On Computationally Tractable Selection of Experiments in Measurement-Constrained Regression Models

Yining Wang   YININGWA@CS.CMU.EDU
Adams Wei Yu   WEIYU@CS.CMU.EDU
Aarti Singh   AARTISINGH@CMU.EDU
Machine Learning Department, School of Computer Science
Carnegie Mellon University
Pittsburgh, PA 15213, USA

Abstract

We derive computationally tractable methods to select a small subset of experiment settings from a large pool of given design points. The primary focus is on linear regression models, while the technique extends to generalized linear models and Delta's method (estimating functions of linear regression models) as well. The algorithms are based on a continuous relaxation of an otherwise intractable combinatorial optimization problem, with sampling or greedy procedures as post-processing steps. Formal approximation guarantees are established for both algorithms, and numerical results on both synthetic and real-world data confirm the effectiveness of the proposed methods.

Keywords: optimal selection of experiments, A-optimality, computationally tractable methods, minimax analysis

1. Introduction

Despite the availability of large datasets, in many applications, collecting labels for all data points is not possible due to measurement constraints. We consider the problem of measurement-constrained regression where we are given a large pool of n data points but can only observe a small set of $k \ll n$ labels. Classical *experimental design* approaches in statistical literature Pukelsheim (1993) have investigated this problem, but the proposed solutions tend to be often combinatorial. In this work, we investigate computationally tractable methods for selecting data points to label from a given pool of design points in measurement-constrained regression.

Despite the simplicity and wide applicability of OLS, in practice it may not be possible to obtain the full n-dimensional response vector y due to measurement constraints. It is then a common practice to select a small subset of rows (e.g., $k \ll n$ rows) in X so that the statistical efficiency of regression on the selected subset of design points is maximized. Compared to the classical *experimental design* problem (Pukelsheim, 1993) where X can be freely designed, in this work we consider the setting where the selected design points must come from an existing (finite) design pool X.

Below we list three example applications where such measurement constraints are relevant:

Example 1 (Material synthesis) Material synthesis experiments are time-consuming and expensive, whose results are sensitive to experimental setting features such as temperature, duration and reactant ratio. Given hundreds, or even thousands of possible experimental settings, it is important to select a handful of representative settings such that a model can be built with maximized statistical efficiency to predict quality of the outcome material from experimental features. In this paper we consider such an application of low-temperature microwave-assisted thin film crystallization (Reeja-Jayan et al., 2012; Nakamura et al., 2017) and demonstrate the effectiveness of our proposed algorithms.

Example 2 (CPU benchmarking) Central processing units (CPU) are vital to the performance of a computer system. It is an interesting statistical question to understand how known manufacturing parameters (clock period, cache size, etc.) of a CPU affect its execution time (performance) on benchmark computing tasks (Ein-Dor and Feldmesser, 1987). As the evaluation of real benchmark execution time is time-consuming and costly, it is desirable to select a subset of CPUs available in the market with diverse range of manufacturing parameters so that the statistical efficiency is maximized by benchmarking for the selected CPUs.

Example 3 (Wind speed prediction) In Chen et al. (2015) a data set is created to record wind speed across a year at established measurement locations on highways in Minnesota, US. Due to instrumentation constraints, wind speed can only be measured at intersections of high ways, and a small subset of such intersections is selected for wind speed measurement in order to reduce data gathering costs.

In this work, we primarily focus on the linear regression model (though extensions to generalized linear models and functions of linear models are also considered later)

$$y = X\beta_0 + \varepsilon,$$

where $X \in \mathbb{R}^{n \times p}$ is the design matrix, $y \in \mathbb{R}^n$ is the response and $\varepsilon \sim \mathcal{N}_n(0, \sigma^2 I_n)$ are homoscedastic Gaussian noise with variance σ^2. β_0 is a p-dimensional regression model that one wishes to estimate. We consider the "large-scale, low-dimensional" setting where both $n, p \to \infty$ and $p < n$, and X has full column rank. A common estimator is the *Ordinary Least Squares* (OLS) estimator:

$$\widehat{\beta}^{\mathrm{ols}} = \mathrm{argmin}_{\beta \in \mathbb{R}^p} \|y - X\beta\|_2^2 = (X^\top X)^{-1} X^\top y.$$

The mean square error of the estimated regression coefficients $\mathbb{E}\|\widehat{\beta}^{\mathrm{ols}} - \beta_0\|_2^2 = \sigma^2 \mathrm{tr}\left[\left(X^\top X\right)^{-1}\right]$. Under measurement constraints, it is well-known that the statistically optimal subset S^* for estimating the regression coefficients is given by the *A-optimality* criterion (Pukelsheim, 1993):

$$S^* = \mathrm{argmin}_{S \subseteq [n], |S| \leq k} \mathrm{tr}\left[\left(X_S^\top X_S\right)^{-1}\right]. \tag{1}$$

Despite the statistical optimality of Eq. (1), the optimization problem is combinatorial in nature and the optimal subset S^* is difficult to compute. A brute-force search over all possible subsets of size k requires $O(n^k k^3)$ operations, which is infeasible for even moderately sized designs X.

In this work, we focus on **computationally tractable** methods for experiment selection that achieve near-optimal statistical efficiency in linear regression models. We consider two experiment selection models: the *with replacement* model where each design point (row of X) can be selected more than once with independent noise involved for each selection, and the *without replacement* model where distinct row subsets are required. We propose two computationally tractable algorithms: one sampling based algorithm that achieves $O(1)$ approximation of the statistically optimal solution for the with replacement model and, when Σ^* is well-conditioned, the algorithm also works for the without replacement model. In the "soft budget" setting $|S| = O_{\mathbb{P}}(k)$, the approximation ratio can be further improved to $1 + O(\epsilon)$. We also propose a greedy method that achieves $O(p^2/k)$ approximation for the without replacement model regardless of the conditioning of the design pool X.

2. Problem formulation and backgrounds

We first give a formal definition of the experiment selection problem in linear regression models:

Definition 1 (experiment selection problem in linear regression models) *Let X be a known $n \times p$ design matrix with full column rank and k be the subset budget, with $p \leq k \leq n$. An experiment selection problem aims to find a subset $S \subseteq [n]$ of size k, either deterministically or randomly, then observes $y_S = X_S \beta_0 + \widetilde{\varepsilon}$, where each coordinate of $\widetilde{\varepsilon}$ is i.i.d. Gaussian random variable with zero mean and equal covariance, and then generates an estimate $\widehat{\beta}$ of the regression coefficients based on (X_S, y_S). Two types of experiment selection algorithms are considered:*

1. *With replacement: S is a multi-set which allows duplicates of indices. Note that fresh (independent) noise is imposed after, and independent of, experiment selection, and hence duplicate design points have independent noise. We use $\mathcal{A}_1(k)$ to denote the class of all with replacement experiment selection algorithms.*

2. *Without replacement: S is a standard set that may not have duplicate indices. We use $\mathcal{A}_2(k)$ to denote the class of all without replacement experiment selection algorithms.*

As evaluation criterion, we consider the mean square error $\mathbb{E}\|\widehat{\beta} - \beta_0\|_2^2$, where $\widehat{\beta}$ is an estimator of β_0 with X_S and y_S as inputs. We study *computationally tractable* algorithms that approximately achieves the *minimax* rate of convergence over $\mathcal{A}_1(k)$ or $\mathcal{A}_2(k)$:

$$\inf_{A' \in \mathcal{A}_b(k)} \sup_{\beta_0 \in \mathbb{R}^p} \mathbb{E}\left[\|\widehat{\beta}_{A'} - \beta_0\|_2^2\right]. \tag{2}$$

Formally, we give the following definition:

---

1. "Additive" means that the statistical error of the resulting estimator cannot be bounded by a multiplicative factor of the minimax optimal error.
2. The leverage score sampling method in Ma et al. (2015) does not have rigorous approximation guarantees in terms of $\|\widehat{\beta} - \beta_0\|_2^2$ or $\|X\widehat{\beta} - X\beta_0\|_2^2$. However, the bounds in that paper establish that leverage score sampling can be worse or better than uniform sampling under different settings.

Table 1: Summary of approximation results. $\Sigma_* = X^\top \mathrm{diag}(\pi^*)X$, where π^* is the optimal solution of Eq. (4). Sampling based algorithms can be with or without replacement and are randomized, which succeed with at least 0.8 probability.

Algorithm	Model	Constraint	$C(n,p,k)$	Assumptions
Ma et al. (2015)	with rep.	additive[1]	-[2]	asymptotic
Avron and Boutsidis (2013)	with rep.	additive	$O(n/k)$	$k = \Omega(p)$
sampling	with rep.	$\|S\| = O_{\mathbb{P}}(k)$	$1 + O(\epsilon)$	$p \log k/k = O(\epsilon^2)$
sampling	without rep.	$\|S\| = O_{\mathbb{P}}(k)$	$1 + O(\epsilon)$	$\|\Sigma_*^{-1}\|_2 \kappa(\Sigma_*)\|X\|_\infty \log p = O(\epsilon^2)$
sampling	with rep.	$\|S\| \le k$	$O(1)$	$p \log k/k = O(1)$
sampling	without rep.	$\|S\| \le k$	$O(1)$	$\|\Sigma_*^{-1}\|_2 \kappa(\Sigma_*)\|X\|_\infty \log p = O(1)$
greedy	without rep.	$\|S\| \le k$	**Rigorous**: $1 + \frac{p(p+1)}{2(k-p+1)}$ **Conjecture**: $1 + O(\frac{p}{k-p})$	$k > p$, or $k = \Omega(p^2/\epsilon)$ if $(1+\epsilon)$-approximation desired

Definition 2 ($C(n,p,k)$**-approximate algorithm**) *Fix $n \ge k \ge p$ and $b \in \{1,2\}$. We say an algorithm (either deterministic or randomized) $A \in \mathcal{A}_b(k)$ is a $C(n,p,k)$-approximate algorithm if for any $X \in \mathbb{R}^{n \times p}$ with full column rank, A produces a subset $S \subseteq [n]$ with size k and an estimate $\widehat{\beta}_A$ in polynomial time such that, with probability at least 0.8,*

$$\sup_{\beta_0 \in \mathbb{R}^p} \mathbb{E}\left[\|\widehat{\beta}_A - \beta_0\|_2^2 \Big| X_S\right] \le C(n,p,k) \cdot \inf_{A' \in \mathcal{A}_b(k)} \sup_{\beta_0 \in \mathbb{R}^p} \mathbb{E}\left[\|\widehat{\beta}_{A'} - \beta_0\|_2^2\right]. \tag{3}$$

Here both expectations are taken over the randomness in the noise variables $\widetilde{\varepsilon}$ and the inherent randomness in A.

Table 1 gives an overview of approximation ratio $C(n,p,k)$ for algorithms proposed in this paper. We remark that the combinatorial A-optimality solution of Eq. (1) upper bounds the minimax risk (since minimaxity is defined over deterministic algorithms as well), hence the approximation guarantees also hold with respect to the combinatorial A-optimality objective.

2.1 Related work

There has been an increasing amount of work on fast solvers for the general least-square problem $\min_\beta \|y - X\beta\|_2^2$. Most of existing work along this direction (Woodruff, 2014; Dhillon et al., 2013; Drineas et al., 2011; Raskutti and Mahoney, 2015) focuses solely on the computational aspects and do not consider statistical constraints such as limited measurements of y. A convex optimization formulation was proposed in Davenport et al. (2015) for a constrained adaptive sensing problem, which is a special case of our setting, but without finite sample guarantees with respect to the combinatorial problem. In Horel et al. (2014) a computationally tractable approximation algorithm was proposed for the D-optimality criterion of the experimental design problem. However, the core idea in Horel et al. (2014) of pipage rounding an SDP solution (Ageev and Sviridenko, 2004) is not applicable in our problem because the objective function we consider in Eq. (4) is not submodular.

Popular subsampling techniques such as leverage score sampling (Drineas et al., 2008) were studied in least square and linear regression problems (Zhu et al., 2015; Ma et al., 2015; Chen et al., 2015). While computation remains the primary subject, measurement

constraints and statistical properties were also analyzed within the linear regression model (Zhu et al., 2015). However, none of the above-mentioned (computationally efficient) methods achieve near minimax optimal statistical efficiency in terms of estimating the underlying linear model β_0, since the methods can be worse than uniform sampling which has a fairly large approximation constant for general X. One exception is Avron and Boutsidis (2013), which proposed a greedy algorithm that achieves an error bound of estimation of β_0 that is within a multiplicative factor of the minimax optimal statistical efficiency. The multiplicative factor is however large and depends on the size of the full sample pool X, as we remark in Table 1 and also in Sec. 3.4.

Another related area is *active learning* (Chaudhuri et al., 2015; Hazan and Karnin, 2015; Sabato and Munos, 2014), which is a stronger setting where feedback from prior measurements can be used to guide subsequent data selection. Chaudhuri et al. (2015) analyzes an SDP relaxation in the context of active maximum likelihood estimation. However, the analysis in Chaudhuri et al. (2015) only works for the with replacement model and the two-stage feedback-driven strategy proposed in Chaudhuri et al. (2015) is not available under the experiment selection model defined in Definition 1 where no feedback is assumed.

2.2 Notations

For a matrix $A \in \mathbb{R}^{n \times m}$, we use $\|A\|_p = \sup_{x \neq 0} \frac{\|Ax\|_p}{\|x\|_p}$ to denote the induced p-norm of A. In particular, $\|A\|_1 = \max_{1 \leq j \leq m} \sum_{i=1}^n |A_{ij}|$ and $\|A\|_\infty = \max_{1 \leq i \leq n} \sum_{j=1}^m |A_{ij}|$. $\|A\|_F = \sqrt{\sum_{i,j} A_{ij}^2}$ denotes the Frobenius norm of A. Let $\sigma_1(A) \geq \sigma_2(A) \geq \cdots \geq \sigma_{\min(n,m)}(A) \geq 0$ be the singular values of A, sorted in descending order. The condition number $\kappa_2(A)$ is defined as $\kappa(A) = \sigma_1(A)/\sigma_{\min(n,m)}(A)$. For sequences of random variables X_n and Y_n, we use $X_n \xrightarrow{p} Y_n$ to denote X_n converges in probability to Y_n. We say $a_n \lesssim b_n$ if $\lim_{n \to \infty} |a_n/b_n| \leq 1$ and $a_n \gtrsim b_n$ if $\lim_{n \to \infty} |b_n/a_n| \leq 1$. For two d-dimensional symmetric matrices A and B, we write $A \preceq B$ if $u^\top(A - B)u \leq 0$ for all $u \in \mathbb{R}^d$, and $A \succeq B$ if $u^\top(A - B)u \geq 0$ for all $u \in \mathbb{R}^d$.

3. Methods and main results

We describe two computationally feasible algorithms for the experiment selection problem, both based on a continuous convex optimization problem. Statistical efficiency bounds are presented for both algorithms, with detailed proofs given in Sec. 7.

3.1 Continuous optimization and minimax lower bounds

We consider the following continuous optimization problem, which is a convex relaxation of the combinatorial A-optimality criterion of Eq. (1):

$$\pi^* = \operatorname{argmin}_{\pi \in \mathbb{R}^n} f(\pi; X) = \operatorname{argmin}_{\pi \in \mathbb{R}^p} \operatorname{tr}\left[\left(X^\top \operatorname{diag}(\pi) X\right)^{-1}\right], \qquad (4)$$

$$s.t. \quad \pi \geq 0, \quad \|\pi\|_1 \leq k,$$
$$\|\pi\|_\infty \leq 1 \quad \text{(only for the without replacement model)}.$$

5

Note that the $\|\pi\|_\infty \leq 1$ constraint is only relevant for the without replacement model and for the with replacement model we drop this constraint in the optimization problem. It is easy to verify that both the objective $f(\pi; X)$ and the feasible set in Eq. (4) are convex, and hence the global optimal solution π^* of Eq. (4) can be obtained using computationally tractable algorithms. In particular, we describe an SDP formulation and a practical projected gradient descent algorithm in Appendix B and D, both provably converge to the global optimal solution of Eq. (4) with running time scaling polynomially in n, p and k.

We first present two facts, which are proved in Sec. 7.

Fact 3.1 *Let π and π' be feasible solutions of Eq. (4) such that $\pi_i \leq \pi_i'$ for all $i = 1, \cdots, n$. Then $f(\pi; X) \geq f(\pi'; X)$, with equality if and only if $\pi = \pi'$.*

Fact 3.2 $\|\pi^*\|_1 = k$.

We remark that the inverse monotonicity of f in π implies second fact, which can potentially be used to understand sparsity of π^* in later sections.

The following theorem shows that the optimal solution of Eq. (4) lower bounds the minimax risk defined in Eq. (2). Its proof is placed in Sec. 7.

Theorem 3 *Let $f_1^*(k; X)$ and $f_2^*(k; X)$ be the optimal objective values of Eq. (4) for with replacement and without replacement, respectively. Then for $b \in \{1, 2\}$,*

$$\inf_{A \in \mathcal{A}_b(k)} \sup_{\beta_0 \in \mathbb{R}^p} \mathbb{E}\left[\|\widehat{\beta}_A - \beta_0\|_2^2\right] \geq \sigma^2 \cdot f_b^*(k; X). \tag{5}$$

Despite the fact that Eq. (4) is computationally tractable, its solution π^* is *not* a valid experiment selection algorithm under a measurement budget of k because there can be much more than k components in π^* that are not zero. In the following sections, we discuss strategies for choosing a subset $\widehat{S}$ of rows with $|\widehat{S}| = O_\mathbb{P}(k)$ (i.e., soft constraint) or $|\widehat{S}| \leq k$ (i.e., hard constraint), using the solution π^* to the above.

