Leveraged volume sampling for linear regression

Michał Dereziński and **Manfred K. Warmuth**
Department of Computer Science
University of California, Santa Cruz
mderezin@berkeley.edu, manfred@ucsc.edu

Daniel Hsu
Computer Science Department
Columbia University, New York
djhsu@cs.columbia.edu

Abstract

Suppose an $n \times d$ design matrix in a linear regression problem is given, but the response for each point is hidden unless explicitly requested. The goal is to sample only a small number $k \ll n$ of the responses, and then produce a weight vector whose sum of squares loss over *all* points is at most $1 + \epsilon$ times the minimum. When k is very small (e.g., $k = d$), jointly sampling diverse subsets of points is crucial. One such method called *volume sampling* has a unique and desirable property that the weight vector it produces is an unbiased estimate of the optimum. It is therefore natural to ask if this method offers the optimal unbiased estimate in terms of the number of responses k needed to achieve a $1 + \epsilon$ loss approximation.

Surprisingly we show that volume sampling can have poor behavior when we require a very accurate approximation – indeed worse than some i.i.d. sampling techniques whose estimates are biased, such as *leverage score sampling*. We then develop a new rescaled variant of volume sampling that produces an unbiased estimate which avoids this bad behavior and has at least as good a tail bound as leverage score sampling: sample size $k = O(d \log d + d/\epsilon)$ suffices to guarantee total loss at most $1 + \epsilon$ times the minimum with high probability. Thus we improve on the best previously known sample size for an unbiased estimator, $k = O(d^2/\epsilon)$.

Our rescaling procedure leads to a new efficient algorithm for volume sampling which is based on a *determinantal rejection sampling* technique with potentially broader applications to determinantal point processes. Other contributions include introducing the combinatorics needed for rescaled volume sampling and developing tail bounds for sums of dependent random matrices which arise in the process.

1 Introduction

Consider a linear regression problem where the input points in $\mathbb{R}^d$ are provided, but the associated response for each point is withheld unless explicitly requested. The goal is to sample the responses for just a small subset of inputs, and then produce a weight vector whose total square loss on all n points is at most $1 + \epsilon$ times that of the optimum.[1] This scenario is relevant in many applications where data points are cheap to obtain but responses are expensive. Surprisingly, with the aid of having all input points available, such multiplicative loss bounds are achievable without any range dependence on the points or responses common in on-line learning [see, e.g., 8].

A natural and intuitive approach to this problem is *volume sampling*, since it prefers "diverse" sets of points that will likely result in a weight vector with low total loss, regardless of what the corresponding responses turn out to be [11]. Volume sampling is closely related to optimal design criteria [18, 26], which are appropriate under statistical models of the responses; here we study a worst-case setting where the algorithm must use randomization to guard itself against worst-case responses.

[1] The total loss being $1 + \epsilon$ times the optimum is the same as the regret being ϵ times the optimum.

32nd Conference on Neural Information Processing Systems (NIPS 2018), Montréal, Canada.

Volume sampling and related determinantal point processes are employed in many machine learning and statistical contexts, including linear regression [11, 13, 26], clustering and matrix approximation [4, 14, 15], summarization and information retrieval [19, 23, 24], and fairness [6, 7]. The availability of fast algorithms for volume sampling [11, 26] has made it an important technique in the algorithmic toolbox alongside i.i.d. leverage score sampling [17] and spectral sparsification [5, 25].

It is therefore surprising that using volume sampling in the context of linear regression, as suggested in previous works [11, 26], may lead to suboptimal performance. We construct an example in which, even after sampling up to half of the responses, the loss of the weight vector from volume sampling is a fixed factor >1 larger than the minimum loss. Indeed, this poor behavior arises because for any sample size $>d$, the marginal probabilities from volume sampling are a mixture of uniform probabilities and leverage score probabilities, and uniform sampling is well-known to be suboptimal when the leverage scores are highly non-uniform.

A possible recourse is to abandon volume sampling in favor of leverage score sampling [17, 33]. However, all i.i.d. sampling methods, including leverage score sampling, suffer from a coupon collector problem that prevents their effective use at small sample sizes [13]. Moreover, the resulting weight vectors are biased (when regarded as estimators for the least squares solution based on all responses). This is a nuisance when averaging multiple solutions (e.g., as produced in distributed settings). In contrast, volume sampling offers multiplicative loss bounds even with sample sizes as small as d and it is the *only* known non-trivial method that gives unbiased weight vectors [11].

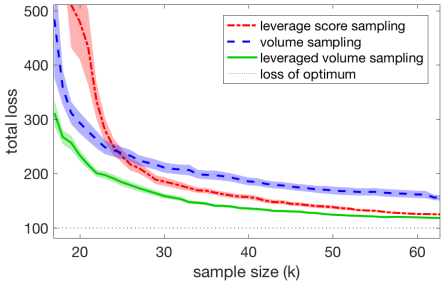

Figure 1: Plots of the total loss for the sampling methods (averaged over 100 runs) versus sample size (shading is standard error) for the libsvm dataset *cpusmall* [9].

We develop a new solution, called *leveraged volume sampling*, that retains the aforementioned benefits of volume sampling while avoiding its flaws. Specifically, we propose a variant of volume sampling based on rescaling the input points to "correct" the resulting marginals. On the algorithmic side, this leads to a new *determinantal rejection sampling* procedure which offers significant computational advantages over existing volume sampling algorithms, while at the same time being strikingly simple to implement. We prove that this new sampling scheme retains the benefits of volume sampling (like unbiasedness) but avoids the bad behavior demonstrated in our lower bound example. Along the way, we prove a new generalization of the Cauchy-Binet formula, which is needed for the rejection sampling denominator. Finally, we develop a new method for proving matrix tail bounds for leveraged volume sampling. Our analysis shows that the unbiased least-squares estimator constructed this way achieves a $1+\epsilon$ approximation factor from a sample of size $O(d \log d + d/\epsilon)$, addressing an open question posed by [11].

Experiments. Figure 1 presents experimental evidence on a benchmark dataset (*cpusmall* from the libsvm collection [9]) that the potential bad behavior of volume sampling proven in our lower bound does occur in practice. Appendix E shows more datasets and a detailed discussion of the experiments. In summary, leveraged volume sampling avoids the bad behavior of standard volume sampling, and performs considerably better than leverage score sampling, especially for small sample sizes k.

Related work. Despite the ubiquity of volume sampling in many contexts already mentioned above, it has only recently been analyzed for linear regression. Focusing on small sample sizes, [11] proved multiplicative bounds for the expected loss of size $k = d$ volume sampling. Because the estimators produced by volume sampling are unbiased, averaging a number of such estimators produced an estimator based on a sample of size $k = O(d^2/\epsilon)$ with expected loss at most $1+\epsilon$ times the optimum. It was shown in [13] that if the responses are assumed to be linear functions of the input points plus white noise, then size $k = O(d/\epsilon)$ volume sampling suffices for obtaining the same expected bounds. These noise assumptions on the response vector are also central to the task of A-optimal design, where volume sampling is a key technique [2, 18, 28, 29]. All of these previous results were concerned with bounds that hold in expectation; it is natural to ask if similar (or better) bounds can also be shown to hold with high probability, without noise assumptions. Concentration bounds for volume sampling and other strong Rayleigh measures were studied in [30], but these results are not sufficient to obtain the tail bounds for volume sampling.

Other techniques applicable to our linear regression problem include leverage score sampling [17] and spectral sparsification [5, 25]. Leverage score sampling is an i.i.d. sampling procedure which achieves tail bounds matching the ones we obtain here for leveraged volume sampling, however it produces biased weight vectors and experimental results (see [13] and Appendix E) show that it has weaker performance for small sample sizes. A different and more elaborate sampling technique based on spectral sparsification [5, 25] was recently shown to be effective for linear regression [10], however this method also does not produce unbiased estimates, which is a primary concern of this paper and desirable in many settings. Unbiasedness seems to require delicate control of the sampling probabilities, which we achieve using determinantal rejection sampling.

Outline and contributions. We set up our task of subsampling for linear regression in the next section and present our lower bound for standard volume sampling. A new variant of rescaled volume sampling is introduced in Section 3. We develop techniques for proving matrix expectation formulas for this variant which show that for any rescaling the weight vector produced for the subproblem is unbiased.

Next, we show that when rescaling with leverage scores, then a new algorithm based on rejection sampling is surprisingly efficient (Section 4): Other than the preprocessing step of computing leverage scores, the runtime does not depend on n (a major improvement over existing volume sampling algorithms). Then, in Section 4.1 we prove multiplicative loss bounds for leveraged volume sampling by establishing two important properties which are hard to prove for joint sampling procedures. We conclude in Section 5 with an open problem and with a discussion of how rescaling with approximate leverage scores gives further time improvements for constructing an unbiased estimator.

2 Volume sampling for linear regression

In this section, we describe our linear regression setting, and review the guarantees that standard volume sampling offers in this context. Then, we present a surprising lower bound which shows that under worst-case data, this method can exhibit undesirable behavior.

2.1 Setting

Suppose the learner is given n input vectors $\mathbf{x}_1, \ldots, \mathbf{x}_n \in \mathbb{R}^d$, which are arranged as the rows of an $n \times d$ input matrix $\mathbf{X}$. Each input vector $\mathbf{x}_i$ has an associated response variable $y_i \in \mathbb{R}$ from the response vector $\mathbf{y} \in \mathbb{R}^n$. The goal of the learner is to find a weight vector $\mathbf{w} \in \mathbb{R}^d$ that minimizes the square loss:

$$\mathbf{w}^* \overset{def}{=} \operatorname*{argmin}_{\mathbf{w}\in\mathbb{R}^d} L(\mathbf{w}), \quad \text{where } L(\mathbf{w}) \overset{def}{=} \sum_{i=1}^{n} (\mathbf{x}_i^\top \mathbf{w} - y_i)^2 = \|\mathbf{X}\mathbf{w} - \mathbf{y}\|^2.$$

Given both matrix $\mathbf{X}$ and vector $\mathbf{y}$, the least squares solution can be directly computed as $\mathbf{w}^* = \mathbf{X}^+\mathbf{y}$, where $\mathbf{X}^+$ is the pseudo-inverse. Throughout the paper we assume w.l.o.g. that $\mathbf{X}$ has (full) rank d.[2]

In our setting, the learner is initially given the entire input matrix $\mathbf{X}$, while response vector $\mathbf{y}$ remains hidden. The learner is then allowed to select a subset S of row indices in $[n] = \{1, \ldots, n\}$ for which the corresponding responses y_i are revealed. The learner next constructs an estimate $\widehat{\mathbf{w}}$ of $\mathbf{w}^*$ using matrix $\mathbf{X}$ and the partial vector of observed responses. Finally, the learner is evaluated by the loss over all rows of $\mathbf{X}$ (including the ones with unobserved responses), and the goal is to obtain a multiplicative loss bound, i.e., that for some $\epsilon > 0$,

$$L(\widehat{\mathbf{w}}) \leq (1+\epsilon)\, L(\mathbf{w}^*).$$

2.2 Standard volume sampling

Given $\mathbf{X} \in \mathbb{R}^{n\times d}$ and a size $k \geq d$, standard volume sampling jointly chooses a set S of k indices in $[n]$ with probability

$$\Pr(S) = \frac{\det(\mathbf{X}_S^\top \mathbf{X}_S)}{\binom{n-d}{k-d} \det(\mathbf{X}^\top \mathbf{X})},$$

[2] Otherwise just reduce $\mathbf{X}$ to a subset of independent columns. Also assume $\mathbf{X}$ has no rows of all zeros (every weight vector has the same loss on such rows, so they can be removed).

where $\mathbf{X}_S$ is the submatrix of the rows from $\mathbf{X}$ indexed by the set S. The learner then obtains the responses y_i, for $i \in S$, and uses the optimum solution $\mathbf{w}_S^* = (\mathbf{X}_S)^+ \mathbf{y}_S$ for the subproblem $(\mathbf{X}_S, \mathbf{y}_S)$ as its weight vector. The sampling procedure can be performed using *reverse iterative sampling* (shown on the right), which, if carefully implemented, takes $O(nd^2)$ time (see [11, 13]).

Reverse iterative sampling

VolumeSample($\mathbf{X}$, k):
  $S \leftarrow [n]$
  while $|S| > k$
    $\forall_{i \in S}: q_i \leftarrow \frac{\det(\mathbf{X}_{S\setminus i}^\top \mathbf{X}_{S\setminus i})}{\det(\mathbf{X}_S^\top \mathbf{X}_S)}$
    Sample $i \propto q_i$ out of S
    $S \leftarrow S \setminus \{i\}$
  end
return S

The key property (unique to volume sampling) is that the subsampled estimator $\mathbf{w}_S^*$ is unbiased, i.e.

$$\mathbb{E}[\mathbf{w}_S^*] = \mathbf{w}^*, \quad \text{where} \quad \mathbf{w}^* = \operatorname*{argmin}_{\mathbf{w}} L(\mathbf{w}).$$

As discussed in [11], this property has important practical implications in distributed settings: Mixtures of unbiased estimators remain unbiased (and can conveniently be used to reduce variance). Also if the rows of $\mathbf{X}$ are in general position, then for volume sampling

$$\mathbb{E}\big[(\mathbf{X}_S^\top \mathbf{X}_S)^{-1}\big] = \frac{n-d+1}{k-d+1}\,(\mathbf{X}^\top \mathbf{X})^{-1}. \tag{1}$$

This is important because in A-optimal design bounding $\operatorname{tr}((\mathbf{X}_S^\top \mathbf{X}_S)^{-1})$ is the main concern. Given these direct connections of volume sampling to linear regression, it is natural to ask whether this distribution achieves a loss bound of $(1+\epsilon)$ times the optimum for small sample sizes k.

