# A Properties of regularized DPPs

In this section we provide proofs omitted from Sections 3 and 4. We start with showing the fact that the regularized DPP distribution $\mathrm{DPP}^p_{\mathrm{reg}}(\mathbf{X},\mathbf{A})$ is a correlation DPP.

**Lemma 7 (restated Lemma 2)** *Given $\mathbf{X}$, $\mathbf{A}$, and $\mathbf{D}_p$ as in Theorem 3, we have*

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

**Proof** First, we show this under the invertibility assumptions of Lemma 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

$$\mathbf{L} = \widetilde{\mathbf{D}} + \widetilde{\mathbf{D}}^{1/2}\mathbf{X}\mathbf{A}^{-1}\mathbf{X}^\top\widetilde{\mathbf{D}}^{1/2} \quad \text{and} \quad \widetilde{\mathbf{D}} = \mathbf{D}_p(\mathbf{I}-\mathbf{D}_p)^{-1}. \tag{6}$$

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 Bernstein (2011). 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 (6). The intuition is that when $\mathbf{A}$ or $\mathbf{I}-\mathbf{D}_p$ is non-invertible, then $\mathrm{DPP}^p_{\mathrm{reg}}(\mathbf{X},\mathbf{A})$ 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 $p_\epsilon$ 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 $\epsilon > 0$):

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

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

We now prove a lemma about combining a determinantal point process with Bernoulli sampling, which itself is a DPP with a diagonal correlation kernel.

**Lemma 8 (restated 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*

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

**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 (Kulesza and Taskar, 2012, Theorem 2.1). If $\mathbf{D}$ has zeros on the diagonal, a similar limit argument as in Lemma 2 with $\mathbf{D}_\epsilon = \mathbf{D} + \epsilon\,\mathbf{I}$ holds. ■

Next, we give a bound on the expected size of a regularized DPP.

**Lemma 9 (restated 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 $S = T \cup \{i : b_i = 1\} \sim \mathrm{DPP}_{\mathrm{reg}}^{p}(\mathbf{X}, \mathbf{A})$ be defined as in Theorem 3. Then*

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

**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 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.

**Lemma 10 (restated Lemma 5)** *Whenever $S \sim \mathrm{DPP}_{\mathrm{reg}}^{p}(\mathbf{X}, \mathbf{A})$ is a well-defined distribution it holds that*

$$\mathbb{E}\Big[\big(\mathbf{X}_S^\top\mathbf{X}_S + \mathbf{A}\big)^{-1}\Big] \preceq \Big(\sum_i p_i \mathbf{x}_i \mathbf{x}_i^\top + \mathbf{A}\Big)^{-1}, \tag{7}$$

$$\mathbb{E}\Big[\det\big(\mathbf{X}_S^\top\mathbf{X}_S + \mathbf{A}\big)^{-1}\Big] \leq \det\Big(\sum_i p_i \mathbf{x}_i \mathbf{x}_i^\top + \mathbf{A}\Big)^{-1}. \tag{8}$$

**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 $j$th row and $i$th 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 Dereziński and Mahoney (2019), 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 (7), and (8) follows even more easily:

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

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

## B Comparison of different effective dimensions

In this section we compare the two notions of effective dimension for Bayesian experimental design considered in this work. Here, we let $\mathbf{X}$ be the full $n \times d$ design matrix and use $k$ to denote the desired subset size. Recall that the effective dimension is defined as a function of the data covariance matrix $\mathbf{\Sigma}_{\mathbf{X}} = \mathbf{X}^\top\mathbf{X}$ and the prior precision matrix $\mathbf{A}$: It is given by $d_{\mathbf{A}} = \mathrm{tr}\big(\mathbf{\Sigma}_{\mathbf{X}}(\mathbf{A}+\mathbf{\Sigma}_{\mathbf{X}})^{-1}\big)$. In Dereziński and Warmuth (2018) it was suggested that $d_{\mathbf{A}}$ should also be used as the effective dimension for the experimental design problem. Theorem 2 suggests it may not reflect the true degrees of freedom of the problem because it does not scale with subset size $k$. Instead we propose to use the *scaled effective dimension* $d_{\frac{n}{k}\mathbf{A}}$. Thus, the two definitions we are comparing can be summarized as follows:

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

**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 3 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 $k \ge C \cdot d_{\frac{n}{k}\mathbf{A}}$ for some small constant $C$ (Theorem 2 has $C = 4$, but we believe this can be improved to 1). While for the identity covariance the condition $k \ge d_{\frac{n}{k}\mathbf{A}}$ is almost equivalent to $k \ge d_{\mathbf{A}}$, in the approximately low-rank case, $k \ge d_{\frac{n}{k}\mathbf{A}}$ holds for $k$ as small as 20, much less than $d_{\mathbf{A}}$.

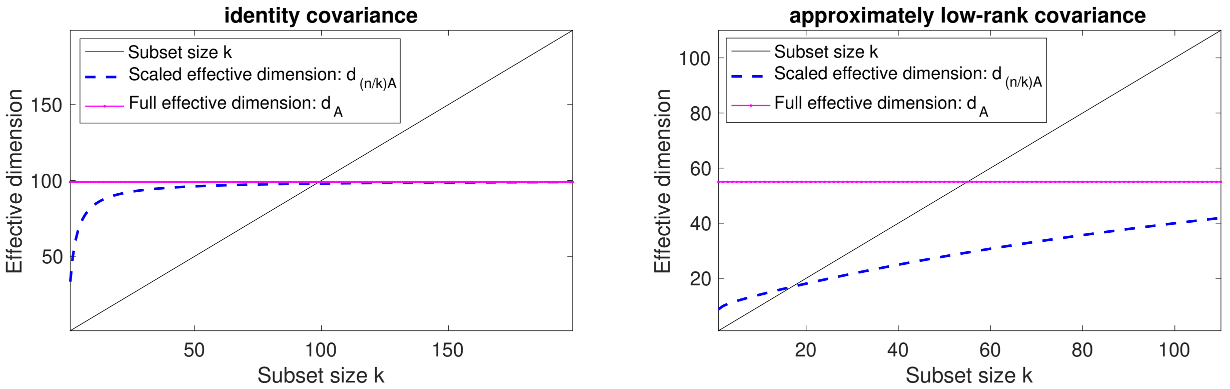

Figure 3: Scaled effective dimension compared to the full effective dimension for two diagonal data covariance matrices, with $\mathbf{A} = 10^{-2}\,\mathbf{I}$.

## C Additional details for the experiments

This section presents additional details and experimental results omitted from the main body of the paper. In addition to the `mg_scale` dataset presented in Section 5, we also benchmarked on three other data sets described in Table 2.

Table 2: Datasets used in the experiments (Chang and Lin, 2011).

| | `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 4. The general trend observed in Section 5 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 5) are also similar across the various datasets. An exception to the rule is on `mg_scale`, 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 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.

