}}_{\boldsymbol{z}}\|_2^2 \\
&= \lambda + \underbrace{\frac{1}{n} \|\boldsymbol{y} - \mathbf{X}_{\tilde{\boldsymbol{z}}} \hat{\boldsymbol{\alpha}}_{\tilde{\boldsymbol{z}}}\|_2^2 - \frac{1}{n} \|\boldsymbol{y} - \mathbf{X}_{\tilde{\boldsymbol{z}}} \boldsymbol{\beta}_{\tilde{\boldsymbol{z}}}^*\|_2^2}_{:=T_1} \\
&\quad + \underbrace{\frac{1}{n} \|\boldsymbol{y} - \mathbf{X}_{\tilde{\boldsymbol{z}}} \boldsymbol{\beta}_{\tilde{\boldsymbol{z}}}^*\|_2^2 - \frac{1}{n} \|\boldsymbol{y} - \mathbf{X}_{\boldsymbol{z}} \boldsymbol{\beta}_{\boldsymbol{z}}^*\|_2^2}_{:=T_2} + \underbrace{\frac{1}{n} \|\boldsymbol{y} - \mathbf{X}_{\boldsymbol{z}} \boldsymbol{\beta}_{\boldsymbol{z}}^*\|_2^2 - \frac{1}{n} \|\boldsymbol{y} - \mathbf{X}_{\boldsymbol{z}} \hat{\boldsymbol{\alpha}}_{\boldsymbol{z}}\|_2^2}_{:=T_3} .
\end{aligned}
\tag{40}
$$

Regarding term T_1 in (40), direct computation shows that

$$
T_1 = -\frac{1}{n} (\mathbf{X}_{-\tilde{\boldsymbol{z}}} \boldsymbol{\beta}_{-\tilde{\boldsymbol{z}}}^* + \boldsymbol{\epsilon})^\top P_{\tilde{\boldsymbol{z}}} (\mathbf{X}_{-\tilde{\boldsymbol{z}}} \boldsymbol{\beta}_{-\tilde{\boldsymbol{z}}}^* + \boldsymbol{\epsilon}) = -\frac{1}{n} \left\| P_{\tilde{\boldsymbol{z}}} (\mathbf{X}_{-\tilde{\boldsymbol{z}}} \boldsymbol{\beta}_{-\tilde{\boldsymbol{z}}}^* + \boldsymbol{\epsilon}) \right\|_2^2 . \tag{41}
$$

By Lemma 3, the rank of $X_{\tilde{\boldsymbol{z}}}$ equals $\|\tilde{\boldsymbol{z}}\|_0$ with probability one. Taking the expectation with T_1, a key observation is that for any $X_{\tilde{\boldsymbol{z}}}$ with full column rank,

$$
\mathbb{E}[T_1 | X_{\tilde{\boldsymbol{z}}}] = -\frac{(\sigma^2 + \|\boldsymbol{\beta}_{-\tilde{\boldsymbol{z}}}^*\|_2^2) \|\tilde{\boldsymbol{z}}\|_0}{n} . \tag{42}
$$

Therefore

$$
\mathbb{E}[T_1] = \mathbb{E}[\mathbb{E}[T_1 | X_{\tilde{\boldsymbol{z}}}]] = -\frac{(\sigma^2 + \|\boldsymbol{\beta}_{-\tilde{\boldsymbol{z}}}^*\|_2^2) \|\tilde{\boldsymbol{z}}\|_0}{n} . \tag{43}
$$

For any given $P_{\tilde{\boldsymbol{z}}}$, we consider its SVD decomposition as $P_{\tilde{\boldsymbol{z}}} = U_{\tilde{\boldsymbol{z}}} \Sigma_{\tilde{\boldsymbol{z}}} V_{\tilde{\boldsymbol{z}}}$ where $\Sigma_{\tilde{\boldsymbol{z}}} = \mathrm{diag}(1, \ldots, 1, 0, \ldots, 0)$ is a diagonal matrix with exactly $\|\tilde{\boldsymbol{z}}\|_0$ entries in the diagonal to be 1. Furthermore, $U_{\tilde{\boldsymbol{z}}}$ and $V_{\tilde{\boldsymbol{z}}}$ are orthonormal matrices. Direct algebra show that

$$
\left\| P_{\tilde{\boldsymbol{z}}} (\mathbf{X}_{-\tilde{\boldsymbol{z}}} \boldsymbol{\beta}_{-\tilde{\boldsymbol{z}}}^* + \boldsymbol{\epsilon}) \right\|_2^2 = \left\| \Sigma_{\tilde{\boldsymbol{z}}} V_{\tilde{\boldsymbol{z}}} (\mathbf{X}_{-\tilde{\boldsymbol{z}}} \boldsymbol{\beta}_{-\tilde{\boldsymbol{z}}}^* + \boldsymbol{\epsilon}) \right\|_2^2 .
$$

