Subsampling for Ridge Regression
via Regularized Volume Sampling[*]

Michał Dereziński
Department of Computer Science
University of California Santa Cruz
mderezin@ucsc.edu

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

Abstract

Given n vectors $\mathbf{x}_i \in \mathbb{R}^d$, we want to fit a linear regression model for noisy labels $y_i \in \mathbb{R}$. The ridge estimator is a classical solution to this problem. However, when labels are expensive, we are forced to select only a small subset of vectors $\mathbf{x}_i$ for which we obtain the labels y_i. We propose a new procedure for selecting the subset of vectors, such that the ridge estimator obtained from that subset offers strong statistical guarantees in terms of the mean squared prediction error over the entire dataset of n labeled vectors. The number of labels needed is proportional to the statistical dimension of the problem which is often much smaller than d. Our method is an extension of a joint subsampling procedure called volume sampling. A second major contribution is that we speed up volume sampling so that it is essentially as efficient as leverage scores, which is the main i.i.d. subsampling procedure for this task. Finally, we show theoretically and experimentally that volume sampling has a clear advantage over any i.i.d. sampling when labels are expensive.

1 Introduction

Given a matrix $\mathbf{X} \in \mathbb{R}^{d \times n}$, we consider the task of fitting a linear model[1] to a vector of labels $\mathbf{y} = \mathbf{X}^\top \mathbf{w}^* + \boldsymbol{\xi}$, where $\mathbf{w}^* \in \mathbb{R}^d$ and the noise $\boldsymbol{\xi} \in \mathbb{R}^n$

is a mean zero random vector with covariance matrix $\mathrm{Var}[\boldsymbol{\xi}] \preceq \sigma^2 \mathbf{I}$ for some $\sigma > 0$. A classical solution to this task is the ridge estimator:

$$\widehat{\mathbf{w}}_\lambda^* = \underset{\mathbf{w} \in \mathbb{R}^d}{\mathrm{argmin}} \ \ \|\mathbf{X}^\top \mathbf{w} - \mathbf{y}\|^2 + \lambda \|\mathbf{w}\|^2$$
$$= (\mathbf{X}\mathbf{X}^\top + \lambda \mathbf{I})^{-1} \mathbf{X}\mathbf{y}.$$

In many settings, obtaining labels y_i is expensive and we are forced to select a subset $S \subseteq \{1..n\}$ of label indices. Let $\mathbf{y}_S \in \mathbb{R}^{|S| \times 1}$ be the sub vector of labels indexed by S and $\mathbf{X}_S \in \mathbb{R}^{d \times |S|}$ be the columns of $\mathbf{X}$ indexed by S. We will show that if S is sampled with a new variant of *volume sampling* [3] on the columns of $\mathbf{X}$, then the ridge estimator for the subproblem $(\mathbf{X}_S, \mathbf{y}_S)$

$$\widehat{\mathbf{w}}_\lambda^*(S) = (\mathbf{X}_S \mathbf{X}_S^\top + \lambda \mathbf{I})^{-1} \mathbf{X}_S \mathbf{y}_S$$

has strong generalization properties with respect to the full problem $(\mathbf{X}, \mathbf{y})$.

Volume sampling is a sampling technique which has received a lot of attention recently [3, 5, 6, 7, 15]. For a fixed size $s \geq d$, the original variant samples $S \subseteq \{1..n\}$ of size s proportional to the squared volume of the parallelepiped spanned by the rows of $\mathbf{X}_S$ [3]:

$$P(S) \propto \det(\mathbf{X}_S \mathbf{X}_S^\top). \qquad (1)$$

A simple approach for implementing volume sampling (just introduced in [5]) is to start with the full set of column indices $S = \{1..n\}$ and then (in reverse order) select an index i in each iteration to be eliminated from set S with probability proportional to the change in matrix volume caused by removing the ith column:

$$\text{Sample } i \sim P(i \,|\, S) = \frac{\det(\mathbf{X}_{S_{-i}} \mathbf{X}_{S_{-i}}^\top)}{(|S| - d) \det(\mathbf{X}_S \mathbf{X}_S^\top)}, \quad (2)$$

Update $S \leftarrow S - \{i\}$.

(Reverse Iterative Volume Sampling)

Note that when $|S| < d$, then all matrices $\mathbf{X}_S \mathbf{X}_S^\top$ are singular, and so the distribution becomes undefined.

---

[1]This setting can easily be extended to "non-linear models" via kernelization.

Motivated by this limitation, we propose a regularized variant, called λ-regularized volume sampling:

$$\text{Sample } i \sim P(i \mid S) \propto \frac{\det(\mathbf{X}_{S_{-i}}\mathbf{X}_{S_{-i}}^{\top} + \lambda\mathbf{I})}{\det(\mathbf{X}_S\mathbf{X}_S^{\top} + \lambda\mathbf{I})}, \quad (3)$$

Update $S \leftarrow S - \{i\}$.

(λ-Regularized Volume Sampling)

Note that in the special case of no regularization (i.e. $\lambda = 0$), then (3) sums to $\frac{1}{|S|-d}$ (see equality (2)). However when $\lambda > 0$, then the sum of (3) depends on all columns of $\mathbf{X}_S$ and not just the size of S. This makes regularized volume sampling more complicated and certain equalities proven in [5] for $\lambda = 0$ become inequalities.

Nevertheless, we were able to show that the proposed λ-regularized distribution exhibits a fundamental connection to ridge regression, and introduce an efficient algorithm to sample from it in time $O((n + d)d^2)$. In particular, we prove that when S is sampled according to λ-regularized volume sampling with $\lambda \leq \frac{\sigma^2}{\|\mathbf{w}^*\|^2}$, then the mean squared prediction error (MSPE) of estimator $\widehat{\mathbf{w}}_\lambda^*(S)$ over the entire dataset $\mathbf{X}$ is bounded:

$$\mathbb{E}_S\mathbb{E}_{\boldsymbol{\xi}} \frac{1}{n}\|\mathbf{X}^{\top}(\widehat{\mathbf{w}}_\lambda^*(S) - \mathbf{w}^*)\|^2 \leq \frac{\sigma^2 d_\lambda}{s - d_\lambda + 1},$$
$$\text{where} \quad d_\lambda = \text{tr}(\mathbf{X}^{\top}(\mathbf{X}\mathbf{X}^{\top} + \lambda\mathbf{I})^{-1}\mathbf{X}) \quad (4)$$

is the statistical dimension. If λ_i are the eigenvalues of $\mathbf{X}\mathbf{X}^{\top}$, then $d_\lambda = \sum_{i=1}^{d} \frac{\lambda_i}{\lambda_i + \lambda}$. Note that d_λ is decreasing with λ and $d_0 = d$. If the spectrum of the matrix $\mathbf{X}\mathbf{X}^{\top}$ decreases quickly then d_λ does so as well with increasing λ. When λ is properly tuned then d_λ is the effective degrees of freedom of $\mathbf{X}$. Our new lower bounds will show that the above upper bound for regularized volume sampling is essentially optimal with respect to the choice of a subsampling procedure.

