Computationally Feasible Near-Optimal Subset Selection for Linear Regression under Measurement Constraints

Yining Wang[1], Adams Wei Yu[1], and Aarti Singh[1]

[1]Machine Learning Department, Carnegie Mellon University

July 5, 2016

Abstract

Computationally feasible and statistically near-optimal subset selection strategies are derived to select a small portion of design (data) points in a linear regression model $\boldsymbol{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$ to reduce measurement cost and data efficiency. We consider two subset selection algorithms for estimating model coefficients $\boldsymbol{\beta}$: the first algorithm is a random subsampling based method that achieves optimal statistical performance with a small $(1+\epsilon)$ relative factor under the with replacement model, and an $O(\log k)$ multiplicative factor under the without replacement model, with k denoting the measurement budget. The second algorithm is fully deterministic and achieves $(1+\epsilon)$ relative approximation under the without replacement model, at the cost of slightly worse dependency of k on the number of variables (data dimension) in the linear regression model. Finally, we show how our method could be extended to the corresponding prediction problem and also remark on interpretable sampling (selection) of data points under random design frameworks.

Keywords: experimental design, A-optimality, subset selection, linear regression

1 Introduction

Linear regression is perhaps the simplest yet most widely used statistical model in scientific and engineering disciplines that involve processing of noisy data. A typical linear regression model is [1]

$$\boldsymbol{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \tag{1}$$

where $\mathbf{X} \in \mathbb{R}^{n\times p}$ is the *design matrix* or *data matrix*, $\boldsymbol{y} \in \mathbb{R}^n$ is the response, $\boldsymbol{\varepsilon} \sim \mathcal{N}_n(\mathbf{0}, \sigma^2\mathbf{I}_n)$ are i.i.d. white Gaussian noise with variance σ^2 and $\boldsymbol{\beta}$ is a fixed but unknown p-dimensional coefficient vector. Under the *fixed design* setting the design matrix $\mathbf{X}$ is fixed a priori. Given the design matrix $\mathbf{X}$ and response $\boldsymbol{y}$, one wishes to estimate the regression model $\boldsymbol{\beta}$ or the "true" responses $\mathbf{X}\boldsymbol{\beta}$. In this paper we focus on the so-called "low-dimensional" setting where the number of data points n is larger than the number of variables p and the design matrix $\mathbf{X}$ has full column rank. Under such settings, an effective estimator of $\boldsymbol{\beta}$ is the *ordinary least squares* (OLS) estimator:

[1] We consider only linear models in this paper. While non-linearity could be introduced via generalized linear model or kernel methods, such extension is left as future work.

$\hat{\boldsymbol{\beta}}^{\text{ols}} = \text{argmin}_{\boldsymbol{\beta}\in\mathbb{R}^p} \|\boldsymbol{y}-\mathbf{X}\boldsymbol{\beta}\|_2^2 = (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\boldsymbol{y}$. The OLS estimator has been extensively studied in prior literature. and is known to be optimal for estimation of either $\boldsymbol{\beta}$ or $\mathbf{X}\boldsymbol{\beta}$ under $n > p$ settings.

Despite the optimality of OLS, in practice it may not be desirable to obtain the full n-dimensional response vector $\boldsymbol{y}$ due to measurement constraints. Below are some example applications:

- **CPU benchmarking**: from known manufacturing parameters (clock period, cache size, etc.) of a CPU ($\mathbf{X}$), predict its execution time ($\boldsymbol{y}$) on benchmark tasks [14]. Limitations on access to CPU device as well as constraints on time needed for benchmarking imply that the number of measurements of execution time for different CPUs are limited.

- **Wind speed prediction**: from wind speeds at established measurement locations ($\mathbf{X}$), predict wind speeds at other locations ($\boldsymbol{y}$) based on road graph and/or distance [7]. In this application, measurements are limited due to instrumentation and data gathering costs.

- **Gene expression completion**: from full sequenced gene expressions of a few patients ($\mathbf{X}$), complete gene expression ($\boldsymbol{y}$) of a *new* patient by probing a small portion of the genes [25].

- **Material synthesis**: from parameters (temperature, pressure, etc.) of experimental settings ($\mathbf{X}$), predict the quality of synthesized materials ($\boldsymbol{y}$). Experimental costs for material synthesis limit the number of measurements in this application.

In all of the above examples, obtaining responses $\boldsymbol{y}$ for all data points in $\mathbf{X}$ is expensive, time-consuming or even infeasible in practice. It is thus an important question to find principled, computationally feasible ways to select a representative subset of rows in $\mathbf{X}$ so that the subsampled regression still accurately estimates $\boldsymbol{\beta}$.

In this paper, we present a systematic approach for data subsampling in linear regression models. The main idea is to consider a convex relaxation of an otherwise computationally intractable combinatorial optimization problem and perform sampling or greedy selections with respect to the optimal solution of the relaxed convex problem. Our main results are as follows:

1. Under the **with replacement** model, [2] a randomized **sampling** based method achieves $(1+\epsilon)$ **relative** approximation of the statistically optimal but computationally infeasible solution.

2. Under the **without replacement** model, the same **sampling** based method achieves an $O(\log k)$ **multiplicative** approximation of the optimal statistical performance, with k being the number of selected data points (measurement budget) allowed.

3. Under the **without replacement** model, a deterministic **greedy** algorithm achieves $(1+\epsilon)$ **relative** approximation of the optimal statistical performance, with slightly worse dependency of k on the number of variables (data dimension) p.

Table 1 summarizes our results and compares them to existing work. Apart from our main results, we also discuss how the algorithms could be extended to linear regression prediction on a test set and interpretable sampling distributions when design (data) points are drawn from some simple distributions.

[2] Detailed definitions of with and without replacement models are given in Sec. 2.

[3] Approximation factor depends on leverage scores of $\mathbf{X}$.

Table 1: Summary of results for subset selection in linear regression $\boldsymbol{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$ for estimation of model coefficients $\boldsymbol{\beta}$. n is the number of data points, p the number of variables (data dimension), and k the subset size. Approximation factor characterizes the statistical performance of the selected subset in linear regression and conditions on k states how large should the selected subset be to validate the approximation bounds. A bound is *additive* if the approximation error depends on $\mathrm{tr}[(\mathbf{X}^\top\mathbf{X})^{-1}]$ and *relative (or multiplicative)* if the error depends on $\min_{|S|\leq k}\mathrm{tr}[(\mathbf{X}_S^\top\mathbf{X}_S)^{-1}]$. $B = \max_{1\leq i\leq n}\|\boldsymbol{x}_i\|_2$ and $\boldsymbol{\Sigma}_* = \mathbf{X}^\top\mathbf{diag}(\boldsymbol{\pi}^*)\mathbf{X}$, where $\boldsymbol{\pi}^*$ is the optimal solution of the continuous problem in Eq. (4).

Algorithm	Model	Bound type	Approx. factor	Condition on k
Leverage score sampling [25]	with replacement	additive	-[3]	asymptotic
Greedy removal [2]	without replacement	additive	$O(n/k)$	$k = \Omega(p)$
Convex A-opt. + sampling	with replacement	relative	$1+\epsilon$	$k = \Omega(\epsilon^{-2}B^2\|\boldsymbol{\Sigma}_*^{-1}\|_2)$
Convex A-opt. + sampling	without replacement	multiplicative	$O(\log k)$	$k = \Omega(B^2\|\boldsymbol{\Sigma}_*^{-1}\|_2)$
Convex A-opt. + greedy	without replacement	relative	$1+\epsilon$	**Rigorous**: $k = \Omega(\epsilon^{-1}p^2)$ **Conjecture**: $k = \Omega(\epsilon^{-1}p)$

1.1 Related work

Subsampled linear regression Our work is mostly inspired by a trend of recent research [39, 7, 25], which studies the popular leverage score sampling (effective resistance sampling [32]) algorithm and its variants for subsampling in least squares problems. While the proposed methods are effective for solving the *deterministic problem* of minimizing least-square error $\|\boldsymbol{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2$, their statistical properties for estimating "underlying" regression coefficients $\boldsymbol{\beta}$ in the *stochastic* regression model of Eq. (1) are less investigated. One exception is [39], which quite surprisingly showed that leverage score sampling equipped with weighted OLS does *not* achieve optimal statistical efficiency in terms of estimating $\boldsymbol{\beta}$.

There is also a rich literature on using sketching or subsampling based methods to accelerate the computation of linear regression models [37, 12, 6, 13, 22, 1, 30]. Most of the work are computationally oriented and typically access *all* response variables $\boldsymbol{y}$. In [12] a uniform subsampling method that only observes a small portion of $\boldsymbol{y}$ was proposed for OLS regression. Though the authors justify their uniform subsampling strategy under an isotropic design setting, such method is far from optimal for general design matrices, as we discuss in Sec. 6.3.

Experimental design and active learning Our work falls in the areas of *statistical experimental design* [29, 2, 24, 23, 9]. However, unlike prior work, our methods are both computationally feasible, and enjoy finite sample guarantees with near-optimal statistical performance under the linear regression model. Also, related is the area of *active learning* [3, 5, 18, 31] which is a stronger setting where feedback from prior measurements can be used to guide subsequent measurements. For active learning the most relevant work is [5], which analyzes an SDP relaxation in the context of active maximum likelihood estimation, very similar to ours. However, there are several differences: [5] focuses on the prediction error of the subsampled estimator while we primarily consider estimation of the underlying regression coefficients; [5] works under the with replacement setting where fresh noise is available for every sample while our analysis is valid for both sampling with and without replacement settings; we also present more practical computationally efficient implementations using projective gradient descent that are faster than SDP; [5] analyzes a two-stage feedback-driven active strategy whereas we consider the one-stage experimental design setting where the measurement

choices are determined prior to obtaining any measurements.

Matrix subset selection One direction of the numerical linear algebra literature is the so-called *matrix subset selection* problems, which aim to select a subset of rows (columns) of a given matrix so as to maximize or minimize certain combinatorial criteria. Common matrix subset selection problems include low-rank matrix reconstruction [15, 11, 17], max volume subset selection [8, 23], A-optimality [2, 10], and many more. Among the existing results for matrix subset selection, the approximate A-optimality results [2, 10] are particularly relevant as optimal subset selection for estimating (linear regression) coefficients can be cast as a combinatorial A-optimality problem (details shown in Sec. 2.2). Nevertheless, existing analysis of computationally efficient approximate A-optimality algorithms shows an unsatisfactory *additive* error dependency over the approximation quality (see Table 1), which could be detrimental for real-world problems. In contrast, our results show *relative* or *multiplicative* approximations of the A-optimality criterion with computationally tractable algorithms.

2 Problem setup and preliminaries

2.1 Subset selection models and minimax formulation

We first present mathematical definitions of the subset selection models considered in this paper:

Definition 2.1 (Linear regression subset selection models)**.** *Let $\mathbf{X}$ be a fixed $n \times p$ design matrix with full column rank and k be the subset budget, with $p \le k \le n$. An algorithm A first observes $\mathbf{X}$ in full and produces, either deterministically or randomly, a matrix $\widetilde{\mathbf{X}} \in \mathbb{R}^{k \times p}$ such that each row of $\widetilde{\mathbf{X}}$ is equal to a particular row in $\mathbf{X}$. Two scenarios are considered:*

1. With replacement*: $\widetilde{\mathbf{X}}$ may have duplicate rows and the algorithm observes $\tilde{\boldsymbol{y}} = \widetilde{\mathbf{X}}\boldsymbol{\beta} + \tilde{\boldsymbol{\varepsilon}}$, $\tilde{\boldsymbol{\varepsilon}} \sim \mathcal{N}_k(\mathbf{0}, \sigma^2\mathbf{I})$. That is, fresh (independent) noise is imposed after subset selection, so that duplicate data points have independent noise.*
2. Without replacement*: $\widetilde{\mathbf{X}}$ may* not *have duplicate rows; $\tilde{\boldsymbol{y}} = \widetilde{\mathbf{X}}\boldsymbol{\beta} + \tilde{\boldsymbol{\varepsilon}}$ is observed.*

We use $\mathcal{A}_1(k)$ and $\mathcal{A}_2(k)$ to denote classes of all algorithms operating under the two subset selection models, respectively.

Our evaluation criteria are the following minimax performance under both subset selection models:

$$\inf_{A \in \mathcal{A}_1(k)} \sup_{\boldsymbol{\beta} \in \mathbb{R}^p} \mathbb{E}\left[R(\hat{\boldsymbol{\beta}}, \boldsymbol{\beta})\right], \qquad \inf_{A \in \mathcal{A}_2(k)} \sup_{\boldsymbol{\beta} \in \mathbb{R}^p} \mathbb{E}\left[R(\hat{\boldsymbol{\beta}}, \boldsymbol{\beta})\right] \tag{2}$$

where the expectation is taken over the noise variables $\boldsymbol{\varepsilon}$ and the possible inherent randomness in A. The loss function $R(\hat{\boldsymbol{\beta}}, \boldsymbol{\beta})$ is the mean-square estimation error $\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\|_2^2$ in most of our algorithms and analysis, except in Sec. 6.2 where we consider the prediction error $\|\mathbf{Z}\hat{\boldsymbol{\beta}} - \mathbf{Z}\boldsymbol{\beta}\|_2^2$ on test matrix $\mathbf{Z}$. Also note that the with replacement model risk $\inf_{A \in \mathcal{A}_1(k)} \sup_{\boldsymbol{\beta} \in \mathbb{R}^p} \mathbb{E}\left[R(\hat{\boldsymbol{\beta}}, \boldsymbol{\beta})\right]$ lower bounds the without replacement model risk $\inf_{A \in \mathcal{A}_2(k)} \sup_{\boldsymbol{\beta} \in \mathbb{R}^p} \mathbb{E}\left[R(\hat{\boldsymbol{\beta}}, \boldsymbol{\beta})\right]$ because $\mathcal{A}_2(k) \subseteq \mathcal{A}_1(k)$ by definition.

2.2 Combinatorial optimality criteria

The conventional approach for subset (data) selection in linear regression is via *combinatorial optimality criteria* [29], where a k-size subset $S \subseteq [n]$ is selected to minimize (or maximize) certain combinatorial functions. More specifically, consider the following combinatorial optimization problem:

$$S^* = \operatorname{argmin}_{S\subseteq[n],|S|\leq k} F(S; \mathbf{X}), \tag{3}$$

where the subset S is allowed to have duplicates (i.e., S being a multi-set) under the with replacement setting and vice versa for the without replacement setting.

There are a number of choices for the combinatorial function $F(\cdot; \mathbf{X})$ in Eq. (3), each exhibiting different objectives and properties. We refer the readers to the seminal book of [29] for a comprehensive discussion of the various optimality criteria analyzed in prior literature. For the purpose of this paper, we focus primarily on the following criteria:

- **A-optimality**: $F(S; \mathbf{X}) = \operatorname{tr}[(\mathbf{X}_S^\top \mathbf{X}_S)^{-1}]$. This criterion is related to the mean-square estimation error $R(\hat{\boldsymbol{\beta}}, \boldsymbol{\beta}) = \|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\|_2^2$, as the OLS estimator on the selected data $\mathbf{X}_S$ and $\mathbf{y}_S$ has mean-square estimation error $\sigma^2 \operatorname{tr}[(\mathbf{X}_S^\top \mathbf{X}_S)^{-1}]$.
- **G-optimality**: $F(S; \mathbf{X}) = \operatorname{tr}[(\mathbf{X}^\top \mathbf{X})(\mathbf{X}_S^\top \mathbf{X}_S)^{-1}]$. This criterion is related to the mean-square prediction error $R(\hat{\boldsymbol{\beta}}, \boldsymbol{\beta}) = \|\mathbf{X}\hat{\boldsymbol{\beta}} - \mathbf{X}\boldsymbol{\beta}\|_2^2$. We also consider a generalized version $F(S; \mathbf{X}, \mathbf{Z}) = \operatorname{tr}[(\mathbf{Z}^\top \mathbf{Z})(\mathbf{X}_S^\top \mathbf{X}_S)^{-1}]$ that corresponds to the prediction error on a test set $\mathbf{Z}$ that might differ from the training set $\mathbf{X}$.

Perhaps not surprisingly, combinatorial optimality is difficult to attain in a computationally efficient way. An exhaustive search over all possible size-k subsets takes exponential computational time and is infeasible even for moderately sized data set. To overcome the computational hurdle, in Sec. 3 we introduce efficient continuous optimization based subset selection algorithms and show that they achieve near-optimal statistical performance under mild conditions.

3 Main results

This section presents our main results, which are two computationally efficient, statistically near-optimal subset selection algorithms for linear regression. We consider solely the parameter estimation problem $R(\hat{\boldsymbol{\beta}}, \boldsymbol{\beta}) = \|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\|_2^2$ in this section, while deferring analysis of prediction and other related problems to Sec. 6. The main idea behind our methods is a continuous, convex optimization problem that can be solved efficiently, followed by post-processing sampling or greedy procedures that select exact k data points in $\mathbf{X}$ to form an effective subset. Complete proofs of all results are avaialble in Appendix C.

3.1 Continuous relaxations of combinatorial optimality criteria

We start by considering a continuous relaxation of the subset selection problem in Eq. (3):

$$f_{\text{opt}} := \min_{\boldsymbol{\pi}=(\pi_1,\cdots,\pi_n)} \operatorname{tr}\left[(\mathbf{X}^\top \mathbf{diag}(\boldsymbol{\pi})\mathbf{X})^{-1})\right], \qquad s.t. \quad \|\boldsymbol{\pi}\|_1 \leq 1, \quad \pi_i \geq 0. \tag{4}$$

Note that Eq. (4) is a convex optimization problem and hence can be solved effectively in polynomial time. We present SDP re-formulations and also practical optimization algorithms for Eq. (4) in Appendix A.1 and Sec. 4. Eq. (4) corresponds to the "multi-set" version of the combinatorial A-optimality problem in Eq. (3). In addition, because from the optimal multiset S^* in Eq. (3) we can construct feasible solution $\boldsymbol{\pi}^*$ in Eq. (4) as $\boldsymbol{\pi}_i^* = |i \in S^*|/k$, the optimal value of Eq. (4), f_{opt} lower bounds its corresponding combinatorial optimization problem as [4] $f_{\text{opt}} \leq k \cdot \min_{|S|\leq k, S \text{ multi-set}} \text{tr}\left[\left(\mathbf{X}_S^\top \mathbf{X}_S\right)^{-1}\right]$.

For the without replacement setting, we consider the same problem with slightly varied constraint set:

$$\tilde{f}_{\text{opt}}(k) := \min_{\boldsymbol{\pi}=(\pi_1,\cdots,\pi_n)} \text{tr}\left[(\mathbf{X}^\top \mathbf{diag}(\boldsymbol{\pi})\mathbf{X})^{-1})\right], \qquad s.t. \quad \|\boldsymbol{\pi}\|_1 \leq 1, \quad 0 \leq \pi_i \leq 1/k. \tag{5}$$

The only difference between Eq. (5) and (4) is the additional constraint $\|\boldsymbol{\pi}\|_\infty \leq 1/k$, which reflects the without replacement requirement. It is then trivially true that $f_{\text{opt}} \leq \tilde{f}_{\text{opt}}(k)$ by definition. Note that the problem remains convex and tractable with the additional infinity norm constraint, as we discuss in Appendix A.1 and Sec. 4. In addition, because one can construct feasible solution $\boldsymbol{\pi}^*$ to Eq. (5) from the optimal standard set solution S^* of Eq. (3) by $\boldsymbol{\pi}_i^* = 1/k$ if $i \in S^*$ and $\boldsymbol{\pi}_i^* = 0$ otherwise, $\tilde{f}_{\text{opt}}(k)$ lower bounds the combinatorial optimality solution under the with replacement model as $\tilde{f}_{\text{opt}}(k) \leq k \cdot \min_{|S|\leq k, S \text{ standard set}} \text{tr}\left[\left(\mathbf{X}_S^\top \mathbf{X}_S\right)^{-1}\right]$.

Theorem 3.1 shows that $f_{\text{opt}}, \tilde{f}_{\text{opt}}(k)$ lower bounds the statistical performance of any deterministic or random subset selection algorithm under corresponding replacement models. Its proof is rather technical and is deferred to Appendix C.1.

Theorem 3.1. *Fix $p \leq k \leq n$ and $\mathbf{X} \in \mathbb{R}^{n\times p}$ with full column rank. We then have*

$$\frac{\sigma^2}{k} f_{\text{opt}} \leq \inf_{A\in\mathcal{A}_1(k)} \sup_{\boldsymbol{\beta}} \mathbb{E}\left[\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\|_2^2\right], \qquad \frac{\sigma^2}{k} \tilde{f}_{\text{opt}}(k) \leq \inf_{A\in\mathcal{A}_2(k)} \sup_{\boldsymbol{\beta}} \mathbb{E}\left[\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\|_2^2\right]. \tag{6}$$

Finally, we emphasize that solutions of the continuous problems in Eq. (4,5) are in general *not* valid subset selection solutions as the optimal solution $\boldsymbol{\pi}^*$ might contain far more than k non-zero coordinates. We show in follow-up sections two effective post-processing procedures that turn $\boldsymbol{\pi}^*$ into a valid size-k subset S while still preserving the optimal statistical performance of $\boldsymbol{\pi}^*$ (implied by Theorem 3.1) under mild conditions.

3.2 The subsampling post-processing procedure

We introduce our first post-processing procedure as a subsampling algorithm, with pseudocode displayed in Algorithm 1. The algorithm is as simple as one can hope for: after obtaining the optimal solution $\boldsymbol{\pi}^*$ of the continuous optimization problems, which by formulation represents a probabilistic distribution over the n design points (rows in $\mathbf{X}$), the algorithm samples with or without replacement (depending on the model settings) according to $\boldsymbol{\pi}^*$ for k times to obtain a size-k subset of $\mathbf{X}$. Finally, the entire post-processing procedure is repeated for T times in order to boost the overall probability of generating a subset of $\mathbf{X}$ with satisfactory statistical performance.

The following theorem provides statistical performance guarantee for Algorithm 1.

[4] Note that we substitute 1 for k in the "signal level" $\sum_{i=1}^n \pi_i$ in Eq. (4) for normalization purposes.