2.3 Lower bound for standard volume sampling

We show that standard volume sampling cannot guarantee $1+\epsilon$ multiplicative loss bounds on some instances, unless over half of the rows are chosen to be in the subsample.

Theorem 1 *Let $(\mathbf{X}, \mathbf{y})$ be an $n \times d$ least squares problem, such that*

$$\mathbf{X} = \begin{pmatrix} \mathbf{I}_{d\times d} \\ \hline \gamma\,\mathbf{I}_{d\times d} \\ \hline \vdots \\ \hline \gamma\,\mathbf{I}_{d\times d} \end{pmatrix}, \quad \mathbf{y} = \begin{pmatrix} \mathbf{1}_d \\ \hline \mathbf{0}_d \\ \hline \vdots \\ \hline \mathbf{0}_d \end{pmatrix}, \qquad \textit{where} \quad \gamma > 0.$$

Let $\mathbf{w}_S^* = (\mathbf{X}_S)^+ \mathbf{y}_S$ be obtained from size k volume sampling for $(\mathbf{X}, \mathbf{y})$. Then,

$$\lim_{\gamma \to 0} \frac{\mathbb{E}[L(\mathbf{w}_S^*)]}{L(\mathbf{w}^*)} \;\geq\; 1 + \frac{n-k}{n-d}, \tag{2}$$

and there is a $\gamma > 0$ such that for any $k \leq \frac{n}{2}$,

$$\Pr\left(L(\mathbf{w}_S^*) \geq \Big(1+\frac{1}{2}\Big) L(\mathbf{w}^*)\right) > \frac{1}{4}. \tag{3}$$

Proof In Appendix A we show (2), and that for the chosen $(\mathbf{X}, \mathbf{y})$ we have $L(\mathbf{w}^*) = \sum_{i=1}^d (1-l_i)$ (see (8)), where $l_i = \mathbf{x}_i^\top (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{x}_i$ is the i-th leverage score of $\mathbf{X}$. Here, we show (3). The marginal probability of the i-th row under volume sampling (as given by [12]) is

$$\Pr(i \in S) = \theta\, l_i + (1-\theta)\, 1 = 1 - \theta\,(1-l_i), \;\; \text{where } \theta = \frac{n-k}{n-d}. \tag{4}$$

Next, we bound the probability that all of the first d input vectors were selected by volume sampling:

$$\Pr\big([d] \subseteq S\big) \overset{(*)}{\leq} \prod_{i=1}^d \Pr(i \in S) = \prod_{i=1}^d \Big(1 - \frac{n-k}{n-d}\,(1-l_i)\Big) \leq \exp\Big(-\frac{n-k}{n-d}\,\overbrace{L(\mathbf{w}^*)}^{\sum_{i=1}^d (1-l_i)}\Big),$$

where $(*)$ follows from negative associativity of volume sampling (see [26]). If for some $i \in [d]$ we have $i \notin S$, then $L(\mathbf{w}_S^*) \geq 1$. So for γ such that $L(\mathbf{w}^*) = \frac{2}{3}$ and any $k \leq \frac{n}{2}$:

$$\Pr\left(L(\mathbf{w}_S^*) \geq \Big(1+\frac{1}{2}\Big)\overbrace{L(\mathbf{w}^*)}^{2/3}\right) \geq 1 - \exp\Big(-\frac{n-k}{n-d}\cdot\frac{2}{3}\Big) \geq 1 - \exp\Big(-\frac{1}{2}\cdot\frac{2}{3}\Big) > \frac{1}{4}. \qquad \blacksquare$$

Note that this lower bound only makes use of the negative associativity of volume sampling and the form of the marginals. However the tail bounds we prove in Section 4.1 rely on more subtle properties of volume sampling. We begin by creating a variant of volume sampling with rescaled marginals.

3 Rescaled volume sampling

Given any size $k \geq d$, our goal is to jointly sample k row indices $\pi_1, \ldots, \pi_k$ with replacement (instead of a *subset* S of $[n]$ of size k, we get a *sequence* $\pi \in [n]^k$). The second difference to standard volume sampling is that we rescale the i-th row (and response) by $\frac{1}{\sqrt{q_i}}$, where $q = (q_1, ..., q_n)$ is any discrete distribution over the set of row indices $[n]$, such that $\sum_{i=1}^n q_i = 1$ and $q_i > 0$ for all $i \in [n]$. We now define q-rescaled size k volume sampling as a joint sampling distribution over $\pi \in [n]^k$, s.t.

$$q\text{-rescaled size } k \text{ volume sampling:} \qquad \Pr(\pi) \;\sim\; \det\Big(\sum_{i=1}^k \frac{1}{q_{\pi_i}} \mathbf{x}_{\pi_i} \mathbf{x}_{\pi_i}^\top\Big) \;\; \prod_{i=1}^k q_{\pi_i}. \tag{5}$$

Using the following rescaling matrix $\mathbf{Q}_\pi \overset{def}{=} \sum_{i=1}^{|\pi|} \frac{1}{q_{\pi_i}} \mathbf{e}_{\pi_i} \mathbf{e}_{\pi_i}^\top \in \mathbb{R}^{n \times n}$, we rewrite the determinant as $\det(\mathbf{X}^\top \mathbf{Q}_\pi \mathbf{X})$. As in standard volume sampling, the normalization factor in rescaled volume sampling can be given in a closed form through a novel extension of the Cauchy-Binet formula (proof in Appendix B.1).

Proposition 2 *For any $\mathbf{X} \in \mathbb{R}^{n \times d}$, $k \geq d$ and $q_1, \ldots, q_n > 0$, such that $\sum_{i=1}^n q_i = 1$, we have*

$$\sum_{\pi \in [n]^k} \det(\mathbf{X}^\top \mathbf{Q}_\pi \mathbf{X}) \prod_{i=1}^k q_{\pi_i} = k(k-1)\cdots(k-d+1)\, \det(\mathbf{X}^\top \mathbf{X}).$$

Given a matrix $\mathbf{X} \in \mathbb{R}^{n \times d}$, vector $\mathbf{y} \in \mathbb{R}^n$ and a sequence $\pi \in [n]^k$, we are interested in a least-squares problem $(\mathbf{Q}_\pi^{1/2}\mathbf{X}, \mathbf{Q}_\pi^{1/2}\mathbf{y})$, which selects instances indexed by π, and rescales each of them by the corresponding $1/\sqrt{q_i}$. This leads to a natural subsampled least squares estimator

$$\mathbf{w}_\pi^* = \operatorname*{argmin}_{\mathbf{w}} \sum_{i=1}^k \frac{1}{q_{\pi_i}} \big(\mathbf{x}_{\pi_i}^\top \mathbf{w} - y_{\pi_i}\big)^2 = (\mathbf{Q}_\pi^{1/2}\mathbf{X})^+ \mathbf{Q}_\pi^{1/2}\mathbf{y}.$$

The key property of standard volume sampling is that the subsampled least-squares estimator is unbiased. Surprisingly this property is retained for any q-rescaled volume sampling (proof in Section 3.1). As we shall see, this will give us great leeway for choosing q to optimize our algorithms.

Theorem 3 *Given a full rank $\mathbf{X} \in \mathbb{R}^{n \times d}$ and a response vector $\mathbf{y} \in \mathbb{R}^n$, for any q as above, if π is sampled according to* (5)*, then*

$$\mathbb{E}[\mathbf{w}_\pi^*] = \mathbf{w}^*, \quad \textit{where} \quad \mathbf{w}^* = \operatorname*{argmin}_{\mathbf{w}} \|\mathbf{X}\mathbf{w} - \mathbf{y}\|^2.$$

The matrix expectation equation (1) for standard volume sampling (discussed in Section 2) has a natural extension to any rescaled volume sampling, but now the equality turns into an inequality (proof in Appendix B.2):

Theorem 4 *Given a full rank $\mathbf{X} \in \mathbb{R}^{n \times d}$ and any q as above, if π is sampled according to* (5)*, then*

$$\mathbb{E}\big[(\mathbf{X}^\top \mathbf{Q}_\pi \mathbf{X})^{-1}\big] \preceq \frac{1}{k-d+1} (\mathbf{X}^\top \mathbf{X})^{-1}.$$

3.1 Proof of Theorem 3

We show that the least-squares estimator $\mathbf{w}_\pi^* = (\mathbf{Q}_\pi^{1/2}\mathbf{X})^+ \mathbf{Q}_\pi^{1/2}\mathbf{y}$ produced from any q-rescaled volume sampling is unbiased, illustrating a proof technique which is also useful for showing Theorem 4, as well as Propositions 2 and 5. The key idea is to apply the pseudo-inverse expectation formula for standard volume sampling (see e.g., [11]) first on the subsampled estimator $\mathbf{w}_\pi^*$, and then again on the full estimator $\mathbf{w}^*$. In the first step, this formula states:

$$\overbrace{(\mathbf{Q}_\pi^{1/2}\mathbf{X})^+ \mathbf{Q}_\pi^{1/2}\mathbf{y}}^{\mathbf{w}_\pi^*} = \sum_{S \in \binom{[k]}{d}} \frac{\det(\mathbf{X}^\top \mathbf{Q}_{\pi_S} \mathbf{X})}{\det(\mathbf{X}^\top \mathbf{Q}_\pi \mathbf{X})} \overbrace{(\mathbf{Q}_{\pi_S}^{1/2}\mathbf{X})^+ \mathbf{Q}_{\pi_S}^{1/2}\mathbf{y}}^{\mathbf{w}_{\pi_S}^*},$$

where $\binom{[k]}{d} \overset{def}{=} \{S \subseteq \{1,\dots,k\} : |S| = d\}$ and π_S denotes a subsequence of π indexed by the elements of set S. Note that since S is of size d, we can decompose the determinant:

$$\det(\mathbf{X}^\top \mathbf{Q}_{\pi_S} \mathbf{X}) = \det(\mathbf{X}_{\pi_S})^2 \prod_{i \in S} \frac{1}{q_{\pi_i}}.$$

Whenever this determinant is non-zero, $\mathbf{w}^*_{\pi_S}$ is the exact solution of a system of d linear equations:

$$\frac{1}{\sqrt{q_{\pi_i}}} \mathbf{x}_{\pi_i}^\top \mathbf{w} = \frac{1}{\sqrt{q_{\pi_i}}} y_{\pi_i}, \qquad \text{for} \quad i \in S.$$

Thus, the rescaling of each equation by $\frac{1}{\sqrt{q_{\pi_i}}}$ cancels out, and we can simply write $\mathbf{w}^*_{\pi_S} = (\mathbf{X}_{\pi_S})^+ \mathbf{y}_{\pi_S}$. (Note that this is not the case for sets larger than d whenever the optimum solution incurs positive loss.) We now proceed with summing over all $\pi \in [n]^k$. Following Proposition 2, we define the normalization constant as $Z = d!\binom{k}{d} \det(\mathbf{X}^\top \mathbf{X})$, and obtain:

$$\begin{aligned}
Z\, \mathbb{E}[\mathbf{w}^*_\pi] &= \sum_{\pi \in [n]^k} \left(\prod_{i=1}^k q_{\pi_i} \right) \det(\mathbf{X}^\top \mathbf{Q}_\pi \mathbf{X})\, \mathbf{w}^*_\pi = \sum_{\pi \in [n]^k} \sum_{S \in \binom{[k]}{d}} \left(\prod_{i \in [k] \setminus S} q_{\pi_i} \right) \det(\mathbf{X}_{\pi_S})^2 (\mathbf{X}_{\pi_S})^+ \mathbf{y}_{\pi_S} \\
&\overset{(1)}{=} \binom{k}{d} \sum_{\bar\pi \in [n]^d} \det(\mathbf{X}_{\bar\pi})^2 (\mathbf{X}_{\bar\pi})^+ \mathbf{y}_{\bar\pi} \sum_{\tilde\pi \in [n]^{k-d}} \prod_{i=1}^{k-d} q_{\tilde\pi_i} \\
&\overset{(2)}{=} \binom{k}{d} d! \sum_{S \in \binom{[n]}{d}} \det(\mathbf{X}_S)^2 (\mathbf{X}_S)^+ \mathbf{y}_S \left(\sum_{i=1}^n q_i \right)^{k-d} \overset{(3)}{=} \overbrace{\binom{k}{d} d! \det(\mathbf{X}^\top \mathbf{X})}^{Z}\, \mathbf{w}^*.
\end{aligned}$$

Note that in (1) we separate π into two parts, $\bar\pi$ and $\tilde\pi$ (respectively, for subsets S and $[k]\setminus S$), and sum over them separately. The binomial coefficient $\binom{k}{d}$ counts the number of ways that S can be "placed into" the sequence π. In (2) we observe that whenever $\bar\pi$ has repetitions, determinant $\det(\mathbf{X}_{\bar\pi})$ is zero, so we can switch to summing over sets. Finally, (3) again uses the standard size d volume sampling unbiasedness formula, now for the least-squares task $(\mathbf{X}, \mathbf{y})$, and the fact that q_i's sum to 1.