3.2 Sampling based experiment selection: soft size constraint

We first consider a weaker setting where *soft* constraint is imposed on the size of the selected subset $\widehat{S}$; in particular, it is allowed that $|\widehat{S}| = O(k)$ with high probability, meaning that a constant fraction of over-selection is allowed. The more restrictive setting of hard constraint is treated in the next section.

A natural idea of obtaining a valid subset S of size k is by *sampling* from a weighted row distribution specified by π^*. Let $\Sigma_* = X^\top \text{diag}(\pi^*) X$ and define distributions $p_j^{(1)}$ and $p_j^{(2)}$ for $j = 1, \cdots, n$ as

$$\begin{aligned} P^{(1)}: \quad & p_j^{(1)} = \pi_j^* x_j^\top \Sigma_*^{-1} x_j / p, && \text{with replacement;} \\ P^{(2)}: \quad & p_j^{(2)} = \pi_j^*/k, && \text{without replacement.} \end{aligned}$$

Note that both $\{p_j^{(1)}\}_{j=1}^n$ and $\{p_j^{(2)}\}_{j=1}^n$ sum to one because $\sum_{j=1}^n \pi_j^* = k$ and $\sum_{j=1}^n \pi_j^* x_j^\top \Sigma_*^{-1} x_j = \text{tr}((\sum_{j=1}^n \pi_j^* x_j x_j^\top)\Sigma_*^{-1}) = \text{tr}(\Sigma_* \Sigma_*^{-1}) = p$.
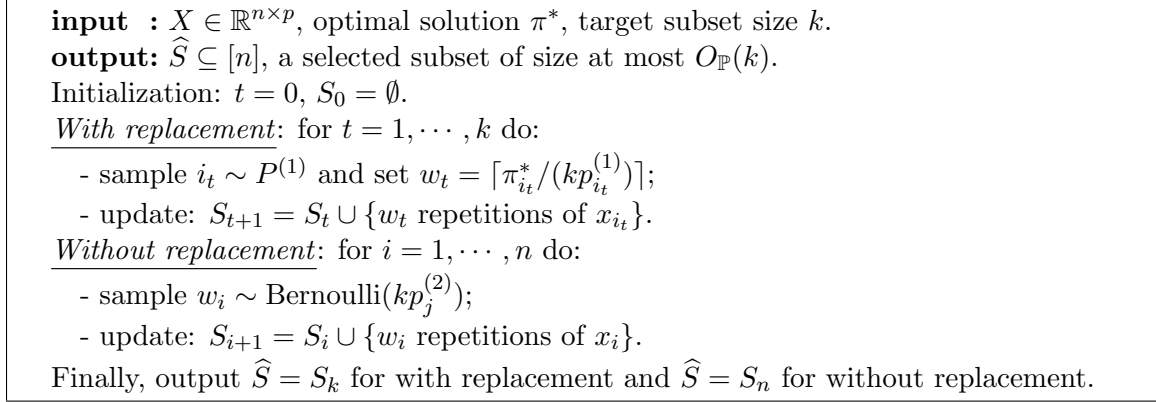
6

---

input : $X \in \mathbb{R}^{n \times p}$, optimal solution π^*, target subset size k.
output: $\widehat{S} \subseteq [n]$, a selected subset of size at most $O_\mathbb{P}(k)$.
Initialization: $t = 0$, $S_0 = \emptyset$.
With replacement: for $t = 1, \cdots, k$ do:
    - sample $i_t \sim P^{(1)}$ and set $w_t = \lceil \pi^*_{i_t}/(kp^{(1)}_{i_t})\rceil$;
    - update: $S_{t+1} = S_t \cup \{w_t \text{ repetitions of } x_{i_t}\}$.
Without replacement: for $i = 1, \cdots, n$ do:
    - sample $w_i \sim \text{Bernoulli}(kp^{(2)}_j)$;
    - update: $S_{i+1} = S_i \cup \{w_i \text{ repetitions of } x_i\}$.
Finally, output $\widehat{S} = S_k$ for with replacement and $\widehat{S} = S_n$ for without replacement.

Figure 1: Sampling based experiment selection (expected size constraint).

Under the without replacement setting the distribution $P^{(2)}$ is straightforward: $p^{(2)}_j$ is proportional to the optimal continuous weights π^*_j; under the with replacement setting, the sampling distribution takes into account leverage scores (effective resistance) of each data point in the conditioned covariance as well. Later analysis (Theorem 5) shows that it helps with the finite-sample condition on k. Figure 1 gives details of the sampling based algorithms for both with and without replacement settings.

The following proposition bounds the size of $\widehat{S}$ in high probability:

Proposition 4 _For any $\delta \in (0, 1/2)$ with probability at least $1 - \delta$ it holds that $|\widehat{S}| \leq 2(1 + 1/\delta)k$. That is, $|\widehat{S}| \leq O_\mathbb{P}(k)$._

Proof Apply Markov's inequality and note that an additional k samples need to be added due to the ceiling operator in with replacement sampling. ∎

The sampling procedure is easy to understand in an asymptotic sense: it is easy to verify that $\mathbb{E}X^\top_{i_t}X_{i_t} = X^\top \text{diag}(\pi^*/k)X$ and $\mathbb{E}[w_{i_t}] = 1$, for both with and without replacement settings. Note that $\|p^{(2)}_{i_t}\|_\infty \leq 1/k$ by feasibility constraints and hence $\text{Bernoulli}(kp^{(2)}_{i_t})$ is a valid distribution for all $i_t \in [n]$. For the with replacement setting, by weak law of large numbers, $X^\top_{\widehat{S}}X_{\widehat{S}} \xrightarrow{p} X^\top \text{diag}(\pi^*)X$ as $k \to \infty$ and hence $\text{tr}[(X^\top_S X_S)^{-1}] \xrightarrow{p} f(\pi^*; X)$ by continuous mapping theorem. A more refined analysis is presented in Theorem 5 to provide explicit conditions under which the asymptotic approximations are valid and on the statistical efficiency of $\widehat{S}$ as well as analysis under the more restrictive without replacement regime.

Theorem 5 _Fix $\epsilon > 0$ as an arbitrarily small accuracy parameter. Suppose the following conditions hold:_

$$\text{With replacement}: \quad p\log k/k = O(\epsilon^2);$$
$$\text{Without replacement}: \quad \|\Sigma_*^{-1}\|_2 \kappa(\Sigma_*)\|X\|^2_\infty \log p = O(\epsilon^2).$$

Here $\Sigma_* = X^\top \text{diag}(\pi^*)X$ and $\kappa(\Sigma_*)$ denotes the conditional number of Σ_*. Then with probability at least 0.9 the subset OLS estimator $\widehat{\beta} = (X_{\widehat{S}}^\top X_{\widehat{S}})^{-1} X_{\widehat{S}}^\top y_{\widehat{S}}$ satisfies

$$\mathbb{E}\left[\|\widehat{\beta} - \beta_0\|_2^2 \Big| X_{\widehat{S}}\right] = \sigma^2 \text{tr}\left[(X_{\widehat{S}}^\top X_{\widehat{S}})^{-1}\right] \leq (1 + O(\epsilon)) \cdot \sigma^2 f_b^*(X; k), \quad b \in \{1, 2\}.$$

We adapt the proof technique of Spielman and Srivastava in their seminal work on spectral sparsification of graphs (Spielman and Srivastava, 2011). More specifically, we prove the following stronger "two-sided" result which shows that $X_{\widehat{S}}^\top X_{\widehat{S}}$ is a *spectral approximation* of Σ_* with high probability, under suitable conditions.

Lemma 6 *Under the same conditions in Theorem 5, it holds that with probability at least 0.9 that*

$$(1 - \epsilon)z^\top \Sigma_* z \leq z^\top \widehat{\Sigma}_{\widehat{S}} z \leq (1 + \epsilon)z^\top \Sigma_* z, \quad \forall z \in \mathbb{R}^p,$$

where $\Sigma_* = X^\top \text{diag}(\pi^*)X$ and $\Sigma_{\widehat{S}} = X_{\widehat{S}}^\top X_{\widehat{S}}$.

Lemma 6 implies that with high probability $\sigma_j(\widehat{\Sigma}_{\widehat{S}}) \geq (1 - \epsilon)\sigma_j(\Sigma_*)$ for all $j = 1, \cdots, p$. Recall that $\text{tr}[(X_S^\top X_S)^{-1}] = \sum_{j=1}^p \sigma_j(\widehat{\Sigma}_{\widehat{S}})^{-1}$ and $f(\pi^*; X) = \sum_{j=1}^p \sigma_j(\Sigma_*)$. Subsequently, for $\epsilon \in (0, 1 - c]$ for some constant $c > 0$, $\text{tr}[(X_{\widehat{S}}^\top X_{\widehat{S}})^{-1}] \leq (1 + O(\epsilon))f(\pi^*; X)$. Theorem 5 is thus proved.

3.3 Sampling based experiment selection: hard size constraint

In some applications it is mandatory to respect a hard subset size constraint; that is, a randomized algorithm is expected to output $\widehat{S}$ that satisfies $|\widehat{S}| \leq k$ almost surely, and no over-sampling is allowed. To handle such hard constraints, we revise the algorithm in Figure 1 as follows:

---

input : $X \in \mathbb{R}^{n \times p}$, optimal solution π^*, target subset size k.
output: $\widehat{S} \subseteq [n]$, a selected subset of size at most k.
Initialization: $t = 0$, $S_0 = \emptyset$, $R_0 = \emptyset$.
1. *With replacement*: sample $i_t \sim P^{(1)}$; set $w_t = \lceil \pi_{i_t}^* / (k p_{i_t}^{(1)}) \rceil$;

   Without replacement: pick random $i_t \notin R_{t-1}$; sample $w_t \sim \text{Bernoulli}(k p_{i_t}^{(2)})$.
2. *Update*: $S_t = S_{t-1} \cup \{w_t \text{ repetitions of } i_t\}$, $R_t = R_{t-1} \cup \{i_t\}$.
3. Repeat steps 1 and 2 until at some $t = T$, $|S_{T+1}| > k$ or $R_{T+1} = [n]$. Output $\widehat{S} = S_T$.

Figure 2: Sampling based experiment selection (deterministic size constraint).

We have the following theorem, which mimics Theorem 5 but with weaker approximation bounds:

Theorem 7 *Suppose the following conditions hold:*

$$\text{with replacement}: \quad p \log k / k = O(1);$$
$$\text{without replacement}: \quad \|\Sigma_*^{-1}\|_2 \kappa(\Sigma_*) \|X\|_\infty^2 \log p = O(1).$$

Then with probability at least 0.8 the subset estimator $\widehat{\beta} = (X_{\widehat{S}}^\top X_{\widehat{S}})^{-1} X_S^\top y_S$ satisfies

$$\mathbb{E}\left[\|\widehat{\beta} - \beta_0\|_2^2 \Big| X_{\widehat{S}}\right] = \sigma^2 \mathrm{tr}\left[(X_{\widehat{S}}^\top X_{\widehat{S}})^{-1}\right] \leq O(1) \cdot \sigma^2 f_b^*(X; k), \quad b \in \{1, 2\}.$$

The following lemma is key to the proof of Theorem 7. Unlike Lemma 6, in Lemma 8 we only prove one side of the spectral approximation relation, which suffices for our purposes. To handle without replacement, we cite matrix Bernstein for combinatorial matrix sums in (Mackey et al., 2014).

Lemma 8 *Define $\widehat{\Sigma}_{\widehat{S}} = X_{\widehat{S}}^\top X_{\widehat{S}}$. Suppose the following conditions hold:*

$$\textit{With replacement}: \quad p \log T / T = O(1);$$
$$\textit{Without replacement}: \quad \|\Sigma_*^{-1}\|_2 \kappa(\Sigma_*) \|X\|_\infty^2 \log p = O(T/n).$$

Then with probability at least 0.9 the following holds:

$$z^\top \widehat{\Sigma}_{\widehat{S}} z \geq K_T z^\top \Sigma_* z, \qquad \forall z \in \mathbb{R}^p, \tag{6}$$

where $K_T = \Omega(T/k)$ for with replacement and $K_T = \Omega(T/n)$ for without replacement.

Finally, we need to relate conditions on T in Lemma 8 to interpretable conditions on subset budget k:

Lemma 9 *Let $\delta > 0$ be an arbitrarily small fixed failure probability. The with probability at least $1 - \delta$ we have that $T \geq \delta k$ for with replacement and $T \geq \delta n$ for without replacement.*

Proof For with replacement we have $\mathbb{E}[\sum_{t=1}^T w_t] = T$ and for without replacement we have $\mathbb{E}[\sum_{t=1}^T w_t] = Tk/n$. Applying Markov's inequality on $\Pr[\sum_{t=1}^T w_t > k]$ for $T = \delta k$ and/or $T = \delta n$ we complete the proof of Lemma 9. ∎

Combining Lemmas 8 and 9 with $\delta = 0.1$ and note that $T \leq k$ almost surely (because $w_t \geq 1$), we prove Theorem 7.

3.4 Greedy experiment selection

---
input : $X \in \mathbb{R}^{n \times p}$, Initial subset $S_0 \subseteq [n]$, target size $k \leq |S_0|$.
output: $\widehat{S} \subseteq [n]$, a selected subset of size k.
Initialization: $t = 0$.
1. Find $j^* \in S_t$ such that $\mathrm{tr}[(X_{S_t \backslash \{j^*\}}^\top X_{S_t \backslash \{j^*\}})^{-1}]$ is minimized.
2. Remove j^* from S_t: $S_{t+1} = S_t \backslash \{j^*\}$.
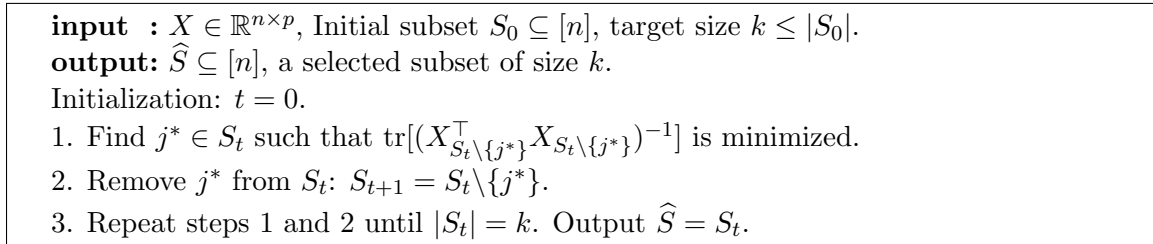3. Repeat steps 1 and 2 until $|S_t| = k$. Output $\widehat{S} = S_t$.

Figure 3: Greedy experiment selection.

Avron and Boutsidis (2013) proposed an interesting greedy removal algorithm (outlined in Figure 3) and established the following result:

Lemma 10 *Suppose $\widehat{S} \subseteq [n]$ of size k is obtained by running algorithm in Figure 3 with an initial subset $S_0 \subseteq [n]$, $|S_0| \geq k$. Both $\widehat{S}$ and S_0 are standard sets (i.e., without replacement). Then*

$$\mathrm{tr}\left[(X_{\widehat{S}}^\top X_{\widehat{S}})^{-1}\right] \leq \frac{|S_0| - p + 1}{k - p + 1} \mathrm{tr}\left[(X_{S_0}^\top X_{S_0})^{-1}\right].$$

In Avron and Boutsidis (2013) the greedy removal procedure in Figure 3 is applied to the entire design set $S_0 = [n]$, which gives approximation guarantee $\mathrm{tr}[(X_{\widehat{S}}^\top X_{\widehat{S}})^{-1}] \leq \frac{n-p+1}{k-p+1}\mathrm{tr}[(X^\top X)^{-1}]$. This results in an approximation ratio of $C(n, p, k) = \frac{n-p+1}{k-p+1}$ as defined in Eq. (2), by applying the trivial bound $\mathrm{tr}[(X^\top X)^{-1}] \leq f_2^*(k; X)$, which is tight for a design that has exactly k non-zero rows.

To further improve the approximation ratio, we consider applying the greedy removal procedure with S_0 equal to the support of π^*; that is, $S_0 = \{j \in [n] : \pi_j^* > 0\}$. Because $\|\pi^*\|_\infty \leq 1$ under the without replacement setting, we have the following corollary:

Corollary 11 *Let S_0 be the support of π^* and suppose $\|\pi^*\|_\infty \leq 1$. Then*

$$\mathrm{tr}[(X_{\widehat{S}}^\top X_{\widehat{S}})^{-1}] \leq \frac{\|\pi^*\|_0 - p + 1}{k - p + 1} f(\pi^*; X) = \frac{\|\pi^*\|_0 - p + 1}{k - p + 1} f_2^*(k; X).$$

It is thus important to upper bound the support size $\|\pi^*\|_0$. With the trivial bound of $\|\pi^*\|_0 \leq n$ we recover the $\frac{n-p+1}{k-p+1}$ approximation ratio by applying Figure 3 to $S_0 = [n]$. In order to bound $\|\pi^*\|_0$ away from n, we consider the following assumption imposed on X:

Assumption 3.1 *Define mapping $\phi : \mathbb{R}^p \to \mathbb{R}^{\frac{p(p+1)}{2}}$ as $\phi(x) = (\xi_{ij}x(i)x(j))_{1\leq i\leq j\leq p}$, where $x(i)$ denotes the ith coordinate of a p-dimensional vector x and $\xi_{ij} = 1$ if $i = j$ and $\xi_{ij} = 2$ otherwise. Denote $\widetilde{\phi}(x) = (\phi(x), 1) \in \mathbb{R}^{\frac{p(p+1)}{2}+1}$ as the affine version of $\phi(x)$. For any $\frac{p(p+1)}{2} + 1$ distinct rows of X, their mappings under $\widetilde{\phi}$ are linear independent.*

Assumption 3.1 is essentially a general-position assumption, which assumes that no $\frac{p(p+1)}{2} + 1$ design points in X lie on a degenerate affine subspace after a specific quadratic mapping. Like other similar assumptions in the literature (Tibshirani, 2013), Assumption 3.1 is very mild and almost always satisfied in practice, for example, if each row of X is independently sampled from absolutely continuous distributions.

