APPENDIX: Practical Nonisotropic Monte Carlo Sampling in High Dimensions via Determinantal Point Processes

8 Experiment Details

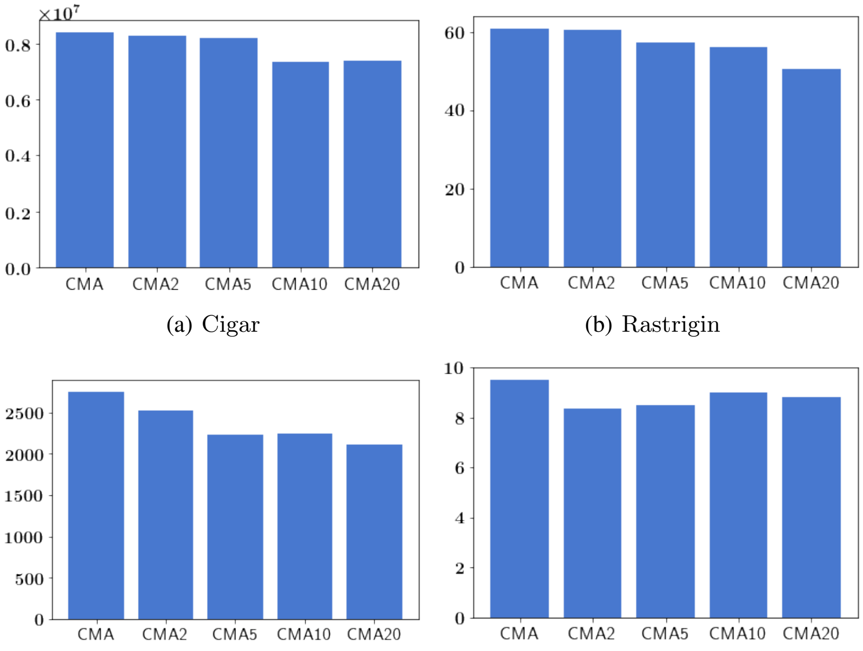
8.1 Code

Here we include some simple code to implement DPPMC using python 3.x.

```
import numpy as np
from pydpp.dpp import DPP

d = 10 # this will be the dimensionality of your problem
rho = 5 # this is a hyper-parameter
cov = np.eye(d) # this will be your nonisotropic covariance matrix
mu = np.repeat(0, d)
A = np.random.multivariate_normal(mu, cov, d * rho)

dpp = DPP(A)
dpp.compute_kernel(kernel_type = 'rbf')
idx = dpp.sample_k(d) # returning to original dimensionality, optional
A = A[idx]

# we now evaluate these samples.
```

This code is simple to include in any setting where samples are drawn from a nonisotropic distribution.

8.2 Optimal Choice of ρ



(a) Cigar

(b) Rastrigin

(c) Rosenbrock

(d) Sphere

Figure 5: Comparison of CMA-ES without DPPMC vs. with DPPMC for $\rho = 2, 5, 10, 20$.

Here we demonstrate the impact of ρ by performing an ablation study using the CMA-ES experiments. In order to measure the importance of this parameter, we test the following values: $\rho = 2, 5, 10, 20$, and measure the mean performance across three seeds after 100 function evaluations.

As we can see in Figure 5, in most cases an increase in ρ leads to a monotonic improvement in performance. This however comes at an increase in computational cost, and as such it is important to consider the trade-off between the cost of evaluating the function vs. the DPPMC algorithm when choosing an optimal ρ for a given problem. In our experiments we choose $\rho = 10$ since this value is sufficient to achieve meaningful performance gains, demonstrating the effectiveness of our approach.

8.3 Reinforcement Learning Experiments

We provide details on the reinforcement learning experiments as follows.

Benchmark Environments. Reinforcement learning tasks are identified by a state space $\mathcal{S}$ and an action space $\mathcal{A}$. The benchmark environments consist of HalfCheetah-v2 ($|\mathcal{S}| = 17, |\mathcal{A}| = 6$), Swimmer-v2 ($|\mathcal{S}| = 8, |\mathcal{A}| = 2$), Reacher-v2 ($|\mathcal{S}| = 11, |\mathcal{A}| = 2$) and Walker2d-v2 ($|\mathcal{S}| = 17, |\mathcal{A}| = 6$). Each task takes the sensory inputs of the robot as states $s_t \in \mathcal{S}$ and motor/position controls as actions $a_t \in \mathcal{A}$. All environments are simulated via OpenAI gym (Brockman et al., 2016).

Policy Architecture. We encode the policy $\pi_\theta : \mathcal{S} \mapsto \mathcal{A}$ with feed-forward network parameter θ. The architecture varies across tasks: for Swimmer-v2 and Reacher-v2, we have two hidden layers each with 16 units; for HalfCheetah-v2 and Walker2d-v2, we have two hidden layers each with 32 units. Each hidden layer is combined with a tanh non-linear function activation. The output layer does not have non-linear function activation. For each hidden layer, instead of a fully-connected structure, we adopt a low displacement rank neural network (Choromanski et al., 2018c) for a compact representation.

Implementations and Common Hyper-parameters. All ES algorithms are implemented with Numpy (Van Der Walt et al., 2011). To make our implementations parallelizable, we have made heavy reference to the Ray open source project (Moritz et al., 2018). At each iteration, the ES algorithms (including Guided ES, Trust Region ES and CMA-ES) all require sampling m perturbation directions for function evaluations. We set m to be the dimension d of the policy parameter θ. Gradient based optimizations are all carried out using Adam Optimizer (Kingma and Ba, 2014) with best learning rates chosen from $\alpha \in \{0.5, 0.1, 0.05, 0.01\}$.

