Large-Margin Determinantal Point Processes

Wei-Lun Chao* U. of Southern California Los Angeles, CA 90089 weilunc@usc.edu

Boqing Gong* U. of Southern California Los Angeles, CA 90089 boqinggo@usc.edu

Kristen Grauman U. of Texas at Austin Austin, TX 78701 grauman@cs.utexas.edu

Fei Sha U. of Southern California Los Angeles, CA 90089 feisha@usc.edu

Abstract

Determinantal point processes (DPPs) offer a powerful approach to modeling diversity in many applications where the goal is to select a diverse subset from a ground set of items. We study the problem of learning the parameters (i.e., the kernel matrix) of a DPP from labeled training data. In this paper, we develop a novel parameter estimation technique particularly tailored for DPPs based on the principle of large margin separation. In contrast to the state-of-the-art method of maximum likelihood estimation of the DPP parameters, our large-margin loss function explicitly models errors in selecting the target subsets, and it can be customized to trade off different types of errors (precision vs. recall). Extensive empirical studies validate our contributions, including applications on challenging document and video summarization, where flexibility in balancing different errors while training the summarization models is indispensable.

1 INTRODUCTION

Imagine we are to design a search engine to retrieve web images that match user queries. In response to the search term JAGUAR, what should we retrieve—the images of the animal jaguar or the images of the automobile jaguar?

This frequently cited example illustrates the need to incorporate the notion of *diversity*. In many tasks, we want to select a subset of items from a "ground set". While the ground set might contain many similar items, our goal is *not* to discover all of the same ones, but rather to find a subset of diverse items that ensure coverage (the exact definition of coverage is task-specific). In the example of retrieving images for JAGUAR, we achieve diversity by including both types of images.

*Equal contribution

Recently, the determinantal point process (DPP) has emerged as a promising technique for modeling diversity [Kulesza and Taskar, 2012]. A DPP defines a probability distribution over the power set of a ground set. Intuitively, subsets of higher diversity are assigned larger probabilities, and thus are more likely to be selected than those with lower diversity. Since its original application to quantum physics, DPP has found many applications in modeling random trees and graphs [Burton and Pemantle, 1993], document summarization [Kulesza and Taskar, 2011b], search and ranking in information retrieval [Kulesza and Taskar, 2011a], and clustering [Kang, 2013]. Various extensions have also been studied, including k-DPP [Kulesza and Taskar, 2011a], structured DPP [Kulesza and Taskar, 2011c], Markov DPP [Affandi et al., 2012], and DPP on continuous spaces [Affandi et al., 2013].

The probability distribution of a DPP depends crucially on its kernel—a square and symmetric, positive semidefinite matrix whose elements specify how similar every pair of items in the ground set are. This kernel matrix is often unknown and needs to be estimated from training data.

This is a very challenging problem for several reasons. First, the number of the parameters, i.e., the number of elements in the kernel matrix, is quadratic in the number of items in the ground set. For many tasks (for instance, image search), the ground set can be very large. Thus it is impractical to directly specify every element of the matrix, and a suitable reparameterization of the matrix is necessary. Secondly, the number of training samples is often limited in many practical applications. One such example is the task of document summarization, where our aim is to select a succinct subset of sentences from a long document. There, acquiring accurate annotations from human experts is costly and difficult. Thirdly, for many tasks, we need to evaluate the performance of the learned DPP not only by its accuracy in predicting whether an item should be selected, but also by other measures like precision and recall. For instance, failing to select key sentences for summarizing documents might be regarded as being more catastrophic than injecting sentences with repetitive information into the

summary.

Existing methods of parameter estimation for DPPs are inadequate to address these challenges. For example, maximum likelihood estimation (MLE) typically requires a large number of training samples in order to estimate the underlying model correctly. This also limits the number of the parameters it can estimate reliably, restricting its use to DPPs whose kernels can be parameterized with few degrees of freedom. It also does not offer fine control over precision and recall.

We propose a two-pronged approach for learning a DPP from labeled data. First, we improve modeling flexibility by reparameterizing the DPP's kernel matrix with multiple base kernels. This representation could easily incorporate domain knowledge and requires learning fewer parameters (instead of the whole kernel matrix). Then, we optimize the parameters such that the probability of the correct subset is larger than other erroneous subsets by a large margin. This margin is task-specific and can be customized to reflect the desired performance measure—for example, to monitor precision and recall. As such, our approach defines objective functions that closely track selection errors and work well with few training samples. While the principle of large margin separation has been widely used in classification [Vapnik, 1998] and structured prediction [Taskar et al., 2005], formulating DPP learning with the large margin principle is novel. Our empirical studies show that the proposed method attains superior performance on two challenging tasks of practical interest: document and video summarization.

The rest of the paper is organized as follows. We provide background on the DPP in section 2, followed by our approach in section 3. We discuss related work in section 4 and report our empirical studies in section 5. We conclude in section 6.

2 BACKGROUND: DETERMINANTAL POINT PROCESSES

We first review background on the determinantal point process (DPP) [Macchi, 1975] and the standard maximum likelihood estimation technique for learning DPP parameters from data. More details can be found in the excellent tutorial [Kulesza and Taskar, 2012].

