Determinantal Point Processes for Machine Learning
Determinantal Point Processes for Machine Learning

Alex Kulesza

University of Michigan
USA
kulesza@umich.edu

Ben Taskar

University of Pennsylvania
USA
taskar@cis.upenn.edu

now
the essence of knowledge

Boston – Delft
Editorial Scope

Foundations and Trends® in Machine Learning will publish survey and tutorial articles in the following topics:

- Adaptive control and signal processing
- Applications and case studies
- Behavioral, cognitive and neural learning
- Bayesian learning
- Classification and prediction
- Clustering
- Data mining
- Dimensionality reduction
- Evaluation
- Game theoretic learning
- Graphical models
- Independent component analysis
- Inductive logic programming
- Kernel methods
- Markov chain Monte Carlo
- Model choice
- Nonparametric methods
- Online learning
- Optimization
- Reinforcement learning
- Relational learning
- Robustness
- Spectral methods
- Statistical learning theory
- Variational inference
- Visualization

Information for Librarians
ISSN paper version 1935-8237. ISSN online version 1935-8245. Also available as a combined paper and online subscription.
Determinantal Point Processes for Machine Learning

Alex Kulesza\textsuperscript{1} and Ben Taskar\textsuperscript{2}

\textsuperscript{1} University of Michigan, USA, kulesza@umich.edu
\textsuperscript{2} University of Pennsylvania, USA, taskar@cis.upenn.edu

Abstract

Determinantal point processes (DPPs) are elegant probabilistic models of repulsion that arise in quantum physics and random matrix theory. In contrast to traditional structured models like Markov random fields, which become intractable and hard to approximate in the presence of negative correlations, DPPs offer efficient and exact algorithms for sampling, marginalization, conditioning, and other inference tasks. We provide a gentle introduction to DPPs, focusing on the intuitions, algorithms, and extensions that are most relevant to the machine learning community, and show how DPPs can be applied to real-world applications like finding diverse sets of high-quality search results, building informative summaries by selecting diverse sentences from documents, modeling nonoverlapping human poses in images or video, and automatically building timelines of important news stories.
## Contents

1 Introduction 1
1.1 Diversity 2
1.2 Outline 4

2 Determinantal Point Processes 7
2.1 Definition 8
2.2 L-ensembles 12
2.3 Properties 16
2.4 Inference 19
2.5 Related Processes 32

3 Representation and Algorithms 41
3.1 Quality versus Diversity 42
3.2 Expressive Power 44
3.3 Dual Representation 52
3.4 Random Projections 57
3.5 Alternative Likelihood Formulas 62

4 Learning 65
4.1 Conditional DPPs 65
4.2 Learning Quality 68
1

Introduction

Probabilistic modeling and learning techniques have become indispensable tools for analyzing data, discovering patterns, and making predictions in a variety of real-world settings. In recent years, the widespread availability of both data and processing capacity has led to new applications and methods involving more complex, structured output spaces, where the goal is to simultaneously make a large number of interrelated decisions. Unfortunately, the introduction of structure typically involves a combinatorial explosion of output possibilities, making inference computationally impractical without further assumptions.

A popular compromise is to employ graphical models, which are tractable when the graph encoding local interactions between variables is a tree. For loopy graphs, inference can often be approximated in the special case when the interactions between variables are positive and neighboring nodes tend to have the same labels. However, dealing with global, negative interactions in graphical models remain intractable, and heuristic methods often fail in practice.

Determinantal point processes (DPPs) offer a promising and complementary approach. Arising in quantum physics and random matrix
Introduction

theory, DPPs are elegant probabilistic models of global, negative correlations, and offer efficient algorithms for sampling, marginalization, conditioning, and other inference tasks. While they have been studied extensively by mathematicians, giving rise to a deep and beautiful theory, DPPs are relatively new in machine learning. We aim to provide a comprehensible introduction to DPPs, focusing on the intuitions, algorithms, and extensions that are most relevant to our community.

1.1 Diversity

A DPP is a distribution over subsets of a fixed ground set, for instance, sets of search results selected from a large database. Equivalently, a DPP over a ground set of \( N \) items can be seen as modeling a binary characteristic vector of length \( N \). The essential characteristic of a DPP is that these binary variables are negatively correlated; that is, the inclusion of one item makes the inclusion of other items less likely. The strengths of these negative correlations are derived from a kernel matrix that defines a global measure of similarity between pairs of items, so that more similar items are less likely to co-occur. As a result, DPPs assign higher probability to sets of items that are diverse; for example, a DPP will prefer search results that cover multiple distinct aspects of a user’s query, rather than focusing on the most popular or salient one.