We are now ready to state the main lemma bounding the support size of π^*.

Lemma 12 $\|\pi^*\|_0 \leq k + \frac{p(p+1)}{2}$ *if Assumption 3.1 holds.*

Lemma 12 is established by an interesting observation into the properties of Karush-Kuhn-Tucker (KKT) conditions of the optimization problem Eq. (4), which involves a linear system with $\frac{p(p+1)}{2} + 1$ variables. The complete proof of Lemma 12 is given in Sec. 7.5. To contrast the results in Lemma 12 with classical rank/support bounds in SDP and/or linear programming (e.g. the Pataki's bound (Pataki, 1998)), note that the number of constraints in the SDP formulation of Eq. (4) (see also Appendix B) is linear in n, and hence analysis similar to (Pataki, 1998) would result in an upper bound of $\|\pi^*\|_0$ that scales with n, which is less useful for our analytical purpose.

Combining results from both Lemma 12 and Corollary 11 we arrive at the following theorem, which upper bounds the approximation ratio of the greedy removal procedure in Figure 3 initialized by the support of π^*.

Theorem 13 *Let π^* be the optimal solution of the without replacement version of Eq. (4) and $\widehat{S}$ be the output of the greedy removal procedure in Figure 3 initialized with $S_0 = \{i \in [n] : \pi_i^* > 0\}$. If $k > p$ and Assumption 3.1 holds then the subset OLS estimator $\widehat{\beta} = (X_{\widehat{S}}^\top X_{\widehat{S}})^{-1} X_{\widehat{S}}^\top y_{\widehat{S}}$ satisfies*

$$\mathbb{E}\left[\|\widehat{\beta} - \beta_0\| \big| X_{\widehat{S}}\right] = \sigma^2 \mathrm{tr}\left[(X_{\widehat{S}}^\top X_{\widehat{S}})^{-1}\right] \leq \left(1 + \frac{p(p+1)}{2(k-p+1)}\right) f_2^*(k; X).$$

Under a slightly stronger condition that $k > 2p$, the approximation ratio $C(n,k,p) = 1 + \frac{p(p+1)}{2(k-p+1)}$ can be simplified to $C(n,k,p) = 1 + O(p^2/k)$. In addition, $C(n,k,p) = 1 + o(1)$ if $p^2/k \to 0$, meaning that near-optimal experiment selection is achievable with computationally tractable methods if $O(p^2)$ design points are allowed in the selected subset.

3.5 Interpretable subsampling example: anisotropic Gaussian design

Even though we consider a fixed pool of design points so far, here we use an anisotropic Gaussian design example to demonstrate that a non-uniform sampling can outperform uniform sampling even under random designs, and to interpret the conditions required in previous analysis. Let $x_1, \cdots, x_n$ be i.i.d. distributed according to an isotropic Gaussian distribution $\mathcal{N}_p(0, \Sigma_0)$.

We first show that non-uniform weights $\pi_i \neq k/n$ could improve the objective $f(\pi; X)$. Let π^{unif} be the uniformly weighted solution of $\pi_i^{\mathrm{unif}} = k/n$, corresponding to selecting each row of X uniformly at random. We then have

$$f(\pi^{\mathrm{unif}}; X) = \frac{1}{k}\mathrm{tr}\left[\left(\frac{1}{n}X^\top X\right)^{-1}\right] \xrightarrow{p} \frac{1}{k}\mathrm{tr}(\Sigma_0^{-1}).$$

On the other hand, let $B^2 = \gamma \mathrm{tr}(\Sigma_0)$ for some universal constant $\gamma > 1$ and $\mathbb{B} = \{x \in \mathbb{R}^p : \|x\|_2^2 \leq B^2\}$, $\mathbb{X} = \{x_1, \cdots, x_n\}$. By Markov inequality, $|\mathbb{X} \cap \mathbb{B}| \gtrsim \frac{\gamma n}{1-\gamma}$. Define weighted solution π^w as $\pi_i^w \propto 1/p(x_i|\mathbb{B}) \cdot I[x_i \in \mathbb{B}]$ normalized such that $\|\pi_i^w\|_1 = k$. [1] Then

$$f(\pi^w; X) = \frac{1}{k}\frac{\mathrm{tr}\left[\left(\frac{1}{n}\sum_{x_i \in \mathbb{X} \cap \mathbb{B}} \frac{x_i x_i^\top}{p(x_i|\mathbb{B})}\right)^{-1}\right]}{\frac{1}{n}\sum_{x_i \in \mathbb{X} \cap \mathbb{B}} 1/p(x_i|\mathbb{B})}$$

$$\xrightarrow{p} \frac{1}{k}\frac{n}{|\mathbb{X} \cap \mathbb{B}|}\frac{\mathrm{tr}\left[\left(\int_{\mathbb{B}} xx^\top \mathrm{d}x\right)^{-1}\right]}{\int_{\mathbb{B}} 1\mathrm{d}x} \lesssim \frac{p^2}{(1-\gamma)\mathrm{tr}(\Sigma_0)}.$$

---

1. Note that π^* may *not* be the optimal solution of Eq. (4); however, it suffices for the purpose of the demonstration of improvement resulting from non-uniform weights.

Here in the last inequality we apply Lemma 17. Because $\frac{\text{tr}(\Sigma_0^{-1})}{p} = \frac{1}{p}\sum_{i=1}^{p}\frac{1}{\sigma_i(\Sigma_0)} \geq \left(\frac{1}{p}\sum_{i=1}^{p}\sigma_i(\Sigma_0)\right)^{-1} = \frac{p}{\text{tr}(\Sigma_0)}$ by Jensen's inequality, we conclude that in general $f(\pi^w; X) < f(\pi^{\text{unif}}; X)$, and the gap is larger for ill-conditioned covariance Σ_0. This example shows that uneven weights in π helps reducing the trace of inverse of the weighted covariance $X^\top \text{diag}(\pi)X$.

Under this model, we also simplify the conditions for the without replacement model in theorem 5 and 7. Because $x_1, \cdots, x_n \overset{i.i.d.}{\sim} \mathcal{N}_p(0, \Sigma_0)$, it holds that $\|X\|_\infty^2 \leq O_\mathbb{P}(\|\Sigma_0\|_2^2 p \log n)$. In addition, by simple algebra $\|\Sigma_*^{-1}\|_2 \leq p^{-1}\kappa(\Sigma_*)\text{tr}(\Sigma_*^{-1}) \leq p^{-1}\kappa(\Sigma_*)f_1^*(k; X)$. Using a very conservative upper bound of $f_1^*(k; X)$ by sampling rows in X uniformly at random and apply weak law of large numbers and the continuous mapping theorem, we have that $f_1^*(k; X) \lesssim \frac{1}{k}\text{tr}(\Sigma_0^{-1})$. In addition, $\text{tr}(\Sigma_0^{-1})\|\Sigma_0\|_2 \leq p\|\Sigma_0^{-1}\|_2\|\Sigma_0\|_2 = p\kappa(\Sigma_0)$. Subsequently, the condition $\|\Sigma_*^{-1}\|_2\kappa(\Sigma_*)\|X\|_\infty^2 \log p = O(\epsilon^2)$ is implied by

$$\frac{p\kappa(\Sigma_*)^2\kappa(\Sigma_0)\log p \log n}{k} = O(\epsilon^2). \tag{7}$$

Essentially, the condition is reduced to $k \gtrsim \kappa(\Sigma_0)\kappa(\Sigma_*) \cdot p \log n \log p$. The linear dependency on p is necessary, as we consider the low-dimensional linear regression problem and $k < p$ would imply an infinite mean-square error in estimation of β_0. We also remark that the condition is scale-invariant, as $X' = \xi X$ and X share the same quantity $\|\Sigma_*^{-1}\|_2\kappa(\Sigma_*)\|X\|_\infty^2 \log k$.

3.6 Extensions

We discuss possible extension of our results beyond estimation of β_0 in the linear regression model.

3.6.1 GENERALIZED LINEAR MODELS

In a generalized linear model $\mu(x) = \mathbb{E}[Y|x]$ satisfies $g(\mu(x)) = \eta = x^\top\beta_0$ for some known link function $g : \mathbb{R} \to \mathbb{R}$. Under regularity conditions (Van der Vaart, 2000), the maximum-likelihood estimator $\widehat{\beta}_n \in \text{argmax}_\beta\{\sum_{i=1}^{n}\log p(y_i|x_i; \beta)\}$ satisfies $\mathbb{E}\|\widehat{\beta}_n - \beta_0\|_2^2 = (1 + o(1))\text{tr}(I(X, \beta_0)^{-1})$, where $I(X, \beta_0)$ is the Fisher's information matrix:

$$I(X, \beta_0) = -\sum_{i=1}^{n}\mathbb{E}\frac{\partial^2 \log p(y_i|x_i; \beta_0)}{\partial\beta\partial\beta^\top} = -\sum_{i=1}^{n}\left(\mathbb{E}\frac{\partial^2 \log p(y_i; \eta_i)}{\partial\eta_i^2}\right)x_i x_i^\top. \tag{8}$$

Here both expectations are taken over y conditioned on X and the last equality is due to the sufficiency of $\eta_i = x_i^\top\beta_0$. The experiment selection problem is then formulated to select a subset $S \subseteq [n]$ of size k, either with or without duplicates, that minimizes $\text{tr}(I(X_S, \beta_0)^{-1})$.

It is clear from Eq. (8) that the optimal subset S^* depends on the unknown parameter β_0, which itself is to be estimated. This issue is known as the *design dependence* problem for generalized linear models (Khuri et al., 2006). One approach is to consider *locally optimal designs* (Khuri et al., 2006; Chernoff, 1953), where a consistent estimate $\check{\beta}$ of β_0 is first obtained on an initial design subset [2] and then $\check{\eta}_i = x_i^\top\check{\beta}$ is supplied to compute a more

---

2. Notice that a consistent estimate can be obtained using much fewer points than an estimate with finite approximation guarantee.

refined design subset to get the final estimate $\widehat{\beta}$. With the initial estimate $\check{\beta}$ available, one may apply transform $x_i \mapsto \widetilde{x}_i$ defined as

$$\widetilde{x}_i = \sqrt{-\mathbb{E}\frac{\partial^2 \log p(y_i; \check{\eta}_i)}{\partial \eta^2}}\, x_i.$$

Note that under regularity conditions $-\mathbb{E}\frac{\partial^2 \log p(y_i; \check{\eta}_i)}{\partial \eta_i^2} = \mathbb{E}\left(\frac{\partial \log(y_i; \check{x}_i)}{\partial \eta_i}\right)^2$ is non-negative and hence the square-root is well-defined. All results in Theorems 5, 7 and 13 are valid with $X = [x_1, \cdots, x_n]^\top$ replaced by $\widetilde{X} = [\widetilde{x}_1, \cdots, \widetilde{x}_n]^\top$ for generalized linear models. Below we consider two generalized linear model examples and derive explicit forms of $\widetilde{X}$.

Example 1: Logistic regression   In a logistic regression model responses $y_i \in \{0, 1\}$ are binary and the likelihood model is

$$p(y_i; \eta_i) = \psi(\eta_i)^{y_i}(1 - \psi(\eta_i))^{1-y_i}, \qquad \text{where} \quad \psi(\eta_i) = \frac{e^{\eta_i}}{1 + e^{\eta_i}}.$$

Simple algebra yields

$$\widetilde{x}_i = \sqrt{\frac{e^{\check{\eta}_i}}{(1 + e^{\check{\eta}_i})^2}}\, x_i,$$

where $\check{\eta}_i = x_i^\top \check{\beta}$.

Example 2: Poisson count model   In a Poisson count model the response variable y_i takes values of non-negative integers and follows a Poisson distribution with parameter $\lambda = e^{\eta_i} = e^{x_i^\top \beta_0}$. The likelihood model is formally defined as

$$p(y_i = r; \eta_i) = \frac{e^{\eta_i r}e^{-e^{\eta_i}}}{r!}, \qquad r = 0, 1, 2, \cdots.$$

Simple algebra yields

$$\widetilde{x}_i = \sqrt{e^{\check{\eta}_i}}\, x_i,$$

where $\check{\eta}_i = x_i^\top \check{\beta}$.

3.6.2 DELTA'S METHOD

Suppose $g(\beta_0)$ is the quantity of interest, where $\beta_0 \in \mathbb{R}^p$ is the parameter in a linear regression model and $g : \mathbb{R}^p \to \mathbb{R}^m$ is some known function. Let $\widehat{\beta}_n = (X^\top X)^{-1}X^\top y$ be the OLS estimate of β_0. If ∇g is continuously differentiable and $\widehat{\beta}_n$ is consistent, then by the classical delta's method (Van der Vaart, 2000) $\mathbb{E}\|g(\widehat{\beta}_n) - g(\beta_0)\|_2^2 = (1 + o(1))\sigma^2 \mathrm{tr}(\nabla g(\beta_0)(X^\top X)^{-1}\nabla g(\beta_0)^\top) = (1 + o(1))\sigma^2 \mathrm{tr}(G_0(X^\top X)^{-1})$, where $G_0 = \nabla g(\beta_0)^\top \nabla g(\beta_0)$. If G_0 depends on the unknown parameter β_0 then the design dependence problem again exists, and a locally optimal solution can be obtained by replacing G_0 in the objective function with $\check{G} = \nabla g(\check{\beta})^\top \nabla g(\check{\beta})$ for some initial estimate $\check{\beta}$ of β_0.

If $\check{G}$ is invertible, then there exists invertible $p \times p$ matrix $\check{P}$ such that $\check{G} = \check{P}\check{P}^\top$ because $\check{G}$ is positive definite. Applying the linear transform

$$x_i \mapsto \widetilde{x}_i = \check{P}^{-1}x_i$$

we have that $\mathrm{tr}[G_0(X^\top X)^{-1}] = \mathrm{tr}[(\widetilde{X}^\top \widetilde{X})^{-1}]$, where $\widetilde{X} = [\widetilde{x}_1, \cdots, \widetilde{x}_n]^\top$. Our results in Theorems 5, 7 and 13 remain valid by operating on the transformed matrix $\widetilde{X} = X\check{P}^{-\top}$.

Table 2: Running time (seconds) / no. of iterations for L_5^* and $L_6^\dagger$.

		Isotropic design (t distribution), df $= 1$.			
$k =$	$2p$	$4p$	$6p$	$8p$	$10p$
L_5^*	1.8/31	0.8/19	1.7/26	0.9/14	0.3/9
$L_6^\dagger$	140.1/90	425.9/156	767.5/216	993.7/245	1077/253
		Isotropic design (t distribution), df $= 3$.			
$k =$	$2p$	$4p$	$6p$	$8p$	$10p$
L_5^*	0.6/14	0.4/8	0.3/7	0.2/5	0.2/5
$L_6^\dagger$	158.3/104	438.0/161	802.5/223	985.2/242	1105/252
		Skewed design (multivariate Gaussian), $\alpha = 3$.			
$k =$	$2p$	$4p$	$6p$	$8p$	$10p$
L_5^*	0.8/16	0.7/12	0.5/9	0.4/8	0.4/8
$L_6^\dagger$	182.9/120	487.1/180	753.4/212	935.8/230	1057/250

Example: prediction error. In some application scenarios the *prediction error* $\|Z\widehat{\beta} - Z\beta_0\|_2^2$ rather than the estimation error $\|\widehat{\beta} - \beta_0\|_2^2$ is of interesting, either because the linear model is used mostly for prediction or component of the underlying model β_0 lack physical interpretations. Another interesting application is the *transfer learning* (Pan and Yang, 2010), in which the training and testing data have different designs (e.g., Z instead of X) but share the same conditional distribution of labels, parameterized by the linear model β_0.

Suppose $Z \in \mathbb{R}^{m \times p}$ is a known full-rank data matrix upon which predictions are seeked, and define $\widehat{\Sigma}_Z = \frac{1}{m} Z^\top Z \succ 0$ to be the sample covariance of Z. Our algorithmic framework as well as its corresponding analysis remain valid for such prediction problems with transform $x_i \mapsto \widehat{\Sigma}_Z^{-1/2} x_i$. In particular, the guarantees for the greedy algorithm and the with replacement sampling algorithm remain unchanged, and the guarantee for the without replacement sampling algorithm is valid as well, except that the $\|\Sigma_*^{-1}\|_2$ and $\kappa(\Sigma_*)$ terms have to be replaced by the (relaxed) optimal sample covariance after the linear transform $x_i \mapsto \widehat{\Sigma}_Z^{-1/2} x_i$.

