Bayesian experimental design using regularized determinantal point processes

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

Feynman Liang
Department of Statistics
University of California, Berkeley
feynman@berkeley.edu

Michael W. Mahoney
ICSI and Department of Statistics
University of California, Berkeley
mmahoney@stat.berkeley.edu

Abstract

We establish a fundamental connection between Bayesian experimental design and determinantal point processes (DPPs). Experimental design is a classical task in combinatorial optimization, where we wish to select a small subset of d-dimensional vectors to minimize a statistical optimality criterion. We show that a new regularized variant of DPPs can be used to design efficient algorithms for finding $(1+\epsilon)$-approximate solutions to experimental design under four commonly used optimality criteria: A-, C-, D- and V-optimality. A key novelty is that we offer improved guarantees under the Bayesian framework. Our algorithm returns a $(1+\epsilon)$-approximate solution when the subset size k is $\Omega(\frac{d_{\mathbf{A}}}{\epsilon}+\frac{\log 1/\epsilon}{\epsilon^2})$, where $d_{\mathbf{A}} \ll d$ is an effective dimension determined by prior knowledge (via a precision matrix $\mathbf{A}$). This is the first approximation guarantee where the dependence on d is replaced by an effective dimension. Moreover, the time complexity of our algorithm significantly improves on existing approaches with comparable guarantees.

1 Introduction

Consider a collection of n experiments parameterized by d-dimensional vectors $\mathbf{x}_1, \dots, \mathbf{x}_n$, 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 $i \in \{1,\dots,n\} = [n]$, 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 $S \subseteq [n]$ of $|S| = k$ experiments. The posterior distribution of $\mathbf{w}$ given $\mathbf{y}_S$ (the vector of outcomes in S) is:

$$\begin{aligned} \mathbf{w} \mid \mathbf{y}_S \;&\sim\; \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma}), \\ \text{where } \boldsymbol{\mu} &= (\mathbf{X}_S^\top\mathbf{X}_S + \mathbf{A})^{-1}\mathbf{X}_S^\top\mathbf{y}_S, \\ \boldsymbol{\Sigma} &= \sigma^2(\mathbf{X}_S^\top\mathbf{X}_S + \mathbf{A})^{-1}. \end{aligned}$$

Here, $\mathbf{X}_S$ is the $k \times d$ matrix with rows $\mathbf{x}_i^\top$ for $i \in S$.

In Bayesian experimental design (Chaloner and Verdinelli, 1995), the prior precision matrix $\mathbf{A}$ is used to encode prior knowledge and our goal is to choose S so as to minimize a function (a.k.a. an optimality criterion) measuring the "size" of the posterior covariance matrix $\boldsymbol{\Sigma}_{\mathbf{w}|\mathbf{y}_S} = \sigma^2(\mathbf{X}_S^\top\mathbf{X}_S + \mathbf{A})^{-1}$. Note that $\boldsymbol{\Sigma}_{\mathbf{w}|\mathbf{y}_S}$ is well defined even if $\mathbf{A}$ is not invertible (i.e., an "improper prior"). In particular, it includes classical experimental design as the special case $\mathbf{A} = \mathbf{0}$, as well as the ridge-regularized case for $\mathbf{A} = \lambda\mathbf{I}$. Denoting $\boldsymbol{\Sigma}$ as the subset covariance $\mathbf{X}_S^\top\mathbf{X}_S$, we will use $f_{\mathbf{A}}(\boldsymbol{\Sigma})$ to represent the following standard Bayesian optimality criteria (Chaloner and Verdinelli, 1995; Pukelsheim, 2006):

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

Applications including clinical trials (Ryan et al., 2015; Ding et al., 2008; Spiegelhalter et al., 2004; Berry et al., 2002; Stangl and Berry, 1998; Flournoy, 1993), medical imaging (Owen et al., 2016), materials science (Frazier and Wang, 2016; Ueno et al., 2016; Terejanu et al.,

Proceedings of the 23$^{\text{rd}}$International Conference on Artificial Intelligence and Statistics (AISTATS) 2020, Palermo, Italy. PMLR: Volume 108. Copyright 2020 by the author(s).

2012), and biological process models (Ryan et al., 2016) all use these optimality criteria and thus stand to benefit from our contributions.

