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 in the sparse label case.

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:

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

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]:

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

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

31st Conference on Neural Information Processing Systems (NIPS 2017), Long Beach, CA, USA.

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:

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

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

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

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

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

(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:

$$\mathbb{E}_S\mathbb{E}_{\boldsymbol{\xi}}\,\frac{1}{n}\|\mathbf{X}^\top(\widehat{\mathbf{w}}_\lambda^*(S) - \mathbf{w}^*)\|^2 = O\left(\frac{\sigma^2 d_\lambda}{s}\right),$$

$$\text{where} \quad d_\lambda = \operatorname{tr}(\mathbf{X}^\top(\mathbf{X}\mathbf{X}^\top + \lambda\mathbf{I})^{-1}\mathbf{X}) \tag{4}$$

FastRegVol: λ-Regularized Volume Sampling

```
Z ← (XX^⊤ + λI)^{-1}
S ← {1..n}
while |S| > s
  repeat
    Sample i uniformly out of S
    h_i ← 1 − x_i^⊤ Z x_i
    Sample A ∼ Bernoulli(h_i)
  until A = 1
  S ← S − {i}
  Z ← Z + h_i^{-1} Z x_i x_i^⊤ Z
end
return S
```

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 good generalization for ridge regression, compared to $O(d_\lambda)$ for regularized volume sampling. We reinforce this claim experimentally in Section 3.

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 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.

[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 $\widetilde{O}(nd + d^3)$.

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, called FastRegVol, 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 $O(n^2d)$ 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

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

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

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], clustering [12], 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 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 $s \geq d$ 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 $\mathbf{X}$.

2 Main results

The main contributions[4] 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 of 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.

[4]Due to space limitations, we omit the proofs of our results in this extended abstract.

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*

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

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

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 which selects 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,*

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

Finally, we propose an algorithm for regularized volume sampling (see FastRegVol in Section 1) 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 sampling, that with probability at least $1 - \delta$ runs in time*[5]

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

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 $\widetilde{O}(nd^2)$.

3 Experiments

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

[5] 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 $O(n^2 d)$ algorithm.

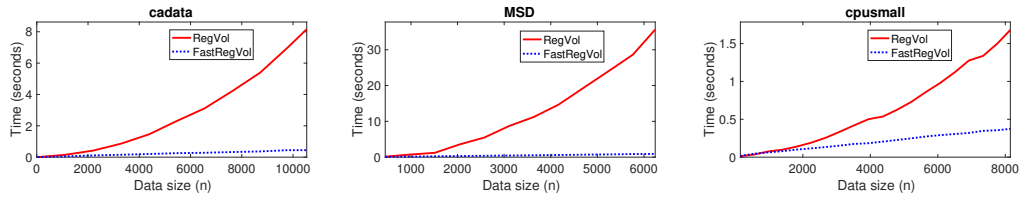

Figure 1: Comparison of runtime between FastRegVol and RegVol (adapted from [5]).

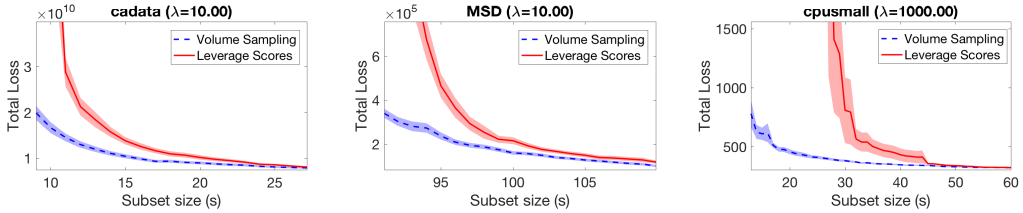

Figure 2: Comparison of loss of the subsampled ridge estimator when using regularized volume sampling vs using leverage score sampling (confidence regions based on standard error of the mean).

1. Regularized Volume Sampling: **FastRegVol** (our algorithm); **RegVol**[6] – adapted from [5];
2. Leverage Score Sampling[7] (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[8].

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.

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

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

3.1 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:

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

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 and leverage score 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). 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).

4 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.

[6] The volume sampling algorithm of [5] can be trivially adapted to the regularized case (i.e. where $\lambda > 0$).

[7] 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.

[8] Dataset MSD was too big for RegVol to finish in reasonable time.

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*, NIPS'15, pages 775–783, Cambridge, MA, USA, 2015. MIT Press.

[2] Zeyuan Allen-Zhu, Yuanzhi Li, Aarti Singh, and Yining Wang. Near-optimal design of experiments via regret minimization. In Doina Precup and Yee Whye Teh, editors, *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 126–135, International Convention Centre, Sydney, Australia, 06–11 Aug 2017. PMLR.

[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 Shai Shalev-Shwartz and Ingo Steinwart, editors, *Proceedings of the 26th Annual Conference on Learning Theory*, volume 30 of *Proceedings of Machine Learning Research*, pages 185–209, Princeton, NJ, USA, 12–14 Jun 2013. PMLR.

[5] Michał Dereziński and Manfred K. Warmuth. Unbiased estimates for linear regression via volume sampling. *CoRR*, abs/1705.06908, 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*, FOCS '10, pages 329–338, Washington, DC, USA, 2010. IEEE Computer Society.

[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*, SODA '06, pages 1117–1126, Philadelphia, PA, USA, 2006. Society for Industrial and Applied Mathematics.

[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] Valeri Vadimovich Fedorov, W.J. Studden, and E.M. Klimko, editors. *Theory of optimal experiments*. Probability and mathematical statistics. Academic Press, New York, 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*, RecSys '16, pages 349–356, New York, NY, USA, 2016. ACM.

[11] Venkatesan Guruswami and Ali Kemal 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.

[12] Byungkon Kang. Fast determinantal point process sampling with application to clustering. In *Proceedings of the 26th International Conference on Neural Information Processing Systems*, NIPS'13, pages 2319–2327, USA, 2013. Curran Associates Inc.

[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. Omnipress, 2011.

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

[15] C. Li, S. Jegelka, and S. Sra. Column Subset Selection via Polynomial Time Dual Volume Sampling. *ArXiv e-prints*, March 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.