4 Leveraged volume sampling: a natural rescaling

Rescaled volume sampling can be viewed as selecting a sequence π of k rank-1 matrices from the covariance matrix $\mathbf{X}^\top \mathbf{X} = \sum_{i=1}^n \mathbf{x}_i \mathbf{x}_i^\top$. If $\pi_1, \dots, \pi_k$ are sampled i.i.d. from q, i.e., $\Pr(\pi) = \prod_{i=1}^k q_{\pi_i}$, then matrix $\frac{1}{k} \mathbf{X}^\top \mathbf{Q}_\pi \mathbf{X}$ is an unbiased estimator of the covariance matrix because $\mathbb{E}[q_{\pi_i}^{-1} \mathbf{x}_{\pi_i} \mathbf{x}_{\pi_i}^\top] = \mathbf{X}^\top \mathbf{X}$. In rescaled volume sampling (5), $\Pr(\pi) \sim \big(\prod_{i=1}^k q_{\pi_i}\big) \frac{\det(\mathbf{X}^\top \mathbf{Q}_\pi \mathbf{X})}{\det(\mathbf{X}^\top \mathbf{X})}$, and the latter volume ratio introduces a bias to that estimator. However, we show that this bias vanishes when q is exactly proportional to the leverage scores (proof in Appendix B.3).

Determinantal rejection sampling

Input: $\mathbf{X} \in \mathbb{R}^{n \times d}, q = (\frac{l_1}{d}, \dots, \frac{l_n}{d}), k \geq d$
$s \leftarrow \max\{k,\, 4d^2\}$
repeat
  Sample $\pi_1, \dots, \pi_s$ i.i.d. $\sim (q_1, \dots, q_n)$
  Sample *Accept* $\sim$ Bernoulli$\Big(\frac{\det(\frac{1}{s}\mathbf{X}^\top \mathbf{Q}_\pi \mathbf{X})}{\det(\mathbf{X}^\top \mathbf{X})}\Big)$
until *Accept* = true
$S \leftarrow$ VolumeSample$\big((\mathbf{Q}^{1/2}_{[1..n]} \mathbf{X})_\pi, k\big)$
return π_S

Proposition 5 *For any q and $\mathbf{X}$ as before, if $\pi \in [n]^k$ is sampled according to* (5)*, then*

$$\mathbb{E}[\mathbf{Q}_\pi] = (k-d)\, \mathbf{I} + \operatorname{diag}\Big(\frac{l_1}{q_1}, \dots, \frac{l_n}{q_n}\Big), \quad \textit{where} \quad l_i \overset{def}{=} \mathbf{x}_i^\top (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{x}_i.$$

In particular, $\mathbb{E}[\frac{1}{k} \mathbf{X}^\top \mathbf{Q}_\pi \mathbf{X}] = \mathbf{X}^\top \mathbb{E}[\frac{1}{k} \mathbf{Q}_\pi] \mathbf{X} = \mathbf{X}^\top \mathbf{X}$ if and only if $q_i = \frac{l_i}{d} > 0$ for all $i \in [n]$.

This special rescaling, which we call *leveraged volume sampling*, has other remarkable properties. Most importantly, it leads to a simple and efficient algorithm we call *determinantal rejection sampling*: Repeatedly sample $O(d^2)$ indices $\pi_1, \dots, \pi_s$ i.i.d. from $q = (\frac{l_1}{d}, \dots, \frac{l_n}{d})$, and accept the sample with probability proportional to its volume ratio. Having obtained a sample, we can further reduce its size via reverse iterative sampling. We show next that this procedure not only returns a q-rescaled volume sample, but also exploiting the fact that q is proportional to the leverage scores, it requires (surprisingly) only a constant number of iterations of rejection sampling with high probability.

Theorem 6 *Given the leverage score distribution $q = (\frac{l_1}{d}, \ldots, \frac{l_n}{d})$ and the determinant $\det(\mathbf{X}^\top\mathbf{X})$ for matrix $\mathbf{X} \in \mathbb{R}^{n\times d}$, determinantal rejection sampling returns sequence π_S distributed according to leveraged volume sampling, and w.p. at least $1-\delta$ finishes in time $O((d^2+k)d^2\ln(\frac{1}{\delta}))$.*

Proof We use a composition property of rescaled volume sampling (proof in Appendix B.4):

Lemma 7 *Consider the following sampling procedure, for $s > k$:*

$$\pi \overset{s}{\sim} \mathbf{X} \qquad \textit{(q-rescaled size s volume sampling)},$$

$$S \overset{k}{\sim} \begin{pmatrix} \frac{1}{\sqrt{q_{\pi_1}}}\mathbf{x}_{\pi_1}^\top \\ \cdots \\ \frac{1}{\sqrt{q_{\pi_s}}}\mathbf{x}_{\pi_s}^\top \end{pmatrix} = \big(\mathbf{Q}_{[1..n]}^{1/2}\mathbf{X}\big)_\pi \qquad \textit{(standard size k volume sampling)}.$$

Then π_S is distributed according to q-rescaled size k volume sampling from $\mathbf{X}$.

First, we show that the rejection sampling probability in line 5 of the algorithm is bounded by 1:

$$\frac{\det(\frac{1}{s}\mathbf{X}^\top\mathbf{Q}_\pi\mathbf{X})}{\det(\mathbf{X}^\top\mathbf{X})} = \det\Big(\frac{1}{s}\mathbf{X}^\top\mathbf{Q}_\pi\mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}\Big) \overset{(*)}{\leq} \Big(\frac{1}{d}\mathrm{tr}\Big(\frac{1}{s}\mathbf{X}^\top\mathbf{Q}_\pi\mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}\Big)\Big)^d$$
$$= \Big(\frac{1}{ds}\mathrm{tr}\big(\mathbf{Q}_\pi\mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\big)\Big)^d = \Big(\frac{1}{ds}\sum_{i=1}^s \frac{d}{l_i}\mathbf{x}_i^\top(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{x}_i\Big)^d = 1,$$

where $(*)$ follows from the geometric-arithmetic mean inequality for the eigenvalues of the underlying matrix. This shows that sequence π is drawn according to q-rescaled volume sampling of size s. Now, Lemma 7 implies correctness of the algorithm. Next, we use Proposition 2 to compute the expected value of acceptance probability from line 5 under the i.i.d. sampling of line 4:

$$\sum_{\pi\in[n]^s}\Big(\prod_{i=1}^s q_{\pi_i}\Big)\frac{\det(\frac{1}{s}\mathbf{X}^\top\mathbf{Q}_\pi\mathbf{X})}{\det(\mathbf{X}^\top\mathbf{X})} = \frac{s(s-1)\ldots(s-d+1)}{s^d} \geq \Big(1-\frac{d}{s}\Big)^d \geq 1-\frac{d^2}{s} \geq \frac{3}{4},$$

where we also used Bernoulli's inequality and the fact that $s \geq 4d^2$ (see line 2). Since the expected value of the acceptance probability is at least $\frac{3}{4}$, an easy application of Markov's inequality shows that at each trial there is at least a 50% chance of it being above $\frac{1}{2}$. So, the probability of at least r trials occurring is less than $(1-\frac{1}{4})^r$. Note that the computational cost of one trial is no more than the cost of SVD decomposition of matrix $\mathbf{X}^\top\mathbf{Q}_\pi\mathbf{X}$ (for computing the determinant), which is $O(sd^2)$. The cost of reverse iterative sampling (line 7) is also $O(sd^2)$ with high probability (as shown by [13]). Thus, the overall runtime is $O((d^2+k)d^2r)$, where $r \leq \ln(\frac{1}{\delta})/\ln(\frac{4}{3})$ w.p. at least $1-\delta$. ■

4.1 Tail bounds for leveraged volume sampling

An analysis of leverage score sampling, essentially following [33, Section 2] which in turn draws from [31], highlights two basic sufficient conditions on the (random) subsampling matrix $\mathbf{Q}_\pi$ that lead to multiplicative tail bounds for $L(\mathbf{w}_\pi^*)$.

It is convenient to shift to an orthogonalization of the linear regression task $(\mathbf{X}, \mathbf{y})$ by replacing matrix $\mathbf{X}$ with a matrix $\mathbf{U} = \mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1/2} \in \mathbb{R}^{n\times d}$. It is easy to check that the columns of $\mathbf{U}$ have unit length and are orthogonal, i.e., $\mathbf{U}^\top\mathbf{U} = \mathbf{I}$. Now, $\mathbf{v}^* = \mathbf{U}^\top\mathbf{y}$ is the least-squares solution for the orthogonal problem $(\mathbf{U}, \mathbf{y})$ and prediction vector $\mathbf{U}\mathbf{v}^* = \mathbf{U}\mathbf{U}^\top\mathbf{y}$ for $(\mathbf{U}, \mathbf{y})$ is the same as the prediction vector $\mathbf{X}\mathbf{w}^* = \mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{y}$ for the original problem $(\mathbf{X}, \mathbf{y})$. The same property holds for the subsampled estimators, i.e., $\mathbf{U}\mathbf{v}_\pi^* = \mathbf{X}\mathbf{w}_\pi^*$, where $\mathbf{v}_\pi^* = (\mathbf{Q}_\pi^{1/2}\mathbf{U})^+\mathbf{Q}_\pi^{1/2}\mathbf{y}$. Volume sampling probabilities are also preserved under this transformation, so w.l.o.g. we can work with the orthogonal problem. Now $L(\mathbf{v}_\pi^*)$ can be rewritten as

$$L(\mathbf{v}_\pi^*) = \|\mathbf{U}\mathbf{v}_\pi^* - \mathbf{y}\|^2 \overset{(1)}{=} \|\mathbf{U}\mathbf{v}^* - \mathbf{y}\|^2 + \|\mathbf{U}(\mathbf{v}_\pi^* - \mathbf{v}^*)\|^2 \overset{(2)}{=} L(\mathbf{v}^*) + \|\mathbf{v}_\pi^* - \mathbf{v}^*\|^2, \qquad (6)$$

where (1) follows via Pythagorean theorem from the fact that $\mathbf{U}(\mathbf{v}_\pi^* - \mathbf{v}^*)$ lies in the column span of $\mathbf{U}$ and the residual vector $\mathbf{r} = \mathbf{U}\mathbf{v}^* - \mathbf{y}$ is orthogonal to all columns of $\mathbf{U}$, and (2) follows from $\mathbf{U}^\top\mathbf{U} = \mathbf{I}$. By the definition of $\mathbf{v}_\pi^*$, we can write $\|\mathbf{v}_\pi^* - \mathbf{v}^*\|$ as follows:

$$\|\mathbf{v}_\pi^* - \mathbf{v}^*\| = \|(\mathbf{U}^\top\mathbf{Q}_\pi\mathbf{U})^{-1}\ \ \mathbf{U}^\top\mathbf{Q}_\pi(\mathbf{y} - \mathbf{U}\mathbf{v}^*)\| \leq \|(\underset{d\times d}{\mathbf{U}^\top\mathbf{Q}_\pi\mathbf{U}})^{-1}\|\ \|\underset{d\times 1}{\mathbf{U}^\top\mathbf{Q}_\pi\,\mathbf{r}}\|, \qquad (7)$$

where $\|\mathbf{A}\|$ denotes the matrix 2-norm (i.e., the largest singular value) of $\mathbf{A}$; when $\mathbf{A}$ is a vector, then $\|\mathbf{A}\|$ is its Euclidean norm. This breaks our task down to showing two key properties:

1. *Matrix multiplication:* Upper bounding the Euclidean norm $\|\mathbf{U}^\top \mathbf{Q}_\pi \mathbf{r}\|$,
2. *Subspace embedding:* Upper bounding the matrix 2-norm $\|(\mathbf{U}^\top \mathbf{Q}_\pi \mathbf{U})^{-1}\|$.

We start with a theorem that implies strong guarantees for approximate matrix multiplication with leveraged volume sampling. Unlike with i.i.d. sampling, this result requires controlling the pairwise dependence between indices selected under rescaled volume sampling. Its proof is an interesting application of a classical Hadamard matrix product inequality from [3] (Proof in Appendix C).