Given a ground set of M items, $\mathcal{Y} = \{1, 2, \ldots, \mathsf{M}\}$, a DPP defines a probabilistic measure over the power set, i.e., all possible subsets (including the empty set) of $\mathcal{Y}$. Concretely, let $\boldsymbol{L}$ denote a symmetric and positive semidefinite matrix in $\mathbb{R}^{\mathsf{M}\times\mathsf{M}}$. The probability of selecting a subset $\boldsymbol{y} \subseteq \mathcal{Y}$ is given by

$$P(\boldsymbol{y}; \boldsymbol{L}) = \det(\boldsymbol{L} + \boldsymbol{I})^{-1} \det(\boldsymbol{L}_{\boldsymbol{y}}), \tag{1}$$

where $\boldsymbol{L}_{\boldsymbol{y}}$ denotes the submatrix of $\boldsymbol{L}$, with rows and columns selected by the indices in $\boldsymbol{y}$. $\boldsymbol{I}$ is the identity matrix with the proper size. We define $\det(\boldsymbol{L}_{\emptyset}) = 1$. The above way of defining a DPP is called an L-ensemble. An equivalent way of defining a DPP is to use a kernel matrix to define the marginal probability of selecting a random subset:

$$P_{\boldsymbol{y}} = \sum_{\boldsymbol{y}' \subseteq \mathcal{Y}} P(\boldsymbol{y}'; \boldsymbol{L}) \mathbb{I}[\, \boldsymbol{y} \subseteq \boldsymbol{y}' \,] = \det(\boldsymbol{K}_{\boldsymbol{y}}), \tag{2}$$

where we sum over all subsets $\boldsymbol{y}'$ that contain $\boldsymbol{y}$ ($\mathbb{I}[\,\cdot\,]$ is an indicator function). The matrix $\boldsymbol{K}$ is another positive semidefinite matrix, computable from the $\boldsymbol{L}$ matrix

$$\boldsymbol{K} = \boldsymbol{L}(\boldsymbol{L} + \boldsymbol{I})^{-1}, \tag{3}$$

and $\boldsymbol{K}_{\boldsymbol{y}}$ is the submatrix of $\boldsymbol{K}$ indexed by $\boldsymbol{y}$. Despite the exponential number of summands in eq. (2), the marginalization is analytically tractable and computable in polynomial time.

2.1 MODELING DIVERSITY

One particularly useful property of the DPP is its ability to model *pairwise repulsion*. Consider the marginal probability of having two items i and j simultaneously in a subset:

$$\begin{aligned} P_{\{i,j\}} &= \det \begin{vmatrix} K_{ii} K_{ij} \\ K_{ji} K_{jj} \end{vmatrix} = K_{ii} K_{jj} - K_{ij}^2 \\ &\leq K_{ii} K_{jj} = P_{\{i\}} P_{\{j\}} \leq \min(P_{\{i\}}, P_{\{j\}}). \end{aligned} \tag{4}$$

Thus, unless $K_{ij} = 0$, the probability of observing i and j jointly is always less than observing either i or j separately. Namely, having i in a subset repulsively excludes j and vice versa. Another extreme case is when i and j are the same; then $K_{ii} = K_{jj} = K_{ij}$, which leads to $P_{\{i,j\}} = 0$. Namely, we should never allow them together in any subset.

Consequently, a subset with a large (marginal) probability cannot have too many items that are similar to each other (i.e., with high values of K_{ij}). In other words, the probability provides a gauge of the diversity of the subset. The most diverse subset, which balances all the pairwise repulsions, is the subset that attains the highest probability

$$\boldsymbol{y}^{\text{MAP}} = \arg\max_{\boldsymbol{y}} P(\boldsymbol{y}; \boldsymbol{L}). \tag{5}$$

Note that this MAP inference is computed with respect to the L-ensemble (instead of $\boldsymbol{K}$) as we are interested in the mode, not the marginal probability of having the subset. Unfortunately, the MAP inference is NP-hard [Ko et al., 1995]. Various approximation algorithms have been investigated [Gillenwater et al., 2012, Kulesza and Taskar, 2012].

2.2 MAXIMUM LIKELIHOOD ESTIMATION

Suppose we are given a training set $\{(\mathcal{Y}_n, \boldsymbol{y}_n)\}$, where each ground set $\mathcal{Y}_n$ is annotated with its most diverse subset $\boldsymbol{y}_n$. How can we discover the underlying parameters $\boldsymbol{L}$ or $\boldsymbol{K}$? Note that different ground sets need not have overlap. Thus, directly specifying kernel values for every pair of items is unlikely to be scalable. Instead, we will need to assume that either $\boldsymbol{L}$ or $\boldsymbol{K}$ for each ground set is represented by a shared set of parameters $\boldsymbol{\theta}$.

For items i and j in $\mathcal{Y}_n$, suppose their kernel values $K_{n_{ij}}$ can be computed as a function of $\boldsymbol{x}_{n_i}$, $\boldsymbol{x}_{n_j}$ and $\boldsymbol{\theta}$, where $\boldsymbol{x}_{n_i}$ and $\boldsymbol{x}_{n_j}$ are features characterizing those items. Our learning objective is to optimize $\boldsymbol{\theta}$ such that $\boldsymbol{y}_n$ is the most diverse subset in $\mathcal{Y}_n$, or attains the highest probability. This gives rise to the following maximum likelihood estimate (MLE) [Kulesza and Taskar, 2011b],

$$\boldsymbol{\theta}^{\text{MLE}} = \arg\max_{\boldsymbol{\theta}} \sum_n \log P(\boldsymbol{y}_n; \boldsymbol{L}_n(\mathcal{Y}_n; \boldsymbol{\theta})), \quad (6)$$

where $\boldsymbol{L}_n(\mathcal{Y}_n; \boldsymbol{\theta})$ converts features in $\mathcal{Y}_n$ to the $\boldsymbol{L}$ matrix for the ground set $\mathcal{Y}_n$. MLE has been a standard approach for estimating DPP parameters. However, as we will discuss in section 3.2, it has important limitations.

Next, we introduce our method for learning the parameters. We first present our multiple kernel based representation of the $\boldsymbol{L}$ matrix and then the large-margin based estimation.

3 OUR APPROACH

Our approach consists of two components that are developed in parallel, yet work in concert: (1) the use of multiple kernel functions to represent the DPP; (2) applying the principle of large margin separation to optimize the parameters. The former reduces the number of parameters to learn and thus is especially advantageous when the number of training samples is limited. The latter strengthens the advantage by optimizing objective functions that closely track subset selection errors.

