Bayesian experimental design using regularized determinantal point processes

Anonymous Author(s)
Affiliation
Address
email

Abstract

In experimental design, we are given n vectors in d dimensions, and our goal is to select $k \ll n$ of them to perform expensive measurements, e.g., to obtain labels/responses, for a linear regression task. Many statistical criteria have been proposed for choosing the optimal design, with popular choices including A- and D-optimality. If prior knowledge is given, typically in the form of a $d \times d$ precision matrix $\mathbf{A}$, then all of the criteria can be extended to incorporate that information via a Bayesian framework. In this paper, we demonstrate a new fundamental connection between Bayesian experimental design and determinantal point processes, the latter being widely used for sampling diverse subsets of data. We use this connection to develop new efficient algorithms for finding $(1+\epsilon)$-approximations of optimal designs under four optimality criteria: A, C, D and V. Our algorithms can achieve this when the desired subset size k is $\Omega(\frac{d_{\mathbf{A}}}{\epsilon} + \frac{\log 1/\epsilon}{\epsilon^2})$, where $d_{\mathbf{A}} \leq d$ is the $\mathbf{A}$-effective dimension, which can often be much smaller than d. Our results offer direct improvements over a number of prior works, for both Bayesian and classical experimental design, in terms of algorithm efficiency, approximation quality, and range of applicable criteria.

1 Introduction

Consider a collection of n experiments parameterized by d-dimensional vectors $\mathbf{x}_1, \ldots, \mathbf{x}_n$, and let $\mathbf{X}$ denote the $n \times d$ matrix with rows $\mathbf{x}_i^\top$. The outcome of the ith experiment is a random variable $y_i = \mathbf{x}_i^\top \mathbf{w} + \xi_i$, where $\mathbf{w}$ is the parameter vector of a linear model with prior distribution $\mathcal{N}(\mathbf{0}, \sigma^2 \mathbf{A}^{-1})$, and $\xi_i \sim \mathcal{N}(0, \sigma^2)$ is independent noise. In experimental design, we have access to the vectors $\mathbf{x}_i^\top$, for $i \in \{1, \ldots, n\}$, but we are allowed to observe only a small number of outcomes y_i for experiments we choose. Suppose that we observe the outcomes from a subset $S \subseteq \{1, ..., n\}$ of k experiments. The posterior distribution of $\mathbf{w}$ given $\mathbf{y}_S$ (the vector of outcomes in S) is:

$$\mathbf{w} \mid \mathbf{y}_S \;\sim\; \mathcal{N}\Big((\mathbf{X}_S^\top \mathbf{X}_S + \mathbf{A})^{-1} \mathbf{X}_S^\top \mathbf{y}_S,\;\; \sigma^2 (\mathbf{X}_S^\top \mathbf{X}_S + \mathbf{A})^{-1} \Big),$$

where $\mathbf{X}_S$ denotes the $k \times d$ matrix with rows $\mathbf{x}_i^\top$ for $i \in S$. In the Bayesian framework of experimental design [CV95], we assume that the prior precision matrix $\mathbf{A}$ of the linear model $\mathbf{w}$ is known, and our goal is to choose S so as to minimize some quantity (a.k.a. an optimality criterion) measuring the "size" of the posterior covariance matrix $\mathbf{\Sigma}_{\mathbf{w}|\mathbf{y}_S} = \sigma^2 (\mathbf{X}_S^\top \mathbf{X}_S + \mathbf{A})^{-1}$. This quantity is a function of the subset covariance $\mathbf{X}_S^\top \mathbf{X}_S$. Note that if matrix $\mathbf{A}$ is non-invertible then, even though the prior distribution is ill-defined, we can still interpret it as having no prior information in the directions with eigenvalue 0. In particular, for $\mathbf{A} = \mathbf{0}$ we recover classical experimental design, where the covariance matrix of $\mathbf{w}$ given $\mathbf{y}_S$ is $\sigma^2 (\mathbf{X}_S^\top \mathbf{X}_S)^{-1}$. We also do not need matrix $\mathbf{X}$ to necessarily be full rank. For the Bayesian experimental design problem to be well-defined it suffices that there exists a subset S of size k such that matrix $\mathbf{X}_S^\top \mathbf{X}_S + \mathbf{A}$ is invertible. We will write

Submitted to 33rd Conference on Neural Information Processing Systems (NeurIPS 2019). Do not distribute.

the Bayesian optimality criteria as functions $f_{\mathbf{A}}(\mathbf{\Sigma})$, where $\mathbf{\Sigma}$ corresponds to the subset covariance $\mathbf{X}_S^\top \mathbf{X}_S$. The following standard criteria [Puk06] are of primary interest to us:

1. A-optimality: $f_{\mathbf{A}}(\mathbf{\Sigma}) = \mathrm{tr}\big((\mathbf{\Sigma}+\mathbf{A})^{-1}\big)$;
2. C-optimality: $f_{\mathbf{A}}(\mathbf{\Sigma}) = \mathbf{c}^\top(\mathbf{\Sigma}+\mathbf{A})^{-1}\mathbf{c}$ for some vector $\mathbf{c}$;
3. D-optimality: $f_{\mathbf{A}}(\mathbf{\Sigma}) = \det(\mathbf{\Sigma}+\mathbf{A})^{-1/d}$;
4. V-optimality: $f_{\mathbf{A}}(\mathbf{\Sigma}) = \frac{1}{n}\mathrm{tr}\big(\mathbf{X}(\mathbf{\Sigma}+\mathbf{A})^{-1}\mathbf{X}^\top\big)$.

Other popular criteria (less relevant to our discussion) include E-optimality, $f_{\mathbf{A}}(\mathbf{\Sigma}) = \|(\mathbf{\Sigma}+\mathbf{A})^{-1}\|$ (here, $\|\cdot\|$ denotes the spectral norm) and G-optimality, $f_{\mathbf{A}}(\mathbf{\Sigma}) = \max \mathrm{diag}(\mathbf{X}(\mathbf{\Sigma}+\mathbf{A})^{-1}\mathbf{X}^\top)$.

The general task we consider is given as follows, where $[n]$ denotes $\{1, ..., n\}$:

Bayesian experimental design. Given an $n \times d$ matrix $\mathbf{X}$, an optimality criterion $f_{\mathbf{A}}(\cdot)$ and $k \in [n]$,

$$\text{efficiently minimize} \quad f_{\mathbf{A}}(\mathbf{X}_S^\top \mathbf{X}_S) \quad \text{over } S \subseteq [n] \text{ s.t. } |S| = k.$$

Optimal value. Given $\mathbf{X}$, $f_{\mathbf{A}}$ and k, we denote the optimum as $\mathrm{OPT}_k = \min_{S:|S|=k} f_{\mathbf{A}}(\mathbf{X}_S^\top \mathbf{X}_S)$.

The prior work around this problem can be grouped into two research questions. The first question asks what can we infer about OPT_k just from the spectral information about the problem, which is contained in the data covariance matrix $\mathbf{\Sigma}_{\mathbf{X}} = \mathbf{X}^\top\mathbf{X} \in \mathbb{R}^{d\times d}$. The second question asks when does there exist a polynomial time algorithm for finding a $(1+\epsilon)$-approximation for OPT_k.