Theorem 8 *Let $\mathbf{U} \in \mathbb{R}^{n\times d}$ be a matrix s.t. $\mathbf{U}^\top\mathbf{U} = \mathbf{I}$. If sequence $\pi \in [n]^k$ is selected using leveraged volume sampling of size $k \geq \frac{2d}{\epsilon}$, then for any $\mathbf{r} \in \mathbb{R}^n$,*

$$\mathbb{E}\left[\left\|\frac{1}{k}\mathbf{U}^\top\mathbf{Q}_\pi\mathbf{r} - \mathbf{U}^\top\mathbf{r}\right\|^2\right] \leq \epsilon\,\|\mathbf{r}\|^2.$$

Next, we turn to the subspace embedding property. The following result is remarkable because standard matrix tail bounds used to prove this property for leverage score sampling are not applicable to volume sampling. In fact, obtaining matrix Chernoff bounds for negatively associated joint distributions like volume sampling is an active area of research, as discussed in [21]. We address this challenge by defining a coupling procedure for volume sampling and uniform sampling without replacement, which leads to a curious reduction argument described in Appendix D.

Theorem 9 *Let $\mathbf{U} \in \mathbb{R}^{n\times d}$ be a matrix s.t. $\mathbf{U}^\top\mathbf{U} = \mathbf{I}$. There is an absolute constant C, s.t. if sequence $\pi \in [n]^k$ is selected using leveraged volume sampling of size $k \geq C\,d\ln(\frac{d}{\delta})$, then*

$$\Pr\left(\lambda_{\min}\Big(\frac{1}{k}\mathbf{U}^\top\mathbf{Q}_\pi\mathbf{U}\Big) \leq \frac{1}{8}\right) \leq \delta.$$

Theorems 8 and 9 imply that the unbiased estimator $\mathbf{w}^*_\pi$ produced from leveraged volume sampling achieves multiplicative tail bounds with sample size $k = O(d\log d + d/\epsilon)$.

Corollary 10 *Let $\mathbf{X} \in \mathbb{R}^{n\times d}$ be a full rank matrix. There is an absolute constant C, s.t. if sequence $\pi \in [n]^k$ is selected using leveraged volume sampling of size $k \geq C\left(d\ln(\frac{d}{\delta}) + \frac{d}{\epsilon\delta}\right)$, then for estimator*

$$\mathbf{w}^*_\pi = \operatorname*{argmin}_{\mathbf{w}} \|\mathbf{Q}_\pi^{1/2}(\mathbf{X}\mathbf{w} - \mathbf{y})\|^2,$$

we have $L(\mathbf{w}^*_\pi) \leq (1+\epsilon)\,L(\mathbf{w}^*)$ with probability at least $1-\delta$.

Proof Let $\mathbf{U} = \mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1/2}$. Combining Theorem 8 with Markov's inequality, we have that for large enough C, $\|\mathbf{U}^\top\mathbf{Q}_\pi\,\mathbf{r}\|^2 \leq \epsilon\,\frac{k^2}{8^2}\|\mathbf{r}\|^2$ w.h.p., where $\mathbf{r} = \mathbf{y} - \mathbf{U}\mathbf{v}^*$. Finally following (6) and (7) above, we have that w.h.p.

$$L(\mathbf{w}^*_\pi) \leq L(\mathbf{w}^*) + \|(\mathbf{U}^\top\mathbf{Q}_\pi\mathbf{U})^{-1}\|^2\,\|\mathbf{U}^\top\mathbf{Q}_\pi\,\mathbf{r}\|^2 \leq L(\mathbf{w}^*) + \frac{8^2}{k^2}\,\epsilon\frac{k^2}{8^2}\,\|\mathbf{r}\|^2 = (1+\epsilon)\,L(\mathbf{w}^*). \quad \blacksquare$$

5 Conclusion

We developed a new variant of volume sampling which produces the first known unbiased subsampled least-squares estimator with strong multiplicative loss bounds. In the process, we proved a novel extension of the Cauchy-Binet formula, as well as other fundamental combinatorial equalities. Moreover, we proposed an efficient algorithm called determinantal rejection sampling, which is to our knowledge the first joint determinantal sampling procedure that (after an initial $O(nd^2)$ preprocessing step for computing leverage scores) produces its k samples in time $\widetilde{O}(d^2+k)d^2)$, independent of the data size n. When n is very large, the preprocessing time can be reduced to $\widetilde{O}(nd + d^5)$ by rescaling with sufficiently accurate approximations of the leverage scores. Surprisingly the estimator stays unbiased and the loss bound still holds with only slightly revised constants. For the sake of clarity we presented the algorithm based on rescaling with exact leverage scores in the main body of the paper. However we outline the changes needed when using approximate leverage scores in Appendix F.

In this paper we focused on tail bounds. However we conjecture that there are also volume sampling based unbiased estimators achieving expected loss bounds $\mathbb{E}[L(\mathbf{w}^*_\pi)] \leq (1+\epsilon)L(\mathbf{w}^*)$ with size $O(\frac{d}{\epsilon})$.

Acknowledgements

Michał Dereziński and Manfred K. Warmuth were supported by NSF grant IIS-1619271. Daniel Hsu was supported by NSF grant CCF-1740833.

References

[1] Nir Ailon and Bernard Chazelle. The fast Johnson–Lindenstrauss transform and approximate nearest neighbors. *SIAM Journal on computing*, 39(1):302–322, 2009.

[2] Zeyuan Allen-Zhu, Yuanzhi Li, Aarti Singh, and Yining Wang. Near-optimal design of experiments via regret minimization. In Doina Precup and Yee Whye Teh, editors, *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 126–135, International Convention Centre, Sydney, Australia, 2017.

[3] T Ando, Roger A. Horn, and Charles R. Johnson. The singular values of a Hadamard product: A basic inequality. *Journal of Linear and Multilinear Algebra*, 21(4):345–365, 1987.

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

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

[6] L Elisa Celis, Amit Deshpande, Tarun Kathuria, and Nisheeth K Vishnoi. How to be fair and diverse? *arXiv:1610.07183*, October 2016.

[7] L Elisa Celis, Vijay Keswani, Damian Straszak, Amit Deshpande, Tarun Kathuria, and Nisheeth K Vishnoi. Fair and diverse dpp-based data summarization. *arXiv:1802.04023*, February 2018.

[8] N. Cesa-Bianchi, P. M. Long, and M. K. Warmuth. Worst-case quadratic loss bounds for on-line prediction of linear functions by gradient descent. *IEEE Transactions on Neural Networks*, 7(3):604–619, 1996. Earlier version in 6th COLT, 1993.

[9] Chih-Chung Chang and Chih-Jen Lin. LIBSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*, 2:27:1–27:27, 2011. Software available at `http://www.csie.ntu.edu.tw/~cjlin/libsvm`.

[10] Xue Chen and Eric Price. Condition number-free query and active learning of linear families. *CoRR*, abs/1711.10051, 2017.

[11] Michał Dereziński and Manfred K Warmuth. Unbiased estimates for linear regression via volume sampling. In *Advances in Neural Information Processing Systems 30*, pages 3087–3096, Long Beach, CA, USA, December 2017.

[12] Michał Dereziński and Manfred K. Warmuth. Reverse iterative volume sampling for linear regression. *Journal of Machine Learning Research*, 19(23):1–39, 2018.

[13] Michał Dereziński and Manfred K. Warmuth. Subsampling for ridge regression via regularized volume sampling. In *Proceedings of the 21st International Conference on Artificial Intelligence and Statistics*, 2018.

[14] Amit Deshpande and Luis Rademacher. Efficient volume sampling for row/column subset selection. In *Proceedings of the 2010 IEEE 51st Annual Symposium on Foundations of Computer Science*, FOCS '10, pages 329–338, Washington, DC, USA, 2010.

[15] Amit Deshpande, Luis Rademacher, Santosh Vempala, and Grant Wang. Matrix approximation and projective clustering via volume sampling. In *Proceedings of the Seventeenth Annual ACM-SIAM Symposium on Discrete Algorithm*, SODA '06, pages 1117–1126, Philadelphia, PA, USA, 2006.

[16] Petros Drineas, Malik Magdon-Ismail, Michael W. Mahoney, and David P. Woodruff. Fast approximation of matrix coherence and statistical leverage. *J. Mach. Learn. Res.*, 13(1):3475–3506, December 2012.

[17] Petros Drineas, Michael W Mahoney, and S Muthukrishnan. Sampling algorithms for ℓ_2 regression and applications. In *Proceedings of the seventeenth annual ACM-SIAM symposium on Discrete algorithm*, pages 1127–1136, 2006.

[18] Valerii V. Fedorov, William J. Studden, and E. M. Klimko, editors. *Theory of optimal experiments*. Probability and mathematical statistics. Academic Press, New York, 1972.

[19] Mike Gartrell, Ulrich Paquet, and Noam Koenigstein. Bayesian low-rank determinantal point processes. In *Proceedings of the 10th ACM Conference on Recommender Systems*, RecSys '16, pages 349–356, New York, NY, USA, 2016.

[20] David Gross and Vincent Nesme. Note on sampling without replacing from a finite collection of matrices. *arXiv:1001.2738*, January 2010.

[21] Nicholas JA Harvey and Neil Olver. Pipage rounding, pessimistic estimators and matrix concentration. In *Proceedings of the twenty-fifth annual ACM-SIAM symposium on Discrete algorithms*, pages 926–945. SIAM, 2014.

[22] Wassily Hoeffding. Probability inequalities for sums of bounded random variables. *Journal of the American statistical association*, 58(301):13–30, 1963.

[23] Alex Kulesza and Ben Taskar. k-DPPs: Fixed-Size Determinantal Point Processes. In *Proceedings of the 28th International Conference on Machine Learning*, pages 1193–1200. Omnipress, 2011.

[24] Alex Kulesza and Ben Taskar. *Determinantal Point Processes for Machine Learning*. Now Publishers Inc., Hanover, MA, USA, 2012.

[25] Yin Tat Lee and He Sun. Constructing linear-sized spectral sparsification in almost-linear time. In *Foundations of Computer Science (FOCS), 2015 IEEE 56th Annual Symposium on*, pages 250–269. IEEE, 2015.

[26] Chengtao Li, Stefanie Jegelka, and Suvrit Sra. Polynomial time algorithms for dual volume sampling. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 5045–5054. 2017.

[27] Michael W. Mahoney. Randomized algorithms for matrices and data. *Found. Trends Mach. Learn.*, 3(2):123–224, February 2011.

[28] Zelda E. Mariet and Suvrit Sra. Elementary symmetric polynomials for optimal experimental design. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 2136–2145. 2017.

[29] Aleksandar Nikolov, Mohit Singh, and Uthaipon Tao Tantipongpipat. Proportional volume sampling and approximation algorithms for A-optimal design. *arXiv:1802.08318*, July 2018.

[30] Robin Pemantle and Yuval Peres. Concentration of Lipschitz functionals of determinantal and other strong rayleigh measures. *Combinatorics, Probability and Computing*, 23(1):140–160, 2014.

[31] Tamas Sarlos. Improved approximation algorithms for large matrices via random projections. In *Proceedings of the 47th Annual IEEE Symposium on Foundations of Computer Science*, FOCS '06, pages 143–152, Washington, DC, USA, 2006.

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

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

A Proof of (2) from Theorem 1

First, let us calculate $L(\mathbf{w}^*)$. Observe that

$$(\mathbf{X}^\top\mathbf{X})^{-1} = \overbrace{\Big(1+\frac{n-d}{d}\gamma^2\Big)^{-1}}^{c}\,\mathbf{I},$$
$$\text{and}\quad \mathbf{w}^* = c\,\mathbf{X}^\top\mathbf{y} = c\,\mathbf{1}_d.$$

The loss $L(\mathbf{w})$ of any $\mathbf{w}\in\mathbb{R}^d$ can be decomposed as $L(\mathbf{w}) = \sum_{i=1}^d L_i(\mathbf{w})$, where $L_i(\mathbf{w})$ is the total loss incurred on all input vectors $\mathbf{e}_i$ or $\gamma\mathbf{e}_i$. For $\mathbf{w}^*$, the i-th component is

$$L_i(\mathbf{w}^*) = (1-c)^2 + \overbrace{\frac{n-d}{d}\gamma^2}^{\frac{1}{c}-1}\,c^2 = 1-c.$$

Note that i-th leverage score of $\mathbf{X}$ is equal $l_i = \mathbf{x}_i^\top(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{x}_i = c$, so we obtain that

$$L(\mathbf{w}^*) = d\,(1-c) = \sum_{i=1}^d (1-l_i). \tag{8}$$

Next, we compute $L(\mathbf{w}_S^*)$. Suppose that $S\subseteq\{1..n\}$ is produced by size k standard volume sampling. Note that if for some $1\le i\le d$ we have $i\notin S$, then $(\mathbf{w}_S^*)_i = 0$ and therefore $L_i(\mathbf{w}_S^*) = 1$. Moreover, denoting $b_i \overset{def}{=} \mathbf{1}_{[i\in S]}$,

$$(\mathbf{X}_S^\top\mathbf{X}_S)^{-1} \succeq (\mathbf{X}^\top\mathbf{X})^{-1} = c\,\mathbf{I},\quad\text{and}\quad \mathbf{X}_S^\top\mathbf{y}_S = (b_1,\dots,b_d)^\top,$$

so if $i\in S$, then $(\mathbf{w}_S^*)_i \ge c$ and

$$L_i(\mathbf{w}_S^*) \ge \frac{n-d}{d}\,\gamma^2\,c^2 = \Big(\frac{1}{c}-1\Big)c^2 = c\,L_i(\mathbf{w}^*).$$

Putting the cases of $i\in S$ and $i\notin S$ together, we get

$$\begin{aligned} L_i(\mathbf{w}_S^*) &\ge c\,L_i(\mathbf{w}^*) + (1-c\,L_i(\mathbf{w}^*))\,(1-b_i)\\ &\ge c\,L_i(\mathbf{w}^*) + c^2(1-b_i).\end{aligned}$$

Applying the marginal probability formula for volume sampling (see (4)), we note that

$$\mathbb{E}[1-b_i] = 1-\Pr(i\in S) = \frac{n-k}{n-d}\,(1-c) = \frac{n-k}{n-d}\,L_i(\mathbf{w}^*).$$

Taking expectation over $L_i(\mathbf{w}_S^*)$ and summing the components over $i\in[d]$, we get

$$\mathbb{E}[L(\mathbf{w}_S^*)] \ge L(\mathbf{w}^*)\Big(c + c^2\frac{n-k}{n-d}\Big).$$

Note that as $\gamma\to 0$, we have $c\to 1$, thus showing (2).