The general task we consider is the following combinatorial optimization problem, where $[n]$ denotes $\{1, ..., n\}$:

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

$$\operatorname*{argmin}_{S \subseteq [n]} f_{\mathbf{A}}(\mathbf{X}_S^\top \mathbf{X}_S) \quad \text{subject to} \quad |S| = k.$$

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

Question 1: Given $\mathbf{X}$, $f_{\mathbf{A}}$ and k, can we efficiently find a $(1+\epsilon)$-approximation for OPT_k?

Question 2: Given only $\mathbf{\Sigma}_{\mathbf{X}}$, $f_{\mathbf{A}}$ and k, what is the upper bound on 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 a baseline, we should expect meaningful results when k is at least $\Omega(d)$ (see discussion in Allen-Zhu et al., 2017), 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 should be able to 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* (Alaoui and Mahoney, 2015).

Definition 1 *For $d \times d$ positive semi-definite (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 $d_{\mathbf{A}}(\mathbf{\Sigma}_{\mathbf{X}})$.*

Goel and Klivans (2017) showed that $d_{\mathbf{A}}$ can be orders of magnitude smaller than the actual dimension d when the eigenvalues of $\mathbf{\Sigma}_{\mathbf{X}}$ exhibit fast decay, which is often the case in real datasets (Gittens and Mahoney, 2016). Recently, Dereziński and Warmuth (2018) 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.

1.1 Main results

Our main results provide new answers to Questions 1 and 2 by proposing a novel algorithm for Bayesian experimental design with strong theoretical guarantees.

Answer to Question 1. We propose an efficient $(1+\epsilon)$-approximation algorithm for A/C/D/V-optimal Bayesian experimental design:

Theorem 1 *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 we can find in polynomial time a subset S of size k s.t.*

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

Remark 1 *The algorithm referred to in Theorem 1 first solves a convex relaxation of the task via a semi-definite program (SDP) to find a weight vector $p \in [0,1]^n$, then uses our new randomized algorithm to round the weights to $\{0,1\}$, obtaining the subset S. The expected cost after SDP is $O(ndk + k^2d^2)$.*

A number of recent works studied $(1+\epsilon)$-approximate SDP-based algorithms for classical and Bayesian experimental design (see Table 1 and Section 2 for a comparison). Unlike *all* prior work on this topic, we are able to eliminate the dependence of the subset size k on the dimension d, replacing it with the potentially much smaller effective dimension $d_{\mathbf{A}}$. Our result also improves over the existing approaches in terms of the computational cost of the rounding procedure that is performed after solving the SDP. A number of different methods can be used to solve the SDP relaxation (see Section 5). For example, Allen-Zhu et al. (2017) 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.

Answer to Question 2. By performing a careful theoretical analysis of the performance of our algorithm, we are able to give an improved upper bound on OPT_k. In the below result, we use a more refined notion of effective dimensionality for Bayesian experimental design, $d_{\frac{n}{k}\mathbf{A}}$ (where the precision matrix $\mathbf{A}$ is scaled by factor $\frac{n}{k}$), which is smaller than $d_{\mathbf{A}}$ and therefore leads to a tighter bound.

Theorem 2 *Let $f_{\mathbf{A}}$ be A/C/D/V-optimality and $\mathbf{X}$ be $n \times d$. For any k such that $k \geq 4 d_{\frac{n}{k}\mathbf{A}}$,*

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

Remark 2 *We give a (randomized) algorithm which (with probability 1) finds the subset S that certifies this bound and has expected time complexity $O(ndk + k^2d^2)$.*

	Criteria	Bayesian	k	Cost after SDP
Wang et al. (2017)	A,V	✗	d^2/ϵ	$n^2 \cdot d$
Allen-Zhu et al. (2017)	A,C,D,E,G,V	✓	d/ϵ^2	$n \cdot kd^2$
Nikolov et al. (2019)	A,D	✗	d/ϵ	$n^4 \cdot k^2 d$
this paper	A,C,D,V	✓	$d_{\mathbf{A}}/\epsilon$	$n \cdot kd + k^2 d^2$

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). In the cost analysis, n could be replaced by the number of non-zero weights in the SDP solution. For simplicity we omit the log terms and assume that $\epsilon = \Omega(\frac{1}{d_{\mathbf{A}}})$. Our approach beats other methods both in terms of the runtime and the dependence of k on d (when $d_{\mathbf{A}} = o(d)$).

