Chapter 1

Bayesian experimental design with regularized determinantal point processes

In this chapter, 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. Some of the results here were initially published in Michał Dereziński, Feynman Liang, and Michael Mahoney. “Bayesian experimental design using regularized determinantal point processes”. In: *International Conference on Artificial Intelligence and Statistics.* 2020, pp. 3197–3207.

1.1 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 [CV95], 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 [CV95; Puk06]:

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 [RDP15; DRM08; Spi+04; Ber+02; SB98; Flo93], medical imaging [Owe+16], materials science [FW16; Uen+16; TUM12], and biological process models [RDP+16] 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

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

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 $\boldsymbol{\Sigma}_{\mathbf{X}} = \mathbf{X}^\top \mathbf{X}$.

Question 1.1 *Given $\mathbf{X}$, $f_{\mathbf{A}}$ and k, can we efficiently find a $(1+\epsilon)$-approximation for OPT_k?*

Question 1.2 *Given only $\boldsymbol{\Sigma}_{\mathbf{X}}$, $f_{\mathbf{A}}$ 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 All+17], 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* [AM15].

Definition 1.1 *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 $d_{\mathbf{A}}(\mathbf{\Sigma}_{\mathbf{X}})$.*

[GK17] 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 [GM16]. 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.

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

Theorem 1.1 *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.*

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

Remark 1.1 *The algorithm referred to in Theorem 1.1 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 1.1 and Section 1.2 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 1.5). For example, [All+17] 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	Cost after SDP
[WYS17]	A,V	✗	d^2/ϵ	$n^2 \cdot d$
[All+17]	A,C,D,E,G,V	✓	d/ϵ^2	$n \cdot kd^2$
[NST19]	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 1.1: 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 $d_{\mathbf{A}} = o(d)$).

Answer to Question 1.2 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 1.2 *Let $f_{\mathbf{A}}$ be A/C/D/V-optimality and $\mathbf{X}$ be $n \times d$. For any k such that $k \geq 4d_{\frac{n}{k}\mathbf{A}}$,*

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

Remark 1.2 *We give a (randomized) algorithm which (with probability 1) finds the subset S that certifies this bound and has expected time complexity $O(ndk + k^2 d^2)$.*

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 [DW18b] in terms of the supported range of sizes k, but also in terms of the obtained bound (see Section 1.2 for a comparison). In Section 1.5, 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.

Comparison of different effective dimensions

Theorem 1.2 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.

Consider the two definitions we are comparing:

Full effective dimension $\quad d_{\mathbf{A}} = \mathrm{tr}\big(\mathbf{\Sigma}_{\mathbf{X}}(\mathbf{A}+\mathbf{\Sigma}_{\mathbf{X}})^{-1}\big)$,

Scaled effective dimension $d_{\frac{n}{k}\mathbf{A}} = \mathrm{tr}\big(\mathbf{\Sigma}_{\mathbf{X}}(\frac{n}{k}\mathbf{A}+\mathbf{\Sigma}_{\mathbf{X}})^{-1}\big)$.

Here, we demonstrate that these two effective dimensions can be very different for some matrices and quite similar on others. For simplicity, we consider two diagonal data covariance matrices as our examples: *identity covariance*, $\mathbf{\Sigma}_1 = \mathbf{I}$, and an *approximately low-rank covariance*, $\mathbf{\Sigma}_2 = (1-\epsilon)\frac{d}{s}\mathbf{I}_S + \epsilon\mathbf{I}$, where $\mathbf{I}_S$ is the diagonal matrix with ones on the entries indexed by subset $S \subseteq [d]$ of size $s < d$ and zeros everywhere else. The second matrix is scaled in such way so that $\mathrm{tr}(\mathbf{\Sigma}_1) = \mathrm{tr}(\mathbf{\Sigma}_2)$. We use $d = 100$, $s = 10$ and $\epsilon = 10^{-2}$. The prior precision matrix is $\mathbf{A} = 10^{-2}\,\mathbf{I}$. Figure 1.1 plots the scaled effective dimension $d_{\frac{n}{k}\mathbf{A}}$ as a function of k, against the full effective dimension for both examples. Unsurprisingly, for the identity covariance the full effective dimension is almost d, and the scaled effective dimension goes up very quickly to match it. On the other hand, for the approximately low-rank covariance, $d_{\mathbf{A}} \approx 55$ is considerably less then $d = 100$. Interestingly, the gap between the $d_{\frac{n}{k}\mathbf{A}}$ and $d_{\mathbf{A}}$ for moderately small values of k is even bigger. Our theory suggests that $d_{\frac{n}{k}\mathbf{A}}$ is a valid indicator of Bayesian degrees of freedom when $k \geq C \cdot d_{\frac{n}{k}\mathbf{A}}$ for some small constant C (Theorem 1.2 has $C = 4$, but we believe this can be improved to 1). While for the identity covariance the condition $k \geq d_{\frac{n}{k}\mathbf{A}}$ is almost equivalent to $k \geq d_{\mathbf{A}}$, in the approximately low-rank case, $k \geq d_{\frac{n}{k}\mathbf{A}}$ holds for k as small as 20, much less than $d_{\mathbf{A}}$.

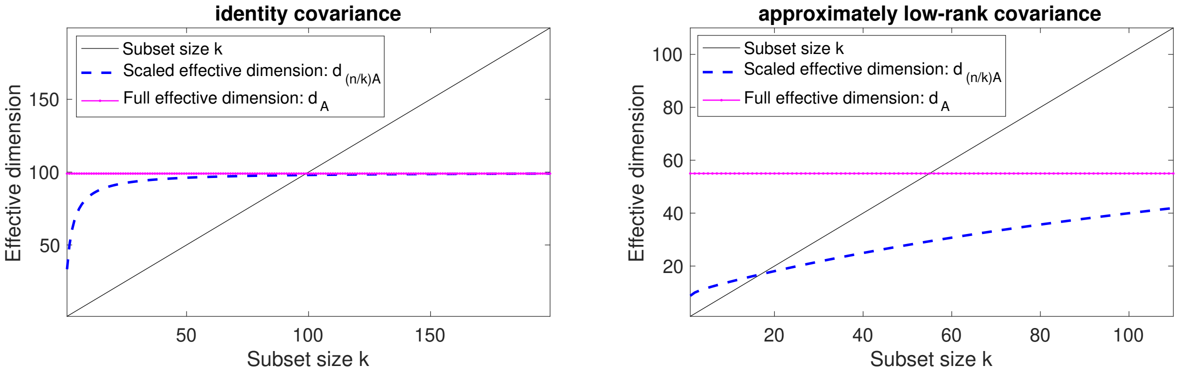

Figure 1.1: Scaled effective dimension compared to the full effective dimension for two diagonal data covariance matrices, with $\mathbf{A} = 10^{-2}\,\mathbf{I}$.