Volume sampling can be viewed as a non-i.i.d. extension of leverage score sampling [8], a widely used method where columns are sampled independently according to their leverage scores. Volume sampling has been shown to return better column subsets than its i.i.d. counterpart in many applications like experimental design, linear regression and graph theory [3, 5]. In this paper we additionally show that any i.i.d. subsampling with respect to any fixed distribution such as leverage score sampling can require $\Omega(d_\lambda \ln(d_\lambda))$ labels to achieve any generalization for ridge regression, compared to $O(d_\lambda)$ for regularized volume sampling. We reinforce this claim experimentally in Section 4.

The main obstacle against using volume sampling in practice has been high computational cost. Only recently, the first polynomial time algorithms have been proposed for exact [5] and approximate [15] volume

	Exact	Approximate
[15]	$O(n^4 s)$	$\widetilde{O}(nd^2 s^3)$
[5]	$O(n^2 d)$	-
here	$O(nd^2)$	-

Table 1: Comparison of runtime for exact and approximate volume sampling algorithms, where $d \leq s \leq n$.

sampling (see Table 1 for comparison). In particular, the fastest algorithm for exact volume sampling[2] is $O(n^2 d)$ whereas exact leverage score sampling[3] is $O(nd^2)$ (in both cases, the dependence on sample size s is not asymptotically significant). In typical settings for experimental design [9] and active learning [18], quality of the sample is more important than the runtime. However for many modern datasets, the number of examples n is much larger than d, which makes existing algorithms for volume sampling infeasible. In this paper, we give an easy-to-implement volume sampling algorithm that runs in time $O(nd^2)$. Thus we give the first volume sampling procedure which is essentially linear in n and matches the time complexity of exact leverage score sampling. For example, dataset MSD from the UCI data repository [16] has $n = 464,000$ examples with dimension $d = 90$. Our algorithm performed volume sampling on this dataset in 39 seconds, whereas the previously best $O(n^2 d)$ algorithm [5] did not finish within 24 hours. Sampling with leverage scores took 12 seconds on this data set. Finally our procedure also achieves regularized volume sampling for any $\lambda > 0$ with the running time of $O((n + d)d^2)$.

1.1 Related work

Many variants of probability distributions based on the matrix determinant have been studied in the literature, including Determinantal Point Processes (DPP) [14], k-DPP's [13] and volume sampling [3, 5, 6, 15], with applications to matrix approximation [7], recommender systems [10], etc. More recently, further theoretical results suggesting applications of volume sampling in linear regression were given by [5], where an expected loss bound for the unregularized least squares estimator was given under volume sampling of size $s = d$. Moreover, Reverse Iterative Volume Sampling – a technique enhanced in this paper with a regularization – was first proposed in [5].

Subset selection techniques for regression have long been studied in the field of experimental design [9]. More recently, computationally tractable techniques

---

[2]The exact time complexity is $O((n - s + d)nd)$ which is $O(n^2 d)$ for $s < n/2$.

[3]Approximate leverage score sampling methods achieve even better runtime of $\widetilde{O}(nd + d^3)$.

have been explored [2]. Statistical guarantees under i.i.d. subsampling in kernel ridge regression have been analyzed for uniform sampling [4] and leverage score sampling [1]. In this paper, we propose the first tractable non-i.i.d. subsampling procedure with strong statistical guarantees for the ridge estimator and show its benefits over using i.i.d. sampling approaches.

For the special case of volume sampling size $s = d$, a polynomial time algorithm was developed by [6], and slightly improved by [11], with runtime $O(nd^3)$. An exact sampling algorithm for arbitrary $s \geq d$ volume sampling was given by [5], with time complexity $O((n-s+d)nd)$ which is $O(n^2d)$ when $s < n/2$. Also, [15] proposed a Markov-chain procedure which generates ϵ-approximate volume samples in time $\widetilde{O}(nd^2s^3)$. The algorithm proposed in this paper, running in time $O(nd^2)$, enjoys a direct asymptotic speed-up over all of the above methods. Moreover, the procedure suffers only a small constant factor overhead over computing exact leverage scores of matrix $\mathbf{X}$.

1.2 Main results

The main contributions of this paper are two-fold:

1. **Statistical**: We define a regularized variant of volume sampling and show that it offers strong generalization guarantees for ridge regression in terms of mean squared error (MSE) and mean squared prediction error (MSPE).

2. **Algorithmic**: We propose a simple implementation of volume sampling, which not only extends the procedure to its regularized variant, but also offers a significant runtime improvement over the existing methods when $n \gg d$.

The key technical result shown in this paper, needed to obtain statistical guarantees for ridge regression, is the following property of regularized volume sampling (where d_λ is defined as in (4)):

Theorem 1 *For any $\mathbf{X} \in \mathbb{R}^{d \times n}$, $\lambda \geq 0$ and $s \geq d_\lambda$, let S be sampled according to λ-regularized size s volume sampling from $\mathbf{X}$. Then,*

$$\mathbb{E}_S \left(\mathbf{X}_S \mathbf{X}_S^\top + \lambda \mathbf{I}\right)^{-1} \preceq \frac{n - d_\lambda + 1}{s - d_\lambda + 1} (\mathbf{X} \mathbf{X}^\top + \lambda \mathbf{I})^{-1},$$

where $\preceq$ denotes a positive semi-definite inequality between matrices.

As a consequence of Theorem 1, we show that ridge estimators computed from volume sampled subproblems offer statistical guarantees with respect to the full regression problem $(\mathbf{X}, \mathbf{y})$, despite observing only a small portion of the labels.

Theorem 2 *Let $\mathbf{X} \in \mathbb{R}^{d \times n}$ and $\mathbf{w}^* \in \mathbb{R}^d$, and suppose that $\mathbf{y} = \mathbf{X}^\top \mathbf{w}^* + \boldsymbol{\xi}$, where $\boldsymbol{\xi}$ is a mean zero vector with $\mathrm{Var}[\boldsymbol{\xi}] \preceq \sigma^2 \mathbf{I}$. Let S be sampled according to λ-regularized size $s \geq d_\lambda$ volume sampling from $\mathbf{X}$ and $\widehat{\mathbf{w}}_\lambda^*(S)$ be the λ-ridge estimator of $\mathbf{w}^*$ computed from subproblem $(\mathbf{X}_S, \mathbf{y}_S)$. Then, if $\lambda \leq \frac{\sigma^2}{\|\mathbf{w}^*\|^2}$, we have*

(MSPE) $\quad \mathbb{E}_S \mathbb{E}_{\boldsymbol{\xi}} \frac{1}{n} \|\mathbf{X}^\top (\widehat{\mathbf{w}}_\lambda^*(S) - \mathbf{w}^*)\|^2 \leq \frac{\sigma^2 d_\lambda}{s - d_\lambda + 1},$

(MSE) $\mathbb{E}_S \mathbb{E}_{\boldsymbol{\xi}} \|\widehat{\mathbf{w}}_\lambda^*(S) - \mathbf{w}^*\|^2 \leq \frac{\sigma^2 n \, \mathrm{tr}((\mathbf{X}\mathbf{X}^\top + \lambda \mathbf{I})^{-1})}{s - d_\lambda + 1}.$

Next, we present two lower-bounds for MSPE of a subsampled ridge estimator which show that the statistical guaranties achieved by regularized volume sampling are nearly optimal for $s \gg d_\lambda$ and better than standard approaches for $s = O(d_\lambda)$. In particular, we show that non-i.i.d. nature of volume sampling is essential if we want to achieve good generalization when the number of labels is close to d_λ. Namely, for certain data matrices, any subsampling procedure selecting examples in an i.i.d. fashion (e.g., leverage score sampling), requires more than $d_\lambda \ln(d_\lambda)$ labels to achieve MSPE below σ^2, whereas volume sampling obtains that bound for any matrix with $2d_\lambda$ labels.