In particular, this means that if $k \geq 4d_{\frac{n}{k}\mathbf{A}}$ then 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 Dereziński and Warmuth (2018) in terms of the supported range of sizes k, but also in terms of the obtained bound (see Section 2 for a comparison). In Section 5, we we provide numerical evidence suggesting that for many real datasets the quantity $f_{\mathbf{A}}(\frac{k}{n}\mathbf{\Sigma}_{\mathbf{X}})$ provides a good estimate for OPT_k to within a factor of 2.

Theorem 2 suggests that the right notion of degrees of freedom for Bayesian experimental design can in fact be smaller than $d_{\mathbf{A}}$. Intuitively, since $d_{\mathbf{A}}$ is computed using the full data covariance $\mathbf{\Sigma}_{\mathbf{X}}$, it is not in the same scale as the smaller covariance $\mathbf{X}_S^\top \mathbf{X}_S$ based on the subset S of size $k \ll n$. In our result this is corrected by increasing the regularization on $\mathbf{\Sigma}_{\mathbf{X}}$ from $\mathbf{A}$ to $\frac{n}{k}\mathbf{A}$ and using $d_{\frac{n}{k}\mathbf{A}} = d_{\frac{n}{k}\mathbf{A}}(\mathbf{\Sigma}_{\mathbf{X}})$ as the degrees of freedom. 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).

1.2 Technical contributions

To establish Theorems 1 and 2, we develop a theoretical framework for a new sampling distribution which can be seen as a *regularized* variant of a determinantal point process (DPP). DPPs are a well-studied family of distributions with numerous applications in sampling diverse subsets of negatively correlated elements (see Kulesza and Taskar, 2012).

Given a psd matrix $\mathbf{A}$ and a weight vector $p = (p_1, ..., p_n) \in [0,1]^n$, we define $\mathrm{DPP}_{\mathrm{reg}}^p(\mathbf{X}, \mathbf{A})$ as a distribution over subsets $S \subseteq [n]$ (of all sizes) such that (see Definition 2):

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

A number of regularized DPPs have been proposed recently (Dereziński, 2019; Dereziński and Warmuth, 2018), mostly within the context of Randomized Numerical Linear Algebra (RandNLA) (Mahoney, 2011; Drineas and Mahoney, 2016, 2017). To our knowledge, ours is the first such definition that strictly falls under the umbrella of traditional DPPs (Kulesza and Taskar, 2012). 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 3). 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 3), which is of independent interest.

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 $S \sim \mathrm{DPP}_{\mathrm{reg}}^p(\mathbf{X}, \mathbf{A})$ and $f_{\mathbf{A}}$ is any one of the A/C/D/V-optimality criteria, then:

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

The proof of Theorem 2 relies on these two inequalities and a concentration bound for the subset size $|S|$, whereas to obtain Theorem 1 we additionally use the SDP relaxation to find the optimal weight vector p. When $\mathbf{A} = \mathbf{0}$, then $\mathrm{DPP}_{\mathrm{reg}}^p(\mathbf{X}, \mathbf{A})$ bears a lot of similarity to *proportional volume sampling* which is an (unregularized) determinantal distribution proposed by Nikolov et al. (2019). Our algorithm not only extends it to the Bayesian setting but also offers a drastic time complexity improvement from the $O(n^4 dk^2 \log k)$ required by Nikolov et al. (2019) down to the $O(nd^2)$ required for sampling from $\mathrm{DPP}_{\mathrm{reg}}^p(\mathbf{X}, \mathbf{A})$, and recent advances in RandNLA for DPP sampling (Dereziński et al., 2018, 2019; Dereziński, 2019) suggest that $O(nd \log n + \mathrm{poly}(d))$ time is also possible.

2 Related work

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). In this line of work we wish to find the smallest k for which a polynomial time approximation algorithm is possible. For example, Wang et al.

(2017) 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. Nikolov et al. (2019) 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 run in polynomial time, they scale very poorly with the number of experiments n (see Table 1). Allen-Zhu et al. (2017) 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 (in Theorem 1) improve on this work in two important ways:

- In terms of the dependence on ϵ for A/C/D/V-optimality,
- In terms of the dependence on the dimension (by replacing d with $d_{\mathbf{A}}$) in the Bayesian setting.

A lower bound shown by Nikolov et al. (2019) implies that our Theorem 1 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 2, however, unlike our bound, none of them apply to the regime of $k \le d$.