4. Numerical results on synthetic data

We report selection performance (measured in terms of $F(S; X) := \text{tr}[(X_S^\top X_S)^{-1}]$ which is the mean squared error of the ordinary least squares estimator based on (X_S, y_S) in estimating the regression coefficients) on synthetic data. Only the without replacement setting is considered, and results for the with replacement setting are similar. In all simulations, the experimental pool (number of given design points) n is set to 1000, number of variables p is set to 50, number of selected design points k ranges from $2p$ to $10p$. For randomized methods, we run them for 20 independent trials under each setting and report the median. Though Fedorov's exchange algorithm is randomized in nature (outcome depending on the initialization), we only run it once for each setting because of its expensive computational requirement. It is observed that in practice, the algorithm's outcome is not sensitive to initializations.
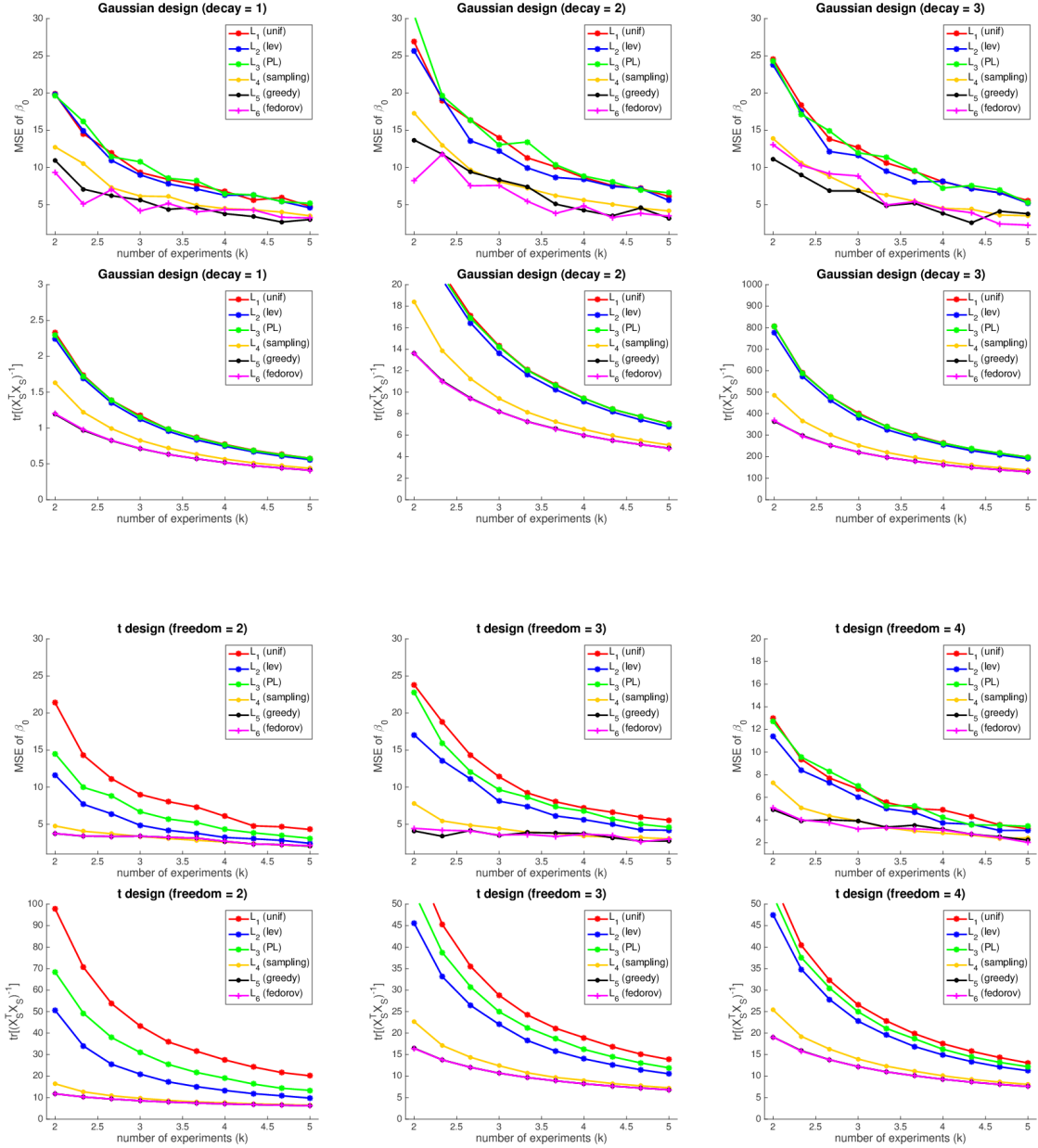
Figure 1: The ratio of MSE $\|\widehat{\beta} - \beta_0\|_2^2$ compared against the MSE of the OLS estimator on the full data pool $\|\widehat{\beta}^{\mathrm{ols}} - \beta_0\|_2^2$, $\widehat{\beta}^{\mathrm{ols}} = (X^\top X)^{-1} X^\top y$, and objective values $F(\widehat{S}; X) = \mathrm{tr}[(X_{\widehat{S}}^\top X_{\widehat{S}})^{-1}]$ on synthetic data sets. Note that the L_5 (greedy) curve is under the L_6 (fedorov) curve when not visible. See Sec. 4 for details of experimental settings.

4.1 Data generation

The design pool X is generated so that each row of X is sampled from some underlying distribution. We consider two distributional settings for generating X. Similar settings were considered in (Ma et al., 2015) for subsampling purposes.

1. *Distributions with skewed covariance*: each row of X is sampled i.i.d. from a multivariate Gaussian distribution $\mathcal{N}_p(0, \Sigma_0)$ with $\Sigma_0 = U\Lambda U^\top$, where U is a random orthogonal matrix and $\Lambda = \mathrm{diag}(\lambda_1, \cdots, \lambda_p)$ controls the skewness or conditioning of X. A power-law decay of λ, $\lambda_j = j^{-\alpha}$, is imposed, with small α corresponding to "flat" distribution and large α corresponding to "skewed" distribution. Rows 1-2 in Figure 1 correspond to this setting.

2. *Distributions with heavy tails*: We use t distribution as the underlying distribution for generating each entry in X, with degrees of freedom ranging from 2 to 4. These distributions are heavy tailed and high-order moments of X typically do not exist. They test the robustness of experiment selection methods. Rows 3-4 in Figure 1 correspond to this setting.

4.2 Methods

The methods that we compare are listed below:

- L_1 (uniform sampling): each row of X is sampled uniformly at random, without replacement.

- L_2 (leverage score sampling): each row of X, x_i, is sampled without replacement, with probability proportional to its leverage score $x_i^\top (X^\top X)^{-1} x_i$. This strategy is considered in Ma et al. (2015) for subsampling in linear regression models.

- L_3 (predictive length sampling): each row of X, x_i, is sampled without replacement, with probability proportional to its ℓ_2 norm $\|x_i\|_2$. This strategy is derived in Zhu et al. (2015).

- L_4^* (sampling based selection with π^*): the sampling based method that is described in Sec. 3.3, based on π^*, the optimal solution of Eq. (4). We consider the hard size constraint algorithm only.

- L_5^* (greedy based selection with π^*): the greedy based method that is described in Sec. 3.4.

- $L_6^\dagger$ (Fedorov's exchange algorithm): the Fedorov's exchange algorithm (Miller and Nguyen, 1994) with the A-optimality objective. Note that this method does not have guarantees, nor does it provably terminate in polynomial number of exchanges. Description of the algorithm is placed in Appendix C.
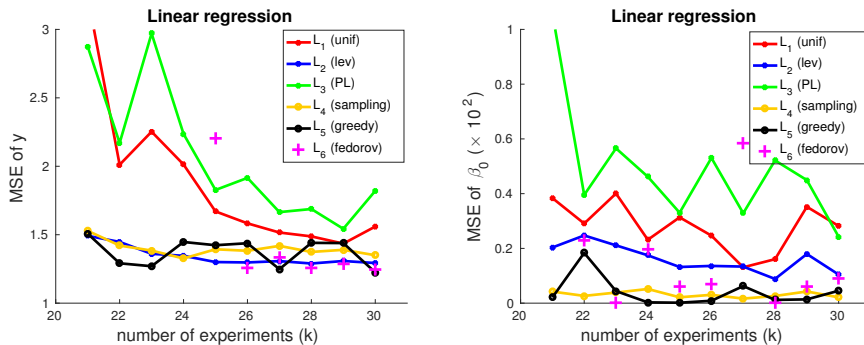
Figure 2: Results on real material synthesis data ($n = 133, p = 9, 20 < k \leq 30$). Left: prediction error $\|y - X\widehat{\beta}\|_2^2$, normalized by dividing by the prediction error $\|y - X\widehat{\beta}^{\mathrm{ols}}\|_2^2$ of the OLS estimator on the full data set; right: estimation error $\|\widehat{\beta} - \widehat{\beta}^{\mathrm{ols}}\|_2^2$. All randomized algorithms ($L_1, L_2, L_3, L_4^*, L_6^\dagger$) are run for 50 times and the median performance is reported.

4.3 Performance

The ratio of the mean square error $\|\widehat{\beta} - \beta_0\|_2^2$ compared to the OLS estimator on the full data set $\|\widehat{\beta}^{\mathrm{ols}} - \beta_0\|_2^2$, $\widehat{\beta}^{\mathrm{ols}} = (X^\top X)^{-1} X^\top y$, and objective values $F(\widehat{S}; X) = \mathrm{tr}((X_{\widehat{S}}^\top X_{\widehat{S}})^{-1})$ are reported in Figure 1.

Table 2 reports the running time and number of iterations of L_5^* (greedy based selection) and $L_6^\dagger$ (Fedorov's exchange algorithm). In general, the greedy method is 100 to 1000 times faster than the exchange algorithm, and also converges in much fewer iterations.

For both synthetic settings, our proposed methods (L_4^* and L_5^*) significantly outperform existing approaches (L_1, L_2, L_3), and their performance is on par with the Fedorov's exchange algorithm, which is much more computationally expensive (cf. Table 2).

5. Numerical results on real data

5.1 The material synthesis dataset

We apply our proposed methods to an experimental design problem of low-temperature microwave-assisted crystallization of ceramic thin films (Reeja-Jayan et al., 2012; Nakamura et al., 2017). The microwave-assisted experiments were controlled by four experimental parameters: temperature T ($120°C$ to $170°C$), hold time H (0 to 60 minutes), ramp power P ($0W$ to $60W$) and tri-ethyl-gallium (TEG) volume ratio R (from 0 to 0.93). A generalized quadratic regression model was employed in (Nakamura et al., 2017) to estimate the coverage percentage of the crystallization CP:

$$CP = \beta_1 + \beta_2 T + \beta_3 H + \beta_4 P + \beta_5 R + \beta_6 T^2 + \beta_7 H^2 + \beta_8 P^2 + \beta_9 R^2. \tag{9}$$

As a proof of concept, we test our proposed algorithms and compare with baseline methods on a data set consisting of 133 experiments, with the coverage percentage fully collected and measured. We consider selection of subsets of k experiments, with k ranging

Table 3: Selected subsets of uniform sampling (Uniform), leverage score sampling (Levscore) and greedy method (Greedy) on the material synthesis dataset.

Uniform (L_1)				Levscore (L_2)				Greedy (L_5^*)			
T	P	H	R	T	P	H	R	T	P	H	R
150	30	7.5	.07	120	30	5	.07	140	15	7.5	.80
165	30	3	.73	140	15	2.5	.80	140	15	2.5	.80
140	15	7.6	.80	170	30	7.5	.93	140	60	2.5	.80
170	30	2.5	.07	140	60	7.5	.80	160	15	7.5	.80
160	15	7.5	.80	150	15	2.5	.80	160	60	7.5	.80
150	15	2.5	.80	150	30	60	.67	140	60	2.5	.80
160	30	2.5	.80	160	30	30	.87	160	60	7.5	.80
150	30	6.1	.87	170	30	2.5	.07	170	30	7.5	.93
160	30	6.1	.87	170	30	0	.93	165	30	7.5	0
150	30	4.1	.67	140	30	5	.80	150	30	5	0
165	30	7.5	.13	140	30	2.5	.87	135	30	2.5	.5
140	30	15	.87	140	30	6.2	.67	150	30	4.1	.87
160	60	7.5	.80	160	30	60	.87	150	30	2.5	.87
150	30	5	.50	140	30	60	.87	135	30	3	.50
140	30	7.5	.80	160	30	7.5	.80	135	30	3	0
150	30	5	.80	160	30	2.5	.87	140	30	30	.67
140	30	6.1	.67	140	30	4.1	.67	140	30	30	.87
140	30	2.5	.67	135	30	3	.50	150	30	30	.87
150	60	5	.80	150	30	30	.67	160	30	30	.87
160	30	5	.80	165	30	5	.67	120	30	5	.07
135	30	3	.07	160	60	7.5	.80	160	30	60	.87
165	30	5	0	165	30	3	.67	140	30	60	.87
120	30	7.5	.93	135	30	3	.67	120	30	7	.07
160	30	5	.80	160	15	7.5	.80	120	30	0	.93
165	30	0	.73	150	60	7.5	.80	170	30	0	.93
165	30	3	.67	160	30	4.1	.67	170	30	0	.07
150	15	7.5	.80	165	30	0	.07	150	30	0	.5
140	30	5	.80	135	30	2.5	.73	150	30	0	.6
150	60	2.5	.80	150	30	15	.87	165	30	0	.5
135	30	3	.67	160	60	2.5	.80	160	30	60	.07

Table 4: Comparison of the "true" model $\beta_0 = [0.49; 0.30; 0.19; 3.78]$ and the subsampled estimators $\widehat{\beta}$ for different components of β. k is the number of subsamples. Random algorithms (L_1 through L_4^*) are repeated for 100 times and the median difference is reported.

	$\Delta\beta_1$	$\Delta\beta_2$	$\Delta\beta_3$	$\Delta\beta_4$	$\|\Delta\beta\|_2$		$\Delta\beta_1$	$\Delta\beta_2$	$\Delta\beta_3$	$\Delta\beta_4$	$\|\Delta\beta\|_2$
$k = 20$						$k = 30$					
L_1	.043	.071	.074	.274	.295	L_1	.038	.064	.070	.171	.199
L_2	.017	.026	.018	.260	.263	L_2	.011	.024	.014	.199	.201
L_3	.017	.043	.026	.243	.249	L_3	.015	.031	.018	.236	.239
L_4^*	.014	.039	.114	.215	.247	L_4^*	.014	.023	.030	.044	.060
L_5^*	.010	.024	.032	.209	.213	L_5^*	.022	.000	.050	.088	.104
$L_6^\dagger$	.010	.024	.032	.209	.213	$L_6^\dagger$	.022	.000	.050	.088	.104
$k = 50$						$k = 75$					
L_1	.025	.035	.037	.130	.142	L_1	.016	.032	.024	.133	.139
L_2	.009	.027	.012	.130	.134	L_2	.006	.034	.010	.126	.131
L_3	.011	.035	.013	.166	.170	L_3	.009	.029	.009	.124	.128
L_4^*	.022	.009	.036	.097	.106	L_4^*	.012	.001	.015	.036	.041
L_5^*	.025	.003	.040	.009	.048	L_5^*	.005	.013	.009	.016	.023
$L_6^\dagger$	.025	.003	.040	.009	.048	$L_6^\dagger$	.005	.012	.010	.011	.020

from 21 to 30, and report in Figure 2 the mean-square error (MSE) of both the prediction error $\|y - X\widehat{\beta}\|_2^2$, normalized by dividing by the prediction error $\|y - X\widehat{\beta}_{\mathrm{ols}}\|_2^2$ of the OLS estimator on the full data set, on the full 133 experiments and the estimation error $\|\widehat{\beta} - \widehat{\beta}^{\mathrm{ols}}\|_2^2$. Figure 2 shows that our proposed methods consistently achieve the best performance and are more stable compared to uniform sampling and Fedorov's exchange methods, even though the linear model assumption may not hold.

We also report the actual subset ($k = 30$) design points selected by uniform sampling (L_1), leverage score sampling (L_3) and our greedy method (L_5^*) in Table 3. Table 3 shows that the greedy algorithm (L_5^*) picked very diverse experimental settings, including several settings of very lower temperature ($120°C$) and very short hold time (0). In contrast, the experiments picked by both uniform sampling (L_1) and leverage score sampling (L_3) are less diverse.

5.2 The CPU performance dataset

CPU relative performance refers to the relative performance of a particular CPU model in terms of a base machine - the IBM 370/158 model. Comprehensive benchmark tests are required to accurately measure the relative performance of a particular CPU, which might be time-consuming or even impossible if the actual CPU has not been on the market yet. Ein-Dor and Feldmesser (1987) considered a linear regression model to characterize the relationship between CPU relative performance and several CPU capacity parameters such as main memory size, cache size, channel units and machine clock cycle time. These parameters are fixed for any specific CPU model and could be known even before the man-

ufacturing process. The learned model can also be used to predict the relative performance of a new CPU model based on its model parameters, without running extensive benchmark tests.

Using domain knowledge, Ein-Dor and Feldmesser (1987) narrow down to three parameters of interest: *average memory size* (X_1), *cache size* (X_2) and *channel capacity* (X_3), all being explicitly computable functions from CPU model parameters. An offset parameter is also involved in the linear regression model, making the number of variables $p = 4$. A total of $n = 209$ CPU models are considered, with all of the model parameters and relative performance collected and no missing data. Stepwise linear regression was applied to obtain the following linear model:

$$0.49X_1 + 0.30X_2 + 0.19X_3 + 3.78. \tag{10}$$

To use this data set as a benchmark for evaluating the experiment selection methods, we synthesize labels Y using model Eq. (10) with standard Gaussian noise and measure the difference between fit $\widehat{\beta}$ and the true model $\beta_0 = [0.49; 0.30; 0.19; 3.78]$. Table 4 shows that under various measurement budget (k) constraints, the proposed methods L_4^* and L_5^* consistently achieve small estimation error, and is comparable to the Fedorov's exchange algorithm $(L_6^\dagger)$. As the data set is small (209×4), all algorithms are computationally efficient.