In order to simplify the presentation, we denote $W_{\tilde{\boldsymbol{z}}} = V_{\tilde{\boldsymbol{z}}} (\mathbf{X}_{-\tilde{\boldsymbol{z}}} \boldsymbol{\beta}_{-\tilde{\boldsymbol{z}}}^* + \boldsymbol{\epsilon})$. Conditioned on $V_{\tilde{\boldsymbol{z}}}$, $W_{\tilde{\boldsymbol{z}}} \sim \mathcal{N}(0, (\sigma^2 + \|\boldsymbol{\beta}_{-\tilde{\boldsymbol{z}}}^*\|_2^2) I_n)$. Furthermore, we find that

$$
\|\Sigma_{\tilde{\boldsymbol{z}}} W_{\tilde{\boldsymbol{z}}}\|_2^2 - \mathbb{E}[\|\Sigma_{\tilde{\boldsymbol{z}}} W_{\tilde{\boldsymbol{z}}}\|_2^2] = \sum_{i=1}^{\|\tilde{\boldsymbol{z}}\|_0} (W_{\tilde{\boldsymbol{z}}})_i^2 - \mathbb{E}[(W_{\tilde{\boldsymbol{z}}})_i^2],
$$

where $(W_{\tilde{z}})_i$ denotes the i-th component of vector $W_{\tilde{z}}$. By conditioning on $V_{\tilde{z}}$, since $(W_{\tilde{z}})_i^2/(\sigma^2 + \|\boldsymbol{\beta}^*_{-\tilde{z}}\|_2^2)$ are i.i.d chi-squared random variables for any $1 \leq i \leq \|\tilde{\boldsymbol{z}}\|_0$, an application of concentration bound for chi-squared random variables leads to

$$\mathbb{P}\left(\frac{1}{\|\tilde{\boldsymbol{z}}\|_0}\left|\sum_{i=1}^{\|\tilde{\boldsymbol{z}}\|_0}(W_{\tilde{z}})_i^2 - \mathbb{E}[(W_{\tilde{z}})_i^2]\right| \geq t\right) \leq 2\exp\left(-\frac{\|\tilde{\boldsymbol{z}}\|_0 t^2}{8(\sigma^2 + \|\boldsymbol{\beta}^*_{-\tilde{z}}\|_2^2)^2}\right).$$

The above result shows that by conditioning on $V_{\tilde{z}}$ or equivalently $X_{\tilde{z}}$, with probability $1 - \delta$, there exists a positive constant C such that

$$\mathbb{P}\left(|(T_1 - \mathbb{E}[T_1|X_{\tilde{z}}]| \geq C\frac{\sqrt{\|\tilde{\boldsymbol{z}}\|_0}}{n}\sqrt{\log(1/\delta)}\right) \leq \delta. \tag{44}$$

By Eqs. (42) and (43), we have $\mathbb{E}[T_1] = \mathbb{E}[T_1|X_{\tilde{z}}]$ for all $X_{\tilde{z}}$ with rank $\|\boldsymbol{z}\|_0$. Using the condition that $\|\tilde{\boldsymbol{z}}\|_0 \leq n - 1$, we arrive at

$$\mathbb{P}\left(|T_1 - \mathbb{E}[T_1]| \geq C\sqrt{\frac{\log(1/\delta)}{n}}\right) \leq \delta. \tag{45}$$

Regarding term T_3 in equation (40), with the similar argument as that of term T_1, we find that

$$\mathbb{P}\left(|T_3 - \mathbb{E}[T_3]| \geq C\sqrt{\frac{\log(1/\delta)}{n}}\right) \leq \delta, \tag{46}$$

where $\mathbb{E}[T_3] = \frac{(\sigma^2 + \|\boldsymbol{\beta}^*_{-z}\|_2^2)\|\boldsymbol{z}\|_0}{n}$.

