Abstract

We study dual volume sampling, a method for selecting $k$ columns from an $n \times m$ short and wide matrix ($n \leq k \leq m$) such that the probability of selection is proportional to the volume spanned by the rows of the induced submatrix. This method was proposed by Avron and Boutsidis (2013), who showed it to be a promising method for column subset selection and its multiple applications. However, its wider adoption has been hampered by the lack of polynomial time sampling algorithms. We remove this hindrance by developing an exact (randomized) polynomial time sampling algorithm as well as its derandomization. Thereafter, we study dual volume sampling via the theory of real stable polynomials and prove that its distribution satisfies the “Strong Rayleigh” property. This result has numerous consequences, including a provably fast-mixing Markov chain sampler that makes dual volume sampling much more attractive to practitioners. This sampler is closely related to classical algorithms for popular experimental design methods that are to date lacking theoretical analysis but are known to empirically work well.

1 Introduction

A variety of applications share the core task of selecting a subset of columns from a short, wide matrix $A$ with $n$ rows and $m > n$ columns. The criteria for selecting these columns typically aim at preserving information about the span of $A$ while generating a well-conditioned submatrix. Classical and recent examples include experimental design, where we select observations or experiments [38]; preconditioning for solving linear systems and constructing low-stretch spanning trees (here $A$ is a version of the node-edge incidence matrix and we select edges in a graph) [6, 4]; matrix approximation [11, 13, 24]; feature selection in $k$-means clustering [10, 12]; sensor selection [25] and graph signal processing [14, 41].

In this work, we study a randomized approach that holds promise for all of these applications. This approach relies on sampling columns of $A$ according to a probability distribution defined over its submatrices: the probability of selecting a set $S$ of $k$ columns from $A$, with $n \leq k \leq m$, is

$$P(S; A) \propto \det(A_S A_S^T),$$

(1.1)

where $A_S$ is the submatrix consisting of the selected columns. This distribution is reminiscent of volume sampling, where $k < n$ columns are selected with probability proportional to the determinant $\det(A_S^T A_S)$ of a $k \times k$ matrix, i.e., the squared volume of the parallelepiped spanned by the selected columns. (Volume sampling does not apply to $k > n$ as the involved determinants vanish.) In contrast, $P(S; A)$ uses the determinant of an $n \times n$ matrix and uses the volume spanned by the rows formed by the selected columns. Hence we refer to $P(S; A)$-sampling as dual volume sampling (DVS).

Contributions. Despite the ostensible similarity between volume sampling and DVS, and despite the many practical implications of DVS outlined below, efficient algorithms for DVS are not known and were raised as open questions in [6]. In this work, we make two key contributions:

- We develop polynomial-time randomized sampling algorithms and their derandomization for DVS. Surprisingly, our proofs require only elementary (but involved) matrix manipulations.
We describe in this section our method to sample from the distribution $P(S; A)$, a remarkable property that captures a specific form of negative dependence. Our proof relies on the theory of real stable polynomials, and the ensuing result implies a provably fast-mixing, practical MCMC sampler. Moreover, this result implies concentration properties for dual volume sampling.

In parallel with our work, [16] also proposed a polynomial time sampling algorithm that works efficiently in practice. Our work goes on to further uncover the hitherto unknown "Strong Rayleigh" property of DVS, which has important consequences, including those noted above.

### 1.1 Connections and implications.

The selection of $k \geq n$ columns from a short and wide matrix has many applications. Our algorithms for DVS hence have several implications and connections; we note a few below.

**Experimental design.** The theory of optimal experiment design explores several criteria for selecting the set of columns (experiments) $S$. Popular choices are $S \in \arg\min_{S \subseteq \{1,\ldots,m\}} J(A_S)$, with $J(A_S) = \|A_S^\dagger\|_F = \|[(A_S A_S^\top)^{-1}]_F\|_F$ (A-optimal design), $J(A_S) = \|A_S^\dagger\|_2$ (E-optimal design), $J(A_S) = -\log \det(A_S A_S^\top)$ (D-optimal design). (1.2)

Here, $A_S^\dagger$ denotes the Moore-Penrose pseudoinverse of $A$, and the minimization ranges over all $S$ such that $A_S$ has full row rank $n$. A-optimal design, for instance, is statistically optimal for linear regression [38].

