Nyström Approximation for Large-Scale Determinantal Processes

Raja Hafiz Affandi
University of Pennsylvania
rajara@wharton.upenn.edu

Alex Kulesza
University of Michigan
kulesza@umich.edu

Emily B. Fox
University of Washington
ebfox@stat.washington.edu

Ben Taskar
University of Washington
taskar@cs.washington.edu

Abstract

Determinantal point processes (DPPs) are appealing models for subset selection problems where diversity is desired. They offer surprisingly efficient inference, including sampling in $O(N^3)$ time and $O(N^2)$ space, where N is the number of base items. However, in some applications, N may grow so large that sampling from a DPP becomes computationally infeasible. This is especially true in settings where the DPP kernel matrix cannot be represented by a linear decomposition of low-dimensional feature vectors. In these cases, we propose applying the Nyström approximation to project the kernel matrix into a low-dimensional space. While theoretical guarantees for the Nyström approximation in terms of standard matrix norms have been previously established, we are concerned with probabilistic measures, like total variation distance between the DPP and its Nyström approximation, that behave quite differently. In this paper we derive new error bounds for the Nyström-approximated DPP and present empirical results to corroborate them. We then demonstrate the Nyström-approximated DPP by applying it to a motion capture summarization task.

1 Introduction

A determinantal point process (DPP) is a probabilistic model that can be used to define a distribution over subsets of a base set $\mathcal{Y} = \{1, \ldots, N\}$. A critical characteristic of the DPP is that it encourages *diversity*: a random subset sampled from a DPP is likely to contain dissimilar items, where similarity is measured by a kernel matrix L that parametrizes the process. The associated sampling algorithm is exact and efficient; it uses an eigendecomposition of the DPP kernel matrix L and runs in time $O(N^3)$ despite sampling from a distribution over 2^N subsets (Hough et al., 2006).

However, when N is very large, an $O(N^3)$ algorithm can be prohibitively slow; for instance, when selecting a subset of frames to summarize a long video. Furthermore, while storing a vector of N items might be feasible, storing an $N \times N$ matrix often is not.

Kulesza and Taskar (2010) offer a solution to this problem when the kernel matrix can be decomposed as $L = B^\top B$, where B is a $D \times N$ matrix and $D \ll N$. In these cases a *dual representation* can be used to perform sampling in $O(D^3)$ time without ever constructing L. If D is finite but large, the complexity of the algorithm can be further reduced by randomly projecting B into a lower-dimensional space. Gillenwater et al. (2012) showed how such random projections yield an approximate model with bounded variational error.

However, linear decomposition of the kernel matrix using low-dimensional (or even finite-dimensional) features may not be possible. Even a simple Gaussian kernel has an infinite-dimensional feature space, and for many applications, including video summarization, the kernel can be even more complex and nonlinear.

Here we address these computational issues by applying the Nyström method to approximate a DPP kernel matrix L as a low rank matrix $\tilde{L}$. This approximation is based on a subset of the N items called *landmarks*; a small number of landmarks can often be sufficient to reproduce the bulk of the kernel's eigenspectrum.

The performance of adaptive Nyström methods has been well documented both empirically and theoretically. However, there are significant challenges in extending these results to the DPP. Most existing theoretical results bound the Frobenius or spectral norm of the kernel error matrix, but we show that these quantities are insufficient to give useful bounds on distributional measures like variational distance. Instead, we derive novel bounds for the Nyström approximation that are specifically tailored to DPPs, nontrivially

Appearing in Proceedings of the 16^{th} International Conference on Artificial Intelligence and Statistics (AISTATS) 2013, Scottsdale, AZ, USA. Volume 31 of JMLR: W&CP 31. Copyright 2013 by the authors.

Algorithm 1 DPP-Sample(L)

```
Input: kernel matrix $L$
$\{(\boldsymbol{v}_n, \lambda_n)\}_{n=1}^N \leftarrow$ eigendecomposition of $L$
$J \leftarrow \emptyset$
for $n = 1, \ldots, N$ do
    $J \leftarrow J \cup \{n\}$ with prob. $\frac{\lambda_n}{\lambda_n + 1}$
$V \leftarrow \{\boldsymbol{v}_n\}_{n \in J}$
$Y \leftarrow \emptyset$
while $|V| > 0$ do
    Select $i$ from $\mathcal{Y}$ with $\Pr(i) = \frac{1}{|V|} \sum_{\boldsymbol{v} \in V} (\boldsymbol{v}^\top e_i)^2$
    $Y \leftarrow Y \cup \{i\}$
    $V \leftarrow V_{\perp e_i}$, an orthonormal basis for the subspace
    of V orthogonal to $e_i$
Output: $Y$
```

Algorithm 2 Dual-DPP-Sample(B)

```
Input: $B$ such that $L = B^\top B$.
$\{(\hat{\boldsymbol{v}}_n, \lambda_n)\}_{n=1}^N \leftarrow$ eigendecomposition of $C = BB^\top$
$J \leftarrow \emptyset$
for $n = 1, \ldots, N$ do
    $J \leftarrow J \cup \{n\}$ with prob. $\frac{\lambda_n}{\lambda_n + 1}$
$\hat{V} \leftarrow \left\{ \frac{\hat{\boldsymbol{v}}_n}{\sqrt{\hat{\boldsymbol{v}}^\top C \hat{\boldsymbol{v}}}} \right\}_{n \in J}$
$Y \leftarrow \emptyset$
while $|\hat{V}| > 0$ do
    Select $i$ from $\mathcal{Y}$ with $\Pr(i) = \frac{1}{|\hat{V}|} \sum_{\hat{\boldsymbol{v}} \in \hat{V}} (\hat{\boldsymbol{v}}^\top B_i)^2$
    $Y \leftarrow Y \cup \{i\}$
    Let $\hat{\boldsymbol{v}}_0$ be a vector in $\hat{V}$ with $B_i^\top \hat{\boldsymbol{v}}_0 \neq 0$
    Update $\hat{V} \leftarrow \left\{ \hat{\boldsymbol{v}} - \frac{\hat{\boldsymbol{v}}^\top B_i}{\hat{\boldsymbol{v}}_0^\top B_i} \hat{\boldsymbol{v}}_0 \mid \hat{\boldsymbol{v}} \in \hat{V} - \{\hat{\boldsymbol{v}}_0\} \right\}$
    Orthonormalize $\hat{V}$ w.r.t. $\langle \hat{\boldsymbol{v}}_1, \hat{\boldsymbol{v}}_2 \rangle = \hat{\boldsymbol{v}}_1^\top C \hat{\boldsymbol{v}}_2$
Output: $Y$
```

characterizing the propagation of the approximation error through the structure of the process.

Our bounds are provably tight in certain cases, and we demonstrate empirically that the bounds are informative for a wide range of real and simulated data. These experiments also show that the proposed method provides a close approximation for DPPs on large sets. Finally, we apply our techniques to select diverse and representative frames from a series of motion capture recordings. Based on a user survey, we find that the frames sampled from a Nyström-approximated DPP form better summaries than randomly chosen frames.

2 Background

In this section we review the determinantal point process (DPP) and its dual representation.We then outline existing Nyström methods and theoretical results.

2.1 Determinantal Point Processes

A random point process $\mathcal{P}$ on a discrete base set $\mathcal{Y} = \{1, \ldots, N\}$ is a probability measure on the set $2^{\mathcal{Y}}$ of all possible subsets of $\mathcal{Y}$. For a positive semidefinite $N \times N$ kernel matrix L, the DPP, $\mathcal{P}_L$, is given by

$$\mathcal{P}_L(A) = \frac{\det(L_A)}{\det(L + I)} \ , \tag{1}$$

where $L_A \equiv [L_{ij}]_{i,j \in A}$ is the submatrix of L indexed by elements in A, and I is the $N \times N$ identity matrix. We use the convention $\det(L_\emptyset) = 1$. This *L-ensemble* formulation of DPPs was first introduced by Borodin and Rains (2005). Hough et al. (2006) showed that sampling from a DPP can be done efficiently in $O(N^3)$, as described in Algorithm 1.

In applications where diverse sets of a fixed size are desired, we can consider instead the kDPP (Kulesza and Taskar, 2011), which only gives positive probability to sets of a fixed cardinality k. The L-ensemble construction of a kDPP, denoted $\mathcal{P}_L^k$, gives probabilities

$$\mathcal{P}_L^k(A) = \frac{\det(L_A)}{\sum_{|A'|=k} \det(L_{A'})} \tag{2}$$

for all sets $A \subseteq \mathcal{Y}$ with cardinality k. Kulesza and Taskar (2011) showed that kDPPs can be sampled with the same asymptotic efficiency as standard DPPs using recursive computation of elementary symmetric polynomials.