For term T_2 in Eq. (40), direct calculation shows that

$$T_2 = (\beta_k^*)^2\frac{1}{n}\|X_k\|^2 + \frac{2}{n}\beta_k^*X_k^\top R_{\boldsymbol{z}}, \quad \mathbb{E}[T_2] = (\beta_k^*)^2$$

where $R_{\boldsymbol{z}} = \mathbf{X}_{-\tilde{z}}\boldsymbol{\beta}^*_{-\tilde{z}} + \boldsymbol{\epsilon}$. Since $X_k \sim \mathcal{N}(0, I_n)$, standard chi-squared bound yields that $\frac{1}{n}\|X_k\|^2 \leq C\sqrt{\log(1/\delta)/n}$ with probability at least $1 - \delta$ where C is some universal constant. Furthermore, since $R_{\boldsymbol{z}} \sim \mathcal{N}(0, (\sigma^2 + \|\boldsymbol{\beta}^*_{-\tilde{z}}\|_2^2)I_n)$, an application of Bernstein's inequality for sub-exponential random variables shows that

$$\mathbb{P}\left(\frac{|X_k^\top R_{\boldsymbol{z}}|}{n} \geq t\right) \leq 2\exp\left(-c\min\left(\frac{n^2t^2}{\sum_{i=1}^n\|X_{ki}(R_{\boldsymbol{z}})_i\|_{\psi_1}^2}, \frac{nt}{\max_i\|X_{ki}(R_{\boldsymbol{z}})_i\|_{\psi_1}}\right)\right),$$

33

where c is some universal constant. Here, $(R_{\boldsymbol{z}})_i$ denotes the i-th component of vector $R_{\boldsymbol{z}}$ and $\|.\|_{\psi_1}$ is the standard sub-exponential norm. From the property of sub-exponential norm, we have $\|X_{ki}(R_{\boldsymbol{z}})_i\|_{\psi_1} \leq \|X_{ki}\|_{\psi_2}\|(R_{\boldsymbol{z}})_i\|_{\psi_2} \leq C(\sigma^2 + \|\boldsymbol{\beta}^*_{-\tilde{\boldsymbol{z}}}\|_2^2)$ where $\|.\|_{\psi_2}$ is sub-Gaussian norm and C is some universal constant. Therefore, the above inequality becomes

$$\mathbb{P}\left(\frac{1}{n}|X_k^\top R_{\boldsymbol{z}}| \geq t\right) \leq 2\exp\left(-c\min\left(\frac{nt^2}{C^2(\sigma^2 + \|\boldsymbol{\beta}^*_{-\tilde{\boldsymbol{z}}}\|_2^2)^2}, \frac{nt}{C(\sigma^2 + \|\boldsymbol{\beta}^*_{-\tilde{\boldsymbol{z}}}\|_2^2)}\right)\right).$$

It demonstrates that for sufficiently large positive constant C_1, $\frac{1}{n}|X_k^\top R_{\boldsymbol{z}}| \leq C_1\sqrt{\log(1/\delta)/n}$ with probability at least $1 - \delta$. Collecting all the previous results, we obtain that

$$\mathbb{P}\left(|T_2 - (\beta_k^*)^2| \geq C\sqrt{\frac{\log(1/\delta)}{n}}\right) \leq 1 - 2\delta, \tag{47}$$

for some universal constant C. Putting the results from equations (45), (46), and (47) together, we find that

$$\left|\Delta_{\boldsymbol{z},k}f - \lambda + (\beta_k^*)^2 - \frac{(\beta_k^*)^2(\|\boldsymbol{z}\|_0 + 1) - \left\|\boldsymbol{\beta}^*_{-\boldsymbol{z}}\right\|_2^2 - \sigma^2}{n}\right| \leq \bar{C}\sqrt{\frac{\log(1/\delta)}{n}}$$

with probability at least $1 - 4\delta$ where $\bar{C}$ is some universal constant.

