Correcting the bias in least squares regression with volume-rescaled sampling

Michał Dereziński
Department of Statistics
University of California, Berkeley
mderezin@berkeley.edu

Manfred K. Warmuth
Dept. of Computer Science
University of California, Santa Cruz
manfred@ucsc.edu

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

Abstract

Consider linear regression where the examples are generated by an unknown distribution on $\mathbb{R}^d \times \mathbb{R}$. Without any assumptions on the noise, the linear least squares solution for any i.i.d. sample will typically be biased w.r.t. the least squares optimum over the entire distribution. However, we show that if an i.i.d. sample of any size k is augmented by a certain small additional sample, then the solution of the combined sample becomes unbiased. We show this when the additional sample consists of d points drawn jointly according to the input distribution that is rescaled by the squared volume spanned by the points. Furthermore, we propose algorithms to sample from this volume-rescaled distribution when the data distribution is only known through an i.i.d sample.

1 INTRODUCTION

Unbiased estimators for linear regression are useful because averaging such estimators gives an unbiased estimator whose prediction variance vanishes as the number of averaged estimators increases. Such estimators might for example be produced in a distributed fashion from multiple small samples. In this paper we develop a unique method for correcting the bias of linear least squares estimators. Our main methodology for producing an unbiased estimator is volume sampling. For a fixed design matrix $\mathbf{X} \in \mathbb{R}^{n\times d}$, the most basic variant of volume sampling chooses a subset $S \subseteq \{1..n\}$ of dimension many rows (i.e. $|S| = d$) with probability proportional to the squared volume spanned by the rows, i.e. $\det(\mathbf{X}_S)^2$, where $\mathbf{X}_S$ is the sub-matrix of rows indexed by S. This procedure generalizes to sampling sets of any fixed size $k \geq d$ [2]:

$$P(S) \overset{def}{=} \frac{\det(\mathbf{X}_S^\top \mathbf{X}_S)}{\binom{n-d}{k-d}\det(\mathbf{X}^\top\mathbf{X})}. \tag{1}$$

Volume sampling has the property that for any design matrix $\mathbf{X}$ with n rows and any real response vector $\mathbf{y} \in \mathbb{R}^n$, the linear least squares solution for the subproblem $(\mathbf{X}_S, \mathbf{y}_S)$ is an unbiased estimator for the solution of the full problem $(\mathbf{X}, \mathbf{y})$ [8].

We propose the following previously unobserved alternate sampling method for size $k > d$ volume sampling: First volume sample a set $S_\circ$ of size d and then pad the sample with a uniform subset R of $k-d$ rows outside of $S_\circ$. Now the probability of the combined size k sample $S = S_\circ \cup R$ is again volume sampling (1):

$$P(S) = \sum_{\substack{S_\circ \subseteq S \\ |S_\circ| = d}} \underbrace{P\big(R = S\setminus S_\circ \mid S_\circ\big)}_{\frac{1}{\binom{n-d}{k-d}}} \underbrace{P(S_\circ)}_{\frac{\det(\mathbf{X}_{S_\circ})^2}{\det(\mathbf{X}^\top\mathbf{X})}} = \frac{\det(\mathbf{X}_S^\top\mathbf{X}_S)}{\binom{n-d}{k-d}\det(\mathbf{X}^\top\mathbf{X})},$$

where the equality is the Cauchy-Binet formula for determinants. Furthermore, we study a more general statistical learning model where the points come from an unknown probability distribution over $\mathbb{R}^d \times \mathbb{R}$, and the goal is to recover the least squares solution w.r.t. the distribution. In this paper we generalize volume sampling to this case by rescaling the i.i.d. sampling distribution by the squared volume of the sampled points.

The simplest way to obtain a linear least squares estimator in the statistical learning model is to find the linear least squares solution for a size k i.i.d. sample. Unfortunately such estimators are generally biased. Note that this is not the kind of bias that we deliberately impose with regularization to reduce the variance of the estimator. Rather, due to the random design, the least squares estimator is typically biased even when it is not regularized at all [16], and we have

Proceedings of the 22nd International Conference on Artificial Intelligence and Statistics (AISTATS) 2019, Naha, Okinawa, Japan. PMLR: Volume 89.

limited control over how large that bias may be (see Section 1.2 for a motivating example). However our alternate sampling procedure for volume sampling (discussed in the previous paragraph) implies the following strategy for correcting the bias: We show that if an i.i.d. sample of any size k is augmented with a size d volume-rescaled sample for this distribution, then the combined sample is a volume-rescaled sample of size $k+d$, and its linear least squares solution is an unbiased estimator of the optimum. In one dimension, this means that if an i.i.d. sample is augmented with just one example, where this additional example is drawn from a distribution whose marginal distribution on x is proportional to the original (unknown) marginal density times x^2, then the resulting least squares estimator becomes unbiased. Curiously enough, for the purpose of correcting the bias it does not matter whether the size d volume-rescaled sample was generated before or after the original size k i.i.d. sample was drawn, since they are independent of each other.

In addition to generalizing volume sampling to the continuous domain and showing that only a subsample of size d needs to be rescaled by the squared volume, we study the time and sample complexity of volume-rescaled sampling when the data distribution is only known through an i.i.d. sample. Specifically:

1. We extend *determinantal rejection sampling* [9] to arbitrary data distributions with bounded support, and our improved analysis reduces its time and sample complexity by a factor of d.

2. When the data distribution is Gaussian with unknown covariance, we propose a new algorithm with $O(d)$ sample complexity.

Related work. Discrete volume sampling of size $k \leq d$ was introduced to computer science literature by [11], with later algorithms by [10, 14]. The extension to sets of size $k > d$ is due to [2], with algorithms by [21, 8, 9], and additional applications in experimental design explored by [1, 24, 22]. Our alternate volume sampling procedure implies that the algorithms by [10, 14] can be used to volume sample larger sets at no additional cost. The unbiasedness of least squares estimators under volume sampling was explored by [8, 9], drawing on observations of [4].

For arbitrary data distributions, volume-rescaled sampling of size d is a special case of a determinantal point process (DPP) (see, e.g. [3, 15]). However for $k > d$ and arbitrary distributions, we are not aware of such sampling appearing in the literature. Related variants of discrete DPPs have been extensively explored in the machine learning community [19, 18, 20, 12, 5, 6].

Notations and assumptions. Throughout the paper, $(\mathbf{x}, y) \in \mathbb{R}^d \times \mathbb{R}$ is a random example drawn from some distribution D. We assume that the point $\mathbf{x}$ and the response y both have finite second moments, i.e. $\mathbb{E}[\|\mathbf{x}\|^2] < \infty$ and $\mathbb{E}[y^2] < \infty$. The marginal probability measure of $\mathbf{x}$ is denoted as $\mathrm{D}_{\mathcal{X}}$, while $\mathrm{D}_{\mathcal{X}}^k$ is the probability measure over $(\mathbb{R}^d)^k$ of k i.i.d. samples $(\mathbf{x}_1, \ldots, \mathbf{x}_k)$ drawn from $\mathrm{D}_{\mathcal{X}}$. We define $\mathbf{\Sigma}_{\mathrm{D}_{\mathcal{X}}} \overset{def}{=} \mathbb{E}[\mathbf{x}\mathbf{x}^\top] \in \mathbb{R}^{d\times d}$ and w.l.o.g. assume that it is invertible. Given a data sample $\mathbb{S} = \{(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_k, y_k)\}$, we denote the least squares estimators for $\mathbb{S}$ and D, respectively, as

$$\mathbf{w}^*(\mathbb{S}) \overset{def}{=} \operatorname*{argmin}_{\mathbf{w}} \sum_{(\mathbf{x}_i, y_i)\in\mathbb{S}} (\mathbf{x}_i^\top \mathbf{w} - y_i)^2 \quad \text{and}$$

$$\mathbf{w}_{\mathrm{D}}^* \overset{def}{=} \operatorname*{argmin}_{\mathbf{w}} \mathbb{E}_D\big[(\mathbf{x}^\top \mathbf{w} - y)^2\big] = \mathbf{\Sigma}_{\mathrm{D}_{\mathcal{X}}}^{-1} \mathbb{E}_D[\mathbf{x}\, y].$$

1.1 Statistical Results

Our results are centered around the following size k *joint sampling distribution.*

Definition 1 *Given distribution* $\mathrm{D}_{\mathcal{X}}$ *and any* $k \geq d$*, we define volume-rescaled size* k *sampling from* $\mathrm{D}_{\mathcal{X}}$ *as the following probability measure: For any event* $A \subseteq (\mathbb{R}^d)^k$ *measurable w.r.t.* $\mathrm{D}_{\mathcal{X}}^k$*, its probability is*

$$\mathrm{VS}_{\mathrm{D}_{\mathcal{X}}}^k(A) \overset{def}{=} \frac{\mathbb{E}_{\mathrm{D}_{\mathcal{X}}^k}\Big[\mathbf{1}_A \overbrace{\det\Big(\sum_{i=1}^k \mathbf{x}_i\mathbf{x}_i^\top\Big)}^{\text{rescaling factor}}\Big]}{d!\binom{k}{d}\det\big(\mathbf{\Sigma}_{\mathrm{D}_{\mathcal{X}}}\big)},$$

where $\mathbf{1}_A$ *is the indicator variable of event* A*.*

This distribution integrates to 1 over its domain $(\mathbb{R}^d)^k$ as a consequence of a continuous version of the classic Cauchy-Binet formula, which has appeared in the literature in various contexts (Lemma 7).