3.1 MULTIPLE KERNEL REPRESENTATION

Learning the $\boldsymbol{L}$ or $\boldsymbol{K}$ matrix for a DPP is an instance of learning kernel functions, as those matrices are positive semidefinite matrices, interpretable as kernel functions being evaluated on the items in the ground set. Thus, our goal is essentially to learn the right kernel function to measure similarity.

However, for many applications, similarity is just one of the criteria for selecting items. For instance, in the previous example of image retrieval, the retrieved images not only need to be diverse (thus different) but also need to have strong relevance to the query term. Similarly, in document summarization, the selected sentences not only need to be succinct and not redundant, but also need to represent the contents of the document [Lin and Bilmes, 2010].

Kulesza and Taskar [2011b] propose to balance these two potentially conflicting forces with a decomposable $\boldsymbol{L}$ matrix:

$$L_{ij} = q_i q_j S_{ij} = q_i q_j \boldsymbol{\phi}_i^{\mathrm{T}} \boldsymbol{\phi}_j,$$
$$q_i = q(\boldsymbol{x}_i) = \exp(\boldsymbol{\theta}^{\mathrm{T}} \boldsymbol{x}_i), \quad \forall\, i, j \in \mathcal{Y}, \quad (7)$$

where q_i is referred to as the *quality* factor, modeling how representative or relevant the selected items are. It depends on item i's feature vector $\boldsymbol{x}_i$, which encodes i's contextual information and its *representativeness* of other items. For example, in document summarization, possible features are the sentence lengths, positions of the sentences in the text, or others. S_{ij}, on the other hand, measures how *similar* two sentences are, computed from a different set of features, $\boldsymbol{\phi}_i$ and $\boldsymbol{\phi}_j$, such as bag-of-words descriptors that represent each item's individual characteristics.

However, prior work [Kulesza and Taskar, 2011b] does not investigate whether this specific definition of similarity could be made optimal and adapted to the data, thus limiting the modeling power of the DPP largely to infer the quality q_i. Our empirical studies show that this limitation can be severe, especially when the modeling choice is erroneous (cf. section 5.2).

In this paper, we retain the aspect of quality modeling but improve the modeling of similarity S_{ij} in two ways. First, we use nonlinear kernel functions such as the Gaussian RBF kernel to determine similarity. Secondly, and more importantly, we combine several base kernels:

$$S_{ij} = \sum_k \alpha_k \exp\{-\left\|\boldsymbol{\phi}_i - \boldsymbol{\phi}_j\right\|_2^2 / \sigma_k^2\} + \beta \boldsymbol{\phi}_i^{\mathrm{T}} \boldsymbol{\phi}_j, \quad (8)$$

where k indexes the base kernels and σ_k is a scaling factor. The combination coefficients are constrained such that $\sum_k \alpha_k + \beta = 1$. They are optimized on the annotated data, either via maximum likelihood estimation or via our novel parameter estimation technique, to be described next.

3.2 LARGE-MARGIN ESTIMATION OF DPP

Maximum likelihood estimation does not closely track discriminative errors [Ng and Jordan, 2002, Vapnik, 1998, Jebara, 2004]. While improving the likelihood of the ground-truth subset $\boldsymbol{y}_n$, MLE could also improve the likelihoods of other competing subsets. Consequentially, a model learned with MLE could have modes that are very different subsets yet are very close to each other in their probability values. Having highly confusable modes is especially problematic for DPP's NP-hard MAP inference—the difference between such modes can fall within the approximation errors of approximate inference algorithms such that the true MAP cannot be easily extracted.

3.2.1 Multiplicative Large Margin Constraints

To address these deficiencies, our large-margin based approach aims to maintain or increase the margin between the correct subset and alternative, incorrect ones. Specifically, we formulate the following large margin constraints

$$\begin{aligned}\log P(\boldsymbol{y}_n; \boldsymbol{L}_n) &\geq \max_{\boldsymbol{y} \subseteq \mathcal{Y}_n} \log \ell(\boldsymbol{y}_n, \boldsymbol{y})\, P(\boldsymbol{y}; \boldsymbol{L}_n) \\ &= \max_{\boldsymbol{y} \subseteq \mathcal{Y}_n} \log \ell(\boldsymbol{y}_n, \boldsymbol{y}) + \log P(\boldsymbol{y}; \boldsymbol{L}_n), \quad (9)\end{aligned}$$

where $\ell(\boldsymbol{y}_n, \boldsymbol{y})$ is a loss function measuring the discrepancy between the correct subset and an alternative $\boldsymbol{y}$. We assume $\ell(\boldsymbol{y}_n, \boldsymbol{y}_n) = 0$.

Intuitively, the more different $\boldsymbol{y}$ is from $\boldsymbol{y}_n$, the larger the gap we want to maintain between the two probabilities. This way, the incorrect one has less chance to be identified as the most diverse one. Note that while similar intuitions have been explored in multi-way classification and structured prediction, the margin here is *multiplicative* instead of additive—this is by design, as it leads to a tractable optimization over the exponential number of constraints, as we will explain later.

3.2.2 Design of the Loss Function

A natural choice for the loss function is the Hamming distance between $\boldsymbol{y}_n$ and $\boldsymbol{y}$, counting the number of disagreements between two subsets:

$$\ell_H(\boldsymbol{y}_n, \boldsymbol{y}) = \sum_{i \in \boldsymbol{y}} \mathbb{I}[i \notin \boldsymbol{y}_n] + \sum_{i \notin \boldsymbol{y}} \mathbb{I}[i \in \boldsymbol{y}_n]. \quad (10)$$

In this loss function, failing to select the right item costs the same as adding an unnecessary item. In many tasks, however, this symmetry does not hold. For example, in summarizing a document, omitting a key sentence has more severe consequences than adding a (trivial) sentence.