Technical contributions

To establish Theorems 1.1 and 1.2, 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 KT12].

Given a psd matrix $\mathbf{A}$ and a weight vector $p = (p_1, ..., p_n) \in [0,1]^n$, we define $\mathrm{DPP}^p_{\mathrm{reg}}(\mathbf{X}, \mathbf{A})$ as a distribution over subsets $S \subseteq [n]$ (of all sizes) such that (see Definition 1.2):

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

A number of regularized DPPs have been proposed recently [Der19; DW18b], mostly within the context of Randomized Numerical Linear Algebra (RandNLA) [Mic11; DM16; DM17]. To our knowledge, ours is the first such definition that strictly falls under the umbrella of traditional DPPs [KT12]. We show this in Section 1.3, where we also prove that regularized DPPs can be decomposed into a low-rank DPP plus i.i.d. Bernoulli sampling (Theorem 1.3). 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 1.3), which is of independent interest.

In Section 1.4 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:

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

The proof of Theorem 1.2 relies on these two inequalities and a concentration bound for the subset size $|S|$, whereas to obtain Theorem 1.1 we additionally use the SDP relaxation to find the optimal weight vector p. 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 [NST19]. 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 [NST19] down to the $O(nd^2)$ required for sampling from $\mathrm{DPP}^p_{\mathrm{reg}}(\mathbf{X}, \mathbf{A})$, and recent advances in RandNLA for DPP sampling [DWH18; DWH19; Der19] suggest that $O(nd \log n + \mathrm{poly}(d))$ time is also possible.

1.2 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 1.1 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, [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. [NST19] 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 1.1). [All+17] 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 1.1) 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 [NST19] implies that our Theorem 1.1 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 1.2, however, unlike our bound, none of them apply to the regime of $k \leq d$.

Non-trivial bounds for the *classical* A-optimality criterion (i.e., OPT_k with $\mathbf{A} = \mathbf{0}$) were first given by [AB13], where they show 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 the subset S attaining the bound can be found in polynomial time. The result was later extended [DW17; DW18b; DW18a] to the case where $\mathbf{A} = \lambda\mathbf{I}$, proving 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. In comparison, our results (in Theorem 1.2) offer the following improvements for upper bounding OPT_k:

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

Additionally, [Der+19] 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 [BGS10]. Also, greedy algorithms for Bayesian A/V-optimality criteria have been considered by [Bia+17] and [CR18]. 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.

1.3 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 [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

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

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:

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

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:

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

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 $O(n\,|S|^2)$ [Hou+06; KT12].

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

Definition 1.2 *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 $\mathrm{DPP}_{\mathrm{reg}}^{p}(\mathbf{X}, \mathbf{A})$ be a distribution over $S \subseteq [n]$:*

$$\Pr(S) = \frac{\det(\mathbf{X}_S^\top \mathbf{X}_S + \mathbf{A})}{\det\big(\sum_i p_i \mathbf{x}_i \mathbf{x}_i^\top + \mathbf{A}\big)} \cdot \prod_{i \in S} p_i \cdot \prod_{i \notin S} (1 - p_i). \tag{1.1}$$

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

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

where $(*)$ follows from Lemma 7 of Dereziński et al. [DM19].

The main theoretical contribution in this section is the following efficient algorithm for $\mathrm{DPP}_{\mathrm{reg}}^{p}(\mathbf{X}, \mathbf{A})$ which reduces it to sampling from a correlation DPP and unioning with i.i.d. Bernoulli samples:

Theorem 1.3 *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*

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

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

Remark 1.3 *Figure 1.2 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 $\mathrm{DPP}_{\mathrm{reg}}^{p}(\mathbf{X}, \mathbf{A})$ 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  $S \sim \mathrm{DPP}_{\mathrm{reg}}^{p}(\mathbf{X}, \mathbf{A})$
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 $T \sim \mathrm{DPP}_{\mathrm{cor}}(\mathbf{B}\mathbf{B}^\top)$                [Hou+06]
Sample $b_i \sim \mathrm{Bernoulli}(p_i)$ for $i \in [n]$
return  $S = T \cup \{i : b_i = 1\}$
```

Figure 1.2: Algorithm which exploits Theorem 1.3 to sample $S \sim \mathrm{DPP}_{\mathrm{reg}}^{p}(\mathbf{X}, \mathbf{A})$ in $O(nd^2)$ time.

Lemma 1.1 *Given $\mathbf{X}$, $\mathbf{A}$ and $\mathbf{D}_p$ as in Theorem 1.3, assume that $\mathbf{A}$ and $\mathbf{I} - \mathbf{D}_p$ are invertible. Then,*

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

Proof Let $S \sim \mathrm{DPP}^p_{\mathrm{reg}}(\mathbf{X},\mathbf{A})$. By Definition 1.2 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 1.2 *Given $\mathbf{X}$, $\mathbf{A}$, and $\mathbf{D}_p$ as in Theorem 1.3 (without any additional assumptions), we have*

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

When $\mathbf{A}$ and $\mathbf{I}-\mathbf{D}_p$ are invertible, then the proof 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 $\epsilon \to 0$.

Proof First, we show this under the invertibility assumptions of Lemma 1.1, i.e., given that $\mathbf{A}$ and $\mathbf{I}-\mathbf{D}_p$ are invertible. In this case $\mathrm{DPP}^p_{\mathrm{reg}}(\mathbf{X},\mathbf{A}) = \mathrm{DPP}_{\mathrm{ens}}(\mathbf{L})$, where

$$\mathbf{L} = \widetilde{\mathbf{D}} + \widetilde{\mathbf{D}}^{1/2}\mathbf{X}\mathbf{A}^{-1}\mathbf{X}^\top\widetilde{\mathbf{D}}^{1/2} \quad \text{and} \quad \widetilde{\mathbf{D}} = \mathbf{D}_p(\mathbf{I}-\mathbf{D}_p)^{-1}. \tag{1.2}$$

Converting this to a correlation kernel $\mathbf{K}$ and denoting $\widetilde{\mathbf{X}} = \mathbf{D}_p^{1/2}\mathbf{X}$, we obtain

$$\begin{aligned}
\mathbf{K} &= \mathbf{I} - (\mathbf{I}+\mathbf{L})^{-1} \\
&= \mathbf{I} - \Big(\mathbf{I} + (\mathbf{I}-\mathbf{D}_p)^{-1}\mathbf{D}_p + (\mathbf{I}-\mathbf{D}_p)^{-1/2}\widetilde{\mathbf{X}}\mathbf{A}^{-1}\widetilde{\mathbf{X}}^\top(\mathbf{I}-\mathbf{D}_p)^{-1/2}\Big)^{-1} \\
&= \mathbf{I} - \Big((\mathbf{I}-\mathbf{D}_p)^{-1} + (\mathbf{I}-\mathbf{D}_p)^{-1/2}\widetilde{\mathbf{X}}\mathbf{A}^{-1}\widetilde{\mathbf{X}}^\top(\mathbf{I}-\mathbf{D}_p)^{-1/2}\Big)^{-1} \\
&= \mathbf{I} - (\mathbf{I}-\mathbf{D}_p)^{1/2}(\mathbf{I} + \widetilde{\mathbf{X}}\mathbf{A}^{-1}\widetilde{\mathbf{X}}^\top)^{-1}(\mathbf{I}-\mathbf{D}_p)^{1/2} \\
&\overset{(*)}{=} \mathbf{I} - (\mathbf{I}-\mathbf{D}_p)^{1/2}\big(\mathbf{I} - \widetilde{\mathbf{X}}\mathbf{A}^{-1/2}(\mathbf{I} + \mathbf{A}^{-1/2}\widetilde{\mathbf{X}}^\top\widetilde{\mathbf{X}}\mathbf{A}^{-1/2})^{-1}\mathbf{A}^{-1/2}\widetilde{\mathbf{X}}^\top\big)(\mathbf{I}-\mathbf{D}_p)^{1/2} \\
&= \mathbf{I} - (\mathbf{I}-\mathbf{D}_p) + (\mathbf{I}-\mathbf{D}_p)^{1/2}\widetilde{\mathbf{X}}(\mathbf{A} + \widetilde{\mathbf{X}}^\top\widetilde{\mathbf{X}})^{-1}\widetilde{\mathbf{X}}^\top(\mathbf{I}-\mathbf{D}_p)^{1/2} \\
&= \mathbf{D}_p + (\mathbf{I}-\mathbf{D}_p)^{1/2}\widetilde{\mathbf{X}}(\mathbf{A} + \widetilde{\mathbf{X}}^\top\widetilde{\mathbf{X}})^{-1}\widetilde{\mathbf{X}}^\top(\mathbf{I}-\mathbf{D}_p)^{1/2},
\end{aligned}$$

where $(*)$ follows from Fact 2.16.19 in [Ber11]. Note that converting from $\mathbf{L}$ to $\mathbf{K}$ got rid of the inverses $\mathbf{A}^{-1}$ and $(\mathbf{I}-\mathbf{D}_p)^{-1}$ appearing in (1.2). The intuition is that when $\mathbf{A}$ or $\mathbf{I}-\mathbf{D}_p$ is non-invertible, then $\mathrm{DPP}_{\mathrm{reg}}^{p}(\mathbf{X},\mathbf{A})$ is not an L-ensemble but it is still a correlation DPP. To show this, we use a limit argument. For $\epsilon \in [0,1]$, let $p_\epsilon = (1-\epsilon)p$ and $\mathbf{A}_\epsilon = \mathbf{A} + \epsilon\mathbf{I}$. Observe that if $\epsilon > 0$ then $\mathbf{A}_\epsilon$ and $\mathbf{I}-\mathbf{D}_{p_\epsilon}$ are always invertible even if $\mathbf{A}$ and $\mathbf{I}-\mathbf{D}_p$ are not. Denote $\mathbf{K}_\epsilon$ as the above correlation kernel with p replaced by p_ϵ and $\mathbf{A}$ replaced by $\mathbf{A}_\epsilon$. Note that all matrix operations defining kernel $\mathbf{K}_\epsilon$ are continuous w.r.t. $\epsilon \in [0,1]$, including the inverse, since $\mathbf{A} + \widetilde{\mathbf{X}}^\top\widetilde{\mathbf{X}}$ is assumed to be invertible. Therefore, the following equalities hold (with limits taken point-wise and $\epsilon > 0$):