Although we define volume-rescaled sampling for any sample size $k \geq d$, we focus primarily on the special case of $k = d$ in the main results below. This is because we show that any $\mathrm{VS}_{\mathrm{D}_{\mathcal{X}}}^k$ can be decomposed into $\mathrm{VS}_{\mathrm{D}_{\mathcal{X}}}^d$ and $\mathrm{D}_{\mathcal{X}}^{k-d}$, the latter being the distribution of a size $k-d$ i.i.d. sample from $\mathrm{D}_{\mathcal{X}}$.

Theorem 1 *Let* $\mathbb{S} \sim \mathrm{D}_{\mathcal{X}}^{k-d}$ *and* $\mathbb{S}_\circ \sim \mathrm{VS}_{\mathrm{D}_{\mathcal{X}}}^d$*. Let* $\widetilde{\mathbb{S}} \in \mathbb{R}^{k\times d}$ *denote a random permutation of the points from* $\mathbb{S}$ *concatenated with* $\mathbb{S}_\circ$*, i.e.* $\widetilde{\mathbb{S}} = \sigma(\langle \mathbb{S}, \mathbb{S}_\circ \rangle)$*, where* σ *is a random permutation. Then* $\widetilde{\mathbb{S}} \sim \mathrm{VS}_{\mathrm{D}_{\mathcal{X}}}^k$*.*

Given the above decomposition, one may wonder what is the purpose of defining volume-rescaled sampling for any size $k > d$. In fact, we will see in the following sections that both in the proofs and in algorithms it is sometimes easier to work with $\mathrm{VS}_{\mathrm{D}_{\mathcal{X}}}^k$ rather than its

decomposed version. For example in the theorem below, we show that for any k, the least squares estimator computed on a volume-rescaled sample is unbiased. Despite the fact that continuous determinantal point processes have been studied extensively in the past, we were not able to find this result for arbitrary $\mathrm{D}_{\mathcal{X}}$ in the literature.

Theorem 2 *Consider the following distribution* $\mathrm{VS}_{\mathrm{D}}^{k}$ *on samples* $\mathbb{S} = \{(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_k, y_k)\}$ *of size* k*:*

$$\begin{aligned} &\text{Sample} && \mathbf{x}_1, \ldots, \mathbf{x}_k \;\sim\; \mathrm{VS}_{\mathrm{D}_{\mathcal{X}}}^{k}, \\ &\text{Query} && y_i \;\sim\; D_{\mathcal{Y}|\mathbf{x}=\mathbf{x}_i} \quad \forall_{i=1..k}. \end{aligned}$$

Then $\mathbb{E}_{\mathrm{VS}_{\mathrm{D}}^{k}}[\mathbf{w}^*(\mathbb{S})] = \mathbf{w}_{\mathrm{D}}^*$.

Combining Theorems 1 and 2, we conclude that an i.i.d. sample only needs to be augmented by a dimension-size volume-rescaled sample (i.e., $k = d$) so that the least squares estimator becomes unbiased.

Corollary 3 *Let* $\mathbb{S} = \{(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_k, y_k)\} \overset{\text{i.i.d.}}{\sim} D^k$*, for any* $k \geq 0$*. Consider the following procedure:*

$$\begin{aligned} &\text{Sample} && \widetilde{\mathbf{x}}_1, \ldots, \widetilde{\mathbf{x}}_d \;\sim\; \mathrm{VS}_{\mathrm{D}_{\mathcal{X}}}^{d}, \\ &\text{Query} && \widetilde{y}_i \;\sim\; D_{\mathcal{Y}|\mathbf{x}=\widetilde{\mathbf{x}}_i} \quad \forall_{i=1..d}. \end{aligned}$$

Then for $\mathbb{S}_\circ = \{(\widetilde{\mathbf{x}}_1, \widetilde{y}_1), \ldots, (\widetilde{\mathbf{x}}_d, \widetilde{y}_d)\}$*,*

$$\begin{aligned} &\mathbb{E}\big[\mathbf{w}^*(\langle \mathbb{S}, \mathbb{S}_\circ \rangle)\big] = \mathbb{E}_{\mathbb{S}\sim D^k}\big[\,\mathbb{E}_{\mathbb{S}_\circ \sim \mathrm{VS}_{\mathrm{D}}^{d}}\big[\,\mathbf{w}^*(\langle \mathbb{S}, \mathbb{S}_\circ \rangle)\,\big]\,\big] \\ &\text{(Theorem 1)} \quad = \mathbb{E}_{\widetilde{\mathbb{S}}\sim \mathrm{VS}_{\mathrm{D}}^{k+d}}\big[\,\mathbf{w}^*\big(\widetilde{\mathbb{S}}\big)\,\big] \\ &\text{(Theorem 2)} \quad = \mathbf{w}_{\mathrm{D}}^*. \end{aligned}$$

To put the above result in context, we note that in the fixed design case it was known that a single volume sampled subset $\mathbb{S}$ of any size $k \geq d$ produces an unbiased least squares estimator (see, e.g., [8]). However this required that all k points be sampled jointly from this special distribution. Thus, Corollary 3 says that volume sampling can be used to correct the bias in existing i.i.d. samples via sample augmentation (requiring labels/responses for only d additional points from $\mathrm{VS}_{\mathrm{D}_{\mathcal{X}}}^{d}$). This is important in active learning scenarios, where samples from $\mathrm{D}_{\mathcal{X}}$ (unlabeled data) are cheaper than draws from $D_{\mathcal{Y}|\mathbf{x}}$ (label queries). We also develop methods for generating the small sample from $\mathrm{VS}_{\mathrm{D}_{\mathcal{X}}}^{d}$ only using additional unlabeled samples from $\mathrm{D}_{\mathcal{X}}$ (see Section 1.3). Indeed, active learning was a motivation for volume sampling in previous works [8, 9].

1.2 A Simple Gaussian Experiment

The bias in least squares estimators is present even when input is a standard Gaussian. As an example, we let $d = 5$ and set:

$$\mathbf{x}^\top = (x_1, \ldots, x_d) \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, 1), \qquad y = \xi(\mathbf{x}) + \epsilon,$$

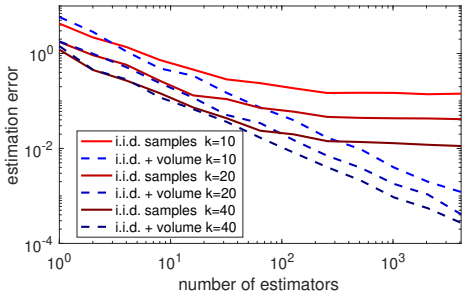

Figure 1: Averaging least squares estimators for Gaussian inputs with dimension $d = 5$.

where the response y is a non-linear function $\xi(\mathbf{x})$ plus independent white noise ϵ. Note that it is crucial that the response contains some non-linearity, and it is something that one would expect in real datasets. For the purposes of the experiment, we wish to make the least squares solution easy to compute algebraically, so we choose the following response model:

$$\xi(\mathbf{x}) = \sum_{i=1}^{d} x_i + \frac{x_i^3}{3}, \quad \epsilon \sim \mathcal{N}(0, 1).$$

We stress that there is nothing special about the choice of this response model other than the fact that it contains a non-linearity and it is easy to solve algebraically for $\mathbf{w}_{\mathrm{D}}^*$. We now compare the bias of the least squares estimator produced for this problem by i.i.d. sampling of k points, with that of an estimator computed from $k-d$ i.i.d. samples augmented by d volume samples (so that the total number of samples is the same in both cases). We used a special formula (Theorem 6 below) to produce the volume-rescaled samples when $\mathrm{D}_{\mathcal{X}}$ is Gaussian. Our strategy is to produce many such estimators $\widehat{\mathbf{w}}_1, \ldots, \widehat{\mathbf{w}}_T$ independently (e.g. by computing them in parallel on separate machines), and look at estimation error of the average of those estimators, i.e.

$$\text{estimation error:} \quad \left\| \left(\frac{1}{T} \sum_{t=1}^{T} \widehat{\mathbf{w}}_t \right) - \mathbf{w}_{\mathrm{D}}^* \right\|^2.$$

Figure 1 shows the above experiment for several values of k and a range of values of T (each presented data point is an average over 50 runs). Since the corrected estimator "i.i.d. + volume" is unbiased, the estimation error of the average estimator exhibits $\frac{1}{T}$ convergence to zero (regardless of k). This type of convergence appears as a straight line on the log-log plot. In contrast, the i.i.d. sampled estimator is biased for any sample size (although the bias decreases with k), and therefore the averaged estimator does not converge to the optimum.