DPPMC Hyper-parameters. We use a fixed RBF-kernel for all experiments: recall that a RBF-kernel takes the form $K(\mathbf{x}, \mathbf{y}) = \exp(\frac{-|\mathbf{x}-\mathbf{y}|^2}{2\sigma^2})$, we set $\sigma = 0.5$. The kernel parameter σ is manually set such that the DPPMC variants achieve good performance while the computations remain numerically stable.

Hyper-parameters for Guided ES. We follow the recipe of Guided ES (Maheswaranathan et al., 2019) to set up hyper-parameters. The DPPMC variant requires constructing a sample pool of size ρm, we choose $\rho = 10$ for our experiments. The Guided ES achieves performance gains over vanilla ES by constructing non-isotropic distribution for gradient sensing, which allows for exploring subspaces where the true gradients lie. We further improve upon Guided ES with significant gains in sample efficiency.

Hyper-parameters for Trust Region ES. We follow the recipe of Trust Region ES (Choromanski et al., 2019b) to set up hyper-parameters. Trust Region ES has two variants: (1) using ridge regression to compute update directions (Ridge); (2) using Monte-Carlo samples to estimate update directions (MC). Both variants require re-using δm samples and function evaluations from the previous iteration, here we set $\delta = 0.2$ so that the algorithm achieves $\approx 20\%$ sample gains. On top of Trust Region ES, the DPPMC variant further improves sample efficiency as demonstrated in the main paper. We refer readers to (Choromanski et al., 2019b) for a detailed description of the algorithm.

9 Variance Reduction for Evolution Strategies using DPPs

The goal of this section is to show that it is possible to use DPPs to reduce the variance of Evolution Strategies gradient estimators.

9.0.1 One dimensional variance reduction using DPPs

We start by showing an auxiliary sequence of one dimensional lemmas. We consider the problem of computing an estimator of the sum $\bar{a}$ of n real numbers $a_1, \cdots, a_n$. In Lemma 1 we first show that using DPPs it is always possible to produce an unbiased estimator of the sum of a sequence of real numbers with less or equal variance than the i.i.d estimator that samples each element a_i of the sequence i.i.d. with probability p_i. We then show in Lemma 2 that it is possible to produce a DPP kernel $\mathbf{K}$ such that the corresponding sum estimator has strictly less variance than the i.i.d. one.

We follow the discussion regarding Determinantal Point Process from Kulesza and Taskar (2012). Recall that a Determinantal Point Pricess (DPP) $\mathcal{P}$ on a ground set $\mathcal{X}$ with $|\mathcal{X}| = N$ is a probability measure over power set $2^{\mathcal{X}}$. When $\mathcal{S}$ is a random subset drawn according to $\mathcal{P}$, we have, for every $A \subset \mathcal{X}$.

$$\mathcal{P}\left(A \subset \mathcal{S}\right) = \det\left(\mathbf{K}_A\right)$$

for some real symmetric $N \times N$ matrix $\mathbf{K}$ indexed by the elements of $\mathcal{X}$. Here $\mathbf{K}_A = [\mathbf{K}_{i,j}]_{i,j \in A}$ and adopt $\det(\mathbf{K}_\emptyset) = 1$. $\mathbf{K}$ is known as the marginal kernel.

Notice that whenever $A = \{i\}$, $\mathbb{P}(i \in \mathbf{S}) = \mathbf{K}_{i,i}$ and that $\mathbb{P}(i, j \in \mathcal{S}) = \mathbb{P}(i \in \mathcal{S})\mathbb{P}(j \in \mathcal{S}) - \mathbf{K}_{i,j}^2$.

We start by showing a basic variance reduction result regarding DPPs. Let $a_1, \cdots, a_n$ be set of real numbers. Let $\bar{a}$ be their sum. We are interested in analyzing the following two estimators of $\bar{a}$:

1. $\hat{a}_{\text{i.i.d}} = \sum_{i=1}^n \frac{a_i \epsilon_i}{p_i}$ where ϵ_i are sampled independent from each other with $\epsilon_i \sim \text{Ber}(p_i)$.
2. $\hat{a}_{\text{DPP}} = \sum_{i \in \mathcal{S}} \frac{a_i \epsilon_i}{p_i}$ where $\mathcal{S}$ is a subset of $[n]$ sampled from a DPP with kernel $\mathbf{K}$ satisfying $\mathbf{K}_{i,i} = p_i$ for all i.

Notice that $\mathbb{E}\left[\hat{a}_{\text{i.i.d}}\right] = \bar{a}$ and $\mathbb{E}\left[\hat{a}_{\text{DPP}}\right] = \bar{a}$ and therefore $\hat{a}_{\text{i.i.d}}$ and $\hat{a}_{\text{DPP}}$ are unbiased estimators of $\bar{a}$.

Lemma 1. *If $a_i \geq 0$ for all i, the estimator $\hat{a}_{DPP}$ has smaller variance than $\hat{a}_{i.i.d}$ whenever $\mathbf{K}_{ii} = p_i$ for all i.*

Proof. Since $\hat{a}_{i.i.d}$ and $\hat{a}_{\text{DPP}}$ are unbiased, it is enough to compare the second moments of the said estimators.

$$\begin{aligned}
\mathbb{E}\left[\hat{a}_{\text{DPP}}^2\right] &= \mathbb{E}\left[\sum_{ij} \frac{a_i a_j \epsilon_i \epsilon_j}{p_i p_j}\right] \\
&= \sum_{i,j} \frac{\mathbb{E}\left[\epsilon_i \epsilon_j\right] a_i a_j}{p_i p_j} \\
&= \sum_{i,j} \frac{(\mathbf{K}_{ii}\mathbf{K}_{jj} - \mathbf{K}_{ij}^2) a_i a_j}{p_i p_j} \\
&= \mathbb{E}\left[\hat{a}_{i.i.d.}^2\right] - \sum_{i \neq j} \frac{\mathbf{K}_{i,j}^2 a_i a_j}{p_i p_j} \\
&\leq \mathbb{E}\left[\hat{a}_{i.i.d.}^2\right]
\end{aligned}$$

The last inequality holds whenever $a_i \geq 0$ for all i.