Non-trivial bounds for the *classical* A-optimality criterion (i.e., OPT_k with $\mathbf{A} = \mathbf{0}$) were first given by Avron and Boutsidis (2013), where they show that for any $k \ge d$, $\mathrm{OPT}_k \le (1 + \frac{d-1}{k-d+1}) \cdot f_{\mathbf{0}}(\frac{k}{n}\mathbf{\Sigma}_{\mathbf{X}})$ and the subset S attaining the bound can be found in polynomial time. The result was later extended (Dereziński and Warmuth, 2017, 2018; Dereziński and Warmuth, 2018) to the case where $\mathbf{A} = \lambda\mathbf{I}$, proving that for any $k \ge d_{\lambda\mathbf{I}}$, we have $\mathrm{OPT}_k \le (1 + \frac{d_{\lambda\mathbf{I}}-1}{k-d_{\lambda\mathbf{I}}+1}) \cdot f_{\frac{k}{n}\lambda\mathbf{I}}(\frac{k}{n}\mathbf{\Sigma}_{\mathbf{X}})$, and also a faster $O(nd^2)$ time algorithm was provided. In comparison, our results (in Theorem 2) offer the following improvements for upper bounding OPT_k:

- We cover a wider range of subset sizes, because $d_{\frac{n}{k}\lambda\mathbf{I}} \le d_{\lambda\mathbf{I}}$,
- Our upper bound can be much tighter because $f_{\lambda\mathbf{I}}(\frac{k}{n}\mathbf{\Sigma}_{\mathbf{X}}) \le f_{\frac{k}{n}\lambda\mathbf{I}}(\frac{k}{n}\mathbf{\Sigma}_{\mathbf{X}})$.

Additionally, Dereziński et al. (2019) 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.

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 (Bouhtou et al., 2010). Also, greedy algorithms for Bayesian A/V-optimality criteria have been considered by Bian et al. (2017) and Chamon and Ribeiro (2018). 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 (Cook and Nachtrheim, 1980). However, in this work we focus on methods that provide theoretical approximation guarantees.

3 A new regularized determinantal point process

In this section we develop the theory for a novel regularized extension of determinantal point processes (DPP) which we use as the sampling distribution for obtaining guarantees in Bayesian experimental design. DPPs form a family of distributions which are used to model repulsion between elements in a random set, with many applications in machine learning (Kulesza and Taskar, 2012). 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

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

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:

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

Crucially, every $\mathrm{DPP}_{\mathrm{ens}}$ is also a $\mathrm{DPP}_{\mathrm{cor}}$, but not the other way around. Specifically, $\mathrm{DPP}_{\mathrm{ens}}(\mathbf{L}) = \mathrm{DPP}_{\mathrm{cor}}(\mathbf{K})$ when:

$$\text{(a)} \quad \mathbf{L} = \mathbf{K}(\mathbf{I}-\mathbf{K})^{-1}, \qquad \text{(b)} \quad \mathbf{K} = \mathbf{I} - (\mathbf{I}+\mathbf{L})^{-1},$$

but (a) 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 $O(n\,|S|^2)$ Hough et al. (2006); Kulesza and Taskar (2012).

We now 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 $p = (p_1, \ldots, p_n) \in [0,1]^n$ 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 $S \subseteq [n]$:*

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

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

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

where $(*)$ follows from Lemma 7 of Dereziński and Mahoney (2019).

The main theoretical contribution 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 and unioning with i.i.d. Bernoulli samples:

Theorem 3 *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*

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

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 3 *Figure 1 illustrates how to exploit this result to build an efficient sampling algorithm. 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. In the process, we establish several novel theoretical properties regarding the representation, decomposition, and closure properties of regularized DPPs which may be of independent interest.

```
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 $T \sim \mathrm{DPP}_{\mathrm{cor}}(\mathbf{B}\mathbf{B}^\top)$          (Hough et al., 2006)
  Sample $b_i \sim \mathrm{Bernoulli}(p_i)$ for $i \in [n]$
  return  $S = T \cup \{i : b_i = 1\}$
```

Figure 1: Algorithm which exploits Theorem 3 to sample $S \sim \mathrm{DPP}^p_{\mathrm{reg}}(\mathbf{X},\mathbf{A})$ in $O(nd^2)$ time.

Lemma 1 *Given $\mathbf{X}$, $\mathbf{A}$ and $\mathbf{D}_p$ as in Theorem 3, assume that $\mathbf{A}$ and $\mathbf{I} - \mathbf{D}_p$ are invertible. Then,*

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

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

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

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 2 *Given $\mathbf{X}$, $\mathbf{A}$, and $\mathbf{D}_p$ as in Theorem 3 (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 2 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 3 *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 3 now follows by combining Lemmas 2 and 3.

