Variable Elimination in Fourier Domain

Yexiang Xue
Cornell University
yexiang@cs.cornell.edu

Stefano Ermon
Stanford University
ermon@cs.stanford.edu

Ronan Le Bras, Carla P. Gomes, Bart Selman
Cornell University
{lebras, gomes, selman}@cs.cornell.edu

Abstract

Probabilistic inference is a key computational challenge in statistical machine learning and artificial intelligence. The ability to represent complex high dimensional probability distributions in a compact form is the most important insight in the field of graphical models.

In this paper, we explore a novel way to exploit compact representations of high-dimensional probability distributions in approximate probabilistic inference algorithms. Our approach is based on discrete Fourier Representation of weighted Boolean Functions, complementing the classical method to exploit conditional independence between the variables. We show that a large class of probabilistic graphical models have a compact Fourier representation. This theoretical result opens up an entirely new way of approximating a probability distribution. We demonstrate the significance of this approach by applying it to the variable elimination algorithm and comparing the results with the bucket representation and other approximate inference algorithms, obtaining very encouraging results.

1 Introduction

Probabilistic inference is a key computational challenge in statistical machine learning and artificial intelligence. Inference methods have an enormous number of applications, from learning models to making predictions and informing decision-making using statistical models of data. Unfortunately, the problem is computationally intractable, and standard exact inference algorithms, such as variable elimination and junction tree algorithms have worst-case exponential complexity.

The ability to represent complex high dimensional probability distributions in a compact form is the most important insight in the field of graphical models. The fundamental idea is to exploit (conditional) independencies between the variables to achieve compact *factored* representations, where a complex global model is represented as a product of simpler, local models. Similar ideas have been considered in the analysis of Boolean functions and logical forms [6], as well as in physics with low rank tensor decompositions and matrix product states representations [8, 14, 15, 23].

Compact representations are also key for the development of efficient inference algorithms, including message-passing ones. Efficient algorithms can be developed when messages representing the interaction among many variables can be decomposed or approximated with the product of several smaller messages, each involving a subset of the original variables. Numerous approximate and exact inference algorithms are based on this idea [6, 7, 18, 9, 24, 5, 13, 11].

Conditional independence (and related factorizations) is not the only type of structure that can be exploited to achieve compactness. For example, consider the regression tree in the left panel of Figure 1. No two variables in this probability distribution are independent with each other. Therefore,

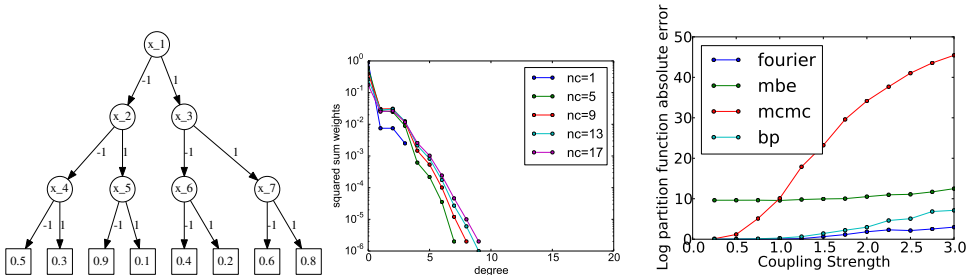

Figure 1: (Left) A decision tree representing a function $f : \{x_1, \ldots, x_7\} \rightarrow \mathcal{R}^+$, which cannot be represented by product of independent terms, but can be described exactly with 8 decision rules. (Middle) Weight concentration on low degree coefficients in the Fourier domain for random weighted 3-SAT instances with 20 variables and nc clauses. (Right) The absolute errors of log-partition function for 15×15 Ising Grids (Field strength 0.1). Fourier is for the Variable Elimination Algorithm in Fourier domain. mbe is for Mini-bucket Elimination. BP is for Belief Propagation. MCMC is using the classical Ogata-Tanemura scheme[20] to estimate the partition function.

the probability distribution cannot be represented by the product of simpler terms and requires a full probability table with 2^7 entries to be represented exactly. Nevertheless, this table can be described exactly by 8 simple decision rules, each corresponding to a path from the root to a leaf in the tree.

In this paper, we explore a novel way to exploit compact representations of high-dimensional probability tables in (approximate) probabilistic inference algorithms. Our approach is based on a (discrete) Fourier representation of the tables, which can be interpreted as a change of basis (see Table 1 for a small example). Crucially, tables that are dense in the canonical basis can have a sparse Fourier representation. Under certain conditions, probability tables can be represented (or well approximated) using a small number of Fourier coefficients. The Fourier representation has found numerous recent applications, including modeling stochastic processes [22, 1], manifolds [4], and permutations [12]. Our approach is based on Fourier representation on Boolean functions, which has found tremendous success in PAC learning [19, 17, 2, 3], but these ideas have not been fully exploited in the fields of probabilistic inference and graphical models.