B Properties of rescaled volume sampling

We give proofs of the properties of rescaled volume sampling which hold for any rescaling distribution q. In this section, we will use $Z = d!\binom{k}{d}\det(\mathbf{X}^\top\mathbf{X})$ as the normalization constant for rescaled volume sampling.

B.1 Proof of Proposition 2

First, we apply the Cauchy-Binet formula to the determinant term specified by a fixed sequence $\pi\in[n]^k$:

$$\det(\mathbf{X}^\top\mathbf{Q}_\pi\mathbf{X}) = \sum_{S\in\binom{[k]}{d}}\det(\mathbf{X}^\top\mathbf{Q}_{\pi_S}\mathbf{X}) = \sum_{S\in\binom{[k]}{d}}\det(\mathbf{X}_{\pi_S})^2\prod_{i\in S}\frac{1}{q_{\pi_i}}.$$

Next, we compute the sum, using the above identity:

$$\begin{aligned}
\sum_{\pi\in[n]^k} \det(\mathbf{X}^\top\mathbf{Q}_\pi\mathbf{X}) \prod_{i=1}^k q_{\pi_i} &= \sum_{\pi\in[n]^k} \sum_{S\in\binom{[k]}{d}} \det(\mathbf{X}_{\pi_S})^2 \prod_{i\in[k]\setminus S} q_{\pi_i} \\
&= \binom{k}{d} \sum_{\bar\pi\in[n]^d} \det(\mathbf{X}_{\bar\pi})^2 \sum_{\tilde\pi\in[n]^{k-d}} \prod_{i=1}^{k-d} q_{\tilde\pi_i} \\
&= \binom{k}{d} \sum_{\bar\pi\in[n]^d} \det(\mathbf{X}_{\bar\pi})^2 \Big(\sum_{i=1}^n q_i\Big)^{k-d} \\
&= \binom{k}{d} d! \sum_{S\in\binom{[n]}{d}} \det(\mathbf{X}_S)^2 = k(k-1)\cdots(k-d+1)\det(\mathbf{X}^\top\mathbf{X}),
\end{aligned}$$

where the steps closely follow the corresponding derivation for Theorem 3, given in Section 3.1.

B.2 Proof of Theorem 4

We will prove that for any vector $\mathbf{v}\in\mathbb{R}^d$,

$$\mathbb{E}\big[\mathbf{v}^\top(\mathbf{X}^\top\mathbf{Q}_\pi\mathbf{X})^{-1}\mathbf{v}\big] \le \frac{\mathbf{v}^\top(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{v}}{k-d+1},$$

which immediately implies the corresponding matrix inequality. First, we use Sylvester's formula, which holds whenever a matrix $\mathbf{A}\in\mathbb{R}^{d\times d}$ is full rank:

$$\det(\mathbf{A}+\mathbf{v}\mathbf{v}^\top) = \det(\mathbf{A})\big(1+\mathbf{v}^\top\mathbf{A}^{-1}\mathbf{v}\big).$$

Note that whenever the matrix is not full rank, its determinant is 0 (in which case we avoid computing the matrix inverse), so we have for any $\pi\in[n]^k$:

$$\begin{aligned}
\det(\mathbf{X}^\top\mathbf{Q}_\pi\mathbf{X})\,\mathbf{v}^\top(\mathbf{X}^\top\mathbf{Q}_\pi\mathbf{X})^{-1}\mathbf{v} &\le \det(\mathbf{X}^\top\mathbf{Q}_\pi\mathbf{X}+\mathbf{v}\mathbf{v}^\top) - \det(\mathbf{X}^\top\mathbf{Q}_\pi\mathbf{X}) \\
&\overset{(*)}{=} \sum_{S\in\binom{[k]}{d-1}} \det(\mathbf{X}_{\pi_S}^\top\mathbf{X}_{\pi_S}+\mathbf{v}\mathbf{v}^\top)\prod_{i\in S}\frac{1}{q_{\pi_i}},
\end{aligned}$$

where $(*)$ follows from applying the Cauchy-Binet formula to both of the determinants, and cancelling out common terms. Next, we proceed in a standard fashion, summing over all $\pi\in[n]^k$:

$$\begin{aligned}
Z\,\mathbb{E}\big[\mathbf{v}^\top(\mathbf{X}^\top\mathbf{Q}_\pi\mathbf{X})^{-1}\mathbf{v}\big] &= \sum_{\pi\in[n]^k} \mathbf{v}^\top(\mathbf{X}^\top\mathbf{Q}_\pi\mathbf{X})^{-1}\mathbf{v}\det(\mathbf{X}^\top\mathbf{Q}_\pi\mathbf{X})\prod_{i=1}^k q_{\pi_i} \\
&\le \sum_{\pi\in[n]^k}\sum_{S\in\binom{[k]}{d-1}} \det(\mathbf{X}_{\pi_S}^\top\mathbf{X}_{\pi_S}+\mathbf{v}\mathbf{v}^\top)\prod_{i\in[k]\setminus S} q_{\pi_i} \\
&= \binom{k}{d-1}\sum_{\bar\pi\in[n]^{d-1}} \det(\mathbf{X}_{\bar\pi}^\top\mathbf{X}_{\bar\pi}+\mathbf{v}\mathbf{v}^\top)\sum_{\tilde\pi\in[n]^{k-d+1}}\prod_{i=1}^{k-d+1} q_{\pi_i} \\
&= \binom{k}{d-1}(d-1)!\sum_{S\in\binom{[n]}{d-1}} \det(\mathbf{X}_S^\top\mathbf{X}_S+\mathbf{v}\mathbf{v}^\top) \\
&= \frac{d!\binom{k}{d}}{k-d+1}\big(\det(\mathbf{X}^\top\mathbf{X}+\mathbf{v}\mathbf{v}^\top)-\det(\mathbf{X}^\top\mathbf{X})\big) = Z\,\frac{\mathbf{v}^\top(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{v}}{k-d+1}.
\end{aligned}$$

B.3 Proof of Proposition 5

First, we compute the marginal probability of a fixed element of sequence π containing a particular index $i \in [n]$ under q-rescaled volume sampling:

$$Z\,\Pr(\pi_k{=}i) \;=\; \sum_{\pi\in[n]^{k-1}} \det(\mathbf{X}^\top\mathbf{Q}_{[\pi,i]}\mathbf{X})\; q_i \prod_{t=1}^{k-1} q_{\pi_t}$$
$$= \underbrace{q_i \sum_{\pi\in[n]^{k-1}} \sum_{S\in\binom{[k-1]}{d}} \det(\mathbf{X}_{\pi_S})^2 \prod_{t\in[k-1]\setminus S} q_{\pi_t}}_{T_1} + \underbrace{\sum_{\pi\in[n]^{k-1}} \sum_{S\in\binom{[k-1]}{d-1}} \det(\mathbf{X}_{\pi_S}^\top\mathbf{X}_{\pi_S} + \mathbf{x}_i\mathbf{x}_i^\top) \prod_{t\in[k-1]\setminus S} q_{\pi_t}}_{T_2},$$

where the first term can be computed by following the derivation in Appendix B.1, obtaining $T_1 = q_i\frac{k-d}{k}\,Z$, and the second term is derived as in Appendix B.2, obtaining $T_2 = \frac{l_i}{k}\,Z$. Putting this together, we get

$$\Pr(\pi_k{=}i) = \frac{1}{k}\big((k{-}d)\,q_i + l_i\big).$$

Note that by symmetry this applies to any element of the sequence. We can now easily compute the desired expectation:

$$\mathbb{E}\big[(\mathbf{Q}_\pi)_{ii}\big] = \frac{1}{q_i}\sum_{t=1}^{k}\Pr(\pi_t{=}i) = (k{-}d) + \frac{l_i}{q_i}.$$

B.4 Proof of Lemma 7

First step of the reverse iterative sampling procedure described in Section 2 involves removing one row from the given matrix with probability proportional to the square volume of that submatrix:

$$\forall_{i\in S}\qquad \Pr(i\,|\,\pi_S) = \frac{\det(\mathbf{X}^\top\mathbf{Q}_{\pi_{S\setminus i}}\mathbf{X})}{(|S|-d)\det(\mathbf{X}^\top\mathbf{Q}_{\pi_S}\mathbf{X})}.$$

Suppose that $k = s-1$ and let $\tilde\pi = \pi_S \in [n]^{s-1}$ denote the sequence obtained after performing one step of the row-removal procedure. Then,

$$\Pr(\tilde\pi) = \sum_{\substack{\pi\in[n]^k:\\ \tilde\pi \text{ is a subsequence of } \pi}} \Pr(\tilde\pi\,|\,\pi)\,\Pr(\pi) \overset{(*)}{=} \sum_{i=1}^{n} s\;\overbrace{\Pr(i\,|\,[\tilde\pi,i])}^{\text{removing one row}}\;\overbrace{\Pr([\tilde\pi,i])}^{\text{rescaled sampling}}$$
$$= \sum_{i=1}^{n} s\,\frac{\det(\mathbf{X}^\top\mathbf{Q}_{\tilde\pi}\mathbf{X})}{(s{-}d)\det(\mathbf{X}^\top\mathbf{Q}_{[\tilde\pi,i]}\mathbf{X})}\,\frac{\det(\mathbf{X}^\top\mathbf{Q}_{[\tilde\pi,i]}\mathbf{X})\,(\prod_{j=1}^{s-1}q_{\tilde\pi_j})\,q_i}{\frac{s!}{(s-d)!}\det(\mathbf{X}^\top\mathbf{X})}$$
$$= \frac{\det(\mathbf{X}^\top\mathbf{Q}_{\tilde\pi}\mathbf{X})(\prod_{j=1}^{s-1}q_{\tilde\pi_j})}{\frac{s-d}{s}\,\frac{s!}{(s-d)!}\det(\mathbf{X}^\top\mathbf{X})}\sum_{i=1}^{n} q_i = \frac{\det(\mathbf{X}^\top\mathbf{Q}_{\tilde\pi}\mathbf{X})\,(\prod_{j=1}^{s-1}q_{\tilde\pi_j})}{\frac{(s-1)!}{(s-1-d)!}\det(\mathbf{X}^\top\mathbf{X})},$$

where $(*)$ follows because the ordering of sequence π does not affect the probabilities, and the factor s next to the sum counts the number of ways to place index i into the sequence $\tilde\pi$ to obtain π. Thus, by induction, for any $k < s$ the algorithm correctly samples from q-rescaled volume sampling.

C Proof of Theorem 8

We rewrite the expected square norm as:

$$\mathbb{E}\left[\left\|\frac{1}{k}\mathbf{U}^\top\mathbf{Q}_\pi\mathbf{r} - \mathbf{U}^\top\mathbf{r}\right\|^2\right] = \mathbb{E}\left[\left\|\mathbf{U}^\top\Big(\frac{1}{k}\mathbf{Q}_\pi{-}\mathbf{I}\Big)\mathbf{r}\right\|^2\right] = \mathbb{E}\left[\mathbf{r}^\top\Big(\frac{1}{k}\mathbf{Q}_\pi{-}\mathbf{I}\Big)\mathbf{U}\mathbf{U}^\top\Big(\frac{1}{k}\mathbf{Q}_\pi{-}\mathbf{I}\Big)\mathbf{r}\right]$$
$$= \mathbf{r}^\top\,\mathbb{E}\left[\Big(\frac{1}{k}\mathbf{Q}_\pi{-}\mathbf{I}\Big)\mathbf{U}\mathbf{U}^\top\Big(\frac{1}{k}\mathbf{Q}_\pi{-}\mathbf{I}\Big)\right]\mathbf{r}$$
$$\le \lambda_{\max}\Big(\underbrace{\big(\mathbb{E}[(z_i{-}1)(z_j{-}1)]\,\mathbf{u}_i^\top\mathbf{u}_j\big)_{ij}}_{\mathbf{M}}\Big)\,\|\mathbf{r}\|^2, \quad \text{where } z_i = \frac{1}{k}(\mathbf{Q}_\pi)_{ii}.$$

It remains to bound $\lambda_{\max}(\mathbf{M})$. By Proposition 5, for leveraged volume sampling $\mathbb{E}[(\mathbf{Q}_\pi)_{ii}] = k$, so

$$\mathbb{E}[(z_i-1)(z_j-1)] = \frac{1}{k^2}\Big(\mathbb{E}\big[(\mathbf{Q}_\pi)_{ii}(\mathbf{Q}_\pi)_{jj}\big] - \mathbb{E}\big[(\mathbf{Q}_\pi)_{ii}\big]\mathbb{E}\big[(\mathbf{Q}_\pi)_{jj}\big]\Big) = \frac{1}{k^2}\operatorname{cov}\big[(\mathbf{Q}_\pi)_{ii},\,(\mathbf{Q}_\pi)_{jj}\big].$$

For rescaled volume sampling this is given in the following lemma, proven in Appendix C.1.