To balance these two types of errors, we introduce the generalized Hamming loss function,

$$\ell_\omega(\boldsymbol{y}_n, \boldsymbol{y}) = \sum_{i \in \boldsymbol{y}} \mathbb{I}[i \notin \boldsymbol{y}_n] + \omega \sum_{i \notin \boldsymbol{y}} \mathbb{I}[i \in \boldsymbol{y}_n]. \quad (11)$$

When ω is greater than 1, the learning biases towards higher *recall* to select as many items in $\boldsymbol{y}_n$ as possible. When ω is significantly less than 1, the learning biases towards high *precision* to avoid incorrect items as much as possible. Our empirical studies demonstrate such flexibility and its advantages in two real-world summarization tasks.

3.2.3 Numerical Optimization

To overcome the challenge of dealing with an exponential number of constraints in eq. (9), we reformulate it as a tractable optimization problem. We first upper-bound the hard-max operation with Jensen's inequality (i.e., softmax):

$$\begin{aligned}\log P(\boldsymbol{y}_n; \boldsymbol{L}_n) &\geq \log \sum_{\boldsymbol{y} \subseteq \mathcal{Y}} e^{\log \ell_\omega(\boldsymbol{y}_n, \boldsymbol{y})\, P(\boldsymbol{y}; \boldsymbol{L}_n)} \\ &= \mathsf{softmax}_{\boldsymbol{y} \subseteq \mathcal{Y}_n} \log \ell_\omega(\boldsymbol{y}_n, \boldsymbol{y}) + \log P(\boldsymbol{y}; \boldsymbol{L}_n). \quad (12)\end{aligned}$$

With the loss function $\ell_\omega(\boldsymbol{y}_n, \boldsymbol{y})$, the right-hand-side is computable in polynomial time,

$$\begin{aligned}&\mathsf{softmax}_{\boldsymbol{y} \subseteq \mathcal{Y}_n} \log \ell_\omega(\boldsymbol{y}_n, \boldsymbol{y}) + \log P(\boldsymbol{y}; \boldsymbol{L}_n) \\ &= \log \left(\sum_{i \notin \boldsymbol{y}_n} K_{n_{ii}} + \omega \sum_{i \in \boldsymbol{y}_n} (1 - K_{n_{ii}}) \right), \quad (13)\end{aligned}$$

where $K_{n_{ii}}$ is the i-th element on the diagonal of $\boldsymbol{K}_n$, the marginal kernel matrix corresponding to $\boldsymbol{L}_n$. The detailed derivation of this result is in the supplementary material. Note that $\boldsymbol{K}_n$ can be computed efficiently from $\boldsymbol{L}_n$ through the identity eq. (3).

The softmax can be seen as a summary of all undesirable subsets (the correct subset $\boldsymbol{y}_n$ does not contribute to the weighted sum as $\ell_\omega(\boldsymbol{y}_n, \boldsymbol{y}_n) = 0$). Our optimization balances this term with the likelihood of the target with the hinge loss function $[z]_+ = \max(0, z)$,

$$\begin{aligned}\min \sum_n \Bigg[&- \log P(\boldsymbol{y}_n; \boldsymbol{L}_n) \\ &+ \lambda \log \left(\sum_{i \notin \boldsymbol{y}_n} K_{n_{ii}} + \omega \sum_{i \in \boldsymbol{y}_n} (1 - K_{n_{ii}}) \right) \Bigg]_+, \quad (14)\end{aligned}$$

where $\lambda \geq 0$ is a tradeoff coefficient, to be tuned on validation datasets. Note that this objective function subsumes maximum likelihood estimation where $\lambda = 0$. We optimize the objective function with subgradient descent. Details are in the supplementary material.

4 RELATED WORK

The DPP arises from random matrix theory and quantum physics [Macchi, 1975, Kulesza and Taskar, 2012]. In machine learning, researchers have proposed different variations to improve its modeling capacity. Kulesza and Taskar [2011a] introduced k-DPP to restrict the sets to have a constant size k. Affandi et al. [2012] proposed a Markov DPP which offers diversity at adjacent time stamps. A structured DPP was presented in [Kulesza and Taskar, 2011c] to model trees and graphs. The MAP inference of DPP is generally NP-hard [Ko et al., 1995]. Gillenwater et al. [2012] developed an 1/4-approximation algorithm. In practice, greedy inference gives rise to decent results [Kulesza and Taskar, 2011b] though it lacks theoretical guarantees.

Another popular alternative is to resort to fast sampling algorithms [Kang, 2013, Kulesza and Taskar, 2012].

In spite of much research activity surrounding DPPs, there is very little work exploring how to effectively learn the model parameters. MLE is the most popular estimator. Compared to MLE, our approach is more robust to the number of training data or mis-specified models, and offers greater flexibility by incorporating customizable error functions. A recent Bayesian approach works with the posterior over the parameters [Affandi et al., 2014]. In contrast to that work, we develop a large-margin training approach for DPPs and directly minimize the set selection errors. The large margin principle has been widely used in classification [Vapnik, 1998] and structured prediction [Taskar et al., 2005, Tsochantaridis et al., 2004, Taskar et al., 2004, Sha and Saul, 2006], but its application to DPP is original. In order to make it tractable for DPPs, we use multiplicative rather than additive margin constraints.

5 EXPERIMENTS

We validate our large-margin approach to learn DPP parameters (DPP^{LME}) with extensive empirical studies on both synthetic data and two real-world summarization tasks with documents and videos. While DPP also has applications beyond summarization, this is a particularly good testbed to illustrate diverse subset selection: a compact summary ought to include high quality items that, taken together, offer good coverage of the source content.