Finding an optimal solution for these design problems is NP-hard; and most discrete algorithms use local search [33]. Avron and Boutsidis [6, Theorem 3.1] show that dual volume sampling yields an approximation guarantee for both A- and E-optimal design: if $S$ is sampled from $P(S; A)$, then

$$
\mathbb{E} \left[ \|A_S^\dagger\|_F^2 \right] \leq \frac{m-n+1}{m-1} \|A\|_F^2; \quad \mathbb{E} \left[ \|A_S^\dagger\|_2^2 \right] \leq \left( 1 + \frac{n(m-k)}{m-1} \right) \|A\|_2^2. \tag{1.3}
$$

Avron and Boutsidis [6] provide a polynomial time sampling algorithm only for the case $k = n$. Our algorithms achieve the bound (1.3) in expectation, and the derandomization in Section 2.3 achieves the bound deterministically. Wang et al. [43] recently (in parallel) achieved approximation bounds for A-optimality via a different algorithm combining convex relaxation and a greedy method. Other methods include leverage score sampling [30] and predictive length sampling [45].

**Low-stretch spanning trees and applications.** Objectives 1.2 also arise in the construction of low-stretch spanning trees, which have important applications in graph sparsification, preconditioning and solving symmetric diagonally dominant (SDD) linear systems [40], among others [18]. In the node-edge incidence matrix $\Pi \in \mathbb{R}^{n \times m}$ of an undirected graph $G$ with $n$ nodes and $m$ edges, the column corresponding to edge $(u, v)$ is $\sqrt{w(u, v)}(e_u - e_v)$. Let $\Pi = US\Sigma V$ be the SVD of $\Pi$ with $Y \in \mathbb{R}^{n-1 \times m}$. The stretch of a spanning tree $T$ in $G$ is then given by $\text{st}_T(G) = \|Y_T^{-1}\|_F^{-2}$. In those applications, we hence search for a set of edges with low stretch.

**Network controllability.** The problem of sampling $k \geq n$ columns in a matrix also arises in network controllability. For example, Zhao et al. [44] consider selecting control nodes $S$ (under certain constraints) over time in complex networks to control a linear time-invariant network. After transforming the problem into a column subset selection problem from a short and wide controllability matrix, the objective becomes essentially an E-optimal design problem, for which the authors use greedy heuristics.

**Notation.** From a matrix $A \in \mathbb{R}^{n \times m}$ with $m \gg n$ columns, we sample a set $S \subseteq [m]$ of $k$ columns ($n \leq k \leq m$), where $[m] := \{1, 2, \ldots, m\}$. We denote the singular values of $A$ by $\{\sigma_i(A)\}_{i=1}^n$, in decreasing order. We will assume $A$ has full row rank $r(A) = n$, so $\sigma_n(A) > 0$. We also assume that $r(A_S) = r(A) = n$ for every $S \subseteq [m]$ where $|S| \geq n$. By $e_k(A)$, we denote the $k$-th elementary symmetric polynomial of $A$, i.e., the $k$-th coefficient of the characteristic polynomial $\det(\lambda I - A) = \sum_{j=0}^N (-1)^j e_j(A) \lambda^{N-j}$.

## 2 Polynomial-time Dual Volume Sampling

We describe in this section our method to sample from the distribution $P(S; A)$. Our first method relies on the key insight that, as we show, the marginal probabilities for DVS can be computed in polynomial time. To demonstrate this, we begin with the partition function and then derive marginals.
2.1 Marginals

The partition function has a conveniently simple closed form, which follows from the Cauchy-Binet formula and was also derived in [6].

**Lemma 1 (Partition Function [6]).** For $A \in \mathbb{R}^{n \times m}$ with $r(A) = n$ and $n \leq |S| = k \leq m$, we have

$$Z_A := \sum_{|S|=k,S \subseteq [m]} \det(A_S A_S^T) = \frac{(m-n)!}{k^n} \det(AA^T).$$

Next, we will need the marginal probability $P(T \subseteq S; A) = \sum_{S: T \subseteq S} P(S; A)$ that a given set $T \subseteq [m]$ is a subset of the random set $S$. In the following theorem, the set $T_e = [m] \setminus T$ denotes the (set) complement of $T$, and $Q^\perp$ denotes the orthogonal complement of $Q$.