Lemma 11 *For any $\mathbf{X}$ and q, if sequence $\pi \in [n]^k$ is sampled from q-rescaled volume sampling then*

$$\operatorname{cov}\big[(\mathbf{Q}_\pi)_{ii},\,(\mathbf{Q}_\pi)_{jj}\big] = \mathbf{1}_{i=j}\frac{1}{q_i}\mathbb{E}\big[(\mathbf{Q}_\pi)_{ii}\big] - (k-d) - \frac{(\mathbf{x}_i^\top(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{x}_j)^2}{q_i q_j}.$$

Since $\|\mathbf{u}_i\|^2 = l_i = dq_i$ and $\mathbf{u}_i^\top(\mathbf{U}^\top\mathbf{U})^{-1}\mathbf{u}_j = \mathbf{u}_i^\top\mathbf{u}_j$, we can express matrix $\mathbf{M}$ as follows:

$$\mathbf{M} = \operatorname{diag}\left(\frac{d\,\mathbb{E}\big[(\mathbf{Q}_\pi)_{ii}\big]}{\|\mathbf{u}_i\|^2 k^2}\|\mathbf{u}_i\|^2\right)_{i=1}^n - \frac{k-d}{k^2}\mathbf{U}\mathbf{U}^\top - \frac{d^2}{k^2}\left(\frac{(\mathbf{u}_i^\top\mathbf{u}_j)^3}{\|\mathbf{u}_i\|^2\|\mathbf{u}_j\|^2}\right)_{ij}.$$

The first term simplifies to $\frac{d}{k}\mathbf{I}$, and the second term is negative semi-definite, so

$$\lambda_{\max}(\mathbf{M}) \le \frac{d}{k} + \frac{d^2}{k^2}\left\|\left(\frac{(\mathbf{u}_i^\top\mathbf{u}_j)^3}{\|\mathbf{u}_i\|^2\|\mathbf{u}_j\|^2}\right)_{ij}\right\|.$$

Finally, we decompose the last term into a Hadamard product of matrices, and apply a classical inequality by [3] (symbol "$\circ$" denotes Hadamard matrix product—i.e., elementwise multiplication):

$$\begin{aligned}
\left\|\left(\frac{(\mathbf{u}_i^\top\mathbf{u}_j)^3}{\|\mathbf{u}_i\|^2\|\mathbf{u}_j\|^2}\right)_{ij}\right\| &= \left\|\left(\frac{\mathbf{u}_i^\top\mathbf{u}_j}{\|\mathbf{u}_i\|\,\|\mathbf{u}_j\|}\right)_{ij} \circ \left(\frac{(\mathbf{u}_i^\top\mathbf{u}_j)^2}{\|\mathbf{u}_i\|\|\mathbf{u}_j\|}\right)_{ij}\right\| \\
&\le \left\|\left(\frac{(\mathbf{u}_i^\top\mathbf{u}_j)^2}{\|\mathbf{u}_i\|\|\mathbf{u}_j\|}\right)_{ij}\right\| = \left\|\left(\frac{\mathbf{u}_i^\top\mathbf{u}_j}{\|\mathbf{u}_i\|\,\|\mathbf{u}_j\|}\right)_{ij} \circ \mathbf{U}\mathbf{U}^\top\right\| \\
&\le \|\mathbf{U}\mathbf{U}^\top\| \;=\; 1.
\end{aligned}$$

Thus, we conclude that $\mathbb{E}[\|\frac{1}{k}\mathbf{U}^\top\mathbf{Q}_\pi\mathbf{r} - \mathbf{U}^\top\mathbf{r}\|^2] \le (\frac{d}{k} + \frac{d^2}{k^2})\|\mathbf{r}\|^2$, completing the proof.

C.1 Proof of Lemma 11

We compute marginal probability of two elements in the sequence π having particular values $i, j \in [n]$:

$$Z\Pr\big((\pi_{k-1}=i)\wedge(\pi_k=j)\big) = \sum_{\pi\in[n]^{k-2}}\sum_{S\in\binom{[k]}{d}} \det(\mathbf{X}_{[\pi,i,j]_S}^\top\mathbf{X}_{[\pi,i,j]_S}) \prod_{t\in[k]\setminus S} q_{[\pi,i,j]_t}.$$

We partition the set $\binom{[k]}{d}$ of all subsets of size d into four groups, and summing separately over each of the groups, we have

$$Z\Pr\big((\pi_{k-1}=i)\wedge(\pi_k=j)\big) = T_{00}+T_{01}+T_{10}+T_{11}, \qquad \text{where:}$$

1. Let $G_{00} = \{S\in\binom{[k]}{d} :\; k-1\notin S,\; k\notin S\}$, and following derivation in Appendix B.1,

$$T_{00} = q_i\,q_j\sum_{\pi\in[n]^{k-2}}\sum_{S\in G_{00}}\det(\mathbf{X}_{\pi_S})^2\prod_{t\in[k-2]\setminus S} q_{\pi_t} = q_i\,q_j\frac{(k-d-1)(k-d)}{(k-1)\,k}\,Z.$$

2. Let $G_{10} = \{S\in\binom{[k]}{d} :\; k-1\in S,\; k\notin S\}$, and following derivation in Appendix B.2,

$$T_{10} = q_j\sum_{\pi\in[n]^{k-1}}\sum_{S\in G_{10}}\det(\mathbf{X}_{[\pi,i]_S})^2\prod_{t\in[k-1]\setminus S} q_{[\pi,i]_t} = l_i\,q_j\frac{(k-d)}{(k-1)\,k}\,Z.$$

3. $G_{01} = \{S\in\binom{[k]}{d} :\; k-1\notin S,\; k\in S\}$, and by symmetry, $T_{01} = l_j\,q_i\frac{(k-d)}{(k-1)\,k}\,Z$.

4. Let $G_{11} = \{S \in \binom{[k]}{d} : \ k-1 \in S, \ k \in S\}$, and the last term is

$$\begin{aligned}
T_{11} &= \sum_{\pi \in [n]^{k-1}} \sum_{S \in G_{11}} \det(\mathbf{X}_{[\pi,i,j]_S})^2 \prod_{t \in [k] \setminus S} q_{[\pi,i,j]_t} \\
&= \binom{k-2}{d-2} \sum_{\pi \in [n]^{d-2}} \det(\mathbf{X}_{[\pi,i,j]})^2 \\
&= \binom{k-2}{d-2} (d-2)! \left(\det(\mathbf{X}^\top \mathbf{X}) - \det(\mathbf{X}_{-i}^\top \mathbf{X}_{-i}) - \det(\mathbf{X}_{-j}^\top \mathbf{X}_{-j}) + \det(\mathbf{X}_{-i,j}^\top \mathbf{X}_{-i,j}) \right) \\
&\overset{(*)}{=} \frac{d!\binom{k}{d}}{k(k-1)} \det(\mathbf{X}^\top \mathbf{X}) \Big(1 - \underbrace{(1-l_i)}_{\frac{\det(\mathbf{X}_{-i}^\top \mathbf{X}_{-i})}{\det(\mathbf{X}^\top \mathbf{X})}} - \underbrace{(1-l_j)}_{\frac{\det(\mathbf{X}_{-j}^\top \mathbf{X}_{-j})}{\det(\mathbf{X}^\top \mathbf{X})}} + \underbrace{(1-l_i)(1-l_j) - l_{ij}^2}_{\frac{\det(\mathbf{X}_{-i,j}^\top \mathbf{X}_{-i,j})}{\det(\mathbf{X}^\top \mathbf{X})}} \Big) \\
&= \frac{Z}{k(k-1)} (\ell_i \ell_j - \ell_{ij}^2),
\end{aligned}$$

where $l_{ij} = \mathbf{x}_i^\top (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{x}_j$, and $(*)$ follows from repeated application of Sylvester's determinant formula (as in Appendix B.2). Putting it all together, we can now compute the expectation for $i \neq j$:

$$\begin{aligned}
\mathbb{E}\big[(\mathbf{Q}_\pi)_{ii} (\mathbf{Q}_\pi)_{jj}\big] &= \frac{1}{q_i \, q_j} \sum_{t_1=1}^{k} \sum_{t_2=1}^{k} \Pr\big((\pi_{k-1}=i) \wedge (\pi_k=j)\big) \\
&= \frac{k(k-1)}{q_i \, q_j} \overbrace{\Pr\big((\pi_{k-1}=i) \wedge (\pi_k=j)\big)}^{\frac{1}{Z}(T_{00}+T_{10}+T_{01}+T_{11})} \\
&= (k-d-1)(k-d) + (k-d)\frac{l_i}{q_i} + (k-d)\frac{l_j}{q_j} + \frac{l_i l_j}{q_i \, q_j} - \frac{l_{ij}^2}{q_i \, q_j} \\
&= \Big((k-d)q_i + \frac{l_i}{q_i}\Big)\Big((k-d)q_j + \frac{l_j}{q_j}\Big) - (k-d) - \frac{l_{ij}^2}{q_i \, q_j} \\
&= \mathbb{E}\big[(\mathbf{Q}_\pi)_{ii}\big] \, \mathbb{E}\big[(\mathbf{Q}_\pi)_{jj}\big] - (k-d) - \frac{l_{ij}^2}{q_i q_j}.
\end{aligned}$$

Finally, if $i = j$, then

$$\begin{aligned}
\mathbb{E}[(\mathbf{Q}_\pi)_{ii} (\mathbf{Q}_\pi)_{ii}] &= \frac{1}{q_i^2} \sum_{t_1=1}^{k} \sum_{t_2=1}^{k} \Pr(\pi_{t_1}=i \ \wedge \ \pi_{t_2}=i) \\
&= \frac{k(k-1)}{q_i^2} \Pr(\pi_{k-1}=i \ \wedge \ \pi_k=i) + \frac{k}{q_i^2} \Pr(\pi_k=i) \\
&= \big(\mathbb{E}\big[(\mathbf{Q}_\pi)_{ii}\big]\big)^2 - (k-d) - \frac{l_i^2}{q_i^2} + \frac{1}{q_i} \mathbb{E}\big[(\mathbf{Q}_\pi)_{ii}\big].
\end{aligned}$$

D Proof of Theorem 9

We break the sampling procedure down into two stages. First, we do leveraged volume sampling of a sequence $\pi \in [n]^m$ of size $m \geq C_0 d^2/\delta$, then we do standard volume sampling size k from matrix $(\mathbf{Q}_{[1..n]}^{1/2} \mathbf{U})_\pi$. Since rescaled volume sampling is closed under this subsampling (Lemma 7), this procedure is equivalent to size k leveraged volume sampling from $\mathbf{U}$. To show that the first stage satisfies the subspace embedding condition, we simply use the bound from Theorem 8 (see details in Appendix D.1):

Lemma 12 *There is an absolute constant C_0, s.t. if sequence $\pi \in [n]^m$ is generated via leveraged volume sampling of size m at least $C_0 \, d^2/\delta$ from $\mathbf{U}$, then*

$$\Pr\left(\lambda_{\min}\Big(\frac{1}{m} \mathbf{U}^\top \mathbf{Q}_\pi \mathbf{U}\Big) \leq \frac{1}{2} \right) \leq \delta.$$

The size of m is much larger than what we claim is sufficient. However, we use it to achieve a tighter bound in the second stage. To obtain substantially smaller sample sizes for subspace embedding than what Theorem 8 can deliver, it is standard to use tail bounds for the sums of independent matrices. However, applying these results to joint sampling is a challenging task. Interestingly, [26] showed that volume sampling is a strongly Raleigh measure, implying that the sampled vectors are negatively correlated. This guarantee is sufficient to show tail bounds for real-valued random variables [see, e.g., 30], however it has proven challenging in the matrix case, as discussed by [21]. One notable exception is uniform sampling without replacement, which is a negatively correlated joint distribution. A reduction argument originally proposed by [22], but presented in this context by [20], shows that uniform sampling without replacement offers the same tail bounds as i.i.d. uniform sampling.