When $\|\boldsymbol{z}\|_0 = n - 1$: Under this setting, we have $\|\tilde{\boldsymbol{z}}\|_0 = n$. By Lemma 3, with probability one, the columns of $\mathbf{X}_{\tilde{\boldsymbol{z}}}$ span the whole space $\mathbb{R}^n$. It demonstrates that $\min_{\boldsymbol{\alpha}} \frac{1}{n}\|\boldsymbol{y} - \mathbf{X}_{\tilde{\boldsymbol{z}}}\boldsymbol{\alpha}\|_2^2 = 0$ with probability one. Hence, with probability one, we obtain that

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

With the similar argument as that of Eq. (45), we have

$$\mathbb{P}\left(\left|\frac{1}{n}\|(I_n - P_z)W_{\tilde{\boldsymbol{z}}}\|_2^2 - \frac{(\sigma^2 + \|\boldsymbol{\beta}^*_{-\boldsymbol{z}}\|_2^2)}{n}\right| \geq C\sqrt{\frac{\log(1/\delta)}{n}}\right) \leq \delta,$$

where C is some universal constant. Putting these results together, we find that

$$\left| \Delta_{\boldsymbol{z},k} f - \lambda + \frac{(\sigma^2 + \left\| \boldsymbol{\beta}^*_{-\boldsymbol{z}} \right\|_2^2)}{n} \right| \leq C \sqrt{\frac{\log(1/\delta)}{n}},$$

with probability at least $1 - \delta$.

When $\|\boldsymbol{z}\|_0 \geq n$: Under this setting, we have $\|\tilde{\boldsymbol{z}}\|_0 \geq n + 1$. By Lemma 3, with probability one, we have $\min_{\boldsymbol{\alpha}} \frac{1}{n} \; \|\boldsymbol{y} - \mathbf{X}_{\boldsymbol{z}} \boldsymbol{\alpha}\|_2^2 = 0$ and $\min_{\boldsymbol{\alpha}} \frac{1}{n} \; \|\boldsymbol{y} - \mathbf{X}_{\tilde{\boldsymbol{z}}} \boldsymbol{\alpha}\|_2^2 = 0$. Therefore, with probability one, we obtain that

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

Collecting the results from the three cases of $\|\boldsymbol{z}\|_0$, we obtain the conclusion.

B   Auxiliary Lemmas

The following lemma describes the linear independence between random Gaussian vectors. Closely following the proof in Tao (2008), which contains a thorough discussion on the singularity of random matrix ensembles, we have the following lemma:

Lemma 3. *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 $V_j := \text{span}(X_1, \cdots, X_{j-1})$. 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
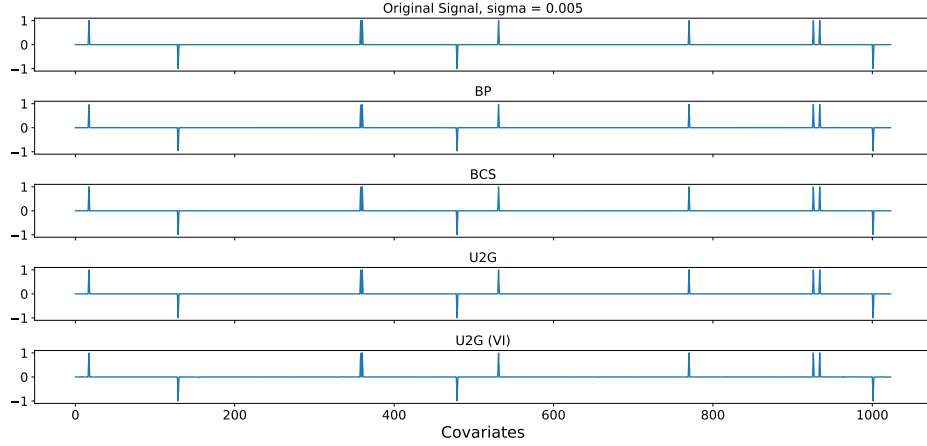
$$p(X_j \in V_j | X_1, \cdots, X_{j-1}) = 0$$

Figure 5: The reconstruction of signals when the SNR_d is high with $n = 500$, $p = 1024$, $\sigma = 0.005$.

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. $\qquad\qquad\square$

C Additional Result

In this section, we provide additional experimental results for compressive sensing. As shown in Figure 5, when SNR_d is high, all methods can recover the strong sparsity in signal. As shown in Figure 4, however, when SNR_d is low, methods with L_1 penalty cannot recover strong sparsity but our method with L_0 penalty can.