2.2 Dual Representation of DPPs

In special cases where L is a linear kernel of low dimension, Kulesza and Taskar (2010) showed that the complexity of sampling from these DPPs can be be significantly reduced. In particular, when $L = B^\top B$, with B a $D \times N$ matrix and $D \ll N$, the complexity of the sampling algorithm can be reduced to $O(D^3)$. This arises from the fact that L and the dual kernel matrix $C = BB^\top$ share the same nonzero eigenvalues, and for each eigenvector v_k of L, Bv_k is the corresponding eigenvector of C. This leads to the sampling algorithm given in Algorithm 2, which takes time $O(D^3 + ND)$ and space $O(ND)$.

2.3 Nyström Method

For many applications, including SVM-based classification, Gaussian process regression, PCA, and, in our case, sampling DPPs, fundamental algorithms require kernel matrix operations of space $O(N^2)$ and time $O(N^3)$. A common way to improve scalability is to create a low-rank approximation to the high-dimensional kernel matrix. One such technique is known as the Nyström method, which involves selecting a small number of landmarks and then using them as the basis for a low rank approximation.

Given a sample W of l landmark items corresponding to a subset of the indices of an $N \times N$ symmetric positive semidefinite matrix L, let $\overline{W}$ be the complement of W (with size $N-l$), let L_W and $L_{\overline{W}}$ denote the principal submatrices indexed by W and $\overline{W}$, respectively, and let $L_{\overline{W}W}$ denote the $(N-l)\times l$ submatrix of L with row indices from $\overline{W}$ and column indices from W. Then we can write L in block form as

$$L = \begin{pmatrix} L_W & L_{W\overline{W}} \\ L_{\overline{W}W} & L_{\overline{W}} \end{pmatrix} . \tag{3}$$

If we denote the pseudo-inverse of L_W as L_W^+, then the Nyström approximation of L using W is

$$\tilde{L} = \begin{pmatrix} L_W & L_{W\overline{W}} \\ L_{\overline{W}W} & L_{W\overline{W}}L_W^+L_{\overline{W}W} \end{pmatrix} . \tag{4}$$

Fundamental to this method is the choice of W. Various techniques have been proposed; some have theoretical guarantees, while others have only been demonstrated empirically. Williams and Seeger (2001) first proposed choosing W by uniform sampling without replacement. A variant of this approach was proposed by Frieze et al. (2004) and Drineas and Mahoney (2005), who sample W *with* replacement, and with probabilities proportional to the squared diagonal entries of L. This produces a guarantee that, with high probability,

$$\|L-\tilde{L}\|_2 \le \|L-L_r\|_2 + \epsilon\sum_{i=1}^{N} L_{ii}^2 . \tag{5}$$

where L_r is the best rank-r approximation to L.

Kumar et al. (2012) later proved that the same rate of convergence applies for uniform sampling without replacement, and argued that uniform sampling outperforms other non-adaptive methods for many real-world problems while being computationally cheaper.

2.4 Adaptive Nyström Method

Instead of sampling elements of W from a fixed distribution, Deshpande et al. (2006) introduced the idea of *adaptive* sampling, which alternates between selecting landmarks and updating the sampling distribution for the remaining items. Intuitively, items whose kernel values are poorly approximated under the existing sample are more likely to be chosen in the next round.

By sampling in each round landmarks W_t chosen according to probabilities $p_i^{(t)} \propto \|L_i - \tilde{L}_i(W_1\cup\cdots\cup W_{t-1})\|_2^2$ (where L_i denotes the ith column of L), we are guaranteed that

$$\mathbb{E}\left(\|L-\tilde{L}(W)\|_F\right) \le \frac{\|L-L_r\|_F}{1-\epsilon} + \epsilon^T\sum_{i=1}^{N} L_{ii}^2 . \tag{6}$$

where L_r is the best rank-r approximation to L and $W = W_1\cup\cdots\cup W_T$.

Algorithm 3 Nyström-based (k)DPP sampling

```
Input: Chosen landmark indices $W = \{i_1,\dots,i_l\}$
$L_{*W} \leftarrow N \times l$ matrix formed by chosen landmarks
$L_W \leftarrow$ principal submatrix of $L$ indexed by $W$
$L_W^+ \leftarrow$ pseudoinverse of $L_W$
$B = (L_{*W})^\top (L_W^+)^{1/2}$
$Y \leftarrow$ Dual-($k$)DPP-Sample($B$)
Output: $Y$
```

Kumar et al. (2012) argue that adaptive Nyström methods empirically outperform the non-adaptive versions in cases where the number of landmarks is small relative to N. In fact, their results suggest that the performance gains of adaptive Nyström methods relative to the non-adaptive schemes are inversely proportional to the percentage of items chosen as landmarks.

3 Nyström Method for DPP/kDPP

As described in Section 2.2, a DPP whose kernel matrix has a known decomposition of rank D can be sampled using the dual representation, reducing the time complexity from $O(N^3)$ to $O(D^3+ND)$ and the space complexity from $O(N^2)$ to $O(ND)$. However, in many settings such a decomposition may not be available, for example if L is generated by infinite-dimensional features. In these cases we propose applying the Nyström approximation to L, building an l-dimensional approximation and applying the dual representation to reduce sampling complexity to $O(l^3+Nl)$ time and $O(Nl)$ space (see Algorithm 3).

To the best of our knowledge, analysis of the error of the Nyström approximation has been limited to the Frobenius and spectral norms of the residual matrix $L-\tilde{L}$, and no bounds exist for volumetric measures of error which are more relevant for DPPs. The challenge here is to study how the Nyström approximation simultaneously affects *all* possible minors of L.

In fact, a small error in the matrix norm can have a large effect on the minors of the matrix:

Example 1. *Consider matrices $L = \mathrm{diag}(M,\dots,M,\epsilon)$ and $\tilde{L} = \mathrm{diag}(M,\dots,M,0)$ for some large M and small ϵ. Although $\|L-\tilde{L}\|_F = \|L-\tilde{L}\|_2 = \epsilon$, for any A that includes the final index, we have $\det(L_A) - \det(\tilde{L}_A) = \epsilon M^{k-1}$, where $k = |A|$.*

It is conceivable that while error on some subsets is large, most subsets are well approximated. Unfortunately, this not generally true.

Definition 1. *The variational distance between the DPP with kernel L and the DPP with the Nystrom-approximated kernel $\tilde{L}$ is given by*

$$\|\mathcal{P}_L - \mathcal{P}_{\tilde{L}}\|_1 = \frac{1}{2}\sum_{A\in 2^{\mathcal{Y}}} |\mathcal{P}_L(A) - \mathcal{P}_{\tilde{L}}(A)| . \tag{7}$$

The variational distance is a natural global measure of approximation that ranges from 0 to 1. Unfortunately, it is not difficult to construct a sequence of matrices where the matrix norms of $L - \tilde{L}$ tend to zero but the variational distance does not.

Example 2. *Let L be a diagonal matrix with entries $1/N$ and $\tilde{L}$ be a diagonal matrix with $N/2$ entries equal to $1/N$ and the rest equal to 0. Note that $||L - \tilde{L}||_F = 1/\sqrt{2N}$ and $||L - \tilde{L}||_2 = 1/N$, which tend to zero as $N \to \infty$. However, the variational distance is bounded away from zero. To see this, note that the normalizers are $\det(L + I) = (1 + 1/N)^N$ and $\det(\tilde{L} + I) = (1 + 1/N)^{N/2}$, which tend to e and $\sqrt{e}$, respectively. Consider all subsets which have zero mass in the approximation, $S = \{A : \det(\tilde{L}_A) = 0\}$. Summing up the unnormalized mass of sets in the complement of S, we have $\sum_{A \notin S} \det(L_A) = \det(\tilde{L}+I)$ and thus $\sum_{A \in S} \det(L_A) = \det(L + I) - \det(\tilde{L} + I)$. Now consider the contribution of just the sets in S to the variational distance:*

$$\|\mathcal{P}_L - \mathcal{P}_{\tilde{L}}\|_1 \geq \frac{1}{2} \sum_{A \in S} \left| \frac{\det(L_A)}{\det(L + I)} - 0 \right| \tag{8}$$

$$= \frac{\det(L + I) - \det(\tilde{L} + I)}{2 \det(L + I)} , \tag{9}$$

which tends to $\frac{e - \sqrt{e}}{2e} \approx 0.1967$ as $N \to \infty$.