□

We can also show that under appropriate conditions there exists a kernel matrix $\mathbf{K}$ such that $\mathrm{Var}(\hat{a}_{\mathrm{i.i.d}}) > \mathrm{Var}(\hat{a}_{\mathrm{DPP}})$ such that the inequality is strict.

Lemma 2. *If $n \geq 3$, $p_i > 0$ for all iand there exists i such that $p_i < 1$, then there exists a matrix $\mathbf{K} \in \mathbb{R}^{n\times n}$ defining a DPP over - not necessarily nonnegative- $a_1, \cdots, a_n \in \mathbb{R}$ satisfying $\mathbf{K}_{i,i} = p_i$ and such that* $\mathrm{Var}(\hat{a}_{\mathrm{i.i.d.}}) > \mathrm{Var}(\hat{a}_{\mathrm{DPP}})$.

Proof. Let $\mathbf{K}$ be a matrix defining a DPP with $\mathbf{K}_{i,i} = p_i$ for all i Following the exact same proof as in Lemma 1, we conclude that $\mathrm{Var}(\hat{a}_{\mathrm{i,i.d}}) > \mathrm{Var}(\hat{a}_{\mathrm{DPP}})$ iff:

$$\sum_{i\neq j} \frac{\mathbf{K}_{i,j}^2 a_i a_j}{p_i p_j} > 0 \tag{12}$$

We show the existence of a kernel matrix $\mathbf{K}$ for which the inequality 12 holds and $\mathbf{K}_{i,i} = p_i$ for all i.

Indeed, let $\mathbf{K} \in \mathbb{R}^{n\times n}$ be such that:

$$\mathbf{K}_{i,j} = \begin{cases} p_i & \text{if } i = j \\ \epsilon & \text{if } \frac{a_i a_j}{p_i p_j} \geq 0 \\ 0 & \text{o.w.} \end{cases}$$

For some $\epsilon > 0$. Under this definition, notice that $\sum_{i,j} \frac{\mathbf{K}_{i,j}^2 a_i a_j}{p_i p_j} > 0$ and notice that since $0 \prec \mathrm{diag}(p_i) \prec I$, there exists a choice of $\epsilon > 0$ such that $0 \prec \mathbf{K} \prec \mathbb{I}_d$, thus defining a valid DPP kernel matrix $\mathbf{K}$.

□

9.0.2 Towards variance reduction for vector estimators using DPPs.

In this section we extend the results of the previous section to the multi dimensional case of Monte Carlo gradient estimators. We start with an auxiliary lemma that will be used in the variance reduction Theorems of the following sections. The following Lemma characterizes the maximum number of vectors that can all be pairwise negatively correlated. This Lemma will be used later on to argue the existence of a DPP kernel $\mathbf{K}$ for which its subsampling estimator of the Evolution Strategies gradient estimator achieves less variance than the i.i.d. subsampling estimator.

Lemma 3. *Let $\mathbf{v}^1, \cdots, \mathbf{v}^M \in \mathbb{R}^d$ vectors such that $\langle \mathbf{v}^i, \mathbf{v}^j \rangle < 0$ for all $i \neq j$. Then $M \leq d+1$.*

Proof. We proceed with a proof by contradiction. Let's assume $M \geq d+2$. Let $\mathbf{v}^1, \cdots, \mathbf{v}^{d+1}$ be a subset of $d+1$ vectors of $\{\mathbf{v}^j\}_{j=1}^M$. There exist $a_1, \cdots, a_{d+1} \in \mathbb{R}$ such that:

$$\sum_{i=1}^{d+1} a_i \mathbf{v}^i = 0$$

If $a_i \geq 0$ for all i then $\langle \mathbf{v}^{d+2}, \sum_i a_i \mathbf{v}^i \rangle = \sum_i a_i \langle \mathbf{v}^{d+2}, \mathbf{v}^i \rangle < 0$ which would result in a contradiction. If a_i are not all nonnegative, there exist disjoint subsets $I \subset [d+2]$ and $K \subset [d+2]$ such that $I \cup J = [d+2]$, and $I \cap J = \emptyset$ and $I, J \neq \emptyset$ and with $a_i \geq 0$ for all $i \in I$ (with at least one $a_i > 0$) and $a_j \leq 0$ (with at least one $a_j < 0$) for all $j \in J$ such that:

$$\underbrace{\sum_{i\in I} a_i \mathbf{v}^i}_{I} = \underbrace{\sum_{j \in J} -a_j \mathbf{v}^j}_{II}$$

Therefore by assumption $\langle I, II \rangle < 0$ which would cause a contradiction since $I = II$.

□

Recall the gradient estimator corresponding to Evolution Strategies. If $f : \mathbb{R}^d \to \mathbb{R}$, the ES gradient estimator $\nabla f_\sigma(\theta)$ at θ equals:

$$\nabla f_\sigma(\theta) = \mathbb{E}_{\mathbf{v}\sim\mathcal{N}(0,\mathbb{I}_d)}\left[\frac{1}{\sigma}f(\theta+\sigma\mathbf{v})\mathbf{v}\right]$$

We denote by $\hat{\nabla} f_\sigma(\theta) = \frac{1}{n\sigma}\sum_{i=1}^n f(\theta+\sigma\mathbf{v}^i)\mathbf{v}^i$ where $\mathbf{v}^i$ are all samples from a standard Gaussian $\mathcal{N}(0,\mathbb{I}_d)$.