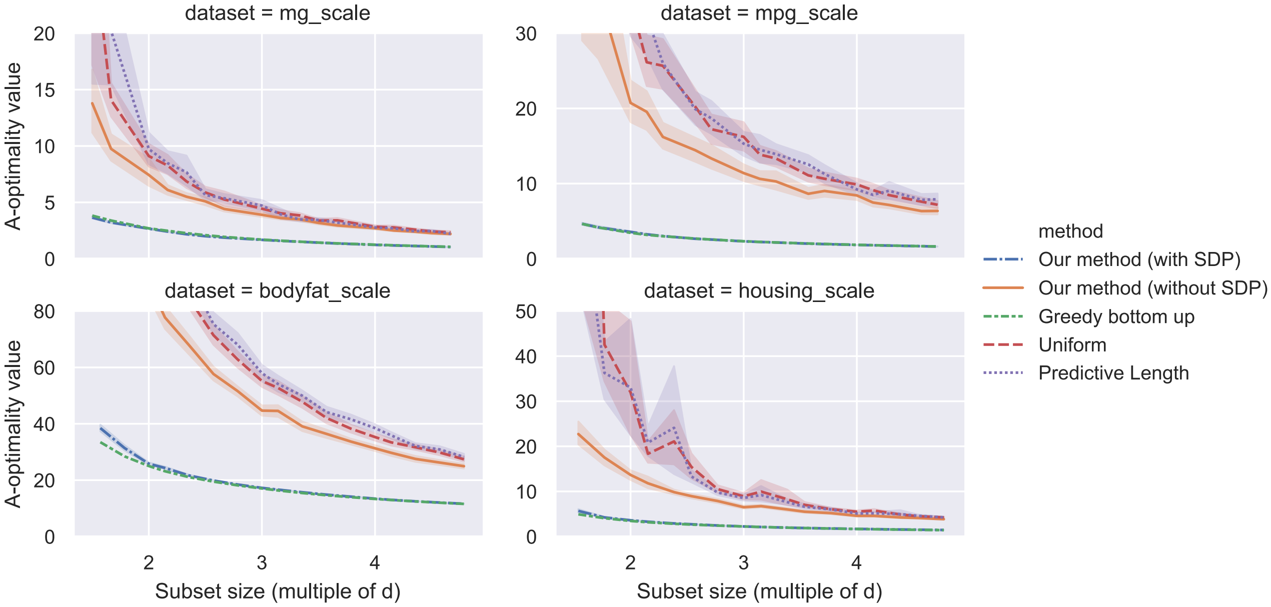

Figure 4: A-optimality values achieved by the methods compared. In all cases considered, we found our method (without SDP) to be superior to independent sampling methods like uniform and predictive length sampling. After paying the price to solve an SDP, our method (with SDP) is able to consistently match the performance of a greedy method which has been noted (Chamon and Ribeiro, 2017) to work well empirically.