Question 1: Given only $\mathbf{\Sigma}_{\mathbf{X}}$, $f_{\mathbf{A}}$ and k, what is the upper bound on OPT_k?
Question 2: Given $\mathbf{X}$, $f_{\mathbf{A}}$ and k, is there an efficient $(1+\epsilon)$-approximation algorithm for OPT_k?

A key aspect of both of these questions is how large the subset size k has to be for us to provide useful answers. As discussed in prior work [AZLSW17], we should expect meaningful results when k is at least $\Omega(d)$, and in fact, for classical experimental design (i.e., when $\mathbf{A} = \mathbf{0}$), the problem becomes ill-defined when $k < d$. In the Bayesian setting we can exploit the additional prior knowledge to achieve strong results even for $k \ll d$. Intuitively, the larger the prior precision matrix $\mathbf{A}$, the fewer degrees of freedom we have in the problem. To measure this, we use the statistical notion of *effective dimension* [AM15].

Definition 1 *For $d \times d$ psd matrices $\mathbf{A}$ and $\mathbf{\Sigma}$, let the $\mathbf{A}$-effective dimension of $\mathbf{\Sigma}$ be defined as $d_{\mathbf{A}}(\mathbf{\Sigma}) = \mathrm{tr}\big(\mathbf{\Sigma}(\mathbf{\Sigma}+\mathbf{A})^{-1}\big) \leq d$. We will use the shorthand $d_{\mathbf{A}}$ when referring to $d_{\mathbf{A}}(\mathbf{\Sigma}_{\mathbf{X}})$.*

Recently, [DW18b] obtained bounds on Bayesian A/V-optimality criteria for $k \geq d_{\mathbf{A}}$, suggesting that $d_{\mathbf{A}}$ is the right notion of degrees of freedom for this problem. We argue that $d_{\mathbf{A}}$ can in fact be far too large of an estimate because it does not take into account the size k when computing the effective dimension. Intuitively, since $d_{\mathbf{A}}$ is computed using the full data covariance $\mathbf{\Sigma}_{\mathbf{X}}$, it is not in the appropriate scale with respect to the smaller covariance $\mathbf{X}_S^\top\mathbf{X}_S$. One way to correct this is to rescale the full covariance and use $d_{\mathbf{A}}(\frac{k}{n}\mathbf{\Sigma}_{\mathbf{X}})$ as the degrees of freedom, which can also be written as $d_{\frac{n}{k}\mathbf{A}}$. Note that $d_{\frac{n}{k}\mathbf{A}} \leq d_{\mathbf{A}}$ and this gap can be very large for some problems (see discussion in Appendix B). The following result supports the above reasoning by showing that for any k such that $k \geq 4d_{\frac{n}{k}\mathbf{A}}$, there is S of size k which satisfies $f_{\mathbf{A}}(\mathbf{X}_S^\top\mathbf{X}_S) = O(1)\cdot f_{\mathbf{A}}(\frac{k}{n}\mathbf{\Sigma}_{\mathbf{X}})$. This not only improves on [DW18b] in terms of the supported range of sizes k, but also in terms of the obtained bound (see Section 2 for a comparison).

Theorem 1 *Let $f_{\mathbf{A}}$ be A/C/D/V-optimality and $\mathbf{X}$ be $n \times d$. For any k such that $k \geq 4d_{\frac{n}{k}\mathbf{A}}$, we have*

$$\mathrm{OPT}_k \leq \left(1 + 8\,\frac{d_{\frac{n}{k}\mathbf{A}}}{k} + 8\sqrt{\frac{\ln(k/d_{\frac{n}{k}\mathbf{A}})}{k}}\right)\cdot f_{\mathbf{A}}\big(\tfrac{k}{n}\mathbf{\Sigma}_{\mathbf{X}}\big).$$

Remark 2 *We give an $O(ndk + k^2d^2)$ time algorithm for finding subset S that certifies this bound.*

To establish Theorem 1, we propose a new sampling distribution $\mathrm{DPP}^p_{\mathrm{reg}}(\mathbf{X}, \mathbf{A})$, where $p = (p_1, ..., p_n) \in [0,1]^n$ is a vector of weights. This is a special *regularized* variant of a determinantal point process (DPP), which is a well-studied family of distributions [KT12] with numerous

applications in sampling diverse subsets of elements. Given a psd matrix $\mathbf{A}$ and a weight vector p, we define $\mathrm{DPP}^p_{\mathrm{reg}}(\mathbf{X}, \mathbf{A})$ as a distribution over subsets $S \subseteq [n]$ (of all sizes) such that:

$$\text{(see Def. 2)} \qquad \Pr(S) \propto \det(\mathbf{X}_S^\top \mathbf{X}_S + \mathbf{A}) \cdot \prod_{i \in S} p_i \cdot \prod_{i \notin S} (1 - p_i).$$

A number of regularized DPPs have been proposed recently [Der18, DW18b], mostly within the context of Randomized Numerical Linear Algebra (RandNLA) [Mah11, DM16, DM17]. To our knowledge, ours is the first such definition that strictly falls under the traditional definition of a DPP [KT12]. We show this in Section 3, where we also prove that regularized DPPs can be decomposed into a low-rank DPP plus i.i.d. Bernoulli sampling (Theorem 5). This decomposition reduces the sampling cost from $O(n^3)$ to $O(nd^2)$, and involves a more general result about DPPs defined via a correlation kernel (Lemma 9), which is of independent interest.

To prove Theorem 1, in Section 4 we demonstrate a fundamental connection between an $\mathbf{A}$-regularized DPP and Bayesian experimental design with precision matrix $\mathbf{A}$. For simplicity of exposition, let the weight vector p be uniformly equal $(\frac{k}{n}, ..., \frac{k}{n})$. If $S \sim \mathrm{DPP}^p_{\mathrm{reg}}(\mathbf{X}, \mathbf{A})$ and $f_{\mathbf{A}}$ is any one of the A/C/D/V-optimality criteria, then:

$$\text{(a)} \ \ \mathbb{E}\big[f_{\mathbf{A}}(\mathbf{X}_S^\top \mathbf{X}_S)\big] \leq f_{\mathbf{A}}\big(\tfrac{k}{n}\boldsymbol{\Sigma}_{\mathbf{X}}\big) \quad \text{and} \quad \text{(b)} \ \ \mathbb{E}\big[|S|\big] \leq d_{\frac{n}{k}\mathbf{A}} + k. \tag{1}$$

Theorem 1 follows by showing an inequality similar to (1a) when $\mathrm{DPP}^p_{\mathrm{reg}}(\mathbf{X}, \mathbf{A})$ is restricted to subsets of size at most k (proof in Section 4). When $\mathbf{A} = \mathbf{0}$, then $\mathrm{DPP}^p_{\mathrm{reg}}(\mathbf{X}, \mathbf{A})$ bears a lot of similarity to *proportional volume sampling* which is an (unregularized) determinantal distribution proposed by [NST18]. This work used an inequality similar to (1a) for obtaining $(1+\epsilon)$-approximate algorithms in A/D-optimal classical experimental design (Question 2 with $\mathbf{A} = \mathbf{0}$). However, the algorithm of [NST18] for proportional volume sampling takes $O(n^4 dk^2 \log k)$ time, making it practically infeasible. On the other hand, the time complexity of sampling from $\mathrm{DPP}^p_{\mathrm{reg}}(\mathbf{X}, \mathbf{A})$ is only $O(nd^2)$, and recent advances in RandNLA for DPP sampling [DWH18b, DWH18a, Der18] suggest that $O(nd \log n + \mathrm{poly}(d))$ time is also possible. Extending the ideas of [NST18] to our new regularized DPP distribution, we obtain efficient $(1+\epsilon)$-approximation algorithms for A/C/D/V-optimal Bayesian experimental design.