9.0.3 Subsampling strategies in ES

In this section we consider subsampling strategies for Evolution strategies when we have a dictionary of N sensing directions $\{\mathbf{v}^i\}_{i=1}^N$. Let $\{p_i\}_{i=1}^N$ be the ensemble of probabilities with which to sample (according to a Bernoulli trial with probability p_i) each sensing i.

We recognize two cases:

1. **Unbiased sampling** In this case we consider a subsampled-importance sampling weighted version of the empirical estimator $\hat{\nabla} f_\sigma(\theta) = \frac{1}{\sigma N}\sum_{i=1}^N \mathbf{v}^i f(\theta+\sigma\mathbf{v}^i)$ of the form $\hat{\nabla}_U f_\sigma(\theta) = \frac{1}{N\sigma}\sum_{i=1}^N \frac{\epsilon_i}{p_i}\mathbf{v}^i f(\theta+\sigma\mathbf{v}^i)$.
2. **Biased** In this case we consider a subsampled version of the empirical estimator $\hat{\nabla} f_\sigma(\theta)$ of the form $\hat{\nabla}_B f_\sigma(\theta) = \frac{1}{\sigma N}\sum_{i=1}^N \frac{\epsilon_i}{w_i}\mathbf{v}^i f(\theta+\sigma\mathbf{v}^i)$ where $\{w_i\}_{i=1}^N$ is a set of importance weights, not necessarily equal to $\{p_i\}$.

The crucial observation behind these estimators is that the evaluation of f need not be performed at the points that are not subsampled. This allows us to trade off computation with variance (or mean squared error). We would like to achieve the optimal tradeoff.

Unbiased subsampling

The goal of this section is to show that for any i.i.d. subsampling strategy to build an unbiased estimator for the ES gradient, there exists a DPP kernel such that the DPP unbiased subsampling estimator achieves less variance than the i.i.d. one.

The main result of this section, Theorem 3 concerns the estimation of functions of the form $F:\mathbb{R}^d\to\mathbb{R}^m$ as defined in Section 1, and shows that for any fixed subsampling i.i.d. strategy (encoded by subsampling probabilities $\{p_i\}$), there exists a marginal kernel $\mathbf{K}$ whose corresponding estimator achieves the same mean but has (strictly) less variance. We prove Theorem 4 which specializes Theorem 3 to the case of ES gradients. A simple notational change would render the proof valid for Theorem 3.

The following corresponds to Theorem 1 in the main text.

Theorem 3. *If $N \geq d+2$ and $p_i < 1$ for all i, there exists a Marginal Kernel $\mathbf{K}\in\mathbb{R}^{N\times N}$ such that:*

$$\begin{aligned}\mathbb{E}_{\{\epsilon_i\}\sim DPP(\mathbf{K})}\left[\hat{F}(\theta)_U^{DPP}\right] &= \mathbb{E}_{\{\epsilon_i\}\sim\{Ber(p_i)\}}\left[\hat{F}(\theta)_U^{iid}\right]\\ &= \frac{1}{N}\sum_{i=1}^N h_\theta(\mathbf{v}^i)\end{aligned}$$

And:

$$\mathrm{Var}(\hat{F}(\theta)_U^{DPP}) < \mathrm{Var}(\hat{F}(\theta)_U^{iid})$$

We show the corresponding result for the case when $\mathbb{F} = \nabla f_\sigma(\theta)$. The proof is exactly the same as in the case when considering any other type of function $F:\mathbb{R}^d\to\mathbb{R}^m$ defined as in Section 1.

Let $\mathbf{K}$ be a marginal kernel matrix defining a DPP whose samples we index as $(\epsilon_1,\cdots,\epsilon_N)$ with $\epsilon_i\in\{0,1\}$ and such that the ensemble follows the DPP process. We consider the following subsampled ES estimator:

$$\hat{\nabla}_U^{DPP} f_\sigma(\theta) = \frac{1}{N\sigma}\sum_{i\in S}\frac{\epsilon_i}{p_i}f(\theta+\sigma\mathbf{v}^i)\mathbf{v}^i$$

Theorem 4. *There exists a marginal kernel $\mathbf{K}\in\mathbb{R}^{N\times N}$ such that $\widehat{\mathrm{MSE}}(\hat{\nabla}_{\mathrm{U}}^{\mathrm{DPP}} f_\sigma(\theta)) < \widehat{\mathrm{MSE}}(\hat{\nabla}_{\mathrm{U}} f_\sigma(\theta))$*

Proof. Since $\mathbb{E}\left[\hat{\nabla}^{\mathrm{DPP}}_{U} f_{\sigma}(\theta)\right] = \mathbb{E}\left[\hat{\nabla}_{\mathrm{U}} f_{\sigma}(\theta)\right]$, it is enough to show the desired statement for the square norms of these vectors.