5.3 The Minnesota Wind Dataset

The Minnesota wind dataset collects wind speed information across $n = 2642$ locations in Minnesota, USA for a period of 24 months (for the purpose of this experiment, we only use wind speed data for one month). The 2642 locations are connected with 3304 bi-directional roads, which form an $n \times n$ sparse unweighted undirected graph G. Let $L = \text{diag}(d) - G$ be the $n \times n$ Laplacian of G, where d is a vector of node degrees and let $V \in \mathbb{R}^{n \times p}$ be an orthonormal eigenbasis corresponding to the smallest p eigenvalues of L. As the wind speed signal $y \in \mathbb{R}^n$ is relatively smooth, it can be well-approximated as $y = V\theta + \varepsilon$, where $\theta \in \mathbb{R}^p$ corresponds to the coefficients of the graph Laplacian basis. The speed signal y has a fast decay on the Laplacian basis V; in particular, OLS on the first $p = 50$ basis accounts for over 99% of the variation in y. (That is, $\|V\widehat{\theta}^{\text{ols}} - y\|_2^2 \leq 0.01\|y\|_2^2$, where $\widehat{\theta}^{\text{ols}} = (V^\top V)^{-1}V^\top y$.)

In our experiments we compare the 6 algorithms (L_1 through L_6) when a very small portion of the full samples is selected. The ratio of the mean-square prediction error $\frac{1}{n}\|V\widehat{\theta} - y\|_2^2$ compared to the MLE of the full-sample OLS estimator $\frac{1}{n}\|V\widehat{\theta}^{\text{ols}} - y\|_2^2$ is reported. Apart from the real speed data, we also report results under a "semi-synthetic" setting similar to Sec. 5.2, where the full OLS estimate $\widehat{\theta}^{\text{ols}}$ is first computed and then y is synthesized as $y = V\widehat{\theta}^{\text{ols}} + \varepsilon$ where ε are i.i.d. standard Normal random variables.

From Figure 3, greedy methods $(L_5^*, L_6^\dagger)$ achieve consistently the lowest MSE and are robust to subset size k. However, the Fedorov's exchange algorithm is very slow and requires more than 10 times the running time than our proposed methods (L_4^*, L_5^*). On the other hand, sampling based methods (L_1 through L_4^*) behave quite badly when subset size k is small and close to the problem dimension p. This is because when k is close to p, even very small changes resulted from randomization could lead to highly singular designs and hence significantly increases the mean-square prediction error. We also observe that for
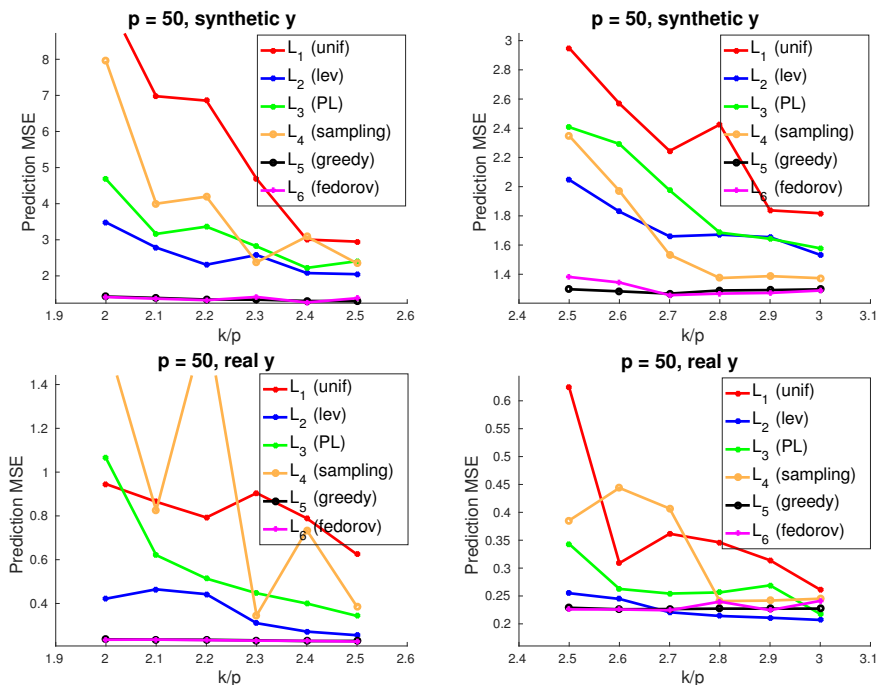
Figure 3: Plots of the ratio of mean square prediction error $\frac{1}{n}\|V\widehat{\theta} - y\|_2^2$ compared against the MSE of the full-sample OLS estimator $\frac{1}{n}\|V\widehat{\theta}^{\text{ols}} - y\|_2^2$ on the Minnesota wind dataset. In the top panel response variables y are synthesized as $y = V\widehat{\theta}^{\text{ols}} + \varepsilon$, where $\widehat{\theta}^{\text{ols}}$ is the full-sample OLS estimator and ε are i.i.d. standard Normal random variables. In the bottom panel the real wind speed response y is used. For randomized algorithms (L_1 to L_4) the experiments are repeated for 50 times and the median MSE ratio is reported.

sufficiently large subset size (e.g., $k = 3p$), the performance gap between all methods is smaller on real signal compared to the synthetic signals. We conjecture that this is because the linear model $y = V\theta + \varepsilon$ only approximately holds in real data.

6. Discussion

We discuss potential improvements in the analysis presented in this paper.

6.1 Sampling based method

To fully understand the finite-sample behavior of the sampling method introduced in Secs. 3.2 and 3.3, it is instructive to relate it to the *graph spectral sparsification* problem (Spielman and Srivastava, 2011) in theoretical computer science: Given a directed weighted graph $G = (V, E, W)$, find a subset $\widetilde{E} \subseteq E$ and new weights $\widetilde{W}$ such that $\widetilde{G} = (V, \widetilde{E}, \widetilde{W})$ is a *spectral sparsification* of G, which means there exists $\epsilon \in (0, 1)$ such that for any vector

$z \in \mathbb{R}^{|V|}$:

$$(1 - \epsilon) \sum_{(u,v) \in E} w_{uv}(z_u - z_v)^2 \leq \sum_{(u,v) \in \widetilde{E}} \widetilde{w}_{uv}(z_u - z_v)^2 \leq (1 + \epsilon) \sum_{u,v \in E} w_{uv}(z_u - z_v)^2.$$

Define $B_G \in \mathbb{R}^{|E| \times |V|}$ to be the *signed edge-vertex incidence matrix*, where each row of B_G corresponds to an edge in E, each column of B_G corresponds to a vertex in G, and $[B_G]_{ij} = 1$ if vertex j is the head of edge i, $[B_G]_{ij} = -1$ if vertex j is the tail of edge i, and $[B_G]_{ij} = 0$ otherwise. The spectral sparsification requirement can then be equivalently written as

$$(1 - \epsilon)z^\top (B_G^\top W B_G)z \leq z^\top (B_{\widetilde{G}}^\top \widetilde{W} B_{\widetilde{G}})z \leq (1 + \epsilon)z^\top (B_G^\top W B_G)z.$$

The similarity of graph sparsification and the experimental selection problem is clear: $B_G \in \mathbb{R}^{n \times p}$ would be the known pool of design points and $W = \text{diag}(\pi^*)$ is a diagonal matrix with the optimal weights π^* obtained by solving Eq. (4). The objective is to seek a small subset of rows in B_G (i.e., the sparsified edge set $\widetilde{E}$) which is a spectral approximation of the original $B_G^\top \text{diag}(\pi^*)B_G$. Approximation of the A-optimality criterion or any other eigen-related quantity immediately follows. One difference is that in linear regression each row of B_G is no longer a $\{\pm 1, 0\}$ vector. Also, the subsampled weight matrix $\widetilde{W}$ needs to correspond to an unweighted graph for linear regression, i.e. diagonal entries of $\widetilde{W}$ must be in $\{0, 1\}$. However, we consider this to be a minor difference as it does not interfere with the spectral properties of B_G.

The spectral sparsification problem where $\widetilde{W}$ can be arbitrarily designed (i.e. not restricted to have $\{0, 1\}$ diagonal entries) is completely solved (Spielman and Srivastava, 2011; Batson et al., 2012), where the size of the selected edge subset is allowed to be *linear* to the number of vertices, or in the terminology of our problem, $k \asymp p$. Unfortunately, both methods require the power of arbitrary designing the weights in $\widetilde{W}$, which is generally not available in experiment selection problems (i.e., cannot set noise variance or signal strength arbitrarily for individual design points). Recently, it was proved that when the original graph is unweighted ($W = I$), it is also possible to find unweighted linear-sized edge sparsifiers ($\widetilde{W} \propto I$) (Marcus et al., 2015a,b; Anderson et al., 2014). This remarkable result leads to the solution of the long-standing Kardison-Singer problem. However, the condition that the *original* weights W are uniform is not satisfied in the linear regression problem, where the optimal solution π^* may be far from uniform. The experiment selection problem somehow falls in between, where an *unweighted* sparsifier is desired for a *weighted* graph. This leads us to the following question:

Question 6.1 *Given a* weighted *graph* $G = (V, E, W)$, *under what conditions are there small edge subset* $\widetilde{E} \subseteq E$ *with* uniform *weights* $\widetilde{W} \propto I$ *such that* $\widetilde{G} = (V, \widetilde{E}, \widetilde{W})$ *is a (one-sided) spectral approximation of* G?

The answer to the above question, especially the smallest possible edge size $|\widetilde{E}|$, would have immediate consequences on finite-sample conditions of k and p in the experiment selection problem.
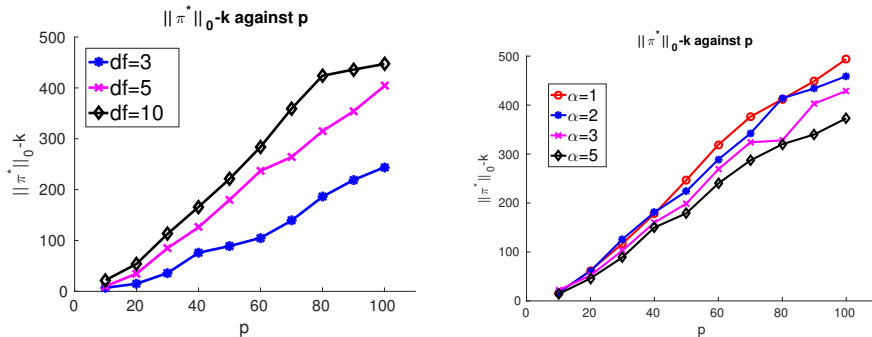
Figure 4: Empirical verification of the rate of $\|\pi^*\|_0$. Left: isotropic design (t distribution); Right: skewed design (transformed multivariate Gaussian) with spectral decay $\lambda_j \propto j^{-\alpha}$.

6.2 Greedy method

Corollary 11 shows that the approximation quality of the greedy based method depends crucially upon $\|\pi^*\|_0$, the support size of the optimal solution π^*. In Lemma 12 we formally established that $\|\pi^*\|_0 \leq k + p(p+1)/2$ under mild conditions; however, we conjecture the $p(p+1)/2$ term is loose and could be improved to $O(p)$ in general cases.

In Fig 4 we plot $\|\pi^*\|_0 - k$ against number of variables p, where p ranges from 10 to 100 and k is set to $3p$. The other simulation settings are kept unchanged. We observe that in all settings $\|\pi^*\|_0 - k$ scales linearly with p, suggesting that $\|\pi^*\|_0 \leq k + O(p)$. Furthermore, the slope of the scalings does not seem to depend on the conditioning of X, as shown in the right panel of Fig 4 where the conditioning of X is controlled by the spectral decay rate α. This is in contrast to the analysis of the sampling based method (Theorem 5), in which the finite sample bound depends crucially upon the conditioning of X for the without replacement setting.

6.3 High-dimensional settings

This paper focuses solely on the so-called *large-sample, low-dimensional* regression setting, where the number of variables p is assumed to be much smaller than both design pool size n and subsampling size k. It is an interesting open question to extend our results to the *high-dimensional* setting, where p is much larger than both n and k, and an assumption like sparsity of β_0 is made to make the problem feasible, meaning that very few components in β_0 are non-zero. In particular, it is desirable to find a sub-sampling algorithm that attains the following minimax estimation rate over high-dimensional sparse models:

$$\inf_{A \in \mathcal{A}_b(k)} \sup_{\|\beta_0\|_0 \leq s} \mathbb{E}_{A,\beta_0} \left[\|\widehat{\beta} - \beta_0\|_2^2 \right].$$

One major obstacle of designing subsampling algorithms for high-dimensional regression is the difficulty of evaluating and optimizing the *restricted eigenvalue* (or other similar criteria) of the subsampled covariance matrix. Unlike the ordinary spectrum, the restricted eigenvalue of a matrix is NP-hard to compute (Dobriban and Fan, 2016) and heavily influences the statistical efficiency of a Lasso-type estimator (Bickel et al., 2009). The question

of optimizing such restricted eigenvalues could be even harder and remains largely open. We also mention that Davenport et al. (2015) suggests a two-step approach where half the budget is used to identify the sparse locations from randomly sampled points and the remaining budget is used to generate a better estimate of regression coefficients using the same convex programming formulation proposed in our paper. Their paper provides some experimental support for this idea, however, no theoretical guarantees are established for the (sub)optimality of such a procedure. Analyzing such a two-step approach could be an interesting future direction.

6.4 Approximate linear models

In cases when the linear model $y = X\beta_0 + \varepsilon$ only approximately holds, we describe here a method that takes into consideration both bias and variance of OLS estimates on sub-sampled data in order to find good sub-samples. Suppose $y = f_0(X) + \varepsilon$ for some un-known underlying function f_0 that might not be linear, and let $\beta^* = (X^\top X)^{-1} X^\top f_0(X)$ be the optimal linear predictor on the full sample X. Suppose $X_S = \Psi X$ is the sub-sampled data, where $\Psi \in \mathbb{R}^{|S| \times n}$ is the subsampling matrix where each row of Ψ is $e_i = (0, \cdots, 0, 1, 0, \cdots, 0) \in \mathbb{R}^n$, with $i \in [n]$ being the subsampled rows in X. Let $\widehat{\beta} = (X_S^\top X_S)^{-1} X_S^\top y_S$ be the OLS on the subsampled data. The error of $\widehat{\beta}$ can then be decomposed and upper bounded as

$$\mathbb{E}\|\widehat{\beta} - \beta^*\|_2^2 = f_0(X)^\top \left[(X^\top \Psi^\top \Psi X)^{-1} X^\top \Psi^\top \Psi - (X^\top X)^{-1} X^\top \right] f_0(X) + \sigma^2 \text{tr} \left[(X^\top \Psi^\top \Psi X)^{-1} \right]$$
(11)

$$\leq \left\| (X^\top \Psi^\top \Psi X)^{-1} X^\top \Psi^\top \Psi - (X^\top X)^{-1} X^\top \right\|_{\text{op}} \cdot \|f_0(X)\|_2^2 + \sigma^2 \text{tr} \left[(X^\top \Psi^\top \Psi X)^{-1} \right].$$
(12)

In the high noise setting, the first term can be ignored and the solution is close to the one considered in this paper. In the low-noise setting, the second term can be ignored and a relaxation similar to Eq. (4) can be derived; however a linear approximation may be undesirable in this case. In general, when $\|f_0(X)\|_2^2 \approx \lambda$ is known or can be estimated, the following continuous optimization problem serves as an approximate objective of subsampled linear regression with approximate linear models:

$$\min_{\pi \in \mathbb{R}^n} \quad \lambda \left\| \left(\sum_{i=1}^n \pi_i x_i x_i^\top \right)^{-1} X^\top \text{diag}(\pi) - (X^\top X)^{-1} X^\top \right\|_{\text{op}} + \sigma^2 \text{tr} \left[\left(\sum_{i=1}^n \pi_i x_i x_i^\top \right)^{-1} \right]$$

$$s.t. \quad \sum_{i=1}^n \pi_i \leq k, \quad 0 \leq \pi_i \leq 1.$$

Unfortunately, the relaxation in Eq. (12) may be loose and hence the optimization problem above fails to serve as a good objective for the optimal subsampling problem with approximate linear models. Exact characterization of the bias term in Eq. (11) without strong assumptions on f_0 is an interesting open question.

7. Proofs

7.1 Proof of facts in Sec. 3.1

Proof [Proof of Fact 3.1] Let $A = X^\top \text{diag}(\pi) X$ and $B = X^\top \text{diag}(\pi') X$. By definition, $B = A + \Delta$ where $\Delta = X^\top \text{diag}(\pi' - \pi) X$ is positive semi-definite. Subsequently, $\sigma_\ell(B) \geq \sigma_\ell(A)$ for all $\ell = 1, \cdots, p$. We then have that

$$f(\pi'; X) = \text{tr}(B^{-1}) = \sum_{\ell=1}^{p} \sigma_\ell(B)^{-1} \leq \sum_{\ell=1}^{p} \sigma_\ell(A)^{-1} = \text{tr}(A^{-1}) = f(\pi; X).$$

Note also that if $\pi' \neq \pi$ then $\Delta \neq 0$ and hence there exists at least one ℓ with $\sigma_\ell(B) < \sigma_\ell(A)$. Therefore, the equality holds if and only if $\pi = \pi'$. ∎

Proof [Proof of Fact 3.2] Suppose $\|\pi^*\|_1 < k$. Then there exists some coordinate $j \in \{1, \cdots, n\}$ such that $\pi_j^* < 1$. Define π' as $\pi_i' = \pi_i^*$ for $i \neq j$ and $\pi_j' = \min\{1, \pi_j^* + k - \|\pi^*\|_1\}$. Then π' is also a feasible solution. On the other hand, By Fact 3.1 we have that $f(\pi'; X) < f(\pi^*; X)$, contradicting the optimality of π^*. Therefore, $\|\pi^*\|_1 = k$. ∎

7.2 Proof of Theorem 3

We only prove Theorem 3 for the with replacement setting ($b = 1$). The proof for the without replacement setting is almost identical.