4 Guarantees for Bayesian experimental design

In this section we prove our main results regarding Bayesian experimental design (Theorems 1 and 2). 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 $d_{\mathbf{A}}(\boldsymbol{\Sigma}) = \mathrm{tr}\big(\boldsymbol{\Sigma}(\mathbf{A}+\boldsymbol{\Sigma})^{-1}\big)$. The omitted proofs are in Appendix A.

Lemma 4 *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 3. Then*

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

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

$$\mathbb{E}\Big[\big(\mathbf{X}_S^\top\mathbf{X}_S+\mathbf{A}\big)^{-1}\Big] \preceq \Big(\sum_i p_i\mathbf{x}_i\mathbf{x}_i^\top + \mathbf{A}\Big)^{-1}, \qquad (2)$$

$$\mathbb{E}\Big[\det\big(\mathbf{X}_S^\top\mathbf{X}_S+\mathbf{A}\big)^{-1}\Big] \le \det\Big(\sum_i p_i\mathbf{x}_i\mathbf{x}_i^\top + \mathbf{A}\Big)^{-1}. \quad (3)$$

Corollary 1 *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 (2). For D-optimality, we apply (3) as follows:

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

which completes the proof. ■

Finally, we present the key lemma that puts everything together. This result is essentially a generalization of Theorem 2 from which also follows Theorem 1.

Lemma 6 *Let $f_{\mathbf{A}}$ be A/C/D/V-optimality and $\mathbf{X}$ be $n\times d$. For some $w=(w_1,\dots,w_n)\in[0,1]^n$, let $\boldsymbol{\Sigma}_w = \sum_i w_i\mathbf{x}_i\mathbf{x}_i^\top$ and assume that $\sum_i w_i = k\in[n]$. If $k \ge 4\,d_{\mathbf{A}}(\boldsymbol{\Sigma}_w)$, then a subset $S\subseteq[n]$ of size k can be found in $O(ndk+k^2d^2)$ time that satisfies*

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

Proof Let $p=(p_1,\dots,p_n)$ 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 Corollary 1, we have

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

Using Lemma 4 we can bound the expected size of S as follows:

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

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 $\mathrm{DPP}_{\mathrm{reg}}^p(\mathbf{X}, \mathbf{A})$ 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

$$\Pr\Big(f_{\mathbf{A}}(\mathbf{X}_S^\top \mathbf{X}_S) \geq (1+\delta) F_k \;\mid\; |S| \leq k\Big) \leq \frac{1}{1+\delta}.$$

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 $\mathrm{DPP}_{\mathrm{reg}}^p(\mathbf{X}, \mathbf{A})$ 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 $\mathrm{DPP}_{\mathrm{reg}}^p(\mathbf{X}, \mathbf{A})$ are needed, so the cost is $O(nd^2)$ for the eigendecomposition, $O(\frac{k}{d_w} \cdot n d_w^2) = O(n d_w k)$ for sampling and $O(\frac{k}{d_w} \cdot k d^2)$ for recomputing $f_{\mathbf{A}}(\mathbf{X}_S^\top \mathbf{X}_S)$. ■

To prove the main results, we use Lemma 6 with appropriately chosen weights w.

Proof of Theorem 1 As discussed by Allen-Zhu et al. (2017) and Boyd and Vandenberghe (2004), the following convex relaxation of experimental design can be written as a semi-definite program and solved using standard SDP solvers:

$$w^* \;=\; \operatorname*{argmin}_{w} \;\; f_{\mathbf{A}}\Big(\sum_{i=1}^n w_i \mathbf{x}_i \mathbf{x}_i^\top\Big), \tag{4}$$

$$\text{subject to} \quad \forall_i \;\; 0 \leq w_i \leq 1, \quad \sum_i w_i = k. \tag{5}$$

The solution w^* satisfies $f_{\mathbf{A}}\big(\mathbf{\Sigma}_{w^*}\big) \leq \mathrm{OPT}_k$. If we use w^* in Lemma 6, 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. ■

Proof of Theorem 2 Let $w = (\frac{k}{n}, ..., \frac{k}{n})$ in Lemma 6. 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. ■