Theorem 3 *Let $f_{\mathbf{A}}$ be A/C/D/V-optimality and $\mathbf{X}$ be $n \times d$. If $k = \Omega\big(\frac{d_{\mathbf{A}}}{\epsilon} + \frac{\log 1/\epsilon}{\epsilon^2}\big)$ for some $\epsilon \in (0,1)$, then there is a polynomial time algorithm that finds $S \subseteq [n]$ of size k such that*

$$f_{\mathbf{A}}\big(\mathbf{X}_S^\top \mathbf{X}_S\big) \leq (1+\epsilon) \cdot \mathrm{OPT}_k.$$

Remark 4 *The algorithm referred to in Theorem 3 first solves a convex relaxation of the task via a semi-definite program (SDP) to find the weights $p \in [0,1]^n$, then samples from the $\mathrm{DPP}^p_{\mathrm{reg}}(\mathbf{X}, \mathbf{A})$ distribution $O(1/\epsilon)$ times. The expected cost in addition to the SDP is $O(ndk + k^2 d^2)$.*

Note that unlike in Theorem 3 we use the unrescaled effective dimension $d_{\mathbf{A}}$ instead of the rescaled one, $d_{\frac{n}{k}\mathbf{A}}$. The actual effective dimension that applies here (given in the proof in Section 4) depends on the SDP solution. It is always upper bounded by $d_{\mathbf{A}}$, but it may be significantly smaller. This result is a direct extension of [NST18] to Bayesian setting and to C/V-optimality criteria. Moreover, in their case, proportional volume sampling is usually the computational bottleneck (because its time dependence on the dimension can reach $O(d^{11})$), whereas for us the cost of sampling is negligible compared to the SDP. A number of different methods can be used to solve the SDP relaxation (see Section 5). For example, [AZLSW17] suggest using an iterative optimizer called entropic mirror descent, which is known to exhibit fast convergence and can run in $O(nd^2T)$ time, where T is the number of iterations.

	Criteria	Bayesian	$k = \Omega(\cdot)$
[WYS17]	A,V	✗	$\frac{d^2}{\epsilon}$
[AZLSW17]	A,C,D,E,G,V	✓	$\frac{d}{\epsilon^2}$
[NST18]	A,D	✗	$\frac{d}{\epsilon} + \frac{\log 1/\epsilon}{\epsilon^2}$
this paper	A,C,D,V	✓	$\frac{d_{\mathbf{A}}}{\epsilon} + \frac{\log 1/\epsilon}{\epsilon^2}$

Table 1: Comparison of SDP-based $(1+\epsilon)$-approximation algorithms for classical and Bayesian experimental design (X-mark means that only the classical setting applies).

2 Related work

We first discuss the prior works that focus on bounding the experimental design optimality criteria without obtaining $(1+\epsilon)$-approximation algorithms. First non-trivial bounds for the *classical* A-optimality criterion (with $\mathbf{A} = \mathbf{0}$) were shown by [AB13]. Their result implies that for any $k \geq d$, $\mathrm{OPT}_k \leq (1 + \frac{d-1}{k-d+1}) \cdot f_{\mathbf{0}}(\frac{k}{n}\mathbf{\Sigma}_\mathbf{X})$ and they provide polynomial time algorithms for finding such solutions. The result was later extended by [DW17, DW18b, DW18a] to the case where $\mathbf{A} = \lambda\mathbf{I}$, obtaining that for any $k \geq d_{\lambda\mathbf{I}}$, we have $\mathrm{OPT}_k \leq (1 + \frac{d_{\lambda\mathbf{I}}-1}{k-d_{\lambda\mathbf{I}}+1}) \cdot f_{\frac{k}{n}\lambda\mathbf{I}}(\frac{k}{n}\mathbf{\Sigma}_\mathbf{X})$, and also a faster $O(nd^2)$ time algorithm was provided. Their result can be easily extended to cover any psd matrix $\mathbf{A}$ and V/C-optimality (but not D-optimality). The key improvements of our Theorem 1 are that we cover a potentially much wider range of subset sizes, because $d_{\frac{n}{k}\lambda\mathbf{I}} \leq d_{\lambda\mathbf{I}}$, and our bound can be much tighter because $f_{\lambda\mathbf{I}}(\frac{k}{n}\mathbf{\Sigma}_\mathbf{X}) \leq f_{\frac{k}{n}\lambda\mathbf{I}}(\frac{k}{n}\mathbf{\Sigma}_\mathbf{X})$. Finally, [DCMW19] propose a new notion of *minimax* experimental design, which is related to A/V-optimality. They also use a determinantal distribution for subset selection, however, due to different assumptions, their bounds are incomparable.

A number of works proposed $(1+\epsilon)$-approximation algorithms for experimental design which start with solving a convex relaxation of the problem, and then use some rounding strategy to obtain a discrete solution (see Table 1 for comparison). For example, [WYS17] gave an approximation algorithm for classical A/V-optimality with $k = \Omega(\frac{d^2}{\epsilon})$, where the rounding is done in a greedy fashion, and some randomized rounding strategies are also discussed. [NST18] suggested *proportional volume sampling* for the rounding step and obtained approximation algorithms for classical A/D-optimality with $k = \Omega(\frac{d}{\epsilon} + \frac{\log 1/\epsilon}{\epsilon^2})$. Their approach is particularly similar to ours (when $\mathbf{A} = \mathbf{0}$). However, as discussed earlier, while their algorithms are polynomial, they are virtually intractable. [AZLSW17] proposed an efficient algorithm with a $(1+\epsilon)$-approximation guarantee for a wide range of optimality criteria, including A/C/D/E/V/G-optimality, both classical and Bayesian, when $k = \Omega(\frac{d}{\epsilon^2})$. Our results improve on this work in two ways: (1) in terms of the dependence on ϵ for A/C/D/V-optimality, and (2) in terms of the dependence on the dimension (by replacing d with $d_\mathbf{A}$) in the Bayesian setting. A lower bound shown by [NST18] implies that our Theorem 3 cannot be directly extended to E-optimality, but a similar lower bound does not exist for G-optimality. We remark that the approximation approaches relying on a convex relaxation can generally be converted to an upper bound on OPT_k akin to our Theorem 1, however none of them apply to the regime of $k \leq d$, which is of primary interest in the Bayesian setting.

Purely greedy approximation algorithms have been shown to provide guarantees in a number of special cases for experimental design. One example is classical D-optimality criterion, which can be converted to a submodular function [BGS10]. Also, greedy algorithms for Bayesian A/V-optimality criteria have been considered [BBKT17, CR17b]. These methods can only provide a constant factor approximation guarantee (as opposed to $1+\epsilon$), and the factor is generally problem dependent (which means it could be arbitrarily large). Finally, a number of heuristics with good empirical performance have been proposed, such as Fedorov's exchange method [CN80]. However, in this work we focus on methods that provide theoretical approximation guarantees.