1.3 Sampling Algorithms

To our knowledge, existing literature on algorithms for DPPs and volume sampling (other than the excep-

tions discussed below) generally assumes full or considerable knowledge of the distribution $\mathrm{D}_{\mathcal{X}}$, which often may not be the case in practice, for example when the data is coming in a stream, or is drawn from a larger population. In this work, we are primarily interested in the setting where access to distribution $\mathrm{D}_{\mathcal{X}}$ is limited to some approximate statistics plus the ability to draw i.i.d. samples from it. Two key concerns in this model are the time and sample complexities of volume-rescaled sampling for a given distribution $\mathrm{D}_{\mathcal{X}}$.

We first consider distributions $\mathrm{D}_{\mathcal{X}}$ with bounded support. We use a standard notion of *conditioning number* for multivariate distributions (see, e.g., [7]):

$$K_{\mathrm{D}_{\mathcal{X}}} \overset{def}{=} \sup_{\widetilde{\mathbf{x}} \in \mathrm{supp}(\mathrm{D}_{\mathcal{X}})} \widetilde{\mathbf{x}}^\top \mathbf{\Sigma}_{\mathrm{D}_{\mathcal{X}}}^{-1} \widetilde{\mathbf{x}}.$$

When $K_{\mathrm{D}_{\mathcal{X}}}$ is known to be bounded and we are given the exact knowledge of the covariance matrix $\mathbf{\Sigma}_{\mathrm{D}_{\mathcal{X}}}$, then it is possible to produce a volume-rescaled sample $\mathrm{VS}^d_{\mathrm{D}_{\mathcal{X}}}$ using a classical algorithm from the DPP literature described in [15] by employing rejection sampling (see also [3]). This approach requires $O(K_{\mathrm{D}_{\mathcal{X}}} \log(d))$ draws from $\mathrm{D}_{\mathcal{X}}$ and runs in time $O(K_{\mathrm{D}_{\mathcal{X}}} d^2 \log(d))$. However, sampled sets produced by that algorithm diverge from the desired distribution unless the given covariance matrix matches the true one exactly. This may be unrealistic when we do not have full access to the distribution $\mathrm{D}_{\mathcal{X}}$. Is it possible to sample from $\mathrm{VS}^d_{\mathrm{D}_{\mathcal{X}}}$ without the exact knowledge of $\mathbf{\Sigma}_{\mathrm{D}_{\mathcal{X}}}$?

We answer the question affirmatively. We show that a recently proposed algorithm from [9] for fixed design volume sampling can be adapted to arbitrary $\mathrm{D}_{\mathcal{X}}$ in such a way that it only requires an approximation of the covariance matrix $\mathbf{\Sigma}_{\mathrm{D}_{\mathcal{X}}}$, while still returning samples exactly from $\mathrm{VS}^d_{\mathrm{D}_{\mathcal{X}}}$. The original algorithm, called *determinantal rejection sampling*, samples from a given finite design matrix (i.e., a discrete distribution $\mathrm{D}_{\mathcal{X}}$ which is fully-known), but it was shown in [9] that the procedure only requires an approximation of the covariance matrix $\widehat{\mathbf{\Sigma}} = (1 \pm \epsilon)\mathbf{\Sigma}_{\mathrm{D}_{\mathcal{X}}}$, where $\epsilon = O(\frac{1}{d})$. We extend this algorithm to handle arbitrary distributions $\mathrm{D}_{\mathcal{X}}$, and also improve the analysis by reducing the required approximation quality to $\epsilon = O(\frac{1}{\sqrt{d}})$.

Theorem 4 *Given any $\widehat{\mathbf{\Sigma}} \in \mathbb{R}^{d\times d}$ s.t.*

$$(1-\epsilon)\mathbf{\Sigma}_{\mathrm{D}_{\mathcal{X}}} \preceq \widehat{\mathbf{\Sigma}} \preceq (1+\epsilon)\mathbf{\Sigma}_{\mathrm{D}_{\mathcal{X}}},$$
$$\text{where } \epsilon = \frac{1}{\sqrt{2d}} \text{ and } K \geq \frac{K_{\mathrm{D}_{\mathcal{X}}}}{1-\epsilon},$$

there is an algorithm which returns $\widetilde{\mathbf{x}}_1, \ldots, \widetilde{\mathbf{x}}_d \sim \mathrm{VS}^d_{\mathrm{D}_{\mathcal{X}}}$, and with probability at least $1-\delta$ its sample and time complexity is $O(Kd(\ln(\frac{1}{\delta}))^2)$ and $O(Kd^3(\ln(\frac{1}{\delta}))^2)$, respectively.

Remark 5 *Our $\epsilon = \frac{1}{\sqrt{2d}}$ condition improves the result from [9] (where $\epsilon = \frac{1}{16d}$ was used). When $\mathrm{D}_{\mathcal{X}}$ is given as a finite set of n vectors in $\mathbb{R}^d$, the main cost of volume sampling is an $\widetilde{O}(nd+d^3/\epsilon^2)$ preprocessing step of computing $\widehat{\mathbf{\Sigma}}$, where $\widetilde{O}(\cdot)$ hides $\mathrm{polylog}(n, d, 1/\epsilon, 1/\delta)$. Setting $\epsilon = \frac{1}{\sqrt{2d}}$ in Appendix F of [9], we reduce that cost from $\widetilde{O}(nd+d^5)$ to $\widetilde{O}(nd+d^4)$.*

In Section 4, we discuss how $\widehat{\mathbf{\Sigma}}$ can be obtained just by sampling from the distribution $\mathrm{D}_{\mathcal{X}}$, which requires $m = O(K_{\mathrm{D}_{\mathcal{X}}} d \ln(d))$ samples with high probability and time $O(md^2) = O(K_{\mathrm{D}_{\mathcal{X}}} d^3 \ln(d))$, nearly the same (up to log terms) as for the algorithm of Theorem 4 (here, the improved ϵ also plays a key role). An upper bound on the conditioning number $K_{\mathrm{D}_{\mathcal{X}}}$ is again needed.

The conditioning number $K_{\mathrm{D}_{\mathcal{X}}}$ can be much larger than the dimension d of the distribution $\mathrm{D}_{\mathcal{X}}$, so obtaining an appropriate estimate of $\mathbf{\Sigma}_{\mathrm{D}_{\mathcal{X}}}$ required for Theorem 4 may still be prohibitively expensive. Thus, it is natural to ask if there are some structural assumptions on distribution $\mathrm{D}_{\mathcal{X}}$ which can allow us to sample from $\mathrm{VS}^d_{\mathrm{D}_{\mathcal{X}}}$ without any estimate of the covariance matrix. In the following result, we exploit a connection between volume-rescaled sampling and the Wishart distribution to show that when $\mathbf{x}$ is a centered multivariate normal, then without any knowledge of $\mathbf{\Sigma}_{\mathrm{D}_{\mathcal{X}}}$, we can produce a volume-rescaled sample from only $2d+2$ samples of $\mathrm{D}_{\mathcal{X}}$ and in $O(d^3)$ running time.

Theorem 6 *Suppose that the point distribution $\mathrm{D}_{\mathcal{X}}$ is a Gaussian $\mathcal{N}(\mathbf{0}, \mathbf{\Sigma}_{\mathrm{D}_{\mathcal{X}}})$ and let $\mathbf{x}_1, \ldots, \mathbf{x}_{2d+2} \sim \mathrm{D}_{\mathcal{X}}^{2d+2}$. Then $\widetilde{\mathbf{x}}_1, \ldots, \widetilde{\mathbf{x}}_d \sim \mathrm{VS}^d_{\mathrm{D}_{\mathcal{X}}}$, where*

$$\widetilde{\mathbf{x}}_i \overset{def}{=} \left(\sum_{j=d+1}^{2d+2} \mathbf{x}_j \mathbf{x}_j^\top\right)^{\frac{1}{2}} \left(\sum_{j=1}^{d} \mathbf{x}_j \mathbf{x}_j^\top\right)^{-\frac{1}{2}} \mathbf{x}_i.$$

Note. For a positive definite matrix $\mathbf{A}$, we define $\mathbf{A}^{\frac{1}{2}}$ as the unique lower triangular matrix with positive diagonal entries s.t. $\mathbf{A}^{\frac{1}{2}}(\mathbf{A}^{\frac{1}{2}})^\top = \mathbf{A}$.

Finding other distribution families which allow for volume-rescaled sampling with bounded sample complexity is an interesting future research direction.