One might still hope that pathological cases occur only for diagonal matrices, or more generally for matrices that have high coherence (Candes and Romberg, 2007). In fact, coherence has previously been used by Talwalkar and Rostamizadeh (2010) to analyze the error of the Nyström approximation. Define the coherence

$$\mu(L) = \sqrt{N} \max_{i,j} |v_{ij}| , \tag{10}$$

where each $\boldsymbol{v}_i$ is a unit-norm eigenvector of L. A diagonal matrix achieves the highest coherence of $\sqrt{N}$ and a matrix with all entries equal to a constant has the lowest coherence of 1. Suppose that $f(N)$ is a sublinear but monotone increasing function with $\lim_{N\to\infty} f(N) = \infty$. We can construct a sequence of kernels L with $\mu(L) = \sqrt{f(N)} = o(\sqrt{N})$ for which matrix norms of the Nyström approximation error tend to zero, but the variational distance tends to a constant.

Example 3. *Let L be a block diagonal matrix with $f(N)$ constant blocks, each of size $N/f(N)$, where each non-zero entry is $1/N$. Let $\tilde{L}$ be structured like L except with half of the blocks set to zero. Note that $\mu^2(L) = f(N)$ by construction and that each block contributes a single eigenvalue of $\frac{1}{f(N)}$; the Frobenius and spectral norms of $L - \tilde{L}$ thus tend to zero as N increases. The DPP normalizers are given by $\det(L + I) = (1 + 1/f(N))^{f(N)} \to e$ and $\det(\tilde{L} + I) = (1 + 1/f(N))^{f(N)/2} \to \sqrt{e}$. By a similar argument to the one for diagonal matrices, we can show that variational distance tends to $\frac{e-\sqrt{e}}{2e}$.*

Unfortunately, in the cases above, the Nyström method will yield poor approximations to the original DPPs. Convergence of the matrix norm error alone is thus generally insufficient to obtain tight bounds on the resulting approximate DPP distribution. It turns out that the gap between the eigenvalues of the kernel matrix and the spectral norm error plays a major role in the effectiveness of the Nyström approximation for DPPs, as we will show in Theorems 1 and 2. In the examples above, this gap is not large enough for a close approximation; in particular, the spectral norm errors are *equal* to the smallest non-zero eigenvalues. In the next subsection, we derive approximation bounds for DPPs that are applicable to any landmark-selection scheme within the Nyström framework.

3.1 Preliminaries

We start with a result for positive semidefinite matrices known as Weyl's inequality:

Lemma 1. *(Bhatia, 1997) Let $L = \tilde{L} + E$, where $L, \tilde{L}$ and E are all positive semidefinite $N \times N$ matrices with eigenvalues $\lambda_1 \geq \ldots \geq \lambda_N \geq 0$, $\tilde{\lambda}_1 \geq \ldots \geq \tilde{\lambda}_N \geq 0$, and $\xi_1 \geq \ldots \geq \xi_N \geq 0$, respectively. Then*

$$\lambda_n \leq \tilde{\lambda}_m + \xi_{n-m+1} \quad \text{for} \quad m \leq n , \tag{11}$$

$$\lambda_n \geq \tilde{\lambda}_m + \xi_{n-m+N} \quad \text{for} \quad m \geq n . \tag{12}$$

Going forward, we use the convention $\lambda_i = 0$ for $i > N$. Weyl's inequality gives the following two corollaries.

Corollary 1. *When $\xi_j = 0$ for $j = r + 1, \ldots, N$, then for $j = 1, \ldots, N$,*

$$\lambda_j \geq \tilde{\lambda}_j \geq \lambda_{j+r} . \tag{13}$$

Proof. For the first inequality, let $n = m = j$ in (12). For the second, let $m = j$ and $n = j + r$ in (11). □

Corollary 2. *For $j = 1, \ldots, N$,*

$$\lambda_j - \xi_N \geq \tilde{\lambda}_j \geq \lambda_j - \xi_1 . \tag{14}$$

Proof. We let $n = m = j$ in (11) and (12), then rearrange terms to get the desired result. □

The following two lemmas pertain specifically to the Nyström method.

Lemma 2. *(Arcolano, 2011) Let $\tilde{L}$ be a Nyström approximation of L. Let $E = L - \tilde{L}$ be the corresponding error matrix. Then E is positive semidefinite with* $\text{rank}(E) = \text{rank}(L) - \text{rank}(\tilde{L})$.

Lemma 3. *Denote the set of indices of the chosen landmarks in the Nyström construction as W. Then if $A \subseteq W$,*

$$\det(L_A) = \det(\tilde{L}_A) \ . \tag{15}$$

Proof. $L_W = \tilde{L}_W$ and $A \subseteq W$; the result follows. □

3.2 Set-wise bounds for DPPs

We are now ready to state set-wise bounds on the Nyström approximation error for DPPs and kDPPs. In particular, for each set $A \subseteq \mathcal{Y}$, we want to bound the probability gap $|\mathcal{P}_L(A) - \mathcal{P}_{\tilde{L}}(A)|$. Going forward, we use $\mathcal{P}_A \equiv \mathcal{P}_L(A)$ and $\tilde{\mathcal{P}}_A \equiv \mathcal{P}_{\tilde{L}}(A)$.

Once again, we denote the set of all sampled landmarks as W. We first consider the case where $A \subseteq W$. In this case, by Lemma 3, $\det(L_A) = \det(\tilde{L}_A)$. Thus the only error comes from the normalization term in Equation (1). Theorem 1 gives the desired bound.

Theorem 1. *Let $\lambda_1 \geq \ldots \geq \lambda_N$ be the eigenvalues of L. If $\tilde{L}$ has rank r, L has rank m, and*

$$\hat{\lambda}_i = \max\left\{\lambda_{i+(m-r)}, \lambda_i - \|L - \tilde{L}\|_2\right\} \ , \tag{16}$$

then for $A \subseteq W$,

$$|\mathcal{P}_A - \tilde{\mathcal{P}}_A| \leq \mathcal{P}_A \left[\frac{\prod_{i=1}^{n}(1+\lambda_i)}{\prod_{i=1}^{n}(1+\hat{\lambda}_i)} - 1\right] \ . \tag{17}$$

Proof.

$$|\mathcal{P}_A - \tilde{\mathcal{P}}_A| = \left[\frac{\det(\tilde{L}_A)}{\det(\tilde{L}+I)} - \frac{\det(L_A)}{\det(L+I)}\right] \tag{18}$$

$$= \mathcal{P}_A \left[\frac{\det(L+I)}{\det(\tilde{L}+I)} - 1\right] = \mathcal{P}_A \left[\frac{\prod_{i=1}^{n}(1+\lambda_i)}{\prod_{i=1}^{n}(1+\tilde{\lambda}_i)} - 1\right] \ ,$$

where $\tilde{\lambda}_1 \geq, \ldots, \tilde{\lambda}_N \geq 0$ represent the eigenvalues of $\tilde{L}$. The first equality follows from the fact that $\lambda_i \geq \tilde{\lambda}_i$, due to the first inequality in Corollary 1. Now note that since $L = \tilde{L} + E$, by Lemma 2, Corollary 1, and Corollary 2 we have

$$\tilde{\lambda}_i \geq \lambda_{i+(m-r)}, \quad \tilde{\lambda}_i \geq \lambda_i - \xi_1 = \lambda_i - \|L - \tilde{L}\|_2 \ . \tag{19}$$

The theorem follows. □

For $A \nsubseteq W$, we must also account for error in the numerator, since it is not generally true that $\det(L_A) = \det(\tilde{L}_A)$. Theorem 2 gives a set-wise probability bound.