3 A new regularized determinantal point process

In this section we introduce the determinantal sampling distribution we use for obtaining guarantees in Bayesian experimental design. Determinantal point processes (DPP) form a family of distributions which are used to model repulsion between elements in a random set, with many applications in machine learning [KT12]. Here, we focus on the setting where we are sampling out of all 2^n subsets $S \subseteq [n]$. Traditionally, a DPP is defined by a correlation kernel, which is an $n \times n$ psd matrix $\mathbf{K}$ with eigenvalues between 0 and 1, i.e., such that $\mathbf{0} \preceq \mathbf{K} \preceq \mathbf{I}$. Given a correlation kernel $\mathbf{K}$, the corresponding DPP is defined as

$$S \sim \mathrm{DPP}_{\mathrm{cor}}(\mathbf{K}) \qquad \text{iff} \qquad \Pr(T \subseteq S) = \det(\mathbf{K}_{T,T}) \;\; \forall_{T \in [n]},$$

where $\mathbf{K}_{T,T}$ is the submatrix of $\mathbf{K}$ with rows and columns indexed by T. Another way of defining a DPP, popular in the machine learning community, is via an ensemble kernel $\mathbf{L}$. Any psd matrix $\mathbf{L}$ is an ensemble kernel of a DPP defined as:

$$S \sim \mathrm{DPP}_{\mathrm{ens}}(\mathbf{L}) \qquad \text{iff} \qquad \Pr(S) \propto \det(\mathbf{L}_{S,S}).$$

Crucially, every $\mathrm{DPP}_{\mathrm{ens}}$ is also a $\mathrm{DPP}_{\mathrm{cor}}$, but not the other way around. Specifically, we have:

$$\text{(a) } \mathrm{DPP}_{\mathrm{cor}}(\mathbf{K}) = \mathrm{DPP}_{\mathrm{ens}}\big(\mathbf{I} - (\mathbf{I}+\mathbf{L})^{-1}\big), \quad \text{and} \quad \text{(b) } \mathrm{DPP}_{\mathrm{ens}}(\mathbf{L}) = \mathrm{DPP}_{\mathrm{cor}}\big(\mathbf{K}(\mathbf{I}-\mathbf{K})^{-1}\big),$$

but (b) requires that $\mathbf{I}-\mathbf{K}$ be invertible which is not true for some DPPs. (This will be important in our analysis.) The classical algorithm for sampling from a DPP requires the eigendecomposition of either matrix $\mathbf{K}$ or $\mathbf{L}$, which in general costs $O(n^3)$, followed by a sampling procedure which costs $O(n\,|S|^2)$ [HKP$^+$06, KT12].

We next define our regularized DPP and describe its connection with correlation and ensemble DPPs.

Definition 2 *Given matrix $\mathbf{X} \in \mathbb{R}^{n\times d}$, a sequence $p = (p_1, \dots, p_n) \in [0,1]^n$ and a psd matrix $\mathbf{A} \in \mathbb{R}^{d\times d}$ such that $\sum_i p_i \mathbf{x}_i \mathbf{x}_i^\top + \mathbf{A}$ is full rank, let $\mathrm{DPP}^p_{\mathrm{reg}}(\mathbf{X}, \mathbf{A})$ be a distribution over $S \subseteq [n]$:*

$$\Pr(S) = \frac{\det(\mathbf{X}_S^\top \mathbf{X}_S + \mathbf{A})}{\det\big(\sum_i p_i \mathbf{x}_i \mathbf{x}_i^\top + \mathbf{A}\big)} \cdot \prod_{i\in S} p_i \cdot \prod_{i\notin S}(1-p_i). \tag{2}$$

The fact that this is a proper distribution (i.e., that it sums to one) can be restated as a determinantal expectation formula: if $b_i \sim \mathrm{Bernoulli}(p_i)$ are independent Bernoulli random variables, then

$$\sum_{S\subseteq[n]} \det(\mathbf{X}_S^\top \mathbf{X}_S + \mathbf{A}) \prod_{i\in S} p_i \prod_{i\notin S}(1-p_i) = \mathbb{E}\bigg[\det\Big(\sum_i b_i \mathbf{x}_i \mathbf{x}_i^\top + \mathbf{A}\Big)\bigg] \overset{(*)}{=} \det\Big(\sum_i p_i \mathbf{x}_i \mathbf{x}_i^\top + \mathbf{A}\Big),$$

where $(*)$ was recently shown by [DM19, Lemma 7]. Our main result in this section is the following efficient algorithm for $\mathrm{DPP}^p_{\mathrm{reg}}(\mathbf{X}, \mathbf{A})$ which reduces it to sampling from a correlation DPP.

Theorem 5 *For any $\mathbf{X} \in \mathbb{R}^{n\times d}$, $p \in [0,1]^n$ and a psd matrix $\mathbf{A}$ s.t. $\sum_i p_i \mathbf{x}_i \mathbf{x}_i^\top + \mathbf{A}$ is full rank, let*

$$T \sim \mathrm{DPP}_{\mathrm{cor}}\big(\mathbf{D}_p^{1/2} \mathbf{X}(\mathbf{A} + \mathbf{X}^\top \mathbf{D}_p \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{D}_p^{1/2}\big), \quad \textit{where} \quad \mathbf{D}_p = \mathrm{diag}(p).$$

If $b_i \sim \mathrm{Bernoulli}(p_i)$ are independent random variables, then $T \cup \{i : b_i = 1\} \sim \mathrm{DPP}^p_{\mathrm{reg}}(\mathbf{X}, \mathbf{A})$.

Remark 6 *Since the correlation kernel matrix has rank at most d, the preprocessing cost of eigendecomposition is $O(nd^2)$. Then, each sample costs only $O(n\,|T|^2)$.*

We prove the theorem in three steps. First, we express $\mathrm{DPP}^p_{\mathrm{reg}}(\mathbf{X}, \mathbf{A})$ as an ensemble DPP, which requires some additional assumptions on $\mathbf{A}$ and p to be possible. Then, we convert the ensemble to a correlation kernel (eliminating the extra assumptions), and finally show that this kernel can be decomposed into a rank d kernel plus Bernoulli sampling.

Sampling $S \sim \mathrm{DPP}^p_{\mathrm{reg}}(\mathbf{X}, \mathbf{A})$

Input: $\mathbf{X} \in \mathbb{R}^{n\times d}$, psd $\mathbf{A} \in \mathbb{R}^{d\times d}$, $p \in [0,1]^n$
Compute $\mathbf{Z} \leftarrow \mathbf{A} + \mathbf{X}^\top \mathbf{D}_p \mathbf{X}$
Compute SVD of $\mathbf{B} = \mathbf{D}_p^{1/2} \mathbf{X} \mathbf{Z}^{-1/2}$
Sample $T \sim \mathrm{DPP}_{\mathrm{cor}}(\mathbf{B}\mathbf{B}^\top)$ using the SVD
Sample $b_i \sim \mathrm{Bernoulli}(p_i)$ for $i \in [n]$
return $S = T \cup \{i : b_i = 1\}$

Lemma 7 *Given $\mathbf{X}$, $\mathbf{A}$ and $\mathbf{D}_p$ as in Theorem 5, if we assume that $\mathbf{A}$ and $\mathbf{I} - \mathbf{D}_p$ are invertible, then*

$$\mathrm{DPP}^p_{\mathrm{reg}}(\mathbf{X}, \mathbf{A}) = \mathrm{DPP}_{\mathrm{ens}}\big(\widetilde{\mathbf{D}} + \widetilde{\mathbf{D}}^{1/2} \mathbf{X} \mathbf{A}^{-1} \mathbf{X}^\top \widetilde{\mathbf{D}}^{1/2}\big), \quad \textit{where} \quad \widetilde{\mathbf{D}} = \mathbf{D}_p(\mathbf{I} - \mathbf{D}_p)^{-1}.$$