**Theorem 2 (Marginals).** Let $T \subseteq [m], |T| \leq k,$ and $\varepsilon > 0$. Let $A_T = Q\Sigma V^T$ be the singular value decomposition of $A_T$ where $Q \in \mathbb{R}^{n \times r(A_T)}$, and $Q^\perp \in \mathbb{R}^{n \times (n-r(A_T))}$. Further define the matrices

$$B = (Q^\perp)^T A_{T_e} \in \mathbb{R}^{(n-r(A_T)) \times (m-|T|)},$$

$$C = \left[ \begin{array}{ccc} \frac{1}{\sqrt{\sigma_i^2(A_T) + \varepsilon}} & 0 & \cdots \\ 0 & \frac{1}{\sqrt{\sigma_i^2(A_T) + \varepsilon}} & \vdots \\ \vdots & \vdots & \ddots \end{array} \right] \in \mathbb{R}^{m-|T| \times m-|T|}.$$  

Let $Q_B \text{diag}(\sigma_i^2(B)) Q_B^T$ be the eigenvalue decomposition of $B^T B$ where $Q_B \in \mathbb{R}^{|T_e| \times r(B)}$. Moreover, let $W^T = (I_{T_e}; C^T)$ and $\Gamma = e_k - (|T_e| - r(B))(W((Q_B^T)^T Q_B^T)W^T)$. Then the marginal probability of $T$ in DVS is

$$P(T \subseteq S; A) = \frac{\prod_{i=1}^{r(A_T)} \sigma_i^2(A_T) \times \prod_{j=1}^{r(B)} \sigma_j^2(B)}{Z_A} \times \Gamma.$$  

We prove Theorem 2 via a perturbation argument that connects DVS to volume sampling. Specifically, observe that for $\varepsilon > 0$ and $|S| \geq n$ it holds that

$$\det(A_S A_S^T + \varepsilon I_n) = \varepsilon^{-k} \det(A_S A_S^T + \varepsilon I_k) = \varepsilon^{-k} \det \left( \left[ \begin{array}{c} A_S \\ \sqrt{\varepsilon}(I_m) \end{array} \right] \left[ \begin{array}{c} A_S^T \\ \sqrt{\varepsilon}(I_m) \end{array} \right] \right). \quad (2.1)$$

Carefully letting $\varepsilon \to 0$ bridges volumes with “dual” volumes. The technical remainder of the proof further relates this equality to singular values, and exploits properties of characteristic polynomials. A similar argument yields an alternative proof of Lemma 1. We show the proofs in detail in Appendix A and B respectively.

**Complexity.** The numerator of $P(T \subseteq S; A)$ in Theorem 2 requires $O(mn^2)$ time to compute the first term, $O(mn^2)$ to compute the second and $O(m^3)$ to compute the third. The denominator takes $O(mn^2)$ time, amounting in a total time of $O(m^3)$ to compute the marginal probability.

2.2 Sampling

The marginal probabilities derived above directly yield a polynomial-time exact DVS algorithm. Instead of $k$-sets, we sample ordered $k$-tuples $\vec{S} = (s_1, \ldots, s_k) \in [m]^k$. We denote the $k$-tuple variant of the DVS distribution by $\vec{P}(-; A)$:

$$\vec{P}(\vec{s} = (i)_{k=1}^k; A) = \frac{1}{k!} P(\{i_1, \ldots, i_k\}; A) = \prod_{j=1}^k \vec{P}(s_j = i_j| s_1 = i_1, \ldots, s_{j-1} = i_{j-1}; A).$$

Sampling $\vec{S}$ is now straightforward. At the $j$th step we sample $s_j$ via $\vec{P}(s_j = i_j| s_1 = i_1, \ldots, s_{j-1} = i_{j-1}; A)$; these probabilities are easily obtained from the marginals in Theorem 2.

**Corollary 3.** Let $T = \{i_1, \ldots, i_{t-1}\}$, and $P(T \subseteq S; A)$ as in Theorem 2. Then,

$$\vec{P}(s_t = i; A| s_1 = i_1, \ldots, s_{t-1} = i_{t-1}) = \frac{P(T \cup \{i\} \subseteq S; A)}{(k-t+1)P(T \subseteq S; A)}.$$  

As a result, it is possible to draw an exact dual volume sample in time $O(km^4)$.

The full proof may be found in the appendix. The running time claim follows since the sampling algorithm invokes $O(mk)$ computations of marginal probabilities, each costing $O(m^3)$ time.
We show the full derivation in Appendix Theorem 5.