Theorem 3 *For any $p \geq 1$ and $\sigma \geq 0$, there is $d \geq p$ such that for any sufficiently large n divisible by d there exists a matrix $\mathbf{X} \in \mathbb{R}^{d \times n}$ such that*

$$d_\lambda(\mathbf{X}) \geq p \quad \text{for any} \quad 0 \leq \lambda \leq \sigma^2,$$

and for each of the following two statements there is a vector $\mathbf{w}^* \in \mathbb{R}^d$ for which the corresponding regression problem $\mathbf{y} = \mathbf{X}^\top \mathbf{w}^* + \boldsymbol{\xi}$ with $\mathrm{Var}[\boldsymbol{\xi}] = \sigma^2 \mathbf{I}$ satisfies that statement:

1. *For any subset $S \subseteq \{1..n\}$ of size s,*

$$\mathbb{E}_{\boldsymbol{\xi}} \frac{1}{n} \|\mathbf{X}^\top (\widehat{\mathbf{w}}_\lambda^*(S) - \mathbf{w}^*)\|^2 \geq \frac{\sigma^2 d_\lambda}{s + d_\lambda};$$

2. *For multiset $S \subseteq \{1..n\}$ of size $s \leq d_\lambda(\ln(d_\lambda) - 1)$, sampled i.i.d. from any distribution over $\{1..n\}$,*

$$\mathbb{E}_S \mathbb{E}_{\boldsymbol{\xi}} \frac{1}{n} \|\mathbf{X}^\top (\widehat{\mathbf{w}}_\lambda^*(S) - \mathbf{w}^*)\|^2 \geq \sigma^2.$$

Finally, we propose an algorithm for regularized volume sampling which runs in time $O((n+d)d^2)$. For the previously studied case of $\lambda = 0$, this algorithm offers a significant asymptotic speed-up over existing volume sampling algorithms (both exact and approximate).

Theorem 4 *For any $\lambda, \delta, s \geq 0$, there is an algorithm sampling according to λ-regularized size s volume sam-*

pling, that with probability at least $1 - \delta$ runs in time[4]

$$O\left(\left(n + d + \log\left(\frac{n}{d}\right)\log\left(\frac{1}{\delta}\right)\right)d^2\right).$$

When $n > d$ the time complexity of our proposed algoirthm is not deterministic, but its dependence on the failure probability δ is very small – even for $\delta = 2^{-n}$ the time complexity is still $\widetilde{O}(nd^2)$.

The remainder of this paper is arranged as follows: in Section 2 we present statistical analysis of regularized volume sampling in the context of ridge regression, in particular proving Theorems 1, 2 and 3; in Section 3 we present two algorithms for regularized volume sampling and use them to prove Theorem 4; in Section 4 we evaluate the runtime of our algorithms on several standard linear regression datasets, and compare the prediction performance of the subsampled ridge estimator under volume sampling versus leverage score sampling; in Section 5 we summarize the results and suggest future research directions.

2 Statistical guarantees

In this section, we show upper and lower bounds for the generalization performance of subsampled ridge estimators, starting with an important property of regularized volume sampling which connects it with ridge regression. We will use $\mathbf{Z}_\lambda(S) = \mathbf{X}_S\mathbf{X}_S^\top + \lambda\mathbf{I}$ as a shorthand in the proofs.

2.1 Proof of Theorem 1

To obtain this result, we will show how the expectation of matrix $(\mathbf{X}_S\mathbf{X}_S^\top + \lambda\mathbf{I})^{-1}$ changes when iteratively removing a column in λ-Regularized Volume Sampling (see (3)):

Lemma 5 *Let* $\mathbf{X} \in \mathbb{R}^{d\times n}$ *and* $S \subseteq \{1..n\}$.

If we sample $i \in S$ *w.p.* $\propto \frac{\det(\mathbf{X}_{S_{-i}}\mathbf{X}_{S_{-i}}^\top + \lambda\mathbf{I})}{\det(\mathbf{X}_S\mathbf{X}_S^\top + \lambda\mathbf{I})}$, *then*

$$\mathbb{E}_i\left(\mathbf{X}_{S_{-i}}\mathbf{X}_{S_{-i}}^\top + \lambda\mathbf{I}\right)^{-1} \preceq \frac{s - d_\lambda + 1}{s - d_\lambda}(\mathbf{X}_S\mathbf{X}_S^\top + \lambda\mathbf{I})^{-1}.$$

Proof We write the unnormalized probability of i as:

$$h_i(S) = \frac{\det(\mathbf{Z}_\lambda(S_{-i}))}{\det(\mathbf{Z}_\lambda(S))} \overset{(*)}{=} 1 - \mathbf{x}_i^\top \mathbf{Z}_\lambda(S)^{-1}\mathbf{x}_i,$$

where $(*)$ follows from Sylvester's theorem. Next, letting $M = \sum_{i\in S} h_i(S)$, we compute unnormalized expectation by applying the Sherman-Morrison formula

---

[4] We are primarily interested in the case where $n \geq d$. However, if $n < d$ and $\lambda > 0$, then our techniques can be adapted to obtain an $O(n^2d)$ algorithm.