5.1 SETUP

5.1.1 Evaluation Metrics

We evaluate the quality of the selected subset $\boldsymbol{y}^{\text{MAP}}$ against the ground-truth $\boldsymbol{y}^\star$ using the F-score, which is the harmonic mean of precision and recall:

$$\text{F-score} = \frac{2\,\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}},$$
$$\text{Precision} = \frac{|\boldsymbol{y}^{\text{MAP}} \cap \boldsymbol{y}^\star|}{|\boldsymbol{y}^{\text{MAP}}|}, \quad \text{Recall} = \frac{|\boldsymbol{y}^{\text{MAP}} \cap \boldsymbol{y}^\star|}{|\boldsymbol{y}^\star|}. \quad (15)$$

All three quantities are between 0 and 1, and higher values are better.

5.1.2 MAP Inference

We conduct the MAP inference of DPP by brute-forth search on the synthetic data, and turn to the so called minimum Bayes risk (MBR) decoding [Goel and Byrne, 2000, Kulesza and Taskar, 2012] for larger ground sets on real data.

The MBR inference samples subsets $\mathcal{S} = \{\boldsymbol{y}_1, \cdots, \boldsymbol{y}_\mathsf{T}\}$ from the learned DPP and outputs the one $\hat{\boldsymbol{y}}$ which achieves the highest consensus with the others, where the consensus can be measured by different evaluation metrics depending on applications. We use the F-score in our case. Particularly,

$$\hat{\boldsymbol{y}} \leftarrow \arg\max_{\boldsymbol{y}_{t'} \in \mathcal{S}} \frac{1}{\mathsf{T}} \sum_{t=1}^{\mathsf{T}} \text{F-SCORE}(\boldsymbol{y}_{t'}, \boldsymbol{y}_t). \quad (16)$$

Note that the MBR inference has actually introduced some degrees of flexibility to DPP (and to other probabilistic models). It allows users to infer the desired output according to different evaluation metrics. As a result, the selected subset is not necessarily the "true" diverse subset, but is biased towards the users' specific interests.

5.2 SYNTHETIC DATASET

5.2.1 Data

Our ground set has 10 items, $\mathcal{Y} = \{\boldsymbol{x}_1, \boldsymbol{x}_2, \cdots, \boldsymbol{x}_{10}\}$. For each item, we sample a 5-dimensional feature vector from a spherical Gaussian: $\boldsymbol{x}_i \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I})$. To generate the $\boldsymbol{L}$ matrix for the DPP, we follow the model in eq. (7); for the parameter vector $\boldsymbol{\theta}$ we sample from a spherical Gaussian, $\boldsymbol{\theta} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I})$, and for the similarity we simply let $\boldsymbol{\phi}_i = \boldsymbol{x}_i$ and compute $S_{ij} = \boldsymbol{\phi}_i^\mathsf{T} \boldsymbol{\phi}_j$.

We identify the most diverse subset $\boldsymbol{y}^\star$ (eq. (5)) via exhaustive search of all subsets, which is possible given the small ground set. The resulting $\boldsymbol{y}^\star$ has 5 items on average. We then add noise by randomly (with probability 0.1) adding or dropping an item to or from $\boldsymbol{y}^\star$. We repeat the process of sampling another pair of the ground set and its most diverse set. We do so 200 times and use 100 pairs for holdout and 100 for testing. We repeat the process to yield training sets of various sizes.

5.2.2 Learning

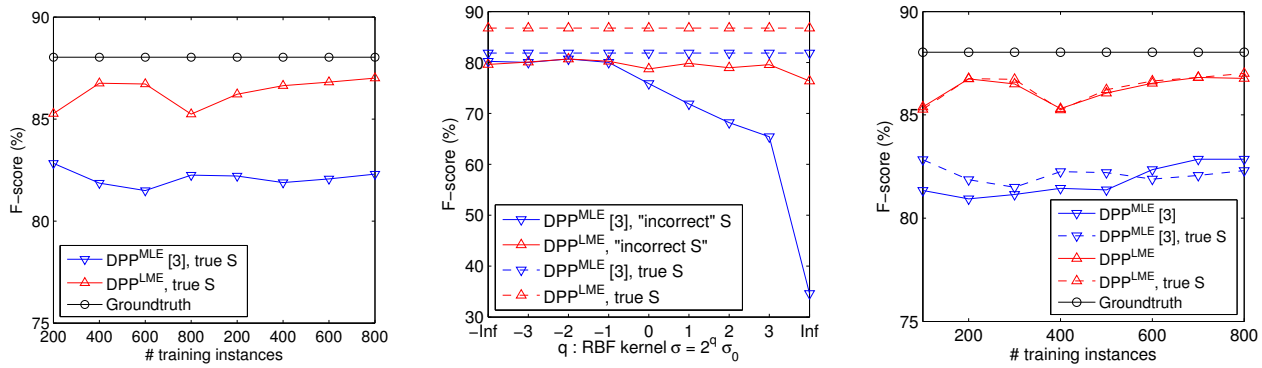
We compare our large-margin approach using the Hamming loss (eq. (10)) to the standard MLE method for learning DPP parameters.[1] All hyperparameters are tuned by cross-validation. After learning, we apply MAP inference to the testing ground sets.

5.2.3 Results

The DPP is parameterized by two things: $\boldsymbol{\theta}$ for the quality of the items, and S_{ij} for the similarity among them. Since the ground-truth parameters are known to us, we conduct experiments to isolate the impact of learning either one.

Fig. 1(a) contrasts the two methods when learning $\boldsymbol{\theta}$ only, assuming all S_{ij} are known and the ground-truths

[1] Adding a zero-mean Gaussian prior over $\boldsymbol{\theta}$ while learning with MLE, as in [Kulesza and Taskar, 2011b], did not yield improvement.



(a) Learning $\boldsymbol{\theta}$ only, with S_{ij} correctly specified

(b) Learning $\boldsymbol{\theta}$ under mis-specified S_{ij} (# training instances = 200)

(c) Learning both $\boldsymbol{\theta}$ and S_{ij}, with multiple kernel parameterization

Figure 1: On synthetic datasets, our method DPP$^{\text{LME}}$ significantly outperforms the state-of-the-art parameter estimation technique DPP$^{\text{MLE}}$ [Kulesza and Taskar, 2011b] in various learning settings. See text for details. Best viewed in color.

are used. Our DPP$^{\text{LME}}$ method significantly outperforms DPP$^{\text{MLE}}$. When the number of training samples is increased, the performance of our method generally improves and gets very close to the oracle's performance, for which the true values of *both* S_{ij} and $\boldsymbol{\theta}$ are used.