The class of SR measures captures a strong notion of negative association and provides a property that occurs more broadly across probability, combinatorics, physics, and other fields—see increasing functions on probability measure. For DVS, the SR property has two major consequences: it leads to a fast mixing practical MCMC sampling in a variety of recent contexts, including recent progress in approximation algorithms. namely, the Strongly Rayleigh (SR) property. This property has proved remarkably fruitful in a variety of recent contexts, including recent progress in approximation algorithms, fast sampling, graph sparsification, extensions to the Kadison-Singer problem, and certain concentration of measure results, among others.

For DVS, the SR property has two major consequences: it leads to a fast mixing practical MCMC sampler, and it implies results on concentration of measure.

Strongly Rayleigh measures. SR measures were introduced in the landmark paper of Borcea et al. [8], who develop a rich theory of negatively associated measures. In particular, we say that a probability measure $\mu : 2^{[n]} \to \mathbb{R}$ is negatively associated if $\int F d\mu \geq \int G d\mu$ for increasing functions on $2^{[n]}$ with disjoint support. This property reflects a “repelling” nature of $\mu$, a property that occurs more broadly across probability, combinatorics, physics, and other fields—see [36, 8, 42] and references therein. The negative association property turns out to be quite subtle in general; the class of SR measures captures a strong notion of negative association and provides a framework for analyzing such measures.

Remark. A potentially more efficient approximate algorithm could be derived by noting the relations between volume sampling and DVS. Specifically, we add a small perturbation to DVS as in Equation 2.1 to transform it into a volume sampling problem, and apply random projection for more efficient volume sampling as in [17]. Please refer to Appendix C for more details.

2.3 Derandomization

Next, we derandomize the above sampling algorithm to deterministically select a subset that satisfies the bound (1.3) for the Frobenius norm, thereby answering another question in [6]. The key insight for derandomization is that conditional expectations can be computed in polynomial time, given the marginals in Theorem 2:

Corollary 4. Let $(i_1, \ldots, i_{t-1}) \in [m]^{t-1}$ be such that the marginal distribution satisfies $\overrightarrow{P}(s_1 = i_1, \ldots, s_{t-1} = i_{t-1}; A) > 0$. The conditional expectation can be expressed as

$$E \left[ \|A^0_s\|_F^2 \mid s_1 = i_1, \ldots, s_{t-1} = i_{t-1} \right] = \sum_{j=1}^n P'({i_1, \ldots, i_{t-1} \in S|S \sim P(S; A_{i_{t-1}}\backslash(j))})$$

where $P'$ are the unnormalized marginal distributions, and it can be computed in $\mathcal{O}(nm^3)$ time.

We show the full derivation in Appendix D.

Corollary 4 enables a greedy derandomization procedure. Starting with the empty tuple $\overrightarrow{S}_0 = \emptyset$, in the $i$th iteration, we greedily select $j^* \in \operatorname{argmax}_j E[\|A^0_{s_{i-1} \cup j}\|_F^2 \mid (s_1, \ldots, s_i) = \overrightarrow{S}_{i-1} \cup j]$ and append it to our selection: $\overrightarrow{S}_i = \overrightarrow{S}_{i-1} \cup j$. The final set is the non-ordered version $S_k$ of $\overrightarrow{S}_k$. Theorem 5 shows that this greedy procedure succeeds, and implies a deterministic version of the bound (1.3).

Theorem 5. The greedy derandomization selects a column set $S$ satisfying

$$\|A^1_S\|_F^2 \leq \frac{m - n + 1}{k - n + 1} \|A^1\|_F^2; \quad \|A^1_S\|_2^2 \leq \frac{n(m - n + 1)}{k - n + 1} \|A^1\|_2^2.$$

In the proof, we construct a greedy algorithm. In each iteration, the algorithm computes, for each column that has not yet been selected, the expectation conditioned on this column being included in the current set. Then it chooses the element with the lowest conditional expectation to actually be added to the current set. This greedy inclusion of elements will only decrease the conditional expectation, thus retaining the bound in Theorem 5. The detailed proof is deferred to Appendix E.

Complexity. Each iteration of the greedy selection requires $\mathcal{O}(nm^3)$ to compute $\mathcal{O}(m)$ conditional expectations. Thus, the total running time for $k$ iterations is $\mathcal{O}(km^4)$. The approximation bound for the spectral norm is slightly worse than that in (1.3), but is of the same order if $k = \mathcal{O}(n)$.