Proof Let $S \sim \mathrm{DPP}^p_{\mathrm{reg}}(\mathbf{X}, \mathbf{A})$. By Definition 2 and the fact that $\det(\mathbf{A}\mathbf{B} + \mathbf{I}) = \det(\mathbf{B}\mathbf{A} + \mathbf{I})$,

$$\begin{aligned}
\Pr(S) &\propto \det(\mathbf{X}_S^\top \mathbf{X}_S + \mathbf{A}) \cdot \prod_{i\in S} p_i \cdot \prod_{i\notin S}(1-p_i) = \det(\mathbf{X}_S^\top \mathbf{X}_S + \mathbf{A}) \cdot \prod_{i\in S} \frac{p_i}{1-p_i} \cdot \prod_{i=1}^n (1-p_i) \\
&\propto \det\big(\mathbf{A}(\mathbf{A}^{-1}\mathbf{X}_S^\top \mathbf{X}_S + \mathbf{I})\big) \det(\widetilde{\mathbf{D}}_{S,S}) = \det(\mathbf{A}) \det(\mathbf{A}^{-1}\mathbf{X}_S^\top \mathbf{X}_S + \mathbf{I}) \det(\widetilde{\mathbf{D}}_{S,S}) \\
&\propto \det(\mathbf{X}_S \mathbf{A}^{-1} \mathbf{X}_S^\top + \mathbf{I}) \det(\widetilde{\mathbf{D}}_{S,S}) = \det\Big(\big[\widetilde{\mathbf{D}}^{1/2} \mathbf{X} \mathbf{A}^{-1} \mathbf{X}^\top \widetilde{\mathbf{D}}^{1/2} + \widetilde{\mathbf{D}}\big]_{S,S}\Big),
\end{aligned}$$

which matches the definition of the L-ensemble DPP. ∎

At this point, to sample from $\mathrm{DPP}^p_{\mathrm{reg}}(\mathbf{X}, \mathbf{A})$, we could simply invoke any algorithm for sampling from an ensemble DPP. However, this would only work for invertible $\mathbf{A}$, which in particular excludes the important case of $\mathbf{A} = \mathbf{0}$ corresponding to classical experimental design. Moreover, the standard algorithm would require computing the eigendecomposition of the ensemble kernel, which (at least if done naïvely) costs $O(n^3)$. Even after this is done, the sampling cost would still be $O(n\,|S|^2)$ which can be considerably more than $O(nd^2)$. We first address the issue of invertibility of matrix $\mathbf{A}$ by expressing our distribution via a correlation DPP.

Lemma 8 *Given* $\mathbf{X}$, $\mathbf{A}$, *and* $\mathbf{D}_p$ *as in Theorem 5 (without any additional assumptions), we have*

$$\mathrm{DPP}^p_{\mathrm{reg}}(\mathbf{X},\mathbf{A}) = \mathrm{DPP}_{\mathrm{cor}}\big(\mathbf{D}_p + (\mathbf{I}-\mathbf{D}_p)^{1/2}\,\mathbf{D}_p^{1/2}\mathbf{X}(\mathbf{A}+\mathbf{X}^\top\mathbf{D}_p\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{D}_p^{1/2}(\mathbf{I}-\mathbf{D}_p)^{1/2}\big).$$

When $\mathbf{A}$ and $\mathbf{I}-\mathbf{D}_p$ are invertible, then the proof (given in Appendix A) is a straightforward calculation. Then, we use a limit argument with $p_\epsilon = (1-\epsilon)p$ and $\mathbf{A}_\epsilon = \mathbf{A}+\epsilon\mathbf{I}$, where $\epsilon \to 0$.

Finally, we show that the correlation DPP arrived at in Lemma 8 can be decomposed into a smaller DPP plus Bernoulli sampling. In fact, in the following lemma we obtain a more general recipe for combining DPPs with Bernoulli sampling, which may be of independent interest. Note that if $b_i \sim \mathrm{Bernoulli}(p_i)$ are independent random variables then $\{i : b_i = 1\} \sim \mathrm{DPP}_{\mathrm{cor}}(\mathbf{D}_p)$.

Lemma 9 *Let* $\mathbf{K}$ *and* $\mathbf{D}$ *be* $n \times n$ *psd matrices with eigenvalues between 0 and 1, and assume that* $\mathbf{D}$ *is diagonal. If* $T \sim \mathrm{DPP}_{\mathrm{cor}}(\mathbf{K})$ *and* $R \sim \mathrm{DPP}_{\mathrm{cor}}(\mathbf{D})$*, then*

$$T \cup R \sim \mathrm{DPP}_{\mathrm{cor}}\big(\mathbf{D} + (\mathbf{I}-\mathbf{D})^{1/2}\mathbf{K}(\mathbf{I}-\mathbf{D})^{1/2}\big).$$

The lemma is proven in Appendix A. Theorem 5 now follows by combining Lemmas 8 and 9.

4 Guarantees for Bayesian experimental design

In this section we prove our main results regarding Bayesian experimental design (Theorems 1 and 3). First, we establish certain properties of the regularized DPP distribution that make it effective in this setting. Even though the size of the sampled subset $S \sim \mathrm{DPP}^p_{\mathrm{reg}}(\mathbf{X},\mathbf{A})$ is random and can be as large as n, it is also highly concentrated around its expectation, which can be bounded in terms of the $\mathbf{A}$-effective dimension. This is crucial, since both of our main results require a subset of deterministically bounded size. Recall that the effective dimension is defined as a function $d_{\mathbf{A}}(\mathbf{\Sigma}) = \mathrm{tr}\big(\mathbf{\Sigma}(\mathbf{A}+\mathbf{\Sigma})^{-1}\big)$. The omitted proofs are in Appendix A.

Lemma 10 *Given any* $\mathbf{X} \in \mathbb{R}^{n\times d}$, $p \in [0,1]^n$ *and a psd matrix* $\mathbf{A}$ *s.t.* $\sum_i p_i\mathbf{x}_i\mathbf{x}_i^\top + \mathbf{A}$ *is full rank, let* $S = T \cup \{i : b_i = 1\} \sim \mathrm{DPP}^p_{\mathrm{reg}}(\mathbf{X},\mathbf{A})$ *be defined as in Theorem 5. Then*

$$\mathbb{E}\big[|S|\big] \le \mathbb{E}\big[|T|\big] + \mathbb{E}\Big[\sum_i b_i\Big] = d_{\mathbf{A}}\Big(\sum_i p_i\mathbf{x}_i\mathbf{x}_i^\top\Big) + \sum_i p_i.$$

Next, we show two expectation inequalities for the matrix inverse and matrix determinant, which hold for the regularized DPP. We use them to bound the Bayesian optimality criteria in expectation.