References

Akaike, H. (1974). A new look at the statistical model identification. In *Selected Papers of Hirotugu Akaike*, pp. 215–222. Springer. (Cited on page 2.)

Akaike, H. (1998). Information theory and an extension of the maximum likelihood

principle. In *Selected papers of Hirotugu Akaike*, pp. 199–213. Springer. (Cited on page 2.)

Barber, R. F. and E. J. Candès (2015). Controlling the false discovery rate via knockoffs. *The Annals of Statistics 43*(5), 2055–2085. (Cited on page 3.)

Beale, E., M. Kendall, and D. Mann (1967). The discarding of variables in multivariate analysis. *Biometrika 54*(3-4), 357–366. (Cited on page 2.)

Belloni, A. and V. Chernozhukov (2013). Least squares after model selection in high-dimensional sparse models. *Bernoulli 19*(2), 521–547. (Cited on page 3.)

Bertsimas, D., A. King, and R. Mazumder (2016). Best subset selection via a modern optimization lens. *The Annals of Statistics 44*(2), 813–852. (Cited on pages 4, 22, 23, 24, 25, and 26.)

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. (Cited on page 21.)

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. (Cited on page 21.)

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*. (Cited on page 9.)

Candès, E. J. and Y. Plan (2009). Near-ideal model selection by ℓ_1 minimization. *The Annals of Statistics 37*(5A), 2145–2177. (Cited on page 3.)

Carvalho, C. M., N. G. Polson, and J. G. Scott (2010). The horseshoe estimator for sparse signals. *Biometrika 97*(2), 465–480. (Cited on page 28.)

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. (Cited on page 2.)

Chen, S. S., D. L. Donoho, and M. A. Saunders (2001). Atomic decomposition by basis pursuit. *SIAM review 43*(1), 129–159. (Cited on page 28.)

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.* (Cited on page 9.)

Dicker, L., B. Huang, and X. Lin (2013). Variable selection and estimation with the seamless-l0 penalty. *Statistica Sinica 23*, 929–962. (Cited on page 3.)

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. (Cited on page 2.)

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. (Cited on pages 3, 23, and 24.)

Fan, J. and J. Lv (2010). A selective overview of variable selection in high dimensional feature space. *Statistica Sinica 20*(1), 101. (Cited on pages 2, 3, and 4.)

Fan, X., Y. Zhang, Z. Wang, and M. Zhou (2020). Adaptive correlated Monte Carlo for contextual categorical sequence generation. In *International Conference on Learning Representations.* (Cited on page 30.)

Foster, D. P. and E. I. George (1994). The risk inflation criterion for multiple regression. *The Annals of Statistics 22*(4), 1947–1975. (Cited on page 3.)

Frank, L. E. and J. H. Friedman (1993). A statistical view of some chemometrics regression tools. *Technometrics 35*(2), 109–135. (Cited on page 3.)

Friedman, J., T. Hastie, and R. Tibshirani (2001). *The Elements of Statistical Learning*, Volume 1. Springer series in statistics New York. (Cited on pages 2 and 4.)

Friedman, J., T. Hastie, and R. Tibshirani (2010). Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software 33*(1), 1. (Cited on page 24.)

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

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

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. (Cited on page 24.)

George, E. I. and R. E. McCulloch (1993). Variable selection via gibbs sampling. *Journal of the American Statistical Association 88*(423), 881–889. (Cited on page 5.)

Grathwohl, W., D. Choi, Y. Wu, G. Roeder, and D. Duvenaud (2017). Backpropagation through the void: Optimizing control variates for black-box gradient estimation. In *ICLR*. (Cited on page 30.)

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. (Cited on page 3.)

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. (Cited on page 24.)

Hastie, T., R. Tibshirani, and R. J. Tibshirani (2017). Extended comparisons of best subset selection, forward stepwise selection, and the lasso. (Cited on pages 4, 22, and 25.)

Hazimeh, H. and R. Mazumder (2018). Fast best subset selection: Coordinate descent and local combinatorial optimization algorithms. (Cited on pages 4, 21, and 22.)

Ho, N., K. Khamaru, R. Dwivedi, M. J. Wainwright, M. I. Jordan, and B. Yu (2020). Instability, computational efficiency and statistical accuracy. (Cited on page 30.)