Algorithm 1 The subsampling post-processing procedure

1: **Input**: full design $\mathbf{X} \in \mathbb{R}^{n\times p}$, sensing budget k, number of repetitions T, convex optimization solution $\boldsymbol{\pi}^*$ of either Eq. (4) or (5).
2: **Subsampling**: sample k rows of $\mathbf{X}$ with or without replacement according to the distribution specified by $\boldsymbol{\pi}^*$ to form $\widetilde{\mathbf{X}} \in \mathbb{R}^{k\times p}$.
3: **Boosting**: Repeat Step 3 for T times to obtain T subsets of size k: $\{\widetilde{\mathbf{X}}_{(t)}\}_{t=1}^T$. Find $t^* = \mathrm{argmin}_{1\le t\le T}\mathrm{tr}[(\widetilde{\mathbf{X}}_{(t)}^\top \widetilde{\mathbf{X}}_{(t)})^{-1}]$.
4: **Output**: selected subset $\widetilde{\mathbf{X}}_{(t^*)}$ and also estimated coefficients $\hat{\boldsymbol{\beta}}$ via OLS on $(\widetilde{\mathbf{X}}_{(t^*)}, \tilde{\boldsymbol{y}})$.

Theorem 3.2. *Fix $\mathbf{X} \in \mathbb{R}^{n\times p}$ with full column rank and error tolerance parameter $\epsilon \in (0, 1/2)$. Suppose $\sup_{1\le i\le n}\|\boldsymbol{x}_i\|_2 \le B < \infty$. Let $\boldsymbol{\Sigma}_* = \sum_{i=1}^n \pi_i^* \boldsymbol{x}_i\boldsymbol{x}_i^\top$, where $\boldsymbol{\pi}^*$ is the optimal solution to Eq. (4). If $k = \Omega(\epsilon^{-2}B^2\|\boldsymbol{\Sigma}_*^{-1}\|_2 \log(n/\epsilon))$ and $T = \Omega(\epsilon^{-1}\log n)$ then with probability at least $1 - O(n^{-1})$ over the randomly produced subset $\widetilde{\mathbf{X}}_{(t^*)}$, we have that*

$$\forall \boldsymbol{\beta} \in \mathbb{R}^p, \qquad \mathbb{E}\left[\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\|_2^2\right] \le \begin{cases} (1+\epsilon)\cdot \sigma^2 f_{\mathrm{opt}}/k, & \textit{with replacement}; \\ O(\log k)\cdot \sigma^2 \tilde{f}_{\mathrm{opt}}(k)/k, & \textit{without replacement}. \end{cases} \tag{7}$$

Compared to Theorem 3.1, it can be seen that Algorithm 1 achieves near-optimal statistical performance in terms of estimation of the underlying regression coefficients $\boldsymbol{\beta}$. In particular, under the with replacement model the MSE error of the estimated coefficients $\hat{\boldsymbol{\beta}}$ is within a $(1+\epsilon)$ relative factor of the optimal performance, and under the without replacement model the MSE is upper bounded by an $O(\log k)$ multiplicative factor of the corresponding optimal error bound.

The near-optimal bounds given in Eq. (7) holds when the subset size k is not too small. Such conditions are necessary as no method could estimate the underlying regression model $\boldsymbol{\beta}$ in an assumption-free manner unless $k = \Omega(p)$ data points (and their responses) are observed. In Theorem 3.2 stronger conditions on k are required: the lower bound for k is always larger than $\Omega(\epsilon^{-1}p\log(n/\epsilon))$ and degrades as $\boldsymbol{\Sigma}_*$ becomes more ill-conditioned. Though $\boldsymbol{\Sigma}_*$ is typically much more well-conditioned than the sample covariance of the full design $\boldsymbol{\Sigma} = \frac{1}{n}\mathbf{X}^\top\mathbf{X}$, the lower bound for k could still be large in extreme cases. We show one potential way to remove the dependency over conditioning in Sec. 3.3 by applying a greedy post-processing procedure.

The proof of Theorem 3.2 is inspired by recent analysis of graph sparsification algorithms [32], which shows that the covariance $\widetilde{\boldsymbol{\Sigma}}^* = \widetilde{\mathbf{X}}^\top\widetilde{\mathbf{X}}$ of the subsampled points is a *spectral approximation* [5] of the optimal covariance $\boldsymbol{\Sigma}^* = \mathbf{X}^\top\mathbf{diag}(\boldsymbol{\pi}^*)\mathbf{X}$ with high probability when k is not too small. The complete proof is deferred to Appendix C.2.

3.3 The greedy removal post-processing procedure

In this section we consider an alternative post-processing procedure to transform the optimal continuous solution $\boldsymbol{\pi}^*$ to a size-k discrete subset while preserving statistical validity. Pseudocodes are displayed in Algorithm 2. The procedure is very simple: we start with a "large" subset $S \subseteq [n]$ that has more than k elements and greedily remove one element at a time from S until $|S| = k$. This

[5] $\mathbf{A}$ is an ϵ-spectral approximation of $\mathbf{B}$ if $(1-\epsilon)\boldsymbol{x}^\top\mathbf{A}\boldsymbol{x} \le \boldsymbol{x}^\top\mathbf{B}\boldsymbol{x} \le (1+\epsilon)\boldsymbol{x}^\top\mathbf{A}\boldsymbol{x}$ for all $\boldsymbol{x}$.

Algorithm 2 The greedy removal post-processing procedure

1: **Input**: full design $\mathbf{X} \in \mathbb{R}^{n\times p}$, sensing budget k, convex optimization solution $\boldsymbol{\pi}^*$ of either Eq. (4) or (5).
2: **Support extraction**: let $S = \{i \in [n] : \boldsymbol{\pi}_i^* > 0\}$ be the support set of $\boldsymbol{\pi}^*$.
3: **Greedy removal**: repeat the following greedy removal algorithm in [2] until $|S| = k$:

- Find $i^* \in S$ with $i^* = \mathrm{argmin}_{i\in S}\mathrm{tr}[(\mathbf{X}_{S\backslash\{i\}}^\top \mathbf{X}_{S\backslash\{i\}})^{-1}]$ and update $S \leftarrow S\backslash\{i^*\}$.

4: **Output**: selected subset $\mathbf{X}_S$ and also estimated coefficients $\hat{\boldsymbol{\beta}}$ via OLS on $(\mathbf{X}_S, \boldsymbol{y}_S)$.

procedure was first proposed and analyzed in [2]. However, in [2] the "initial" subset S is taken to be the entire index set $[n]$, which results in weak additive-error bounds (cf. Table 1). In contrast, we apply the greedy removal algorithm on the support set of $\boldsymbol{\pi}^*$ and hence achieve much favorable theoretical guarantees.

Unlike the subsampling post-processing method that requires essentially no conditions on the design matrix $\mathbf{X}$ (except the necessary condition that $\mathbf{X}$ has full column rank), we need some additional conditions on $\mathbf{X}$ in order to establish theoretical guarantee for Algorithm 2. At a higher level, we assume "second-order expansions" of each row of $\mathbf{X}$ are in "general condition", meaning that no unnecessary linear relationship in the expanded domain is satisfied. Formal definition of the assumption is articulated as follows:

Assumption 3.1. *For any $\boldsymbol{x} \in \mathbb{R}^p$, define mapping $\phi : \mathbb{R}^p \to \mathbb{R}^{\frac{p(p+1)}{2}}$ as $\phi(\boldsymbol{x}) = (\xi_{ij}\boldsymbol{x}(i)\boldsymbol{x}(j))_{1\le i\le j\le p}$, where $\boldsymbol{x}(i)$ represents the ith coordinate of $\boldsymbol{x}$, $\xi_{ij} = 1$ if $i = j$ and $\xi_{ij} = 2$ otherwise. Denote $\tilde{\phi}(\boldsymbol{x}) = (\phi(\boldsymbol{x}), 1) \in \mathbb{R}^{\frac{p(p+1)}{2}+1}$ as the affine version of $\phi(\boldsymbol{x})$. Assume that for any $p(p+1)/2+1$ distinct rows of $\mathbf{X}$ (denoted as $\boldsymbol{x}_1, \cdots, \boldsymbol{x}_{p(p+1)/2+1}$), the mapped vectors $\tilde{\phi}(\boldsymbol{x}_1), \cdots, \tilde{\phi}(\boldsymbol{x}_{p(p+1)/2+1})$ are linear independent.*

Like other general condition assumptions in the literature [35], Assumption 3.1 is very mild and is almost always satisfied in practice. As a simple example, Proposition 3.1 shows that Assumption 3.1 is satisfied with probability 1 if each row of $\mathbf{X}$ is sampled i.i.d. from any reasonable distribution. Its proof is placed in Appendix C.3.

Proposition 3.1. *Suppose each row of $\mathbf{X}$ is sampled i.i.d. from any distribution with bounded pdf. Then Assumption 3.1 holds almost surely.*

We now state the main result that bounds the statistical estimation error of $\hat{\boldsymbol{\beta}}$ produced by Algorithm 2.

Theorem 3.3. *Let $\hat{S} \subseteq [n]$, $|\hat{S}| = k$ be the subset selected by Algorithm 2 and suppose $k > p$. Under Assumption 3.1, we have that*

$$\mathrm{tr}\left[\left(\mathbf{X}_{\hat{S}}^\top \mathbf{X}_{\hat{S}}\right)^{-1}\right] \;\le\; \left(1 + \frac{p(p+1)}{2(k-p+1)}\right)\cdot \frac{\tilde{f}_{\mathrm{opt}}(k)}{k}.$$

As a simple remark, $k = \Omega(p^2/\epsilon)$ is sufficient to guarantee that

$$\forall \boldsymbol{\beta} \in \mathbb{R}^p, \qquad \mathbb{E}\left[\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\|_2^2\right] \le (1+\epsilon)\cdot \frac{\tilde{f}_{\mathrm{opt}}(k)}{k},$$

Algorithm 3 Optimization procedure of Eq. (4) and (8) via projective gradient descent

Input: $\mathbf{X}$, $\alpha \in (0, 1/2]$, $\beta, \delta \in (0, 1)$, objective function $\mathcal{L}$.
Initialization: $\boldsymbol{\pi}_i^{(0)} = 1/n$ for all data point $\boldsymbol{x}_i$; $t = 0$. Let $\mathcal{S}$ be the feasible solution set.
while $|\mathcal{L}(\boldsymbol{\pi}^{(t)}) - \mathcal{L}(\boldsymbol{\pi}^{(t-1)})|/\mathcal{L}(\boldsymbol{\pi}^{(t-1)}) > \delta$ **do**
  Gradient computation: $\frac{\partial \mathcal{L}}{\partial \pi_i} = \|\widetilde{\boldsymbol{\Sigma}}_X^{-1}\boldsymbol{x}_i\|_2^2$, where $\widetilde{\boldsymbol{\Sigma}}_X = \mathbf{X}^\top \mathbf{diag}(\boldsymbol{\pi}^{(t)})\mathbf{X}$.
  Backtracking line search: find the smallest integer $s \geq 0$ such that $\mathcal{L}(\tilde{\boldsymbol{\pi}}^{(t)}) - \mathcal{L}(\boldsymbol{\pi}^{(t)}) \leq \alpha \nabla \mathcal{L}^\top (\tilde{\boldsymbol{\pi}}^{(t)} - \boldsymbol{\pi}^{(t)})$, where $\tilde{\boldsymbol{\pi}}^{(t)} = \mathcal{P}_{\mathcal{S}}(\boldsymbol{\pi}^{(t)} - \beta^s \nabla \mathcal{L})$.
  Projective gradient descent: $\boldsymbol{\pi}^{(t+1)} = \mathcal{P}_{\mathcal{S}}(\boldsymbol{\pi}^{(t)} - \beta^s \nabla \mathcal{L})$; $t \leftarrow t + 1$.
end while
Output: optimal sampling probabilities $\boldsymbol{\pi}^{(t)}$.

where the expectation is taken conditioned on $\hat{S}$ under the without replacement model.

Compared to subsampling based post-processing method, the greedy removal processing achieves relative $(1 + \epsilon)$ approximation error bound under the *without* replacement setting. Furthermore, it is a fully deterministic algorithm and the condition (lower bound) on subset size k does *not* depend on conditioning of any covariance matrices, which makes the algorithm applicable to a wider range of applications where design matrices are ill-conditioned. On the other hand, the lower bound of k depends *quadratically* on data dimension p, which is an $O(p)$ factor worse than the ideal bound (i.e., k scales linearly in p). We give a more in-depth discussion on this issue in Sec. 6.4.

The key idea behind the proof of Theorem 3.3 is to bound the support size of $\boldsymbol{\pi}^*$. Afterwards, existing analysis of the greedy removal procedure in [2] could be directly applied to achieve the desired error bounds. To upper bound $\|\boldsymbol{\pi}^*\|_0$, we analyze the KKT condition of the continuous optimization problem Eq. (5). Compete proofs of Theorem 3.3 as well as Proposition 3.1 are deferred to Appendix C.3.

4 Practical optimization methods

Although Eq. (4,5) can be solved in polynomial time via SDP (details in Appendix A.1), such algorithms are not practical because SDP does not scale well. To overcome such difficulties, we implement first-order methods for optimizing these problems. As the objective functions are complicated and there are many constraints in our problems, standard approaches like ADMM or Frank-Wolfe might be difficult to implement and also slow to converge. Therefore, we adopt the *projective gradient descent* method with step sizes chosen by backtracking line search. Pseudocode of the optimization procedure is given in Algorithm 3 and parameters are set as $\alpha = \beta = 0.5$ and $\delta = 10^{-5}$.

One key step in Algorithm 3 it to compute the projection of an n-dimensional vector $\boldsymbol{\pi}$ onto the intersection of ℓ_1 and ℓ_∞ balls, which corresponds to the following quadratic programming problem:

$$\boldsymbol{x}^* = \mathrm{argmin}_{\boldsymbol{x}} \|\boldsymbol{x} - \boldsymbol{\pi}\|_2^2 \quad \text{s.t. } \|\boldsymbol{x}\|_1 \leq c_1,\ \|\boldsymbol{x}\|_\infty \leq c_2.$$

Here, we solve its associated KKT system by adopting a similar strategy as [38]. In a nutshell, we want to find a intermediate point $\boldsymbol{z}$ such that $\boldsymbol{z} = \max(\boldsymbol{\pi} - \boldsymbol{\lambda}_1, \mathbf{0})$ and $\boldsymbol{x}^* = \mathcal{P}_\infty^{c_2}(\boldsymbol{z})$, where $\mathcal{P}_\infty^{c_2}(\cdot)$ is the projection onto the ℓ_∞ ball with radius c_2. Such a procedure (determining $\boldsymbol{\lambda}_1$) could be reduced to finding a zero point of a decreasing function of $\boldsymbol{\lambda}_1$, which can then be solved by binary search.

The time complexity of the projection step is $O(n \log n \log(\|\boldsymbol{x}\|_\infty/\delta'))$, where n is the dimension of $\boldsymbol{\pi}$ and δ' is the projection precision. Due to space limits, details of the projection step and its associated analysis are deferred to Appendix A.2.

Assuming the projection steps $\mathcal{P}_\mathcal{S}$ in Algorithm 3 can be evaluated exactly, [6] the convergence of Algorithm 3 can be established by the following proposition (complete proof in Appendix A.2):

Proposition 4.1. *Suppose* $\mathrm{rank}(\mathbf{X}) = p$. *Algorithm 3 converges in that* $\lim_{t\to\infty} \mathcal{L}(\boldsymbol{\pi}^{(t)}) = \mathcal{L}(\boldsymbol{\pi}^*)$.

5 Simulation results

We validate our theoretical results on synthetic data sets and compare our proposed methods (continuous A-optimality plus subsampling/greedy-removal post-processing) with algorithms considered in prior literature. In particular, we compare to the following methods that are commonly used for subset selection in linear regression problems:

- **Leverage score sampling**: $\boldsymbol{\pi}_{\mathrm{wlev}}(i) \propto \boldsymbol{x}_i^\top \boldsymbol{\Sigma}^{-1} \boldsymbol{x}_i$, where $\boldsymbol{\Sigma} = \frac{1}{n}\mathbf{X}^\top\mathbf{X}$ is the sample covariance matrix. As suggested by [25], the leverage score sampling distribution $\boldsymbol{\pi}_{\mathrm{wlev}}$ is combined with uniform sampling $\boldsymbol{\pi}_{\mathrm{unif}}(i) = 1/n$ to achieve improved sample complexity. More specifically, the hybrid sampling distribution $\boldsymbol{\pi}_{\mathrm{hybrid}}$ is taken as a linear combination $\boldsymbol{\pi}_{\mathrm{hybrid}} = \alpha\boldsymbol{\pi}_{\mathrm{wlev}} + (1-\alpha)\boldsymbol{\pi}_{\mathrm{unif}}$, with α taking values in $\{0.1, 0.5, 0.9\}$.

- **Predictive-Length (PL) sampling**: the PL sampling is derived in [39] as a computationally-efficient alternative to leverage score sampling. It samples each row of $\mathbf{X}$ with probability proportional to its ℓ_2 norm: $\boldsymbol{\pi}_{\mathrm{pl}}(i) \propto \|\boldsymbol{x}_i\|_2$.

- **D-optimality**: a commonly-used combinatorial criterion for subset selection problems in the statistical experimental design literature [29]. In general, D-optimality seeks a subset S that maximizes $\det(\mathbf{X}_S^\top \mathbf{X}_S)$. We invoke the built-in `candexch` routine in Matlab to approximately compute D-optimal subsets. The `candexch` routine implements the Fedorov exchange algorithm [28].

The regression model $\boldsymbol{\beta} \in \mathbb{R}^p$ is generated from the standard normal distribution $\mathcal{N}_p(\mathbf{0}, \mathbf{I})$. Following [25], the full design matrix $\mathbf{X} \in \mathbb{R}^{n\times p}$ is sampled from an element-wise t distribution, with degrees of freedom varying in $\{2, 3, 10\}$. When the t distribution has a large degrees of freedom (e.g., $\mathrm{df} = 10$), the resulting design has a near-uniform distribution of leverage scores and each row of $\mathbf{X}$ is of similar importance; in constrast, when the t distribution has a small degree of freedom the design would be unbalanced and there are data points that are of significant importance compared to the other rows of $\mathbf{X}$. Finally, the noise variables $\boldsymbol{\varepsilon}$ are sampled from i.i.d. normal distribution with zero mean and $\sigma^2 = 0.01$ variance.

Figures 1, 2, 3 plots the mean-square error $\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\|_2^2$ against the number of subsamples k under various experimental settings (data dimension p, degrees of freedom df and with/without replacement models). We have the following observations:

1. Previous sampling methods (leverage score sampling, hybrid of leverage score and uniform sampling, PL sampling) performs much worse than optimization based subset selection al-

[6] In our implementation we set projection precision at $\delta' = 10^{-9}$, which is sufficiently negligible.

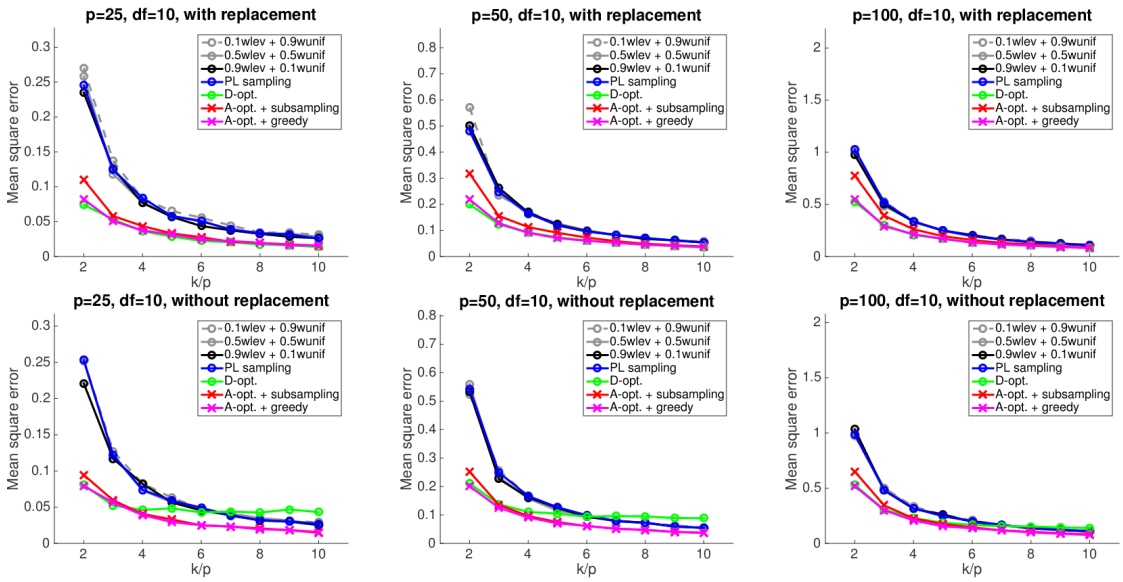

Figure 1: Designs with near-uniformly distributed leverage scores: $\mathbf{X}$ is sampled from an element-wise t distribution with 10 degrees of freedom and is gone through a linear transform such that the resulting covariance $\mathbf{\Sigma}$ satisfies $\mathbf{\Sigma}_{ij} \propto 0.5^{|i-j|}$. Each subset selection algorithm is performed for 50 trials under each setting and the mean of the MSE is reported.

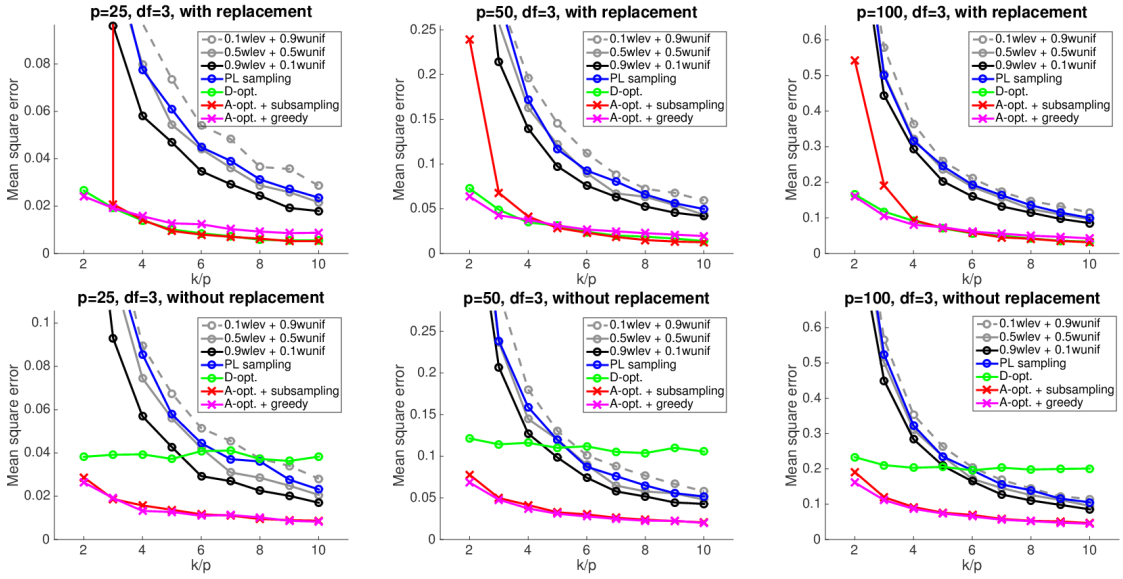

Figure 2: Designs with moderately non-uniformly distributed leverage scores: $\mathbf{X}$ is sampled from an element-wise t distribution with 3 degrees of freedom and is gone through a linear transform such that the resulting covariance $\mathbf{\Sigma}$ satisfies $\mathbf{\Sigma}_{ij} \propto 0.5^{|i-j|}$.