to $\mathbf{Z}_\lambda(S_{-i})^{-1}$:

$$M\,\mathbb{E}_i\left(\mathbf{X}_{S_{-i}}\mathbf{X}_{S_{-i}}^\top + \lambda\mathbf{I}\right)^{-1} = \sum_{i\in S} h_i(S)\mathbf{Z}_\lambda(S_{-i})^{-1}$$

$$= \sum_{i\in S} h_i(S)\left(\mathbf{Z}_\lambda(S)^{-1} + \frac{\mathbf{Z}_\lambda(S)^{-1}\mathbf{x}_i\mathbf{x}_i^\top\mathbf{Z}_\lambda(S)^{-1}}{1 - \mathbf{x}_i^\top\mathbf{Z}_\lambda(S)^{-1}\mathbf{x}_i}\right)$$

$$= M\,\mathbf{Z}_\lambda(S)^{-1} + \mathbf{Z}_\lambda(S)^{-1}\left(\sum_{i\in S}\mathbf{x}_i\mathbf{x}_i^\top\right)\mathbf{Z}_\lambda(S)^{-1}$$

$$= M\,\mathbf{Z}_\lambda(S)^{-1} + \mathbf{Z}_\lambda(S)^{-1}(\mathbf{Z}_\lambda(S) - \lambda\mathbf{I})\mathbf{Z}_\lambda(S)^{-1}$$

$$= M\,\mathbf{Z}_\lambda(S)^{-1} + \mathbf{Z}_\lambda(S)^{-1} - \lambda\mathbf{Z}_\lambda(S)^{-2}$$

$$\preceq (M+1)\,\mathbf{Z}_\lambda(S)^{-1}.$$

Finally, we compute the normalization factor:

$$M = \sum_{i\in S}(1 - \mathbf{x}_i^\top\mathbf{Z}_\lambda(S)^{-1}\mathbf{x}_i)$$

$$= s - \text{tr}(\mathbf{X}_S^\top\mathbf{Z}_\lambda(S)^{-1}\mathbf{X}_S) = s - \text{tr}(\mathbf{X}_S\mathbf{X}_S^\top\mathbf{Z}_\lambda(S)^{-1})$$

$$= s - \text{tr}((\mathbf{Z}_\lambda(S) - \lambda\mathbf{I})\mathbf{Z}_\lambda(S)^{-1})$$

$$= s - \text{tr}(\mathbf{I}) + \lambda\text{tr}(\mathbf{Z}_\lambda(S)^{-1})$$

$$\geq s - d + \lambda\text{tr}(\mathbf{Z}_\lambda(\{1..n\})^{-1}) = s - d_\lambda,$$

where we used the fact that d_λ can we rewritten as $d - \lambda\text{tr}(\mathbf{Z}_\lambda(\{1..n\})^{-1})$. Putting the bounds together, we obtain the result. ∎

To prove Theorem 1 it remains to chain the conditional expectations along the sequence of subsets obtained by λ-Regularized Volume Sampling:

$$\mathbb{E}_S\,\mathbf{Z}_\lambda(S)^{-1} \preceq \left(\prod_{t=s+1}^n \frac{t - d_\lambda + 1}{t - d_\lambda}\right)\mathbf{Z}_\lambda(\{1..n\})^{-1}$$

$$= \frac{n - d_\lambda + 1}{s - d_\lambda + 1}(\mathbf{XX}^\top + \lambda\mathbf{I})^{-1}.$$

2.2 Proof of Theorem 2

Standard analysis for the ridge regression estimator follows by performing bias-variance decomposition of the error, and then selecting λ so that bias can be appropriately bounded. We will recall this calculation for a fixed subproblem $(\mathbf{X}_S, \mathbf{y}_S)$. First, we compute the bias of the ridge estimator for a fixed set S:

$$\text{Bias}_{\boldsymbol{\xi}}[\widehat{\mathbf{w}}_\lambda^*(S)] = \mathbb{E}[\widehat{\mathbf{w}}_\lambda^*(S)] - \mathbf{w}^*$$

$$= \mathbb{E}_{\boldsymbol{\xi}}\left[\mathbf{Z}_\lambda(S)^{-1}\mathbf{X}_S\mathbf{y}_S\right] - \mathbf{w}^*$$

$$= \mathbf{Z}_\lambda(S)^{-1}\mathbf{X}_S\left(\mathbf{X}_S^\top\mathbf{w}^* + \mathbb{E}_{\boldsymbol{\xi}}[\boldsymbol{\xi}_S]\right) - \mathbf{w}^*$$

$$= (\mathbf{Z}_\lambda(S)^{-1}\mathbf{X}_S\mathbf{X}_S^\top - \mathbf{I})\mathbf{w}^*$$

$$= -\lambda\mathbf{Z}_\lambda(S)^{-1}\mathbf{w}^*.$$

Similarly, the covariance matrix of $\widehat{\mathbf{w}}_\lambda^*(S)$ is given by:

$$\text{Var}_{\boldsymbol{\xi}}[\widehat{\mathbf{w}}_\lambda^*(S)] = \mathbf{Z}_\lambda(S)^{-1}\mathbf{X}_S\text{Var}_{\boldsymbol{\xi}}[\boldsymbol{\xi}_S]\mathbf{X}_S^\top\mathbf{Z}_\lambda(S)^{-1}$$

$$\preceq \sigma^2\mathbf{Z}_\lambda(S)^{-1}\mathbf{X}_S\mathbf{X}_S^\top\mathbf{Z}_\lambda(S)^{-1}$$

$$= \sigma^2(\mathbf{Z}_\lambda(S)^{-1} - \lambda\mathbf{Z}_\lambda(S)^{-2}).$$

Mean squared error of the ridge estimator for a fixed subset S can now be bounded by:

$$\mathbb{E}_{\boldsymbol{\xi}}\|\widehat{\mathbf{w}}_\lambda^*(S) - \mathbf{w}^*\|^2 = \mathrm{tr}(\mathrm{Var}_{\boldsymbol{\xi}}[\widehat{\mathbf{w}}_\lambda^*(S)]) + \|\mathrm{Bias}_{\boldsymbol{\xi}}[\widehat{\mathbf{w}}_\lambda^*(S)]\|^2$$
$$\leq \sigma^2 \mathrm{tr}(\mathbf{Z}_\lambda(S)^{-1} - \lambda \mathbf{Z}_\lambda(S)^{-2}) + \lambda^2 \mathrm{tr}(\mathbf{Z}_\lambda(S)^{-2}\mathbf{w}^*\mathbf{w}^{*\top})$$
$$\leq \sigma^2 \mathrm{tr}(\mathbf{Z}_\lambda(S)^{-1}) + \lambda \mathrm{tr}(\mathbf{Z}_\lambda(S)^{-2})(\lambda\|\mathbf{w}^*\|^2 - \sigma^2) \quad (5)$$
$$\leq \sigma^2 \mathrm{tr}(\mathbf{Z}_\lambda(S)^{-1}), \quad (6)$$

where in (5) we applied Cauchy-Schwartz inequality for matrix trace, and in (6) we used the assumption that $\lambda \leq \frac{\sigma^2}{\|\mathbf{w}^*\|^2}$. Thus, taking expectation over the sampling of set S, we get

$$\mathbb{E}_S \mathbb{E}_{\boldsymbol{\xi}}\|\widehat{\mathbf{w}}_\lambda^*(S) - \mathbf{w}^*\|^2 \leq \sigma^2 \mathbb{E}_S \mathrm{tr}(\mathbf{Z}_\lambda(S)^{-1})$$

$$(\text{Thm. 1}) \leq \sigma^2 \frac{n - d_\lambda + 1}{s - d_\lambda + 1} \mathrm{tr}(\mathbf{Z}_\lambda(\{1..n\})^{-1}) \quad (7)$$

$$\leq \frac{\sigma^2 n \, \mathrm{tr}((\mathbf{X}\mathbf{X}^\top + \lambda \mathbf{I})^{-1})}{s - d_\lambda + 1}.$$