Lemma 13 *Assume that* $\lambda_{\min}\big(\frac{1}{m}\mathbf{U}^\top\mathbf{Q}_\pi\mathbf{U}\big) \geq \frac{1}{2}$. *Suppose that set* T *is a set of fixed size sampled uniformly without replacement from* $[m]$. *There is a constant* C_1 *s.t. if* $|T| \geq C_1\, d\ln(d/\delta)$, *then*

$$\Pr\Big(\lambda_{\min}\Big(\frac{1}{|T|}\mathbf{U}^\top\mathbf{Q}_{\pi_T}\mathbf{U}\Big) \leq \frac{1}{4}\Big) \leq \delta.$$

The proof of Lemma 13 (given in appendix D.2) is a straight-forward application of the argument given by [20]. We now propose a different reduction argument showing that a subspace embedding guarantee for uniform sampling without replacement leads to a similar guarantee for volume sampling. We achieve this by exploiting a volume sampling algorithm proposed recently by [13], shown in Algorithm 3, which is a modification of the reverse iterative sampling procedure introduced in [11]. This procedure relies on iteratively removing elements from the set S until we are left with k elements. Specifically, at each step, we sample an index i from a conditional distribution, $i \sim \Pr(i \,|\, S) = (1 - \frac{1}{q_{\pi_i}}\mathbf{u}_{\pi_i}^\top(\mathbf{U}^\top\mathbf{Q}_{\pi_S}\mathbf{U})^{-1}\mathbf{u}_{\pi_i})/(|S| - d)$. Crucially for us, each step proceeds via rejection sampling with the proposal distribution being uniform. We can easily modify the algorithm, so that the samples from the proposal distribution are used to construct a uniformly sampled set T, as shown in Algorithm 4. Note that sets S returned by both algorithms are identically distributed, and furthermore, T is a subset of S, because every index taken out of S is also taken out of T.

Algorithm 3: Volume sampling

$S \leftarrow [m]$
while $|S| > k$
  repeat
    Sample i unif. out of S
    $p \leftarrow 1 - \frac{1}{q_{\pi_i}}\mathbf{u}_{\pi_i}^\top(\mathbf{U}^\top\mathbf{Q}_{\pi_S}\mathbf{U})^{-1}\mathbf{u}_{\pi_i}$
    Sample *Accept* $\sim$ Bernoulli(p)
  until *Accept* $=$ true
  $S \leftarrow S\backslash\{i\}$
end
return S

Algorithm 4: Coupled sampling

1: $S, T \leftarrow [m]$
2: **while** $|S| > k$
3: Sample i unif. out of $[m]$
4: $T \leftarrow T - \{i\}$
5: **if** $i \in S$
6: $p \leftarrow 1 - \frac{1}{q_{\pi_i}}\mathbf{u}_{\pi_i}^\top(\mathbf{U}^\top\mathbf{Q}_{\pi_S}\mathbf{U})^{-1}\mathbf{u}_{\pi_i}$
7: Sample *Accept* $\sim$ Bernoulli(p)
8: **if** *Accept* $=$ true, $S \leftarrow S\backslash\{i\}$ **end**
9: **end**
10: **end**
11: **return** S, T

By Lemma 13, if size of T is at least $C_1\, d\log(d/\delta)$, then this set offers a subspace embedding guarantee. Next, we will show that in fact set T is not much smaller than S, implying that the same guarantee holds for S. Specifically, we will show that $|S \setminus T| = O(d\log(d/\delta))$. Note that it suffices to bound the number of times that a uniform sample is rejected by sampling $A = 0$ in line 7 of Algorithm 4. Denote this number by R. Note that $R = \sum_{t=k+1}^{m} R_t$, where $m = |Q|$ and R_t is the number of times that $A = 0$ was sampled while the size of set S was t. Variables R_t are independent, and each is distributed according to the geometric distribution (number of failures until success), with the success probability

$$r_t = \frac{1}{t}\sum_{i\in S}\Big(1 - \frac{1}{q_{\pi_i}}\mathbf{u}_{\pi_i}^\top(\mathbf{U}^\top\mathbf{Q}_{\pi_S}\mathbf{U})^{-1}\mathbf{u}_{\pi_i}\Big) = \frac{1}{t}\Big(t - \mathrm{tr}\big((\mathbf{U}^\top\mathbf{Q}_{\pi_S}\mathbf{U})^{-1}\mathbf{U}^\top\mathbf{Q}_{\pi_S}\mathbf{U}\big)\Big) = \frac{t-d}{t}.$$

Now, as long as $\frac{m-d}{k-d} \leq C_0\, d^2/\delta$, we can bound the expected value of R as follows:

$$\mathbb{E}[R] = \sum_{t=k+1}^{m}\mathbb{E}[R_t] = \sum_{t=k+1}^{m}\Big(\frac{t}{t-d} - 1\Big) = d\sum_{t=k-d+1}^{m-d}\frac{1}{t} \leq d\,\ln\Big(\frac{m-d}{k-d}\Big) \leq C_2\, d\ln(d/\delta).$$

In this step, we made use of the first stage sampling, guaranteeing that the term under the logarithm is bounded. Next, we show that the upper tail of R decays very rapidly given a sufficiently large gap between m and k (proof in Appendix D.3):

Lemma 14 *Let $R_t \sim \mathrm{Geom}(\frac{t-d}{t})$ be a sequence of independent geometrically distributed random variables (number of failures until success). Then, for any $d < k < m$ and $a > 1$,*

$$\Pr\big(R \geq a\,\mathbb{E}[R]\big) \leq e^{\frac{a}{2}}\left(\frac{k-d}{m-d}\right)^{\frac{a}{2}-1} \quad \textit{for} \quad R = \sum_{t=k+1}^{m} R_t.$$

Let $a = 4$ in Lemma 14. Setting $C = C_1 + 2a\,C_2$, for any $k \geq C\,d\ln(d/\delta)$, using $m = \max\{C_0\,\frac{d^2}{\delta},\ d + \mathrm{e}^2\,\frac{k}{\delta}\}$, we obtain that

$$R \leq a\,C_2\,d\ln(d/\delta) \leq k/2, \quad \text{w.p.} \quad \geq 1 - \mathrm{e}^2\,\frac{k-d}{m-d} \geq 1-\delta,$$

showing that $|T| \geq k - R \geq C_1\,d\ln(d/\delta)$ and $k \leq 2|T|$.

Therefore, by Lemmas 12, 13 and 14, there is a $1-3\delta$ probability event in which

$$\lambda_{\min}\Big(\frac{1}{|T|}\mathbf{U}^\top\mathbf{Q}_{\pi_T}\mathbf{U}\Big) \geq \frac{1}{4} \quad \text{and} \quad k \leq 2|T|.$$

In this same event,

$$\lambda_{\min}\Big(\frac{1}{k}\mathbf{U}^\top\mathbf{Q}_{\pi_S}\mathbf{U}\Big) \geq \lambda_{\min}\Big(\frac{1}{k}\mathbf{U}^\top\mathbf{Q}_{\pi_T}\mathbf{U}\Big) \geq \lambda_{\min}\Big(\frac{1}{2|T|}\mathbf{U}^\top\mathbf{Q}_{\pi_T}\mathbf{U}\Big) \geq \frac{1}{2}\cdot\frac{1}{4} = \frac{1}{8},$$

which completes the proof of Theorem 9.

D.1 Proof of Lemma 12

Replacing vector $\mathbf{r}$ in Theorem 8 with each column of matrix $\mathbf{U}$, we obtain that for $m \geq C\,\frac{d}{\epsilon}$,

$$\mathbb{E}\big[\|\mathbf{U}^\top\mathbf{Q}_\pi\mathbf{U} - \mathbf{U}^\top\mathbf{U}\|_F^2\big] \leq \epsilon\,\|\mathbf{U}\|_F^2 = \epsilon\,d.$$

We bound the 2-norm by the Frobenius norm and use Markov's inequality, showing that w.p. $\geq 1-\delta$

$$\|\mathbf{U}^\top\mathbf{Q}_\pi\mathbf{U} - \mathbf{I}\| \leq \|\mathbf{U}^\top\mathbf{Q}_\pi\mathbf{U} - \mathbf{I}\|_F \leq \sqrt{\epsilon\,d/\delta}.$$

Setting $\epsilon = \frac{\delta}{4d}$, for $m \geq C_0\,d^2/\delta$, the above inequality implies that

$$\lambda_{\min}\Big(\frac{1}{m}\mathbf{U}^\top\mathbf{Q}_\pi\mathbf{U}\Big) \geq \frac{1}{2}.$$

D.2 Proof of Lemma 13

Let π denote the sequence of m indices selected by volume sampling in the first stage. Suppose that $i_1, ..., i_k$ are independent uniformly sampled indices from $[m]$, and let $j_1, ..., j_k$ be indices sampled uniformly without replacement from $[m]$. We define matrices

$$\mathbf{Z} \overset{def}{=} \sum_{t=1}^{k} \overbrace{\frac{1}{kq_{i_t}}\mathbf{u}_{i_t}\mathbf{u}_{i_t}^\top}^{\mathbf{Z}_t}, \quad \text{and} \quad \widehat{\mathbf{Z}} \overset{def}{=} \sum_{t=1}^{k} \overbrace{\frac{1}{kq_{j_t}}\mathbf{u}_{j_t}\mathbf{u}_{j_t}^\top}^{\widehat{\mathbf{Z}}_t}.$$

Note that $\|\mathbf{Z}_t\| = \frac{d}{k\,l_i}\|\mathbf{u}_{i_t}\|^2 = \frac{d}{k}$ and, similarly, $\|\widehat{\mathbf{Z}}_t\| = \frac{d}{k}$. Moreover,

$$\mathbb{E}[\mathbf{Z}] = \sum_{t=1}^{k}\left[\frac{1}{m}\sum_{i=1}^{m}\frac{1}{kq_i}\mathbf{u}_i\mathbf{u}_i^\top\right] = k\,\frac{1}{k}\frac{1}{m}\mathbf{U}^\top\mathbf{Q}_\pi\mathbf{U} = \frac{1}{m}\mathbf{U}^\top\mathbf{Q}_\pi\mathbf{U}.$$

Combining Chernoff's inequality with the reduction argument described in [20], for any λ, and $\theta > 0$,

$$\Pr\big(\lambda_{\max}(-\widehat{\mathbf{Z}}) \geq \lambda\big) \leq \mathrm{e}^{-\theta\lambda}\,\mathbb{E}\Big[\mathrm{tr}\big(\exp(\theta(-\widehat{\mathbf{Z}}))\big)\Big] \leq \mathrm{e}^{-\theta\lambda}\,\mathbb{E}\Big[\mathrm{tr}\big(\exp(\theta(-\mathbf{Z}))\big)\Big].$$

Using matrix Chernoff bound of [32] applied to $-\mathbf{Z}_1, ..., -\mathbf{Z}_k$ with appropriate θ, we have

$$\mathrm{e}^{-\theta\lambda}\,\mathbb{E}\Big[\mathrm{tr}\big(\exp(\theta(-\mathbf{Z}))\big)\Big] \leq d\,\exp\Big(-\frac{k}{16d}\Big), \quad \text{for} \quad \lambda = \frac{1}{2}\,\lambda_{\max}\Big(-\frac{1}{m}\mathbf{U}^\top\mathbf{Q}_\pi\mathbf{U}\Big) \leq -\frac{1}{4}.$$

Thus, there is a constant C_1 such that for $k \geq C_1\,d\ln(d/\delta)$, w.p. at least $1-\delta$ we have $\lambda_{\min}(\widehat{\mathbf{Z}}) \geq \frac{1}{4}$.

D.3 Proof of Lemma 14

We compute the moment generating function of the variable $R_t \sim \mathrm{Geom}(r_t)$, where $r_t = \frac{t-d}{t}$:

$$\mathbb{E}\big[\mathrm{e}^{\theta R_t}\big] = \frac{r_t}{1-(1-r_t)\mathrm{e}^{\theta}} = \frac{\frac{t-d}{t}}{1-\frac{d}{t}\,\mathrm{e}^{\theta}} = \frac{t-d}{t-d\,\mathrm{e}^{\theta}}.$$

Setting $\theta = \frac{1}{2d}$, we observe that $d\mathrm{e}^{\theta} \leq d+1$, and so $\mathbb{E}[\mathrm{e}^{\theta R_t}] \leq \frac{t-d}{t-d-1}$. Letting $\mu = \mathbb{E}[R]$, for any $a > 1$ using Markov's inequality we have

$$\Pr(R \geq a\mu) \leq \mathrm{e}^{-a\theta\mu}\,\mathbb{E}\big[\mathrm{e}^{\theta R}\big] \leq \mathrm{e}^{-a\theta\mu}\prod_{t=k+1}^{m}\frac{t-d}{t-d-1} = \mathrm{e}^{-a\theta\mu}\,\frac{m-d}{k-d}.$$

Note that using the bounds on the harmonic series we can estimate the mean:

$$\mu = d\sum_{t=k-d+1}^{m-d}\frac{1}{t} \geq d\,(\ln(m-d) - \ln(k-d) - 1) = d\,\ln\Big(\frac{m-d}{k-d}\Big) - d,$$

$$\text{so}\quad \mathrm{e}^{-a\theta\mu} \leq \mathrm{e}^{a/2}\,\exp\bigg(-\frac{a}{2}\ln\Big(\frac{m-d}{k-d}\Big)\bigg) = \mathrm{e}^{a/2}\Big(\frac{m-d}{k-d}\Big)^{-a/2}.$$

Putting the two inequalities together we obtain the desired tail bound.