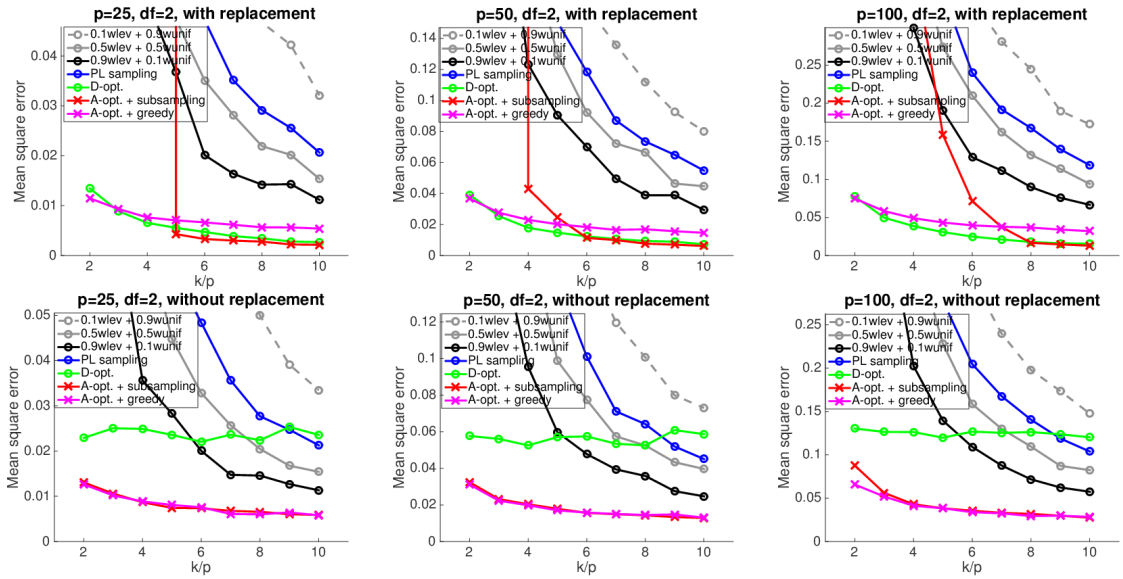

Figure 3: Designs with very non-uniformly distributed leverage scores: $\mathbf{X}$ is sampled from an element-wise t distribution with 2 degrees of freedom and is gone through a linear transform such that the resulting covariance $\mathbf{\Sigma}$ satisfies $\mathbf{\Sigma}_{ij} \propto 0.5^{|i-j|}$.

gorithms (D-optimality, continuous A-optimality with post-processing) under virtually all experimental settings. The performance gap is even larger when the design $\mathbf{X}$ has non-uniformly distributed leverage scores (that is, when df of the t distribution is small).

2. The D-optimality criterion (implemented via the Fedorov exchange algorithm) suffers a lot under the selection without replacement settings. For the selection with replacement setting, it is slightly outperformed by the greedy removal procedure when k is small and by the subsampling post-processing procedure when k is large.

3. The continuous A-optimality with subsampling post-processing algorithm achieves the best performance among all methods when subset size k is large. However, it suffers huge performance degradation when k is as small as the data dimension p, especially under the selection with replacement setting when the full design $\mathbf{X}$ has skewed leverage score distributions. This is explained by our theoretical results in Theorem 3.2, which shows that the subsampling procedure needs more samples (subset size k) on ill-conditioned design matrices.

4. The continuous A-optimality with greedy removal post-processing algorithm is the most robust method and delivers reasonably good subsets under all settings of data dimension, subset size and selection models. It is slightly outperformed by D-optimality and subsampling processed continuous A-optimality under the with replacement setting when k is large. This is because the greedy removal procedure *always* selects distinct data points and hence cannot benefit from repeatedly selecting a few informative design points under the with replacement setting.

Table 2: Number of iterations required to achieve $\delta = 10^{-5}$ relative objective approximation on designs $\mathbf{X}$ with $n = 10000$ rows and $p = 100$ columns. The subset size k (used in the without-replacement objective) and degrees of freedom of the t distributions used to generate $\mathbf{X}$ vary.

$k =$	200	300	400	500	600	700	800	900	1000
$\mathrm{df} = 10$	11	11	11	11	11	11	11	10	9
$\mathrm{df} = 5$	15	16	15	14	10	10	8	8	8
$\mathrm{df} = 3$	15	16	12	10	9	9	8	8	8
$\mathrm{df} = 2$	24	20	19	15	19	17	10	9	11

6 Discussions

6.1 Comparison of computational complexity of subset selection algorithms

Before we remark on the computational efficiency of our proposed subset selection algorithms, we would like to first clarify that our subset selection formulation concerns primarily *sample efficiency* rather than computational efficiency. More specifically, the subset selection step is employed to reduce the number of responses $\boldsymbol{y}$ observed under a measurement constraint framework, rather than to efficiently compute the OLS estimator with full design and responses available.

Nevertheless, our proposed method implemented by the projective gradient descent algorithm in Sec. 4 is computationally efficient, compared to existing subset selection methods in the literature. The per-iteration time complexity of the projective gradient descent implementation of our subset selection algorithm is $O(np^2 + p^3)$, which matches the time complexity for leverage score sampling and is much cheaper than Fedorov exchange algorithm, which requires $O(knp^2)$ operations per iteration. For the greedy removal post-precessing procedure, the continuous A-optimality solution also greatly improves its time efficiency: if the greedy algorithm is directly applied to the full design $\mathbf{X}$ as in [2], its time complexity is $O(n^2p^2)$, which scales quadratically with the total number of data points n; on the other hand, after the support of $\boldsymbol{\pi}^*$ is extracted the number of candidate columns is drastically reduced (independent of n) and hence the greedy removal procedure is much more computationally efficient.

The above paragraph discusses the *per-iteration* time complexity of the continuous A-optimaization problem, while the total number of iterations of the projective gradient descent algorithm could be large (see, for example, Proposition A.2 in Appendix A.3). However, in practice the algorithm typically converges in a very small number of iterations. In Table 2 we report the number of iterations required to achieve $\delta = 10^{-5}$ relative error in the objective function. It can be seen that within 25 iterations the optimization algorithm converges under all experimental settings. Furthermore, the algorithm converges even faster when the subset size k is large, contrary to the Fedorov exchange algorithm whose per-iteration complexity scales linearly with k. It remains an open question to understand, from a theoretical perspective, the behavior of the projective gradient descent method for the conitinuous A-optimality objective.

6.2 Extension to the prediction problem

Our subset selection algorithms can be easily extended to target *prediction error* of response variables. To be fully general, consider the prediction loss function $R(\hat{\boldsymbol{\beta}}, \boldsymbol{\beta}) = \frac{1}{m}\|\mathbf{Z}\hat{\boldsymbol{\beta}} - \mathbf{Z}\boldsymbol{\beta}\|_2^2$, where

$\mathbf{Z} \in \mathbb{R}^{m\times p}$ is a *test matrix* that might differ from the data $\mathbf{X}$ the estimated regression model $\hat{\boldsymbol{\beta}}$ is obtained. [7] As we remark in Sec. 2.2, the relevant combinatorial optimality criterion for prediction purpose would be the (extended) G-optimality criterion $F(S;\mathbf{X},\mathbf{Z}) = \mathrm{tr}[(\mathbf{Z}^\top\mathbf{Z})(\mathbf{X}_S^\top\mathbf{X}_S)^{-1}]$. We thus consider the following convex relaxations:

$$g_{\mathrm{opt}} = \min_{\boldsymbol{\pi}=\pi_1,\cdots,\pi_n} \frac{1}{m}\mathrm{tr}\left((\mathbf{Z}^\top\mathbf{Z})(\mathbf{X}^\top\mathbf{diag}(\boldsymbol{\pi})\mathbf{X})^{-1}\right), \;\; s.t. \;\; \sum_{i=1}^n \pi_i \le 1, \;\; \pi_i \ge 0; \tag{8}$$

$$\tilde{g}_{\mathrm{opt}}(k) = \min_{\boldsymbol{\pi}=\pi_1,\cdots,\pi_n} \frac{1}{m}\mathrm{tr}\left((\mathbf{Z}^\top\mathbf{Z})(\mathbf{X}^\top\mathbf{diag}(\boldsymbol{\pi})\mathbf{X})^{-1}\right), \;\; s.t. \;\; \sum_{i=1}^n \pi_i \le 1, \;\; 0 \le \pi_i \le \frac{1}{k}. \tag{9}$$

Using similar arguments and subsampling procedures in Algorithm 1, we arrive at the following theorem:

Theorem 6.1. *Fix $\mathbf{X} \in \mathbb{R}^{n\times p}$ with full column rank and $\epsilon \in (0, 1/2)$. Let $\mathbf{Z} \in \mathbb{R}^{m\times p}$ be a test data matrix. Then Eq. (6) holds with $f_{\mathrm{opt}}, \tilde{f}_{\mathrm{opt}}$ replaced by $g_{\mathrm{opt}}, \tilde{g}_{\mathrm{opt}}$, $\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\|_2^2$ replaced by $\|\mathbf{Z}\hat{\boldsymbol{\beta}} - \mathbf{Z}\boldsymbol{\beta}\|_2/m$ and $\mathrm{tr}[(\mathbf{X}^\top\mathbf{X})^{-1}]$ replaced by $\mathrm{tr}[(\mathbf{Z}^\top\mathbf{Z})(\mathbf{X}^\top\mathbf{X})^{-1}]/m$. Furthermore, Suppose $k = \Omega(\epsilon^{-2}B^2\|(\boldsymbol{\Sigma}_{*,p})^{-1}\|_2\log(n/\epsilon))$, where $B = \max_i \|\boldsymbol{x}_i\|_2$ and $\boldsymbol{\Sigma}_{*,p} = \frac{1}{m}(\mathbf{Z}^\top\mathbf{Z})\sum_{i=1}^n \pi_i^*\boldsymbol{x}_i\boldsymbol{x}_i^\top$ for the optimal solution $\{\pi_i^*\}_{i=1}^n$ to Eq. (8) or (9). Then with probability at least $1 - O(n^{-1})$ Algorithm 1 returns $\hat{\boldsymbol{\beta}}$ that satisfies* [8]

$$\forall \boldsymbol{\beta} \in \mathbb{R}^p, \qquad \mathbb{E}\left[\frac{\|\mathbf{Z}\hat{\boldsymbol{\beta}} - \mathbf{Z}\boldsymbol{\beta}\|_2^2}{m}\right] \le \begin{cases} (1+\epsilon)\cdot\sigma^2 g_{\mathrm{opt}}/k, & \textit{with replacement}; \\ O(\log k)\cdot\sigma^2\tilde{g}_{\mathrm{opt}}(k)/k, & \textit{without replacement}, \end{cases}$$

where the expecation is taken with respect to $\boldsymbol{y}_S$ and conditioned on the produced subset $\widetilde{\mathbf{X}}_{(t)}$.

The proof of Theorem 6.1 is trivial by substituting $\boldsymbol{\Sigma}_* = \sum_{i=1}^n \pi_i^*\boldsymbol{x}_i\boldsymbol{x}_i^\top$ with $\boldsymbol{\Sigma}_{*,p} = (\mathbf{Z}^\top\mathbf{Z})\sum_{i=1}^n \pi_i^*\boldsymbol{x}_i\boldsymbol{x}_i^\top$ everywhere in the proof of Theorem 3.2.

Finally, we remark that the problem of predicting on a test set different from the training one (i.e., $\mathbf{Z} \neq \mathbf{X}$) is usually referred to as transfer learning in machine learning literature [34]. One major difference is that in transfer learning it is conventional to assume that rows in both $\mathbf{X}$ and $\mathbf{Z}$ are i.i.d. sampled from some underlying data distributions P_X and P_Z, while in this section $\mathbf{X}$ and $\mathbf{Z}$ are assumed to be fixed. In addition, transfer learning generally deals with fully observed labels either for $\mathbf{X}$ (covariance-shift) or for both $\mathbf{X}$ and $\mathbf{Z}$, while our setting concerns the case when we can only observe k out of the n labels for $\mathbf{X}$ and no labels for $\mathbf{Z}$.

6.3 Intepretable bounds for subset selection with random designs

To gain more insights into continuous optimization problems considered in this paper, we discuss other potential solutions $\hat{\boldsymbol{\pi}^*}$ of Eq. (4) that are more interpretable and derive explicit upper bounds on the statistical error of the subsampling based subset selection algorithm, which is strictly better than one that is achieved by uniformly sampling the rows of the full design matrix $\mathbf{X}$. In particular, we have the following theorem that considers the case when each row of $\mathbf{X}$ is sampled i.i.d. from some underlying multivariate Gaussian distribution.

[7] The more narrowed scenario of $\mathbf{Z} = \mathbf{X}$ is not particularly interesting because the prediction risk $O(\sigma^2 p/n)$ does not depend on any properties of the design matrix $\mathbf{X}$, and uniform subsampling would result in a rate of $O(\sigma^2 p/k)$ is asymptotically tight.

[8] The $\mathrm{tr}[(\mathbf{X}_S^\top\mathbf{X}_S)^{-1}]$ criterion in Algorithm 1 should be replaced by $\mathrm{tr}[(\mathbf{Z}^\top\mathbf{Z})(\mathbf{X}_S^\top\mathbf{X}_S)^{-1}]$.

Theorem 6.2. *Let $\boldsymbol{x}_1, \cdots, \boldsymbol{x}_n \sim \mathcal{N}_p(\mathbf{0}, \boldsymbol{\Sigma}_0)$ be n i.i.d. sampled data points and $\mathbf{X} = (\boldsymbol{x}_1, \cdots, \boldsymbol{x}_n) \in \mathbb{R}^{n\times p}$ be the stacked full design matrix. Let $\hat{S} \subseteq [n]$, $|\hat{S}| \leq k$ be the subset returned by Algorithm 1 sampled with replacement, with distribution $\hat{\boldsymbol{\pi}}^*$ specified as*

$$\hat{\boldsymbol{\pi}}_i^* = \frac{1}{Z} \begin{cases} \exp(\boldsymbol{x}_i^\top \boldsymbol{\Sigma}_0^{-1} \boldsymbol{x}_i), & \text{if } \|\boldsymbol{x}_i\|_2^2 \leq 2\mathrm{tr}(\boldsymbol{\Sigma}_0); \\ 0, & \text{otherwise}, \end{cases} \tag{10}$$

where Z is a normalization factor such that $\hat{\boldsymbol{\pi}}^*$ sums to one. Suppose $\hat{\boldsymbol{\beta}}$ is the OLS estimator imposed upon $\mathbf{X}_{\hat{S}}$ and $\boldsymbol{y}_{\hat{S}} = \mathbf{X}_{\hat{S}}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$. If n satisfies

$$n = \Omega\left\{(2\pi ep)^{p/2} \cdot |\boldsymbol{\Sigma}_0| \cdot \exp\left(\mathrm{tr}(\boldsymbol{\Sigma}_0)\|\boldsymbol{\Sigma}_0^{-1}\|_2 - 2p\log \mathrm{tr}(\boldsymbol{\Sigma}_0)\right)\right\}$$

and $k = \Omega(p \log n)$, then with probability at least $1 - O(n^{-1})$ we have that

$$\forall \boldsymbol{\beta} \in \mathbb{R}^p, \qquad \mathbb{E}\left[\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\|_2^2\right] \leq \gamma \cdot \frac{\sigma^2 p^2}{k\mathrm{tr}(\boldsymbol{\Sigma}_0)}, \tag{11}$$

where $\gamma > 0$ is an absolute constant.

We compare the obtained rates in Eq. (11) with the one achieved by uniformly sampling the rows in $\mathbf{X}$. It is straightforward to see that if $\hat{S} \subseteq [n]$, $|\hat{S}| = k$ is obtained by uniformly sampling rows in $\mathbf{X}$, the asymptotic mean-square error of the OLS estimator $\hat{\boldsymbol{\beta}}$ imposed on $\mathbf{X}_{\hat{S}}, \boldsymbol{y}_{\hat{S}}$ admits the form of

$$\mathbb{E}\left[\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\|_2^2\right] \asymp \frac{\sigma^2 \mathrm{tr}(\boldsymbol{\Sigma}_0^{-1})}{k}.$$

Let $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p > 0$ be eigenvalues of $\boldsymbol{\Sigma}_0$. By Jensen's inequality we have

$$\frac{\mathrm{tr}(\boldsymbol{\Sigma}_0^{-1})}{p} = \frac{1}{p}\sum_{i=1}^{p} \frac{1}{\lambda_i} \geq \left(\frac{1}{p}\sum_{i=1}^{n} \lambda_i\right)^{-1} = \frac{p}{\mathrm{tr}(\boldsymbol{\Sigma}_0)}.$$

Subsequently, we have $\sigma^2 \mathrm{tr}(\boldsymbol{\Sigma}_0^{-1})/k \geq \sigma^2 p^2 / k\mathrm{tr}(\boldsymbol{\Sigma}_0)$ and hence the rate of convergence in Eq. (11) is provably faster than that of uniform subsampling . In addition, by Jensen's inequality one expects the gap to be even larger when the underlying Gaussian covariance matrix $\boldsymbol{\Sigma}_0$ is ill-conditioned.

The above argument is not constrained to the Gaussian design setting alone and could be extended easily to general data distributions p_X and also the prediction problem. We give more detailed results in Appendix B, as well as complete proofs in Appendix C.4.

6.4 On conditions of the subset size k

Both subsampling and greedy removal based algorithms require the subset size k to be sufficiently large for the relative/multiplicative approximation bound to hold. Below we again summarize the conditions on k for each of our proposed algorithms (approximation ratio ϵ treated as a constant here):

- **Subsampling**: $k = \Omega(B^2 \|\boldsymbol{\Sigma}_*^{-1}\|_2 \log n)$. Note that the bound is always larger than $\Omega(p \log n)$ and degrades when $\boldsymbol{\Sigma}_*$ is ill-conditioned.

- **Greedy removal**: $k = \Omega(p^2)$.

On the other hand, we only know that $k = \Omega(p)$ is absolute necessary to effectively learn linear regression coefficients on a subset of k observations. We can see that there is an $O(p)$ gap against the lower bound for the greedy removal procedure and a conditioning-dependent gap for the subsampling procedure, which might even be unbounded. This naturally leads us to raise the following important question:

Open question 1. *Given $\mathbf{X} \in \mathbb{R}^{n\times p}$ with full column rank and $k = \Omega(p)$, is it possible to find a subset $\hat{S} \subseteq [n]$, $|\hat{S}| = k$ in polynomial time such that $\mathrm{tr}[(\mathbf{X}_{\hat{S}}^\top \mathbf{X}_{\hat{S}})^{-1}] \leq \gamma \cdot \min_{|S|\leq k} \mathrm{tr}[(\mathbf{X}_S^\top \mathbf{X}_S)^{-1}]$ for some absolute constant $\gamma > 0$?*

We remark on potential improvement of conditions on k following our arguments for subsampling and greedy removal procedures. For the subsampling procedure, the key step is to select *exact* k rows of $\mathbf{X}$ so that the sample covariance of the selected rows approximates the sample covariance of the full design matrix $\mathbf{X}$ with respect to certain distribution specified by $\boldsymbol{\pi}^*$. When k is sufficiently large this can always be done by matrix concentration type inequalities; however, optimal dependency over p on k remains a mystery. In fact, this problem alone is interesting enough to be stated as another open question:

Open question 2. *Given $\mathbf{X} \in \mathbb{R}^{n\times p}$ and $\boldsymbol{\pi}^* \in \mathbb{R}^n$ with $\boldsymbol{\pi}^* \geq \mathbf{0}$, $\|\boldsymbol{\pi}^*\|_1 = 1$, what is the smallest k such that there exists $\hat{S} \subseteq [n]$, $|\hat{S}| = k$ such that $\frac{1}{|\hat{S}|}\mathbf{X}_{\hat{S}}^\top \mathbf{X}_{\hat{S}}$ is a* spectral approximation *(cf. footnote 5) of $\mathbf{X}^\top \mathbf{diag}(\boldsymbol{\pi}^*)\mathbf{X}$?*

Question 2 has interesting connections to *unweighted graph sparsification*, which is known to be related to the famous Kadison-Singer problem [26, 27]. However, two important differences apply. First, in Question 2 each row $\boldsymbol{x}_i$ of $\mathbf{X}$ is weighted by a known weight vector $\boldsymbol{\pi}^*$ and the selected rows are unweighted, while in the standard formulation of unweighted graph sparsification the rows are unweighted (i.e., $\boldsymbol{\pi}^* \equiv 1/n$). In addition, to reduce unweighted graph sparsification to the Kadison-Singer problem the matrix $\mathbf{X}$ (corresponding to the Laplacian of a graph) must be *incoherent*, meaning that the leverage scores of $\mathbf{X}$ are near uniformly distributed. This is a very strong assumption that is not expected to hold in real applications.

For the greedy removal procedure, one might expect improvement of conditions on k to be easier as the current bound $k = \Omega(p^2)$ does not depend on conditioning of any matrix at least. The key component in our argument is an upper bound on the support size of $\boldsymbol{\pi}^*$, the optimal solution to the infinity-norm regularized continuous optimization problem in Eq. (5). To gain deeper insights into this key quantity, we plot in Figure 4 the values of $\|\boldsymbol{\pi}^*\|_0 - k$, $k = p$ against different settings of data dimension p. The resulting plot clearly indicates a linear dependency of $\|\boldsymbol{\pi}^*\|_0 - k$ on p, which is much better than the $\|\boldsymbol{\pi}^*\|_0 = O(k + p^2)$ bound we are able to prove in this paper. We thus formulate the following conjecture based on our simulation findings.