This focus on diversity places DPPs alongside a number of recently developed techniques for working with diverse sets, particularly in the information retrieval community. However, unlike these methods, DPPs are fully probabilistic, opening the door to a wider variety of potential applications, without compromising algorithmic tractability.

The general concept of diversity can take on a number of forms depending on context and application. Including multiple kinds of search results might be seen as covering or summarizing relevant interpretations of the query or its associated topics; see Figure 1.1. Alternatively, items inhabiting a continuous space may exhibit diversity as a result of repulsion, as in Figure 1.2. In fact, certain repulsive quantum particles are known to be precisely described by a DPP; however, a DPP can also serve as a model for general repulsive phenomena, such
1.1 Diversity

Fig. 1.1 Diversity is used to generate a set of summary timelines describing the most important events from a large news corpus.

Fig. 1.2 On the left, points are sampled randomly; on the right, repulsion between points leads to the selection of a diverse set of locations.

Fig. 1.3 On the left, the output of a human pose detector is noisy and uncertain; on the right, applying diversity as a filter leads to a clean, separated set of predictions.

as the locations of trees in a forest, which appear diverse due to physical and resource constraints. Finally, diversity can be used as a filtering prior when multiple selections must be based on a single detector or scoring metric. For instance, in Figure 1.3 a weak pose detector favors large clusters of poses that are nearly identical, but filtering through a DPP ensures that the final predictions are well separated.

Throughout this survey we demonstrate applications for DPPs in a variety of settings, including:

- The DUC 2003/2004 text summarization task, where we form extractive summaries of news articles by choosing diverse subsets of sentences (Section 4.2.1);
Introduction

• An image search task, where we model human judgments of diversity for image sets returned by Google Image Search (Section 5.3),
• A multiple pose estimation task, where we improve the detection of human poses in images from television shows by incorporating a bias toward nonoverlapping predictions (Section 6.4), and
• A news threading task, where we automatically extract timelines of important news stories from a large corpus by balancing intra-timeline coherence with inter-timeline diversity (Section 6.6.4).

1.2 Outline

In this monograph we present general mathematical background on DPPs along with a range of modeling extensions, efficient algorithms, and theoretical results that aim to enable practical modeling and learning. The material is organized as follows.

Section 2 Determinantal Point Processes. We begin with an introduction to determinantal point processes tailored to the interests of the machine learning community. We focus on discrete DPPs, emphasizing intuitions and including new, simplified proofs for some theoretical results. We provide descriptions of known efficient inference algorithms and characterize their computational properties.

Section 3 Representation and Algorithms. We describe a decomposition of the DPP that makes explicit its fundamental trade-off between quality and diversity. We compare the expressive power of DPPs and MRFs, characterizing the trade-offs in terms of modeling power and computational efficiency. We also introduce a dual representation for DPPs, showing how it can be used to perform efficient inference over large ground sets. When the data are high-dimensional and dual inference is still too slow, we show that random projections can be used to maintain a provably close approximation to the original model while greatly reducing computational requirements.
Section 4: Learning. We derive an efficient algorithm for learning the parameters of a quality model when the diversity model is held fixed. We employ this learning algorithm to perform extractive summarization of news text.

Section 5: K-DPPs. We present an extension of DPPs that allows for explicit control over the number of items selected by the model. We show not only that this extension solves an important practical problem, but also that it increases expressive power: a k-DPP can capture distributions that a standard DPP cannot. The extension to k-DPPs necessitates new algorithms for efficient inference based on recursions for the elementary symmetric polynomials. We validate the new model experimentally on an image search task.

Section 6: Structured DPPs. We extend DPPs to model diverse sets of structured items, such as sequences or trees, where there are combinatorially many possible configurations. In this setting the number of possible subsets is doubly exponential, presenting a daunting computational challenge. However, we show that a factorization of the quality and diversity models together with the dual representation for DPPs makes efficient inference possible using second-order message passing. We demonstrate structured DPPs on a toy geographical paths problem, a still-image multiple pose estimation task, and two high-dimensional text threading tasks.
References


References


References


References


References


References