Fig. 1(b) examines the two methods in the setting of model mis-specification, where the S_{ij} values deliberately deviate from the true values. Specifically, we set them to $\exp(-\|\boldsymbol{x}_i - \boldsymbol{x}_j\|_2^2/\sigma^2)$ where the bandwidth σ varies from small to large, while the true values are $\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{x}_j$. All methods generally suffer. However, our method is fairly robust to the mis-specification while DPP$^{\text{MLE}}$ quickly deteriorates. Our advantage is likely due to our method's focus on learning to reduce subset selection errors, whereas MLE focuses on learning the right probabilistic model (even if it is already mis-specified).

Fig. 1(c) compares the two methods when both $\boldsymbol{\theta}$ and S_{ij} need to be learned from the data. We apply our multiple kernel parameterization technique to model S_{ij}, as in eq. (8), except β is set to be zero to avoid including the ground-truth. We see that our parameterization overcomes the problems of model mis-specification in Fig. 1(b), demonstrating its effectiveness in approximating unknown similarities. In fact, both learning methods match the performance of the corresponding methods with ground-truth similarity values, respectively. Nonetheless, our large-margin estimation still outperforms MLE significantly.

In summary, our results on synthetic data are very encouraging. Our multiple kernel parameterization avoids the pitfall of model mis-specification, and the large-margin estimation outperforms MLE due to its ability to track selection errors more closely.

5.3 DOCUMENT SUMMARIZATION

Next we apply DPP to the task of extractive multi-document summarization [Dang, 2005, Kulesza and Taskar, 2011b, Lin and Bilmes, 2010]. In this task, the input is a document cluster consisting of several documents on a single topic. The desired output is a subset of the sentences in the cluster that serve as a summary for the entire cluster. Naturally, we want the sentences in this subset to be both representative and diverse.

5.3.1 Experimental Setting

We use the text data from Document Understanding Conference (DUC) 2003 and 2004 [Dang, 2005] as the training and testing sets, respectively. There are 60 document clusters in DUC 2003 and 50 in DUC 2004, each collected over a short time period on a single topic. A cluster includes 10 news articles and on average 250 sentences. Four human reference summaries are provided along with each cluster. Following prior work, we generate the oracle/ground-truth summary by identifying a subset of the original sentences that best agree with the human reference summaries [Kulesza and Taskar, 2011b]. On average, the oracle summary consists of 5 sentences. As is standard practice, we use the oracles only during training. During testing, the algorithm output is evaluated against each of the four human reference summaries separately, and we report the average accuracy [Dang, 2005, Kulesza and Taskar, 2011b, Lin and Bilmes, 2010].

We use the widely-used evaluation package ROUGE [Lin, 2004], which scores document summaries based on n-gram overlap statistics. We use ROUGE 1.5.5 along with WordNet 2.0, and report the F-score (F), Precision (P), and Recall (R) of both unigram and bigram matchings, denoted by ROUGE-1X and ROUGE-2X respectively (X $\in$ {F, P, R}). Additionally, we limit the maximum length of each summary to be 665 characters to be consistent with existing work [Dang, 2005]. This yields 5 sentences on average for subsets generated by our algorithm.

To allow the fairest comparison to existing DPP work for

Table 1: Accuracy on document summarization. Our methods outperform others with statistical significance.

Method	ROUGE-1F	ROUGE-1P	ROUGE-1R	ROUGE-2F	ROUGE-2P	ROUGE-2R
PEER 35 [Dang, 2005]	37.54	37.69	37.45	8.37	–	–
PEER 104 [Dang, 2005]	37.12	36.79	37.48	8.49	–	–
PEER 65 [Dang, 2005]	37.87	37.58	38.20	9.13	–	–
DPP^{MLE}+COS [Kulesza and Taskar, 2011b]	37.89±0.08	37.37±0.08	38.46±0.08	7.72±0.06	7.63±0.06	7.83±0.06
Ours (DPP^{LME}+COS)	38.36±0.09	37.72±0.10	39.07±0.08	8.20±0.07	8.07±0.07	8.35±0.07
Ours (DPP^{MLE}+MKR)	39.14±0.08	39.03±0.09	39.31±0.09	9.25±0.08	9.24±0.08	9.27±0.08
Ours (DPP^{LME}+MKR)	**39.71**±0.05	**39.61**±0.08	**39.87**±0.06	**9.40**±0.08	**9.38**±0.08	**9.43**±0.08

this task, we use the same features designated in [Kulesza and Taskar, 2011b]. To model quality, the features are the sentence length, position in the original document, mean cluster similarity, LexRank [Erkan and Radev, 2004], and personal pronouns. To model the similarity, the features are the standard normalized term frequency-inverse document frequency (tf-idf) vectors.

5.3.2 Learning

We consider two ways of modeling similarities. The first one is to use the cosine similarity (COS) between feature vectors, as in [Kulesza and Taskar, 2011b]. The second is our multiple kernel based similarity (MKR, eq. (8)). For MKR, the bandwidths are $\sigma = 2^q$, $q = -6, -5, \cdots, 6$, and the combination coefficients are learned on the data. We implement the method in [Kulesza and Taskar, 2011b] as a baseline (DPP^{MLE}+COS). We also test an enhanced variant of that method by replacing its cosine similarity with our multiple kernel based similarity (DPP^{MLE}+MKR).