Define $\widetilde{\mathcal{A}}_1(k)$ as the class of *deterministic* algorithms that proceed as follows:

1. The algorithm deterministically outputs pairs $\{(w_i, x_i)\}_{i=1}^{M}$, where x_i is one of the rows in X and $\{w_i\}_{i=1}^{n}$ satisfies $w_i \geq 0$, $\sum_{i=1}^{M} w_i \leq k$. Here M is an arbitrary finite integer.

2. The algorithm observes $\{y_i\}_{i=1}^{M}$ with $y_i = \sqrt{w_i} x_i^\top \beta + \varepsilon_i$. Here β is a fixed but unknown regression model and $\varepsilon_i \overset{i.i.d.}{\sim} \mathcal{N}(0, \sigma^2)$ is the noise.

3. The algorithm outputs $\widehat{\beta}$ as an estimation of β, based on the observations $\{y_i\}_{i=1}^{M}$.

Because all algorithms in $\widetilde{\mathcal{A}}_1(k)$ are deterministic and the design matrix still has full column rank, the optimal estimator of $\widehat{\beta}$ *given* $\{(w_i, x_i)\}_{i=1}^{M}$ is the OLS estimator. For a specific data set weighted through $\sqrt{w_1}, \cdots, \sqrt{w_M}$, the minimax estimation error (which is achieved by OLS) is given by

$$\inf_{\widehat{\beta}} \sup_{\beta} \mathbb{E}\left[\|\widehat{\beta} - \beta\|_2^2\right] = \sup_{\beta} \mathbb{E}\left[\|\widehat{\beta}^{\text{ols}} - \beta\|_2^2\right] = \sigma^2 \text{tr}\left[(\widetilde{X}^\top \widetilde{X})^{-1}\right] = \sigma^2 \text{tr}\left[\left(\sum_{i=1}^{n} \widetilde{w}_i x_i x_i^\top\right)^{-1}\right],$$

where $\widetilde{w}_i$ is the aggregated weight of data point x_i in all the M weighted pairs. Subsequently,

$$\inf_{\widetilde{A} \in \widetilde{\mathcal{A}}_1(k)} \sup_{\beta} \mathbb{E}\left[\|\widehat{\beta}_{\widetilde{A}} - \beta\|_2^2\right] = \inf_{w_1 + \cdots + w_M \leq k} \sigma^2 \text{tr}\left[\left(\sum_{i=1}^{M} w_i x_i x_i^\top\right)^{-1}\right] = \sigma^2 f_1^*(k; X).$$

25

It remains to prove that

$$\inf_{\widetilde{A} \in \widetilde{\mathcal{A}}_1(k)} \sup_\beta \mathbb{E}\left[\|\widehat{\beta}_{\widetilde{A}} - \beta\|_2^2\right] \leq \inf_{A \in \mathcal{A}_1(k)} \sup_\beta \mathbb{E}\left[\|\widehat{\beta}_A - \beta\|_2^2\right].$$

We prove this inequality by showing that for every (possibly random) algorithm $A \in \mathcal{A}_1(k)$, there exists $\widetilde{A} \in \widetilde{\mathcal{A}}_1(k)$ such that $\sup_\beta \mathbb{E}[\|\widehat{\beta}_A - \beta\|_2^2] \geq \sup_\beta \mathbb{E}[\|\widehat{\beta}_{\widetilde{A}} - \beta\|_2^2]$. To see this, we construct $\widetilde{A}$ based on A as follows:

1. For every k-subset (duplicates allowed) of all possible outputs of A (which by definition are all subsets of X) and its corresponding weight vector w, add $(w_i', \widetilde{x}_i)$ to the design set of $\widetilde{A}$, where $w_i' = w_i \Pr_A(\widetilde{X})$.

2. The algorithm $\widetilde{A}$ observes all responses $\{y_i\}$ for $\{(w_i', \widetilde{x}_i)\}$.

3. $\widetilde{A}$ outputs the *expected* estimation of A; that is, $\widetilde{A}(X, y) = \mathbb{E}_{\widetilde{X}}[\mathbb{E}_y[\widehat{\beta}_A|\widetilde{X}]]$. Note that by definition of the estimator class $\mathcal{A}_1(k)$, all estimators $\widehat{\beta}_A$ *conditioned on subsampled data points* $\widetilde{X}$ are deterministic.

We claim that $\widetilde{A} \in \widetilde{\mathcal{A}}_1(k)$ because

$$\sum_i w_i' = \sum_{\widetilde{X}} \Pr(\widetilde{X}) \sum_{i=1}^k w_i \leq k.$$

Furthermore, by Jensen's inequality we have

$$\mathbb{E}_{\widetilde{X},y}\left[\|\widehat{\beta}_A - \beta\|_2^2\right] \geq \mathbb{E}_y\left[\|\mathbb{E}_{\widetilde{X}}[\widehat{\beta}_A] - \beta\|_2^2\right] = \mathbb{E}_y\left[\|\widehat{\beta}_{\widetilde{A}} - \beta\|_2^2\right].$$

Taking supreme over β we complete the proof.

7.3 Proof of Lemma 6

With replacement setting Define $\Phi = \mathrm{diag}(\pi^*)$ and $\Pi = \Phi^{1/2} X \Sigma_*^{-1} X^\top \Phi^{1/2} \in \mathbb{R}^{n \times n}$. The following proposition lists properties of Π:

Proposition 14 (Properties of projection matrix) *The following properties for Π hold:*

1. *Π is a projection matrix. That is, $\Pi^2 = \Pi$.*

2. *$\mathrm{Range}(\Pi) = \mathrm{Range}(\Phi^{1/2} X)$.*

3. *The eigenvalues of Π are 1 with multiplicity p and 0 with multiplicity $n - p$.*

4. *$\Pi_{ii} = \|\Pi_{i,\cdot}\|_2^2 = \pi_i^* x_i^\top \Sigma_*^{-1} x_i$.*

Proof Proof of 1: By definition, $\Sigma_* = X^\top \Phi X$ and subsequently

$$\begin{aligned} \Pi^2 &= \Phi^{1/2} X \Sigma_*^{-1} X^\top \Phi^{1/2} \Phi^{1/2} X \Sigma_*^{-1} X^\top \Phi^{1/2} \\ &= \Phi^{1/2} X (X^\top \Phi X)^{-1} X^\top \Phi X (X^\top \Phi X)^{-1} \Phi^{1/2} \\ &= \Phi^{1/2} X (X^\top \Phi X)^{-1} X^\top \Phi^{1/2} = \Pi. \end{aligned}$$

Proof of 2: First note that $\text{Range}(\Pi) = \text{Range}(\Phi^{1/2} X \Sigma_*^{-1} X^\top \Phi^{1/2}) \subseteq \text{Range}(\Phi^{1/2} X)$. For the other direction, take arbirary $u \in \text{Range}(\Phi^{1/2} X)$ and express u as $u = \Phi^{1/2} X v$ for some $v \in \mathbb{R}^p$. We then have

$$
\begin{aligned}
\Pi u &= \Phi^{1/2} X \Sigma_*^{-1} X^\top \Phi^{1/2} u \\
&= \Phi^{1/2} X (X^\top \Phi X)^{-1} X^\top \Phi^{1/2} \Phi^{1/2} X v \\
&= \Phi^{1/2} X v = u
\end{aligned}
$$

and hence $u \in \text{Range}(\Pi)$.

Proof of 3: Because $\Sigma_* = X^\top \Phi X$ is invertible, the $n \times p$ matrix $\Phi^{1/2} X$ must have full column rank and hence $\ker(\Phi^{1/2} X) = \{0\}$. Consequently, $\dim(\text{Range}(\Pi)) = \dim(\text{Range}(\Phi^{1/2} X)) = p - \dim(\ker(\Phi^{1/2} X)) = p$. On the other hand, the eigenvalues of Π must be either 0 or 1 because Π is a projection matrix. So the eigenvalues of Π are 1 with multiplicity p and 0 with multiplicity $n - p$.

Proof of 4: By definition,

$$
\Pi_{ii} = \sqrt{\pi_i^*} x_i^\top \Sigma_*^{-1} x_i \sqrt{\pi_i^*} = \pi_i^* x_i^\top \Sigma_*^{-1} x_i.
$$

In addition, Π is a symmetric projection matrix. Therefore,

$$
\Pi_{ii} = [\Pi^2]_{ii} = \Pi_{i,\cdot}^\top \Pi_{i,\cdot} = \|\Pi_{i,\cdot}\|_2^2.
$$

$\blacksquare$

The following lemma shows that a spectral norm bound over deviation of the projection matrix implies spectral approximation of the underlying (weighted) covariance matrix.

Lemma 15 (Spielman and Srivastava (2011), Lemma 4) *Let $\Pi = \Phi^{1/2} X \Sigma_*^{-1} X^\top \Phi^{1/2}$ and W be an $n \times n$ non-negative diagonal matrix. If $\|\Pi W \Pi - \Pi\|_2 \leq \epsilon$ for some $\epsilon \in (0, 1/2)$ then*

$$
(1 - \epsilon) u^\top \Sigma_* u \leq u^\top \widetilde{\Sigma}_* u \leq (1 + \epsilon) u^\top \Sigma_* u, \quad \forall u \in \mathbb{R}^p,
$$

where $\Sigma_* = X^\top \Phi X$ and $\widetilde{\Sigma}_* = X^\top W^{1/2} \Phi W^{1/2} X$.

We next proceed to find an appropriate diagonal matrix W and validate Lemma 15. Define $w_j^* = \pi_j^*/(k p_j^{(2)})$ and $\widehat{\Sigma}_{\widehat{W}} = \sum_{t=1}^k w_{i_t}^* x_{i_t} x_{i_t}^\top$. It is obvious that $\widehat{\Sigma}_{\widehat{W}} \preceq \widehat{\Sigma}_{\widehat{S}}$ because $w_i = \lceil w_i^* \rceil \geq w_i^*$. It then suffices to lower bound the spectrum of $\widehat{\Sigma}_{\widehat{W}}$ by the spectrum of $\widehat{\Sigma}_*$. Define random diagonal matrix $W^{(1)}$ as ($\mathbb{I}[\cdot]$ is the indicator function)

$$
W_{jj}^{(1)} = \frac{\sum_{t=1}^k w_{i_t}^* \mathbb{I}[i_t = j]}{\pi_j^*}, \quad j = 1, \cdots, n.
$$

Then by definition, $\widetilde{\Sigma}_* = X^\top W^{1/2} \Phi W^{1/2} X = \sum_{t=1}^k w_{i_t}^* x_{i_t} x_{i_t}^\top = \widehat{\Sigma}_{\widehat{R}}$. The following lemma bounds the perturbation $\|\Pi W \Pi - \Pi\|_2$ for this particular choice of W.

Lemma 16 *For any $\epsilon > 0$,*

$$
\Pr\left[\|\Pi W^{(1)} \Pi - \Pi\|_2 > \epsilon\right] \leq 2 \exp\left\{-C \cdot \frac{k \epsilon^2}{p \log k}\right\},
$$

where $C > 0$ is an absolute constant.

Proof Define n-dimensional random vector v as [3]

$$\Pr\left[v = \sqrt{\frac{kw_j^*}{\pi_j^*}}\Pi_{j\cdot}\right] = p_j^{(2)}, \qquad j = 1, \cdots, n.$$

Let $v_1, \cdots, v_k$ be i.i.d. copies of v and define $A_t = v_t v_t^\top$. By definition, $\Pi W^{(1)}\Pi$ is equally distributed with $\frac{1}{k}\sum_{t=1}^k A_t$. In addition,

$$\mathbb{E}A_t = \sum_{j=1}^n \frac{kw_j^*}{\pi_j^*}p_j^{(2)}\Pi_{j\cdot}.\Pi_{j\cdot}^\top = \Pi^2 = \Pi,$$

which satisfies $\|\mathbb{E}A_t\|_2 = 1$, and

$$\|A_t\|_2 = \|v_t\|_2^2 \le \sum_{1 \le j \le n} \frac{kw_j^*}{\pi_j^*}\|\Pi_{j\cdot}\|_2^2 \le \sup_{1 \le j \le n} \frac{p}{x_j^\top \Sigma_*^{-1} x_j} \cdot x_j^\top \Sigma_*^{-1} x_j = p.$$

Applying Lemma 18 we have that

$$\Pr\left[\|\Pi W^{(1)}\Pi - \Pi\|_2 > \epsilon\right] \le 2\exp\left\{-C \cdot \frac{k\epsilon^2}{p\log k}\right\}.$$

∎

With Lemma 16, we know that $\|\Pi W\Pi - \Pi\|_2 \le \epsilon$ holds with probability at least 0.9 if $\frac{p\log k}{k} = O(\epsilon^2)$. Eq. (6) then holds with high probability by Lemma 15.

Without replacement setting Define independently distributed random matrices $A_1, \cdots, A_n$ as

$$A_j = (w_j - \pi_j^*)x_j x_j^\top, \qquad j = 1, \cdots, n.$$

Note that w_j is a random Bernoulli variable with $\Pr[w_j = 1] = kp_j^{(2)} = \pi_j^*$. Therefore, $\mathbb{E}A_j = 0$. In addition,

$$\sup_{1 \le j \le n} \|A_j\|_2 \le \sup_{1 \le j \le n} \|x_j\|_2^2 \le \|X\|_\infty^2 \quad a.s.$$

and

$$\left\|\sum_{j=1}^n \mathbb{E}A_j^2\right\|_2 = \left\|\sum_{j=1}^n \pi_j^*(1 - \pi_j^*)\|x_j\|_2^2 x_j x_j^\top\right\|_2 \le \|X\|_\infty^2\|\Sigma_*\|_2.$$

Noting that $\sum_{j=1}^n A_j = X_{\widehat{S}}^\top X_{\widehat{S}} - X^\top \text{diag}(\pi^*)X = \widehat{\Sigma}_{\widehat{S}} - \Sigma_*$ and invoking Lemma 19 with $t = \epsilon\lambda_{\min}(\Sigma_*)$ we have that

$$\Pr\left[\|\widehat{\Sigma}_{\widehat{S}} - \Sigma_*\|_2 > \epsilon\lambda_{\min}(\Sigma_*)\right] \le 2p\exp\left\{-\frac{\epsilon^2\lambda_{\min}(\Sigma_*)^2}{3\|\Sigma_*\|_2\|X\|_\infty^2 + 2\|X\|_\infty^2 \cdot \epsilon\lambda_{\min}(\Sigma_*)}\right\}.$$

---

3. For those j with $\pi_j^* = 0$, we have by definition that $p_j^{(2)} = 0$.

Equating the right-hand side with $O(1)$ we have that

$$\|\Sigma_*^{-1}\|_2 \kappa(\Sigma_*) \|X\|_\infty^2 \log p = O(\epsilon^2).$$

Finally, by Weyl's theorem we have that

$$\left| \lambda_\ell(\widehat{\Sigma}_{\widehat{S}}) - \lambda_\ell(\Sigma_*) \right| \le \|\widehat{\Sigma}_{\widehat{S}} - \Sigma_*\|_2 \le \epsilon \lambda_{\min}(\Sigma_*) \le \epsilon \lambda_\ell(\Sigma_*),$$

and hence the proof of Lemma 6.

7.4 Proof of Lemma 8

With replacement setting   Define $\widetilde{w}_j^* = k/T \cdot w_j^* = \pi_j^*/(Tp_j^{(1)})$ and let $\widehat{\Sigma}_T = \sum_{t=1}^T \widetilde{w}_{i_t}^* x_{i_t} x_{i_t}^\top$. Because $\widehat{\Sigma}_T = \frac{k}{T} \widehat{\Sigma}_{\widehat{S}_W} \preceq \widehat{\Sigma}_{\widehat{S}}$, we have that $\widehat{\Sigma}_{\widehat{S}} \succeq \frac{T}{k} \widehat{\Sigma}_T$ and hence $z^\top \widehat{\Sigma}_{\widehat{S}} z \ge \frac{T}{k} z^\top \widehat{\Sigma}_T z$ for all $z \in \mathbb{R}^p$. Therefore, to lower bound the spectrum of $\widehat{\Sigma}_{\widehat{S}}$ it suffices to lower bound the spectrum of $\widehat{\Sigma}_T$.

Define diagonal matrix $W^{(2)}$ as

$$W_{jj}^{(2)} = \frac{\sum_{t=1}^T \widetilde{w}_{i_t}^* \mathbb{I}[i_t = j]}{\pi_j^*}, \qquad j = 1, \cdots, n.$$

We have that $\widetilde{\Sigma}_* = \widehat{\Sigma}_T$ for this particular choice of W. Following the same analysis in the proof of Lemma 16, we have that for every $t > 0$

$$\Pr\left[\|\Pi W^{(2)} \Pi - \Pi\|_2 > t \right] \le 2 \exp\left\{ -C \cdot \frac{T\epsilon^2}{p \log T} \right\}.$$

Set $t = O(1)$ and equate the right-hand side of the above inequality with $O(1)$. We then have

$$\Pr\left[\widehat{\Sigma}_T \succeq \Omega(1) \cdot \widehat{\Sigma}_* \right] = \Omega(1) \qquad \text{if } p \log T/T = O(1).$$

Subsequently, under the condition that $p \log T/T = O(1)$, with probability at least 0.9 it holds that

$$\widehat{\Sigma}_{\widehat{S}} \succeq \Omega(T/k) \cdot \Sigma_*,$$

which completes the proof of Lemma 8 for the without replacement setting.