Theorem 2. *Assume $\tilde{L}$ has rank r, L has rank m, $|A| = k$, and L_A has eigenvalues $\lambda_1^A \geq \ldots \geq \lambda_k^A$. Let*

$$\hat{\lambda}_i = \max\left\{\lambda_{i+(m-r)}, \lambda_i - \|L - \tilde{L}\|_2\right\} \tag{20}$$

$$\hat{\lambda}_i^A = \max\left\{\lambda_i^A - \|L - \tilde{L}\|_2, 0\right\} \ . \tag{21}$$

Then for $A \nsubseteq W$,

$$|\mathcal{P}_A - \tilde{\mathcal{P}}_A| \tag{22}$$

$$\leq \mathcal{P}_A \max\left\{\left[1 - \frac{\prod_{i=1}^{k}\hat{\lambda}_i^A}{\prod_{i=1}^{k}\lambda_i^A}\right], \left[\frac{\prod_{i=1}^{n}(1+\lambda_i)}{\prod_{i=1}^{n}(1+\hat{\lambda}_i)} - 1\right]\right\} \ .$$

Proof.

$$\tilde{\mathcal{P}}_A - \mathcal{P}_A = \mathcal{P}_A \left[\frac{\det(L+I)\det(\tilde{L}_A)}{\det(\tilde{L}+I)\det(L_A)} - 1\right] \tag{23}$$

$$= \mathcal{P}_A \left[\left(\frac{\prod_{i=1}^{n}(1+\lambda_i)}{\prod_{i=1}^{n}(1+\tilde{\lambda}_i)}\right)\left(\frac{\prod_{i=1}^{k}\tilde{\lambda}_i^A}{\prod_{i=1}^{k}\lambda_i^A}\right) - 1\right] \ .$$

Here $\tilde{\lambda}_1^A, \geq, \ldots, \geq \tilde{\lambda}_k^A$ are the eigenvalues of $\tilde{L}_A$. Now note that $L_A = \tilde{L}_A + E_A$. Since E is positive semidefinite, E_A is also positive semidefinite. Thus by Corollary 1 we have $\lambda_i^A \geq \tilde{\lambda}_i^A$, and so

$$\tilde{\mathcal{P}}_A - \mathcal{P}_A \leq \mathcal{P}_A \left[\frac{\prod_{i=1}^{n}(1+\lambda_i)}{\prod_{i=1}^{n}(1+\hat{\lambda}_i)} - 1\right] \ . \tag{24}$$

For the reverse inequality, we multiply Equation (23) by -1 and use the fact that $\lambda_i \geq \tilde{\lambda}_i$ and $\lambda_i^A \geq 0$. By Corrolary 2,

$$\tilde{\lambda}_i^A \geq \lambda_i^A - \xi_1^A \geq \lambda_i^A - \xi_1 = \lambda_i^A - \|L - \tilde{L}\|_2 \ , \tag{25}$$

resulting in the inequality

$$\mathcal{P}_A - \tilde{\mathcal{P}}_A \leq \mathcal{P}_A \left[1 - \frac{\prod_{i=1}^{k}\hat{\lambda}_i^A}{\prod_{i=1}^{k}\lambda_i^A}\right] \ . \tag{26}$$

The theorem follows by combining the two inequalities. □

Both theorems are tight if the approximation is exact ($\|L - \tilde{L}\|_2 = 0$). It can also be shown that these bounds are tight for the diagonal matrix examples discussed at the beginning of this section, where the spectral norm error is equal to the non-zero eigenvalues. Moreover, these bounds are convenient since they are expressed in terms of the spectral norm of the error matrix and therefore can be easily combined with existing approximation bounds for the Nyström method. Note that the eigenvalues of L and the size of the set A both play important roles in the bound. In fact, these two quantities are closely related; it is possible to show that the expected size of a set sampled from a DPP is

$$\mathbb{E}\left[|A|\right] = \sum_{n=1}^{N} \frac{\lambda_n}{\lambda_n + 1} \ . \tag{27}$$

Thus, if L has large eigenvalues, we expect the Nyström approximation error to be large as well since the DPP associated with L gives high probability to large sets.

3.3 Set-wise bounds for kDPPs

We can obtain similar results for Nyström-approximated kDPPs. In this case, for each set A with $|A| = k$ we want to bound the probability gap $|\mathcal{P}^k_A - \tilde{\mathcal{P}}^k_A|$. Using Equation (2) for $\mathcal{P}^k_A$, it is easy to generalize the theorems in the preceding section. The proofs for the following theorems are provided in the supplementary material.

Theorem 3. *Let e_k denote the kth elementary symmetric polynomial of L:*

$$e_k(\lambda_1, \ldots, \lambda_N) = \sum_{|J|=k} \prod_{n \in J} \lambda_n \ . \qquad (28)$$

Under the conditions of Theorem 1, for $A \subseteq W$,

$$|\mathcal{P}^k_A - \tilde{\mathcal{P}}^k_A| \leq \mathcal{P}^k_A \left[\frac{e_k(\lambda_1, \ldots, \lambda_N)}{e_k(\hat{\lambda}_1, \ldots, \hat{\lambda}_N)} - 1 \right] \ . \qquad (29)$$

Theorem 4. *Under the conditions of Theorem 2, for $A \not\subseteq W$,*

$$|\mathcal{P}^k_A - \tilde{\mathcal{P}}^k_A| \qquad (30)$$

$$\leq \mathcal{P}^k_A \max \left\{ \left[\frac{e_k(\lambda_1, \ldots, \lambda_N)}{e_k(\hat{\lambda}_1, \ldots, \hat{\lambda}_N)} - 1 \right], \left[1 - \frac{\prod_{i=1}^k \hat{\lambda}^A_i}{\prod_{i=1}^k \lambda^A_i} \right] \right\} \ .$$

Note that the scale of the eigenvalues has no effect on the kDPP; we can directly observe from Equation (2) that scaling L does not change the kDPP distribution since any constant factor appears to the kth power in both the numerator and denominator.

4 Empirical Results

In this section we present empirical results on the performance of the Nyström approximation for kDPPs using three datasets small enough for us to perform ground-truth inference in the original kDPP. Two of the datasets are derived from real-world applications available on the UCI repository[1]—the first is a linear kernel matrix constructed from 1000 MNIST images, and the second an RBF kernel matrix constructed from 1000 Abalone data points—while the third is synthetic and comprises a 1000 × 1000 diagonal kernel matrix with exponentially decaying diagonal elements. Figure 1 displays the log-eigenvalues for each dataset.

On each dataset, we perform the Nyström approximation with three different sampling schemes: stochastic adaptive, greedy adaptive, and uniform. The stochastic adaptive sampling technique is a simplified version of the scheme used in Deshpande et al. (2006), where, on each iteration of landmark selection, we update $E = L - \tilde{L}$ and then sample landmarks with probabilities proportional to E^2_{ii}. In the greedy scheme, we perform a similar update, but always choose the landmarks with the maximum diagonal value E_{ii}. Finally, for the uniform method, we simply sample the landmarks uniformly without replacement.

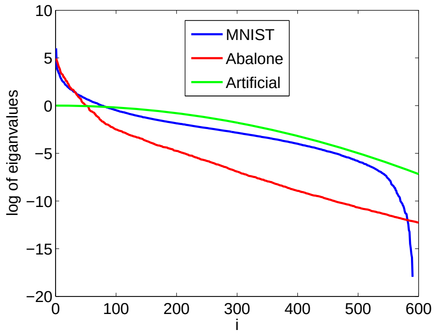


Figure 1: The first 600 log-eigenvalues for each dataset.

In Figure 2 (top), we plot $\log ||L - \tilde{L}||_2$ for each dataset as a function of the number of landmarks sampled. For the MNIST data all sampling algorithms initially perform equally well, but uniform sampling becomes relatively worse after about 550 landmarks are sampled. For the Abalone data the adaptive methods perform much better than uniform sampling over the entire range of sampled landmarks. This phenomenon is perhaps explained by the analysis of Talwalkar and Rostamizadeh (2010), which suggests that uniform sampling works well for the MNIST data due to its relatively low coherence ($\mu(L) = 0.5\sqrt{N}$), while performing poorly on the higher-coherence Abalone dataset ($\mu(L) = 0.8\sqrt{N}$). For both of the UCI datasets, the stochastic and greedy adaptive methods perform similarly. However, for our artificial dataset it is easy to see that the greedy adaptive scheme is optimal since it chooses the top remaining eigenvalues in each iteration.

In Figure 2 (bottom), we plot $\log ||P - \tilde{P}||_1$ for $k = 10$ (estimated by sampling), as well as the theoretical bounds from Section 3. The bounds track the actual variational error closely for both the MNIST and Abalone datasets. For the artificial dataset uniform sampling can do arbitrarily poorly, so we see looser bounds in this case. We note that the variational distance correlates strongly with the spectral norm error for each dataset.