Next, we bound the mean squared prediction error. As before, we start with the standard bias-variance decomposition for fixed set S:

$$\mathbb{E}_{\boldsymbol{\xi}}\|\mathbf{X}^\top(\widehat{\mathbf{w}}_\lambda^*(S) - \mathbf{w}^*)\|^2$$
$$= \mathrm{tr}(\mathrm{Var}_{\boldsymbol{\xi}}[\mathbf{X}^\top\widehat{\mathbf{w}}_\lambda^*(S)]) + \|\mathbf{X}^\top(\mathbb{E}_{\boldsymbol{\xi}}[\widehat{\mathbf{w}}_\lambda^*(S)] - \mathbf{w}^*)\|^2$$
$$\leq \sigma^2 \mathrm{tr}(\mathbf{X}^\top(\mathbf{Z}_\lambda(S)^{-1} - \lambda \mathbf{Z}_\lambda(S)^{-2})\mathbf{X})$$
$$\quad + \lambda^2 \mathrm{tr}(\mathbf{Z}_\lambda(S)^{-1}\mathbf{X}\mathbf{X}^\top \mathbf{Z}_\lambda(S)^{-1}\mathbf{w}^*\mathbf{w}^{*\top})$$
$$\leq \sigma^2 \mathrm{tr}(\mathbf{X}^\top \mathbf{Z}_\lambda(S)^{-1}\mathbf{X})$$
$$\quad + \lambda \mathrm{tr}(\mathbf{X}^\top \mathbf{Z}_\lambda(S)^{-2}\mathbf{X})(\lambda\|\mathbf{w}^*\|^2 - \sigma^2)$$
$$\leq \sigma^2 \mathrm{tr}(\mathbf{X}^\top \mathbf{Z}_\lambda(S)^{-1}\mathbf{X}).$$

Once again, taking expectation over subset S, we have

$$\mathbb{E}_S \mathbb{E}_{\boldsymbol{\xi}} \frac{1}{n}\|\mathbf{X}^\top(\widehat{\mathbf{w}}_\lambda^*(S) - \mathbf{w}^*)\|^2$$
$$\leq \frac{\sigma^2}{n} \mathbb{E}_S \, \mathrm{tr}(\mathbf{X}^\top \mathbf{Z}_\lambda(S)^{-1}\mathbf{X})$$
$$= \frac{\sigma^2}{n} \mathrm{tr}(\mathbf{X}^\top \mathbb{E}_S[\mathbf{Z}_\lambda(S)^{-1}]\mathbf{X})$$
$$(\text{Thm. 1}) \leq \frac{\sigma^2}{n} \frac{n - d_\lambda + 1}{s - d_\lambda + 1} \mathrm{tr}(\mathbf{X}^\top \mathbf{Z}_\lambda(\{1..n\})^{-1}\mathbf{X}) \quad (8)$$
$$\leq \frac{\sigma^2 d_\lambda}{s - d_\lambda + 1}.$$

The key part of proving both bounds is the application of Theorem 1. For MSE, we only used the trace version of the inequality (see (7)), however to obtain the bound on MSPE we used the more general positive semi-definite inequality in (8).

2.3   Proof of Theorem 3

Let $d = \lceil p \rceil + 1$ and $n \geq \lceil \sigma^2 \rceil d(d-1)$ be divisible by d. We define

$$\mathbf{X} \stackrel{def}{=} [\mathbf{I}, ..., \mathbf{I}] \in \mathbb{R}^{d \times n}$$

$$\mathbf{w}^{*\top} \stackrel{def}{=} [a\sigma, ..., a\sigma] \in \mathbb{R}^d$$

for some $a > 0$. For any $\lambda \leq \sigma^2$, the λ-statistical dimension of $\mathbf{X}$ is

$$d_\lambda = \mathrm{tr}(\mathbf{X}^\top \mathbf{Z}_\lambda(\{1..n\})^{-1}\mathbf{X})$$
$$\geq \frac{\lceil \sigma^2 \rceil d(d-1)}{\lceil \sigma^2 \rceil(d-1) + \lambda} \geq \frac{d(d-1)}{d-1+1} \geq p.$$

Let $S \subseteq \{1..n\}$ be any set of size s, and for $i \in \{1..d\}$ let

$$s_i \stackrel{def}{=} |\{i \in S : \mathbf{x}_i = \mathbf{e}_i\}|.$$

The prediction variance of estimator $\widehat{\mathbf{w}}_\lambda^*(S)$ is equal to

$$\mathrm{tr}(\mathrm{Var}_{\boldsymbol{\xi}}[\mathbf{X}^\top\widehat{\mathbf{w}}_\lambda^*(S)]) = \sigma^2 \mathrm{tr}(\mathbf{X}^\top(\mathbf{Z}_\lambda(S)^{-1} - \lambda \mathbf{Z}_\lambda(S)^{-2})\mathbf{X})$$
$$= \frac{\sigma^2 n}{d} \sum_{i=1}^d \left(\frac{1}{s_i + \lambda} - \frac{\lambda}{(s_i + \lambda)^2} \right)$$
$$= \frac{\sigma^2 n}{d} \sum_{i=1}^d \frac{s_i}{(s_i + \lambda)^2}.$$

The prediction bias of estimator $\widehat{\mathbf{w}}_\lambda^*(S)$ is equal to

$$\|\mathbf{X}^\top(\mathbb{E}_{\boldsymbol{\xi}}[\widehat{\mathbf{w}}_\lambda^*(S)] - \mathbf{w}^*)\|^2$$
$$= \lambda^2 \mathbf{w}^{*\top}\mathbf{Z}_\lambda(S)^{-1}\mathbf{X}\mathbf{X}^\top \mathbf{Z}_\lambda(S)^{-1}\mathbf{w}^{*\top}$$
$$= \frac{\lambda^2 a^2 \sigma^2 n}{d} \mathrm{tr}(\mathbf{Z}_\lambda(S)^{-2})$$
$$= \frac{\lambda^2 a^2 \sigma^2 n}{d} \sum_{i=1}^d \frac{1}{(s_i + \lambda)^2}.$$

Thus, MSPE of estimator $\widehat{\mathbf{w}}_\lambda^*(S)$ is equal to:

$$\mathbb{E}_{\boldsymbol{\xi}} \frac{1}{n}\|\mathbf{X}^\top(\widehat{\mathbf{w}}_\lambda^*(S) - \mathbf{w}^*)\|^2$$
$$= \frac{1}{n}\mathrm{tr}(\mathrm{Var}_{\boldsymbol{\xi}}[\mathbf{X}^\top\widehat{\mathbf{w}}_\lambda^*(S)]) + \frac{1}{n}\|\mathbf{X}^\top(\mathbb{E}_{\boldsymbol{\xi}}[\widehat{\mathbf{w}}_\lambda^*(S)] - \mathbf{w}^*)\|^2$$
$$= \frac{\sigma^2}{d} \sum_{i=1}^d \left(\frac{s_i}{(s_i + \lambda)^2} + \frac{a^2\lambda^2}{(s_i + \lambda)^2} \right)$$
$$= \frac{\sigma^2}{d} \sum_{i=1}^d \frac{s_i + a^2\lambda^2}{(s_i + \lambda)^2}.$$