Without replacement setting   Define $\widehat{\Sigma}_{\widehat{R}} = X_{R_T}^\top X_{R_T} = \sum_{t=1}^T \pi_{i_t} x_{i_t} x_{i_t}^\top$. Conditioned on R_T, the subset $\widehat{S} = S_T$ is selected using the same procedure of the soft-constraint algorithm Figure 1 on $X_{R_T} = \{x_i\}_{i \in R_T}$. Subsequently, following analysis in the proof of Lemma 6 we have

$$\Pr\left[\|\widehat{\Sigma}_{\widehat{S}} - \widehat{\Sigma}_T\|_2 > t \big| R_T \right] \le 2p \cdot \exp\left\{ -\frac{t^2}{3\|\widehat{\Sigma}_T\|_2 \|X\|_\infty^2 + 2\|X\|_\infty^2 t} \right\}.$$

Setting $t = O(1) \cdot \lambda_{\min}(\widehat{\Sigma}_T)$ we have that, if $\|\widehat{\Sigma}_T^{-1}\|_2 \kappa(\widehat{\Sigma}_T) \|X\|_\infty^2 \log p = O(1)$, then with probability at least 0.95 conditioned on $\widehat{\Sigma}_T$

$$\Omega(1) \cdot \widehat{\Sigma}_T \preceq \widehat{\Sigma}_{\widehat{S}} \preceq O(1) \cdot \widehat{\Sigma}_T. \tag{13}$$

It remains to establish spectral similarity between $\widehat{\Sigma}_T$ and $\frac{T}{n}\Sigma_*$, a scaled version of Σ_*. Define deterministic matrices $A_1, \cdots, A_n$ as

$$A_j = \pi_j^* x_j x_j^\top - \frac{1}{n}\Sigma_*, \qquad j = 1, \cdots, n.$$

By definition, $\sum_{j=1}^n A_j = 0$ and $\sum_{t=1}^T A_{\sigma(t)} = \widehat{\Sigma}_T - \frac{T}{n}\Sigma_*$, where σ is a random permutation from $[n]$ to $[n]$. In addition,

$$\sup_{1 \le j \le n} \|A_j\|_2 \le \frac{1}{n}\|\Sigma_*\|_2 + \sup_{1 \le j \le n} \|x_j\|_2^2 \le 2\|X\|_\infty^2$$

and

$$
\begin{aligned}
\frac{T}{n}\left\|\sum_{j=1}^n A_j^2\right\|_2 &\le \frac{2T}{n}\left(\left\|\sum_{j=1}^n (\pi_j^*)^2 \|x_i\|_2^2 x_i x_i^\top\right\|_2 + \frac{1}{n^2}\|\Sigma_*\|_2^2\right)\\
&\le \frac{2T}{n}\left(\|X\|_\infty^2 \left\|\sum_{j=1}^n \pi_j^* x_i x_i^\top\right\|_2 + \frac{1}{n^2}\|\Sigma_*\|_2^2\right)\\
&\le \frac{2T}{n}\left(\|X\|_\infty^2\|\Sigma_*\|_2 + \frac{1}{n^2}\|\Sigma_*\|_2^2\right)\\
&\le \frac{4T}{n}\|X\|_\infty^2\|\Sigma_*\|_2.
\end{aligned}
$$

Invoking Lemma 20, we have that

$$\Pr\left[\left\|\widehat{\Sigma}_T - \frac{T}{n}\Sigma_*\right\|_2 > t\right] \le p\exp\left\{-t^2\left[\frac{48T}{n}\|X\|_\infty^2\|\Sigma_*\|_2 + 8\sqrt{2}\|X\|_\infty^2 t\right]^{-1}\right\}.$$

Set $t = O(T/n) \cdot \lambda_{\min}(\Sigma_*)$. We then have that, if $\|\Sigma_*^{-1}\kappa(\Sigma_*)\|X\|_\infty^2 \log p = O(T/n)$ holds, then with probability at least 0.95

$$\Omega(T/n) \cdot \Sigma_* \preceq \widehat{\Sigma}_T \preceq O(T/n) \cdot \Sigma_*. \tag{14}$$

Combining Eqs. (13,14) and noting that $\|\widehat{\Sigma}_T^{-1}\|_2 \le O(\frac{n}{T})\|\Sigma_*^{-1}\|_2$, $\kappa(\Sigma_T) \le O(1)\kappa(\Sigma_*)$, we complete the proof of Lemma 8 under the without replacement setting.

7.5 Proof of Lemma 12

Let $f(\pi; \lambda, \widetilde{\lambda}, \mu)$ be the Lagrangian muliplier function of the without replacement formulation of Eq. (4):

$$f(\pi; \lambda, \widetilde{\lambda}, \mu) = f(\pi; X) - \sum_{i=1}^n \lambda_i \pi_i + \sum_{i=1}^n \widetilde{\lambda}_i\left(\pi_i - \frac{1}{k}\right) + \mu\left(\sum_{i=1}^n \pi_i - 1\right).$$

Here $\{\lambda_i\}_{i=1}^n \ge 0$, $\{\widetilde{\lambda}_i\}_{i=1}^n \ge 0$ and $\mu \ge 0$ are Lagrangian multipliers for constraints $\pi_i \ge 0$, $\pi_i \le 1$ and $\sum_i \pi_i \le k$, respectively. By KKT condition, $\frac{\partial f}{\partial \pi_i}\big|_{\pi^*} = 0$ and hence

$$-\frac{\partial f}{\partial \pi_i}\bigg|_{\pi^*} = x_i^\top \Sigma_*^{-2} x_i = \widetilde{\lambda}_i - \lambda_i + \mu, \qquad i = 1, \cdots, n,$$

where $\Sigma_* = X^\top \mathrm{diag}(\pi^*)X$ is a $p \times p$ positive definite matrix.

Split the index set $[n]$ into three disjoint sets defined as $A = \{i \in [n] : \pi_i^* = 1\}$, $B = \{i \in [n] : 0 < \pi_i^* < 1\}$ and $C = \{i \in [n] : \pi_i^* = 0\}$. Note that $\|\pi^*\|_0 = |A| + |B|$ and $|A| \leq k$. Therefore, to upper bound $\|\pi^*\|_0$ it suffices to upper bound $|B|$. By complementary slackness, for all $i \in B$ we have that $\widetilde{\lambda}_i = \lambda_i = 0$; that is,

$$x_i^\top \Sigma_*^{-2} x_i = \langle \phi(x_i), \psi(\Sigma_*^{-2}) \rangle = \mu, \qquad \forall i \in B, \tag{15}$$

where $\phi : \mathbb{R}^p \to \mathbb{R}^{p(p+1)/2}$ is the mapping defined in Assumption 3.1 and $\psi(\cdot)$ takes the upper triangle of a symmetric matrix and vectorizes it into a $\frac{p(p-1)}{2}$-dimensional vector. Assume by way of contradiction that $|B| > p(p+1)/2$ and let $x_1, \cdots, x_{p(p+1)/2+1}$ be arbitrary distinct $\frac{p(p+1)}{2} + 1$ rows whose indices belong to B. Eq. (15) can then be cast as a homogenous linear system with $\frac{p(p+1)}{2} + 1$ variables and equations as follows:

$$\begin{bmatrix} \widetilde{\phi}(x_1) \\ \widetilde{\phi}(x_2) \\ \vdots \\ \widetilde{\phi}(x_{p(p+1)/2+1}) \end{bmatrix} \begin{bmatrix} \psi(\Sigma_*^{-2}) \\ -\mu \end{bmatrix} = 0.$$

Under Assumption 3.1, $\widetilde{\Phi} = [\widetilde{\phi}(x_1); \cdots ; \widetilde{\phi}(x_{p(p+1)/2+1})]^\top$ is invertible and hence both $\psi(\Sigma_*^{-2})$ and μ must be zero. This contradicts the fact that Σ_*^{-2} is positive definite.

Appendix A. Technical lemmas

Lemma 17 *Let $\mathbb{B} = \{x \in \mathbb{R}^p : \|x\|_2^2 \leq B^2\}$ and $\mathrm{vol}(\mathbb{B}) = \int_\mathbb{B} 1\mathrm{d}x$ be the volume of $\mathbb{B}$. Then $\int_\mathbb{B} xx^\top \mathrm{d}x = \frac{B^2}{p+2}\mathrm{vol}(\mathbb{B})I_{p \times p}$.*

Proof Let U be the uniform distribution in the p-dimensional ball of radius B. By definition, $\int_\mathbb{B} xx^\top \mathrm{d}x = \mathrm{vol}(\mathbb{B})\mathbb{E}_{x \sim U}[xx^\top]$. By symmetry, $\mathbb{E}_{x \sim U}[xx^\top] = c \cdot I_{p \times p}$ for some constant c that depends on p and B. To determine the constant c, note that

$$c = \mathbb{E}[x_1^2] = \frac{1}{p}\mathbb{E}\|x\|_2^2 = \frac{1}{p}\frac{\int_0^B r^{p-1} \cdot r^2 \mathrm{d}r}{\int_0^B r^{p-1}\mathrm{d}r} = \frac{B^2}{p+2}.$$

■

Lemma 18 (Rudelson and Vershynin (2007)) *Let x be a p-dimensional random vector such that $\|x\|_2 \leq M$ almost surely and $\|\mathbb{E}xx^\top\|_2 \leq 1$. Let $x_1, \cdots, x_n$ be i.i.d. copies of x. Then for every $t \in (0,1)$*

$$\Pr\left[\left\|\frac{1}{n}\sum_{i=1}^n x_i x_i^\top - \mathbb{E}xx^\top\right\|_2 > t\right] \leq 2\exp\left\{-C \cdot \frac{nt^2}{M^2 \log n}\right\},$$

where $C > 0$ is some universal constant.

Lemma 19 (Corollary 5.2 of (Mackey et al., 2014), Matrix Bernstein) *Let $(Y_k)_{k \geq 1}$ be a sequence of random d-dimensional Hermitian matrices that satisfy*

$$\mathbb{E}Y_k = 0 \quad and \quad \|Y_k\|_2 \leq R \quad a.s.$$

Define $X = \sum_{k \geq 1} Y_k$. The for any $t > 0$,

$$\Pr\left[\|X\|_2 \geq t\right] \leq d \cdot \exp\left\{-\frac{t^2}{3\sigma^2 + 2Rt}\right\} \quad for \ \sigma^2 = \left\|\sum_{k \geq 1} \mathbb{E}Y_k^2\right\|_2.$$

Lemma 20 (Corollary 10.3 of (Mackey et al., 2014)) *Let $A_1, \cdots, A_n$ be a sequence of deterministic d-dimensional Hermitian matrices that satisfy*

$$\sum_{k=1}^n A_k = 0 \quad and \quad \sup_{1 \leq k \leq n} \|A_k\|_2 \leq R.$$

Define random matrix $X = \sum_{j=1}^m A_{\sigma(j)}$ for $m \leq n$, where σ is a random permutation from $[n]$ to $[n]$. Then for all $t > 0$,

$$\Pr\left[\|X\|_2 \geq t\right] \leq d \exp\left\{-\frac{t^2}{12\sigma^2 + 4\sqrt{2}Rt}\right\} \quad for \ \sigma^2 = \frac{m}{n}\left\|\sum_{k=1}^n A_k^2\right\|_2.$$

Appendix B. Optimization methods

Two algorithms for optimizing Eq. (4) are described. The SDP formulation is of theoretical interest only and the projected gradient descent algorithm is practical, which also enjoys theoretical convergence guarantees.

SDP formulation For $\pi \in \mathbb{R}^n$ define $A(\pi) = \sum_{i=1}^n \pi_i x_i x_i^\top$, which is a $p \times p$ positive semidefinite matrix. By definition, $f(\pi; X) = \sum_{j=1}^p e_j^\top A(\pi)^{-1} e_j$, where e_j is the p-dimensional vector with only pth coordinate being 1. Subsequently, Eq. (4) is equivalent to the following SDP problem:

$$\min_{\pi, t} \sum_{j=1}^p t_j \quad \text{subject to} \quad 0 \leq \pi_i \leq 1, \ \sum_{i=1}^n \pi_i \leq k, \ \text{diag}(B_1, \cdots, B_p) \succeq 0,$$

where

$$B_j = \begin{bmatrix} A(\pi) & e_j \\ e_j^\top & t_j \end{bmatrix}, \quad j = 1, \cdots, p.$$

Global optimal solution of an SDP can be computed in polynomial time (Vandenberghe and Boyd, 1996). However, this formulation is not intended for practical computation because of the large number of variables in the SDP system. First-order methods such as projected gradient descent is a more appropriate choice for practical computation.

Projected gradient descent For any convex set S and point x let $\mathcal{P}_S(x) = \operatorname{argmin}_{y \in S} \|x - y\|_2$ denote the ℓ_2 projection of x onto S. The projected gradient descent algorithm is a general purpose method to solve convex constrained smooth convex optimization problems of the form

$$\min_x f(x) \quad \text{subject to} \quad x \in S.$$

The algorithm (with step size selected via backtracking line search) iterates until desired optimization accuracy is reached:

---

input : backtracking parameters $\alpha \in (0, 1/2]$, $\beta \in (0, 1)$.
output: $\widehat{x}$, approximate solution of the optimization problem.
Initialization: x_0, $t = 0$.
1. Compute gradient $g_t = \nabla f(x_t)$.
2. Find the smallest integer $s \geq 0$ such that $f(x') - f(x_t) \leq \alpha g_t^\top (x' - x_t)$, where $x' = \mathcal{P}_S(x_t - \beta^s g_t)$.
3. Set $x_{t+1} = \mathcal{P}_S(x_t - \beta^s g_t)$, $t \leftarrow t + 1$ and repeat steps 1 and 2, until the desired accuracy is achieved. Output $\widehat{x} = x_t$.

Figure 4: The projected gradient descent algorithm.

The gradient $\nabla_\pi f(\pi; X)$ in Eq. (4) is easy to compute:

$$\frac{\partial f(\pi; X)}{\partial \pi_i} = -x_i^\top (X^\top \operatorname{diag}(\pi) X)^{-1} x_i, \quad i = 1, \cdots, n.$$

Because $X^\top \operatorname{diag}(\pi) X$ is a shared term, computing $\nabla_\pi f(\pi; X)$ takes $O(np^2 + p^3)$ operations. The projection step onto the intersection of ℓ_1 and ℓ_∞ balls is complicated and non-trivial, which we describe in details in Appendix D. In general, the projection step can be done in $O(n \log \|\pi\|_\infty)$ time, where π is the point to be projected.

The following proposition establishes convergence guarantee for the projected gradient descent algorithm. Its proof is given in the appendix.

Proposition 21 *Let $\pi^{(0)}$ be the "flat" initialization (i.e., $\pi_i^{(0)} = k/n$) and $\pi^{(t)}$ be the solution after t projected gradient iterations. Then*

$$f(\pi^{(t)}; X) - f(\pi^*; X) \leq \frac{\|X\|_2^4 [\operatorname{tr}(\widehat{\Sigma}^{-1})]^3 \|\pi^{(0)} - \pi^*\|_2^2}{\beta t},$$

where $\widehat{\Sigma} = \frac{1}{n} X^\top X$.

We also remark that the provided convergence speed is very conservative, especially in the cases when k is large where practical evidence suggests that the algorithm converges in very few iterations (cf. Sec. 4).

Appendix C. Fedorov's exchange algorithm

The algorithm starts with a initial subset $S \subseteq [n]$, $|S| \leq k$, usually initialized with random indices. A "best" pair of exchanging indices are computed as

$$i^*, j^* = \operatorname{argmin}_{i \in S, j \notin S} \ell(S \backslash \{i\} \cup \{j\}),$$

where $\ell(S)$ is the objective function to be minimized. In our case it would be the A-optimality objective $\ell(S) = F(S; X) = \text{tr}((X_S^\top X_S)^{-1})$. The algorithm then "exchanges" i^* and j^* by setting $S' \leftarrow S\backslash\{i^*\} \cup \{j^*\}$ and continues such exchanges until no exchange can lower the objective value. Under without replacement settings, special care needs to be taken to ensure that S consists of distinct indices.

Computing the objective function $F(S; X) = \text{tr}((X_S^\top X_S)^{-1})$ requires inverting a $p \times p$ matrix, which could be computationally slow. A more computationally efficient approach is to perform rank-1 update of the inverse of $X_S^\top X_S$ after each iteration, via the Sherman-Morrison formula:

$$(A + uv^\top)^{-1} = A^{-1} - \frac{A^{-1}uv^\top A^{-1}}{1 + u^\top A^{-1}v}.$$

Each exchange would then take $O(nkp^2)$ operations. The total number of exchanges, however, is unbounded and could be as large as $\binom{n}{k}$ in theory. In practice we do observe that a large number of exchanges are required in order to find a local optimal solution.

Appendix D. Projection onto the intersection of ℓ_1 and ℓ_∞ norm balls

Similar to (Yu et al., 2012), we only need to consider the case that π lies in the first quadrant, i.e., $\pi_i \geq 0, i = 1, 2, ..., n$. Then the projection x also lies in the first quadrant. Furthermore, we assume the point to be project lies in the area which is out of both norm balls and projection purely onto either ℓ_1 or ℓ_∞ ball is not the intersection of both. Otherwise, it is trivial to conduct the projection. The projection problem is formulated as follows:

$$\begin{aligned} \min_{x} \quad & \frac{1}{2}\|\pi - x\|_2^2 \\ \text{s.t.} \quad & \|x\|_1 \leq c_1, \quad \|x\|_\infty \leq c_2. \end{aligned}$$

By introducing an auxiliary variable d, the problem above has the following equivalent form:

$$\begin{aligned} \min_{x,d} \quad & \frac{1}{2}\|\pi - x\|_2^2 & (16) \\ \text{s.t.} \quad & \|x\|_1 \leq c_1, \\ & x_i \leq d, \ i = 1, 2, ...n, \\ & d \leq c_2. \end{aligned}$$

The Lagrangian of problem (16) is

$$\mathcal{L}(x, d, \lambda_1, \lambda_2) = \frac{1}{2}\|\pi - x\|_2^2 + \sum_i \omega_i(x_i - d) + \lambda_1(\|x\|_1 - c_1) + \lambda_2(d - c_2) \quad (17)$$

34

```
input  : π, c₁, c₂ and precision parameter δ.
output: x, the projection of π onto {x : ‖x‖₁ ≤ c₁} ∩ {x : ‖x‖∞ ≤ c₂}.
Initialization: λ₁ = ‖π‖∞.
while |h(λ₁)| > δ do
    if h(λ₁) > δ then
    |   l = λ₁;
    else
    |   r = λ₁;
    end
    λ₁ = (l + r)/2;
    for each i do xᵢ = min(max(πᵢ − λ₁, 0), c₂);
end
```

Algorithm 5: Projecting π onto $\{x : \|x\|_1 \le c_1\} \cap \{x : \|x\|_\infty \le c_2\}$.