Ishwaran, H., J. S. Rao, et al. (2005). Spike and slab variable selection: frequentist and bayesian strategies. *The Annals of Statistics 33*(2), 730–773. (Cited on page 5.)

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. (Cited on page 30.)

Ji, S., Y. Xue, and L. Carin (2008). Bayesian compressive sensing. *IEEE Transactions on signal processing 56*(6), 2346. (Cited on page 28.)

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. (Cited on page 3.)

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. (Cited on page 2.)

Kool, W., H. van Hoof, and M. Welling (2019). Buy 4 reinforce samples, get a baseline for free! (Cited on page 13.)

Liang, Y., C. Liu, X.-Z. Luan, K.-S. Leung, T.-M. Chan, Z.-B. Xu, and H. Zhang (2013). Sparse logistic regression with a l 1/2 penalty for gene selection in cancer classification. *BMC bioinformatics 14*(1), 198. (Cited on page 26.)

Linero, A. R. (2018). Bayesian regression trees for high-dimensional prediction and variable selection. *Journal of the American Statistical Association 113*(522), 626–636. (Cited on page 23.)

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. (Cited on pages 3 and 21.)

Maddison, C. J., A. Mnih, and Y. W. Teh (2016). The concrete distribution: A continuous relaxation of discrete random variables. In *ICLR*. (Cited on page 30.)

Mallat, S. G. and Z. Zhang (1993). Matching pursuits with time-frequency dictionaries. *IEEE Transactions on Signal Processing 41*(12), 3397–3415. (Cited on page 2.)

Mitchell, T. J. and J. J. Beauchamp (1988). Bayesian variable selection in linear regression. *Journal of the american statistical association 83*(404), 1023–1032. (Cited on page 5.)

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

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. (Cited on page 21.)

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. (Cited on page 2.)

Polson, N. G. and L. Sun (2019). Bayesian l 0-regularized least squares. *Applied Stochastic Models in Business and Industry 35*(3), 717–731. (Cited on pages 4, 5, and 21.)

Ross, S. M. (2014). *Introduction to probability models*. Academic press. (Cited on page 14.)

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

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. (Cited on page 3.)

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. (Cited on page 26.)

Tao, T. (2008). Singularity and determinant of random matrices. (Cited on page 35.)

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological) 58*(1), 267–288. (Cited on page 23.)

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. (Cited on page 30.)

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. (Cited on page 3.)

Wainwright, M. J. (2019). *High-dimensional statistics: A non-asymptotic viewpoint*, Volume 48. Cambridge University Press. (Cited on page 28.)

Wang, Y. and D. M. Blei (2018). Frequentist consistency of variational Bayes. *Journal of the American Statistical Association 114*(527), 1147–1161. (Cited on page 21.)

Williams, R. J. (1992). Simple statistical gradient-following algorithms for connectionist reinforcement learning. In *Reinforcement Learning*, pp. 5–32. Springer. (Cited on page 7.)

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*. (Cited on page 21.)

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. (Cited on page 30.)

Yin, M. and M. Zhou (2019). ARM: Augment-REINFORCE-merge gradient for stochastic binary networks. In *International Conference on Learning Representations*. (Cited on pages 8, 9, and 14.)

Yue, Y., Y. Tang, M. Yin, and M. Zhou (2020). Discrete action on-policy learning with action-value critic. In *Artificial Intelligence and Statistics*. (Cited on pages 9 and 30.)

Zhang, A. Y. and H. H. Zhou (2017). Theoretical and computational guarantees of mean field variational inference for community detection. (Cited on page 21.)

Zhang, C.-H. (2010). Nearly unbiased variable selection under minimax concave penalty. *The Annals of Statistics 38*(2), 894–942. (Cited on page 3.)

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. (Cited on page 3.)

Zhao, P. and B. Yu (2006). On model selection consistency of lasso. *Journal of Machine learning research 7*(Nov), 2541–2563. (Cited on page 3.)

Zhou, M., H. Chen, L. Ren, G. Sapiro, L. Carin, and J. W. Paisley (2009). Nonparametric Bayesian dictionary learning for sparse image representations. In *Advances in neural information processing systems*, pp. 2295–2303. (Cited on page 5.)