$$
\begin{aligned}
\|\hat{\nabla}^{\mathrm{DPP}}_{\mathrm{U}} f_{\sigma}(\theta)\|^2 &= \sum_{j=1}^{d} \left(\frac{1}{\sigma N} \sum_{i \in S} \frac{\epsilon_i}{p_i} f(\theta + \sigma \mathbf{v}^i) \mathbf{v}^i(j) \right)^2 \\
&= \frac{1}{\sigma^2 N^2} \sum_{j=1}^{d} \left(\sum_{i \in S} \frac{\epsilon_i}{p_i} f(\theta + \sigma \mathbf{v}^i) \mathbf{v}^i(j) \right)^2 \\
&= \frac{1}{\sigma^2 N^2} \sum_{j=1}^{d} \left(\sum_{i,k \in S} \frac{\epsilon_i \epsilon_k}{p_i p_k} f(\theta + \sigma \mathbf{v}^i) f(\theta + \sigma \mathbf{v}^k) \mathbf{v}^i(j) \mathbf{v}^k(j) \right)
\end{aligned}
$$

Therefore:

$$
\begin{aligned}
\mathbb{E}\left[\|\hat{\nabla}^{\mathrm{DPP}}_{\mathrm{U}} f_{\sigma}(\theta)\|^2\right] &= \mathbb{E}\left[\frac{1}{\sigma^2 N^2} \sum_{j=1}^{d} \left(\sum_{i,k \in \mathcal{S}} \frac{\epsilon_i \epsilon_k}{p_i p_k} f(\theta + \sigma \mathbf{v}^i) f(\theta + \sigma \mathbf{v}^k) \mathbf{v}^i(j) \mathbf{v}^k(j) \right) \right] \\
&= \frac{1}{\sigma^2 N^2} \sum_{j=1}^{d} \left(\sum_{i,k} \frac{\mathbb{E}[\epsilon_i \epsilon_k]}{p_i p_k} f(\theta + \sigma \mathbf{v}^i) f(\theta + \sigma \mathbf{v}^k) \mathbf{v}^i(j) \mathbf{v}^k(j) \right) \\
&= \frac{1}{\sigma^2 N^2} \sum_{j=1}^{d} \left(\sum_{i \neq k} \frac{\mathbf{K}_{i,i}\mathbf{K}_{k,k} - \mathbf{K}_{i,k}^2}{p_i p_k} f(\theta + \sigma \mathbf{v}^i) f(\theta + \sigma \mathbf{v}^k) \mathbf{v}^i(j) \mathbf{v}^k(j) \right) + \\
&\quad \frac{1}{\sigma^2 N^2} \sum_{j=1}^{d} \left(\sum_{i=1}^{N} \frac{\mathbf{K}_{i,i}}{p_i^2} f^2(\theta + \sigma \mathbf{v}^i) \left(\mathbf{v}^i\right)^2 (j) \right)
\end{aligned}
$$

Let $K_{i,i} = p_i$ for all i. The expression above becomes:

$$
\begin{aligned}
\mathbb{E}\left[\|\hat{\nabla}^{\mathrm{DPP}}_{\mathrm{U}} f_{\sigma}(\theta)\|^2\right] &= \mathbb{E}\left[\|\hat{\nabla}_{\mathrm{U}} f_{\sigma}(\theta)\|^2\right] - \frac{1}{\sigma^2 N^2} \sum_{j=1}^{d} \left(\sum_{i \neq k} \frac{\mathbf{K}_{i,k}^2}{p_i p_k} f(\theta + \sigma \mathbf{v}_i) f(\theta + \sigma \mathbf{v}_k) \mathbf{v}_i(j) \mathbf{v}_k(j) \right) \\
&= \mathbb{E}\left[\|\hat{\nabla}_{\mathrm{U}} f_{\sigma}(\theta)\|^2\right] - \frac{1}{\sigma^2 N^2} \sum_{i \neq k} \frac{\mathbf{K}_{i,k}^2}{p_i p_k} f(\theta + \sigma \mathbf{v}^i) f(\theta + \sigma \mathbf{v}^j) \left(\sum_{j} \mathbf{v}^i(j) \mathbf{v}^k(j) \right) \\
&= \mathbb{E}\left[\|\hat{\nabla}_{\mathrm{U}} f_{\sigma}(\theta)\|^2\right] - \underbrace{\frac{1}{\sigma^2 N^2} \sum_{i \neq k} \frac{\mathbf{K}_{i,k}^2}{p_i p_k} f(\theta + \sigma \mathbf{v}^i) f(\theta + \sigma \mathbf{v}^j) \langle \mathbf{v}^i, \mathbf{v}^k \rangle}_{I}
\end{aligned}
$$

Let $\mathbf{V} \in \mathbb{R}^{d \times N}$ where the $i-$th column of $\mathbf{V}$ equals $\mathbf{v}^i$, and let $\mathbf{D} \in \mathbb{R}^{N \times N}$ a diagonal matrix such that $\mathbf{D}_{i,i} = \frac{f(\theta + \sigma \mathbf{v}^i)}{p_i \sigma N}$. Let $\mathbf{K}^0 \in \mathbb{R}^{N \times N}$ be a matrix having zero diagonal entries and such that $\mathbf{K}^0_{i,j} = \mathbf{K}_{i,j}$ with $i \neq j$. Similarly to the proof of Lemma 2, let's focus on term I.

$$
\begin{aligned}
\frac{1}{\sigma^2 N^2} \sum_{i \neq k} \frac{\mathbf{K}_{i,k}^2}{p_i p_k} f(\theta + \sigma \mathbf{v}^i) f(\theta + \sigma \mathbf{v}^j) \langle \mathbf{v}^i, \mathbf{v}^k \rangle &= \sum_{i \neq k} \frac{\mathbf{K}_{i,k}^2}{\sigma^2 N^2 p_i p_k} f(\theta + \sigma \mathbf{v}^i) f(\theta + \sigma \mathbf{v}^j) \langle \mathbf{v}^i, \mathbf{v}^k \rangle \\
&= \langle \left(\mathbf{K}^0\right)^2, \mathbf{D}^\top \mathbf{V}^\top \mathbf{D} \mathbf{V} \rangle
\end{aligned}
$$