Let x^*, d^* and λ_1^*, λ_2^* respectively be the primal and dual solution of (16), then its KKT condition is:

$$0 \in \partial \mathcal{L}_x(x^*, d^*, \lambda_1^*, \lambda_2^*), \tag{18}$$

$$0 \in \partial \mathcal{L}_d(x^*, d^*, \lambda_1^*, \lambda_2^*), \tag{19}$$

$$w_i^*(x_i^* - d^*) = 0, \quad i = 1, 2, ..., n \tag{20}$$

$$\lambda_2^*(d^* - c_2) = 0, \tag{21}$$

$$\lambda_1^*(\|x^*\|_1 - c_1) = 0, \tag{22}$$

$$x_i^* \le d^*, \quad i = 1, 2, ..., n \tag{23}$$

$$d^* \le c_2, \tag{24}$$

$$\|x^*\|_1 \le c_1, \tag{25}$$

$$x_i^*, d^*, \lambda_1^*, \lambda_2^*, w_i^* \ge 0. \tag{26}$$

The following lemmas illustrate the relation between the primal and dual solution of problem (16):

Lemma 22 *Either one of the following holds: 1) $d^* > 0$ and $\sum_i(\max(\pi_i - \lambda_1^*, 0) - x_i^*) = \lambda_2^*$; 2) $d^* = 0$ and $\sum_i \max(\pi_i - \lambda_1^*, 0) \le \lambda_2^*$.*

Lemma 23 *For $i = 1, 2, ..., n$, the optimal x_i^* satisfies*

$$x_i^* = \min(\max(\pi_i - \lambda_1^*, 0), d^*). \tag{27}$$

Proof Lemma 22 and Lemma 23 can be both viewed as the special cases of Lemma 6 and 7 in Yu et al. (2012). ∎

Now the KKT conditions can be reduced to finding $\lambda_1^*, \lambda_2^*, d^*$ that satisfy the following equations:

$$\sum_i \min(\max(\pi_i - \lambda_1^*, 0), d^*) - c_1 \;=\; 0, \tag{28}$$

$$d^* - c_2 \;=\; 0, \tag{29}$$

$$\sum_i \max(\max(\pi_i - \lambda_1^*, 0) - d^*, 0) \;=\; \lambda_2^* \;\; \text{iff} \;\; d^* > 0, \tag{30}$$

$$\sum_i \max(\pi_i - \lambda_1^*, 0) \;\leq\; \lambda_2^* \;\; \text{iff} \;\; d^* = 0, \tag{31}$$

$$d^* \geq 0, \lambda_2^* \geq 0, \lambda_1^* \geq 0. \tag{32}$$

Further simplification induces the following result:

Lemma 24 *Suppose d^* and λ_1^* are the primal and dual solution respectively, then*

$$d^* \;=\; c_2, \tag{33}$$

$$x_i^* \;=\; \min(\max(\pi_i - \lambda_1^*, 0), c_2). \tag{34}$$

Proof Direct result of (27) and (29). ∎

(34) shows that the solution x^* is determined once the optimal λ_1^* is found. Given a $\lambda_1 \in [0, \max_i(\pi_i)]$, we only need to check whether (28) holds by looking into the function

$$h(\lambda_1) = \sum_i \min(\max(\pi_i - \lambda_1, 0), c_2) - c_1, \tag{35}$$

and λ_1^* is simply the zero point of $h(\cdot)$. The following theorem shows that $h(\cdot)$ is a strictly monotonically decreasing function, so a binary search is sufficient to find λ_1^*, and x_i^* can be determined accordingly.

Theorem 25 *1) $h(\lambda_1)$ is a continuous piecewise linear function in $[0, \max_i\{\pi_i\}]$; 2) $h(\lambda_1)$ is strictly monotonically decreasing and it has a unique root in $[0, \max_i\{\pi_i\}]$.*

Proof 1) is obviously true. It is easy to check that $h'(\lambda_1) < 0$ in each piece, $h(\max_i\{\pi_i\}) < 0$ and $h(0) > 0$, so 2) also holds. ∎

Complexity Analysis: Algorithm 5 is proposed based on the observation above to solve the projection problem, which is essentially a bisection to search λ_1. Given a precision δ, the iteration complexity of bisection is $O(\log(\max_i(\pi_i)/\delta))$. Besides, the time complexity of evaluating $h(\cdot)$ is $O(n)$, so the total complexity of Algorithm 5 is $O(n\log(\max_i(\pi_i)/\delta))$.

Appendix E. Convergence analysis of PGD

We provide a convergence analysis of the projected gradient descent algorithm used in optimizing Eq. (4). The analysis shows that the PGD algorithm approximately computes the global optimum of Eq. (4) in polynomial time. In simulation studies, much fewer iterations are required for convergence than predicted by the theoretical results.

Because we're using exact projected gradient descent algorithms, the objective function shall decay monotonically and hence convergence of such algorithms can be established by showing Lipschitz continuity of ∇f on a specific *level set*; that is, for some Lipschitz constant $L > 0$ the following holds for all π, π' such that $f(\pi), f(\pi') \leq f(\pi^{(0)})$:

$$\|\nabla f(\pi) - \nabla f(\pi')\|_2 \leq L\|\pi - \pi'\|_2, \tag{36}$$

where $\pi^{(0)} = (1/n, 1/n, \cdots, 1/n)$ is the initialization point. Once Eq. (36) holds, linear convergence (i.e., $f(\pi^{(t)}) - f(\pi^*) = O(1/t)$) can be established via standard projected gradient analysis.

The main idea of establishing Eq. (36) is to upper bound the spectral norm of the Hessian matrix $H = \nabla^2 f(\pi)$ *uniformly* over all points π that satisfies $f(\pi) \leq f(\pi^{(0)})$. As a first step, we derive analytic forms of H in the following proposition:

Proposition 26 *Let $\widetilde{\Sigma} = X^\top \mathrm{diag}(\pi) X$. We then have that*

$$H = 2(X^\top \widetilde{\Sigma}^{-2} X) \circ (X^\top \widetilde{\Sigma}^{-1} X),$$

where $\circ$ denotes the element-wise Hadamard product between two matrices of same dimensions.

Proof We first derive the gradient of f. Fix arbitrary $i \in [n]$. The partial derivative can be computed as

$$\frac{\partial f}{\partial \pi_i} = \frac{\partial \mathrm{tr}(\widetilde{\Sigma}^{-1})}{\partial \pi_i} = \left\langle \frac{\partial \mathrm{tr}(\widetilde{\Sigma}^{-1})}{\partial \widetilde{\Sigma}}, \frac{\partial \widetilde{\Sigma}}{\partial \pi_i} \right\rangle = \left\langle -\widetilde{\Sigma}^{-2}, x_i x_i^\top \right\rangle = -x_i^\top \widetilde{\Sigma}^{-2} x_i.$$

Here $\langle A, B \rangle = \mathrm{tr}(B^\top A)$ is the element-wise multiplication inner product between two matrices. The second-order partial derivatives can then be computed as

$$\frac{\partial^2 f}{\partial \pi_i \partial \pi_j} = -\frac{\partial (x_i^\top \widetilde{\Sigma}^{-2} x_i)}{\partial \pi_j} = -x_i^\top \frac{\partial \widetilde{\Sigma}^{-2}}{\partial \pi_j} x_i = x_i^\top \widetilde{\Sigma}^{-2} \frac{\partial \widetilde{\Sigma}^2}{\partial \pi_j} \widetilde{\Sigma}^{-2} x_i$$

$$= x_i^\top \widetilde{\Sigma}^{-2} (x_j x_j^\top \widetilde{\Sigma} + \widetilde{\Sigma} x_j x_j^\top) \widetilde{\Sigma}^{-2} x_i = 2(x_i^\top \widetilde{\Sigma}^{-2} x_j) \cdot (x_i^\top \widetilde{\Sigma}^{-1} x_j).$$

Subsequently,

$$H = \nabla^2 f = \left[\frac{\partial^2 f}{\partial \pi_i \partial \pi_j} \right]_{i,j=1}^n = 2(X^\top \widetilde{\Sigma}^{-2} X) \circ (X^\top \widetilde{\Sigma}^{-1} X).$$

∎

Corollary 27 *Suppose π satisfies $f(\pi) \leq f(\pi^{(0)}) = \mathrm{tr}(\Sigma_0^{-1})$, where $\Sigma_0 = \frac{1}{n} X^\top X$. We then have*

$$\|H\|_2 = \|\nabla^2 f(\pi)\|_2 \leq 2\|X\|_2^4 \cdot [\mathrm{tr}(\Sigma_0^{-1})]^3.$$

Proof Let $\rho(A) = \max_i |\sigma_i(A)|$ denote the *spectral range* of matrix A. Clearly, $\|A\|_2 \leq \rho(A)$ and $\|A\|_2 = \rho(A)$ for positive semi-definite matrices. In (Hom and Johnson, 1991) it is established that $\rho(A \circ B) \leq \rho(A)\rho(B)$. Subsequently,

$$\|H\|_2 \leq 2\|X\|_2^4 \|\widetilde{\Sigma}^{-1}\|_2^3 \leq 2\|X\|_2^4 \cdot [\mathrm{tr}(\widetilde{\Sigma}^{-1})]^3 \leq 2\|X\|_2^4 \cdot [\mathrm{tr}(\Sigma_0^{-1})]^3,$$

where the last inequality is due to the condition that $\mathrm{tr}(\widetilde{\Sigma}^{-1}) = f(\pi) \leq f(\pi^{(0)}) = \mathrm{tr}(\Sigma_0^{-1})$. ∎

From Corollary 27, we can prove the convergence of the optimization procedures outlined in Appendix. B following standard analysis of projected gradient descent on objective functions with Lipschitz continuous gradient:

Theorem 28 *Suppose $\pi^{(t)}$ is the solution at the tth iteration of the projected gradient descent algorithm and π^* is the optimal solution. We then have*

$$f(\pi^{(t)}) - f(\pi^*) \leq \frac{\|X\|_2^4 \cdot [\mathrm{tr}(\Sigma_0^{-1})]^3 \cdot \|\pi^{(0)} - \pi^*\|_2^2}{\beta t},$$

where $\beta \in (0, 1)$ is the backtracking parameter in backtracking line search.

Acknowledgments

References

Alexander A Ageev and Maxim I Sviridenko. Pipage rounding: A new method of constructing algorithms with proven performance guarantee. *Journal of Combinatorial Optimization*, 8(3):307–328, 2004.

David G Anderson, Ming Gu, and Christopher Melgaard. An efficient algorithm for unweighted spectral graph sparsification. *arXiv preprint arXiv:1410.4273*, 2014.

Haim Avron and Christos Boutsidis. Faster subset selection for matrices and applications. *SIAM Journal on Matrix Analysis and Applications*, 34(4):1464–1499, 2013.

Joshua Batson, Daniel A Spielman, and Nikhil Srivastava. Twice-ramanujan sparsifiers. *SIAM Journal on Computing*, 41(6):1704–1721, 2012.

Peter J Bickel, Ya'acov Ritov, and Alexandre B Tsybakov. Simultaneous analysis of lasso and dantzig selector. *The Annals of Statistics*, 37(4):1705–1732, 2009.

Kamalika Chaudhuri, Sham Kakade, Praneeth Netrapalli, and Sujay Sanghavi. Convergence rates of active learning for maximum likelihood estimation. In *Proceedings of Advances in Neural Information Processing Systems (NIPS)*, 2015.

Siheng Chen, Rohan Varma, Aarti Singh, and Jelena Kovačević. Signal representations on graphs: Tools and applications. *arXiv:1512.05406*, 2015.

Herman Chernoff. Locally optimal designs for estimating parameters. *The Annals of Statistics*, pages 586–602, 1953.

Mark A Davenport, Andrew K Massimino, Deanna Needell, and Tina Woolf. Constrained adaptive sensing. *IEEE Transactions on Signal Processing*, 64(20):5437–5449, 2015.

Paramveer Dhillon, Yichao Lu, Dean Foster, and Lyle Ungar. New sampling algorithms for fast least squares regression. In *Proceedings of Advances in Neural Information Processing Systems (NIPS)*, 2013.

Edgar Dobriban and Jianqing Fan. Regularity properties for sparse regression. *Communications in Mathematics and Statistics*, 4(1):1–19, 2016.

Petros Drineas, Michael W Mahoney, and S Muthukrishnan. Relative-error CUR matrix decompositions. *SIAM Journal on Matrix Analysis and Applications*, 30(2):844–881, 2008.

Petros Drineas, Michael W Mahoney, S Muthukrishnan, and Tamás Sarlós. Faster least squares approximation. *Numerische Mathematik*, 117(2):219–249, 2011.

Phillip Ein-Dor and Jacob Feldmesser. Attributes of the performance of central processing units: a relative performance prediction model. *Communications of the ACM*, 30(4): 308–317, 1987.

Elad Hazan and Zohar Karnin. Hard margin active linear regression. In *Proceedings of International Conference on Machine Learning (ICML)*, 2015.

RA Hom and CR Johnson. *Topics in Matrix Analysis*. Cambridge UP, New York, 1991.

Thibaut Horel, Stratis Ioannidis, and S Muthukrishnan. Budget feasible mechanisms for experimental design. In *Latin American Symposium on Theoretical Informatics (LATIN)*, pages 719–730. Springer, 2014.

Andre Khuri, Bhramar Mukherjee, Bikas Sinha, and Malay Ghosh. Design issues for generalized linear models: a review. *Statistical Science*, 21(3):376–399, 2006.

Ping Ma, Michael W Mahoney, and Bin Yu. A statistical perspective on algorithmic leveraging. *Journal of Machine Learning Research*, 16(1):861–911, 2015.

Lester Mackey, Michael I Jordan, Richard Y Chen, Brendan Farrell, and Joel A Tropp. Matrix concentration inequalities via the method of exchangeable pairs. *The Annals of Probability*, 42(3):906–945, 2014.

Adam Marcus, Daniel Spielman, and Nikhil Srivastava. Interlacing families I: bipartite ramanujan graphs of all degrees. *Annals of Mathematics*, 182:307–325, 2015a.

Adam Marcus, Daniel Spielman, and Nikhil Srivastava. Interlacing families II: Mixed characteristic polynomials and the kadison-singer problem. *Annals of Mathematics*, 182: 327–350, 2015b.

Alan Miller and Nam-Ky Nguyen. A fedorov exchange algorithm for d-optimal design. *Journal of the Royal Statistical Society, Series C (Applied Stiatistics)*, 43(4):669–677, 1994.

Nathan Nakamura, Jason Seepaul, Joseph B Kadane, and B Reeja-Jayan. Design for low-temperature microwave-assisted crystallization of ceramic thin films. *Applied Stochastic Models in Business and Industry*, 2017.

Sinno Jialin Pan and Qiang Yang. A survey on transfer learning. *IEEE Transactions on Knowledge and Data Engineering*, 22(10):1345–1359, 2010.

Gabor Pataki. On the rank of extreme matrices in semidefinite programs and the multiplicity of optimal eigenvalues. *Mathematics of Operations Research*, 23(2):339–358, 1998.

Friedrich Pukelsheim. *Optimal design of experiments*, volume 50. SIAM, 1993.

Garvesh Raskutti and Michael Mahoney. Statistical and algorithmic perspectives on randomized sketching for ordinary least-squares. In *Proceedings of International Conference on Machine Learning (ICML)*, 2015.

B Reeja-Jayan, Katharine L Harrison, K Yang, Chih-Liang Wang, AE Yilmaz, and Arumugam Manthiram. Microwave-assisted low-temperature growth of thin films in solution. *Scientific reports*, 2(1003):1–8, 2012.

Mark Rudelson and Roman Vershynin. Sampling from large matrices: An approach through geometric functional analysis. *Journal of the ACM*, 54(4):21, 2007.

Sivan Sabato and Remi Munos. Active regression by stratification. In *Proceedings of Advances in Neural Information Processing Systems (NIPS)*, 2014.

Daniel A Spielman and Nikhil Srivastava. Graph sparsification by effective resistances. *SIAM Journal on Computing*, 40(6):1913–1926, 2011.

Ryan Tibshirani. The lasso problem and uniqueness. *Electronic Journal of Statistics*, 7: 1456–1490, 2013.

Aad W Van der Vaart. *Asymptotic statistics*, volume 3. Cambridge university press, 2000.

Lieven Vandenberghe and Stephen Boyd. Semidefinite programming. *SIAM Review*, 38(1): 49–95, 1996.

David Woodruff. Sketching as a tool for numerical linear algebra. *Foundations and Trends in Theoretical Computer Science*, 10(1–2):1–157, 2014.

Adams Wei Yu, Hao Su, and Fei-Fei Li. Efficient euclidean projections onto the intersection of norm balls. In *Proceedings of International Conference on Machine Learning (ICML)*, 2012.

Rong Zhu, Ping Ma, Michael Mahoney, and Bin Yu. Optimal subsampling approaches for large sample linear regression. *arXiv:1509.05111*, 2015.