Conjecture 1. *Given $\mathbf{X} \in \mathbb{R}^{n\times p}$ with each row of $\mathbf{X}$ i.i.d. sampled from an underlying p-dimensional distribution with bounded pdf and $p \leq k \leq n$, the optimal solution $\boldsymbol{\pi}^*$ to the continuous problem in Eq. (5) with infinity norm constraint $\|\boldsymbol{\pi}\|_\infty \leq 1/k$ satisfies $\|\boldsymbol{\pi}^*\|_0 = O(k + p)$ almost surely.*

A positive solution to Conjecture 1 would immediately result in polynomial-time algorithm that approximates A-optimality with optimal approximation guarantees (up to constant factors) by invoking analysis of the same greedy removal procedure in [2] and Algorithm 2.

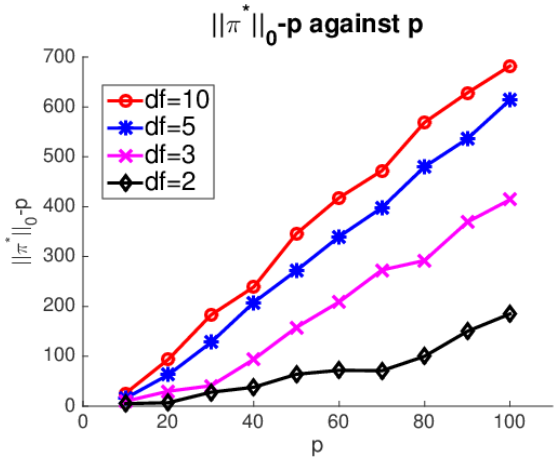

Figure 4: Plot of $\|\boldsymbol{\pi}^*\|_0 - p$ against p, where $\boldsymbol{\pi}^*$ is the optimal solution to the continuous optimization problem in Eq. (5) with infinity norm constraint $\|\boldsymbol{\pi}\|_\infty \leq 1/p$. The number of data points in the full design is $n = 1000$ and data dimension p ranges from 10 to 100. $\mathbf{X}$ is sampled from a t distribution with degrees of freedom taking values in $\{2, 3, 5, 10\}$.

References

[1] A. Alaoui and M. Mahoney. Fast randomized kernel ridge regression with statistical guarantees. In *NIPS*, 2014.

[2] H. Avron and C. Boutsidis. Faster subset selection for matrices and applications. *SIAM Journal on Matrix Analysis and Applications*, 34(4):1464–1499, 2013.

[3] M.-F. Balcan and P. Long. Active and passive learning of linear separators under log-concave distributions. In *COLT*, 2013.

[4] F. Bunea and L. Xiao. On the sample covariance matrix estimator of reduced effective rank population matrices, with applications to fPCA. *Bernoulli*, 21(2):1200–1230, 2015.

[5] K. Chaudhuri, S. Kakade, P. Netrapalli, and S. Sanghavi. Convergence rates of active learning for maximum likelihood estimation. In *NIPS*, 2015.

[6] S. Chen, Y. Liu, M. Lyu, I. King, and S. Zhang. Fast relative-error approximation algorithm for ridge regression. In *UAI*, 2015.

[7] S. Chen, R. Varma, A. Singh, and J. Kovačević. Signal representations on graphs: Tools and applications. *arXiv:1512.05406*, 2015.

[8] A. Çivril and M. Magdon-Ismail. On selecting a maximum volume sub-matrix of a matrix and related problems. *Theoretical Computer Science*, 410(47):4801–4811, 2009.

[9] M. A. Davenport, A. K. Massimino, D. Needell, and T. Woolf. Constrained adaptive sensing. *arXiv preprint arXiv:1506.05889*, 2015.

[10] F. de Hoog and R. Mattheij. A note on subset selection for matrices. *Linear Algebra and its Applications*, 434(8):1845–1850, 2011.

[11] A. Deshpande, L. Rademacher, S. Vempala, and G. Wang. Matrix approximation and projective clustering via volume sampling. *Theory of Computing*, 2:225–247, 2006.

[12] P. Dhillon, Y. Lu, D. Foster, and L. Ungar. New sampling algorithms for fast least squares regression. In *NIPS*, 2013.

[13] P. Drineas, M. W. Mahoney, S. Muthukrishnan, and T. Sarlós. Faster least squares approximation. *Numerische Mathematik*, 117(2):219–249, 2011.

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

[15] A. Frieze, R. Kannan, and S. Vempala. Fast monte-carlo algorithms for finding low-rank approximations. *Journal of the ACM*, 51(6):1025–1041, 2004.

[16] W. Gautschi. The incomplete gamma functions since tricomi. In *Tricomi's Ideas and Contemporary Applied Mathematics, Atti dei Convegni Lincei, n. 147, Accademia Nazionale dei Lincei*. Citeseer, 1998.

[17] M. Gu and S. C. Eisenstat. Efficient algorithms for computing a strong rank-revealing QR factorization. *SIAM Journal on Scientific Computing*, 17(4):848–869, 1996.

[18] E. Hazan and Z. Karnin. Hard margin active linear regression. In *ICML*, 2015.

[19] W. Hoeffding. Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association*, 58(301):13–30, 1963.

[20] R. Hom and C. Johnson. *Topics in Matrix Analysis*. Cambridge UP, New York, 1991.

[21] D. Hsu, S. Kakade, and T. Zhang. A tail inequality for quadratic forms of subgaussian random vectors. *Electronic Communications in Probability*, 17(52):1–6, 2012.

[22] D. Hsu, S. M. Kakade, and T. Zhang. Random design analysis of ridge regression. *Foundations of Computational Mathematics*, 14(3):569–600, 2014.

[23] S. Joshi and S. Boyd. Sensor selection via convex optimization. *IEEE Transactions on Signal Processing*, 57(2):451–462, 2009.

[24] A. Krause, A. Singh, and C. Guestrin. Near-optimal sensor placements in gaussian processes: Theory, efficient algorithms and empirical studies. *Journal of Machine Learning Research*, 9:235–284, 2008.

[25] P. Ma, M. W. Mahoney, and B. Yu. A statistical perspective on algorithmic leveraging. *Journal of Machine Learning Research*, 16(1):861–911, 2015.

[26] A. Marcus, D. Spielman, and N. Srivastava. Interlacing families I: bipartite ramanujan graphs of all degrees. *Annals of Mathematics*, 182:307–325, 2015.

[27] A. Marcus, D. Spielman, and N. Srivastava. Interlacing families II: Mixed characteristic polynomials and the kadison-singer problem. *Annals of Mathematics*, 182:327–350, 2015.

[28] A. Miller and N.-K. Nguyen. A fedorov exchange algorithm for d-optimal design. *Journal of the Royal Statistical Society, Series C (Applied)*, 43(4):669–677, 1994.

[29] F. Pukelsheim. *Optimal design of experiments*, volume 50. SIAM, 1993.

[30] G. Raskutti and M. Mahoney. Statistical and algorithmic perspectives on randomized sketching for ordinary least-squares. In *ICML*, 2015.

[31] S. Sabato and R. Munos. Active regression by stratification. In *NIPS*, 2014.

[32] D. A. Spielman and N. Srivastava. Graph sparsification by effective resistances. *SIAM Journal on Computing*, 40(6):1913–1926, 2011.

[33] G. Stewart and J.-G. Sun. *Matrix Perturbation Theory*. Academic Press, 1990.

[34] M. Sugiyama, M. Krauledat, and K.-R. Müller. Covariance shift adaptation by importance weighted cross validation. *Journal of Machine Learning Research*, 8:985–1005, 2007.

[35] R. Tibshirani. The lasso problem and uniqueness. *Electronic Journal of Statistics*, 7:1456–1490, 2013.

[36] J. A. Tropp. User-friendly tail bounds for sums of random matrices. *Foundations of Computational Mathematics*, 12(4):389–434, 2012.

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

[38] A. W. Yu, H. Su, and F. Li. Efficient euclidean projections onto the intersection of norm balls. In *ICML*, 2012.

[39] R. Zhu, P. Ma, M. Mahoney, and B. Yu. Optimal subsampling approaches for large sample linear regression. *arXiv:1509.05111*, 2015.

A Supplementary materials on computational aspects

A.1 An SDP formulation of the continuously relaxed problem

We consider optimization problems of the form

$$\min_{\boldsymbol{\alpha}} \sum_{j=1}^{N} \boldsymbol{c}_j^\top \mathbf{A}(\boldsymbol{\alpha})^{-1} \boldsymbol{c}_j, \tag{12}$$

where $\{\boldsymbol{c}_j\}_{j=1}^N$ are fixed vectors and $\mathbf{A}(\boldsymbol{\alpha}) = \sum_{i=1}^M \alpha_i \mathbf{A}_i$ for some fixed positive semi-definite matrices $\mathbf{A}_1, \cdots, \mathbf{A}_M$. Both Eq. (4) and (8) can be re-formulated in the form of Eq. (12) by noting that $\mathrm{tr}(\mathbf{Z}^\top \mathbf{A}^{-1} \mathbf{Z}) = \sum_i \boldsymbol{z}_i^\top \mathbf{A}^{-1} \boldsymbol{z}_i$. In this section we show that Eq. (12) can be cast into a semi-definite programming (SDP) problem and can hence be solved in polynomial time. In fact, the problem is equivalent to the following SDP problem:

$$\begin{aligned} \min_{\boldsymbol{\alpha}, t_1, \cdots, t_N} \quad & \sum_{j=1}^{N} t_j \\ s.t. \quad & \mathbf{diag}(\mathbf{B}_1, \cdots, \mathbf{B}_N) \succeq \mathbf{0}, \end{aligned}$$

where

$$\mathbf{B}_j = \begin{bmatrix} \mathbf{A}(\boldsymbol{\alpha}) & \boldsymbol{c}_j \\ \boldsymbol{c}_j^\top & t_j \end{bmatrix}.$$

However, we remark that the SDP formulation is for theoretical purposes only and is not suitable for practical implementation due to its poor scalability. In real-world applications, coordinate descent or even first-order methods like gradient descent might be more appropriate for large-scale data sets.

A.2 Projection onto the intersection of ℓ_1 and ℓ_∞-norm Ball

Similar to [38], 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. The projection problem could be formulated as follows:

$$\begin{aligned} \min_{x,d} \quad & \frac{1}{2}\|\pi - x\|_2^2 \\ \text{s.t.} \quad & x_i \leq d, \ i = 1, 2, ...n, \\ & \|x\|_1 \leq c_1, \\ & d \leq c_2. \end{aligned} \tag{13}$$

The following lemmas illustrate the relation between the primal and dual solution of this problem:

Lemma A.1. *There is a constant λ_2^* at the optimal solution of problem*(13) *such that: If $d^* > 0$, then $\sum_i (\max(\pi_i - \lambda_1^*, 0) - x_i^*) = \lambda_2^*$; If $d^* = 0$, then $\sum_i \max(\pi_i - \lambda_1^*, 0) \leq \lambda_2^*$.*

Lemma A.2. *If x^*, d^* and λ_1^* are the optimal solution of problem* (13)*, then $x_i^* = \min(\max(\pi_i - \lambda_1^*, 0), d^*)$.*

Algorithm 4 Projection onto ℓ_1 and ℓ_∞ norm ball

Input: π, c_1, c_2, δ.
Output: x.
Initialization;
while $|h(\lambda_2)| > \epsilon$ **do**
  if $h(\lambda_2) > \epsilon$ **then**
    $left = \lambda_2$;
  else
    $right = \lambda_2$;
  end if
  $\lambda_2 = (left + right)/2$;
  for each i **do** $x_i = \max(\pi_i - \lambda_2, 0)$;
  $[x, d, \lambda_1, \lambda_2] = \mathcal{P}_\infty^{c_2}(x)$;
end while

Proof. Lemma A.1 and Lemma A.2 can be view as both the special cases of Lemma 6 and 7 of [38]. □

Now we proceed to solve λ_1, λ_2, d that satisfy the following equations:

$$\sum_i \min(\max(\pi_i - \lambda_1, 0), d) - c_1 = 0 \tag{14}$$

$$d - c_2 = 0 \tag{15}$$

$$\sum_i \max(\max(\pi_i - \lambda_1, 0) - d, 0) = \lambda_2 \iff d > 0 \tag{16}$$

$$\sum_i \max(\pi_i - \lambda_1, 0) \leq \lambda_2 \iff d = 0 \tag{17}$$

$$d \geq 0, \lambda_2 \geq 0, \lambda_1 \geq 0. \tag{18}$$

Lemma A.3. *Suppose x^*, d^* and λ_1^*, λ_2^* are the primal and dual solution respectively, then*

$$x_i^* = \min(\max(\pi_i - \lambda_1^*, 0), d^*), \tag{19}$$

$$d^* = \frac{\sum_{i \in \mathcal{S}_{\lambda_1^*}} (\pi_i - \lambda_1^*) - \lambda_2^*}{|\mathcal{S}_{\lambda_1^*}|}, \tag{20}$$

$$\lambda_2^* = \sum_{i \in \mathcal{S}_{\lambda_1^*}} (\pi_i - \lambda_1^*) - c_2 |\mathcal{S}_{\lambda_1^*}|, \tag{21}$$

where $\mathcal{S}_{\lambda_1} = \{i | \pi_i - \lambda_1 > d\}$. .

Proof. (20) and (21) can be obtained by simply solving (15) and (16). (19) is the direct result of Lemma A.2. □

Given a $\lambda_1 \in [0, \max_i(\pi_i)]$, $\forall i, \max(\pi_i - \lambda_1, 0)$ is determined. Hence we can apply the projection operator $\mathcal{P}_\infty^{\tau_2}(\cdot)$ to solve (15) to (18) and obtain the corresponding λ_2, d. However, (14) might not be satisfied. Note that d_i in (14) is dependent on λ_1, so we can define

$$h(\lambda_1) = \sum_i \min(\max(\pi_i - \lambda_1, 0), d(\lambda_1)) - c_1 \tag{22}$$

and the problem has been further distilled to finding the zero point of $h(\lambda_1)$. Next, we will prove that $h(\lambda_1)$ is a strictly monotonically decreasing function, so a binary search is sufficient to solve λ_1.

Theorem A.1. *1) $h(\lambda_1)$ is a continuous piecewise smooth 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. Since $d(\lambda_1)$ is a piecewise smooth function in $(0, \max_i(\pi_i))$, and so does $h(\lambda_1)$. Within each interval, the index set $\mathcal{S}_{\lambda_1}$ remains unchanged. Then it is easy to check that $h'(\lambda_1) < 0$ in each piece of interval. □

Complexity Analysis: Based upon the above theorem, we can determine λ_1^* using the bisection Algorithm 4. For a given λ_1, $\forall i, \max(\pi_i - \lambda_1, 0)$ is determined, λ_2 and d can be solved with time complexity $O(n \log n)$. So the total complexity is $O(n \log n \log(\max_i(\pi_i)/\delta'))$.

A.3 Proof of convergence of Algorithm 3

In this section we prove the convergence of the optimization algorithms presented in Sec. 4. We only prove convergence for the estimation problem with objective

$$\mathcal{L}(\boldsymbol{\pi}) = \operatorname{tr}\left[(\mathbf{X}^\top \mathbf{diag}(\boldsymbol{\pi})\mathbf{X})^{-1}\right],$$

as the convergence of the prediction objective follows by essentially the same argument. Because we're using exact projective gradient descent algorithms, the objective function shall decay monotonically and hence convergence of such algorithms can be established by showing Lipschitz continuity of $\nabla\mathcal{L}$ on a specific *level set*; that is, for some Lipschitz constant $L > 0$ the following holds:

$$\|\nabla\mathcal{L}(\boldsymbol{\pi}) - \nabla\mathcal{L}(\boldsymbol{\pi}')\|_2 \leq L\|\boldsymbol{\pi} - \boldsymbol{\pi}'\|_2, \quad \forall \boldsymbol{\pi}, \boldsymbol{\pi}' \text{ such that } \mathcal{L}(\boldsymbol{\pi}), \mathcal{L}(\boldsymbol{\pi}') \leq \mathcal{L}(\boldsymbol{\pi}^{(0)}), \tag{23}$$

where $\boldsymbol{\pi}^{(0)} = (1/n, 1/n, \cdots, 1/n)$ is the initialization point. Once Eq. (23) holds, linear convergence (i.e., $\mathcal{L}(\boldsymbol{\pi}^{(t)}) - \mathcal{L}(\boldsymbol{\pi}^*) = O(1/t)$) can be established via standard projective gradient analysis.

The main idea of establishing Eq. (23) is to upper bound the spectral norm of the Hessian matrix $\mathbf{H} = \nabla^2\mathcal{L}(\boldsymbol{\pi})$ *uniformly* over all points $\boldsymbol{\pi}$ that satisfies $\mathcal{L}(\boldsymbol{\pi}) \leq \mathcal{L}(\boldsymbol{\pi}^{(0)})$. As a first step, we derive analytic forms of $\mathbf{H}$ in the following proposition:

Proposition A.1. *Let $\widetilde{\boldsymbol{\Sigma}} = \mathbf{X}^\top \mathbf{diag}(\boldsymbol{\pi})\mathbf{X}$. We then have that*

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

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

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

$$\frac{\partial \mathcal{L}}{\partial \pi_i} = \frac{\partial \mathrm{tr}(\widetilde{\boldsymbol{\Sigma}}^{-1})}{\partial \pi_i} = \left\langle \frac{\partial \mathrm{tr}(\widetilde{\boldsymbol{\Sigma}}^{-1})}{\partial \widetilde{\boldsymbol{\Sigma}}}, \frac{\partial \widetilde{\boldsymbol{\Sigma}}}{\partial \pi_i} \right\rangle = \left\langle -\widetilde{\boldsymbol{\Sigma}}^{-2}, \boldsymbol{x}_i \boldsymbol{x}_i^\top \right\rangle = -\boldsymbol{x}_i^\top \widetilde{\boldsymbol{\Sigma}}^{-2} \boldsymbol{x}_i.$$

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

$$\begin{aligned}\frac{\partial^2 \mathcal{L}}{\partial \pi_i \partial \pi_j} = -\frac{\partial (\boldsymbol{x}_i^\top \widetilde{\boldsymbol{\Sigma}}^{-2} \boldsymbol{x}_i)}{\partial \pi_j} &= -\boldsymbol{x}_i^\top \frac{\partial \widetilde{\boldsymbol{\Sigma}}^{-2}}{\partial \pi_j} \boldsymbol{x}_i = \boldsymbol{x}_i^\top \widetilde{\boldsymbol{\Sigma}}^{-2} \frac{\partial \widetilde{\boldsymbol{\Sigma}}^2}{\partial \pi_j} \widetilde{\boldsymbol{\Sigma}}^{-2} \boldsymbol{x}_i \\ &= \boldsymbol{x}_i^\top \widetilde{\boldsymbol{\Sigma}}^{-2} (\boldsymbol{x}_j \boldsymbol{x}_j^\top \widetilde{\boldsymbol{\Sigma}} + \widetilde{\boldsymbol{\Sigma}} \boldsymbol{x}_j \boldsymbol{x}_j^\top) \widetilde{\boldsymbol{\Sigma}}^{-2} \boldsymbol{x}_i = 2(\boldsymbol{x}_i^\top \widetilde{\boldsymbol{\Sigma}}^{-2} \boldsymbol{x}_j) \cdot (\boldsymbol{x}_i^\top \widetilde{\boldsymbol{\Sigma}}^{-1} \boldsymbol{x}_j).\end{aligned}$$

Subsequently,

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

□

Corollary A.1. *Suppose $\boldsymbol{\pi}$ satisfies $\mathcal{L}(\boldsymbol{\pi}) \leq \mathcal{L}(\boldsymbol{\pi}^{(0)}) = \mathrm{tr}(\boldsymbol{\Sigma}_0^{-1})$, where $\boldsymbol{\Sigma}_0 = \frac{1}{n} \mathbf{X}^\top \mathbf{X}$. We then have*

$$\|\mathbf{H}\|_2 = \|\nabla^2 \mathcal{L}(\boldsymbol{\pi})\|_2 \leq 2 \|\mathbf{X}\|_2^4 \cdot [\mathrm{tr}(\boldsymbol{\Sigma}_0^{-1})]^3.$$

Proof. Let $\rho(\mathbf{A}) = \max_i |\sigma_i(\mathbf{A})|$ denote the *spectral range* of matrix $\mathbf{A}$. Clearly, $\|\mathbf{A}\|_2 \leq \rho(\mathbf{A})$ and $\|\mathbf{A}\|_2 = \rho(\mathbf{A})$ for positive semi-definite matrices. In [20] it is established that $\rho(\mathbf{A} \circ \mathbf{B}) \leq \rho(\mathbf{A}) \rho(\mathbf{B})$. Subsequently,

$$\|\mathbf{H}\|_2 \leq 2 \|\mathbf{X}\|_2^4 \|\widetilde{\boldsymbol{\Sigma}}^{-1}\|_2^3 \leq 2 \|\mathbf{X}\|_2^4 \cdot [\mathrm{tr}(\widetilde{\boldsymbol{\Sigma}}^{-1})]^3 \leq 2 \|\mathbf{X}\|_2^4 \cdot [\mathrm{tr}(\boldsymbol{\Sigma}_0^{-1})]^3,$$

where the last inequality is due to the condition that $\mathrm{tr}(\widetilde{\boldsymbol{\Sigma}}^{-1}) = \mathcal{L}(\boldsymbol{\pi}) \leq \mathcal{L}(\boldsymbol{\pi}^{(0)}) = \mathrm{tr}(\boldsymbol{\Sigma}_0^{-1})$. □

From Corollary A.1, we can prove the convergence of the optimization procedures outline in Sec. 4 following standard analysis of projective gradient descent on objective functions with Lipschitz continuous gradient:

Proposition A.2. *Suppose $\boldsymbol{\pi}^{(t)}$ is the solution at the tth iteration of the projective gradient descent algorithm and $\boldsymbol{\pi}^*$ is the optimal solution. We then have*

$$\mathcal{L}(\boldsymbol{\pi}^{(t)}) - \mathcal{L}(\boldsymbol{\pi}^*) \leq \frac{\|\mathbf{X}\|_2^4 \cdot [\mathrm{tr}(\boldsymbol{\Sigma}_0^{-1})]^3 \cdot \|\boldsymbol{\pi}^{(0)} - \boldsymbol{\pi}^*\|_2^2}{\beta t},$$

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

Note that in our case the objective function $\mathcal{L}$ is only Lipschitz continuous on a level set $\{\boldsymbol{\pi} : \mathcal{L}(\boldsymbol{\pi}) \leq \mathcal{L}(\boldsymbol{\pi}^{(0)})\}$. This, however, does not affect the convergence result because we are using batch projective gradient descent with step size chosen via backtracking line search, and hence the objective function monotonically decreases and the solutions $\boldsymbol{\pi}^{(t)}$ will never get outside of the level set.