5 Experiments

We confirm our theoretical results with experiments on real world data from `libsvm` datasets (Chang and Lin, 2011) (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,

$$f_{\mathbf{A}}(\mathbf{X}_S^\top \mathbf{X}_S) = \mathrm{tr}\left((\mathbf{X}_S^\top \mathbf{X}_S + \mathbf{A})^{-1}\right),$$

is compared against several baselines and recently proposed methods for A-optimal design that have been shown to perform well in practice. Note that none of these algorithms come with theoretical guarantees as strong as those offered by our approach. The list of implemented methods is as follows:

Our method (with SDP) uses the efficient algorithms developed in proving Theorem 1 to sample $\mathrm{DPP}_{\mathrm{reg}}^p(\mathbf{X}, \mathbf{A})$ 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, see O'Donoghue et al., 2016). We chose SCS because it can handle the

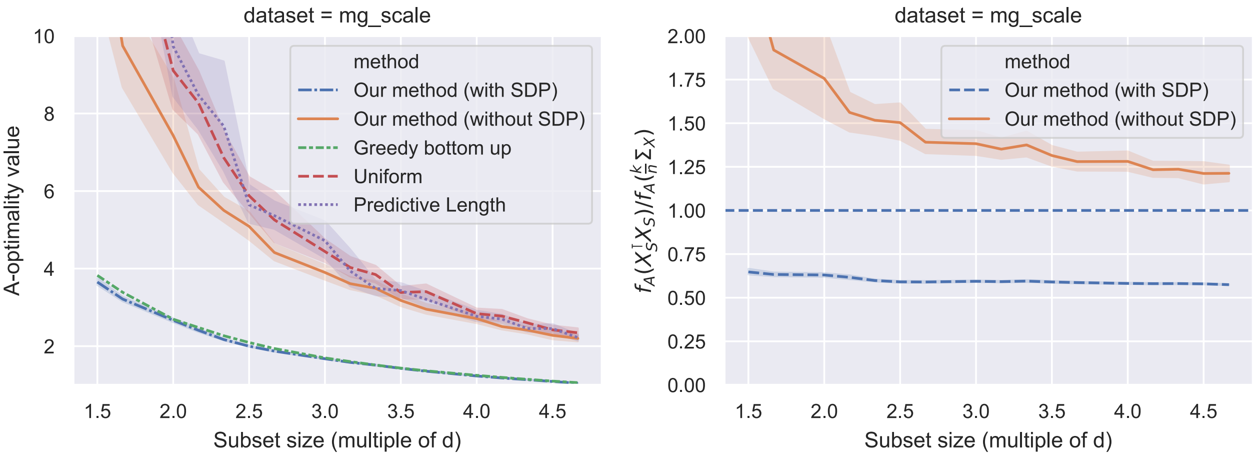

Figure 2: (left) A-optimality value obtained by the various methods on the `mg_scale` dataset (Chang and Lin, 2011) with prior precision $\mathbf{A} = 10^{-5}\,\mathbf{I}$, (right) A-optimality value for our method (with and without SDP) divided by $f_{\mathbf{A}}(\frac{k}{n}\mathbf{\Sigma_X})$, the baseline estimate suggested by Theorem 2.

SDP constraints in (5) and has provable termination guarantees, while also finding solutions faster (O'Donoghue et al., 2016) 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 $p \equiv \frac{k}{n}$.

Greedy bottom-up adds an index $i \in [n]$ to the sample S maximizing the increase in A-optimality criterion (Bian et al., 2017; Chamon and Ribeiro, 2017).

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

Predictive length sampling (Zhu et al., 2015) samples each row $\mathbf{x}_i$ of $\mathbf{X}$ with probability $\propto \|\mathbf{x}_i\|$.

Figure 2 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 3 shows that our method (without SDP) only differs from uniform sampling by an additional DPP sample with controlled expected size (see Lemma 4), we may conclude that adding even a small DPP sample can improve a uniformly sampled design.

Consistent with prior observations (Wang et al., 2017; Chamon and Ribeiro, 2017), the greedy bottom up method achieves surprisingly good performance, despite the limited theoretical guarantees it offers. 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 2 (right) numerically evaluates the tightness of the bound obtained in Theorem 2 by plotting the ratio:

$$\frac{f_{\mathbf{A}}(\mathbf{X}_S^\top \mathbf{X}_S)}{f_{\mathbf{A}}(\frac{k}{n}\mathbf{\Sigma_X})}$$

for subsets returned by our method (with and without SDP). Note that the line for our method with SDP on Figure 2 (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 OPT_k is within only a small constant factor away from $f_{\mathbf{A}}(\frac{k}{n}\mathbf{\Sigma_X})$, matching the upper bound of Theorem 2.