$$\mathrm{DPP}_{\mathrm{reg}}^{p}(\mathbf{X},\mathbf{A}) = \lim_{\epsilon\to 0}\mathrm{DPP}_{\mathrm{reg}}^{p_\epsilon}(\mathbf{X},\mathbf{A}_\epsilon) = \lim_{\epsilon\to 0}\mathrm{DPP}_{\mathrm{cor}}(\mathbf{K}_\epsilon) = \mathrm{DPP}_{\mathrm{cor}}(\mathbf{K}),$$

where we did not have to assume invertibility of $\mathbf{A}$ or $\mathbf{I}-\mathbf{D}_p$. ■

Finally, we show that the correlation DPP arrived at in Lemma 1.2 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 1.3 *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*

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

Proof For this proof we will use the shorthand $\mathbf{K}_A$ for $\mathbf{K}_{A,A}$. If $\mathbf{D}$ has no zeros on the diagonal then $\det(\mathbf{D}_A) > 0$ for all $A \subseteq [n]$ and

$$\begin{aligned}
\Pr(A \subset T \cup R) &= \sum_{B \subset A} \Pr(R \cap A = A \setminus B)\ \Pr(B \subseteq T)\\
&= \sum_{B \subset A} \det(\mathbf{D}_{A \setminus B}) \det\big([\mathbf{I} - \mathbf{D}]_B\big)\ \det(\mathbf{K}_B)\\
&= \sum_{B \subset A} \det(\mathbf{D}_{A \setminus B}) \det\Big(\big[(\mathbf{I} - \mathbf{D})^{1/2}\mathbf{K}(\mathbf{I} - \mathbf{D})^{1/2}\big]_B\Big)\\
&= \det(\mathbf{D}_A) \sum_{B \subset A} \det\Big(\big[\mathbf{D}^{-1/2}(\mathbf{I} - \mathbf{D})^{1/2}\mathbf{K}(\mathbf{I} - \mathbf{D})^{1/2}\mathbf{D}^{-1/2}\big]_B\Big)\\
&\overset{(*)}{=} \det(\mathbf{D}_A) \det\Big(\mathbf{I} + \big[\mathbf{D}^{-1/2}(\mathbf{I} - \mathbf{D})^{1/2}\mathbf{K}(\mathbf{I} - \mathbf{D})^{1/2}\mathbf{D}^{-1/2}\big]_A\Big)\\
&= \det\Big(\big[\mathbf{D} + (\mathbf{I} - \mathbf{D})^{1/2}\mathbf{K}(\mathbf{I} - \mathbf{D})^{1/2}\big]_A\Big),
\end{aligned}$$

where $(*)$ follows from a standard determinantal identity used to compute the L-ensemble partition function [KT12, Theorem 2.1]. If $\mathbf{D}$ has zeros on the diagonal, a similar limit argument as in Lemma 1.2 with $\mathbf{D}_\epsilon = \mathbf{D} + \epsilon\,\mathbf{I}$ holds. ■

Theorem 1.3 now follows by combining Lemmas 1.2 and 1.3.

1.4 Guarantees for Bayesian experimental design

In this section we prove our main results regarding Bayesian experimental design (Theorems 1.1 and 1.2). 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 $d_{\mathbf{A}}(\mathbf{\Sigma}) = \mathrm{tr}\big(\mathbf{\Sigma}(\mathbf{A} + \mathbf{\Sigma})^{-1}\big)$.

Lemma 1.4 *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 1.3. Then*

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

Proof For correlation kernels it is known that the expected size of $\mathrm{DPP}_{\mathrm{cor}}(\mathbf{K})$ is $\mathrm{tr}(\mathbf{K})$. Thus, using $\mathbf{D}_p = \mathrm{diag}(p)$, we can invoke Lemma 1.2 to obtain

$$\begin{aligned}\mathbb{E}\big[|S|\big] &= \mathrm{tr}\Big(\mathbf{D}_p + (\mathbf{I}-\mathbf{D}_p)^{1/2}\,\mathbf{D}_p^{1/2}\mathbf{X}(\mathbf{A}+\mathbf{X}^\top\mathbf{D}_p\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{D}_p^{1/2}(\mathbf{I}-\mathbf{D}_p)^{1/2}\Big)\\ &\leq \mathrm{tr}(\mathbf{D}_p) + \mathrm{tr}\Big(\mathbf{D}_p^{1/2}\mathbf{X}(\mathbf{A}+\mathbf{X}^\top\mathbf{D}_p\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{D}_p^{1/2}\Big)\\ &= \mathrm{tr}(\mathbf{D}_p) + \mathrm{tr}\Big(\mathbf{X}^\top\mathbf{D}_p\mathbf{X}(\mathbf{A}+\mathbf{X}^\top\mathbf{D}_p\mathbf{X})^{-1}\Big) = \mathrm{tr}(\mathbf{D}_p) + d_{\mathbf{A}}(\mathbf{X}^\top\mathbf{D}_p\mathbf{X}),\end{aligned}$$

from which the claim follows. ■

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

$$\mathbb{E}\Big[\big(\mathbf{X}_S^\top\mathbf{X}_S+\mathbf{A}\big)^{-1}\Big] \preceq \Big(\sum_i p_i\mathbf{x}_i\mathbf{x}_i^\top + \mathbf{A}\Big)^{-1}, \tag{1.3}$$

$$\mathbb{E}\Big[\det\big(\mathbf{X}_S^\top\mathbf{X}_S+\mathbf{A}\big)^{-1}\Big] \leq \det\Big(\sum_i p_i\mathbf{x}_i\mathbf{x}_i^\top + \mathbf{A}\Big)^{-1}. \tag{1.4}$$

Proof For a square matrix $\mathbf{M}$, define its adjugate, denoted $\mathrm{adj}(\mathbf{M})$, as a matrix whose i,j-th entry is $(-1)^{i+j}\det(\mathbf{M}_{-j,-i})$, where $\mathbf{M}_{-j,-i}$ is the matrix $\mathbf{M}$ without jth row and ith column. If $\mathbf{M}$ is invertible, then $\mathrm{adj}(\mathbf{M}) = \det(\mathbf{M})\mathbf{M}^{-1}$. Now, let $b_i \sim \mathrm{Bernoulli}(p_i)$ be independent random variables. As seen in previous section, the identity $\mathbb{E}[\det(\sum_i b_i\mathbf{x}_i\mathbf{x}_i^\top + \mathbf{A})] = \det(\sum_i p_i\mathbf{x}_i\mathbf{x}_i^\top + \mathbf{A})$ gives us the normalization constant for $\mathrm{DPP}_{\mathrm{reg}}^{p}(\mathbf{X},\mathbf{A})$. Moreover, as noted in a different context by [DM19], when applied entrywise to the adjugate matrix, this identity implies that $\mathbb{E}[\mathrm{adj}(\sum_i b_i\mathbf{x}_i\mathbf{x}_i^\top + \mathbf{A})] = \mathrm{adj}(\sum_i p_i\mathbf{x}_i\mathbf{x}_i^\top + \mathbf{A})$. Let $\mathcal{I}$ denote the set of all subsets $S \subseteq [n]$ such that $\mathbf{X}_S^\top\mathbf{X}_S + \mathbf{A}$ is invertible. We have