We denote by $\left(\mathbf{K}^0\right)^2$ be the matrix $\mathbf{K}^0$ with entries squared. Where $\langle \left(\mathbf{K}^0\right)^2, \mathbf{D}^\top \mathbf{V}^\top \mathbf{D}\mathbf{V}\rangle = \text{trace}(\left(\mathbf{K}^0\right)^2 \mathbf{D}^\top \mathbf{V}^\top \mathbf{D}\mathbf{V})$. Define $\mathbf{K}^0$ in this way, for $i \neq j$. Let $\epsilon > 0$:

$$(\mathbf{K}^0)_{i,j} = \begin{cases} \epsilon & \text{if } f(\theta + \sigma \mathbf{v}^i) f(\theta + \sigma \mathbf{v}^j) \langle \mathbf{v}^i, \mathbf{v}^k \rangle > 0 \\ 0 & \text{o.w.} \end{cases}$$

Let $\mathbf{VD}$ be the matrix with columns equal to $\mathbf{v}^i f(\theta + \sigma \mathbf{v}^i)$ and define $\mathbf{W} = \mathbf{VD}$. Consider $\mathbf{J} = \mathbf{W}^\top \mathbf{W}$ and define $\mathbf{J}^0$ be the matrix $\mathbf{J}$ without its diagonal entries. Since $N \geq d + 2$, Lemma 3 there must be at least two positive non diagonal entries of J and therefore in this case $\langle \left(\mathbf{K}^0\right)^2, \mathbf{D}^\top \mathbf{V}^\top \mathbf{D}\mathbf{V}\rangle > 0$.

If $\mathbf{K}_{i,i} = p_i < 1$ for all i then following an argument similar to the proof of 2, we conclude there exists $\epsilon > 0$ such that $0 \prec \mathbf{K} \prec \mathbb{I}_d$ such that $\widehat{\text{MSE}}(\hat{\nabla}^{\text{DPP}}_{\text{U}} f_\sigma(\theta)) < \widehat{\text{MSE}}(\hat{\nabla}_{\text{U}} f_\sigma(\theta))$ as desired.

□

Theorem 3, yields the following corollary (corresponding to Corollary 1 in the main text). Under i.i.d. uniform sampling ($p_i = p$ for all i):

Corollary 2. *Let $\mathbf{v}^1, \cdots, \mathbf{v}^N \sim \mathcal{N}(0, I_d)$ be normally distributed sensings sampled i.i.d. Let $\hat{\nabla}_U f_\sigma(\theta)$ and $\hat{\nabla}^{DPP}_U f_\sigma(\theta)$ be subsampled gradients with $p_i = p < 1$ for all i where $\hat{\nabla}^{DPP}_U f_\sigma(\theta)$ is produced with a kernel as in Theorem 1. The following hold:*

$$\mathbb{E}\left[\hat{\nabla}^{DPP}_U f_\sigma(\theta)\right] = \mathbb{E}\left[\hat{\nabla}_U f_\sigma(\theta)\right] = \nabla f_\sigma(\theta)$$

And:

$$\widehat{\text{MSE}}(\hat{\nabla}^{DPP}_U f_\sigma(\theta)) < \widehat{\text{MSE}}(\hat{\nabla}_U f_\sigma(\theta))$$

This corollary implies that picking the right Kernel, subsampling perturbations from a DPP process when these perturbations are all i.i.d. Gaussian vectors, yields an unbiased estimator of the smoothed gradient $\nabla f_\sigma(\theta)$ with less variance (in this case equal to the mean squared error) than a naive subsampled gradient estimator that subsamples the $\{\mathbf{v}^i\}$ perturbations each with probability p.

Biased subsampling

The goal of this section is to show that for any i.i.d. subsampling strategy to build a biased estimator for the ES gradient, there exists a DPP kernel such that the DPP unbiased subsampling estimator achieves less mean squared error (MSE) than the i.i.d. one.

Define the biased downsampled estimator as:

$$\hat{F}(\theta)_B = \frac{1}{N}\sum_{i=1}^N \frac{\epsilon_i}{w_i} h_\theta(\mathbf{v}^i). \tag{13}$$

Theorem 1 and Corollary 1 of the main text can be generalized to the case of biased estimators. Borrowing notation from the previous section, and assuming access to an ensemble $\{w_i\}$ of importance weights, we get as a biased equivalent version of Theorem 1:

Theorem 5. *If $N \geq d + 2$ and $p_i < 1$ there exists a Marginal Kernel $\mathbf{K} \in \mathbb{R}^{N \times N}$ such that:*

$$\mathbb{E}_{\{\epsilon_i\} \sim DPP(\mathbf{K})}\left[\hat{F}(\theta)^{\text{DPP}}_B\right] = \mathbb{E}_{\{\epsilon_i\} \sim \{Ber(p_i)\}}\left[\hat{F}(\theta)^{\text{iid}}_B\right],$$

and furthermore $\text{MSE}(\hat{F}(\theta)^{\text{DPP}}_B) \leq \text{MSE}(\hat{F}(\theta)^{\text{iid}}_B)$, where the comparison mean equals $\mu = \hat{F}(\theta) = \frac{1}{N}\sum_{i=1}^N h_\theta(\mathbf{v}^i)$.