2 SAMPLE AUGMENTATION

Let $\mathbf{a}_i^\top$ denote the ith row of a matrix $\mathbf{A}$. First, we extend a classic lemma by [26], which was originally used to show the expected value of a metric in multivariate statistics known as "generalized variance".

Lemma 7 (based on [26]) *If the (transposed) rows of the random matrices $\mathbf{A}, \mathbf{B} \in \mathbb{R}^{k\times d}$ are sampled as pairs of vectors $(\mathbf{a}_1, \mathbf{b}_1), \ldots, (\mathbf{a}_k, \mathbf{b}_k)$ i.i.d. from a distribution over random vectors $(\mathbf{a}, \mathbf{b}) \in \mathbb{R}^{d\times 2}$ such that*

$\mathbb{E}[\mathbf{a}\mathbf{b}^\top]$ *exists, then*

$$\mathbb{E}\big[\det(\mathbf{A}^\top\mathbf{B})\big] = d!\binom{k}{d}\det\big(\mathbb{E}[\mathbf{a}\mathbf{b}^\top]\big).$$

The above result is slightly different than what was presented in [26] (the original one had $\mathbf{A}=\mathbf{B}$, and the sample mean was subtracted from the vectors before constructing the matrix $\mathbf{A}^\top\mathbf{A}$), but the analysis is similar (see proof in Appendix A). Note that for $\mathbf{a}=\mathbf{b}=\mathbf{x}$, Lemma 7 shows that $\mathrm{VS}^k_{\mathrm{D}_\mathcal{X}}$ integrates to 1, making it a well-defined probability distribution:

$$\mathbb{E}_{\mathrm{D}^k_\mathcal{X}}\bigg[\det\Big(\sum_{i=1}^k \mathbf{x}_i\mathbf{x}_i^\top\Big)\bigg] = d!\binom{k}{d}\det(\mathbf{\Sigma}_{\mathrm{D}_\mathcal{X}}).$$

The asymmetry of Lemma 7 is crucial for showing the unbiasedness property of volume-rescaled sampling.

Proof of Theorem 2 For $k=d$, the least squares estimator is simply the unique solution to a system of linear equations[1], so Cramer's rule states that the ith component of that solution is given by:

$$\big(\mathbf{w}^*(\mathbb{S})\big)_i = \frac{\det(\mathbf{X}\overset{i}{\leftarrow}\mathbf{y})}{\det(\mathbf{X})},$$

where $\mathbf{X}\overset{i}{\leftarrow}\mathbf{y}$ is matrix $\mathbf{X}$ with column i replaced by $\mathbf{y}$. We first prove unbiasedness of $\mathbf{w}^*(\mathbb{S})$ for samples of size d:

$$\begin{aligned}
\mathbb{E}_{\mathrm{VS}^d_\mathrm{D}}\big[\big(\mathbf{w}^*(\mathbb{S})\big)_i\big] &= \frac{\mathbb{E}_{D^d}[\det(\mathbf{X})^2\big(\mathbf{w}^*(\mathbb{S})\big)_i]}{d!\det(\mathbf{\Sigma}_{\mathrm{D}_\mathcal{X}})}\\
&= \frac{\mathbb{E}_{D^d}\big[\det(\mathbf{X})\det(\mathbf{X}\overset{i}{\leftarrow}\mathbf{y})\big]}{d!\det(\mathbf{\Sigma}_{\mathrm{D}_\mathcal{X}})}\\
\text{(Lemma 7)}\quad &= \frac{\det\big(\mathbb{E}_D[\mathbf{x}\,(\mathbf{x}\overset{i}{\leftarrow}y)^\top]\big)}{\det(\mathbf{\Sigma}_{\mathrm{D}_\mathcal{X}})}\\
&= \frac{\det\big(\mathbf{\Sigma}_{\mathrm{D}_\mathcal{X}}\overset{i}{\leftarrow}\mathbb{E}_D[\mathbf{x}\,y]\big)}{\det(\mathbf{\Sigma}_{\mathrm{D}_\mathcal{X}})} = \big(\mathbf{w}^*_\mathrm{D}\big)_i,
\end{aligned}$$

where we applied Lemma 7 to the pair of $d\times d$ matrices $\mathbf{A}=\mathbf{X}$ and $\mathbf{B}=\mathbf{X}\overset{i}{\leftarrow}\mathbf{y}$. The case of $k>d$ follows by induction based on a formula shown in [8]:

$$\begin{aligned}
\mathbb{E}_{\mathrm{VS}^k_\mathrm{D}}\big[\mathbf{w}^*(\mathbb{S})\big] &= \frac{\mathbb{E}_{\mathrm{D}^k}\big[\det(\mathbf{X}^\top\mathbf{X})\,\mathbf{w}^*(\mathbb{S})\big]}{d!\binom{k}{d}\det(\mathbf{\Sigma}_{\mathrm{D}_\mathcal{X}})}\\
&\overset{(1)}{=} \frac{\mathbb{E}_{\mathrm{D}^k}\Big[\frac{1}{k-d}\sum_{i=1}^k \det(\mathbf{X}_{-i}^\top\mathbf{X}_{-i})\,\mathbf{w}^*\big(\mathbb{S}\backslash\{(\mathbf{x}_i,y_i)\}\big)\Big]}{d!\binom{k}{d}\det(\mathbf{\Sigma}_{\mathrm{D}_\mathcal{X}})}\\
&= \frac{1}{k-d}\frac{\sum_{i=1}^k \mathbb{E}_{\mathrm{D}^k}\big[\det(\mathbf{X}_{-i}^\top\mathbf{X}_{-i})\mathbf{w}^*\big(\mathbb{S}\backslash\{(\mathbf{x}_i,y_i)\}\big)\big]}{d!\binom{k}{d}\det(\mathbf{\Sigma}_{\mathrm{D}_\mathcal{X}})}\\
&\overset{(2)}{=} \frac{k}{k-d}\frac{d!\binom{k-1}{d}}{d!\binom{k}{d}}\mathbb{E}_{\mathrm{VS}^{k-1}_\mathrm{D}}\big[\mathbf{w}^*(\mathbb{S})\big] = \mathbb{E}_{\mathrm{VS}^{k-1}_\mathrm{D}}\big[\mathbf{w}^*(\mathbb{S})\big],
\end{aligned}$$

where $\mathbf{X}_{-i}$ denotes matrix $\mathbf{X}$ without the ith row, (1) follows from the formula shown in [8] (given in Lemma 15 of Appendix A), while (2) follows because the samples $\mathbf{x}_1,\dots,\mathbf{x}_k \sim \mathrm{D}^k_\mathcal{X}$ are exchangeable, i.e. $\mathbf{x}_1,\dots,\cancel{\mathbf{x}_i},\dots,\mathbf{x}_k$ is distributed identically to $\mathbf{x}_1,\dots,\mathbf{x}_{k-1}$. ■

[1] Unless $\det(\mathbf{X})=0$, in which case we let $\mathbf{w}^*(\mathbb{S})=\mathbf{X}^+\mathbf{y}$.

Finally, our key observation given in Theorem 1 is that size k volume-rescaled sampling can be decomposed into size d volume-rescaled sampling plus i.i.d. sampling of $k-d$ points. Note that a version of this already occurs for discrete volume sampling (see Section 1). However it was not previously known even in that case.

Proof of Theorem 1 Let $\mathrm{DVS}^k_{\mathrm{D}_\mathcal{X}}$ denote the distribution of a matrix $\mathbf{X}\in\mathbb{R}^{k\times d}$ whose transposed rows are $\{\mathbf{x}_1,\dots,\mathbf{x}_k\} = \sigma(\langle\mathbb{S},\mathbb{S}_\circ\rangle)$. The probability of a measurable event A w.r.t. $\mathrm{DVS}^k_{\mathrm{D}_\mathcal{X}}$ is:

$$\begin{aligned}
\mathbb{E}_{\mathrm{DVS}^k_{\mathrm{D}_\mathcal{X}}}\big[\mathbf{1}_A\big] &= \frac{1}{\binom{k}{d}}\sum_{T\subseteq[k]:\,|T|=d}\frac{\mathbb{E}_{\mathrm{D}^k_\mathcal{X}}[\mathbf{1}_A\det(\mathbf{X}_T^\top\mathbf{X}_T)]}{d!\det(\mathbf{\Sigma}_{\mathrm{D}_\mathcal{X}})}\\
&= \frac{1}{d!\binom{k}{d}\det(\mathbf{\Sigma}_{\mathrm{D}_\mathcal{X}})}\mathbb{E}_{\mathrm{D}^k_\mathcal{X}}\bigg[\mathbf{1}_A\sum_{T\subseteq[k]:\,|T|=d}\det(\mathbf{X}_T^\top\mathbf{X}_T)\bigg]\\
&\overset{(*)}{=} \frac{1}{d!\binom{k}{d}\det(\mathbf{\Sigma}_{\mathrm{D}_\mathcal{X}})}\mathbb{E}_{\mathrm{D}^k_\mathcal{X}}\big[\mathbf{1}_A\det(\mathbf{X}^\top\mathbf{X})\big]\\
&= \mathbb{E}_{\mathrm{VS}^k_{\mathrm{D}_\mathcal{X}}}[\mathbf{1}_A],
\end{aligned}$$

where $[k]=\{1..k\}$, matrix $\mathbf{X}_T$ consists of the rows of $\mathbf{X}$ indexed by set T, and $(*)$ follows from the Cauchy-Binet formula. ■