$$\begin{aligned}\mathbb{E}\Big[\big(\mathbf{X}_S^\top\mathbf{X}_S+\mathbf{A}\big)^{-1}\Big] &= \sum_{S\in\mathcal{I}}\big(\mathbf{X}_S^\top\mathbf{X}_S+\mathbf{A}\big)^{-1}\frac{\det(\mathbf{X}_S^\top\mathbf{X}_S+\mathbf{A})}{\det(\sum_i p_i\mathbf{x}_i\mathbf{x}_i^\top+\mathbf{A})}\prod_{i\in S}p_i\prod_{i\notin S}(1-p_i)\\ &= \sum_{S\in\mathcal{I}}\frac{\mathrm{adj}(\mathbf{X}_S^\top\mathbf{X}_S+\mathbf{A})}{\det(\sum_i p_i\mathbf{x}_i\mathbf{x}_i^\top+\mathbf{A})}\prod_{i\in S}p_i\prod_{i\notin S}(1-p_i)\\ &\preceq \sum_{S\subseteq[n]}\frac{\mathrm{adj}(\mathbf{X}_S^\top\mathbf{X}_S+\mathbf{A})}{\det(\sum_i p_i\mathbf{x}_i\mathbf{x}_i^\top+\mathbf{A})}\prod_{i\in S}p_i\prod_{i\notin S}(1-p_i)\\ &= \frac{\mathbb{E}\big[\mathrm{adj}(\sum_i b_i\mathbf{x}_i\mathbf{x}_i^\top+\mathbf{A})\big]}{\det(\sum_i p_i\mathbf{x}_i\mathbf{x}_i^\top+\mathbf{A})}\\ &= \frac{\mathrm{adj}(\sum_i p_i\mathbf{x}_i\mathbf{x}_i^\top+\mathbf{A})}{\det(\sum_i p_i\mathbf{x}_i\mathbf{x}_i^\top+\mathbf{A})} = \Big(\sum_i p_i\mathbf{x}_i\mathbf{x}_i^\top+\mathbf{A}\Big)^{-1}.\end{aligned}$$

Note that if $\mathcal{I}$ contains all subsets of $[n]$, for example when $\mathbf{A} \succ \mathbf{0}$, then the inequality turns into equality. Thus, we showed (1.3), and (1.4) follows even more easily:

$$\mathbb{E}\Big[\det\big(\mathbf{X}_S^\top \mathbf{X}_S + \mathbf{A}\big)^{-1}\Big] = \sum_{S \in \mathcal{I}} \frac{1}{\det(\sum_i p_i \mathbf{x}_i \mathbf{x}_i^\top + \mathbf{A})} \prod_{i \in S} p_i \prod_{i \notin S} (1 - p_i) \leq \det\Big(\sum_i p_i \mathbf{x}_i \mathbf{x}_i^\top\Big)^{-1},$$

where the equality holds if $\mathcal{I}$ consists of all subsets of $[n]$. ■

Corollary 1.1 *Let $f_{\mathbf{A}}$ be A/C/D/V-optimality. Whenever $S \sim \mathrm{DPP}_{\mathrm{reg}}^{p}(\mathbf{X}, \mathbf{A})$ is well-defined,*

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

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 (1.3). For D-optimality, we apply (1.4) as follows:

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

which completes the proof. ■

Finally, we present the key lemma that puts everything together. This result is essentially a generalization of Theorem 1.2 from which also follows Theorem 1.1.

Lemma 1.6 *Let $f_{\mathbf{A}}$ be A/C/D/V-optimality and $\mathbf{X}$ be $n \times d$. For some $w = (w_1, \ldots, w_n) \in [0,1]^n$, 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*

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

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 Corollary 1.1, we have

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

Using Lemma 1.4 we can bound the expected size of S as follows:

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

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

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

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

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

Denoting $\mathbb{E}\big[f_{\mathbf{A}}(\mathbf{X}_S^\top \mathbf{X}_S) \mid |S| \le k\big]$ as F_k, Markov's inequality implies that

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

Also, we showed that $\Pr(|S| \le k) \ge 1 - \frac{d_w}{k} \ge \frac{3}{4}$. 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

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

We can sample from $\mathrm{DPP}_{\mathrm{reg}}^p(\mathbf{X}, \mathbf{A})$ conditioned on $|S| \le k$ 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, $O(\frac{k}{d_w} \cdot nd_w^2) = O(nd_wk)$ 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 1.6 with appropriately chosen weights w.

Proof of Theorem 1.1 As discussed by [All+17] and [BV04], the following convex relaxation of experimental design can be written as a semi-definite program and solved using standard SDP solvers:

$$w^* \;=\; \operatorname*{argmin}_{w} \quad f_{\mathbf{A}}\Big(\sum_{i=1}^{n} w_i \mathbf{x}_i \mathbf{x}_i^\top\Big), \tag{1.5}$$

$$\text{subject to} \quad \forall_i \;\; 0 \le w_i \le 1, \quad \sum_i w_i = k. \tag{1.6}$$

The solution w^* satisfies $f_{\mathbf{A}}\big(\mathbf{\Sigma}_{w^*}\big) \le \mathrm{OPT}_k$. If we use w^* in Lemma 1.6, then observing that $d_{\mathbf{A}}(\mathbf{\Sigma}_{w^*}) \le d_{\mathbf{A}}$, and setting $k \ge C(\frac{d_{\mathbf{A}}}{\epsilon} + \frac{\log 1/\epsilon}{\epsilon^2})$ for sufficiently large C, the algorithm in the lemma finds subset S such that

$$f_{\mathbf{A}}(\mathbf{X}_S^\top \mathbf{X}_S) \le (1+\epsilon) \cdot f_{\mathbf{A}}(\mathbf{\Sigma}_{w^*}) \le (1+\epsilon) \cdot \mathrm{OPT}_k.$$

Note that we did not need to solve the SDP exactly, so approximate solvers could be used instead. ■

Proof of Theorem 1.2 Let $w = (\frac{k}{n}, ..., \frac{k}{n})$ in Lemma 1.6. 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 $\mathrm{OPT}_k \le f_{\mathbf{A}}(\mathbf{X}_S^\top \mathbf{X}_S)$, the result follows. ■