Proof. The following equalities hold:

$$\begin{aligned}\mathrm{MSE}(\hat{F}(\theta)_B^{\mathrm{DPP}}) &= \mathrm{Var}(\hat{F}(\theta)_B^{\mathrm{DPP}}) + \\ &\left\|\mathbb{E}\left[\hat{F}(\theta)_B^{\mathrm{DPP}} - \hat{F}(\theta)\right]\right\|^2 \\ \mathrm{MSE}(\hat{F}(\theta)_B^{\mathrm{iid}}) &= \mathrm{Var}(\hat{F}(\theta)_B^{\mathrm{iid}}) + \\ &\left\|\mathbb{E}\left[\hat{F}(\theta)_B^{\mathrm{iid}} - \hat{F}(\theta)\right]\right\|^2\end{aligned}$$

Since the expectations of $\hat{F}(\theta)_B^{\mathrm{iid}}$ and $\hat{F}(\theta)_B^{\mathrm{DPP}}$ agree, and as a consequence of Theorem 1, we can produce a kernel $\mathbf{K}$ such that:

$$\mathrm{Var}(\hat{F}(\theta)_B^{\mathrm{DPP}}) < \mathrm{Var}(\hat{F}(\theta)_B^{\mathrm{iid}}),$$

the result follows. □

As a consequence of Theorem 5, the biased downsampled versions $\hat{\nabla}_B^{\mathrm{iid}} f_\sigma(\theta)$ and $\hat{\nabla}_B^{\mathrm{DPP}} f_\sigma(\theta)$ of the ES gradient estimator $\nabla f_\sigma(\theta)$ satisfy an analogous version of Corollary 1 where Var is substituted by MSE.

The proofs of Theorem 3 and 5 can be used to produce an algorithm to find kernel matrix $\mathbf{K}$ reducing MSE. The results of the previous section can be extended to the case of biased sampling estimators. These result from the case when the importance weights are different from p_i.

Similarly to the previous section, the following theorem holds. Defining $\hat{\nabla}_{\mathrm{B}} f_\sigma(\theta)$ and $\hat{\nabla}_{\mathrm{B}}^{\mathrm{DPP}} f_\sigma(\theta)$ as:

1. $\hat{\nabla}_{\mathrm{B}} f_\sigma(\theta) = \frac{1}{\sigma N}\sum_{i=1}^N \frac{\epsilon_i}{w_i}\mathbf{v}^i f(\theta + \sigma\mathbf{v}^i)$ where $\{w_i\}_{i=1}^N$ is a set of importance weights and $\epsilon_i \sim \mathrm{Ber}(p_i)$ for some probabilities ensemble $\{p_i\}$

2. $\hat{\nabla}_{\mathrm{B}}^{\mathrm{DPP}} f_\sigma(\theta) = \frac{1}{N\sigma}\sum_{i \in \mathcal{S}} \frac{\epsilon}{w_i} f(\theta + \sigma\mathbf{v}^i)\mathbf{v}^i$.

In this case, the corresponding version of Theorem 3 is:

Theorem 6. *There exists a marginal kernel* $\mathbf{K} \in \mathbb{R}^{N\times N}$ *such that* $\widehat{\mathrm{MSE}}(\hat{\nabla}_{\mathrm{B}}^{\mathrm{DPP}} f_\sigma(\theta)) < \widehat{\mathrm{MSE}}(\hat{\nabla}_{\mathrm{B}} f_\sigma(\theta))$.

Proof. The mean squared errors $\widehat{\mathrm{MSE}}(\hat{\nabla}_{\mathrm{B}}^{\mathrm{DPP}} f_\sigma(\theta))$ and $\widehat{\mathrm{MSE}}(\hat{\nabla}_{\mathrm{B}} f_\sigma(\theta))$ can be written as:

$$\widehat{\mathrm{MSE}}(\hat{\nabla}_{\mathrm{B}}^{\mathrm{DPP}} f_\sigma(\theta)) = \mathrm{Var}(\hat{\nabla}_{\mathrm{B}}^{\mathrm{DPP}} f_\sigma(\theta)) + \underbrace{\left\|\mathbb{E}\left[\hat{\nabla}_{\mathrm{B}}^{\mathrm{DPP}} f_\sigma(\theta))\right] - \nabla f_\sigma(\theta)\right\|^2}_{I}$$

$$\widehat{\mathrm{MSE}}(\hat{\nabla}_{\mathrm{B}} f_\sigma(\theta)) = \mathrm{Var}(\hat{\nabla}_{\mathrm{B}} f_\sigma(\theta)) + \underbrace{\left\|\mathbb{E}\left[\hat{\nabla}_{\mathrm{B}} f_\sigma(\theta))\right] - \nabla f_\sigma(\theta)\right\|^2}_{II}$$

The bias terms I and II are always equal since $\mathbb{E}\left[\hat{\nabla}_{\mathrm{B}} f_\sigma(\theta))\right] = \mathbb{E}\left[\hat{\nabla}_{\mathrm{B}}^{\mathrm{DPP}} f_\sigma(\theta))\right]$.

The remainder of the proof is exactly the same as in Theorem 1.

□

9.1 DPP Connections with orthogonality

In this section we flesh out some connections between structured sampling via DPPs and structured sampling via orthogonal directions such as in Rowland et al. (2018). We show that in some way DPP structured sampling subsumes orthogonal sampling. We start showing Lemma 4, leading to Theorem 7, (Theorem 2 in the main text).

In what follows assume $\mathcal{X} = \{\mathbf{x}^1, \cdots, \mathbf{x}^N\}$ with $\mathbf{x}^i \in \mathbb{R}^d$ and let $\phi : \mathbb{R}^d \to \mathbb{R}^D$ be a possibly infinite feature map ϕ.