3 VOLUME-RESCALED GAUSSIAN

In this section, we obtain a simple formula for producing volume-rescaled samples when $\mathrm{D}_\mathcal{X}$ is a centered multivariate Gaussian with any (non-singular) covariance matrix. We achieve this by making a connection to the Wishart distribution. Thus, for this section, assume that $\mathbf{x}\sim\mathcal{N}(\mathbf{0},\mathbf{\Sigma}_{\mathrm{D}_\mathcal{X}})$, and let $\mathbf{x}_1,\dots,\mathbf{x}_k\sim\mathrm{D}^k_\mathcal{X}$ be the transposed rows of matrix $\mathbf{X}$. Then matrix $\mathbf{\Sigma}=\mathbf{X}^\top\mathbf{X}\in\mathbb{R}^{d\times d}$ is distributed according to Wishart distribution $W_d(k,\mathbf{\Sigma}_{\mathrm{D}_\mathcal{X}})$ with k degrees of freedom. The density function of this random matrix is proportional to $\det(\mathbf{\Sigma})^{(k-d-1)/2}\exp(-\frac{1}{2}\mathrm{tr}(\mathbf{\Sigma}_{\mathrm{D}_\mathcal{X}}^{-1}\mathbf{\Sigma}))$. On the other hand, if $\widetilde{\mathbf{\Sigma}}=\widetilde{\mathbf{X}}^\top\widetilde{\mathbf{X}}$ is constructed from vectors $\widetilde{\mathbf{x}}_1,\dots,\widetilde{\mathbf{x}}_k\sim\mathrm{VS}^k_{\mathrm{D}_\mathcal{X}}$, then its density function is multiplied by an additional $\det(\widetilde{\mathbf{\Sigma}})$, thus increasing the value of k in the exponent of the determinant. This observation leads to the following result:

Theorem 8 *If* $\mathrm{D}_{\mathcal{X}} \sim \mathcal{N}(\mathbf{0}, \mathbf{\Sigma}_{\mathrm{D}_{\mathcal{X}}})$ *and* $\widetilde{\mathbf{x}}_1, \ldots, \widetilde{\mathbf{x}}_k \sim \mathrm{VS}^k_{\mathrm{D}_{\mathcal{X}}}$ *are rows of a random matrix* $\widetilde{\mathbf{X}} \in \mathbb{R}^{k \times d}$*, then*

$$\widetilde{\mathbf{X}}^\top \widetilde{\mathbf{X}} \sim W_d(k+2, \mathbf{\Sigma}_{\mathrm{D}_{\mathcal{X}}}).$$

Proof Let $\mathbf{\Sigma} = \mathbf{X}^\top \mathbf{X} \sim W_d(k, \mathbf{\Sigma}_{\mathrm{D}_{\mathcal{X}}})$ and $\widetilde{\mathbf{\Sigma}} \sim W_d(k+2, \mathbf{\Sigma}_{\mathrm{D}_{\mathcal{X}}})$. For any measurable event A over the random matrix $\widetilde{\mathbf{X}}^\top \widetilde{\mathbf{X}}$, we have

$$\begin{aligned}\Pr\big(\widetilde{\mathbf{X}}^\top \widetilde{\mathbf{X}} \in A\big) &= \frac{\mathbb{E}[\mathbf{1}_{[\mathbf{X}^\top \mathbf{X} \in A]} \det(\mathbf{X}^\top \mathbf{X})]}{\mathbb{E}[\det(\mathbf{X}^\top \mathbf{X})]} \\ &= \frac{\mathbb{E}[\mathbf{1}_{[\mathbf{\Sigma} \in A]} \det(\mathbf{\Sigma})]}{\mathbb{E}[\det(\mathbf{\Sigma})]} \overset{(*)}{=} \Pr\big(\widetilde{\mathbf{\Sigma}} \in A\big),\end{aligned}$$

where $(*)$ follows because the density function of Wishart distribution $\widetilde{\mathbf{\Sigma}} \sim W_d(k+2, \mathbf{\Sigma}_{\mathrm{D}_{\mathcal{X}}})$ is proportional to $\det(\widetilde{\mathbf{\Sigma}}) \det(\widetilde{\mathbf{\Sigma}})^{(k-d-1)/2} \exp(-\frac{1}{2}\mathrm{tr}(\mathbf{\Sigma}^{-1}_{\mathrm{D}_{\mathcal{X}}} \widetilde{\mathbf{\Sigma}}))$. ■

This gives us an easy way to produce the total covariance matrix $\widetilde{\mathbf{X}}^\top \widetilde{\mathbf{X}}$ of volume-rescaled samples in the Gaussian case. We next show that the individual vectors can also be recovered easily.

Proof of Theorem 6 The proof relies on the following two lemmas.

Lemma 9 *For any* $\mathbf{\Sigma} \in \mathbb{R}^{d \times d}$*, the conditional distribution of* $\widetilde{\mathbf{X}} \sim \mathrm{VS}^k_{\mathrm{D}_{\mathcal{X}}}$ *given* $\widetilde{\mathbf{X}}^\top \widetilde{\mathbf{X}} = \mathbf{\Sigma}$ *is the same as the conditional distribution of* $\mathbf{X} \sim \mathrm{D}^k_{\mathcal{X}}$ *given* $\mathbf{X}^\top \mathbf{X} = \mathbf{\Sigma}$*.*

While this lemma (proven in Appendix B) relies primarily on the definition of conditional probability, the second one uses properties of the matrix variate Beta and Dirichlet distributions.

Lemma 10 *For* $\mathbf{\Sigma} \in \mathbb{R}^{d \times d}$ *and vectors* $\mathbf{x}_1, \ldots, \mathbf{x}_k \sim \mathcal{N}(\mathbf{0}, \mathbf{\Sigma}_{\mathrm{D}_{\mathcal{X}}})$ *forming the transposed rows of a matrix* $\mathbf{X}$*, let*

$$\widetilde{\mathbf{x}}_i = \mathbf{\Sigma}^{\frac{1}{2}} (\mathbf{X}^\top \mathbf{X})^{-\frac{1}{2}} \mathbf{x}_i.$$

Then $\widetilde{\mathbf{x}}_1, \ldots, \widetilde{\mathbf{x}}_k$ *are jointly distributed as* k *Gaussians* $\mathcal{N}(\mathbf{0}, \mathbf{\Sigma}_{\mathrm{D}_{\mathcal{X}}})$ *conditioned on* $\widetilde{\mathbf{X}}^\top \widetilde{\mathbf{X}} = \mathbf{\Sigma}$*.*

Putting Theorem 8 together with the two lemmas, we observe that for any $k \geq d$, constructing $\mathbf{\Sigma} \sim W_d(k+2, \mathbf{\Sigma}_{\mathrm{D}_{\mathcal{X}}})$, and plugging it into Lemma 10, we obtain that $\widetilde{\mathbf{x}}_1, \ldots \widetilde{\mathbf{x}}_k \sim \mathrm{VS}^k_{\mathrm{D}_{\mathcal{X}}}$, completing the proof of Theorem 6. ■

We conclude this section with the proof of Lemma 10, which demonstrates an interesting application for classical results in matrix variate statistics.