1.5 Experiments

We confirm our theoreticala results with experiments on real world data from `libsvm` datasets [CL11] (more details in Appendix **??**). For all our experiments, the prior precision matrix is

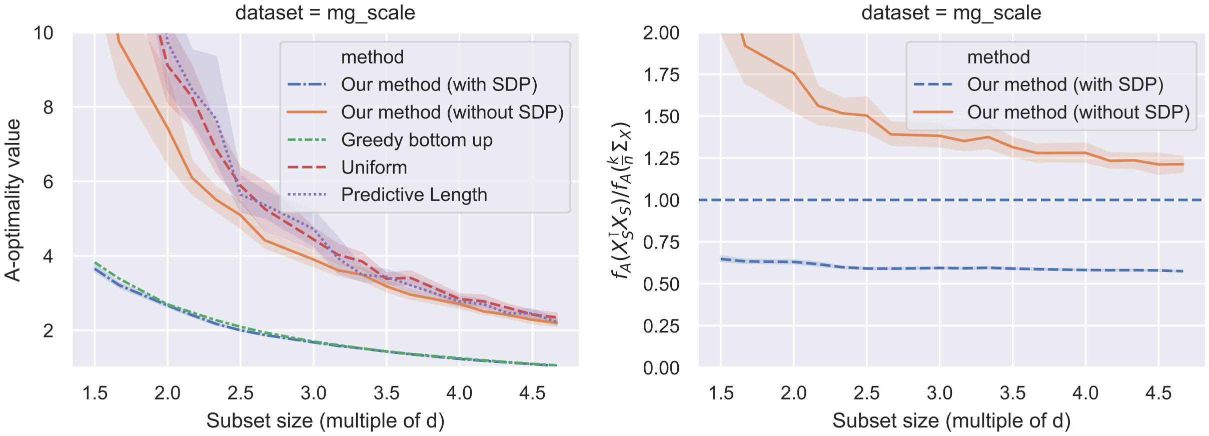

Figure 1.3: (left) A-optimality value obtained by the various methods on the `mg_scale` dataset [CL11] with prior precision $\mathbf{A} = 10^{-5}\,\mathbf{I}$, (right) A-optimality value for our method (with and without SDP) divided by $f_{\mathbf{A}}(\frac{k}{n}\mathbf{\Sigma}_{\mathbf{X}})$, the baseline estimate suggested by Theorem 1.2.

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 1.1 to sample $\mathrm{DPP}_{\mathrm{reg}}^p(\mathbf{X}, \mathbf{A})$ constrained to subset size k with $p = w^*$, see (1.6), obtained using a recently developed first order convex cone solver called Splitting Conical Solver [SCS, see ODo+16]. We chose SCS because it can handle the SDP constraints in (1.6) and has provable termination guarantees, while also finding solutions faster [ODo+16] than alternative off-the-shelf optimization software libraries such as SDPT3 and Sedumi.

Our method (without SDP) samples $\mathrm{DPP}_{\mathrm{reg}}^p(\mathbf{X}, \mathbf{A})$ 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 [Bia+17; CR17].

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

Predictive length sampling [Zhu+15] samples each row $\mathbf{x}_i$ of $\mathbf{X}$ with probability $\propto \|\mathbf{x}_i\|$.

Figure 1.3 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 1.3 shows that our method (without SDP) only differs from uniform sampling by an additional DPP sample with controlled expected size (see Lemma 1.4), we may conclude that adding even a small DPP sample can improve a uniformly sampled design.

Consistent with prior observations [WYS17; CR17], 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 **??**) 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 1.3 (right) numerically evaluates the tightness of the bound obtained in Theorem 1.2 by plotting the ratio:

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

for subsets returned by our method (with and without SDP). Note that the line for our method with SDP on Figure 1.3 (right) shows that the ratio never goes below 0.5, and we saw similar behavior across all examined datasets (see Appendix **??**). This evidence suggests that for many real datasets OPT_k is within only a small constant factor away from $f_{\mathtt{A}}(\frac{k}{n}\mathbf{\Sigma}_{\mathbf{X}})$, matching the upper bound of Theorem 1.2.

In addition to the `mg_scale` dataset presented in Section **??**, we also benchmarked on three other data sets described in Table 1.2.

Table 1.2: Datasets used in the experiments [CL11].

	`mg_scale`	`bodyfat_scale`	`mpg_scale`	`housing_scale`
n	1385	252	392	506
d	6	14	7	13

The A-optimality values obtained are illustrated in Figure 1.4. The general trend observed in Section **??** of our method (without SDP) outperforming independent sampling methods (uniform and predictive length) and our method (with SDP) matching the performance of the greedy bottom up method continues to hold across the additional datasets considered.

The relative ranking and overall order of magnitude differences between runtimes (Figure 1.5) are also similar across the various datasets. An exception to the rule is on `mg_scale`,

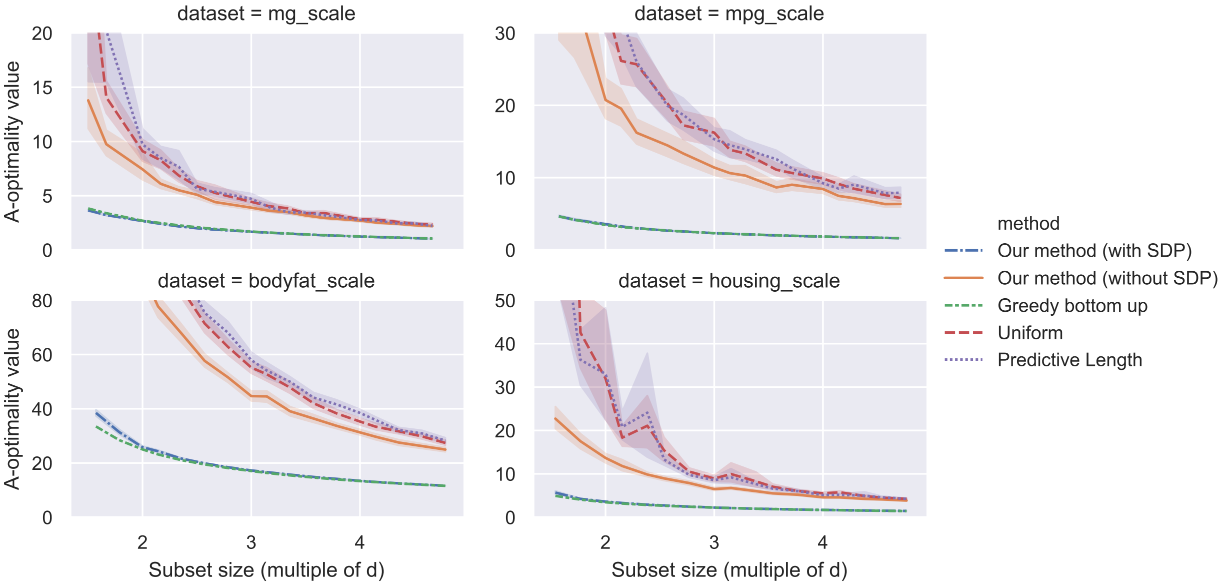

Figure 1.4: A-optimality values achieved by the methods compared. In all cases considered, we found our method (without SDP) to be superior to independent sampling methods like uniform and predictive length sampling. After paying the price to solve an SDP, our method (with SDP) is able to consistently match the performance of a greedy method which has been noted [CR17] to work well empirically.

where we see that our method (without SDP) costs more than the greedy method (whereas everywhere else it costs less).

The claim that $f_{\mathbf{A}}(\frac{k}{n}\mathbf{\Sigma}_{\mathbf{X}})$ is an appropriate quantity to summarize the contribution of problem-dependent factors on the performance of Bayesian A-optimal designs is further evidenced in Figure 1.6. Here, we see that after normalizing the A-optimality values by this quantity, the remaining quantities are all on the same scale and close to 1.

1.6 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.

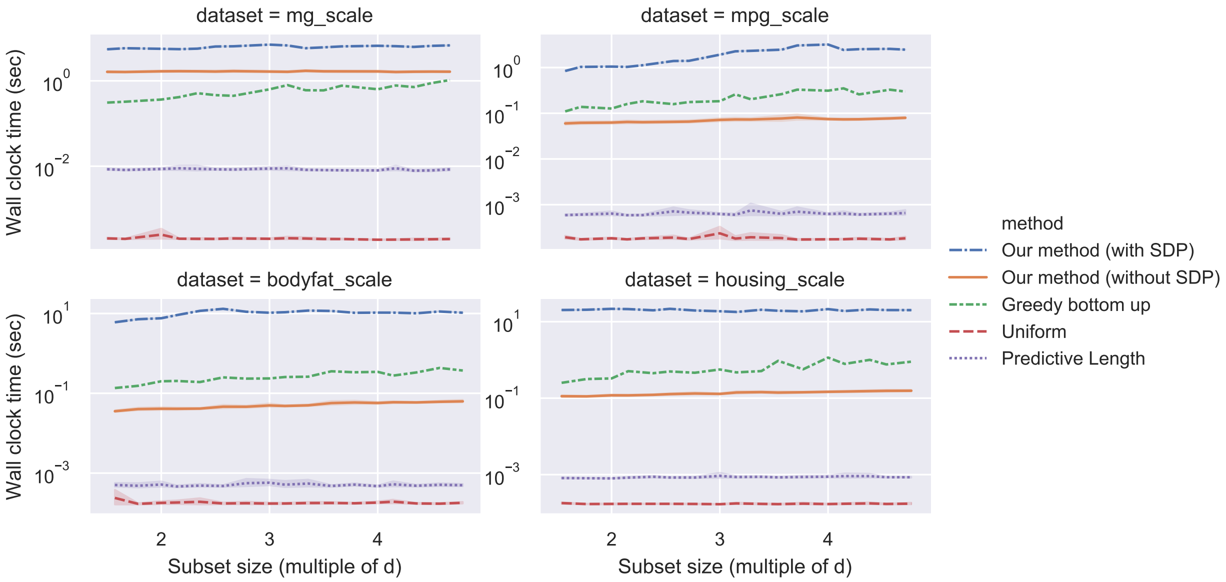

Figure 1.5: Runtimes of the methods compared. Our method (without SDP) is within an order of magnitude of greedy bottom up and faster in 3 out of 4 cases. The gap between our method with and without SDP is attributable to the SDP solver, making investigation of more efficient solvers and approximate solutions an interesting direction for future work.