B Subset selection with random designs

B.1 Linear estimation

Let $\mathbf{X} = (\boldsymbol{x}_1, \cdots, \boldsymbol{x}_n) \in \mathbb{R}^{n\times p}$ be the design matrix. We assume $\boldsymbol{x}_1, \cdots, \boldsymbol{x}_n \overset{i.i.d.}{\sim} F_0$ for some underlying data distribution F_0. We also make the following assumption on F_0:

Assumption B.1. *The covariance matrix $\boldsymbol{\Sigma}_0$ of F_0 exists and is invertible. In addition, there exists a constant $B > 0$ such that for some measurable set $S \subseteq \mathbb{R}^p$ with $\sup_{\boldsymbol{x}\in S} \|\boldsymbol{x}\|_2 \le B$, we have $\Pr[\boldsymbol{x} \in S] \ge 1/2$ and $\int_S \boldsymbol{x}\boldsymbol{x}^\top \mathrm{d}\boldsymbol{x} = C_1(B,p)\mathbf{I}_{p\times p}$, $\int_S 1\mathrm{d}\boldsymbol{x} = C_2(B,p)$ [9] for some constants $C_1(B,p)$ and $C_2(B,p)$ that only depend on B and p. Furthermore, F_0 has a density $p_X(\cdot)$ that satisfies $\inf_{\boldsymbol{x}\in S} p(\boldsymbol{x}) \ge \psi > 0$ for some positive constant ψ.*

Intuitively, Assumption B.1 requires that the underlying distribution F has a density that is bounded away from below on a small isotropic set S. We would then have the following theorem that upper bounds the minimax mean square error of subsampled linear regression estimators.

Theorem B.1. *Fix $\delta \in (0, 1/2)$. Suppose $\boldsymbol{x}_1, \cdots, \boldsymbol{x}_n \sim F_0$ for some underlying data distribution F_0 on $\mathbb{R}^p$ that satisfies Assumption B.1 with parameters B, ψ and constants $C_1(B,p)$, $C_2(B,p)$. Suppose also that*

$$n = \Omega\left(\frac{(1 + B^2\psi^{-1} + \psi^{-2})\log(p/\delta)}{\min\{1, C_1(B,p), C_2(B,p)^2\}}\right).$$

Suppose $\hat{\boldsymbol{\beta}}$ is the OLS estimate on a subset $\hat{S} \subseteq [n]$ of size k produced by Algorithm 1 under the with replacement model. If $k = \Omega(B^2 \log n \cdot C_2(B,p)/C_1(B,p))$ then with probability at least $1 - \delta$ over the random draw of $\mathbf{X}$ and $\hat{S}$, we have

$$\forall \boldsymbol{\beta} \in \mathbb{R}^p, \qquad \mathbb{E}\left[\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\|_2^2\right] \le \gamma \cdot \frac{C_2(B,p)}{C_1(B,p)}\frac{\sigma^2 p}{k} \tag{24}$$

for some absolute constant $\gamma > 0$.

The main intuition behind Theorem B.1 is an explicit weight assignment construction that achieves near-optimal solution for Eq. (4). More specifically, for $\boldsymbol{x}_1, \cdots, \boldsymbol{x}_n$ i.i.d. sampled from F_0, we define a near-optimal solution $\hat{\boldsymbol{\pi}}^*$ as [10]

$$\hat{\pi}_i^* = \frac{1}{Z}\begin{cases} 1/(n' p_X(\boldsymbol{x}_i | \boldsymbol{x}_i \in S)), & \boldsymbol{x}_i \in S; \\ 0, & \text{otherwise.} \end{cases} \tag{25}$$

Here $n' = |\{\boldsymbol{x}_i : \boldsymbol{x}_i \in S\}| \le n$ is the total number of data points in S and Z is a normalization constant to make $\sum_{i=1}^n \hat{\pi}_i^* = 1$; that is, $Z = \frac{1}{n'}\sum_{\boldsymbol{x}_i\in S} 1/p_X(\boldsymbol{x}_i|\boldsymbol{x}_i \in S)$. When the number of data points n is sufficiently large, we have

$$Z\boldsymbol{\Sigma}^* = Z\sum_{i=1}^n \hat{\pi}_i^* \boldsymbol{x}_i \boldsymbol{x}_i^\top \to \mathbb{E}\left[\frac{1}{p_X(\boldsymbol{x}|\boldsymbol{x}\in S)}\boldsymbol{x}\boldsymbol{x}^\top \middle| \boldsymbol{x} \in S\right] = C_2(B,p)\mathbf{I}_{p\times p};$$

[9] If S is a finite set then $\int_S \boldsymbol{x}\boldsymbol{x}^\top \mathrm{d}\boldsymbol{x}$ and $\int_S 1\mathrm{d}\boldsymbol{x}$ should be replaced by $\sum_{\boldsymbol{x}\in S} \boldsymbol{x}\boldsymbol{x}^\top$ and $|S|$.

[10] We assume the underlying data distribution F_0 is known for theoretical analysis. In practice one may consider using density estimation techniques first (e.g., kernel density estimation) to approximately estimate the density of the data distribution.

$$Z = \frac{1}{n'} \sum_{\boldsymbol{x}_i \in S} \frac{1}{p_X(\boldsymbol{x}_i | \boldsymbol{x}_i \in S)} \to \mathbb{E}\left[\frac{1}{p_X(\boldsymbol{x}|\boldsymbol{x} \in S)} \middle| \boldsymbol{x} \in S \right] = C_1(B, p).$$

Applying Theorem 3.2, we immediately arrive at Theorem B.1. A complete version of the proof, which formally establishes the finite-sample guarantees, is given in Appendix C.4.

B.2 Linear prediction

This section concerns the prediction loss function $R(\hat{\boldsymbol{\beta}}, \boldsymbol{\beta}) = \frac{1}{m}\|\mathbf{Z}\hat{\boldsymbol{\beta}} - \mathbf{Z}\boldsymbol{\beta}\|_2^2$. Assume each row of $\mathbf{X}$ or $\mathbf{Z}$ are i.i.d. sampled from some underlying data distributions P_X and P_Z. To simplify presentations, we assume both P_X and P_Z are centered (i.e., have zero mean). In practice one may subtract the empirical mean from both $\mathbf{X}$ and $\mathbf{Z}$ to achieve near-centered data distributions.

To facilitate our analysis, we impose the following three assumptions on the underlying data distributions P_X and P_Z. Assumption B.2 requires both P_X and P_Z to be *sub-Gaussian*; this is essentially needed because we want to bound the difference between the sample covariance matrix and the true covariance of P_X and P_Z, which would be otherwise difficult if the distributions have heavy tails. Assumption B.3 is a technical moment assumption that is used for covariance estimation in [4]. This assumption trivially holds with $c_0 = \pi/2$ [4]. Finally, Assumption B.4 bounds the relative discrepancy on a compact set between the densities of training and testing distributions p_X and p_Z, since in general we cannot hope to do well when P_Z is far from P_X.

Assumption B.2 (Sub-gaussian design)**.** *Both P_X and P_Z are sub-Gaussian with parameter ν^2. That is, for every $\boldsymbol{a} \in \mathbb{R}^p$ we have*

$$\mathbb{E}\left[\exp(\boldsymbol{a}^\top \boldsymbol{x})\right] \le \exp\left\{\|\boldsymbol{a}\|_2^2 \nu^2 / 2\right\}.$$

Assumption B.3 (Polynomial bounds of higher moments)**.** *Let $\boldsymbol{x} \sim P_X$ or $\boldsymbol{x} \sim P_Z$ be a p-dimensional random vector. There exists a constant $c_0 > 0$ such that*

$$\mathbb{E}\left[(\boldsymbol{u}^\top \boldsymbol{x})^2\right] \ge c_0 \|\boldsymbol{u}^\top \boldsymbol{x}\|_{\psi_2}^2$$

for all $\boldsymbol{u} \in \mathbb{R}^p$, where the ψ_2 norm of a p-dimensional random vector $\boldsymbol{Y} \in \mathbb{R}^p$ is defined as

$$\|\boldsymbol{Y}\|_{\psi_2} := \sup_{\boldsymbol{u} \in \mathbb{R}^p \setminus \{\mathbf{0}\}, k \ge 1} \frac{\left(\mathbb{E}|\boldsymbol{u}^\top \boldsymbol{Y}|^k\right)^{1/k}}{\sqrt{k}\|\boldsymbol{u}\|_2}.$$

Assumption B.4 (Bounded discrepancy on a compact set)**.** *There exists $B > 0$ such that $\Pr_X[\|\boldsymbol{x}\|_2 \le B], \Pr_Z[\|\boldsymbol{z}\|_2 \le B] \ge 1/2$. Furthermore, there exists a constant $\xi > 0$ such that*

$$\sup_{\|\boldsymbol{x}\|_2 \le B} \frac{p_Z(\boldsymbol{x})}{p_X(\boldsymbol{x})} \le \xi < \infty,$$

where $p_Z(\cdot)$ and $p_X(\cdot)$ are densities of P_X and P_Z.

We consider the following explicit construction of subsampling distriutions as solution to the optimization problem in Eq. (8):

$$\hat{\pi}_i^* = \frac{1}{Z}\begin{cases} p_Z(\boldsymbol{x}_i)/(n' p_X(\boldsymbol{x}_i)), & \text{if } \|\boldsymbol{x}_i\|_2 \le B; \\ 0, & \text{otherwise.} \end{cases} \tag{26}$$

Here $n' = \{\boldsymbol{x}_i : \|\boldsymbol{x}_i\|_2 \le B\} \le n$ is the number of data points $\boldsymbol{x}_i$ that have bounded norm and Z is a normalization factor so that $\sum_{i=1}^n \hat{\pi}_i^* = 1$. We remark that the subsampling distribution specified in Eq. (26) is closely related to the *importance à la importance reweighting* in the covariance shifting and transfer learning literature [34], where the covariance of the training and testing distributions change while the conditional distribution $y|\boldsymbol{x}$ is assumed to be fixed. Though our setting is arguably similar to covariance shifting, two important differences apply: first, we focus on selecting a few training data points to reduce number of measurements/samples, while covariance shifting and/or transfer learning typically assume all labels on the training data set are known and can be accessed; In addition, the weights specified in Eq. (26) are used for subsampling the data points and afterwards *unweighted* OLS $\mathbf{Z}\hat{\boldsymbol{\beta}} = \mathbf{Z}(\mathbf{X}_S^\top \mathbf{X}_S)^{-1}\mathbf{X}_S^\top \boldsymbol{y}_S$ is carried out to make linear predictions on $\mathbf{Z}$; in contrast, in transfer learning/covariance shifting the relative importance weights are typically used to perform *weighted* empirical risk minimization.

We then have the following theorem that bounds the expected risk using subsampling distributions specified in Eq. (26). Its proof is presented in Appendix C.4.

Theorem B.2. *Fix $\epsilon, \delta \in (0, 1/2)$. Suppose both data distributions P_X and P_Z satisfy Assumptions B.2 through B.4 with $B = \omega(\nu p \log(1/\epsilon))$. Assume also that*

$$\begin{aligned} n &= \Omega\left(\epsilon^{-2}\xi^2 B^2 \|\boldsymbol{\Sigma}_Z\|_2^2 \log(p/\delta)\right); \\ m &= \Omega\left(\epsilon^{-2}\mathrm{tr}(\boldsymbol{\Sigma}_Z)\|\boldsymbol{\Sigma}_Z\|_2 \log(p/\delta)\right). \end{aligned}$$

Suppose $\hat{\boldsymbol{\beta}}$ is the OLS estimate on $\mathbf{X}_{\hat{S}}$ with subset $\hat{S} \subseteq [n]$, $|\hat{S}| = k$ selected by Algorithm 1 under the with replacement model. If $k = \Omega(\epsilon^{-2}B^2 \log(n/\epsilon))$ then with probability at least $1 - \delta$ we have that

$$\forall \boldsymbol{\beta} \in \mathbb{R}^p, \qquad \mathbb{E}\left[\frac{1}{m}\|\mathbf{Z}\hat{\boldsymbol{\beta}} - \mathbf{Z}\boldsymbol{\beta}\|_2^2\right] \le \frac{(1+\gamma\epsilon)\sigma^2 p}{k}.$$

for some absolute constant $\gamma > 0$.

Theorem B.2 shows that when the total number of samples (n and m) is sufficiently large, by employing relative-probability based subsampling one can achieve the near-optimal $\sigma^2 p/k$ rate for prediction even when the testing distribution P_Z is different from the training distribution P_X. On the other hand, without subsampling the asymptotic rate for prediction would be $\sigma^2 \mathrm{tr}(\boldsymbol{\Sigma}_Z \boldsymbol{\Sigma}_X^{-1})/k$, which could be very large when $\boldsymbol{\Sigma}_X$ is not close to $\boldsymbol{\Sigma}_Z$.

C Technical proofs

C.1 Proof of Theorem 3.1

To facilitate the proof of Theorem 3.1, we first formulate two generalized models of the subwith replacement model considered in the main text.

Definition C.1 (Weighted selection model)**.** *An algorithm A first observes the full $n \times p$ design matrix $\mathbf{X}$ and produces, either deterministically or randomly, a submatrix $\widetilde{\mathbf{X}} \in \mathbb{R}^{k\times p}$ whose rows belong to $\mathbf{X}$. The algorithm then* deterministically *outputs k non-negative weights $w_1, \cdots, w_k$ such that $\sum_{i=1}^k w_i = k$. The algorithm then observes $y_i = \sqrt{w_i}\boldsymbol{x}_i^\top \boldsymbol{\beta} + \varepsilon_i$ for $i = 1, \cdots k$ and $\varepsilon_i \overset{i.i.d.}{\sim} \mathcal{N}(0, \sigma^2)$ and estimates the underlying model $\boldsymbol{\beta}$. Furthermore,*

1. *In the* with replacement *model, there may be duplicate rows in $\widetilde{\mathbf{X}}$;*
2. *In the* without replacement *model, rows in $\widetilde{\mathbf{X}}$ must be distinct.*

We use $\mathcal{A}_1^w(k)$ and $\mathcal{A}_2^w(k)$ to denote the classes of all algorithms in this model for sampling with and without replacement settings, respectively.

Definition C.2 (Extended weighted selection model)**.** *An algorithm A first observes the full $n \times p$ design matrix and* deterministically *outputs pairs $\{(w_i, \boldsymbol{x}_i)\}_{i=1}^M \subseteq \mathbb{R} \times \mathbb{R}^p$, where $\boldsymbol{x}_i$ is one of the rows in $\mathbf{X}$ and $\{w_i\}_{i=1}^M$ satisfies $w_i \geq 0$, $\sum_{i=1}^M w_i \leq k$. M is a finite positive integer that could be much larger than n. The algorithm then observes $y_i = \sqrt{w_i}\boldsymbol{x}_i^\top \boldsymbol{\beta} + \varepsilon_i$ for $i = 1, \cdots, M$ and $\varepsilon_i \overset{i.i.d.}{\sim} \mathcal{N}(0, \sigma^2)$ and estimates the underlying model $\boldsymbol{\beta}$. Furthermore,*

1. *In the* with replacement *model, there is no additional constraints on $\{w_i\}_{i=1}^M$;*
2. *In the* without replacement *model, $\{w_i\}_{i=1}^M$ must further satisfy that for any particular row $\boldsymbol{x}_i$ in the original design $\mathbf{X}$, its aggregated weight does not exceed 1.*

We use $\widetilde{\mathcal{A}}_1^w(k)$ and $\widetilde{\mathcal{A}}_2^w(k)$ to denote the classes of all algorithms in this model for sampling with and without replacement settings, respectively.

We are now ready to prove Theorem 3.1.

Proof of Theorem 3.1. We first consider the with replacement model $\mathcal{A}_1(k)$. Note that $\mathcal{A}_1(k) \subseteq \mathcal{A}_1^w(k)$ because any algorithm $A \in \mathcal{A}_1(k)$ can be transformed to an algorithm in $\mathcal{A}_1^w(k)$ by outputing constant weights $w_i \equiv 1$. As a result, we have that

$$\inf_{A\in\mathcal{A}_1^w(k)} \sup_{\boldsymbol{\beta}} \mathbb{E}\left[\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\|_2^2\right] \leq \inf_{A\in\mathcal{A}_1(k)} \sup_{\boldsymbol{\beta}} \mathbb{E}\left[\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\|_2^2\right].$$

On the other hand, $\widetilde{A} \in \widetilde{\mathcal{A}}_1^w(k)$ produces deterministic pairs $\{(w_i, \boldsymbol{x}_i)\}_{i=1}^M$ given $\mathbf{X}$, the optimal estimator of $\boldsymbol{\beta}$ is still the OLS estimator and its minimax error is lower bounded as $\sigma^2 \mathrm{tr}\left((\widetilde{\mathbf{X}}^\top\widetilde{\mathbf{X}})^{-1}\right) = \sigma^2 \mathrm{tr}\left(\left(\sum_{i=1}^n \tilde{w}_i \boldsymbol{x}_i \boldsymbol{x}_i^\top\right)^{-1}\right)$, where $\tilde{w}_i$ is the aggregated weight of data point $\boldsymbol{x}_i$ in all the M weighted pairs. As a result, we have

$$\inf_{\widetilde{A}\in\widetilde{\mathcal{A}}_1^w(k)} \sup_{\boldsymbol{\beta}} \mathbb{E}\left[\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\|_2^2\right] = \inf_{\tilde{w}_1+\cdots+\tilde{w}_n\leq k} \sigma^2 \mathrm{tr}\left(\left(\sum_{i=1}^n \tilde{w}_i \boldsymbol{x}_i \boldsymbol{x}_i^\top\right)^{-1}\right) = \frac{\sigma^2}{k} f_{\mathrm{opt}}.$$

It remains to show that

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

We prove this inequality by showing that for every (possibly random) algorithm $A \in \mathcal{A}_1^w(k)$, there exists $\widetilde{A} \in \widetilde{\mathcal{A}}_1^w(k)$ such that $\sup_{\boldsymbol{\beta}} \mathbb{E}[\|A(\mathbf{X}, \boldsymbol{y}) - \boldsymbol{\beta}\|_2^2] \geq \sup_{\boldsymbol{\beta}} \mathbb{E}[\|\widetilde{A}(\mathbf{X}, \boldsymbol{y}) - \boldsymbol{\beta}\|_2^2]$. 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 $\mathbf{X}$) and its corresponding weight vector $\boldsymbol{w}$, add $(w_i', \tilde{\boldsymbol{x}}_i)$ to the design set of $\widetilde{A}$, where $w_i' = w_i \Pr_A(\widetilde{\mathbf{X}})$. Here $\Pr_A(\widetilde{\mathbf{X}})$ is the probability of A outputing $\widetilde{\mathbf{X}}$.
2. $\widetilde{A}$ observes all responses $\{y_i\}$ for $\{(w_i', \tilde{\boldsymbol{x}}_i)\}$.
3. $\widetilde{A}$ outputs the *expected* estimation of A; that is, $\widetilde{A}(\mathbf{X}, \boldsymbol{y}) = \sum_{\widetilde{\mathbf{X}}} \Pr_A[\widetilde{\mathbf{X}}] \cdot A(\widetilde{\mathbf{X}}, \widetilde{\boldsymbol{y}})$.

We claim that $\widetilde{A} \in \widetilde{\mathcal{A}}_1^w(k)$ because $\sum_i w_i' = \sum_{\widetilde{\mathbf{X}}} \Pr(\widetilde{\mathbf{X}}) \sum_{i=1}^k w_i \leq k$. Furthermore, $\widetilde{A}(\mathbf{X}, \boldsymbol{y}) = \mathbb{E}_{\widetilde{\mathbf{X}}}[\mathbb{E}[A(\mathbf{X}, \boldsymbol{y})|\widetilde{\mathbf{X}}]]$, where the first expectation is taken over the inherent randomness in A that produces $\widetilde{\mathbf{X}}$. by Jensen's inequality we have $\mathbb{E}_{\widetilde{\mathbf{X}}, \boldsymbol{y}}\left[\|A(\mathbf{X}, \boldsymbol{y}) - \boldsymbol{\beta}\|_2^2\right] \geq \mathbb{E}_{\boldsymbol{y}}\left[\|\mathbb{E}_{\widetilde{\mathbf{X}}}[A(\mathbf{X}, \boldsymbol{y})] - \boldsymbol{\beta}\|_2^2\right] = \mathbb{E}_{\boldsymbol{y}}\left[\|\widetilde{A}(\mathbf{X}, \boldsymbol{y}) - \boldsymbol{\beta}\|_2^2\right]$. Taking supreme over $\boldsymbol{\beta}$ we complete the proof.

We next prove the lower bound for the without replacement model. Consider any algorithm A in $\mathcal{A}_2(k)$. We can then construct a deterministic algorithm $\widetilde{A}$ following the same procedure as mentioned above, with weights $w_i = 1$ for each data point $\boldsymbol{x}_i$ sampled in A. Because $\mathcal{A}_2(k) \subseteq \mathcal{A}_1^w(k)$, by above analysis we immediately have that $\widetilde{A} \in \widetilde{\mathcal{A}}_1^w(k)$. To further show $\widetilde{A} \in \widetilde{\mathcal{A}}_2^w(k)$ we only need to prove that in the output of $\widetilde{A}$, the aggregated weight $\tilde{w}_i$ for any data point $\boldsymbol{x}_i$ in the original design does not exceed 1. This is true because $A \in \mathcal{A}_2(k)$ and hence by definition each data point $\boldsymbol{x}_i$ only appears at most once in each of the k-data point samples produced. The aggregated weight $\tilde{w}_i$ for $\boldsymbol{x}_i$ then, by definition, cannot exceed 1. Finally, note that

$$\inf_{A \in \widetilde{\mathcal{A}}_2^w(k)} \sup_{\boldsymbol{\beta}} \mathbb{E}\left[\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\|_2^2\right] = \inf_{\substack{\tilde{w}_1 + \cdots + \tilde{w}_n \leq k, \\ \tilde{w}_1, \cdots, \tilde{w}_n \leq 1}} \sigma^2 \mathrm{tr}\left(\left(\sum_{i=1}^n \tilde{w}_i \boldsymbol{x}_i \boldsymbol{x}_i^\top\right)^{-1}\right) = \frac{\sigma^2 \tilde{f}_{\mathrm{opt}}(k)}{k}.$$

□

C.2 Proof of results for subsampling post-processing

In this section we give the complete proof of Theorem 3.2. Most of the proof is based on the seminal analysis of effective resistance sampling for graph sparsification [32]. We also apply matrix Berstein inequality in order to get smaller failure probability.