4.1 Related methods

The Nystöm technique is, of course, not the only possible means of finding low-rank kernel approximations. One alternative for shift-invariant kernels is random Fourier features (RFFs), which were recently proposed by Rahimi and Recht (2007). RFFs map each item onto a random direction drawn from the Fourier transform of the kernel function; this results in a uniform approximation of the kernel matrix. In practice, however, reasonable RFF approximations seem to require

[1] http://archive.ics.uci.edu/ml/

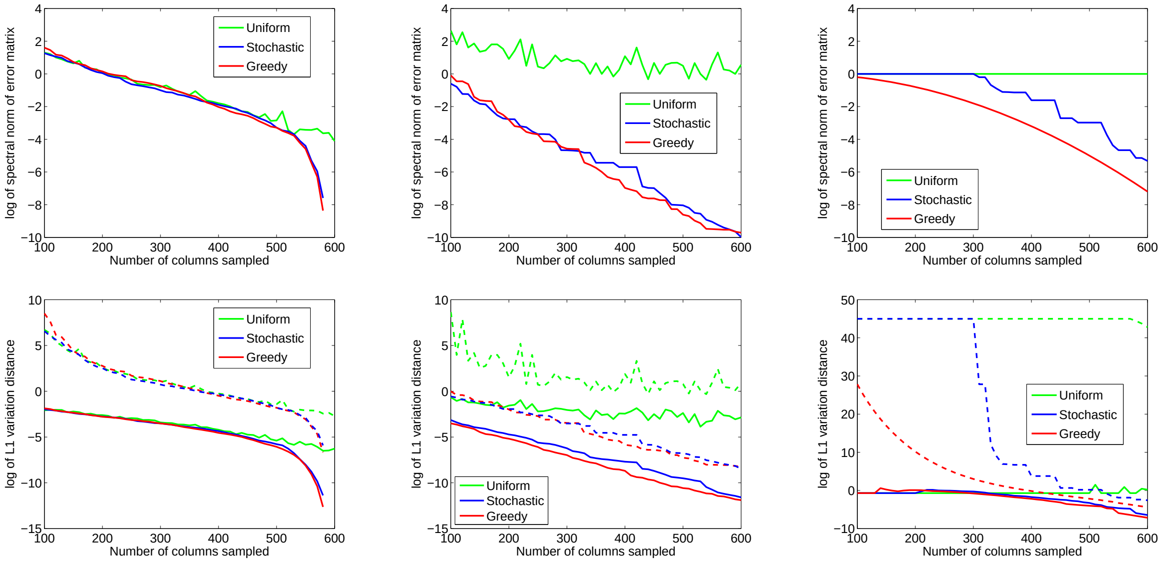


Figure 2: Error of Nyström approximations. *Top:* $\log(\|L - \tilde{L}\|_2)$ as a function of number of landmarks sampled. *Bottom:* $\log(\|\mathcal{P} - \tilde{\mathcal{P}}\|_1)$ as a function of number of landmarks sampled. The dashed lines show the bounds derived in Sec 3. From left to right, the datasets used are MNIST, Abalone and Artificial.

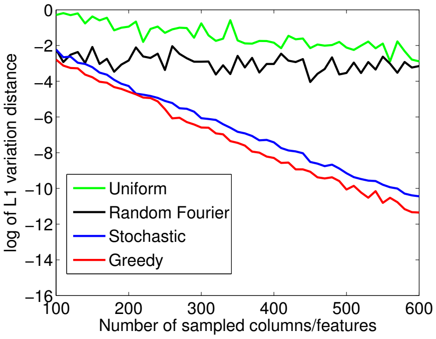


Figure 3: Error of Nyström and random Fourier features approximations on Abalone data: $\log(\|\mathcal{P} - \tilde{\mathcal{P}}\|_1)$ as a function of the number of landmarks or random features sampled.

a large number of random features, which can reduce the computational benefits of this technique.

We performed empirical comparisons between the Nyström methods and random Fourier features (RFFs) by approximating DPPs on the Abalone dataset. While RFFs generally match or outperform uniform sampling of Nyström landmarks, they result in significantly higher error compared to the adaptive versions, especially when there is high correlation between items, as shown in Figure 3. These results are consistent with those previously reported for kernel learning (Yang et al., 2012), where the Nyström method was shown to perform significantly better in the presence of large eigengaps. We provide a more detailed empirical comparison with RFFs in the supplementary material.

5 Experiments

Finally, we demonstrate the Nyström approximation on a motion summarization task that is too large to permit tractable inference in the original DPP. As input, we are given a series of motion capture recordings, each of which depicts human subjects performing motions related to a particular activity, such as dancing or playing basketball. In order to aid browsing and retrieval of these recordings in the future, we would like to choose, from each recording, a small number of frames that summarize its motions in a visually intuitive way. Since a good summary should contain a diverse set of frames, a DPP is a natural model for this task.

We obtained test recordings from the CMU motion capture database[2], which offers motion captures of over 100 subjects performing a variety of actions. Each capture involves 31 sensors attached to the subject's body and sampled 120 times per second. For each of nine activity categories—basketball, boxing, dancing, exercise, jumping, martial arts, playground, running, and soccer—we made a large input recording by concatenating all available captures in that category. On average, the resulting recordings are about $N = 24{,}000$ frames long (min 3,358; max 56,601). At this scale, storage of a full $N \times N$ DPP kernel matrix would be highly impractical (requiring up to 25GB of memory), and $O(N^3)$ SVD would be prohibitively expensive.

In order to model the summarization problem as a DPP, we designed a simple kernel to measure the similarity between pairs of poses recorded in different frames. We first computed the variance for the location of each

[2] `http://mocap.cs.cmu.edu/`

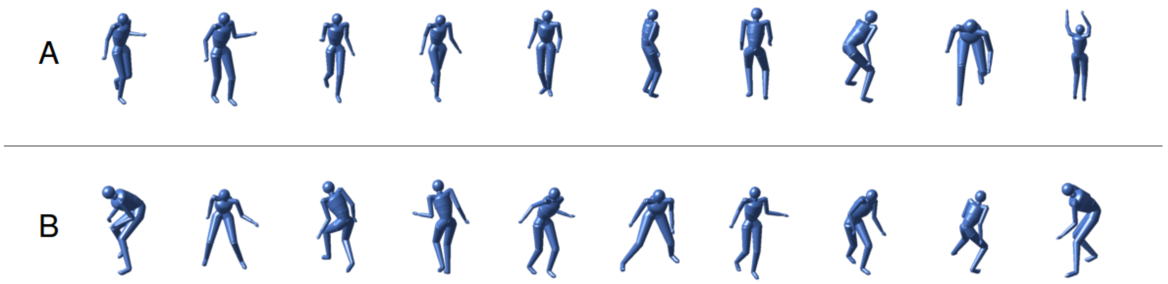


Figure 4: A sample pair of frame sets for the activity basketball. The top set is chosen randomly, while the bottom is sampled from the Nyström-approximated DPP.

sensor for each activity; this allowed us to tailor the kernel to the specific motion being summarized. For instance, we might expect a high variance for foot locations in dancing, and a relatively smaller variance in boxing. We then used these variance measurements to specify a Gaussian kernel over the position of each sensor, and finally combined the Gaussian kernels with a set of weights chosen manually to approximately reflect the importance of each sensor location to human judgments of pose similarity. Specifically, for poses $\mathcal{A} = (\boldsymbol{a}_1, \boldsymbol{a}_2, \ldots, \boldsymbol{a}_{31})$ and $\mathcal{B} = (\boldsymbol{b}_1, \boldsymbol{b}_2, \ldots, \boldsymbol{b}_{31})$, where $\boldsymbol{a}_1$ is the three dimensional location of the first sensor in pose $\mathcal{A}$, etc., the kernel value is given by

$$L(\mathcal{A}, \mathcal{B}) = \sum_{i=1}^{31} w_i \exp\left(-\frac{\|\boldsymbol{a}_i - \boldsymbol{b}_i\|_2^2}{2\sigma_i^2}\right) , \tag{31}$$

where σ_i^2 is the variance measured for sensor i, and $\boldsymbol{w} = (w_1, w_2, \ldots, w_{31})$ is the importance weight vector. We chose a weight of 1 for the head, wrists, and ankles, a weight of 0.5 for the elbows and knees, and a weight of 0 for the remaining 22 sensors.

This kind of spatial kernel is natural for this task, where the items have inherent geometric relationships. However, because the feature representation is infinite-dimensional, it does not readily admit use of the dual methods of Kulesza and Taskar (2010). Instead, we applied the stochastic adaptive Nyström approximation developed above, sampling a total of 200 landmark frames from each recording in 20 iterations (10 frames per iteration), bringing the intractable task of sampling from the high dimensional DPP down to an easily manageable size: sampling a set of ten summary frames from the longest recording took less than one second.