3 Strong Rayleigh Property and Fast Markov Chain Sampling

Next, we investigate DVS more deeply and discover that it possesses a remarkable structural property, namely, the Strongly Rayleigh (SR) property. This property has proved remarkably fruitful in a variety of recent contexts, including recent progress in approximation algorithms, fast sampling, graph sparsification, extensions to the Kadison-Singer problem, and certain concentration of measure results, among others.

For DVS, the SR property has two major consequences: it leads to a fast mixing practical MCMC sampler, and it implies results on concentration of measure.
Specifically, SR measures are defined via their connection to real stable polynomials [36, 8, 42]. A multivariate polynomial $f \in \mathbb{C}[z]$ where $z \in \mathbb{C}^m$ is called real stable if all its coefficients are real and $f(z) \neq 0$ whenever $\Re(z_i) > 0$ for $1 \leq i \leq m$. A measure is called an SR measure if its multivariate generating polynomial $p_c(z) := \sum_{S \subseteq [m]} \mu(S) \prod_{i \in S} z_i$ is real stable. Notable examples of SR measures are Determinantal Point Processes [31, 29, 9, 26], balanced matroids [19, 37], Bernoulli conditioned on their sum, among others. It is known (see [8, pg. 523]) that the class of SR measures is exponentially larger than the class of determinantal measures.

### 3.1 Strong Rayleigh Property of DVS

Theorem 6 establishes the SR property for DVS and is the main result of this section. Here and in the following, we use the notation $z^S = \prod_{i \in S} z_i$.

**Theorem 6.** Let $A \in \mathbb{R}^{n \times m}$ and $n \leq k \leq m$. Then the multiaffine polynomial

$$p(z) := \sum_{|S|=k, S \subseteq [m]} \det(A_S A_S^T) \prod_{i \in S} z_i = \sum_{|S|=k, S \subseteq [m]} \det(A_S A_S^T) z^S,$$

(3.1)

is real stable. Consequently, $P(S; A)$ is an SR measure.

The proof of Theorem 6 relies on key properties of real stable polynomials and SR measures established in [8]. Essentially, the proof demonstrates that the generating polynomial of $\overline{P}(S; c; A)$ can be obtained by applying a few carefully chosen stability preserving operations to a polynomial that we know to be real stable. Stability, although easily destroyed, is closed under several operations noted in the important proposition below.

**Proposition 7** (Prop. 2.1 [8]). Let $f : \mathbb{C}^m \to \mathbb{C}$ be a stable polynomial. The following properties preserve stability: (i) **Substitution**: $f(\mu, z_2, \ldots, z_m)$ for $\mu \in \mathbb{R}$; (ii) **Differentiation**: $\partial^S f(z_1, \ldots, z_m)$ for any $S \subseteq [m]$; (iii) **Diagonalization**: $f(z, z, z_3, \ldots, z_m)$ is stable, and hence $f(z, z, \ldots, z)$; and (iv) **Inversion**: $z_1 \cdots z_m f(z_1^{-1}, \ldots, z_m^{-1})$.

In addition, we need the following two propositions for proving Theorem 6.

**Proposition 8** (Prop. 2.4 [7]). Let $B$ be Hermitian, $z \in \mathbb{C}^m$ and $A_i (1 \leq i \leq m)$ be Hermitian semidefinite matrices. Then, the following polynomial is stable:

$$f(z) := \det(B + \sum_i z_i A_i).$$

(3.2)

**Proposition 9.** For $n \leq |S| \leq m$ and $L := A^T A$, we have $\det(A_S A_S^T) = e_n(L_{S,S})$.

**Proof.** Let $Y = \text{Diag}([y_i]_{i=1}^m)$ be a diagonal matrix. Using the Cauchy-Binet identity we have

$$\det(A Y A^T) = \sum_{|T| = n, T \subseteq [m]} \det((A Y)_{T,:}) \det((A^T)_{T,:}) = \sum_{|T| = n, T \subseteq [m]} \det(A_T^T A_T) y^T.$$

Thus, when $Y = I_S$, the (diagonal) indicator matrix for $S$, we obtain $A Y A^T = A_S A_S^T$. Consequently, in the summation above only terms with $T \subseteq S$ survive, yielding

$$\det(A_S A_S^T) = \sum_{|T| = n, T \subseteq S} \det(A_T^T A_T) = \sum_{|T| = n, T \subseteq S} \det(L_{T,T}) = e_n(L_{S,S}). \quad \square$$

We are now ready to sketch the proof of Theorem 6.