Recall that $\boldsymbol{\Sigma}^* = \sum_{i=1}^n \pi_i^* \boldsymbol{x}_i \boldsymbol{x}_i^\top$ is the $p \times p$ covariance matrix of the optimally reweighting obtained from solving Eq. (4). Since we assume the original design matrix $\mathbf{X}$ has full column rank, Eq. (4) has feasible solutions and $\boldsymbol{\Sigma}^*$ is invertible. Let $\boldsymbol{\Phi} = \mathbf{diag}(\boldsymbol{\pi}^*)$ be the optimal reweighting diagnal matrix and define $\boldsymbol{\Pi} = \boldsymbol{\Phi}^{1/2}\mathbf{X}(\boldsymbol{\Sigma}^*)^{-1}\mathbf{X}^\top\boldsymbol{\Phi}^{1/2}$. We have the following proposition that lists properties of the $n \times n$ matrix $\boldsymbol{\Pi}$:

Proposition C.1 (Properties of projection matrix)**.** *The following properties hold for* $\boldsymbol{\Pi}$*:*

1. $\boldsymbol{\Pi}$ *is a projection matrix. That is,* $\boldsymbol{\Pi}^2 = \boldsymbol{\Pi}$.
2. $\mathrm{im}(\boldsymbol{\Pi}) = \mathrm{im}(\boldsymbol{\Phi}^{1/2}\mathbf{X})$.

3. *The eigenvalues of* $\mathbf{\Pi}$ *are 1 with multiplicity* p *and 0 with multiplicity* $n - p$.

4. $\mathbf{\Pi}_{ii} = \|\mathbf{\Pi}_{i,\cdot}\|_2^2 = \pi_i^* \boldsymbol{x}_i^\top (\mathbf{\Sigma}^*)^{-1} \boldsymbol{x}_i$.

Proof. Proof of 1: By definition, $\mathbf{\Sigma}^* = \mathbf{X}^\top \mathbf{\Phi} \mathbf{X}$ and subsequently

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

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

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

and hence $\boldsymbol{u} \in \mathrm{im}(\mathbf{\Pi})$.

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

Proof of 4: By definition,

$$\mathbf{\Pi}_{ii} = \sqrt{\pi_i^*}\boldsymbol{x}_i^\top (\mathbf{\Sigma}^*)^{-1}\boldsymbol{x}_i\sqrt{\pi_i^*} = \pi_i^* \boldsymbol{x}_i^\top (\mathbf{\Sigma}^*)^{-1}\boldsymbol{x}_i.$$

In addition, $\mathbf{\Pi}$ is a symmetric projection matrix. Therefore,

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

□

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 C.1 ([32], Lemma 4)**.** *Let* $\mathbf{\Pi} = \mathbf{\Phi}^{1/2}\mathbf{X}(\mathbf{\Sigma}^*)^{-1}\mathbf{X}^\top \mathbf{\Phi}^{1/2}$ *and* $\mathbf{S}$ *be an* $n \times n$ *non-negative diagonal matrix. If* $\|\mathbf{\Pi}\mathbf{S}\mathbf{\Pi} - \mathbf{\Pi}\|_2 \le \epsilon$ *for some error parameter* $\epsilon \in (0, 1/2)$ *then*

$$(1-\epsilon)\boldsymbol{u}^\top \mathbf{\Sigma}^* \boldsymbol{u} \le \boldsymbol{u}^\top \widetilde{\mathbf{\Sigma}}^* \boldsymbol{u} \le (1+\epsilon)\boldsymbol{u}^\top \mathbf{\Sigma}^* \boldsymbol{u}, \quad \forall \boldsymbol{u} \in \mathbb{R}^p,$$

where $\mathbf{\Sigma}^* = \mathbf{X}^\top \mathbf{\Phi}\mathbf{X}$ *and* $\widetilde{\mathbf{\Sigma}}^* = \mathbf{X}^\top \mathbf{\Phi}\mathbf{S}\mathbf{X}$.

We next introduce a key lemma, which shows that with high probability the covariance matrix $\widetilde{\mathbf{\Sigma}}^*$ of the subsampled data points $\widetilde{\mathbf{X}}$ is a spectral approximation of the covariance $\mathbf{\Sigma}^*$ of the optimal solution to the continuous convex optimization problem Eq. (4).

Lemma C.2. *Fix $\epsilon \in (0, 1/2)$, $t \in \{1, \cdots, T\}$ and suppose $k = \Omega(\epsilon^{-2} B^2 \|(\mathbf{\Sigma}^*)^{-1}\|_2 \log(n/\epsilon))$. Let $\widetilde{\mathbf{\Sigma}}^{(t)} = \frac{1}{k}\sum_{i=1}^{k} \boldsymbol{x}_i^{(t)} \boldsymbol{x}_i^{(t)\top}$ be the covariance matrix of the subsampled data points $\{\widetilde{\mathbf{X}}^{(t)}\}_{t=1}^{T}$. Then with probability at least $1 - 0.01\epsilon$ the following holds for all $\boldsymbol{u} \in \mathbb{R}^p$:*

$$(1 - 0.1\epsilon)\boldsymbol{u}^\top \mathbf{\Sigma}^* \boldsymbol{u} \leq \boldsymbol{u}^\top \widetilde{\mathbf{\Sigma}}^{(t)} \boldsymbol{u} \leq (1 + 0.1\epsilon)\boldsymbol{u}^\top \mathbf{\Sigma}^* \boldsymbol{u}.$$

Proof. Let $\widetilde{\mathbf{X}} \in \mathbb{R}^{k \times p}$ be the subsample design matrix. Define $\mathbf{S}$ to be an $n \times n$ non-negative diagonal matrix with $\mathbf{S}_{ii} = k_i / k p_i$, where $p_i = \pi_i^*$ is the probability of choosing data point $\boldsymbol{x}_i$ and k_i is the number of times $\boldsymbol{x}_i$ is chosen in $\widetilde{\mathbf{X}}$. Since $\mathbf{\Sigma}^* = \mathbf{X}^\top \mathbf{\Phi} \mathbf{X}$ and $\widetilde{\mathbf{\Sigma}}^* = \frac{1}{k}\sum_{i=1}^{k} \tilde{\boldsymbol{x}}_i \tilde{\boldsymbol{x}}_i^\top = \sum_{i=1}^{k} \frac{\pi_i^*}{k p_i} \tilde{\boldsymbol{x}}_i \tilde{\boldsymbol{x}}_i^\top = \mathbf{X}^\top \mathbf{\Phi} \mathbf{S} \mathbf{X}$, by Lemma C.1 we only need to show $\|\mathbf{\Pi}\mathbf{S}\mathbf{\Pi} - \mathbf{\Pi}\|_2 \leq 0.1\epsilon$ with probability at least $1 - 0.1\epsilon$ to obtain spectral approximation between $\mathbf{\Sigma}^*$ and $\widetilde{\mathbf{\Sigma}}^*$.

Let $\boldsymbol{v}$ be an n-dimensional random vector defined as

$$\Pr\left[\boldsymbol{v} = \frac{1}{\sqrt{p_i}} \mathbf{\Pi}_{i,\cdot}\right] = p_i, \quad \forall i = 1, \cdots, n.$$

Also define $\mathbf{A}_\ell = \frac{1}{k} \boldsymbol{v}_\ell \boldsymbol{v}_\ell^\top$, where $\boldsymbol{v}_1, \cdots, \boldsymbol{v}_k$ are i.i.d. sampled from the above specified distribution. Note that

$$\mathbb{E}[\mathbf{\Pi}\mathbf{S}\mathbf{\Pi}] = \mathbb{E}\left[\frac{1}{k}\sum_{t=1}^{k} \boldsymbol{v}_\ell \boldsymbol{v}_\ell^\top\right] = \mathbb{E}[\boldsymbol{v}\boldsymbol{v}^\top] = \sum_{i=1}^{n} p_i \frac{1}{p_i} \mathbf{\Pi}_{i,\cdot} \mathbf{\Pi}_{i,\cdot}^\top = \mathbf{\Pi}^2 = \mathbf{\Pi}.$$

In addition, we have

$$\begin{aligned}\|\boldsymbol{v}\|_2^2 \leq \max_i \frac{1}{p_i} \|\mathbf{\Pi}_{i,\cdot}\|_2^2 &= \max_i \frac{\pi_i^* \boldsymbol{x}_i^\top (\mathbf{\Sigma}^*)^{-1} \boldsymbol{x}_i}{\pi_i^*} \\ &= \max_i \boldsymbol{x}_i^\top (\mathbf{\Sigma}^*)^{-1} \boldsymbol{x}_i \leq \max_i \|\boldsymbol{x}_i\|_2^2 \cdot \|(\mathbf{\Sigma}^*)^{-1}\|_2 \leq B^2 \|(\mathbf{\Sigma}^*)^{-1}\|_2\end{aligned}$$

with probability one. Consequently, we have that

$$\begin{aligned} R = \sup \|\mathbf{A}_\ell\|_2 &\leq \frac{2B^2 \|(\mathbf{\Sigma}^*)^{-1}\|_2}{k}; \\ \nu^2 = \left\|\sum_{\ell=1}^{k} \mathbb{E}[\mathbf{A}_\ell^2]\right\|_2 &\leq \frac{4B^2 \|(\mathbf{\Sigma}^*)^{-1}\|_2}{k}. \end{aligned}$$

Applying matrix Bernstein inequality (Theorem D.3) the following holds

$$\Pr\left[\|\mathbf{\Pi}\mathbf{S}\mathbf{\Pi} - \mathbf{\Pi}\|_2 > \epsilon\right] \leq n \exp\left(-\frac{k\epsilon^2}{(8 + 4\epsilon/3) B^2 \|(\mathbf{\Sigma}^*)^{-1}\|_2}\right). \tag{27}$$

Equating the right-hand side of Eq. (27) with $(1 - 0.01\epsilon)$ and applying Lemma C.1 we obtain $k = \Omega(\epsilon^{-2} B^2 \|(\mathbf{\Sigma}^*)^{-1}\|_2 \log(n/\epsilon))$. □

We are now ready to prove Theorem 3.2.

Proof of Theorem 3.2. We first consider the with replacement model. By Lemma C.2, we know that if $k = \Omega(\epsilon^{-2}B^2\|(\boldsymbol{\Sigma}^*)^{-1}\|_2 \log(n/\epsilon))$ then

$$(1-0.1\epsilon)\boldsymbol{u}^\top \boldsymbol{\Sigma}^* \boldsymbol{u} \le \boldsymbol{u}^\top \widetilde{\boldsymbol{\Sigma}}^{(t)} \boldsymbol{u} \le (1+0.1\epsilon)\boldsymbol{u}^\top \boldsymbol{\Sigma}^* \boldsymbol{u} \tag{28}$$

holds for every $\boldsymbol{u} \in \mathbb{R}^p$ and any fixed $t \in \{1, \cdots, T\}$ with probability at least $1 - 0.01\epsilon$. Setting $T = \Theta(\log n / \log(1+0.04\epsilon)) = \Theta(\epsilon^{-1}\log n)$ the probability that none of the T subsets $\widetilde{\mathbf{X}}_{(t)}$ are spectral approximations of $\boldsymbol{\Sigma}_*$ is at most $O(1/n)$. The bound then follows.

We next analyze the statistical performance of Algorithm 1 under the without replacement setting. Suppose $\boldsymbol{\pi}^*$ is the optimal solution to the optimization problem $\tilde{g}_{\mathrm{opt}}(k)$. We consider the following sampling procedure, which is equivalent to sampling k rows from $\mathbf{X}$ according to $\boldsymbol{\pi}^*$ without replacement:

1. Sample $\boldsymbol{x}_i$ *with replacement* according to the distribution specified by $\boldsymbol{\pi}^*$.
2. If the obtained sample $\boldsymbol{x}_i$ is a duplicate of existing samples, discard this sample and repeat step 1; otherwise, store the newly obtained sample, which is different from all existing samples.
3. Repeat steps 1 and 2 multiple times until k distinct samples are obtained.

Let $\boldsymbol{x}_1, \cdots, \boldsymbol{x}_k$ be the samples obtained through *the first* k *executions* of step 1 in the above procedure and let $\tilde{\boldsymbol{x}}_1, \cdots, \tilde{\boldsymbol{x}}_t$ denote the distinct samples from $\boldsymbol{x}_1, \cdots, \boldsymbol{x}_k$. Note that t may be smaller than k as there might be duplicates in $\boldsymbol{x}_1, \cdots, \boldsymbol{x}_k$. Suppose M is the maximum number of duplicates; that is, a data point appears at most M times in $\boldsymbol{x}_1, \cdots, \boldsymbol{x}_k$. We then have the following inequality by elementary algebra:

$$\mathrm{tr}\left[\left(\sum_{i=1}^{k} \tilde{\boldsymbol{x}}_i \tilde{\boldsymbol{x}}_i^\top\right)^{-1}\right] \le \mathrm{tr}\left[\left(\sum_{i=1}^{t} \tilde{\boldsymbol{x}}_i \tilde{\boldsymbol{x}}_i^\top\right)^{-1}\right] \le M \cdot \mathrm{tr}\left[\left(\sum_{i=1}^{k} \boldsymbol{x}_i \boldsymbol{x}_i^\top\right)^{-1}\right], \tag{29}$$

where $\tilde{\boldsymbol{x}}_1, \cdots, \tilde{\boldsymbol{x}}_k$ are the *distinct* k samples obtained by the without replacement procedure. In addition, because $\boldsymbol{x}_1, \cdots, \boldsymbol{x}_k$ are sampled *with* replacement according to $\boldsymbol{\pi}^*$, by invoking the existing analysis we have that

$$\sigma^2 \cdot \mathrm{tr}\left[\left(\sum_{i=1}^{k} \boldsymbol{x}_i \boldsymbol{x}_i^\top\right)^{-1}\right] \le \frac{(1+0.2\epsilon)\sigma^2}{k} \tilde{f}_{\mathrm{opt}}(k)$$

holds with constant probability. Subsequently, applying Eq. (29) we have that

$$\sigma^2 \cdot \mathrm{tr}\left[\left(\sum_{i=1}^{k} \tilde{\boldsymbol{x}}_i \tilde{\boldsymbol{x}}_i^\top\right)^{-1}\right] \le M \cdot \frac{(1+0.2\epsilon)\sigma^2}{k} \tilde{f}_{\mathrm{opt}}(k) \tag{30}$$

The final part of the proof is to upper bound M, the maximum number of duplicates with high probability. Let $M_0 \in \{0, 1, \cdots, k-1\}$ be an arbitrary threshold. Applying union bound we get

$$\begin{aligned}
\Pr[M > M_0] &\le \sum_{1 \le i_1 < \cdots < i_{M_0+1} \le k} \Pr[\boldsymbol{x}_{i_1} = \boldsymbol{x}_{i_2} = \cdots = \boldsymbol{x}_{i_{M_0+1}}] \\
&\le \binom{k}{M_0+1} \left(\frac{1}{k}\right)^{M_0} = k\binom{k}{M_0+1} \left(\frac{1}{k}\right)^{M_0+1}.
\end{aligned}$$

Here the second inequality is due to the constraint that $\|\boldsymbol{\pi}^*\|_\infty \le 1/k$, which yields $\Pr[\boldsymbol{x}_{i_1} = \cdots = \boldsymbol{x}_{i_{M_0+1}}] \le k^{-M_0}$. Now consider i.i.d. binary random variables $Z_1, \cdots, Z_k$ distributed according to $\Pr[Z_i = 1] = 1/k$ and $\Pr[Z_i = 0] = 1 - 1/k$. By binary theorem,

$$\begin{aligned}
\Pr\left[Z_1 + \cdots + Z_k > M_0\right] &= \sum_{m=M_0+1} k\binom{k}{m}\left(\frac{1}{k}\right)^m\left(1-\frac{1}{k}\right)^{k-m} \\
&\ge \binom{k}{M_0+1}\left(\frac{1}{k}\right)^{M_0+1}\left(1-\frac{1}{k}\right)^{k-M_0-1} \\
&\ge \left(1-\frac{1}{k}\right)^k \cdot \binom{k}{M_0+1}\left(\frac{1}{k}\right)^{M_0+1} \\
&\ge \frac{1}{e}\cdot\binom{k}{M_0+1}\left(\frac{1}{k}\right)^{M_0+1}. \qquad (31)
\end{aligned}$$

On the other hand, applying Chernoff bound (Theorem D.1) with $\mu = \mathbb{E}[Z_1 + \cdots + Z_k] = 1$ we have that

$$\Pr\left[Z_1 + \cdots + Z_k > M_0\right] = \Pr\left[Z_1 + \cdots + Z_k \ge M_0 + 1\right] \le e^{-M_0/3}. \qquad (32)$$

Combining Eq. (31,31) with we arrive at

$$\Pr[M > M_0] \le k\exp\left(-\frac{M_0}{3} + 1\right).$$

Setting $M_0 = \Omega(\log k)$ we can upper bound the right-hand side of the above inequality by a small constant (e.g., 0.01). This means we have $M = O(\log k)$ with constant probability. Applying this result and Eq. (30) we have

$$\sigma^2 \cdot \mathrm{tr}\left[\left(\sum_{i=1}^k \tilde{\boldsymbol{x}}_i\tilde{\boldsymbol{x}}_i^\top\right)^{-1}\right] \le O(\log k)\cdot\frac{\sigma^2}{k}\tilde{f}_{\mathrm{opt}}(k)$$

with constant probability. Finally, repeating the sampling step for $T = \Omega(\log n)$ times we complete the proof. □

C.3 Proof of results for greedy post-processing

Proof of Proposition 3.1. We only need to show that for *arbitrary* $\boldsymbol{x}_1, \cdots, \boldsymbol{x}_{p(p+1)/2} \in \mathbb{R}^p$,

$$\Pr\left[\boldsymbol{x}\middle|\tilde{\phi}(\boldsymbol{x}) \in \mathrm{span}\{\tilde{\phi}(\boldsymbol{x}_1), \cdots, \tilde{\phi}(\boldsymbol{x}_{p(p+1)/2})\}\right] = 0.$$

Denote $\mathcal{E}(\boldsymbol{x}_1, \cdots, \boldsymbol{x}_{p(p+1)/2})$ as the event of $\tilde{\phi}(\boldsymbol{x}) \in \mathrm{span}\{\tilde{\phi}(\boldsymbol{x}_1), \cdots, \tilde{\phi}(\boldsymbol{x}_{p(p+1)/2})\}$ and let $f(\cdot)$ be the pdf associated with the underlying distribution of $\boldsymbol{x}$. Note that $\tilde{\phi}(\boldsymbol{x}) \in \mathrm{span}\{\tilde{\phi}(\boldsymbol{x}_1), \cdots, \tilde{\phi}(\boldsymbol{x}_{p(p+1)/2})\}$ holds if and only if $\tilde{\phi}(\boldsymbol{x})$ satisfies a linear equation with coefficients depending on $\boldsymbol{x}_1, \cdots, \boldsymbol{x}_{p(p+1)/2}$. Since $\tilde{\phi}(\cdot)$ is a quadratic function, we have that

$$\boldsymbol{x} \in \mathcal{E}(\boldsymbol{x}_1, \cdots, \boldsymbol{x}_{p(p+1)/2}) \iff F(\boldsymbol{x}) = 0,$$

where $F : \mathbb{R}^{p(p+1)/2+1} \to \mathbb{R}^p$ is some quadratic function that depends on $\boldsymbol{x}_1, \cdots, \boldsymbol{x}_{p(p+1)/2}$. Subsequently,

$$\Pr\left[\mathcal{E}\middle|\boldsymbol{x}_1, \cdots, \boldsymbol{x}_{p(p+1)/2}\right] = \int_{\mathbb{R}} \cdots \int_{\mathbb{R}} f(\boldsymbol{x}(1), \cdots, \boldsymbol{x}(p)) \cdot 1[F(\boldsymbol{x}) = 0] \mathrm{d}\boldsymbol{x}(1) \cdots \mathrm{d}\boldsymbol{x}(p),$$

where $\boldsymbol{x}(i)$ denotes the ith coordinate of the p-dimensional vector $\boldsymbol{x}$. Because F is a quadratic function, $\boldsymbol{x}(p)$ may take at most two values given $\boldsymbol{x}(1), \cdots, \boldsymbol{x}(p-1)$. Therefore,

$$\Pr\left[\mathcal{E}\middle|\boldsymbol{x}_1, \cdots, \boldsymbol{x}_{p(p+1)/2}\right] = \int_{\mathbb{R}} \cdots \int_{\mathbb{R}} \int_{S(\boldsymbol{x}(1), \cdots, \boldsymbol{x}(p-1))} f(\boldsymbol{x}(1), \cdots, \boldsymbol{x}(p)) \mathrm{d}\boldsymbol{x}(1) \cdots \mathrm{d}\boldsymbol{x}(p),$$

where $|S(\boldsymbol{x}(1), \cdots, \boldsymbol{x}(p-1))| \leq 2$ for all $\boldsymbol{x}(1), \cdots, \boldsymbol{x}(p-1)$. Because f is bounded, we conclude that $\Pr[\mathcal{E}|\boldsymbol{x}_1, \cdots, \boldsymbol{x}_{p(p+1)/2}] = 0$. □

We next prove Theorem 3.3. Our proof can be divided into two steps: at the first step, we show that under Assumption 3.1, the support size of the optimal solution $\boldsymbol{\pi}^*$ can be upper bounded by a function that does not depend on n:

Lemma C.3. *Under Assumption 3.1, the optimal solution $\boldsymbol{\pi}^*$ of Eq. (5) must satisfy $\|\boldsymbol{\pi}^*\|_0 \leq k + p(p+1)/2$.*

Proof. Let $f(\boldsymbol{\pi}) = \mathrm{tr}((\mathbf{X}^\top \mathbf{diag}(\boldsymbol{\pi})\mathbf{X})^{-1})$ be the objective function and denote $\mathcal{L}(\boldsymbol{\pi}; \boldsymbol{\lambda}, \tilde{\boldsymbol{\lambda}}, \mu)$ as the Lagrangian function:

$$\mathcal{L}(\boldsymbol{\pi}; \boldsymbol{\lambda}, \tilde{\boldsymbol{\lambda}}, \mu) = f(\boldsymbol{\pi}) - \sum_{i=1}^n \lambda_i \pi_i + \sum_{i=1}^n \tilde{\lambda}_i \left(\pi_i - \frac{1}{k}\right) + \mu \left(\sum_{i=1}^n \pi_i - 1\right), \qquad \lambda_i, \tilde{\lambda}_i, \mu \geq 0.$$

By KKT condition, $\frac{\partial \mathcal{L}}{\partial \pi_i}\big|_{\boldsymbol{\pi}^*} = \mathbf{0}$ and hence

$$-\frac{\partial f}{\partial \pi_i}\bigg|_{\boldsymbol{\pi}^*} \;=\; \boldsymbol{x}_i^\top \boldsymbol{\Sigma}_*^{-2} \boldsymbol{x}_i \;=\; \tilde{\lambda}_i - \lambda_i + \mu,$$

where $\boldsymbol{\Sigma}_* = \mathbf{X}^\top \mathbf{diag}(\boldsymbol{\pi}^*)\mathbf{X}$.