Proof of Lemma 10 Let $\mathbf{\Sigma}_1 \sim W_d(k_1, \mathbf{\Sigma}_{\mathrm{D}_{\mathcal{X}}})$ and $\mathbf{\Sigma}_2 \sim W_d(k_2, \mathbf{\Sigma}_{\mathrm{D}_{\mathcal{X}}})$ be independent Wishart matrices (where $k_1 + k_2 \geq d$). Then matrix

$$\mathbf{U} = (\mathbf{\Sigma}_1 + \mathbf{\Sigma}_2)^{-\frac{1}{2}} \mathbf{\Sigma}_1 \big((\mathbf{\Sigma}_1 + \mathbf{\Sigma}_2)^{-\frac{1}{2}}\big)^\top$$

is matrix variate beta distributed, written as $\mathbf{U} \sim B_d(k_1, k_2)$. The following was shown by [23]:

Lemma 11 ([23], Lemma 3.5) *If* $\mathbf{\Sigma} \sim W_d(k, \mathbf{\Sigma}_{\mathrm{D}_{\mathcal{X}}})$ *is distributed independently of* $\mathbf{U} \sim B_d(k_1, k_2)$*, and if* $k = k_1 + k_2$*, then*

$$\mathbf{B} = \mathbf{\Sigma}^{\frac{1}{2}} \mathbf{U} \big(\mathbf{\Sigma}^{\frac{1}{2}}\big)^\top \quad \textit{and} \quad \mathbf{C} = \mathbf{\Sigma}^{\frac{1}{2}} (\mathbf{I} - \mathbf{U}) \big(\mathbf{\Sigma}^{\frac{1}{2}}\big)^\top$$

are independently distributed and $\mathbf{B} \sim W_d(k_1, \mathbf{\Sigma}_{\mathrm{D}_{\mathcal{X}}})$*,* $\mathbf{C} \sim W_d(k_2, \mathbf{\Sigma}_{\mathrm{D}_{\mathcal{X}}})$*.*

Now, suppose that we are given a matrix $\mathbf{\Sigma} \sim W_d(k, \mathbf{\Sigma}_{\mathrm{D}_{\mathcal{X}}})$. We can decompose it into components of degree one via a splitting procedure described in [23], namely taking $\mathbf{U}_1 \sim B_d(1, k-1)$ and computing $\mathbf{B}_1 \sim \mathbf{\Sigma}^{\frac{1}{2}} \mathbf{U}_1 \big(\mathbf{\Sigma}^{\frac{1}{2}}\big)^\top$, $\mathbf{C}_1 = \mathbf{\Sigma} - \mathbf{\Sigma}_1$ as in Lemma 11, then recursively repeating the procedure on $\mathbf{C}_1$ (instead of $\mathbf{\Sigma}$) with $\mathbf{U}_2 \sim B_d(1, k-2)$, …, until we get k Wishart matrices of degree one summing to $\mathbf{\Sigma}$:

$$\begin{aligned}
\mathbf{B}_1 &= \mathbf{\Sigma}^{\frac{1}{2}} \mathbf{U}_1 \big(\mathbf{\Sigma}^{\frac{1}{2}}\big)^\top \\
\mathbf{B}_2 &= \underbrace{\mathbf{\Sigma}^{\frac{1}{2}} (1 - \mathbf{U}_1)^{\frac{1}{2}}}_{\mathbf{C}_1^{1/2}} \mathbf{U}_2 \underbrace{\big((\mathbf{I} - \mathbf{U}_1)^{\frac{1}{2}}\big)^\top \big(\mathbf{\Sigma}^{\frac{1}{2}}\big)^\top}_{(\mathbf{C}_1^{1/2})^\top} \\
&\vdots \\
\mathbf{B}_k &= \underbrace{\mathbf{\Sigma}^{\frac{1}{2}} (1 - \mathbf{U}_{k-1})^{\frac{1}{2}} \ldots}_{\mathbf{C}_{k-1}^{1/2}} \mathbf{U}_k \underbrace{\ldots \big((1 - \mathbf{U}_{k-1})^{\frac{1}{2}}\big)^\top \big(\mathbf{\Sigma}^{\frac{1}{2}}\big)^\top}_{(\mathbf{C}_{k-1}^{1/2})^\top}.
\end{aligned}$$

The above collection of matrices can be described more simply via the matrix variate Dirichlet distribution. Given independent matrices $\mathbf{\Sigma}_i \sim W_d(k_i, \mathbf{\Sigma}_{\mathrm{D}_{\mathcal{X}}})$ for $i = 1..s$, the matrix variate Dirichlet distribution $D_d(k_1, \ldots, k_s)$ corresponds to a sequence of matrices

$$\mathbf{V}_i = \mathbf{\Sigma}^{-\frac{1}{2}} \mathbf{\Sigma}_i \big(\mathbf{\Sigma}^{-\frac{1}{2}}\big)^\top, \; i = 1..s, \quad \mathbf{\Sigma} = \sum_{i=1}^{s} \mathbf{\Sigma}_i.$$

Now, Theorem 6.3.14 from [13] states that matrices $\mathbf{B}_i$ defined recursively as above can also be written as

$$\mathbf{B}_i = \mathbf{\Sigma}^{\frac{1}{2}} \mathbf{V}_i \big(\mathbf{\Sigma}^{\frac{1}{2}}\big)^\top, \quad (\mathbf{V}_1, \ldots, \mathbf{V}_k) \sim D_d(1, \ldots, 1).$$

In particular, we can construct them as

$$\mathbf{B}_i = \widetilde{\mathbf{x}}_i \widetilde{\mathbf{x}}_i^\top = \mathbf{\Sigma}^{\frac{1}{2}} (\mathbf{X}^\top \mathbf{X})^{-\frac{1}{2}} \mathbf{x}_i \mathbf{x}_i^\top \big((\mathbf{X}^\top \mathbf{X})^{-\frac{1}{2}}\big)^\top \big(\mathbf{\Sigma}^{\frac{1}{2}}\big)^\top.$$

Note that since matrix $\mathbf{\Sigma}$ is independent of vectors $\mathbf{x}_i$, we can condition on it without altering the distribution of the vectors. It remains to observe that the conditional distribution of matrix $\mathbf{B}_i$ determines the distribution of $\widetilde{\mathbf{x}}_i$ up to multiplying by ± 1, and since both $\widetilde{\mathbf{x}}_i$ and $-\widetilde{\mathbf{x}}_i$ are identically distributed, we recover the correct distribution of $\widetilde{\mathbf{x}}_1, \ldots, \widetilde{\mathbf{x}}_k$ conditioned on $\widetilde{\mathbf{X}}^\top \widetilde{\mathbf{X}} = \mathbf{\Sigma}$, completing the proof. ■

4 GENERAL ALGORITHM

In this section, we present a general algorithm for volume-rescaled sampling, which uses approximate leverage score sampling to generate a larger pool of points from which the smaller volume-rescaled sample can be drawn. The method relies on a technique called "determinantal rejection sampling", introduced recently in [9] for a variant of volume sampling of finite subsets of points from a fixed set. Also, as in [9] our algorithm uses the most standard volume sampling distribution (see (1) and the associated discussion in the introduction) as a subroutine which samples a subset of points/rows from a fixed set. This is done via an efficient implementation of "reverse iterative sampling" [8] (See Algorithm 2 for a high-level description of this sampling method). Curiously enough, the efficient implementation of reverse iterative sampling given by [8] (denoted here as "VolSamp($\{\mathbf{x}_1,\dots,\mathbf{x}_n\},k$)" and not repeated here for lack of space) is again based on rejection sampling: It samples a set of k points out of n in time $O(nd^2)$ (independent of k). The runtime bound for this implementation only holds with high probability because of its use of rejection sampling.

For our algorithm we assume that an estimate $\widehat{\mathbf{\Sigma}} \approx \mathbf{\Sigma}_{\mathrm{D}_\mathcal{X}}$ of the covariance matrix is available, along with an upper-bound on the conditioning number.

Algorithm 1 **Determinantal rejection sampling for arbitrary distributions** $\mathrm{D}_\mathcal{X}$

Input: $\widehat{\mathbf{\Sigma}}, K, t$
repeat
  $k \to 0$
  while $k < t$
    Sample $\mathbf{x} \sim \mathrm{D}_\mathcal{X}$
    $a \sim \mathrm{Bernoulli}\Big(\min\big\{1,\ \frac{\mathbf{x}^\top\widehat{\mathbf{\Sigma}}^{-1}\mathbf{x}}{K}\big\}\Big)$
    if $a=$true, **then**
      $k \leftarrow k+1$
      $\mathbf{x}_k \leftarrow \mathbf{x}$
      $\widetilde{\mathbf{x}}_k \leftarrow \frac{\sqrt{d}}{\sqrt{\mathbf{x}_k^\top\widehat{\mathbf{\Sigma}}^{-1}\mathbf{x}_k}}\,\mathbf{x}_k$
    end
  end
  $\widetilde{\mathbf{\Sigma}} \leftarrow \frac{1}{t}\sum_{j=1}^t \widetilde{\mathbf{x}}_j\widetilde{\mathbf{x}}_j^\top$
  Sample $Acc \sim \mathrm{Bernoulli}\big(\min\{1,\ \det(\widetilde{\mathbf{\Sigma}}\widehat{\mathbf{\Sigma}}^{-1})\}\big)$
until $Acc =$ true
$\{\widetilde{\mathbf{x}}_{i_1},\dots,\widetilde{\mathbf{x}}_{i_d}\} \leftarrow \mathrm{VolSamp}\big(\{\widetilde{\mathbf{x}}_1,\dots,\widetilde{\mathbf{x}}_t\},d\big)$
return $\mathbf{x}_{i_1},\dots,\mathbf{x}_{i_d}$

Algorithm 1 has one additional hyperparameter t, which controls the number of inner-loop iterations. Our analysis works for any $t > d^2$, although for simplicity we use $t = 2d^2$ in the main result.