**Proof.** (Theorem 6). Notationally, it is more convenient to prove that the “complement” polynomial $p_c(z) := \sum_{|S|=k, S \subseteq [m]} \det(A_S A_S^T) z^{S^c}$ is stable; subsequently, an application of Prop. 7-(iv) yields stability of (3.1). Using matrix notation $W = \text{Diag}(w_1, \ldots, w_m)$, $Z = \text{Diag}(z_1, \ldots, z_m)$, our starting stable polynomial (this stability follows from Prop. 8) is

$$h(z, w) := \det(L + W + Z), \quad w \in \mathbb{C}^m, \ z \in \mathbb{C}^m,$$

which can be expanded as

$$h(z, w) = \sum_{S \subseteq [m]} \det(W_S + L_S) z^{S^c} = \sum_{S \subseteq [m]} \left( \sum_{T \subseteq S} w^{S \setminus T} \det(L_{T,T}) \right) z^{S^c}.$$
Thus, \( h(z, w) \) is real stable in \( 2m \) variables, indexed below by \( S \) and \( R \) where \( R := S^c \cap T \). Instead of the form above, We can sum over \( S, R \subseteq [m] \) but then have to constrain the support to the case when \( S_c \cap T = \emptyset \) and \( S_c \cap R = \emptyset \). In other words, we may write (using Iverson-brackets \([]\))

\[
h(z, w) = \sum_{S, R \subseteq [m]} \left[ S_c \cap R = \emptyset \wedge S_c \cap T = \emptyset \right] \det(L_{T,T}) z^{S_c} w^R.
\] (3.3)

Next, we truncate polynomial (3.3) at degree \((m-k)+(k-n) = m-n\) by restricting \(|S_c \cup R| = m-n\).

By [8, Corollary 4.18] this truncation preserves stability, whence

\[
H(z, w) := \sum_{S, R \subseteq [m]} \left[ S_c \cap R = \emptyset \right] \det(L_{S \setminus R, S \setminus R}) z^{S_c} w^R,
\]

is also stable. Using Prop. 7-(iii), setting \( w_1 = \ldots = w_m = y \) retains stability; thus

\[
g(z, y) := H(z, (y, y, \ldots, y)) = \sum_{S, R \subseteq [m]} \left[ S_c \cap R = \emptyset \right] \det(L_{S \setminus R, S \setminus R}) z^{S_c} y^{|R|}
\]

\[
= \sum_{S \subseteq [m]} \left( \sum_{|T|=n, T \subseteq S} \det(L_{T,T}) y^{|S| - |T|} z^{S_c} = \sum_{S \subseteq [m]} e_n(L_{S,S}) y^{|S| - n} z^{S_c},
\]

is also stable. Next, differentiating \( g(z, y) \), \( k-n \) times with respect to \( y \) and evaluating at \( 0 \) preserves stability (Prop. 7-(ii) and (i)). In doing so, only terms corresponding to \(|S| = k\) survive, resulting in

\[
\frac{\partial^{k-n} g(z, y)}{\partial y^{k-n}} \bigg|_{y=0} = (k-n)! \sum_{|S|=k, S \subseteq [m]} e_n(L_{S,S}) z^{S_c} = (k-n)! \sum_{|S|=k, S \subseteq [m]} \det(A_S A_S^T) z^{S_c},
\]

which is just \( p_c(z) \) (up to a constant); here, the last equality follows from Prop. 9. This establishes stability of \( p_c(z) \) and hence of \( p(z) \). Since \( p(z) \) is in addition multiaffine, it is the generating polynomial of an SR measure, completing the proof.

### 3.2 Implications: MCMC

The SR property of \( P(S; A) \) established in Theorem 6 implies a fast mixing Markov chain for sampling \( S \). The states for the Markov chain are all sets of cardinality \( k \). The chain starts with a randomly-initialized active set \( S \), and in each iteration we swap an element \( s^{in} \in S \) with an element \( s^{out} \notin S \) with a specific probability determined by the probability of the current and proposed set. The stationary distribution of this chain is the one induced by DVS, by a simple detailed-balance argument. The chain is shown in Algorithm 1.