5.3.3 Results

Table 1 compares several DPP-based methods, as well as the top three results (PEER 35, 104, 65) from the DUC 2004 competition, which are not DPP-based ("-" indicates results not available). Since the DPP MAP inference is NP-hard, we use a sampling technique to extract the most diverse subset [Kulesza and Taskar, 2012]. We run inference 10 times and report the mean accuracy and standard error.

The state-of-the-art MLE-trained DPP model (DPP^{MLE}+COS) [Kulesza and Taskar, 2011b] achieves about the same performance as the best PEER results of DUC 2004. We obtain a noticeable improvement by applying our large-margin estimation (DPP^{LME}+COS). By applying multiple kernels to model similarity, we obtain significant improvements (above the standard errors) for both parameter estimation techniques. In particular, our complete method, DPP^{LME}+MKR, attains the best performance across all the evaluation metrics.

5.4 VIDEO SUMMARIZATION

Finally, we demonstrate the broad applicability of our method by applying it to video summarization. In this case, the goal is to select a set of representative and diverse frames from a video sequence.

5.4.1 Experimental Setting

The dataset consists of 50 videos from the Open Video Project (OVP)[2]. They are 30fps, 352×240 pixels, vary from 1 to 4 minutes, and are distributed across several genres including documentary, educational, historical, etc. We use the provided ground truth key frame summaries [de Avila et al., 2011], where each video is labeled by five annotators independently. We perform 5-fold validation and report the average result. We apply several preprocessing steps to remove frames that are trivially redundant (due to high temporal correlation) or of low visual quality. We use a similar procedure as in the document summarization task to generate the oracle/ground-truth subsets. On average, the ground-truth has 9 frames (in contrast, our method yields subsets from 5 to 20 frames). We use the public evaluation package VSUMM to evaluate the system-generated summary frames and again compute Precision, Recall and F-score [de Avila et al., 2011]. More details are in the supplementary material.

5.4.2 Features

We extract from each frame a color histogram and SIFT-based Fisher vector [Lowe, 2004, Perronnin and Dance, 2007] to model pairwise frame similarity S_{ij}. The two features are combined via our multiple kernel representation. To model the quality of each frame, we extract both intra-frame and inter-frame representativeness features. They are computed on the saliency maps [Rahtu et al., 2010, Hou et al., 2012] and include the mean, standard deviation, median, and quantiles of the maps as well as the the visual similarities between a frame and its neighbors. We z-score them within each video sequence.

5.4.3 Results

Table 2 compares several methods for selecting key frames: an unsupervised clustering method VSUMM [de Avila et al., 2011] (we implemented its two variants, offering a degree of tradeoff between precision and recall, and finely tuned the parameters), DPP^{MLE} with a multiple kernel parameterization of S_{ij}, and our margin-based approach. For

[2]The Open Video Project: www.open-video.org

Table 2: Accuracy on video summarization. Our method performs the best and allows precision-recall control.

Metric	VSUMM1 [de Avila et al., 2011]	VSUMM2 [de Avila et al., 2011]	DPP^{MLE}+MKR	Ours (DPP^{LME}+MKR)		
				$\omega = 1/64$	$\omega = 1$	$\omega = 64$
F-score	70.25	68.20	72.94±0.08	71.25±0.09	**73.46**±0.07	72.39±0.10
Precision	70.57	73.14	68.40±0.08	**74.00**±0.09	69.68±0.08	67.19±0.11
Recall	75.77	69.14	82.51±0.11	72.71±0.11	81.39±0.09	**83.24**±0.09

our method, we illustrate its flexibility to target different operating points, by varying the tradeoff constant ω in the generalized Hamming distance loss function eq. (11). Recall that higher values of ω will promote higher recall, while lower promote higher precision.

The results clearly demonstrate the advantage of our approach, particularly in how it offers finer control of the tradeoff between precision and recall. By adjusting ω, our method performs the best in each of the three metrics and outperforms the baselines by a statistically significant margin measured in the standard errors. Controlling the tradeoff is quite valuable in this application; for example, high precision may be preferable to a user summarizing a video he himself captured (he knows what appeared in the video, and wants a noise-free summary), whereas high recall may be preferable to a user summarizing a video taken by a third party (he has not seen the original video, and prefers some noise to dropped frames).

More details are illustrated in Fig. 2, in which by varying ω from 2^{-6} to 2^{8} we obtain 8 pairs of (precision, recall) values. We apply uniform interpolation among them and draw the precision-recall curve. One can see that DPP^{LME} is able to control the characteristics of the DPP generated summaries, biasing them to either high precision or high recall and without sacrificing the other too much. Though MLE or VSUMM does not supply such modeling flexibility, we also include them in the figure for reference.

We also present qualitative results on video summarization in Fig. 3. For this particular video, DPP^{MLE}, DPP^{LME} with $\omega = 1$, and DPP^{LME} with $\omega = 64$ all give rise to high recalls. Their output summaries are pretty lengthy, and may be boring to some users who just want to grasp something interesting to watch. By turning down the weight to $\omega = 1/64$, our DPP^{LME} dramatically improves the precision to 76% (in contrast to the 48% of DPP^{MLE}).

5.5 COMPUTATIONAL COMPLEXITY

The computational complexities of MLE and our large-margin approach are the same, $\mathcal{O}(\mathsf{N} \times \mathsf{D}^3)$, for computing the objective functions. Here N is the number of training instances, and D is the size of the largest ground set. Our softmax trick allows us to handle tractably an exponentially large number of constraints. Using gradient descent in parameter learning, the computational complexity in each iteration is thus also $\mathcal{O}(\mathsf{N} \times \mathsf{D}^3)$ for both methods. MLE is slightly faster (20% measured in wall-clock time).