E Experiments

We present experiments comparing leveraged volume sampling to standard volume sampling and to leverage score sampling, in terms of the total square loss suffered by the subsampled least-squares estimator. The three estimators can be summarized as follows:

$$\textit{volume sampling:}\quad \mathbf{w}_S^* = (\mathbf{X}_S)^{+}\mathbf{y}_S, \qquad \Pr(S) \sim \det(\mathbf{X}_S^{\top}\mathbf{X}_S), \quad S \in \binom{[n]}{k};$$

$$\textit{leverage score sampling:}\quad \mathbf{w}_\pi^* = (\mathbf{Q}_\pi^{1/2}\mathbf{X})^{+}\mathbf{Q}_\pi^{1/2}\mathbf{y}, \quad \Pr(\pi) = \prod_{i=1}^{k}\frac{l_{\pi_i}}{d}, \qquad \pi \in [n]^k;$$

$$\textit{leveraged volume sampling:}\quad \mathbf{w}_\pi^* = (\mathbf{Q}_\pi^{1/2}\mathbf{X})^{+}\mathbf{Q}_\pi^{1/2}\mathbf{y}, \quad \Pr(\pi) \sim \det(\mathbf{X}^{\top}\mathbf{Q}_\pi\mathbf{X})\prod_{i=1}^{k}\frac{l_{\pi_i}}{d}.$$

Both the volume sampling-based estimators are unbiased, however the leverage score sampling estimator is not. Recall that $\mathbf{Q}_\pi = \sum_{i=1}^{|\pi|} q_{\pi_i}^{-1}\mathbf{e}_{\pi_i}\mathbf{e}_{\pi_i}^{\top}$ is the selection and rescaling matrix as defined for q-rescaled volume sampling with $q_i = \frac{l_i}{d}$. For each estimator we plotted its average total loss, i.e., $\frac{1}{n}\|\mathbf{X}\mathbf{w} - \mathbf{y}\|^2$, for a range of sample sizes k, contrasted with the loss of the best least-squares estimator $\mathbf{w}^*$ computed from all data.

Dataset	Instances (n)	Features (d)
bodyfat	252	14
housing	506	13
mg	1,385	21
abalone	4,177	36
cpusmall	8,192	12
cadata	20,640	8
MSD	463,715	90

Table 1: Libsvm regression datasets [9] (to increase dimensionality of *mg* and *abalone*, we expanded features to all degree 2 monomials, and removed redundant ones).

Plots shown in Figures 1 and 2 were averaged over 100 runs, with shaded area representing standard error of the mean. We used seven benchmark datasets from the libsvm repository [9] (six in this section and one in Section 1), whose dimensions are given in Table 1. The results confirm that leveraged volume sampling is as good or better than either of the baselines for any sample size k. We can see that in some of the examples standard volume sampling exhibits bad behavior for larger sample sizes, as suggested by the lower bound of Theorem 1 (especially noticeable on *bodyfat* and *cpusmall* datasets). On the other hand, leverage score sampling exhibits poor performance for small sample sizes due to the coupon collector problem, which is most noticeable for *abalone* dataset, where we can see a very sharp transition after which leverage score sampling becomes effective. Neither of the variants of volume sampling suffers from this issue.

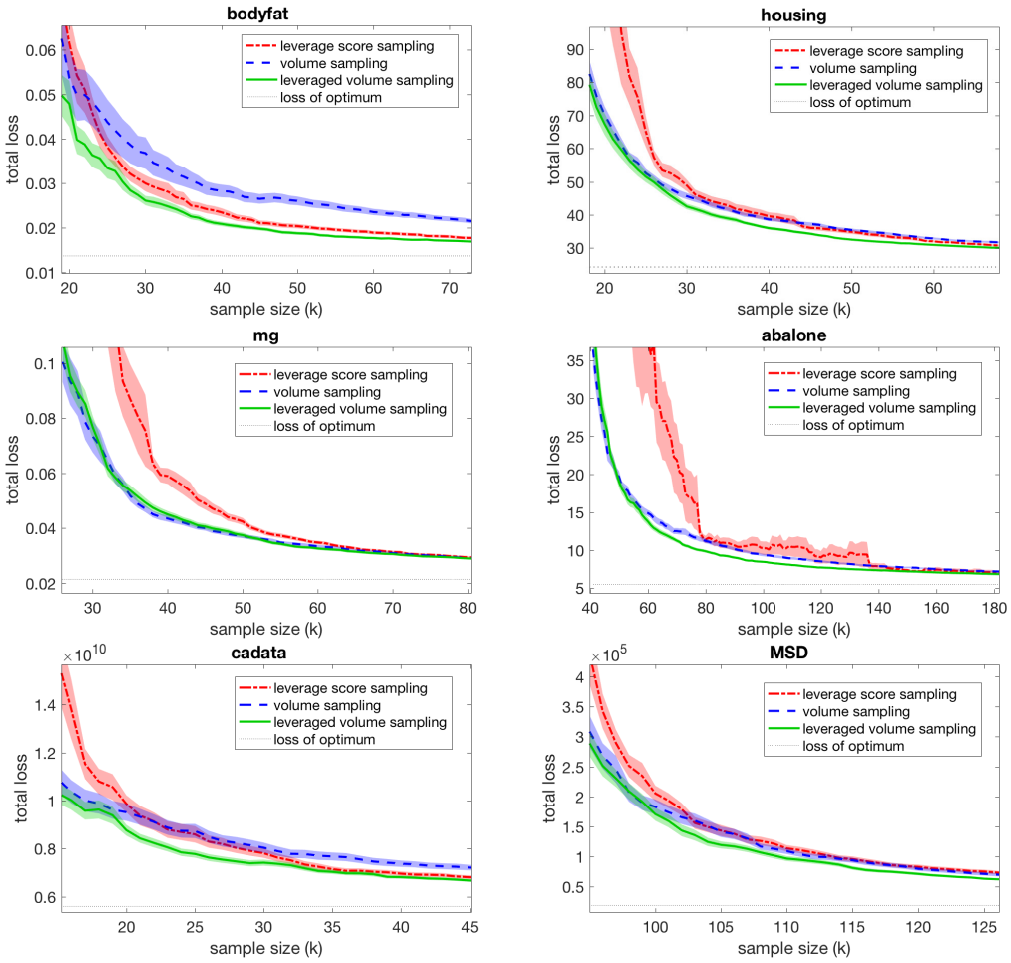

Figure 2: Comparison of loss of the subsampled estimator when using *leveraged volume sampling* vs using *leverage score sampling* and standard *volume sampling* on six datasets.

F Faster algorithm via approximate leverage scores

In some settings, the primary computational cost of deploying leveraged volume sampling is the preprocessing cost of computing exact laverage scores for matrix $\mathbf{X} \in \mathbb{R}^{n\times d}$, which takes $O(nd^2)$. There is a large body of work dedicated to fast estimation of leverage scores (see, e.g., [16, 27]), and in this section we examine how these approaches can be utilized to make leveraged volume sampling more efficient. The key challenge here is to show that the determinantal rejection sampling step remains effective when distribution q consists of approximate leverage scores. Our strategy, which is described in the algorithm *fast leveraged volume sampling*, will be to compute an approximate covariance matrix $\mathbf{A} = (1 \pm \epsilon)\mathbf{X}^\top\mathbf{X}$ and use it to compute the rescaling distribution $q_i \sim \mathbf{x}_i^\top \mathbf{A}^{-1}\mathbf{x}_i$. As we see in the lemma below, for sufficiently small ϵ, this rescaling still retains the runtime guarantee of determinantal rejection sampling from Theorem 6.

Fast leveraged volume sampling

Input: $\mathbf{X}\in\mathbb{R}^{n\times d}$, $k \geq d$, $\epsilon \geq 0$
Compute $\mathbf{A} = (1 \pm \epsilon)\,\mathbf{X}^\top\mathbf{X}$
Compute $\tilde{l}_i = (1 \pm \frac{1}{2})\,l_i \quad \forall_{i\in[n]}$
$s \leftarrow \max\{k,\, 8d^2\}$
repeat
  $\pi \leftarrow$ empty sequence
  while $|\pi| < s$
    Sample $i \sim (\tilde{l}_1, \ldots, \tilde{l}_n)$
    $a \sim \text{Bernoulli}\Big((1-\epsilon)\frac{\mathbf{x}_i^\top \mathbf{A}^{-1}\mathbf{x}_i}{2\tilde{l}_i}\Big)$
    if $a =$ true, **then** $\pi \leftarrow [\pi, i]$
  end
  $\mathbf{Q}_\pi \leftarrow \sum_{j=1}^{s} d\,(\mathbf{x}_{\pi_j}^\top \mathbf{A}^{-1}\mathbf{x}_{\pi_j})^{-1}\mathbf{e}_{\pi_j}\mathbf{e}_{\pi_j}^\top$
  Sample $Acc \sim \text{Bernoulli}\Big(\frac{\det(\frac{1}{s}\mathbf{X}^\top\mathbf{Q}_\pi\mathbf{X})}{\det(\mathbf{A})}\Big)$
until $Acc =$ true
$S \leftarrow \text{VolumeSample}\big((\mathbf{Q}_{[1..n]}^{1/2}\mathbf{X})_\pi, k\big)$
return π_S

Lemma 15 *Let $\mathbf{X} \in \mathbb{R}^{n\times d}$ be a full rank matrix, and suppose that matrix $\mathbf{A} \in \mathbb{R}^{d\times d}$ satisfies*

$$(1-\epsilon)\,\mathbf{X}^\top\mathbf{X} \preceq \mathbf{A} \preceq (1+\epsilon)\,\mathbf{X}^\top\mathbf{X}, \quad \textit{where} \quad \frac{\epsilon}{1-\epsilon} \leq \frac{1}{16d}.$$

Let $\pi_1, \ldots, \pi_s$ be sampled i.i.d. $\sim (\hat{l}_1, \ldots, \hat{l}_n)$, where $\hat{l}_i = \mathbf{x}_i^\top \mathbf{A}^{-1}\mathbf{x}_i$. If $s \geq 8d^2$, then

$$\textit{for} \quad \mathbf{Q}_\pi = \sum_{j=1}^{s} \frac{d}{\hat{l}_{\pi_j}} \mathbf{e}_{\pi_j}\mathbf{e}_{\pi_j}^\top, \qquad \frac{\det(\frac{1}{s}\mathbf{X}^\top\mathbf{Q}_\pi\mathbf{X})}{\det(\mathbf{A})} \leq 1 \quad \textit{and} \quad \mathbb{E}\left[\frac{\det(\frac{1}{s}\mathbf{X}^\top\mathbf{Q}_\pi\mathbf{X})}{\det(\mathbf{A})}\right] \geq \frac{3}{4}.$$

Proof of Lemma 15 follows along the same lines as the proof of Theorem 6. We can compute matrix $\mathbf{A}^{-1}$ efficiently in time $\widetilde{O}(nd + d^3/\epsilon^2)$ using a sketching technique called Fast Johnson-Lindenstraus Transform [1], as described in [16]. However, the cost of computing the entire rescaling distribution is still $O(nd^2)$. Standard techniques circumvent this issue by performing a second matrix sketch. We cannot afford to do that while at the same time preserving the sufficient quality of leverage score estimates needed for leveraged volume sampling. Instead, we first compute weak estimates $\tilde{l}_i = (1 \pm \frac{1}{2})l_i$ in time $\widetilde{O}(nd + d^3)$ as in [16], then use rejection sampling to sample from the more accurate leverage score distribution, and finally compute the correct rescaling coefficients just for the obtained sample. Note that having produced matrix $\mathbf{A}^{-1}$, computing a single leverage score estimate $\hat{l}_i$ takes $O(d^2)$. The proposed algorithm with high probability only has to compute $O(s)$ such estimates, which introduces an additional cost of $O(sd^2) = O((k+d^2)\,d^2)$. Thus, as long as $k = O(d^3)$, dominant cost of the overall procedure still comes from the estimation of matrix $\mathbf{A}$, which takes $\widetilde{O}(nd + d^5)$ when ϵ is chosen as in Lemma 15.

It is worth noting that *fast leveraged volume sampling* is a valid q-rescaled volume sampling distribution (and not an approximation of one), so the least-squares estimators it produces are exactly unbiased. Moreover, proofs of Theorems 8 and 9 can be straightforwardly extended to the setting where q is constructed from approximate leverage scores, so our loss bounds also hold in this case.