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

$$\boldsymbol{x}_i^\top \boldsymbol{\Sigma}_*^{-2} \boldsymbol{x}_i = \langle \phi(\boldsymbol{x}_i), \psi(\boldsymbol{\Sigma}_*^{-2}) \rangle = \mu, \qquad \forall i \in B, \tag{33}$$

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 $\boldsymbol{x}_1, \cdots, \boldsymbol{x}_{p(p+1)/2+1}$ be arbitrary distinct $\frac{p(p+1)}{2} + 1$ rows whose indices belong to B. Eq. (33) can then be cast as a homogenous linear system with $\frac{p(p+1)}{2} + 1$ variables and equations as follows:

$$\begin{bmatrix} \tilde{\phi}(\boldsymbol{x}_1) \\ \tilde{\phi}(\boldsymbol{x}_2) \\ \vdots \\ \tilde{\phi}(\boldsymbol{x}_{p(p+1)/2+1}) \end{bmatrix} \begin{bmatrix} \psi(\boldsymbol{\Sigma}_*^{-2}) \\ -\mu \end{bmatrix} \;=\; \mathbf{0}.$$

Under Assumption 3.1, $\mathbf{\Phi} = [\phi(\boldsymbol{x}_1); \cdots; \phi(\boldsymbol{x}_{p(p+1)/2+1})]^\top$ is invertible and hence both $\psi(\mathbf{\Sigma}_*^{-2})$ and μ must be zero. This contradicts the fact that $\mathbf{\Sigma}_*^{-2}$ is positive definite. □

The second step is to show that the greedy removal procedure (originally proposed in [2]) produces a size-k subset $\hat{S}$ that approximates the larger subset $S^* = \{i \in [n] : \boldsymbol{\pi}_i^* > 0\}$ induced by the optimal continuous solution $\boldsymbol{\pi}^*$. For this purpose, we cite results from [2] which shows that $\mathrm{tr}((\mathbf{X}_{\hat{S}}^\top \mathbf{X}_{\hat{S}})^{-1})$ is within a multiplicative factor of $\mathrm{tr}((\mathbf{X}_{S^*}^\top \mathbf{X}_{S^*})^{-1})$.

Lemma C.4 ([2], Theorem 3.1)**.** *Suppose $|S^*| = m$. The greedy row removal algorithm with S^* as input produces a size-k subset $\hat{S} \subseteq S^*$ that satisfies*

$$\mathrm{tr}\left[\left(\mathbf{X}_{\hat{S}}^\top \mathbf{X}_{\hat{S}}\right)^{-1}\right] \;\leq\; \frac{m-p+1}{k-p+1} \cdot \mathrm{tr}\left[\left(\mathbf{X}_{S^*}^\top \mathbf{X}_{S^*}\right)^{-1}\right].$$

We are now ready to give the complete proof of the main theorem.

Proof of Theorem 3.3. Let $S^* = \{i \in [n] : \boldsymbol{\pi}^*\}$ be the support of the optimal continuous solution $\boldsymbol{\pi}^*$. Because $\|\boldsymbol{\pi}^*\|_\infty \leq 1/k$, we have that

$$\mathrm{tr}\left[\left(\mathbf{X}_{S^*}^\top \mathbf{X}_{S^*}\right)^{-1}\right] \leq \frac{1}{k}\mathrm{tr}\left[\left(\mathbf{X}^\top \mathbf{diag}(\boldsymbol{\pi})\mathbf{X}\right)^{-1}\right] = \frac{f(\boldsymbol{\pi}^*)}{k} = \frac{\tilde{f}_{\mathrm{opt}}(k)}{k}.$$

Due to Lemma C.3, the size of S^* can be upper bounded as $|S^*| \leq k + p(p-1)/2$. Applying Lemma C.4 we then have

$$\mathrm{tr}\left[\left(\mathbf{X}_{\hat{S}}^\top \mathbf{X}_{\hat{S}}\right)^{-1}\right] \leq \left(1 + \frac{k + \frac{p(p+1)}{2} - k}{k-p+1}\right) \cdot \mathrm{tr}\left[\left(\mathbf{X}_{S^*}^\top \mathbf{X}_{S^*}\right)^{-1}\right] \leq \left(1 + \frac{p(p+1)}{2(k-p+1)}\right) \cdot \frac{\tilde{f}_{\mathrm{opt}}(k)}{k}.$$

□

C.4 Proof of results for interpretable bounds with random designs

C.4.1 Proof of Theorem B.1

We first prove the following lemma, which shows that assigning $\pi_i^* \propto 1/p(\boldsymbol{x}_i | \boldsymbol{x}_i \in S)$ will produce a near-identity weighted covariance matrix with high probability.

Lemma C.5. *Let F_0 be an underlying data distribution that satisfies Assumption B.1. Fix $\epsilon, \delta \in (0, 1/2)$ and suppose*

$$n = \Omega\left(\frac{(1 + B^2\psi^{-1})\log(p/\delta)}{\min\{\epsilon, \epsilon^2 C_1(B,p)\}}\right).$$

Draw n data points $\boldsymbol{x}_1, \cdots, \boldsymbol{x}_n$ i.i.d. from F_0 and obtain $\boldsymbol{x}_1, \cdots, \boldsymbol{x}_{n'}$ by discarding all data points that are not in S. Then with probability at least $1 - \delta$ the following holds:

$$\left\| \frac{1}{n'} \sum_{i=1}^{n'} \frac{\boldsymbol{x}_i \boldsymbol{x}_i^\top}{p_X(\boldsymbol{x}_i | \boldsymbol{x}_i \in S)} - C_1(B,p) \cdot \mathbf{I}_{p\times p} \right\|_2 \leq \epsilon C_1(B,p). \tag{34}$$

Proof. We use matrix Bernstein inequality (Theorem D.3) as the main tool of proving this lemma. First observe that

$$\mathbb{E}\left[\frac{1}{n'}\sum_{i=1}^{n'}\frac{\boldsymbol{x}_i\boldsymbol{x}_i^\top}{p_X(\boldsymbol{x}_i|\boldsymbol{x}_i\in S)}\right] = \mathbb{E}\left[\frac{\boldsymbol{x}\boldsymbol{x}^\top}{p_X(\boldsymbol{x}|\boldsymbol{x}\in S)}\middle|\boldsymbol{x}\in S\right]$$
$$= \int_S p_X(\boldsymbol{x}|S)\frac{\boldsymbol{x}\boldsymbol{x}^\top}{p_X(\boldsymbol{x}|S)}\mathrm{d}\boldsymbol{x} = \int_S \boldsymbol{x}\boldsymbol{x}^\top\mathrm{d}\boldsymbol{x} = C_1(B,p)\mathbf{I}.$$

Here the last equation is due to Assumption B.1. Define

$$\mathbf{T}_i = \frac{1}{n'}\frac{\boldsymbol{x}_i\boldsymbol{x}_i^\top}{p_X(\boldsymbol{x}_i|\boldsymbol{x}_i\in S)} - \frac{C_1(B,p)}{m}\mathbf{I}.$$

By definition, $\mathbb{E}[\mathbf{T}_i] = \mathbf{0}$. In addition, by Assumption B.1, for any $\boldsymbol{x}\in S$ we have $\|\boldsymbol{x}\|_2\le B$ and $p_X(\boldsymbol{x}|\boldsymbol{x}\in S) = \frac{p_X(\boldsymbol{x})}{p_X(\boldsymbol{x}\in S)} \le \frac{2}{\psi}$. Consequently,

$$R := \max_i \|\mathbf{T}_i\|_2 \le \frac{1}{n'}\left(\frac{B^2}{\psi/2} + C_1(B,p)\right) = O\left(\frac{B^2\psi^{-1} + C_1(B,p)}{n'}\right)$$

and

$$\sigma^2 := \left\|\sum_{i=1}^{n'}\mathbb{E}[\mathbf{T}_i^2]\right\|_2 = n'\left\|\mathbb{E}[\mathbf{T}_1^2]\right\|_2 = \frac{1}{n'}\left\|\mathbb{E}\left[\left(\frac{\boldsymbol{x}_i\boldsymbol{x}_i^\top}{p_X(\boldsymbol{x}_i|S)} - C_1(B,p)\mathbf{I}\right)^2\right]\right\|_2$$
$$= \frac{1}{n'}\left\|\mathbb{E}\left[\frac{\boldsymbol{x}_i\boldsymbol{x}_i^\top\boldsymbol{x}_i\boldsymbol{x}_i^\top}{p_X(\boldsymbol{x}_i|S)^2}\middle|S\right] - C_1(B,p)\mathbf{I}\right\|_2 \le \frac{1}{n'}\left(\left\|\int_S\frac{\boldsymbol{x}\boldsymbol{x}^\top\boldsymbol{x}\boldsymbol{x}^\top}{p_X(\boldsymbol{x}|S)}\mathrm{d}\boldsymbol{x}\right\|_2 + C_1(B,p)\right)$$
$$\le \frac{1}{n'}\left(\frac{B^2}{\psi}\left\|\int_S\boldsymbol{x}\boldsymbol{x}^\top\mathrm{d}\boldsymbol{x}\right\|_2 + C_1(B,p)\right) = O\left(\frac{C_1(B,p)(1+B^2\psi^{-1})}{n'}\right).$$

Matrix Berstein inequality (Theorem D.3) states that Eq. (34) holds with probability $\ge 1-\delta$ if

$$R\log(p/\delta) = O(\epsilon C_1(B,p)), \quad \sigma^2\log(p/\delta) = O(\epsilon^2 C_1(B,p)^2),$$

which yields

$$n' = \Omega\left(\max\left\{\frac{B^2\psi^{-1}\log(p/\delta)}{\epsilon C_1(B,p)}, \frac{\log(p/\delta)}{\epsilon}, \frac{(1+B^2\psi^{-1})\log(p/\delta)}{\epsilon^2 C_1(B,p)}\right\}\right)$$
$$= \Omega\left(\frac{(1+B^2\psi^{-1})\log(p/\delta)}{\min\{\epsilon,\epsilon^2 C_1(B,p)\}}\right).$$

The proof is then completed by noting that $n' = \Omega(n)$ with high probability because $\Pr[\boldsymbol{x}\in S]\ge 1/2$. □

The next lemma upper bounds the normalization constant Z:

Lemma C.6. *Fix $\delta \in (0, 1/2)$ and suppose $n = \Omega(C_2(B,p)^{-2}\psi^{-2}\log(1/\delta))$. Draw n data points $\boldsymbol{x}_1, \cdots, \boldsymbol{x}_n$ i.i.d. from F_0 and obtain $\boldsymbol{x}_1, \cdots, \boldsymbol{x}_{n'}$ by discarding all data points that are not in S. Then with probability at least $1-\delta$ the following holds:*

$$Z = \frac{1}{n'}\sum_{i=1}^{n'} \frac{1}{p_X(\boldsymbol{x}_i | \boldsymbol{x} \in S)} \le 2C_2(B,p). \tag{35}$$

Proof. Define $X_i = 1/p_X(\boldsymbol{x}_i | \boldsymbol{x}_i \in S)$. It is clear by definition that X_i are independent random variables and

$$\mathbb{E}[X_i] = \mathbb{E}\left[\frac{1}{p_X(\boldsymbol{x}|\boldsymbol{x}\in S)}\middle| \boldsymbol{x} \in S\right] = \int_S p_X(\boldsymbol{x}|S)\cdot\frac{1}{p_X(\boldsymbol{x}|S)}\mathrm{d}\boldsymbol{x} = \int_S 1\mathrm{d}\boldsymbol{x} = C_2(B,p).$$

Furthermore, by Assumption B.1, $p_X(\boldsymbol{x}|\boldsymbol{x}\in S) = \frac{p_X(\boldsymbol{x})}{p_X(\boldsymbol{x}\in S)} \le \frac{2}{\psi}$ and hence $0 \le X_i \le 2\psi^{-1}$ almost surely. Applying Hoeffding's inequality (Theorem D.2) we obtain

$$\Pr\left[\frac{1}{n'}\sum_{i=1}^{n'}\frac{1}{p_X(\boldsymbol{x}_i|\boldsymbol{x}_i\in S)} > C_2(B,p) + t\right] \le \exp\left(-\frac{n't^2}{2\psi^{-2}}\right). \tag{36}$$

Setting the right-hand side of Eq. (36) to δ with respect to $t = C_2(B,p)$, we have

$$\Pr\left[\frac{1}{n'}\sum_{i=1}^{n'}\frac{1}{p_X(\boldsymbol{x}_i|\boldsymbol{x}_i\in S)} > 2C_2(B,p)\right] \le \delta$$

if

$$n' = \Omega\left(\frac{\log(1/\delta)}{C_2(B,p)^2\psi^2}\right).$$

The proof is then completed by noting that $n' = \Omega(n)$ with high probability because $\Pr[\boldsymbol{x}\in S] \ge 1/2$. □

We are now ready to prove Theorem B.1.

Proof of Theorem B.1. Fix failure probability $\delta \in (0, 1/2)$ and suppose

$$n = \Omega\left(\frac{(1 + B^2\psi^{-1} + \psi^{-2})\log(p/\delta)}{\min\{1, C_1(B,p), C_2(B,p)^2\}}\right).$$

Applying Lemma C.5 with $\epsilon = 0.5$, we have

$$\left\|\frac{1}{n'}\sum_{i=1}^{n'}\frac{\boldsymbol{x}_i\boldsymbol{x}_i^\top}{p_X(\boldsymbol{x}_i|\boldsymbol{x}_i\in S)} - C_1(B,p)\cdot\mathbf{I}_{p\times p}\right\|_2 \le 0.5C_1(B,p)$$

with probability at least $1-\delta/2$. By Weyl's theorem (Theorem D.4), this implies

$$\sigma_j\left(\frac{1}{n'}\sum_{i=1}^{n'}\frac{\boldsymbol{x}_i\boldsymbol{x}_i^\top}{p_X(\boldsymbol{x}_i|\boldsymbol{x}_i\in S)}\right) \ge 0.5C_1(B,p), \quad \forall j = 1, \cdots, p. \tag{37}$$

On the other hand, by Lemma C.6 the following holds with probability at least $1-\delta/2$:

$$Z = \frac{1}{n'}\sum_{i=1}^{n'} \frac{1}{p_X(\boldsymbol{x}_i|\boldsymbol{x}\in S)} \le 2C_2(B,p). \tag{38}$$

Combine Eq. (37), (38) using union bound and recall the definition of $\hat{\boldsymbol{\pi}}$. We then obtain the following bound on the trace and spectral norm of the weighted covariance matrix with probability at least $1-\delta$:

$$\mathrm{tr}\left(\left(\sum_{i=1}^{n}\hat{\pi}_i^*\boldsymbol{x}_i\boldsymbol{x}_i^\top\right)^{-1}\right) = O\left(p\cdot\frac{C_2(B,p)}{C_1(B,p)}\right), \quad \left\|\left(\sum_{i=1}^{n}\hat{\pi}_i^*\boldsymbol{x}_i\boldsymbol{x}_i^\top\right)^{-1}\right\|_2 = O\left(\frac{C_2(B,p)}{C_1(B,p)}\right).$$

Finally, applying Theorem 3.2 with $\boldsymbol{\Sigma}_* = \sum_{i=1}^n \hat{\pi}_i^*\boldsymbol{x}_i\boldsymbol{x}_i^\top$ we complete the proof of Theorem B.1. □

C.4.2 Proof of Theorem 6.2

Proof of Theorem 6.2. First, note that for any orthogonal matrix $\mathbf{W}\in\mathbb{R}^{p\times p}$, $\mathbf{X}\boldsymbol{\beta} = (\mathbf{X}\mathbf{W})\tilde{\boldsymbol{\beta}}$ with $\tilde{\boldsymbol{\beta}} = \mathbf{W}^\top\boldsymbol{\beta}$ and one may estimate $\boldsymbol{\beta}$ by estimating $\tilde{\boldsymbol{\beta}}$ and then taking the orthogonal transform $\hat{\boldsymbol{\beta}}\mapsto\mathbf{W}\hat{\boldsymbol{\beta}}$. As a result, we may assume without loss of generality that $\boldsymbol{\Sigma}_0 = \mathbf{diag}(\lambda_1,\cdots,\lambda_p)$ with $\lambda_1\ge\cdots\ge\lambda_p>0$.

Let $B^2 = 2\mathrm{tr}(\boldsymbol{\Sigma}_0)$ and $S = \{\boldsymbol{x} : \|\boldsymbol{x}\|_2\le B\}$. Since $\mathbb{E}[\|\boldsymbol{x}\|^2] = \mathrm{tr}(\boldsymbol{\Sigma}_0)$, by Markov's inequality we have $\Pr[\boldsymbol{x}\notin S] = \Pr[\|\boldsymbol{x}\|^2 > 2\mathrm{tr}(\boldsymbol{\Sigma}_0)]\le 1/2$. The other quantities in Assumption B.1 can then be computed by

$$\psi = (2\pi)^{-p/2}\cdot|\boldsymbol{\Sigma}_0|^{-1/2}\cdot\exp\left(-\mathrm{tr}(\boldsymbol{\Sigma}_0)\|\boldsymbol{\Sigma}_0\|_2\right)$$

and

$$C_2(B,p) = \int_{\|\boldsymbol{x}\|_2\le B} 1\mathrm{d}\boldsymbol{x} \approx \frac{1}{\sqrt{p\pi}}\left(\frac{2\pi e}{p}\right)^{p/2}B^p = \frac{1}{\sqrt{p\pi}}\left(\frac{4\pi e}{p}\right)^{p/2}\mathrm{tr}(\boldsymbol{\Sigma}_0)^{p/2}.$$

Here we applied Sterling's formula to approximate the volume of a p-dimensional ball of radius B. To compute the two constants $C_1(B,p)$, note that the uniform distribution over a p-dimensional ball of radius B has covariance matrix $\frac{B^2}{p+2}\mathbf{I}_{p\times p}$. We then have

$$\int_{\|\boldsymbol{x}\|_2\le B}\boldsymbol{x}\boldsymbol{x}^\top\mathrm{d}\boldsymbol{x} = C_2(B,p)\mathbb{E}[\boldsymbol{x}\boldsymbol{x}^\top] = \frac{B^2}{p+2}C_2(B,p)\cdot\mathbf{I}_{p\times p}.$$

Subsequently,

$$C_1(B,p) = \frac{B^2}{p+2}C_2(B,p) = \frac{2\mathrm{tr}(\boldsymbol{\Sigma}_0)}{p+2}C_2(B,p).$$

Finally, applying Theorem B.1 we complete the proof. □

C.4.3 Proof of Theorem B.2

Let $S = \{\boldsymbol{x} \in \mathbb{R}^p : \|\boldsymbol{x}\|_2 \le B\}$, where B is the parameter defined in Assumption B.4. Throughout the proof we shall define the following three quantities:

$$\begin{aligned}
\boldsymbol{\Delta}_1 &:= \frac{1}{m}\mathbf{Z}^\top\mathbf{Z} - \boldsymbol{\Sigma}_Z; \\
\boldsymbol{\Delta}_2 &:= \frac{1}{n'}\sum_{\boldsymbol{x}_i \in S} \frac{p_Z(\boldsymbol{x}_i)}{p_X(\boldsymbol{x}_i|S)}\boldsymbol{x}_i\boldsymbol{x}_i^\top - \int_S p_Z(\boldsymbol{x})\boldsymbol{x}\boldsymbol{x}^\top \mathrm{d}\boldsymbol{x}; \\
\boldsymbol{\Delta}_3 &:= \int_{S^c} p_Z(\boldsymbol{x})\boldsymbol{x}\boldsymbol{x}^\top \mathrm{d}\boldsymbol{x}.
\end{aligned}$$

The following lemmas bound the spectral norm of $\boldsymbol{\Delta}_1, \boldsymbol{\Delta}_2$ and $\boldsymbol{\Delta}_3$ with high probability:

Lemma C.7. *Fix error tolerance parameter $\epsilon \in (0, 1/2)$. Suppose Assumptions B.2 and B.3 hold with $B = \omega(p\nu\log(1/\epsilon))$. Then we have*

$$\|\boldsymbol{\Delta}_3\|_2 = O(\epsilon).$$

Proof. By definition,

$$\begin{aligned}
\|\boldsymbol{\Delta}_3\|_2 = \left\|\int_{\|\boldsymbol{x}\|_2 \ge B} p_Z(\boldsymbol{x})\boldsymbol{x}\boldsymbol{x}^\top \mathrm{d}\boldsymbol{x}\right\|_2 &\le \int_{\|\boldsymbol{x}\|_2 \ge B} p_Z(\boldsymbol{x})\|\boldsymbol{x}\|_2^2 \mathrm{d}\boldsymbol{x} \\
&= S_p \int_B^\infty a^{2+p-1} p_Z(\|\boldsymbol{x}\|_2 = a)\mathrm{d}a \le S_p \int_B^\infty a^{p+1} p_Z(\|\boldsymbol{x}\|_2 \ge a)\mathrm{d}a.
\end{aligned}$$