Next, we find the λ that minimizes this expression. Taking the derivative with respect to λ we get:

$$\frac{\partial}{\partial \lambda}\left(\frac{\sigma^2}{d} \sum_{i=1}^d \frac{s_i + a^2\lambda^2}{(s_i + \lambda)^2} \right) = \frac{\sigma^2}{d} \sum_{i=1}^d \frac{2s_i(\lambda - a^{-2})}{(s_i + \lambda)^3}.$$

Thus, since at least one s_i has to be greater than 0, for any set S the derivative is negative for $\lambda < a^{-2}$ and positive for $\lambda > a^{-2}$, and the unique minimum of MSPE is achieved at $\lambda = a^{-2}$, regardless of which subset S is chosen. So, as we are seeking a lower bound, we can focus on the case of $\lambda = a^{-2}$.

2.3.1 Proof of Part 1

Let $a = 1$. As shown above, we can assume that $\lambda = 1$. In this case the formula simplifies to:

$$\mathbb{E}_{\boldsymbol{\xi}}\frac{1}{n}\|\mathbf{X}^\top(\widehat{\mathbf{w}}_\lambda^*(S) - \mathbf{w}^*)\|^2 = \frac{\sigma^2}{d}\sum_{i=1}^d \frac{s_i + 1}{(s_i + 1)^2}$$

$$= \frac{\sigma^2}{d}\sum_{i=1}^d \frac{1}{s_i + 1} \overset{(*)}{\geq} \frac{\sigma^2}{\frac{s}{d} + 1} = \frac{\sigma^2 d}{s + d} \geq \frac{\sigma^2 d_\lambda}{s + d_\lambda},$$

where $(*)$ follows by applying Jensen's inequality to convex function $\phi(x) = \frac{1}{x+1}$.

2.3.2 Proof of Part 2

Let $a = \sqrt{2d}$. As shown above, we can assume that $\lambda = 1/(2d)$. Suppose that set S is sampled i.i.d. from some distribution over set $\{1..n\}$. Using standard analysis for the Coupon Collector's problem, it can be shown that if $|S| \leq d(\ln(d) - 1)$, then with probability at least $1/2$ there is $i \in \{1..d\}$ such that $s_i = 0$ (ie, one of the unit vectors $\mathbf{e}_i$ was never selected). Thus, MSPE can be lower-bounded as follows:

$$\mathbb{E}_S\mathbb{E}_{\boldsymbol{\xi}}\frac{1}{n}\|\mathbf{X}^\top(\widehat{\mathbf{w}}_\lambda^*(S) - \mathbf{w}^*)\|^2$$

$$\geq \frac{1}{2}\frac{\sigma^2}{d}\frac{s_i + a^2\lambda^2}{(s_i + \lambda)^2} = \frac{\sigma^2}{2d}\frac{2d\lambda^2}{\lambda^2} = \sigma^2.$$

3 Algorithms

In this section, we present two algorithms that implement λ-regularized volume sampling (3). The first one, RegVol (Algorithm 1), simply adds regularization to the algorithm proposed in [5] for Reverse Iterative Volume Sampling (2). Note that in this algorithm the conditional distribution $P(i\,|\,S)$ is updated at every iteration, which leads to $O((n^2 + d^2)d)$ time complexity. The second algorithm, FastRegVol (Algorithm 2), is our new algorithm which avoids recomputing the conditional distribution at every step, making it essentially as fast as exact leverage score sampling.

The correctness of RegVol follows from Lemma 6, which is a straight-forward application of the Sherman-Morrison formula (see [5] for more details).

Lemma 6 *For any matrix* $\mathbf{X} \in \mathbb{R}^{d \times n}$, *set* $S \subseteq \{1..n\}$ *and two distinct indices* $i, j \in S$, *we have*

$$1 - \mathbf{x}_j^\top \mathbf{Z}_\lambda(S_{-i})^{-1}\mathbf{x}_j = 1 - \mathbf{x}_j^\top \mathbf{Z}_\lambda(S)^{-1}\mathbf{x}_j - a^2,$$

where

$$a = \frac{\mathbf{x}_j^\top \mathbf{Z}_\lambda(S)^{-1}\mathbf{x}_i}{\sqrt{1 - \mathbf{x}_i^\top \mathbf{Z}_\lambda(S)^{-1}\mathbf{x}_i}}.$$

Algorithm 1 RegVol$(\mathbf{X}, s, \lambda)$ (adapted from [5])

1: $\mathbf{Z} \leftarrow (\mathbf{X}\mathbf{X}^\top + \lambda\mathbf{I})^{-1}$
2: $\forall_{i \in \{1..n\}} \quad h_i \leftarrow 1 - \mathbf{x}_i^\top \mathbf{Z}\mathbf{x}_i$
3: $S \leftarrow \{1..n\}$
4: **while** $|S| > s$
5:     Sample $i \propto h_i$ out of S
6:     $S \leftarrow S - \{i\}$
7:     $\mathbf{v} \leftarrow \mathbf{Z}\mathbf{x}_i/\sqrt{h_i}$
8:     $\forall_{j \in S} \quad h_j \leftarrow h_j - (\mathbf{x}_j^\top \mathbf{v})^2$
9:     $\mathbf{Z} \leftarrow \mathbf{Z} + \mathbf{v}\mathbf{v}^\top$
10: **end**
11: **return** S

Algorithm 2 FastRegVol$(\mathbf{X}, s, \lambda)$

1: $\mathbf{Z} \leftarrow (\mathbf{X}\mathbf{X}^\top + \lambda\mathbf{I})^{-1}$
2: $S \leftarrow \{1..n\}$
3: **while** $|S| > s$
4:     **repeat**
5:         Sample i uniformly out of S
6:         $h_i \leftarrow 1 - \mathbf{x}_i^\top \mathbf{Z}\mathbf{x}_i$
7:         Sample $A \sim \text{Bernoulli}(h_i)$
8:     **until** $A = 1$
9:     $S \leftarrow S - \{i\}$
10:     $\mathbf{Z} \leftarrow \mathbf{Z} + h_i^{-1}\mathbf{Z}\mathbf{x}_i\mathbf{x}_i^\top \mathbf{Z}$
11: **end**
12: **return** S

We propose a new volume sampling algorithm, which runs in time $O((n+d)d^2)$, significantly faster than RegVol when $n \gg d$. Our key observation is that updating the full conditional distribution $P(i|S)$ is wasteful, since the distribution changes very slowly throughout the procedure. Moreover, the unnormalized weights h_i, which are computed in the process are all bounded by 1. Thus, to sample from the correct distribution at any given iteration, we can employ rejection sampling as follows:

1. Sample i uniformly from set S,

2. Compute h_i,

3. Accept with probability h_i,

4. Otherwise, draw another sample.

Note that this rejection sampling can be employed locally, within each iteration of the algorithm. Thus, one rejection does not revert us back to the beginning of the algorithm. Moreover, if the probability of acceptance is high, then this strategy requires computing only a small number of weights per iteration of the algorithm, as opposed to updating all of them. This turns out to be the case, as shown below. The full pseudo-code of the sampling procedure is given in Algorithm 2, called FastRegVol.