Lemma 11 *Whenever* $S \sim \mathrm{DPP}^p_{\mathrm{reg}}(\mathbf{X},\mathbf{A})$ *is a well-defined distribution it holds that*

$$\mathbb{E}\Big[\big(\mathbf{X}_S^\top\mathbf{X}_S + \mathbf{A}\big)^{-1}\Big] \preceq \Big(\sum_i p_i\mathbf{x}_i\mathbf{x}_i^\top + \mathbf{A}\Big)^{-1}, \tag{3}$$

$$\mathbb{E}\Big[\det\big(\mathbf{X}_S^\top\mathbf{X}_S + \mathbf{A}\big)^{-1}\Big] \le \det\Big(\sum_i p_i\mathbf{x}_i\mathbf{x}_i^\top + \mathbf{A}\Big)^{-1}. \tag{4}$$

Corollary 12 *Let* $f_{\mathbf{A}}$ *be A/C/D/V-optimality. Whenever* $S \sim \mathrm{DPP}^p_{\mathrm{reg}}(\mathbf{X},\mathbf{A})$ *is well-defined,*

$$\mathbb{E}\big[f_{\mathbf{A}}(\mathbf{X}_S^\top\mathbf{X}_S)\big] \le f_{\mathbf{A}}\Big(\sum_i p_i\mathbf{x}_i\mathbf{x}_i^\top\Big).$$

Proof In the case of A-, C-, and V-optimality, the function $f_{\mathbf{A}}$ is a linear transformation of the matrix $(\mathbf{X}_S^\top\mathbf{X}_S + \mathbf{A})^{-1}$ so the bound follows from (3). For D-optimality, we apply (4) as follows:

$$\begin{aligned}\mathbb{E}\big[f_{\mathbf{A}}(\mathbf{X}_S^\top\mathbf{X}_S)\big] &= \mathbb{E}\Big[\det\big(\mathbf{X}_S^\top\mathbf{X}_S+\mathbf{A}\big)^{-1/d}\Big] \le \mathbb{E}\Big[\Big(\det\big(\mathbf{X}_S^\top\mathbf{X}_S+\mathbf{A}\big)^{-1/d}\Big)^d\Big]^{1/d}\\ &= \mathbb{E}\Big[\det\big(\mathbf{X}_S^\top\mathbf{X}_S+\mathbf{A}\big)^{-1}\Big]^{1/d} = \det\Big(\sum_i p_i\mathbf{x}_i\mathbf{x}_i^\top\Big)^{-1/d},\end{aligned}$$

which completes the proof. ∎

Finally, we present the key lemma that puts everything together. This result is essentially a generalization of Theorem 1 from which also follows Theorem 3.

Lemma 13 *Let $f_{\mathbf{A}}$ be A/C/D/V-optimality and $\mathbf{X}$ be $n \times d$. For some $w = (w_1, \ldots, w_n) \in [0,1]$, let $\mathbf{\Sigma}_w = \sum_i w_i \mathbf{x}_i \mathbf{x}_i^\top$ and assume that $\sum_i w_i = k \in [n]$. If $k \geq 4\, d_{\mathbf{A}}(\mathbf{\Sigma}_w)$, then a subset $S \subseteq [n]$ of size k can be found in $O(ndk + k^2 d^2)$ time that satisfies*

$$f_{\mathbf{A}}\big(\mathbf{X}_S^\top \mathbf{X}_S\big) \leq \left(1 + 8\,\frac{d_{\mathbf{A}}(\mathbf{\Sigma}_w)}{k} + 8\sqrt{\frac{\ln(k/d_{\mathbf{A}}(\mathbf{\Sigma}_w))}{k}}\right) \cdot f_{\mathbf{A}}\big(\mathbf{\Sigma}_w\big).$$

Proof Let $p = (p_1, \ldots, p_n)$ be defined so that $p_i = \frac{w_i}{1+\epsilon}$, and suppose that $S \sim \mathrm{DPP}^p_{\mathrm{reg}}(\mathbf{X}, \mathbf{A})$. Then, using Theorem 11, we have

$$\begin{aligned}\Pr\big(|S| \leq k\big)\, \mathbb{E}\Big[f_{\mathbf{A}}(\mathbf{X}_S^\top \mathbf{X}_S) \mid |S| \leq k\Big] &\leq \mathbb{E}\big[f_{\mathbf{A}}(\mathbf{X}_S^\top \mathbf{X}_S)\big] \leq f_{\mathbf{A}}\Big(\sum_i p_i \mathbf{x}_i \mathbf{x}_i^\top\Big) \\ &\leq (1+\epsilon) \cdot f_{\mathbf{A}}\Big(\sum_i w_i \mathbf{x}_i \mathbf{x}_i^\top\Big).\end{aligned}$$

Using Lemma 10 we can bound the expected size of S as follows:

$$\mathbb{E}\big[|S|\big] \leq d_{\mathbf{A}}(\mathbf{\Sigma}_w) + \sum_i p_i = d_{\mathbf{A}}(\mathbf{\Sigma}_w) + \frac{k}{1+\epsilon} = k \cdot \Big(1 + \frac{d_{\mathbf{A}}(\mathbf{\Sigma}_w)}{k} - \frac{\epsilon}{1+\epsilon}\Big).$$

Let $d_w = d_{\mathbf{A}}(\mathbf{\Sigma}_w)$ and $\alpha = 1 + \frac{d_w}{k} - \frac{\epsilon}{1+\epsilon}$. If $1 \geq \epsilon \geq \frac{4d_w}{k}$, then $\alpha \leq 1 + \frac{\epsilon}{4} - \frac{\epsilon}{2} = 1 - \frac{\epsilon}{4}$. Since $\mathrm{DPP}^p_{\mathrm{reg}}(\mathbf{X}, \mathbf{A})$ is a determinantal point process, $|S|$ is a Poisson binomial r.v. so for $\epsilon \geq 6\sqrt{\frac{\ln(k/d_w)}{k}}$,

$$\Pr(|S| > k) \leq \mathrm{e}^{-\frac{(k-\alpha k)^2}{2k}} = \mathrm{e}^{-\frac{k}{2}(1-\alpha)^2} \leq \mathrm{e}^{-\frac{k\epsilon^2}{32}} \leq \frac{d_w}{k}.$$

For any $\epsilon \geq 4\,\frac{d_w}{k} + 6\sqrt{\frac{\ln(k/d_w)}{k}}$, we have

$$\begin{aligned}\mathbb{E}\big[f_{\mathbf{A}}(\mathbf{X}_S^\top \mathbf{X}_S) \mid |S| \leq k\big] &\leq \frac{1+\epsilon}{1-\frac{d_w}{k}} \cdot f_{\mathbf{A}}(\mathbf{\Sigma}_w) \leq \left(1 + \frac{\epsilon + \frac{d_w}{k}}{1 - \frac{d_w}{k}}\right) \cdot f_{\mathbf{A}}(\mathbf{\Sigma}_w) \\ &\leq \left(1 + 7\,\frac{d_w}{k} + 8\sqrt{\frac{\ln(k/d_w)}{k}}\right) \cdot f_{\mathbf{A}}(\mathbf{\Sigma}_w).\end{aligned}$$