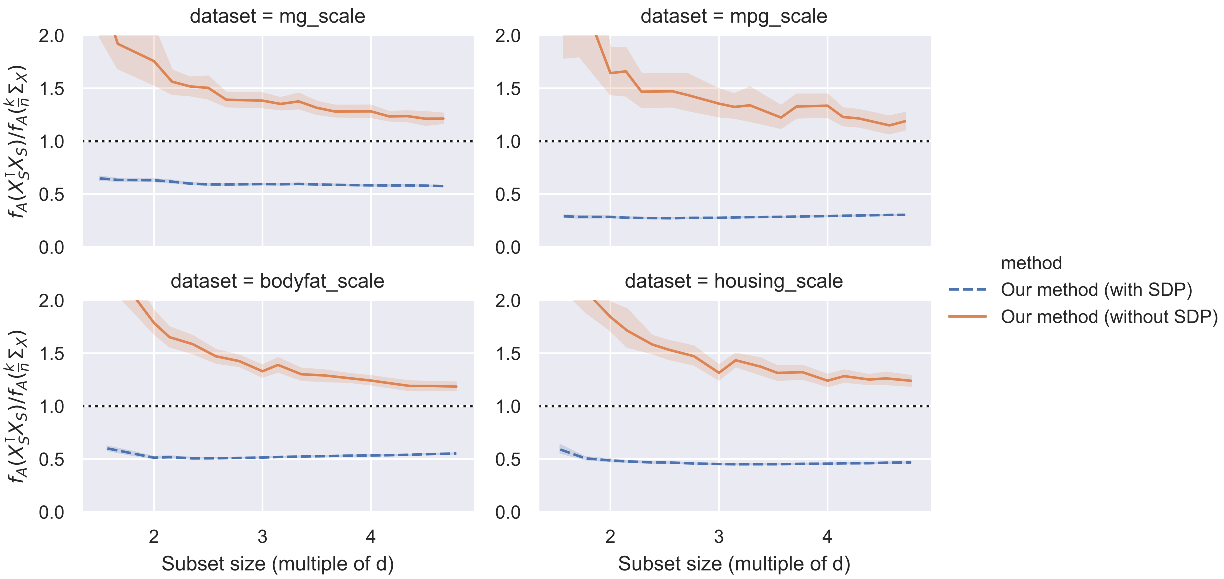

Figure 1.6: The ratio controlled by Lemma 1.6. This ratio converges to 1 as $k \to n$ and is close to 1 across all real world datasets, suggesting that $f_{\mathbf{A}}(\frac{k}{n}\mathbf{\Sigma_X})$ is an appropriate problem-dependent scale for Bayesian A-optimal experimental design.

Bibliography

[AB13] Haim Avron and Christos Boutsidis. “Faster Subset Selection for Matrices and Applications”. In: *SIAM Journal on Matrix Analysis and Applications* 34.4 (2013), pp. 1464–1499.

[All+17] 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*. Vol. 70. Proceedings of Machine Learning Research. Sydney, Australia, Aug. 2017, pp. 126–135. URL: http://proceedings.mlr.press/v70/allen-zhu17e.html.

[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*. 2015, pp. 775–783.

[Ber+02] Donald A Berry, Peter Mueller, Andy P Grieve, Michael Smith, Tom Parke, Richard Blazek, Neil Mitchard, and Michael Krams. “Adaptive Bayesian designs for dose-ranging drug trials”. In: *Case studies in Bayesian statistics*. Springer, 2002, pp. 99–181.

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

[BGS10] Mustapha Bouhtou, Stéphane Gaubert, and Guillaume Sagnol. “Submodularity and Randomized rounding techniques for Optimal Experimental Design”. In: *Electronic Notes in Discrete Mathematics* 36 (Aug. 2010), pp. 679–686. DOI: 10.1016/j.endm.2010.05.086.

[Bia+17] Andrew An Bian, Joachim M. Buhmann, Andreas Krause, and Sebastian Tschiatschek. “Guarantees for Greedy Maximization of Non-submodular Functions with Applications”. In: *Proceedings of the 34th International Conference on Machine Learning*. Ed. by Doina Precup and Yee Whye Teh. Vol. 70. Proceedings of Machine Learning Research. International Convention Centre, Sydney, Australia: PMLR, June 2017, pp. 498–507. URL: http://proceedings.mlr.press/v70/bian17a.html.

[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". In: *ACM Transactions on Intelligent Systems and Technology* 2 (3 2011), 27:1–27:27.

[CN80] R Dennis Cook and Christopher J Nachtrheim. "A comparison of algorithms for constructing exact D-optimal designs". In: *Technometrics* 22.3 (1980), pp. 315–324.

[CR17] Luiz Chamon and Alejandro Ribeiro. "Approximate supermodularity bounds for experimental design". In: *Advances in Neural Information Processing Systems*. 2017, pp. 5403–5412.

[CR18] L. F. O. Chamon and A. Ribeiro. "Greedy Sampling of Graph Signals". In: *IEEE Transactions on Signal Processing* 66.1 (Jan. 2018), pp. 34–47.

[CV95] Kathryn Chaloner and Isabella Verdinelli. "Bayesian Experimental Design: A Review". In: *Statist. Sci.* 10.3 (Aug. 1995), pp. 273–304. DOI: `10.1214/ss/1177009939`. URL: `https://doi.org/10.1214/ss/1177009939`.

[Der+19] 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". In: *Proceedings of the Thirty-Second Conference on Learning Theory*. Ed. by Alina Beygelzimer and Daniel Hsu. Vol. 99. Proceedings of Machine Learning Research. Phoenix, USA, 25–28 Jun 2019, pp. 1050–1069.

[Der19] Michał Dereziński. "Fast determinantal point processes via distortion-free intermediate sampling". In: *Proceedings of the Thirty-Second Conference on Learning Theory*. 2019, pp. 1029–1049.

[DLM20] Michał Dereziński, Feynman Liang, and Michael Mahoney. "Bayesian experimental design using regularized determinantal point processes". In: *International Conference on Artificial Intelligence and Statistics*. 2020, pp. 3197–3207.

[DM16] Petros Drineas and Michael W. Mahoney. "RandNLA: Randomized Numerical Linear Algebra". In: *Communications of the ACM* 59 (2016), pp. 80–90.

[DM17] Petros Drineas and Michael W. Mahoney. *Lectures on Randomized Numerical Linear Algebra*. Tech. rep. Preprint: arXiv:1712.08880; To appear in: *Lectures of the 2016 PCMI Summer School on Mathematics of Data*. 2017.

[DM19] Michał Dereziński and Michael W Mahoney. "Distributed estimation of the inverse Hessian by determinantal averaging". In: *Advances in Neural Information Processing Systems 32*. Ed. by H. Wallach, H. Larochelle, A. Beygelzimer, F. d Alché-Buc, E. Fox, and R. Garnett. Curran Associates, Inc., 2019, pp. 11401–11411.

[DRM08] Meichun Ding, Gary L Rosner, and Peter Müller. "Bayesian optimal design for phase II screening trials". In: *Biometrics* 64.3 (2008), pp. 886–894.

[DW17] Michał Dereziński and Manfred K. Warmuth. "Unbiased estimates for linear regression via volume sampling". In: *Advances in Neural Information Processing Systems 30*. Long Beach, CA, USA, 2017, pp. 3087–3096.

[DW18a] Michał Dereziński and Manfred K. Warmuth. "Reverse Iterative Volume Sampling for Linear Regression". In: *Journal of Machine Learning Research* 19.23 (2018), pp. 1–39.

[DW18b] Michał Dereziński and Manfred K. Warmuth. "Subsampling for Ridge Regression via Regularized Volume Sampling". In: *Proceedings of the Twenty-First International Conference on Artificial Intelligence and Statistics*. Ed. by Amos Storkey and Fernando Perez-Cruz. Playa Blanca, Lanzarote, Canary Islands, Apr. 2018, pp. 716–725.