**Algorithm 1** Markov Chain for Dual Volume Sampling

**Input:** \( A \in \mathbb{R}^{n \times m} \) the matrix of interest, \( k \) the target cardinality, \( T \) the number of steps

**Output:** \( S \sim P(S; A) \)

Initialize \( S \subseteq [m] \) such that \(|S| = k \) and \( \det(A_S A_S^T) > 0 \)

for \( i = 1 \) to \( T \) do

\[ \text{draw } b \in \{0, 1\} \text{ uniformly} \]

if \( b = 1 \) then

\[ \text{Pick } s^{in} \in S \text{ and } s^{out} \in [m] \setminus S \text{ uniformly randomly} \]

\[ \begin{align*}
q(s^{in}, s^{out}, S) & \leftarrow \min \left\{ 1, \frac{\det(A_{S \cup \{s^{out}\} \setminus \{s^{in}\}} A_{S \cup \{s^{out}\} \setminus \{s^{in}\}}^T)}{\det(A_S A_S^T)} \right\} \\
S & \leftarrow S \cup \{s^{out}\} \setminus \{s^{in}\} \text{ with probability } q(s^{in}, s^{out}, S)
\end{align*} \]

end if

end for

The convergence of the Markov chain is measured via its mixing time: The *mixing time* of the chain indicates the number of iterations \( t \) that we must perform (starting from \( S_0 \)) before we can consider \( S_t \) as an approximately valid sample from \( P(S; A) \). Formally, if \( \delta_{S_0}(t) \) is the total variation distance between the distribution of \( S_t \) and \( P(S; A) \) after \( t \) steps, then

\[ \tau_{S_0}(\varepsilon) := \min \left\{ t : \delta_{S_0}(t') \leq \varepsilon, \forall t' \geq t \right\} \]
We report selection performance of DVS on real regression data (CompAct, CompAct(s), Abalone) and compare against various baselines, including uniform sampling (Unif), leverage score sampling (Lev) [30], predictive length sampling (PL) [45], the sampling (Smpl)/greedy (Greedy) selection methods in [43] and Fedorov’s exchange algorithm [20]. We initialize the MCMC sampler with kmeans++ [5] for DVS and run for 10,000 iterations, which empirically yields selections that are

\[^1\text{http://www.dcc.fc.up.pt/~ltorgo/Regression/DataSets.html}\]
sufficiently good. We measure performances via (1) the prediction error $\|y - X\hat{\alpha}\|$, and 2) running times. Figure 1 shows the results for these three measures with sample sizes $k$ varying from 60 to 200. Further experiments (including for the interpolation $\beta < 1$), may be found in the appendix.

![Figure 1: Results on the CompAct(s) dataset. Results are the median of 10 runs, except Greedy and Fedorov. Note that Unif, Lev, PL and DVS use less than 1 second to finish experiments.](image)

In terms of prediction error, DVS performs well and is comparable with Lev. Its strength compared to the greedy and relaxation methods (Smpl, Greedy, Fedorov) is running time, leading to good time-error tradeoffs. These tradeoffs are illustrated in Figure 1 for $k = 120$.

In other experiments (shown in Appendix G) we observed that in some cases, the optimization and greedy methods (Smpl, Greedy, Fedorov) yield better results than sampling, however with much higher running times. Hence, given time-error tradeoffs, DVS may be an interesting alternative in situations where time is a very limited resource and results are needed quickly.

5 Conclusion

In this paper, we study the problem of DVS and develop an exact (randomized) polynomial time sampling algorithm as well as its derandomization. We further study dual volume sampling via the theory of real-stable polynomials and prove that its distribution satisfies the “Strong Rayleigh” property. This result has remarkable consequences, especially because it implies a provably fast-mixing Markov chain sampler that makes dual volume sampling much more attractive to practitioners. Finally, we observe connections to classical, computationally more expensive experimental design methods (Fedorov’s method and SA); together with our results here, these could be a first step towards a better theoretical understanding of those methods.

Acknowledgement

This research was supported by NSF CAREER award 1553284, NSF grant IIS-1409802, DARPA grant N66001-17-1-4039, DARPA FunLoL grant (W911NF-16-1-0551) and a Siebel Scholar Fellowship. The views, opinions, and/or findings contained in this article are those of the author and should not be interpreted as representing the official views or policies, either expressed or implied, of the Defense Advanced Research Projects Agency or the Department of Defense.

References