Denoting $\mathbb{E}\big[f_{\mathbf{A}}(\mathbf{X}_S^\top \mathbf{X}_S) \mid |S| \leq k\big]$ as F_k, Markov's inequality implies that $\Pr(f_{\mathbf{A}}(\mathbf{X}_S^\top \mathbf{X}_S) \geq (1+\delta)F_k \mid |S| \leq k) \leq \frac{1}{1+\delta}$. Also, we showed that $\Pr(|S| \leq k) \geq 1 - \frac{d_w}{k} \geq \frac{3}{4}$. Setting $\delta = \frac{d_w}{Ck}$ for sufficiently large C we obtain that with probability $\Omega(\frac{d_w}{k})$, the random set S has size at most k and

$$\begin{aligned}f_{\mathbf{A}}(\mathbf{X}_S^\top \mathbf{X}_S) &\leq \left(1 + \frac{d_w}{Ck}\right) \cdot \left(1 + 7\,\frac{d_w}{k} + 8\sqrt{\frac{\ln(k/d_w)}{k}}\right) \cdot f_{\mathbf{A}}(\mathbf{\Sigma}_w) \\ &\leq \left(1 + 8\,\frac{d_w}{k} + 8\sqrt{\frac{\ln(k/d_w)}{k}}\right) \cdot f_{\mathbf{A}}(\mathbf{\Sigma}_w).\end{aligned}$$

We can sample from $\mathrm{DPP}^p_{\mathrm{reg}}(\mathbf{X}, \mathbf{A})$ conditioned on $|S| \leq k$ and $f_{\mathbf{A}}(\mathbf{X}_S^\top \mathbf{X}_S)$ bounded as above by rejection sampling. When $|S| < k$, the set is completed to k with arbitrary indices. On average, $O(\frac{k}{d_w})$ samples from $\mathrm{DPP}^p_{\mathrm{reg}}(\mathbf{X}, \mathbf{A})$ are needed, so the cost is $O(nd^2)$ for the eigendecomposition, $O(\frac{k}{d_w} \cdot nd_w^2) = O(nd_w k)$ for sampling and $O(\frac{k}{d_w} \cdot kd^2)$ for recomputing $f_{\mathbf{A}}(\mathbf{X}_S^\top \mathbf{X}_S)$. ■

To prove the main results, we use Lemma 13 with appropriately chosen weights w.

Proof of Theorem 1 Let $w = (\frac{k}{n}, ..., \frac{k}{n})$ in Lemma 13. Then, we have $\mathbf{\Sigma}_w = \frac{k}{n}\mathbf{\Sigma}_{\mathbf{X}}$ and also $d_{\mathbf{A}}(\mathbf{\Sigma}_w) = d_{\frac{n}{k}\mathbf{A}}$. Since for any set S of size k, we have $\mathrm{OPT}_k \leq f_{\mathbf{A}}(\mathbf{X}_S^\top \mathbf{X}_S)$, the result follows. ■

Proof of Theorem 3 As discussed in [AZLSW17, BV04], the following convex relaxation of experimental design can be written as a semi-definite program and solved using standard SDP solvers:

$$w^* \;=\; \operatorname*{argmin}_{w} \;\; f_{\mathbf{A}}\Big(\sum_{i=1}^n w_i \mathbf{x}_i \mathbf{x}_i^\top\Big), \quad \text{subject to} \quad \forall_i \;\; 0 \leq w_i \leq 1, \quad \sum_i w_i = k. \tag{5}$$

The solution w^* satisfies $f_{\mathbf{A}}\big(\mathbf{\Sigma}_{w^*}\big) \leq \mathrm{OPT}_k$. If we use w^* in Lemma 13, then observing that $d_{\mathbf{A}}(\mathbf{\Sigma}_{w^*}) \leq d_{\mathbf{A}}$, and setting $k \geq C(\frac{d_{\mathbf{A}}}{\epsilon} + \frac{\log 1/\epsilon}{\epsilon^2})$ for sufficiently large C, the algorithm in the lemma finds subset S such that $f_{\mathbf{A}}(\mathbf{X}_S^\top \mathbf{X}_S) \leq (1+\epsilon) \cdot f_{\mathbf{A}}(\mathbf{\Sigma}_w) \leq (1+\epsilon) \cdot \mathrm{OPT}_k$. Note that we did not need to solve the SDP exactly, so approximate solvers could be used instead. ■

5 Experiments

We confirm our theoretical results with experiments on real world data from `libsvm` datasets [CL11] (more details in Appendix C). For all our experiments, the prior precision matrix is set to $\mathbf{A} = n^{-1}\mathbf{I}$ and we consider sample sizes $k \in [d, 5d]$. Each experiment is averaged over 25 trials and bootstrap 95% confidence intervals are shown. The quality of our method, as measured by the A-optimality criterion $f_{\mathbf{A}}(\mathbf{X}_S^\top \mathbf{X}_S) = \operatorname{tr}\left((\mathbf{X}_S^\top \mathbf{X}_S + \mathbf{A})^{-1}\right)$ is compared against the following references and recently proposed methods for A-optimal design:

Greedy bottom-up adds an index $i \in [n]$ to the sample S maximizing the increase in A-optimality criterion [BBKT17, CR17a].

Our method (with SDP) uses the efficient algorithms developed in proving Theorem 3 to sample $\mathrm{DPP}^p_{\mathrm{reg}}(\mathbf{X}, \mathbf{A})$ constrained to subset size k with $p = w^*$, see (5), obtained using a recently developed first order convex cone solver called Splitting Conical Solver (SCS) [OCPB16]. We chose SCS because it can handle the SDP constraints in (5) and has provable termination guarantees, while also finding solutions faster [OCPB16] than alternative off-the-shelf optimization software libraries such as SDPT3 and Sedumi.

Our method (without SDP) samples $\mathrm{DPP}^p_{\mathrm{reg}}(\mathbf{X}, \mathbf{A})$ with uniform probabilities $p \equiv \frac{k}{n}$.

Uniform samples every size k subset $S \subseteq [n]$ with equal probability.

Predictive length sampling [ZMMY15] samples each row $\mathbf{x}_i$ of $\mathbf{X}$ with probability $\propto \|\mathbf{x}_i\|$.

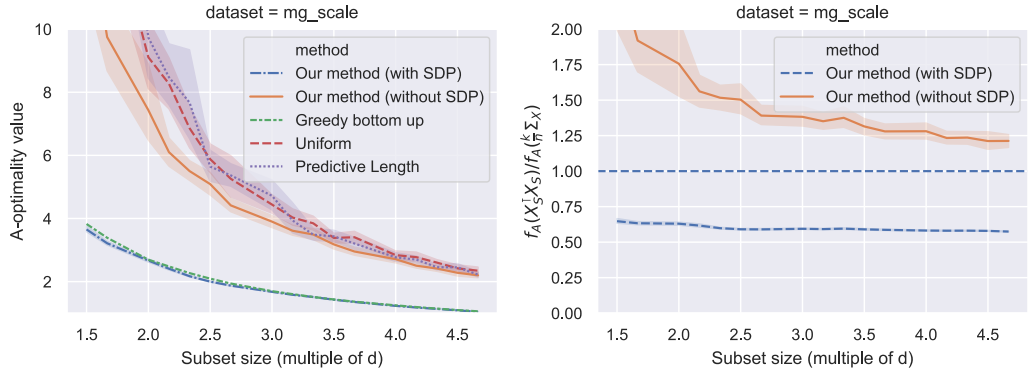

Figure 1: (left) A-optimality value obtained by the various methods on the `mg_scale` dataset [CL11] with prior precision $\mathbf{A} = 10^{-5}\,\mathbf{I}$, (right) A-optimality value for our method (with and without SDP) divided by $f_{\mathbf{A}}(\frac{k}{n}\mathbf{\Sigma}_{\mathbf{X}})$, the estimate suggested by Theorem 1.