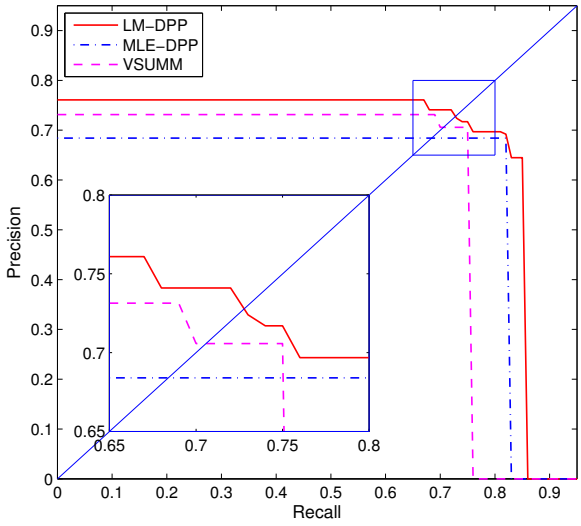


Figure 2: Balancing precision and recall. Through our large-margin DPPs (DPP^{LME}), we can balance precision and recall by varying ω in the generalized Hamming distance. In contrast, neither MLE nor VSUMM (the two variants in [de Avila et al., 2011] are plotted together) is readily able to support such flexibility.

6 CONCLUSION

The determinantal point process (DPP) offers a powerful and probabilistically grounded approach for selecting diverse subsets. We proposed a novel technique for learning DPPs from annotated data. In contrast to the status quo of maximum likelihood estimation, our method is more flexible in modeling pairwise similarity and avoids the pitfall of model mis-specification. Empirical results demonstrate its advantages on both synthetic datasets and challenging real-world summarization applications.

Acknowledgements

F. S., W. C., and B. G. are supported by ARO Award #W911NF-12-1-0241, DARPA Award #D11AP00278, NSF IIS Award #1065243, ONR #N000141210066, and Alfred P. Sloan Fellowship. K. G. is supported by ONR YIP #N00014-12-1-0754.

References

R. H. Affandi, A. Kulesza, and E. B. Fox. Markov determinantal point processes. In *UAI*, 2012.

Figure 3: Video summaries generated by DPP$^{\text{MLE}}$ and our DPP$^{\text{LME}}$ with $\omega = 1, \omega = 2^6$, and $\omega = 2^{-6}$, respectively. The oracle summary is also included for reference.

R. H. Affandi, E. B. Fox, and B. Taskar. Approximate inference in continuous determinantal point processes. In *NIPS*, 2013.

R. H. Affandi, E. B. Fox, R. P. Adams, and B. Taskar. Learning the parameters of determinantal point process kernels. In *ICML*, 2014.

R. Burton and R. Pemantle. Local characteristics, entropy and limit theorems for spanning trees and domino tilings via transfer-impedances. *The Annals of Probability*, pages 1329–1371, 1993.

H. T. Dang. Overview of duc 2005. In *Document Understanding Conf.*, 2005.

S. E. F. de Avila, A. P. B. Lopes, et al. Vsumm: A mechanism designed to produce static video summaries and a novel evaluation method. *Pattern Recognition Letters*, 32(1):56–68, 2011.

G. Erkan and D. R. Radev. Lexrank: Graph-based lexical centrality as salience in text summarization. *JAIR*, 22(1): 457–479, 2004.

J. Gillenwater, A. Kulesza, and B. Taskar. Near-optimal map inference for determinantal point processes. In *NIPS*, 2012.

V. Goel and W. J. Byrne. Minimum bayes-risk automatic speech recognition. *Computer Speech & Language*, 14 (2):115–135, 2000.

X. Hou, J. Harel, and C. Koch. Image signature: Highlighting sparse salient regions. *T-PAMI*, 34(1):194–201, 2012.

T. Jebara. *Machine learning: discriminative and generative*. Springer, 2004.

B. Kang. Fast determinantal point process sampling with application to clustering. In *NIPS*, 2013.

C.-W. Ko, J. Lee, and M. Queyranne. An exact algorithm for maximum entropy sampling. *Operations Research*, 43(4):684–691, 1995.

A. Kulesza and B. Taskar. k-dpps: Fixed-size determinantal point processes. In *ICML*, 2011a.

A. Kulesza and B. Taskar. Learning determinantal point processes. In *UAI*, 2011b.

A. Kulesza and B. Taskar. Structured determinantal point processes. In *NIPS*, 2011c.

A. Kulesza and B. Taskar. Determinantal point processes for machine learning. *Foundations and Trends® in Machine Learning*, 5(2-3):123–286, 2012.

C.-Y. Lin. Rouge: A package for automatic evaluation of summaries. In *Text Summarization Branches Out: Proc. of the ACL-04 Workshop*, 2004.

H. Lin and J. Bilmes. Multi-document summarization via budgeted maximization of submodular functions. In *NAACL/HLT*, 2010.

D. G. Lowe. Distinctive image features from scale-invariant keypoints. *IJCV*, 60(2):91–110, 2004.

O. Macchi. The coincidence approach to stochastic point processes. *Advances in Applied Probability*, 7(1):83–122, 1975.

A. Y. Ng and M. I. Jordan. On discriminative vs. generative classifiers: A comparison of logistic regression and naive bayes. In *NIPS*, 2002.

F. Perronnin and C. Dance. Fisher kernels on visual vocabularies for image categorization. In *CVPR*, 2007.

E. Rahtu, J. Kannala, M. Salo, and J. Heikkil. Segmenting salient objects from images and videos. In *ECCV*, 2010.

F. Sha and L. K. Saul. Large margin hidden markov models for automatic speech recognition. In *NIPS*, 2006.

B. Taskar, C. Guestrin, and D. Koller. Max-margin markov networks. In *NIPS*, 2004.

B. Taskar, V. Chatalbashev, D. Koller, and C. Guestrin. Learning structured prediction models: A large margin approach. In *ICML*, 2005.

I. Tsochantaridis, T. Hofmann, T. Joachims, and Y. Altun. Support vector machine learning for interdependent and structured output spaces. In *ICML*, 2004.

V. Vapnik. *Statistical learning theory. 1998*. Wiley, New York, 1998.