3.1 Proof of Theorem 4

To simplify the analysis, we combine RegVol and FastRegVol, by running FastRegVol while subset S has size at least $2d$, and then, if a smaller subset is needed, switching to RegVol:

1: $S \leftarrow \text{FastRegVol}(\mathbf{X}, \max\{s, 2d\}, \lambda)$
2: **if** $s < 2d$
3:     $S \leftarrow \text{RegVol}(\mathbf{X}_S, s, \lambda)$
4: **end**
5: **return** S

Following the analysis of [5], time complexity of the RegVol portion of the procedure is $O((2\,d)^2 d) = O(d^3)$. Next, we analyze the efficiency of rejection sampling in the FastRegVol portion. Let R_t be a random variable corresponding to the number of trials needed in the **repeat** loop from line 4 in FastRegVol at the point when $|S| = t$. Note that conditioning on the algorithm's history, R_t is distributed according to geometric distribution $\text{Ge}(q_t)$ with success probability:

$$q_t = \frac{1}{t} \sum_{i \in S} (1 - \mathbf{x}_i^\top \mathbf{Z}_\lambda(S)^{-1} \mathbf{x}_i) \geq \frac{t - d}{t} \geq \frac{1}{2}.$$

Thus, even though variables R_t are not themselves independent, they can be upper-bounded by a sequence of independent variables $\widehat{R}_t \sim \text{Ge}(\frac{t-d}{t})$. The expectation of the total number of trials in FastRegVol, $\bar{R} = \sum_t R_t$, can thus be bounded as follows:

$$\mathbb{E}[\bar{R}] \leq \sum_{t=2d}^{n} \mathbb{E}[\widehat{R}_t] = \sum_{t=2d}^{n} \frac{t}{t - d} \leq 2n.$$

Next, we will obtain a similar bound with high probability instead of in expectation. Here, we will have to use the fact that the variables $\widehat{R}_t$ are independent, which means that we can upper-bound their sum with high probability using standard concentration bounds for geometric distribution. For example, using Corollary 2.2 from [12] one can immediately show that with probability at least $1 - \delta$ we have $\bar{R} = O(n \ln \delta^{-1})$. However, more careful analysis shows an even better dependence on δ.

Lemma 7 *Let $\widehat{R}_t \sim \text{Ge}(\frac{t-d}{t})$ be independent random variables. Then w.p. at least $1 - \delta$*

$$\sum_{t=2d}^{n} \widehat{R}_t = O\left(n + \log\left(\frac{n}{d}\right) \log\left(\frac{1}{\delta}\right)\right).$$

Proof As observed by [12], tail-bounds for the sum of geometric random variables depend on the minimum acceptance probability among those variables. Note that for the vast majority of $\widehat{R}_t$'s the acceptance probability is very close to 1, so intuitively we should be able

to take advantage of this to improve our tail bounds. To that end, we partition the variables into groups of roughly similar acceptance probability and then separately bound the sum of variables in each group. Let $J = \log(\frac{n}{d})$ (w.l.o.g. assume that J is an integer). For $1 \leq j \leq J$, let $I_j = \{d2^j, d2^j + 1, .., d2^{j+1}\}$ represent the j-th partition. We use the following notation for each partition:

$$\bar{R}_j \overset{def}{=} \sum_{t \in I_j} R_t, \qquad \mu_j \overset{def}{=} \mathbb{E}[\bar{R}_j],$$

$$r_j \overset{def}{=} \min_{t \in I_j} \frac{t - d}{t}, \qquad \gamma_j \overset{def}{=} \frac{\log(\delta^{-1})}{d2^{j-2}} + 3.$$

Now, we apply Theorem 2.3 of [12] to $\bar{R}_j$, obtaining

$$P(\bar{R}_j \geq \gamma_j \mu_j) \leq \gamma_j^{-1} (1 - r_j)^{(\gamma_j - 1 - \ln \gamma_j) \mu_j}$$

$$\overset{(1)}{\leq} (1 - r_j)^{\gamma_j \mu_j / 4} \overset{(2)}{\leq} 2^{-j \gamma_j d2^{j-2}},$$

where (1) follows since $\gamma_j \geq 3$, and (2) holds because $\mu_j \geq d2^j$ and $r_j \geq 1 - 2^{-j}$. Moreover, for the chosen γ_j we have

$$j\gamma_j d2^{j-2} = j \log(\delta^{-1}) + 3jd2^{j-2}$$

$$\geq \log(\delta^{-1}) + j = \log(2^j \delta^{-1}).$$

Let A denote the event that $\bar{R}_j \leq \gamma_j \mu_j$ for all $j \leq J$. Applying union bound, we get

$$P(A) \geq 1 - \sum_{j=1}^{J} P(\bar{R}_j \geq \gamma_j \mu_j)$$

$$\geq 1 - \sum_{j=1}^{J} 2^{-\log(2^j \delta^{-1})} = 1 - \sum_{j=1}^{J} \frac{\delta}{2^j} \geq 1 - \delta.$$

If A holds, then we obtain the desired bound:

$$\sum_{t=2d}^{n} \widehat{R}_t \leq \sum_{j=1}^{J} \gamma_j \mu_j \leq \sum_{j=1}^{J} \left(\frac{\log(\delta^{-1})}{d2^{j-2}} + 3\right) d2^{j+1}$$

$$= 8J \log(\delta^{-1}) + 6 \sum_{j=1}^{J} d2^j$$

$$= O\left(\log\left(\frac{n}{d}\right) \log\left(\frac{1}{\delta}\right) + n\right). \qquad \blacksquare$$

Returning to the proof of Theorem 4, we note that each trial of rejection sampling requires computing one weight h_i in time $O(d^2)$. The overall time complexity of FastVol also includes computation and updating of matrix $\mathbf{Z}$ (in time $O((n + d)d^2)$), rejection sampling which takes $O\left(\left(n + \log\left(\frac{n}{d}\right) \log\left(\frac{1}{\delta}\right)\right) d^2\right)$ time, and (if $s < 2d$) the RegVol portion, taking $O(d^3)$ (see [5] for details). This concludes the proof of Theorem 4.