6 Conclusions

We proposed a new algorithm for finding $(1+\epsilon)$-approximate Bayesian experimental designs by leveraging a fundamental connection with determinantal point processes. Compared to the state-of-the-art approaches, our method provides stronger theoretical guarantees in terms of the allowed range of subset sizes, as well as offering significantly better time complexity guarantees. At the same time, our experiments show that on the task of A-optimal design the proposed algorithm performs as well as or better than several methods that are used in practice.

Acknowledgements

MWM would like to acknowledge ARO, DARPA, NSF and ONR for providing partial support of this work. Also, MWM and MD thank the NSF TRIPODS program and FL thanks the NPSC for funding. The authors thank Uthaipon T. Tantipongpipat for valuable discussions.

References

Alaoui, A. E. and Mahoney, M. W. (2015). 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.

Allen-Zhu, Z., Li, Y., Singh, A., and Wang, Y. (2017). 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.

Avron, H. and Boutsidis, C. (2013). Faster subset selection for matrices and applications. *SIAM Journal on Matrix Analysis and Applications*, 34(4):1464–1499.

Bernstein, D. S. (2011). *Matrix Mathematics: Theory, Facts, and Formulas*. Princeton University Press, second edition.

Berry, D. A., Mueller, P., Grieve, A. P., Smith, M., Parke, T., Blazek, R., Mitchard, N., and Krams, M. (2002). Adaptive bayesian designs for dose-ranging drug trials. In *Case studies in Bayesian statistics*, pages 99–181. Springer.

Bian, A. A., Buhmann, J. M., Krause, A., and Tschiatschek, S. (2017). Guarantees for greedy maximization of non-submodular functions with applications. In Precup, D. and Teh, Y. W., 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. PMLR.

Bouhtou, M., Gaubert, S., and Sagnol, G. (2010). Submodularity and randomized rounding techniques for optimal experimental design. *Electronic Notes in Discrete Mathematics*, 36:679–686.

Boyd, S. and Vandenberghe, L. (2004). *Convex optimization*. Cambridge university press.

Chaloner, K. and Verdinelli, I. (1995). Bayesian experimental design: A review. *Statist. Sci.*, 10(3):273–304.

Chamon, L. and Ribeiro, A. (2017). Approximate supermodularity bounds for experimental design. In *Advances in Neural Information Processing Systems*, pages 5403–5412.

Chamon, L. F. O. and Ribeiro, A. (2018). Greedy sampling of graph signals. *IEEE Transactions on Signal Processing*, 66(1):34–47.

Chang, C.-C. and Lin, C.-J. (2011). LIBSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*, 2:27:1–27:27.

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

Dereziński, M. (2019). Fast determinantal point processes via distortion-free intermediate sampling. In Beygelzimer, A. and Hsu, D., editors, *Proceedings of the Thirty-Second Conference on Learning Theory*, volume 99 of *Proceedings of Machine Learning Research*, pages 1029–1049, Phoenix, USA.

Dereziński, M., Clarkson, K. L., Mahoney, M. W., and Warmuth, M. K. (2019). Minimax experimental design: Bridging the gap between statistical and worst-case approaches to least squares regression. In Beygelzimer, A. and Hsu, D., editors, *Proceedings of the Thirty-Second Conference on Learning Theory*, volume 99 of *Proceedings of Machine Learning Research*, pages 1050–1069, Phoenix, USA.

Dereziński, M. and Mahoney, M. W. (2019). Distributed estimation of the inverse hessian by determinantal averaging. In Wallach, H., Larochelle, H., Beygelzimer, A., d Alché-Buc, F., Fox, E., and Garnett, R., editors, *Advances in Neural Information Processing Systems 32*, pages 11401–11411. Curran Associates, Inc.

Dereziński, M. and Warmuth, M. K. (2017). Unbiased estimates for linear regression via volume sampling. In *Advances in Neural Information Processing Systems 30*, pages 3087–3096, Long Beach, CA, USA.

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

Dereziński, M. and Warmuth, M. K. (2018). Subsampling for ridge regression via regularized volume sampling. In Storkey, A. and Perez-Cruz, F., editors, *Proceedings of the Twenty-First International Conference on Artificial Intelligence and Statistics*, pages 716–725, Playa Blanca, Lanzarote, Canary Islands.