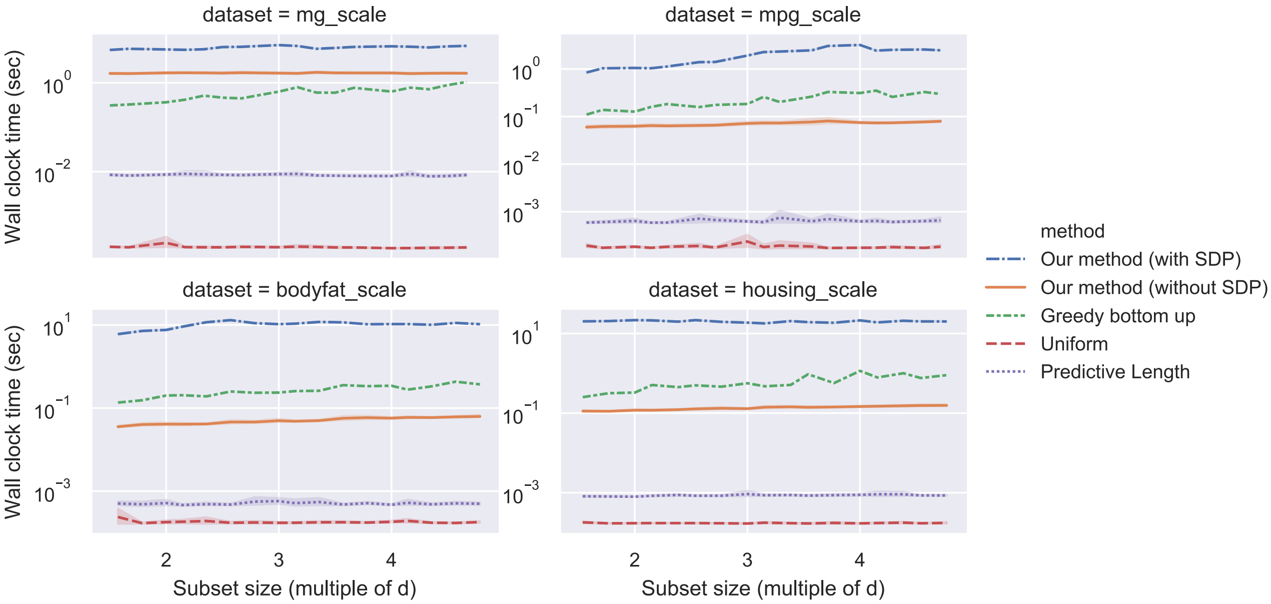

Figure 5: Runtimes of the methods compared. Our method (without SDP) is within an order of magnitude of greedy bottom up and faster in 3 out of 4 cases. The gap between our method with and without SDP is attributable to the SDP solver, making investigation of more efficient solvers and approximate solutions an interesting direction for future work.

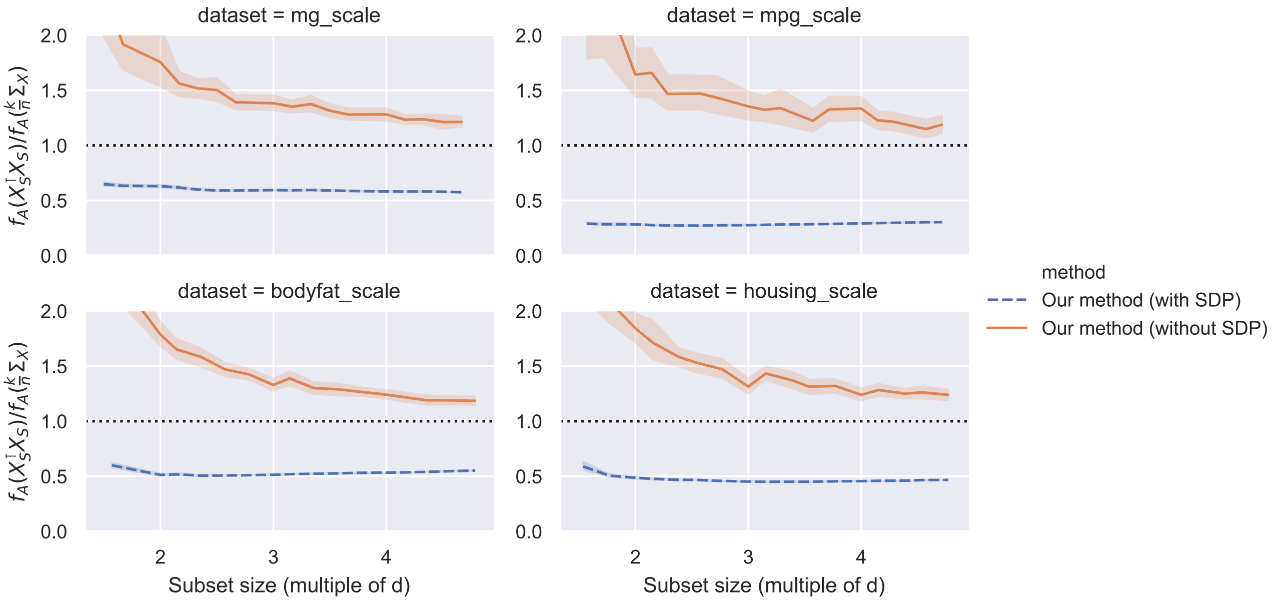

Figure 6: The ratio controlled by Lemma 6. This ratio converges to 1 as $k \to n$ and is close to 1 across all real world datasets, suggesting that $f_{\mathbf{A}}(\frac{k}{n}\mathbf{\Sigma_X})$ is an appropriate problem-dependent scale for Bayesian A-optimal experimental design.