Here in the second to last equation we apply change of variables from the Cartesian system to the spherical (polar) system and S_p is the area of the unit sphere in $\mathbb{R}^p$. In addition, because P_Z is subgaussian with parameter ν^2, applying Theorem 1 in [21] we obtain

$$\Pr_Z\left[\|\boldsymbol{x}\|_2 > a\right] \le \exp\left\{-\frac{a^2}{\nu^2 p}\right\}$$

for all $a > \nu\sqrt{p}$. Subsequently, when $\nu = O(B/\sqrt{p})$ we have

$$\|\boldsymbol{\Delta}_3\|_2 \le S_p \int_B^\infty a^{p+1} e^{-a^2/\nu^2 p}\mathrm{d}a = S_p \cdot \frac{1}{2} p^{p/2+1}\nu^{p+2}\Gamma\left(\frac{p}{2}+1, \frac{B^2}{p\nu^2}\right),$$

where $\Gamma(a, x) = \int_x^\infty t^{a-1}e^{-t}\mathrm{d}t$ is the *incomplete Gamma function*. By Theorem D.5, when $B = \omega(p\nu)$ we have

$$\Gamma\left(\frac{p}{2}+1, \frac{B^2}{p\nu^2}\right) = O\left(e^{-B^2/p\nu^2}\left(\frac{B^2}{p\nu^2}\right)^{p/2+1}\right).$$

In addition, S_p adopts the following closed-form expression and asymptotic approximation:

$$S_p = \frac{2\pi^{\frac{p+1}{2}}}{\Gamma(\frac{p+1}{2})} = \frac{2\pi^{\frac{p+1}{2}}}{\sqrt{\frac{4\pi}{p+1}}\left(\frac{p+1}{2e}\right)^{\frac{p+1}{2}}\left(1+O\left(\frac{1}{p}\right)\right)} = \sqrt{\frac{p+1}{\pi}}\left(\frac{2\pi e}{p+1}\right)^{\frac{p+1}{2}}\left(1+O\left(\frac{1}{p}\right)\right).$$

Subsequently,

$$\|\mathbf{\Delta}_3\|_2 = O\left(\sqrt{\frac{p+1}{\pi}}\left(\frac{2\pi e}{p+1}\right)^{\frac{p+1}{2}} B^{p+2} e^{-B^2/p\nu^2}\right).$$

Under the assumption that $B = \omega(\nu p \log(1/\epsilon))$ we have $\|\mathbf{\Delta}_3\|_2 = O(\epsilon)$. □

Lemma C.8. *Fix error tolerance parameter $\epsilon \in (0, 1/2)$ and failure parameter $\delta \in (0, 1/2)$. Assume also $\epsilon < \|\mathbf{\Sigma}_Z\|_2$. Suppose Assumption B.2 through B.4 hold and Lemma C.7 is valid. If*

$$\begin{aligned} m &= \Omega\left(\epsilon^{-2}\mathrm{tr}(\mathbf{\Sigma}_Z)\|\mathbf{\Sigma}_Z\|_2 \log(p/\delta)\right), \\ n &= \Omega\left(\epsilon^{-2}\xi B^2 \|\mathbf{\Sigma}_Z\|_2^2 \log(p/\delta)\right), \end{aligned}$$

then with probability at least $1-\delta$ (over the random draw of $\mathbf{X} \in \mathbb{R}^{n\times p}$ and $\mathbf{Z} \in \mathbb{R}^{m\times p}$) the following holds:

$$\max\left\{\|\mathbf{\Delta}_1\|_2, \|\mathbf{\Delta}_2\|_2\right\} \le 0.1\epsilon.$$

Proof. We first prove the bound on $\|\mathbf{\Delta}_1\|_2$. Theorem 2.2 in [4] states that when Assumptions B.2 and B.3 hold, with probability at least $1-\delta/3$ we have

$$\left\|\frac{1}{m}\mathbf{Z}^\top\mathbf{Z} - \mathbf{\Sigma}_Z\right\|_2 \le C \cdot \|\mathbf{\Sigma}_Z\|_2 \cdot \max\left\{\sqrt{\frac{r_e(\mathbf{\Sigma}_Z)\ln(p/\delta)}{m}}, \frac{r_e(\mathbf{\Sigma}_Z)\ln(p/\delta)}{m}\right\},$$

where $r_e(\mathbf{\Sigma}) = \mathrm{tr}(\mathbf{\Sigma})/\|\mathbf{\Sigma}\|_2$ denotes the *stable rank* of a covariance matrix $\mathbf{\Sigma}$. Here C is a constant that only depends on the sub-gaussian parameters ν^2 and c_0 in Assumptions B.2 and B.3. As a result, we have $\|\frac{1}{m}\mathbf{Z}^\top\mathbf{Z} - \mathbf{\Sigma}_Z\|_2 \le 0.1\epsilon$ with probability $\ge 1-\delta/3$ provided that $m = \Omega(\epsilon^{-2}\mathrm{tr}(\mathbf{\Sigma}_Z)\|\mathbf{\Sigma}_Z\|_2\log(p/\delta))$.

We next turn to the bound on $\|\mathbf{\Delta}_2\|_2$. We apply matrix Bernstein inequality (Theorem D.3) to prove this bound. By Lemma C.7, when $B = \omega(p\nu\log(1/\epsilon))$ we have $\|\mathbf{\Sigma}_3\|_2 \le 0.1\epsilon$. This gives us $\|\mathbf{\Sigma}_Z^S\|_2 \le \|\mathbf{\Sigma}_Z\|_2 + \epsilon$, where $\mathbf{\Sigma}_Z^S = \int_S p_Z(\boldsymbol{x})\boldsymbol{x}\boldsymbol{x}^\top \mathrm{d}\boldsymbol{x} = \mathbf{\Sigma}_Z + \mathbf{\Delta}_3$.

Let $\mathbf{T}_i$ be a $p \times p$ random matrix defined as

$$\mathbf{T}_i = \frac{1}{n'} \cdot \frac{p_Z(\boldsymbol{x}_i)}{p_X(\boldsymbol{x}_i|S)}\boldsymbol{x}_i\boldsymbol{x}_i^\top - \frac{1}{n'}\int_S p_Z(\boldsymbol{x})\boldsymbol{x}\boldsymbol{x}^\top \mathrm{d}\boldsymbol{x}$$

with probability $p(\boldsymbol{x}_i|S)$ for some $\boldsymbol{x}_i \in S$. Here $n' = |\{\boldsymbol{x}_i : \|\boldsymbol{x}_i\|_2 \le B\}| = |\{\boldsymbol{x}_i : \boldsymbol{x}_i \in S\}|$ is the number of data points in $\mathbf{X}$ that belong to S. We then have

$$\begin{aligned} \mathbb{E}[\mathbf{T}_i] &= \frac{1}{n'}\mathbb{E}\left[\frac{p_Z(\boldsymbol{x})}{p_X(\boldsymbol{x}|S)}\boldsymbol{x}\boldsymbol{x}^\top\right] - \frac{1}{n'}\int_S p_Z(\boldsymbol{x})\boldsymbol{x}\boldsymbol{x}^\top \mathrm{d}\boldsymbol{x} \\ &= \frac{1}{n'}\int_S \left(p_X(\boldsymbol{x}|S)\cdot\frac{p_Z(\boldsymbol{x})}{p_X(\boldsymbol{x}|S)} - p_Z(\boldsymbol{x})\right)\boldsymbol{x}\boldsymbol{x}^\top \mathrm{d}\boldsymbol{x} \\ &= \mathbf{0}. \end{aligned}$$

In addition, by Assumption B.4 we have

$$\|\mathbf{T}_i\|_2 \le \frac{2\xi B^2 + \|\mathbf{\Sigma}_Z^S\|_2}{n'}, \quad a.s.$$

and

$$\begin{aligned}
\sigma^2 &= n' \|\mathbb{E}[\mathbf{T}_1^2]\|_2 \\
&= \frac{1}{n'} \left\| \mathbb{E}\left[\left(\frac{p_Z(\boldsymbol{x})}{p_X(\boldsymbol{x}|S)} \boldsymbol{x}\boldsymbol{x}^\top - \int_S p_Z(\boldsymbol{x}) \boldsymbol{x}\boldsymbol{x}^\top \mathrm{d}\boldsymbol{x} \right)^2 \right] \right\|_2 \\
&= \frac{1}{n'} \left\| \mathbb{E}\left[\frac{p_Z(\boldsymbol{x})^2}{p_X(\boldsymbol{x}|S)^2} \boldsymbol{x}\boldsymbol{x}^\top \boldsymbol{x}\boldsymbol{x}^\top \right] - \left(\int_S p_Z(\boldsymbol{x}) \boldsymbol{x}\boldsymbol{x}^\top \mathrm{d}\boldsymbol{x} \right)^2 \right\|_2 \\
&\le \frac{1}{n'} \left(\left\| \int_S \frac{p_Z(\boldsymbol{x})^2}{p_X(\boldsymbol{x}|S)} \boldsymbol{x}\boldsymbol{x}^\top \boldsymbol{x}\boldsymbol{x}^\top \mathrm{d}\boldsymbol{x} \right\|_2 + \|\boldsymbol{\Sigma}_Z^S\|_2^2 \right) \\
&\le \frac{1}{n'} \left(2\xi B^2 \left\| \int_S p_Z(\boldsymbol{x}) \boldsymbol{x}\boldsymbol{x}^\top \mathrm{d}\boldsymbol{x} \right\|_2 + \|\boldsymbol{\Sigma}_Z^S\|_2^2 \right) \\
&= \frac{2\xi B^2 \|\boldsymbol{\Sigma}_Z^S\|_2 + \|\boldsymbol{\Sigma}_Z^S\|_2^2}{n'}.
\end{aligned}$$

Applying matrix Bernstein inequality (Theorem D.3) with $R = (2\xi B^2 + \|\boldsymbol{\Sigma}_Z^S\|_2)/n'$ and $\sigma^2 = (2\xi B^2 \|\boldsymbol{\Sigma}_Z^S\|_2 + \|\boldsymbol{\Sigma}_Z^S\|_2^2)/n'$ we obtain $\|\boldsymbol{\Delta}_2\|_2 \le \epsilon$ with probability $\ge 1 - \delta$, provided that

$$\begin{aligned}
n' &= \Omega\left(\xi B^2 \|\boldsymbol{\Sigma}_Z^S\|_2 \log(p/\delta) \cdot \max\left\{ \frac{1}{\epsilon}, \frac{\|\boldsymbol{\Sigma}_Z^S\|_2}{\epsilon^2} \right\} \right) \\
&= \Omega\left(\xi B^2 \|\boldsymbol{\Sigma}_Z\|_2 \log(p/\delta) \cdot \max\left\{ \frac{1}{\epsilon}, \frac{\|\boldsymbol{\Sigma}_Z\|_2}{\epsilon^2} \right\} \right),
\end{aligned}$$

where the last inequality is due to the assumption that $\epsilon \le \|\boldsymbol{\Sigma}\|_2$. Finally, note that $n' = \Omega(n)$ with high probability because $\Pr_X[\boldsymbol{x} \in S] \ge 1/2$. □

The next lemma bounds the spectral norm as well as trace of the weighted covariance matrix $\widetilde{\boldsymbol{\Sigma}}_{\mathrm{pred}}^*$ under the subsampling probabilities $\hat{\boldsymbol{\pi}}^*$ specified in Eq. (26).

Lemma C.9. *Let $\widetilde{\boldsymbol{\Sigma}}_{\mathrm{pred}}^* = \frac{1}{m}(\mathbf{Z}^\top \mathbf{Z})(\sum_{i=1}^n \hat{\pi}_i^* \boldsymbol{x}_i \boldsymbol{x}_i^\top)^{-1}$ be the weighted covariance matrix as in Eq. (8), where the subsampling probabilities $\{\hat{\pi}_i^*\}_{i=1}^n$ are specified in Eq. (26). Suppose $\max(\|\boldsymbol{\Delta}_1\|_2, \|\boldsymbol{\Delta}_2\|_2, \|\boldsymbol{\Delta}_3\|_2) \le \epsilon\sigma_p(\boldsymbol{\Sigma}_Z)$, where $\epsilon \in (0, 1/6)$ is some error tolerance parameter and $\sigma_p(\boldsymbol{\Sigma}_Z)$ is the least singular value of the underlying covariance matrix $\boldsymbol{\Sigma}_Z$. We then have*

$$\|\widetilde{\boldsymbol{\Sigma}}_{\mathrm{pred}}^*\|_2 \le (1 + 3.5\epsilon)Z \quad \textit{and} \quad g_{\mathrm{opt}} = \mathrm{tr}(\widetilde{\boldsymbol{\Sigma}}_{\mathrm{pred}}^*) \le (1 + 3.5\epsilon)pZ$$

where Z is the normalization constant defined in Eq. (26).

Proof. By definition, we have $Z^{-1}\widetilde{\boldsymbol{\Sigma}}_{\mathrm{pred}}^* = (\boldsymbol{\Sigma}_Z + \boldsymbol{\Delta}_1)(\boldsymbol{\Sigma}_Z + \boldsymbol{\Delta}_2 + \boldsymbol{\Delta}_3)^{-1}$. Let $\boldsymbol{\Delta} = \boldsymbol{\Delta}_1$ and $\boldsymbol{\Delta}' = \boldsymbol{\Delta}_2 + \boldsymbol{\Delta}_3$. It then amounts to bound the spectral and trace norm of $(\boldsymbol{\Sigma}_Z + \boldsymbol{\Delta})(\boldsymbol{\Sigma}_Z + \boldsymbol{\Delta}')^{-1}$. Write $\mathbf{A} = \boldsymbol{\Delta}(\boldsymbol{\Sigma}_Z + \boldsymbol{\Delta}')^{-1}$ and $\mathbf{B} = \boldsymbol{\Sigma}_Z(\boldsymbol{\Sigma}_Z + \boldsymbol{\Delta}')^{-1}$. We can then bound the norm (or singular values) of $\mathbf{A}$ and $\mathbf{B}$ separately.

We start with bounding $\|\mathbf{A}\|_2$. Because $\|\boldsymbol{\Delta}'\|_2 \le 2\epsilon\sigma_p(\boldsymbol{\Sigma}_Z)$, applying Weyl's theorem (Theorem D.4) we have

$$\sigma_p(\boldsymbol{\Sigma}_Z + \boldsymbol{\Delta}') \ge \sigma_p(\boldsymbol{\Sigma}_Z) - \|\boldsymbol{\Delta}'\|_2 \ge (1 - 2\epsilon)\sigma_p(\boldsymbol{\Sigma}_Z).$$

Subsequently,

$$\|\mathbf{A}\|_2 \le \|\mathbf{\Delta}\|_2 \|(\mathbf{\Sigma}_Z + \mathbf{\Delta}')^{-1}\|_2 \le \epsilon \sigma_p(\mathbf{\Sigma}_Z) \cdot \frac{\sigma_p(\mathbf{\Sigma}_Z)^{-1}}{1-2\epsilon} \le \epsilon(1+3\epsilon) \le 1.5\epsilon, \tag{39}$$

where the last two inequalities is due to the assumption that $\epsilon \in (0, 1/6)$.

We next turn to bound the singular values of $\mathbf{B}$. Since both $\mathbf{\Sigma}_Z$ and $(\mathbf{\Sigma}_Z + \mathbf{\Delta}')$ are invertible, we have $\mathbf{B} = (\mathbf{I} + \mathbf{\Delta}'\mathbf{\Sigma}_Z^{-1})^{-1}$. In addition, $\|\mathbf{\Delta}'\mathbf{\Sigma}_Z^{-1}\|_2 \le \|\mathbf{\Delta}'\|_2 \|\mathbf{\Sigma}_Z^{-1}\|_2 \le \epsilon \sigma_p(\mathbf{\Sigma}_Z)\sigma_p(\mathbf{\Sigma}_Z)^{-1} \le \epsilon$. Consequently, Weyl's theorem yields

$$\sigma_r(\mathbf{I} + \mathbf{\Delta}'\mathbf{\Sigma}_Z^{-1}) \ge 1 - \epsilon, \quad \forall r = 1, \cdots, p. \tag{40}$$

The proof is then completed by combining Eq. (39) and (40) and noting that $1/(1-\epsilon) \le 1 + 2\epsilon$ by conditions on ϵ. □

The following lemma upper bounds the normalization constant Z in Eq. (26):

Lemma C.10. *Fix $\epsilon, \delta \in (0, 1/2)$. Let $\boldsymbol{x}_1, \cdots, \boldsymbol{x}_n$ be i.i.d. sampled from the underlying data distribution F_0 and let $\boldsymbol{x}_1, \cdots, \boldsymbol{x}_{n'}$ be the data points after discarding all $\boldsymbol{x}_i$ with $\|\boldsymbol{x}_i\|_2 > B$. Suppose*

$$n = \Omega(\epsilon^{-2}\xi^2 \log(1/\delta)).$$

Then with probability at least $1 - \delta$ the following bound on Z holds:

$$Z = \frac{1}{n'} \sum_{i=1}^{n'} \frac{p_Z(\boldsymbol{x}_i)}{p_X(\boldsymbol{x}_i | \boldsymbol{x}_i \in S)} \le 1 + \epsilon. \tag{41}$$

Proof. We use Hoeffding's inequality (Theorem D.2) to prove this lemma. Define random variable A_i as

$$A_i = \frac{p_Z(\boldsymbol{x}_i)}{p_X(\boldsymbol{x}_i | S)}.$$

By definition, A_i are i.i.d. random variables and

$$\mathbb{E}[A_i] = \int_{\|\boldsymbol{x}\|_2 \le B} p_X(\boldsymbol{x}|S) \cdot \frac{p_Z(\boldsymbol{x})}{p_X(\boldsymbol{x}|S)} \mathrm{d}\boldsymbol{x} = \int_{\|\boldsymbol{x}\|_2 \le B} p_Z(\boldsymbol{x}) \mathrm{d}\boldsymbol{x}.$$

On the other hand, by Assumption B.4 we have

$$|A_i| = \left| \frac{p_Z(\boldsymbol{x}_i)}{p_X(\boldsymbol{x}_i|S)} \right| \le 2 \left| \frac{p_Z(\boldsymbol{x}_i)}{p_X(\boldsymbol{x}_i)} \right| \le 2\xi$$

with probability 1. Hoeffding's inequality (Theorem D.2) then yields

$$\Pr\left[\left| Z - \int_{\|\boldsymbol{x}\|_2 \le B} p_Z(\boldsymbol{x}) \mathrm{d}\boldsymbol{x} \right| \ge \epsilon \right] \le 2 \exp\left(-\frac{2n'\epsilon^2}{4\xi^2} \right). \tag{42}$$

Equating the right-hand side of Eq. (42) with δ we get

$$Z \le \epsilon + \int_{\|\boldsymbol{x}\|_2 \le B} p_Z(\boldsymbol{x}) \mathrm{d}\boldsymbol{x}$$

with probability at least $1-\delta$, provided that

$$n' = \Omega\left(\epsilon^{-2}\xi^2 \log(1/\delta)\right).$$

Finally, note that

$$\int_{\|\boldsymbol{x}\|_2 \le B} p_Z(\boldsymbol{x})\mathrm{d}\boldsymbol{x} \le \int_{\mathbb{R}^p} p_Z(\boldsymbol{x})\mathrm{d}\boldsymbol{x} = 1$$

and $n' = \Omega(n)$ by Assumption B.4. □

We are now ready to prove Theorem B.2.

Proof of Theorem B.2. By Lemma C.9 and C.10, we have (with high probability) that

$$\|\widetilde{\boldsymbol{\Sigma}}^*_{\mathrm{pred}}\|_2 = (1+O(\epsilon)) \quad \text{and} \quad g_{\mathrm{opt}} = p(1+O(\epsilon)).$$

Applying Theorem 6.1 we complete the proof. □

D Some concentration inequalities

Theorem D.1 (Chernoff bound, [19]). *Suppose $X_1, \cdots, X_n$ are independent random variables taking values in $\{0,1\}$. Let $\mu = \mathbb{E}[X_1 + \cdots X_n]$. Then for any $\delta > 1$,*

$$\Pr\left[X \ge (1+\delta)\mu\right] \le \exp\left(-\frac{\delta\mu}{3}\right).$$

Theorem D.2 (Hoeffding's inequality, [19]). *Suppose $X_1, \cdots, X_n$ are independent random variables that satisfies $|X_i| \le R$ almost surely. Then the following holds for all $t > 0$:*

$$\Pr\left[\frac{1}{n}\sum_{i=1}^n (X_i - \mathbb{E}[X_i]) > t\right] \le \exp\left(-\frac{2nt^2}{R^2}\right).$$

Theorem D.3 (Matrix Bernstein Inequality, [36]). *Let $\{\mathbf{A}_k\}_{k=1}^T$ be a sequence of independent, random and self-adjoint matrices with dimension p. Suppose*

$$\mathbb{E}[\mathbf{A}_k] = \mathbf{0} \quad \textit{and} \quad \|\mathbf{A}_k\|_2 \le R \ \ a.s.$$

Then for all $t > 0$ the following holds:

$$\Pr\left[\left\|\sum_{k=1}^T \mathbf{A}_k\right\| \ge t\right] \le p\exp\left(-\frac{t^2/2}{\nu^2 + Rt/3}\right),$$

where

$$\nu^2 := \left\|\sum_{k=1}^T \mathbb{E}[\mathbf{A}_k^2]\right\|_2.$$

Theorem D.4 (Weyl's theorem, [33]). *Let* $\mathbf{A}$ *and* $\mathbf{E}$ *be* $p \times p$ *matrices. We then have*

$$\sup_{i=1,\cdots,p} \left|\sigma_i(\mathbf{A}+\mathbf{E}) - \sigma_i(\mathbf{A})\right| \leq \|\mathbf{E}\|_2,$$

where $\sigma_i(\mathbf{A})$ *is the* i*th singular value of* $\mathbf{A}$.

Theorem D.5 (Asymptotics of the incomplete Gamma function, Eq. (2.12) in [16]). *Let* $\Gamma(a,x)$ *be the* incomplete Gamma function *defined as*

$$\Gamma(a,x) = \int_x^\infty t^{a-1}e^{-t}\mathrm{d}t, \quad a,x>0.$$

We then have the following asymptotic approximation:

$$\Gamma(a+1,x) = \frac{e^{-x}x^{a+1}}{x-a}\left[1+O\left(\frac{a}{(x-a)^2}\right)\right].$$