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

Dereziński, M., Warmuth, M. K., and Hsu, D. (2019). Correcting the bias in least squares regression with volume-rescaled sampling. In Chaudhuri, K. and

Sugiyama, M., editors, *Proceedings of the 22nd International Conference on Artificial Intelligence and Statistics*, volume 89 of *Proceedings of Machine Learning Research*, pages 944–953. PMLR.

Ding, M., Rosner, G. L., and Müller, P. (2008). Bayesian optimal design for phase ii screening trials. *Biometrics*, 64(3):886–894.

Drineas, P. and Mahoney, M. W. (2016). RandNLA: Randomized numerical linear algebra. *Communications of the ACM*, 59:80–90.

Drineas, P. and Mahoney, M. W. (2017). Lectures on randomized numerical linear algebra. Technical report. Preprint: arXiv:1712.08880; To appear in: *Lectures of the 2016 PCMI Summer School on Mathematics of Data*.

Flournoy, N. (1993). A clinical experiment in bone marrow transplantation: Estimating a percentage point of a quantal response curve. In *case studies in Bayesian Statistics*, pages 324–336. Springer.

Frazier, P. I. and Wang, J. (2016). Bayesian optimization for materials design. In *Information Science for Materials Discovery and Design*, pages 45–75. Springer.

Gittens, A. and Mahoney, M. W. (2016). Revisiting the Nyström method for improved large-scale machine learning. *J. Mach. Learn. Res.*, 17(1):3977–4041.

Goel, S. and Klivans, A. (2017). Eigenvalue decay implies polynomial-time learnability for neural networks. In Guyon, I., Luxburg, U. V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., and Garnett, R., editors, *Advances in Neural Information Processing Systems 30*, pages 2192–2202. Curran Associates, Inc.

Hough, J. B., Krishnapur, M., Peres, Y., Virág, B., et al. (2006). Determinantal processes and independence. *Probability surveys*, 3:206–229.

Kulesza, A. and Taskar, B. (2012). *Determinantal Point Processes for Machine Learning*. Now Publishers Inc., Hanover, MA, USA.

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

Nikolov, A., Singh, M., and Tao Tantipongpipat, U. (2019). Proportional volume sampling and approximation algorithms for a -optimal design. In *Proceedings of the Thirtieth Annual ACM-SIAM Symposium on Discrete Algorithms*, pages 1369–1386.

Owen, D., Melbourne, A., Thomas, D., De Vita, E., Rohrer, J., and Ourselin, S. (2016). Optimisation of arterial spin labelling using bayesian experimental design. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 511–518. Springer.

O'Donoghue, B., Chu, E., Parikh, N., and Boyd, S. (2016). Conic optimization via operator splitting and homogeneous self-dual embedding. *Journal of Optimization Theory and Applications*, 169(3):1042–1068.

Pukelsheim, F. (2006). *Optimal Design of Experiments*. Society for Industrial and Applied Mathematics, Philadelphia, PA, USA.

Ryan, C. M., Drovandi, C. C., Pettitt, A. N., et al. (2016). Optimal bayesian experimental design for models with intractable likelihoods using indirect inference applied to biological process models. *Bayesian Analysis*, 11(3):857–883.

Ryan, E., Drovandi, C., and Pettitt, A. (2015). Fully bayesian experimental design for pharmacokinetic studies. *Entropy*, 17(3):1063–1089.

Spiegelhalter, D. J. et al. (2004). Incorporating bayesian ideas into health-care evaluation. *Statistical Science*, 19(1):156–174.

Stangl, D. K. and Berry, D. A. (1998). Bayesian statistics in medicine: Where are we and where should we be going? *Sankhyā: The Indian Journal of Statistics, Series B*, pages 176–195.

Terejanu, G., Upadhyay, R. R., and Miki, K. (2012). Bayesian experimental design for the active nitridation of graphite by atomic nitrogen. *Experimental Thermal and Fluid Science*, 36:178–193.

Ueno, T., Rhone, T. D., Hou, Z., Mizoguchi, T., and Tsuda, K. (2016). Combo: an efficient bayesian optimization library for materials science. *Materials discovery*, 4:18–21.

Wang, Y., Yu, A. W., and Singh, A. (2017). On computationally tractable selection of experiments in measurement-constrained regression models. *J. Mach. Learn. Res.*, 18(1):5238–5278.

Zhu, R., Ma, P., Mahoney, M. W., and Yu, B. (2015). Optimal subsampling approaches for large sample linear regression. *arXiv preprint arXiv:1509.05111*.