Of course, this speedup naturally comes at some approximation cost. In order to evaluate empirically whether the Nyström samples retained the advantages of the original DPP, which is too expensive for direct comparison, we performed a user study. Each subject in the study was shown, for each of the original nine recordings, a set of ten poses (rendered graphically) sampled from the approximated DPP model alongside a set of ten poses sampled uniformly at random (see Figure 4). We asked the subjects to evaluate the two pose sets with respect to the motion capture recording, which was provided in the form of a rendered video. The subjects chose the set they felt better represented the characteristic poses from the video (quality), the set they felt was more diverse, and the set they felt made the better overall summary. The order of the two sets was randomized, and the samples were different for each user. 18 subjects completed the study, for a total of 162 responses to each question.

Evaluation measure	% DPP	% Random
Quality	66.7	33.3
Diversity	64.8	35.2
Overall	67.3	32.7

Table 1: The percentage of subjects choosing each method in a user study of motion capture summaries.

The results of the user study are shown in Table 1. Overall, the subjects felt that the samples from the Nyström-approximated DPP were significantly better on all three measures, $p < 0.001$.

6 Conclusion

The Nyström approximation is an appealing technique for managing the otherwise intractable task of sampling from high-dimensional DPPs. We showed that this appeal is theoretical as well as practical: we proved upper bounds for the variational error of Nyström-approximated DPPs and presented empirical results to validate them. We also demonstrated that Nyström-approximated DPPs can be usefully applied to the task of summarizing motion capture recordings. Future work includes incorporating the structure of the kernel matrix to derive potentially tighter bounds.

Acknowledgements

RHA and EBF were supported in part by AFOSR Grant FA9550-12-1-0453 and DARPA Grant FA9550-12-1-0406 negotiated by AFOSR. BT was partially supported by NSF CAREER Grant 1054215 and by STARnet, a Semiconductor Research Corporation program sponsored by MARCO and DARPA.

References

N.F. Arcolano. *Approximation of Positive Semidefinite Matrices Using the Nyström Method*. PhD thesis, Harvard University, 2011.

R. Bhatia. *Matrix Analysis*, volume 169. Springer Verlag, 1997.

A. Borodin and E.M. Rains. Eynard-Mehta Theorem, Schur Process, and Their Pfaffian Analogs. *Journal of Statistical Physics*, 121(3):291–317, 2005.

E. Candes and J. Romberg. Sparsity and Incoherence in Compressive Sampling. *Inverse Problems*, 23(3): 969, 2007.

A. Deshpande, L. Rademacher, S. Vempala, and G. Wang. Matrix Approximation and Projective Clustering via Volume Sampling. *Theory of Computing*, 2: 225–247, 2006.

P. Drineas and M. W. Mahoney. On the Nyström Method for Approximating a Gram Matrix for Improved Kernel-based Learning. *Journal of Machine Learning Research*, 6:2153–2175, 2005.

A. Frieze, R. Kannan, and S. Vempala. Fast Monte-Carlo Algorithms for Finding Low-rank Approximations. *Journal of the ACM (JACM)*, 51(6):1025–1041, 2004.

J. Gillenwater, A. Kulesza, and B. Taskar. Discovering Diverse and Salient Threads in Document Collections. *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 710–720, 2012.

J.B. Hough, M. Krishnapur, Y. Peres, and B. Virág. Determinantal Processes and Independence. *Probability Surveys*, 3:206–229, 2006.

A. Kulesza and B. Taskar. Structured Determinantal Point Processes. *Advances in Neural Information Processing Systems*, 23:1171–1179, 2010.

A. Kulesza and B. Taskar. k-DPPs: Fixed-size Determinantal Point Processes. *Proceedings of the 28th International Conference on Machine Learning*, pages 1193–1200, 2011.

S. Kumar, M. Mohri, and A. Talwalkar. Sampling Methods for the Nyström Method. *Journal of Machine Learning Research*, 13:981–1006, 2012.

A. Rahimi and B. Recht. Random Features for Large-scale Kernel Machines. *Advances in Neural Information Processing Systems*, 20:1177–1184, 2007.

A. Talwalkar and A. Rostamizadeh. Matrix Coherence and the Nyström Method. *arXiv preprint arXiv:1004.2008*, 2010.

C. Williams and M. Seeger. Using the Nyström Method to Speed Up Kernel Machines. *Advances in Neural Information Processing Systems*, 13:682–688, 2001.

Ti. Yang, Y. Li, M. Mahdavi, R. Jin, and Z. Zhou. Nystrom Method vs Random Fourier Features: A Theoretical and Empirical Comparison. *Advances in Neural Information Processing Systems*, 25:485–493, 2012.

A Appendix-Supplementary Material

A.1 Proofs to Theorem 5 and Theorem 6

Lemma 4. *Let e_k denote the kth elementary symmetric polynomial of L:*

$$e_k(\lambda_1, \ldots, \lambda_N) = \sum_{|J|=k} \prod_{n \in J} \lambda_n \ , \tag{32}$$

and

$$\hat{\lambda}_i = \max\left\{\lambda_{i+(m-r)}, \lambda_i - \|L - \tilde{L}\|_2\right\} \ , \tag{33}$$

where m is the rank of L and r is the rank of $\tilde{L}$. Then

$$e_k(\lambda_1, \ldots, \lambda_N) \geq e_k(\tilde{\lambda}_1, \ldots, \tilde{\lambda}_N) \geq e_k(\hat{\lambda}_1, \ldots, \hat{\lambda}_N) \ . \tag{34}$$

Proof.

$$e_k(\lambda_1, \ldots, \lambda_N) = \sum_{|J|=k} \prod_{n \in J} \lambda_n \geq \sum_{|J|=k} \prod_{n \in J} \tilde{\lambda}_n = e_k(\tilde{\lambda}_1, \ldots, \tilde{\lambda}_N) \ ,$$

by Corollary 1.

On the other hand,

$$e_k(\tilde{\lambda}_1, \ldots, \tilde{\lambda}_N) = \sum_{|J|=k} \prod_{n \in J} \tilde{\lambda}_n \geq \sum_{|J|=k} \prod_{n \in J} \hat{\lambda}_n = e_k(\hat{\lambda}_1, \ldots, \hat{\lambda}_N) \ ,$$

by Corollary 1 and Corollary 2. □

Since

$$\mathcal{P}_L^k(A) = \frac{\det(L_A)}{\sum_{|A'|=k} \det(L_{A'})} = \frac{\det(L_A)}{\sum_{|J|=k} \prod_{n \in J} \lambda_n} = \frac{\det(L_A)}{e_k(\lambda_1, \ldots, \lambda_N)} \ , \tag{35}$$

using Lemma 4, we can now prove Theorem 3 and Theorem 4.

Proof of Theorem 3.

$$|\mathcal{P}_A^k - \tilde{\mathcal{P}}_A^k| = \left[\frac{\det(\tilde{L}_A)}{e_k(\tilde{\lambda}_1, \ldots, \tilde{\lambda}_N)} - \frac{\det(L_A)}{e_k(\lambda_1, \ldots, \lambda_N)}\right] = \ \mathcal{P}_A^k \left[\frac{e_k(\lambda_1, \ldots, \lambda_N)}{e_k(\tilde{\lambda}_1, \ldots, \tilde{\lambda}_N)} - 1\right] \leq \ \mathcal{P}_A^k \left[\frac{e_k(\lambda_1, \ldots, \lambda_N)}{e_k(\hat{\lambda}_1, \ldots, \hat{\lambda}_N)} - 1\right] \ ,$$

where the last inequality follows from Lemma 4. □

Proof of Theorem 4.

$$\tilde{\mathcal{P}}_A^k - \mathcal{P}_A^k = \mathcal{P}_A^k \left[\frac{e_k(\lambda_1, \ldots, \lambda_N) \det(\tilde{L}_A)}{e_k(\tilde{\lambda}_1, \ldots, \tilde{\lambda}_N) \det(L_A)} - 1\right] = \mathcal{P}_A^k \left[\left(\frac{e_k(\lambda_1, \ldots, \lambda_N)}{e_k(\tilde{\lambda}_1, \ldots, \tilde{\lambda}_N)}\right)\left(\frac{\prod_{i=1}^k \tilde{\lambda}_i^A}{\prod_{i=1}^k \lambda_i^A}\right) - 1\right] \ .$$

Here $\tilde{\lambda}_1^A, \geq, \ldots, \geq \tilde{\lambda}_k^A$ are the eigenvalues of $\tilde{L}_A$. Now note that $L_A = \tilde{L}_A + E_A$. Since E is positive semidefinite, it follows that E_A is also positive semidefinite. Thus by Corollary 1, we have $\lambda_i^A \geq \tilde{\lambda}_i^A$ and so

$$\tilde{\mathcal{P}}_A^k - \mathcal{P}_A^k \leq \mathcal{P}_A^k \left[\frac{e_k(\lambda_1, \ldots, \lambda_N)}{e_k(\tilde{\lambda}_1, \ldots, \tilde{\lambda}_N)} - 1\right] \leq \ \mathcal{P}_A^k \left[\frac{e_k(\lambda_1, \ldots, \lambda_N)}{e_k(\hat{\lambda}_1, \ldots, \hat{\lambda}_N)} - 1\right] \ ,$$

where the last inequality follows from Lemma 4.