[DWH18] Michał Dereziński, Manfred K. Warmuth, and Daniel Hsu. "Leveraged volume sampling for linear regression". In: *Advances in Neural Information Processing Systems 31*. Ed. by S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett. Curran Associates, Inc., 2018, pp. 2510–2519.

[DWH19] Michał Dereziński, Manfred K. Warmuth, and Daniel Hsu. "Correcting the bias in least squares regression with volume-rescaled sampling". In: *Proceedings of the 22nd International Conference on Artificial Intelligence and Statistics*. Ed. by Kamalika Chaudhuri and Masashi Sugiyama. Vol. 89. Proceedings of Machine Learning Research. PMLR, 16–18 Apr 2019, pp. 944–953.

[Flo93] Nancy Flournoy. "A clinical experiment in bone marrow transplantation: Estimating a percentage point of a quantal response curve". In: *case studies in Bayesian Statistics*. Springer, 1993, pp. 324–336.

[FW16] Peter I Frazier and Jialei Wang. "Bayesian optimization for materials design". In: *Information Science for Materials Discovery and Design*. Springer, 2016, pp. 45–75.

[GK17] Surbhi Goel and Adam Klivans. "Eigenvalue Decay Implies Polynomial-Time Learnability for Neural Networks". In: *Advances in Neural Information Processing Systems 30*. Ed. by I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett. Curran Associates, Inc., 2017, pp. 2192–2202. URL: `http://papers.nips.cc/paper/6814-eigenvalue-decay-implies-polynomial-time-learnability-for-neural-networks.pdf`.

[GM16] Alex Gittens and Michael W. Mahoney. "Revisiting the Nyström Method for Improved Large-scale Machine Learning". In: *J. Mach. Learn. Res.* 17.1 (Jan. 2016), pp. 3977–4041. ISSN: 1532-4435.

[Hou+06] J. Ben Hough, Manjunath Krishnapur, Yuval Peres, Bálint Virág, et al. "Determinantal processes and independence". In: *Probability surveys* 3 (2006), pp. 206–229.

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

[Mic11] Michael W. Mahoney. “Randomized algorithms for matrices and data”. In: *Foundations and Trends in Machine Learning* 3.2 (2011). Also available at: arXiv:1104.5557, pp. 123–224.

[NST19] Aleksandar Nikolov, Mohit Singh, and Uthaipon Tao Tantipongpipat. “Proportional Volume Sampling and Approximation Algorithms for A -Optimal Design”. In: *Proceedings of the Thirtieth Annual ACM-SIAM Symposium on Discrete Algorithms*. Jan. 2019, pp. 1369–1386.

[ODo+16] Brendan O’Donoghue, Eric Chu, Neal Parikh, and Stephen Boyd. “Conic optimization via operator splitting and homogeneous self-dual embedding”. In: *Journal of Optimization Theory and Applications* 169.3 (2016), pp. 1042–1068.

[Owe+16] David Owen, Andrew Melbourne, David Thomas, Enrico De Vita, Jonathan Rohrer, and Sebastien Ourselin. “Optimisation of arterial spin labelling using bayesian experimental design”. In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer. 2016, pp. 511–518.

[Puk06] Friedrich Pukelsheim. *Optimal Design of Experiments*. Philadelphia, PA, USA: Society for Industrial and Applied Mathematics, 2006. ISBN: 0898716047.

[RDP+16] Caitriona M Ryan, Christopher C Drovandi, Anthony N Pettitt, et al. “Optimal Bayesian experimental design for models with intractable likelihoods using indirect inference applied to biological process models”. In: *Bayesian Analysis* 11.3 (2016), pp. 857–883.

[RDP15] Elizabeth Ryan, Christopher Drovandi, and Anthony Pettitt. “Fully Bayesian experimental design for pharmacokinetic studies”. In: *Entropy* 17.3 (2015), pp. 1063–1089.

[SB98] Dalene K Stangl and Donald A Berry. “Bayesian statistics in medicine: Where are we and where should we be going?” In: *Sankhyā: The Indian Journal of Statistics, Series B* (1998), pp. 176–195.

[Spi+04] David J Spiegelhalter et al. “Incorporating Bayesian ideas into health-care evaluation”. In: *Statistical Science* 19.1 (2004), pp. 156–174.

[TUM12] Gabriel Terejanu, Rochan R Upadhyay, and Kenji Miki. “Bayesian experimental design for the active nitridation of graphite by atomic nitrogen”. In: *Experimental Thermal and Fluid Science* 36 (2012), pp. 178–193.

[Uen+16] Tsuyoshi Ueno, Trevor David Rhone, Zhufeng Hou, Teruyasu Mizoguchi, and Koji Tsuda. “COMBO: an efficient Bayesian optimization library for materials science”. In: *Materials discovery* 4 (2016), pp. 18–21.

[WYS17] Yining Wang, Adams W. Yu, and Aarti Singh. “On Computationally Tractable Selection of Experiments in Measurement-constrained Regression Models”. In: *J. Mach. Learn. Res.* 18.1 (Jan. 2017), pp. 5238–5278. ISSN: 1532-4435. URL: `http://dl.acm.org/citation.cfm?id=3122009.3208024`.

[Zhu+15] Rong Zhu, Ping Ma, Michael W Mahoney, and Bin Yu. “Optimal subsampling approaches for large sample linear regression”. In: *arXiv preprint arXiv:1509.05111* (2015).