Algorithm 2 **Reverse iterative sampling [8]**

Input $\{\mathbf{x}_1,\dots,\mathbf{x}_n\} \subset \mathbb{R}^d$ and $k \geq d$
  $S \leftarrow \{1..n\}$
  while $|S| > k$
    For each $i \in S$: $q_i \leftarrow \frac{\det(\sum_{j\in S\setminus i}\mathbf{x}_j\mathbf{x}_j^\top)}{(|S|-d)\det(\sum_{j\in S}\mathbf{x}_j\mathbf{x}_j^\top)}$
    Sample i from distribution $(q_i)_{i\in S}$
    $S \leftarrow S\setminus\{i\}$
  end
return $\{\mathbf{x}_i\}_{i\in S}$

Our analysis of Algorithm 1 uses the following two lemmas, both of which are extensions of results from [9].

Lemma 12 *For $\widehat{\mathbf{\Sigma}} \succ 0$, let $l_{\widehat{\mathbf{\Sigma}}}(\mathbf{x}) = \mathbf{x}^\top\widehat{\mathbf{\Sigma}}^{-1}\mathbf{x}$. Define the following probability measure over $\mathbb{R}^d$:*

$$\mathrm{Lev}_{\widehat{\mathbf{\Sigma}},\mathcal{X}}(A) \overset{def}{=} \mathbb{E}_{\mathrm{D}_\mathcal{X}}\bigg[\mathbf{1}_A\ \frac{l_{\widehat{\mathbf{\Sigma}}}(\mathbf{x})}{\mathrm{tr}(\mathbf{\Sigma}_{\mathrm{D}_\mathcal{X}}\widehat{\mathbf{\Sigma}}^{-1})}\bigg].$$

If $\mathbf{x}_1,\dots,\mathbf{x}_t \overset{\text{i.i.d.}}{\sim} \mathrm{Lev}_{\widehat{\mathbf{\Sigma}},\mathcal{X}}$, and $\widetilde{\mathbf{x}}_i = \frac{\sqrt{d}}{\sqrt{l_{\widehat{\mathbf{\Sigma}}}(\mathbf{x}_i)}}\,\mathbf{x}_i$, then

$$\det(\widetilde{\mathbf{\Sigma}}\widehat{\mathbf{\Sigma}}^{-1}) \leq 1, \quad \textit{where} \quad \widetilde{\mathbf{\Sigma}} = \frac{1}{t}\sum_{i=1}^t \widetilde{\mathbf{x}}_i\widetilde{\mathbf{x}}_i^\top,$$

$$\textit{and} \quad \mathbb{E}\big[\det(\widetilde{\mathbf{\Sigma}}\widehat{\mathbf{\Sigma}}^{-1})\big] \geq \Big(1-\frac{d^2}{t}\Big)\frac{\det(\mathbf{\Sigma}_{\mathrm{D}_\mathcal{X}}\widehat{\mathbf{\Sigma}}^{-1})}{(\frac{1}{d}\mathrm{tr}(\mathbf{\Sigma}_{\mathrm{D}_\mathcal{X}}\widehat{\mathbf{\Sigma}}^{-1}))^d}.$$

Lemma 13 *Let $\mathbf{x}_1,\dots,\mathbf{x}_k \sim \mathrm{VS}^k_{\mathrm{D}_\mathcal{X}}$ be a volume-rescaled sample, and suppose that $\{\mathbf{x}_{i_1},\dots,\mathbf{x}_{i_d}\}$ is a subset produced from it by standard volume sampling, i.e. by calling* $\mathrm{VolSamp}(\{\mathbf{x}_1,\dots,\mathbf{x}_k\},d)$. *Then $\mathbf{x}_{i_1},\dots,\mathbf{x}_{i_d} \sim \mathrm{VS}^d_{\mathrm{D}_\mathcal{X}}$.*

We now show that Algorithm 1 with $t = 2d^2$ satisfies the conditions of Theorem 4. Our key contribution compared to the analysis of [9] is the use of the Kantorovich inequality, which allows us to significantly relax the ϵ-approximation condition on $\widehat{\mathbf{\Sigma}}$.

Proof of Theorem 4 From the assumptions, we have

$$K \geq \frac{K_{\mathrm{D}_\mathcal{X}}}{1-\epsilon} \geq \max_{\widetilde{\mathbf{x}}\in\mathrm{supp}(\mathrm{D}_\mathcal{X})} \widetilde{\mathbf{x}}^\top\widehat{\mathbf{\Sigma}}^{-1}\widetilde{\mathbf{x}},$$

so the sequence $\widetilde{\mathbf{x}}_1,\dots,\widetilde{\mathbf{x}}_t$ obtained by the algorithm at the point of exiting the **while** loop is distributed as in Lemma 12, and let $D_{\widetilde{\mathcal{X}}}$ be the distribution of one such vector. The lemma ensures that $\det\big(\widetilde{\mathbf{\Sigma}}\widehat{\mathbf{\Sigma}}^{-1}\big) \leq 1$ is a valid Bernoulli success probability so after exiting the **repeat** loop, $\widetilde{\mathbf{x}}_1,\dots,\widetilde{\mathbf{x}}_t$ is distributed so that the probability of any event A is proportional to

$$\mathbb{E}_{D^t_{\widetilde{\mathcal{X}}}}\bigg[\mathbf{1}_A\frac{\det(\widetilde{\mathbf{\Sigma}})}{\det(\widehat{\mathbf{\Sigma}})}\bigg] \propto \mathbb{E}_{D^t_{\widetilde{\mathcal{X}}}}\bigg[\mathbf{1}_A\det\Big(\sum_{i=1}^t \widetilde{\mathbf{x}}_i\widetilde{\mathbf{x}}_i^\top\Big)\bigg] \propto \mathrm{VS}^t_{D_{\widetilde{\mathcal{X}}}},$$

i.e., volume-rescaled sampling from $D_{\widetilde{\mathcal{X}}}$. Now Lemma 13 implies that $\widetilde{\mathbf{x}}_{i_1}, \ldots, \widetilde{\mathbf{x}}_{i_d} \sim \mathrm{VS}^d_{D_{\widetilde{\mathcal{X}}}}$. In particular, it means that the distribution of $\mathbf{x}_{i_1}, \ldots, \mathbf{x}_{i_d}$ is the same for any choice of $t \geq d$. We use this observation to compute the probability of an event A w.r.t. sampling of $\mathbf{x}_{i_1}, \ldots, \mathbf{x}_{i_d}$ (up to constant factors) by setting $t = d$ (in the below, $\widetilde{\boldsymbol{\Sigma}}$ is treated as a function of $\mathbf{x}_1, \ldots, \mathbf{x}_d$):

$$\begin{aligned}
\Pr(A) &\propto \mathbb{E}_{\mathrm{D}^k_{\mathcal{X}}}\Big[\mathbf{1}_A \det(\widetilde{\boldsymbol{\Sigma}})\Big(\prod_{i=1}^d l_{\widehat{\boldsymbol{\Sigma}}}(\mathbf{x}_i)\Big)\Big] \\
&\overset{(*)}{=} \mathbb{E}_{\mathrm{D}^k_{\mathcal{X}}}\Big[\mathbf{1}_A \frac{\det(\sum_i \mathbf{x}_i\mathbf{x}_i^\top)}{(\frac{d}{t})^d \prod_i l_{\widehat{\boldsymbol{\Sigma}}}(\mathbf{x}_i)}\Big(\prod_{i=1}^d l_{\widehat{\boldsymbol{\Sigma}}}(\mathbf{x}_i)\Big)\Big] \\
&\propto \mathbb{E}_{\mathrm{D}^k_{\mathcal{X}}}\Big[\mathbf{1}_A \det\Big(\sum_{i=1}^d \mathbf{x}_i\mathbf{x}_i^\top\Big)\Big] \\
&\propto \mathrm{VS}^d_{\mathrm{D}_{\mathcal{X}}}(A),
\end{aligned}$$

where $(*)$ uses the fact that for $t = d$, $\det(\widetilde{\boldsymbol{\Sigma}})$ is the squared volume of the parallelopiped spanned by $\mathbf{x}_1, \ldots, \mathbf{x}_d$ and stretched with the appropriate scaling factors. Thus, we established the correctness of Algorithm 1 for any $t \geq d$, and we move on to complexity analysis. If we think of each iteration of the **repeat** loop as a single Bernoulli trial, the success probability $\Pr(Acc = \text{true})$ equals $\mathbb{E}[\det(\widehat{\boldsymbol{\Sigma}}\widetilde{\boldsymbol{\Sigma}}^{-1})]$ with the expectation defined as in Lemma 12. Let $\lambda_1, \ldots, \lambda_d$ be the eigenvalues of matrix $\widehat{\boldsymbol{\Sigma}}\boldsymbol{\Sigma}^{-1}_{\mathrm{D}_{\mathcal{X}}}$. The approximation guarantee for $\widehat{\boldsymbol{\Sigma}}$ implies that all of these eigenvalues lie in the range $[1-\epsilon, 1+\epsilon]$. To lower-bound the success probability, we use the Kantorovich arithmetic-harmonic mean inequality. Letting $A(\cdot)$, $G(\cdot)$ and $H(\cdot)$ denote the arithmetic, geometric and harmonic means respectively:

$$\begin{aligned}
\frac{\det(\boldsymbol{\Sigma}_{\mathrm{D}_{\mathcal{X}}}\widehat{\boldsymbol{\Sigma}}^{-1})}{(\frac{1}{d}\mathrm{tr}(\boldsymbol{\Sigma}_{\mathrm{D}_{\mathcal{X}}}\widehat{\boldsymbol{\Sigma}}^{-1}))^d} &= \frac{\prod_{i=1}^d \frac{1}{\lambda_i}}{(\frac{1}{d}\sum_{i=1}^d \frac{1}{\lambda_i})^d} \\
&= \left(\frac{H(\lambda_1, \ldots, \lambda_d)}{G(\lambda_1, \ldots, \lambda_d)}\right)^d \overset{(1)}{\geq} \left(\frac{H(\lambda_1, \ldots, \lambda_d)}{A(\lambda_1, \ldots, \lambda_d)}\right)^d \\
&\overset{(2)}{\geq} \big((1-\epsilon)(1+\epsilon)\big)^d \overset{\epsilon=1/\sqrt{2d}}{=} \Big(1 - \frac{1}{2d}\Big)^d \geq \frac{1}{2},
\end{aligned}$$

where (1) is the geometric-arithmetic mean inequality and (2) is the Kantorovich inequality ([17]) with $a = 1-\epsilon$ and $b = 1+\epsilon$:

$$\text{For } 0 < a \leq \lambda_1, \ldots, \lambda_d \leq b, \quad \frac{A(\lambda_1, \ldots, \lambda_d)}{H(\lambda_1, \ldots, \lambda_d)} \leq \left(\frac{A(a,b)}{G(a,b)}\right)^2.$$

Now setting $t = 2d^2$ in Lemma 12, we obtain that

$$\Pr(Acc = \text{true}) = \mathbb{E}\big[\det(\widehat{\boldsymbol{\Sigma}}\widetilde{\boldsymbol{\Sigma}}^{-1})\big] \geq \Big(1 - \frac{d^2}{t}\Big)\frac{1}{2} = \frac{1}{4}.$$

So a simple tail bound on a geometric random variable shows that the number of iterations of **repeat** loop is $r \leq \ln(\frac{1}{\delta})/\ln(\frac{4}{3})$ w.p. at least $1-\delta$. It remains to bound the number of samples needed from $\mathrm{D}_{\mathcal{X}}$. Note that we can lower bound this success probability

$$\begin{aligned}
\Pr(a = \text{true}) = \mathbb{E}_{\mathrm{D}_{\mathcal{X}}}\left[\frac{\mathbf{x}^\top\widehat{\boldsymbol{\Sigma}}^{-1}\mathbf{x}}{K}\right] &= \frac{\mathrm{tr}(\boldsymbol{\Sigma}_{\mathrm{D}_{\mathcal{X}}}\widehat{\boldsymbol{\Sigma}}^{-1})}{K} \\
&\geq \frac{\mathrm{tr}(\boldsymbol{\Sigma}_{\mathrm{D}_{\mathcal{X}}}\boldsymbol{\Sigma}^{-1}_{\mathrm{D}_{\mathcal{X}}})}{(1+\epsilon)K} = \frac{d}{(1+\epsilon)K}.
\end{aligned}$$

Similarly as before we conclude that the number of samples needed for a single iteration of **repeat** loop is $O(2d^2\frac{K}{d}\ln(\frac{1}{\delta})) = O(Kd\ln(\frac{1}{\delta}))$ w.p. at least $1-\delta$. Note that the computational cost per sample is $O(d^2)$ and the cost of VolSamp is $O(d^4)$, obtaining the desired complexities. ∎

Finally, we discuss the time and sample complexity of obtaining $\widehat{\boldsymbol{\Sigma}}$ with desired accuracy under the model where access to $\mathrm{D}_{\mathcal{X}}$ is given only through sampling from the distribution. For this we can rely on standard matrix Chernoff bounds given by [25]. The below version is adapted from [7]:

Lemma 14 ([25, 7]) *If* $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_m \overset{\text{i.i.d.}}{\sim} \mathrm{D}_{\mathcal{X}}$ *and* $m \geq C\frac{K_{\mathrm{D}_{\mathcal{X}}}}{\epsilon^2}\ln(\frac{d}{\delta})$ *for some absolute constant* C*, then*

$$(1-\epsilon)\boldsymbol{\Sigma}_{\mathrm{D}_{\mathcal{X}}} \preceq \frac{1}{m}\sum_{i=1}^m \mathbf{x}_i\mathbf{x}_i^\top \preceq (1+\epsilon)\boldsymbol{\Sigma}_{\mathrm{D}_{\mathcal{X}}} \quad \textit{w.p.} \geq 1-\delta.$$

Setting $\epsilon = \frac{1}{\sqrt{2d}}$ in Lemma 14, we note that the sample complexity of obtaining $\widehat{\boldsymbol{\Sigma}}$ that would satisfy the assumptions of Theorem 4 is $m = O(K_{\mathrm{D}_{\mathcal{X}}}d\ln(\frac{d}{\delta}))$, and computing it takes $O(md^2) = O(K_{\mathrm{D}_{\mathcal{X}}}d^3\ln(\frac{d}{\delta}))$.

5 CONCLUSIONS

We show that for the least squares estimator, the bias which occurs in random design linear regression can be corrected by augmenting the dataset with dimension many points sampled from a special joint distribution - an extension of discrete volume sampling. We present two methods for performing this augmentation when the underlying data distribution is only known through i.i.d. samples. In the process we improve the time complexity of a recently proposed algorithm for discrete volume sampling.

An important future research direction is providing a random design error analysis for the least squares estimator of the augmented sample. Furthermore, it is natural to ask if there are distribution families other than multivariate normal which offer better complexity guarantees for producing volume-rescaled samples.

Acknowledgements

Michał Dereziński and Manfred K. Warmuth were supported by NSF grant IIS-1619271. Daniel Hsu was supported by NSF grant CCF-1740833 and a Sloan Research Fellowship. Part of this work was done while Michał Dereziński was visiting the Simons Institute for the Theory of Computing.

References

[1] 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.

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

[3] Rémi Bardenet, Frédéric Lavancier, Xavier MARY, and Aurélien Vasseur. On a few statistical applications of determinantal point processes. *ESAIM: Proceedings and Surveys*, 60, 2017.

[4] Aharon Ben-Tal and Marc Teboulle. A geometric property of the least squares solution of linear equations. *Linear Algebra and its Applications*, 139:165 – 170, 1990.

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

[6] 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.

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

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

[9] Michał Dereziński, Manfred K. Warmuth, and Daniel Hsu. Leveraged volume sampling for linear regression. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems 31*, pages 2510–2519. Curran Associates, Inc., 2018.

[10] 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.

[11] 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.

[12] 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.

[13] A.K. Gupta and D.K. Nagar. *Matrix Variate Distributions*. PMS Series. Addison-Wesley Longman, Limited, 1999.

[14] Venkatesan Guruswami and Ali K. Sinop. Optimal column-based low-rank matrix reconstruction. In *Proceedings of the Twenty-third Annual ACM-SIAM Symposium on Discrete Algorithms*, SODA '12, pages 1207–1214, Philadelphia, PA, USA, 2012. Society for Industrial and Applied Mathematics.

[15] J. Ben Hough, Manjunath Krishnapur, Yuval Peres, and Bálint Virág. Determinantal processes and independence. *Probab. Surveys*, 3:206–229, 2006.

[16] Daniel Hsu, Sham M. Kakade, and Tong Zhang. Random design analysis of ridge regression. *Foundations of Computational Mathematics*, 14(3):569–600, 2014.

[17] Leonid V. Kantorovich. Functional analysis and applied mathematics. *Uspekhi Matematicheskikh Nauk*, 3(6):89–185, 1948.

[18] 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.

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

[20] Chengtao Li, Stefanie Jegelka, and Suvrit Sra. Efficient sampling for k-determinantal point processes. In Arthur Gretton and Christian C. Robert, editors, *Proceedings of the 19th International Conference on Artificial Intelligence and Statistics*, volume 51 of *Proceedings of Machine Learning Research*, pages 1328–1337, Cadiz, Spain, 09–11 May 2016.

[21] 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.

[22] 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.

[23] Sujit Kumar Mitra. A density-free approach to the matrix variate beta distribution. *Sankhyā: The Indian Journal of Statistics, Series A (1961-2002)*, 32(1):81–88, 1970.

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

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

[26] H. Robert van der Vaart. A note on wilks' internal scatter. *Ann. Math. Statist.*, 36(4):1308–1312, 08 1965.