4   Experiments

In this section we experimentally evaluate the proposed algorithms for regularized volume sampling, in terms of runtime and the quality of subsampled ridge estimators. The list of implemented algorithms is:

1. Regularized Volume Sampling:
   (a) FastRegVol – Our new approach (see Alg. 2);
   (b) RegVol – Adapted from [5] (see Alg. 1);

2. Leverage Score Sampling[5] (LSS) – a popular i.i.d. sampling technique [17], where examples are selected w.p. $P(i) = (\mathbf{x}_i^\top (\mathbf{X}\mathbf{X}^\top)^{-1} \mathbf{x}_i)/d$.

The experiments were performed on several benchmark linear regression datasets [16]. Table 2 lists those datasets along with running times for sampling dimension many columns with each method. Dataset MSD was too big for RegVol to finish in reasonable time.

Dataset	$d \times n$	RegVol	FastRegVol	LSS
cadata	$8 \times 21k$	33.5s	0.9s	0.1s
MSD	$90 \times 464k$	>24hr	39s	12s
cpusmall	$12 \times 8k$	1.7s	0.4s	0.07s
abalone	$8 \times 4k$	0.5s	0.2s	0.03s

Table 2: A list of used regression datasets, with runtime comparison between RegVol (see Alg. 1 and [5]) and FastRegVol (see Alg. 2). We also provide the runtime for obtaining exact leverage score samples (LSS).

In Figure 1 we plot the runtime against varying values of n (using portions of the datasets), to compare how FastRegVol and RegVol scale with respect to the datasize. We observe that unlike RegVol, our new algorithm exhibits linear dependence on n, thus it is much better suited for running on large datasets.

4.1   Subset selection for ridge regression

We applied volume sampling to the task of subset selection for linear regression, by evaluating the subsampled ridge estimator $\widehat{\mathbf{w}}_\lambda^*(S)$ using the total loss over the full dataset:

$$L(\widehat{\mathbf{w}}_\lambda^*(S)) \stackrel{def}{=} \frac{1}{n}\|\mathbf{X}^\top \widehat{\mathbf{w}}_\lambda^*(S) - \mathbf{y}\|^2.$$

We computed $L(\widehat{\mathbf{w}}_\lambda^*(S))$ for a range of subset sizes and values of λ, when the subsets are sampled according to λ-regularized volume sampling[6] and leverage score

---

[5]Regularized variants of leverage scores have also been considered in context of kernel ridge regression [1]. However, in our experiments regularizing leverage scores did not provide any improvements.

[6]Our experiments suggest that using the same λ for sampling and for computing the ridge estimator works best.
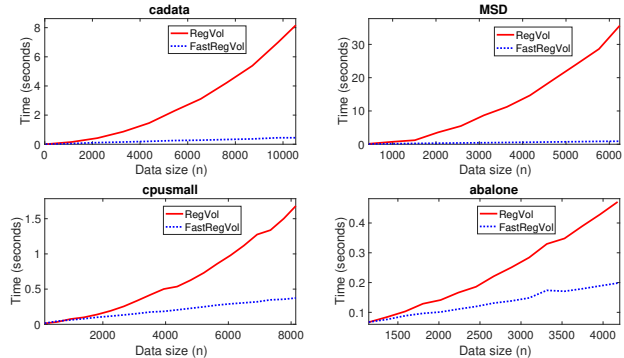

Figure 1: Comparison of runtime between FastRegVol (see Alg. 2) and RegVol [5] on four datasets, with the methods ran on data subsets of varying size (n).
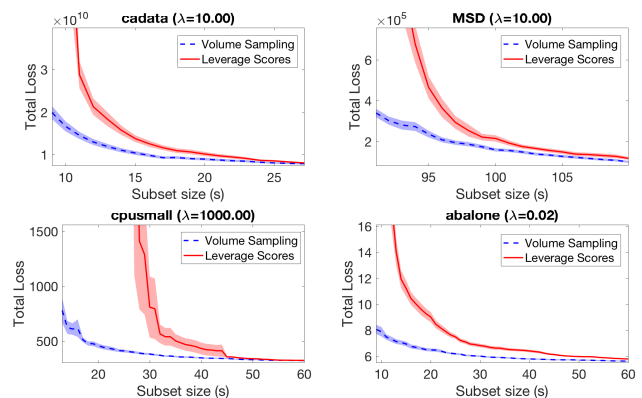

Figure 2: Comparison of loss of the subsampled ridge estimator when using regularized volume sampling vs using leverage score sampling on four datasets.

sampling. The results were averaged over 20 runs of each experiment. For clarity, Figure 2 shows the results only with one value of λ for each dataset, chosen so that the subsampled ridge estimator performed best (on average over all samples of preselected size s). Note that for leverage scores we did the appropriate rescaling of the instances before solving for $\widehat{\mathbf{w}}_\lambda^*(S)$ for the sampled subproblems (see [17] for details). Volume sampling does not require any rescaling. The results on all datasets show that when only a small number of labels s is obtainable, then regularized volume sampling offers better estimators than leverage score sampling (as predicted by Theorems 2 and 3).

5   Conclusions

We proposed a sampling procedure called regularized volume sampling, which offers near-optimal statistical guarantees for subsampled ridge estimators. We also gave a new algorithm for volume sampling which is essentially as efficient as i.i.d. leverage score sampling.

References

[1] Ahmed El Alaoui and Michael W. Mahoney. Fast randomized kernel ridge regression with statistical guarantees. In *Proceedings of the 28th International Conference on Neural Information Processing Systems*, pages 775–783, Montreal, Canada, December 2015.

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

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

[4] Francis Bach. Sharp analysis of low-rank kernel matrix approximations. In *Proceedings of the 26th Annual Conference on Learning Theory*, volume 30 of *Proceedings of Machine Learning Research*, pages 185–209, Princeton, NJ, USA, June 2013.

[5] 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.

[6] 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*, pages 329–338, Las Vegas, USA, October 2010.

[7] 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*, pages 1117–1126, Miami, FL, USA, January 2006.

[8] 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.

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

[10] Mike Gartrell, Ulrich Paquet, and Noam Koenigstein. Bayesian low-rank determinantal point processes. In *Proceedings of the 10th ACM Conference on Recommender Systems*, pages 349–356, Boston, MA, USA, September 2016.

[11] 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*, pages 1207–1214, Kyoto, Japan, January 2012.

[12] Svante Janson. Tail bounds for sums of geometric and exponential variables. *Statistics and Probability Letters*, 135:1 – 6, 2018.

[13] 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, Bellevue, WA, USA, June 2011.

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

[15] Chengtao Li, Stefanie Jegelka, and Suvrit Sra. Polynomial time algorithms for dual volume sampling. In *Advances in Neural Information Processing Systems 30*, pages 5045–5054, Long Beach, CA, USA, December 2017.

[16] M. Lichman. UCI machine learning repository, 2013.

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

[18] Masashi Sugiyama and Shinichi Nakajima. Pool-based active learning in approximate linear regression. *Mach. Learn.*, 75(3):249–274, June 2009.