On the other hand,

$$\mathcal{P}_A^k - \tilde{\mathcal{P}}_A^k = \mathcal{P}_A^k \left[1 - \left(\frac{e_k(\lambda_1, \ldots, \lambda_N)}{e_k(\tilde{\lambda}_1, \ldots, \tilde{\lambda}_N)}\right)\left(\frac{\prod_{i=1}^k \tilde{\lambda}_i^A}{\prod_{i=1}^k \lambda_i^A}\right)\right] \ . \tag{36}$$

By Corrolary 2,

$$\tilde{\lambda}_i^A \geq \lambda_i^A - \xi_1^A \geq \lambda_i^A - \xi_1 = \lambda_i^A - \|L - \tilde{L}\|_2 \ . \tag{37}$$

We also note that $\tilde{\lambda}_i^A \geq 0$. Since $e_k(\lambda_1, \ldots, \lambda_N) \geq e_k(\tilde{\lambda}_1, \ldots, \tilde{\lambda}_N)$ by Lemma 4, we have

$$\mathcal{P}_A^k - \tilde{\mathcal{P}}_A^k \leq \mathcal{P}_A^k \left[1 - \frac{\prod_{i=1}^k \hat{\lambda}_i^A}{\prod_{i=1}^k \lambda_i^A} \right] \ . \tag{38}$$

The theorem follows by combining the two inequalities. □

A.2 Empirical Comparisons to Random Fourier Features

In cases where the kernel matrix L is generated from a shift-invariant kernel function $k(\boldsymbol{x}, \boldsymbol{y}) = k(\boldsymbol{x} - \boldsymbol{y})$, we can construct a low-rank approximation using random Fourier features (RFFs) (Rahimi and Recht, 2007). This involves mapping each data point $\boldsymbol{x} \in \mathbb{R}^d$ onto a random direction $\boldsymbol{\omega}$ drawn from the Fourier transform of the kernel function. In particular, we draw $\boldsymbol{\omega} \sim p(\boldsymbol{\omega})$, where

$$p(\boldsymbol{\omega}) = \int_{\mathbb{R}^d} k(\boldsymbol{\Delta}) \exp(-i\boldsymbol{\omega}^\top \boldsymbol{\Delta}) d\boldsymbol{\Delta} \ , \tag{39}$$

draw b uniformly from $[0, 2\pi]$, and set $z_{\boldsymbol{\omega}}(\boldsymbol{x}) = \sqrt{2}\cos(\boldsymbol{\omega}^\top \boldsymbol{x} + b)$. It can be shown then that $z_{\boldsymbol{\omega}}(\boldsymbol{x}) z_{\boldsymbol{\omega}}(\boldsymbol{y})$ is an unbiased estimator of $k(\boldsymbol{x} - \boldsymbol{y})$. Note that the shift-invariant property of the kernel function is crucial to ensure that $p(\boldsymbol{\omega})$ is a valid probability distribution, due to Bochner's Theorem. The variance of the estimate can be improved by drawing D random direction, $\boldsymbol{\omega}_1, \ldots, \boldsymbol{\omega}_D \sim p(\boldsymbol{\omega})$ and estimating the kernel function with $k(\boldsymbol{x} - \boldsymbol{y})$ as $\frac{1}{D} \sum_{j=1}^D z_{\boldsymbol{\omega}_j}(\boldsymbol{x}) z_{\boldsymbol{\omega}_j}(\boldsymbol{y})$.

To use RFFs for approximating DPP kernel matrices, we assume that the matrix L is generated from a shift-invariant kernel function, so that if $\boldsymbol{x}_i$ is the vector representing item i then

$$L_{ij} = k(\boldsymbol{x}_i - \boldsymbol{x}_j) \ . \tag{40}$$

We construct a $D \times N$ matrix B with

$$B_{ij} = \frac{1}{\sqrt{D}} z_{\boldsymbol{\omega}_i}(\boldsymbol{x}_j) \qquad i = 1, \ldots, D, j = 1, \ldots, N \ . \tag{41}$$

An unbiased estimator of the kernel matrix L is now given by $\tilde{L}^{\text{RFF}} = B^\top B$. Furthermore, note that an approximation to the dual kernel matrix C is given by $\tilde{C}^{\text{RFF}} = BB^\top$; this allows use of the sampling algorithm given in Algorithm 2.

We apply the RFF approximation method to the Abalone data from Section 4. We use a Gaussian RBF kernel,

$$L_{ij} = \exp(-\frac{\|\boldsymbol{x}_i - \boldsymbol{x}_j\|^2}{\sigma^2}) \qquad i, j = 1, \ldots, 1000 \ , \tag{42}$$

with σ^2 taking values 0.1,1, and 10. In this case, the Fourier transform of the kernel function, $p(\boldsymbol{\omega})$ is also a multivariate Gaussian.

In Figure 5 we plot the empirically estimated $\log(\|\mathcal{P}^k - \tilde{\mathcal{P}}^k\|_1)$ for $k = 10$. While RFFs compare favorably to the uniform random sampling of landmarks, their performance is significantly worse than that of the adaptive Nyström methods, especially in the case where there are strong correlations between items ($\sigma^2 = 1$ and 10). In the extreme case where there is little to no correlation, the Nyström methods suffer because a small sample of landmarks cannot reconstruct the other items accurately. Yang et al. (2012) have previously demonstrated that, in kernel learning tasks, the Nyström methods perform favorably compared to RFFs in cases where there are large eigengaps in the kernel matrix. The plot of the eigenvalues in Figure 6 suggests that a similar result holds for approximating DPPs as well. In practice, for kernel learning tasks, the RFF approach typically requires more features than the number of landmarks needed for Nystroöm methods. However, due the fact that sampling from a DPP requires $O(D^3)$ time, we are constrained by the number of landmarks that can be used.

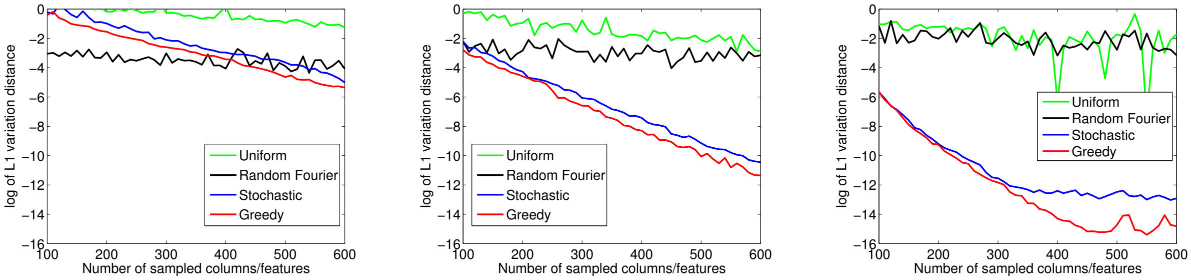


Figure 5: Error of Nyström and random Fourier features approximations: $\log(\|\mathcal{P} - \tilde{\mathcal{P}}\|_1)$ as a function of the number of landmarks sampled/random features used. From left to right, the values of σ^2 are 0.1, 1, and 10.