Figure 1 (left) reveals that our method (without SDP) is superior to both uniform and predictive length sampling, producing designs which achieve lower A-optimality criteria values for all sample sizes. As Theorem 5 shows that our method (without SDP) only differs from uniform sampling by an additional DPP sample with controlled expected size (see Lemma 10), we may conclude that adding even a small DPP sample can improve a uniformly sampled design.

Consistent to prior observations [CR17a, WYS17], the greedy bottom up method achieves surprisingly good performance. However, if our method is used in conjunction with an SDP solution, then we are able to match and even slightly exceed the performance of the greedy bottom up method. Furthermore, the overall run-time costs (see Appendix C) between the two are comparable. As the majority of the runtime of our method (with SDP) is occupied by solving the SDP, an interesting future direction is to investigate alternative solvers such as interior point methods as well as terminating the solvers early once an approximate solution is reached.

Figure 1 (right) displays the ratio $f_{\mathbf{A}}(\mathbf{X}_S^\top \mathbf{X}_S) \,/\, f_{\mathbf{A}}(\frac{k}{n}\mathbf{\Sigma}_{\mathbf{X}})$ for subsets returned by our method (with and without SDP). Note that the line for our method with SDP on Figure 1 (right) shows that the ratio never goes below 0.5, and we saw similar behavior across all examined datasets (see Appendix C). This evidence suggests that for many real datasets $\mathrm{OPT}_k = \Theta(1) \cdot f_{\mathbf{A}}(\frac{k}{n}\mathbf{\Sigma}_{\mathbf{X}})$, matching the upper bound of Theorem 1.

References

[AB13] Haim Avron and Christos Boutsidis. Faster subset selection for matrices and applications. *SIAM Journal on Matrix Analysis and Applications*, 34(4):1464–1499, 2013.

[AM15] 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.

[AZLSW17] 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.

[BBKT17] Andrew An Bian, Joachim M. Buhmann, Andreas Krause, and Sebastian Tschiatschek. Guarantees for greedy maximization of non-submodular functions with applications. 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 498–507, International Convention Centre, Sydney, Australia, 06–11 Aug 2017. PMLR.

[Ber11] Dennis S. Bernstein. *Matrix Mathematics: Theory, Facts, and Formulas*. Princeton University Press, second edition, 2011.

[BGS10] Mustapha Bouhtou, Stéphane Gaubert, and Guillaume Sagnol. Submodularity and randomized rounding techniques for optimal experimental design. *Electronic Notes in Discrete Mathematics*, 36:679–686, 08 2010.

[BV04] Stephen Boyd and Lieven Vandenberghe. *Convex optimization*. Cambridge university press, 2004.

[CL11] Chih-Chung Chang and Chih-Jen Lin. LIBSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*, 2:27:1–27:27, 2011.

[CN80] R Dennis Cook and Christopher J Nachtrheim. A comparison of algorithms for constructing exact d-optimal designs. *Technometrics*, 22(3):315–324, 1980.

[CR17a] Luiz Chamon and Alejandro Ribeiro. Approximate supermodularity bounds for experimental design. In *Advances in Neural Information Processing Systems*, pages 5403–5412, 2017.

[CR17b] Luiz F. O. Chamon and Alejandro Ribeiro. Greedy sampling of graph signals. *CoRR*, abs/1704.01223, 2017.

[CV95] Kathryn Chaloner and Isabella Verdinelli. Bayesian experimental design: A review. *Statist. Sci.*, 10(3):273–304, 08 1995.

[DCMW19] Michał Dereziński, Kenneth L. Clarkson, Michael W. Mahoney, and Manfred K. Warmuth. Minimax experimental design: Bridging the gap between statistical and worst-case approaches to least squares regression. *arXiv e-prints*, page arXiv:1902.00995, Feb 2019.

[Der18] Michał Dereziński. Fast determinantal point processes via distortion-free intermediate sampling. *CoRR*, abs/1811.03717, 2018.

[DM16] Petros Drineas and Michael W. Mahoney. RandNLA: Randomized numerical linear algebra. *Communications of the ACM*, 59:80–90, 2016.

[DM17] Petros Drineas and Michael W. Mahoney. Lectures on randomized numerical linear algebra. Technical report, 2017. Preprint: arXiv:1712.08880; To appear in: *Lectures of the 2016 PCMI Summer School on Mathematics of Data*.

[DM19] Michał Dereziński and Michael W. Mahoney. Distributed estimation of the inverse Hessian by determinantal averaging. *arXiv e-prints*, May 2019.

[DW17] 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.

[DW18a] Michał Dereziński and Manfred K. Warmuth. Reverse iterative volume sampling for linear regression. *Journal of Machine Learning Research*, 19(23):1–39, 2018.

[DW18b] Michał Dereziński and Manfred K. Warmuth. Subsampling for ridge regression via regularized volume sampling. In Amos Storkey and Fernando Perez-Cruz, editors, *Proceedings of the Twenty-First International Conference on Artificial Intelligence and Statistics*, pages 716–725, Playa Blanca, Lanzarote, Canary Islands, April 2018.

[DWH18a] Michał Dereziński, Manfred K. Warmuth, and Daniel Hsu. Correcting the bias in least squares regression with volume-rescaled sampling. *CoRR*, abs/1810.02453, 2018.

[DWH18b] Michał Dereziński, Manfred K. Warmuth, and Daniel Hsu. Leveraged volume sampling for linear regression. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems 31*, pages 2510–2519. Curran Associates, Inc., 2018.

[HKP$^+$06] J. Ben Hough, Manjunath Krishnapur, Yuval Peres, Bálint Virág, et al. Determinantal processes and independence. *Probability surveys*, 3:206–229, 2006.

[KT12] Alex Kulesza and Ben Taskar. *Determinantal Point Processes for Machine Learning*. Now Publishers Inc., Hanover, MA, USA, 2012.

[Mah11] Michael W. Mahoney. Randomized algorithms for matrices and data. *Foundations and Trends in Machine Learning*, 3(2):123–224, 2011. Also available at: arXiv:1104.5557.

[NST18] Aleksandar Nikolov, Mohit Singh, and Uthaipon T. Tantipongpipat. Proportional volume sampling and approximation algorithms for A-optimal design. *CoRR*, abs/1802.08318, 2018.

[OCPB16] Brendan O'Donoghue, Eric Chu, Neal Parikh, and Stephen Boyd. Conic optimization via operator splitting and homogeneous self-dual embedding. *Journal of Optimization Theory and Applications*, 169(3):1042–1068, 2016.

[Puk06] Friedrich Pukelsheim. *Optimal Design of Experiments (Classics in Applied Mathematics) (Classics in Applied Mathematics, 50)*. Society for Industrial and Applied Mathematics, Philadelphia, PA, USA, 2006.

[WYS17] Yining Wang, Adams W. Yu, and Aarti Singh. On computationally tractable selection of experiments in measurement-constrained regression models. *J. Mach. Learn. Res.*, 18(1):5238–5278, January 2017.

[ZMMY15] Rong Zhu, Ping Ma, Michael W Mahoney, and Bin Yu. Optimal subsampling approaches for large sample linear regression. *arXiv preprint arXiv:1509.05111*, 2015.