Lemma 4. *Let $\mathbf{W} \in \mathbb{R}^{N\times N}$ such that $\mathbf{W}_{i,j} = \langle \phi(\mathbf{x}^i), \phi(\mathbf{x}^j)\rangle$ for some a $D-$dimensional feature map ϕ. Let $A \subseteq [N]$. The nonzero eigenvalues of the principal minor W_A equal the nonzero eigenvalues of $\sum_{i\in A}\phi(\mathbf{x}^i)\phi^\top(\mathbf{x}^i)$.*

Proof. Let $A = \{i_1, \cdots, i_{|A|}\}$ and define $B_A = \left[\phi(\mathbf{x}^{i_1})\cdots\phi(\mathbf{x}^{|A|})\right] \in \mathbb{R}^{D\times|A|}$. It follows immediately that:

$$\mathbf{W}_A = \mathbf{B}_A^\top \mathbf{B}_A$$

Assume the SVD decomposition of $\mathbf{B}_A = \mathbf{U}_A^\top \mathbf{D}_A \mathbf{V}_A$ with $\mathbf{U}_A \in \mathbb{R}^{D\times D}$, $\mathbf{D}_A \in \mathbb{R}^{D\times|A|}$, and $\mathbf{V}_A \in \mathbb{R}^{|A|\times|A|}$. And thus:

$$\mathbf{W}_A = \mathbf{V}_A^\top \underbrace{\mathbf{D}_A\mathbf{D}_A^\top}_{I} \mathbf{V}_A$$

Observe that:

$$\sum_{i\in A}\phi(\mathbf{x}^i)\phi^\top(\mathbf{x}^i) = \mathbf{B}_A\mathbf{B}_A^\top$$

And substituting the SVD decomposition of B_A yields:

$$\sum_{i\in A}\phi(\mathbf{x}^i)\phi^\top(\mathbf{x}^i) = \mathbf{U}_A^\top \underbrace{\mathbf{D}_A^\top\mathbf{D}_A}_{II} \mathbf{U}_A$$

Since the nonzero entries of I and II are the same, we conclude the nonzero eigenvalues of $\mathbf{W}_A$ and of $\sum_{i\in A}\phi(\mathbf{x}^i)\phi^\top(\mathbf{x}^i)$ coincide. □

We now show a relationship between orthogonality and DPPs.

Theorem 7. *Let $\mathbf{L} \in \mathbb{R}^{N\times N}$ be an $\mathbf{L}-$ensemble such that $\mathbf{L}_{i,j} = \langle \phi(\mathbf{x}^i), \phi(\mathbf{x}^j)\rangle$, where $\|\Phi(\mathbf{x}^i)\| = 1$ for all $i \in [N]$. Let $k \in \mathbb{N}$ with $k \leq N$ and assume there exist k samples $\mathbf{x}^{i_1}, \cdots, \mathbf{x}^{i_k}$ in $\mathcal{X}$ satisfying $\langle \phi(\mathbf{x}^{i_j}), \phi(\mathbf{x}^{i_l})\rangle = 0$ for all $j, l \in [k]$. If $\mathbb{P}_k$ denotes the DPP measure over subsets of size k of $[N]$ defined by $\mathbf{L}$, the most likely outcomes from $\mathbb{P}_k$ are the size k pairwise orthogonal subsets of $\mathcal{X}$.*

Proof. Recall that $\mathbb{P}_k \propto \det(\mathbf{L}_A)$. Observe also that, since all eigenvalues of $\mathbf{L}_A$ are nonnegative, if we assume the determinant of $\mathbf{L}_A$ to be nonnegative, by the arithmetic-geometric inequality:

$$\left(\det(\mathbf{L}_A)\right)^{1/k} \leq \frac{\mathrm{tr}(\mathbf{L}_A)}{k} = 1 \tag{14}$$

Since the determinant equals the product of the eigenvalues while the trace is the sum. Equality holds iff all of the eigenvalues are equal to 1. Let A be a subset of size k such that all points are pairwise orthogonal after the map Φ, then $\det(\mathbf{L}_A = 1$. Furthermore, if $\det(\mathbf{L}_A) = 1$, then the set of points $\{\phi(\mathbf{x}^i)\}_{i\in A}$ must be orthogonal.

As a consequence of inequality 14, the equality $\det(\mathbf{L}_A) = t^k$ can only hold if all eigenvalues of $\mathbf{L}_A$ equal 1. We show this implies all the vectors must be orthogonal.

Let $A = \{i_1, \cdots, i_{|A|}\}$ and write $L_A^{(t)} = (\mathbf{B}_A)^\top \mathbf{B}_A$ where $B_A = \left[\phi(\mathbf{x}^{i_1})\cdots\phi(\mathbf{x}^{i_{|A|}})\right]$. As a consequence of Lemma 4, the nonzero eigenvalues of $L_A^{(t)}$agree with the nonzero eigenvalues of $\Sigma = \sum_{i\in A}\phi(\mathbf{x}^i)\phi^\top(\mathbf{x}^i)$.

Since by assumption $\|\Sigma\| = t$, and $\|\phi(x_i)\| = 1$ for all i:

$$\phi^\top(\mathbf{x}^i)\Sigma\phi(\mathbf{x}^i) \leq 1$$

Expanding this equation by substituting the value of Σ, we get: $\phi^\top(\mathbf{x}^i)\Sigma\phi(\mathbf{x}^i) = \sum_{j\in A}\langle\Phi(x_j), \Phi(x_i)\rangle^2 \leq 1$

Since the term corresponding to $j = i$ already equals 1, the remaining terms must be zero. This finishes the proof.

This result implies that the subsets of points of size k with the largest mass are those corresponding to pairwise orthogonal ensembles. This finishes the proof.

$\square$