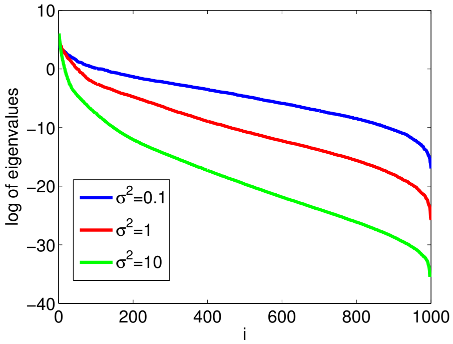
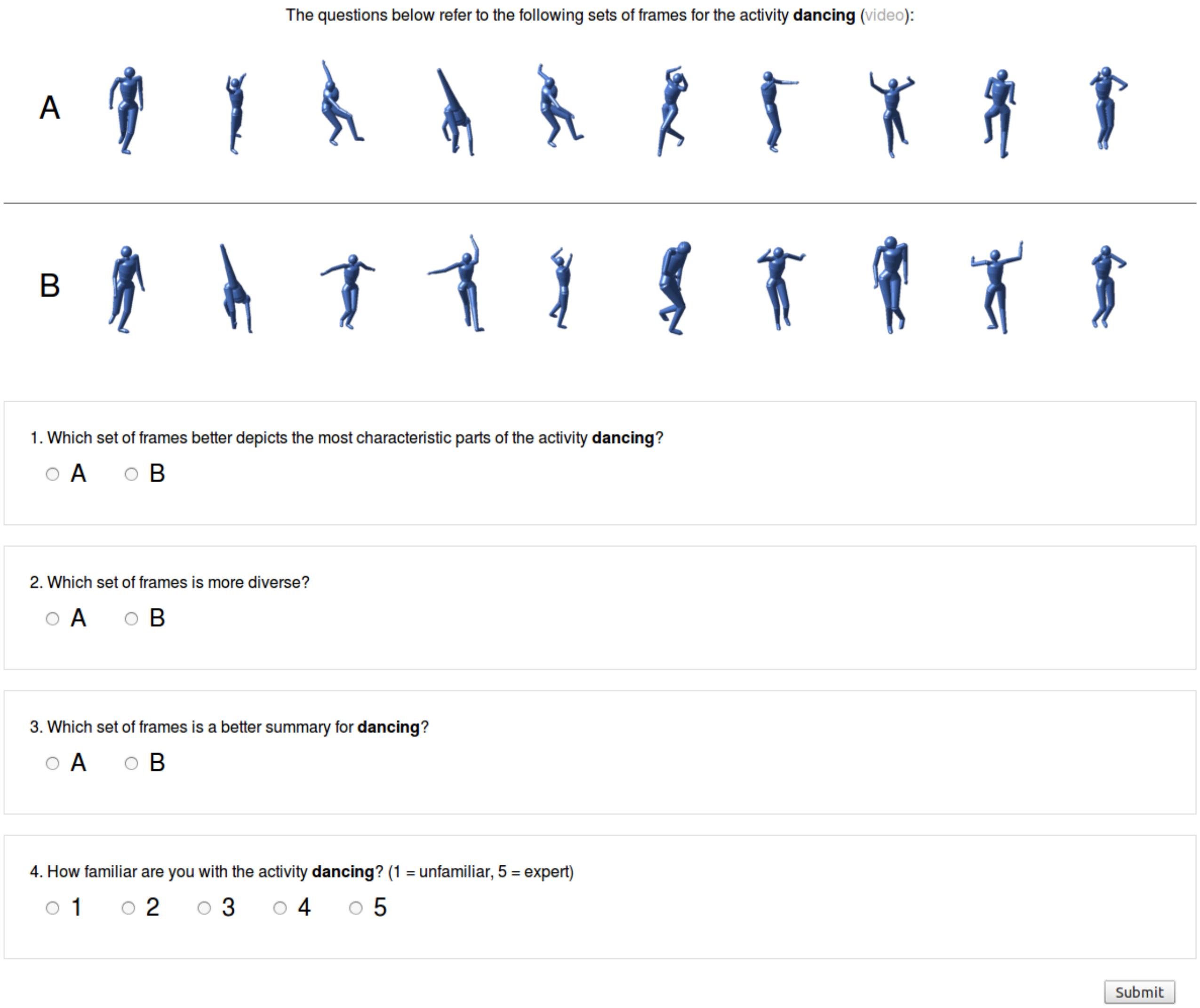


Figure 6: The log-eigenvalues of RBF kernel applied on the Abalone datset.

A.3 Sample User Study

Figure 7 shows a sample screen from our user study. Each subject completed four questions for each of the nine pairs of sets they saw (one pair for each of the nine activities). There was no significant correlation between a user's preference for the DPP set and their familiarity with the activity.

Figure 8 shows motion capture summaries sampled from the Nyström-approximated kDPP (k=10).



The questions below refer to the following sets of frames for the activity **dancing** (video):

1. Which set of frames better depicts the most characteristic parts of the activity **dancing**?

○ A ○ B

2. Which set of frames is more diverse?

○ A ○ B

3. Which set of frames is a better summary for **dancing**?

○ A ○ B

4. How familiar are you with the activity **dancing**? (1 = unfamiliar, 5 = expert)

○ 1 ○ 2 ○ 3 ○ 4 ○ 5

Figure 7: Sample screen from the user study.

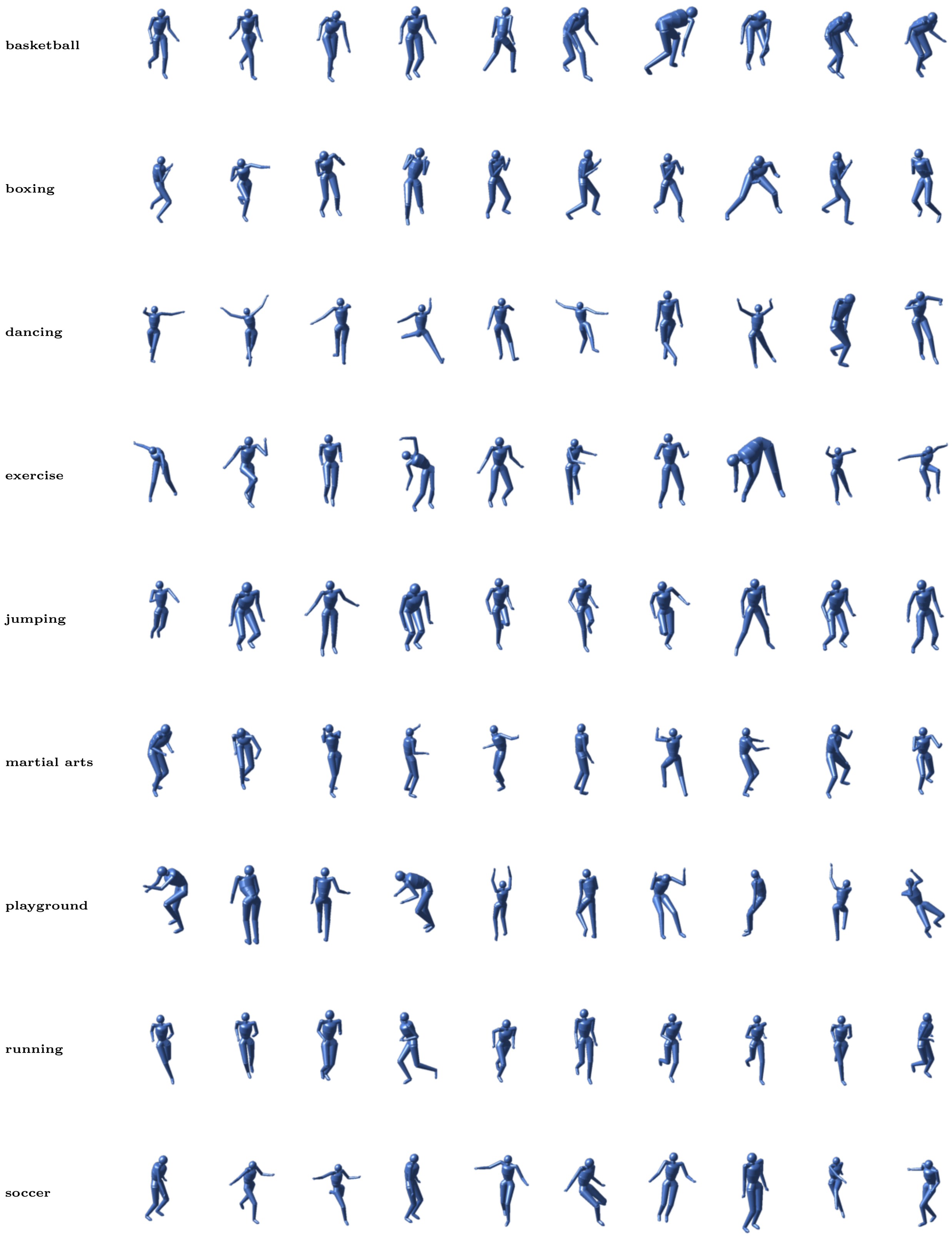


Figure 8: DPP samples ($k = 10$) for each activity.