In general, a general factor over n Boolean variables requires $O(2^n)$ entries to be specified, and similarly the corresponding Fourier representation is dense in general, i.e., it has $O(2^n)$ non-zero coefficients. However, a rather surprising fact which was first discovered by Linial [16] is that factors corresponding to fairly general classes of logical forms admit a compact Fourier representation. Linial discovered that formulas in Conjunctive (or Disjunctive) Normal Form (CNF or DNF) with bounded width (the number of variables in each clause) have compact Fourier representations.

In this paper, we introduce a novel approach for using approximate Fourier representations in the field of probabilistic inference. We generalize the work of Linial to the probability distributions (the weighted case where the entries are not necessarily 0 or 1), showing that a large class of probabilistic graphical models have compact Fourier representation, if they are composed with factors with bounded domain size and discriminative weights (see the middle panel of Figure 1). This includes interesting graphical models, such as Markov Logic Networks [21] with discriminative weights. The proof extends the famous Hastad's Switching Lemma[10] to the weighted case. At a high level, a compact Fourier representation often means the weighted probabilistic distribution can be captured by a small set of critical decision rules. Hence, this notion is closely related to decision trees with bounded depth.

Sparse (low-degree) Fourier representations provide an entirely new way of approximating a probability distribution. We demonstrate the power of this idea by applying it to the variable elimination algorithm and comparing the result with bucket representation and various approximate inference algorithms on a set of benchmarks, and obtain very encouraging results (see the right panel of Figure 1).

x	y	$\phi(x,y)$
-1	-1	f_1
-1	1	f_2
1	-1	f_3
1	1	f_4

$$\begin{aligned}\phi(x,y) = & \\ & \frac{1-x}{2}\cdot\frac{1-y}{2}\cdot f_1 \\ &+\frac{1-x}{2}\cdot\frac{1+y}{2}\cdot f_2 \\ &+\frac{1+x}{2}\cdot\frac{1-y}{2}\cdot f_3 \\ &+\frac{1+x}{2}\cdot\frac{1+y}{2}\cdot f_4\end{aligned} \qquad \begin{aligned}\phi(x,y) = & \\ & \frac{1}{4}(f_1+f_2+f_3+f_4) \\ &\frac{1}{4}(-f_1-f_2+f_3+f_4)x \\ &\frac{1}{4}(-f_1+f_2-f_3+f_4)y \\ &\frac{1}{4}(f_1-f_2-f_3+f_4)xy\end{aligned}$$

Table 1: (Left) Function $\phi : \{-1,1\}^2 \to \mathbb{R}$ is represented in a contingency table. (Middle) ϕ is re-written using interpolation. (Right) The terms of the equation in the middle is re-arranged, which yields the Fourier expansion of function ϕ. Mathematically, Every function $f : \{-1,1\}^n \to \mathbb{R}$ can be uniquely expressed as a multilinear polynomial, $f(\mathbf{x}) = \sum_{S\subseteq[n]} c_S \prod_{i\in S} x_i$, where each $c_S \in \mathbb{R}$. This representation is known as the Fourier expansion.

References

[1] Jaap H Abbring and Tim Salimans. The likelihood of mixed hitting times. Technical report, working paper, 2012.

[2] Avirim Blum, Carl Burch, and John Langford. On Learning Monotone Boolean Functions. In *FOCS*, pages 408–415, 1998.

[3] David Buchman, Mark W. Schmidt, Shakir Mohamed, David Poole, and Nando de Freitas. On sparse, spectral and other parameterizations of binary probabilistic models. In *AISTATS*, 2012.

[4] Taco Cohen and Max Welling. Harmonic exponential families on manifolds. In *Proceedings of the 32nd International Conference on Machine Learning, ICML*, 2015.

[5] Adnan Darwiche and Pierre Marquis. A knowledge compilation map. *J. Artif. Int. Res.*, 2002.

[6] Rina Dechter. Mini-buckets: A general scheme for generating approximations in automated reasoning. In *Proceedings of the Fifteenth International Joint Conference on Artificial Intelligence*, 1997.

[7] Natalia Flerova, Er Ihler, Rina Dechter, and Lars Otten. Mini-bucket elimination with moment matching. In *In NIPS Workshop DISCML*, 2011.

[8] Abram L. Friesen and Pedro Domingos. Recursive decomposition for nonconvex optimization. In *Proceedings of the 24th International Joint Conference on Artificial Intelligence*, 2015.

[9] Vibhav Gogate and Pedro M. Domingos. Structured message passing. In *UAI*, 2013.

[10] Johan Håstad. *Computational Limitations of Small-depth Circuits*. MIT Press, Cambridge, MA, USA, 1987.

[11] Tamir Hazan and Tommi S. Jaakkola. On the partition function and random maximum a-posteriori perturbations. In *ICML*, 2012.

[12] Jonathan Huang, Carlos Guestrin, and Leonidas J. Guibas. Fourier theoretic probabilistic inference over permutations. *Journal of Machine Learning Research*, 10:997–1070, 2009.

[13] Alexander T. Ihler, Natalia Flerova, Rina Dechter, and Lars Otten. Join-graph based cost-shifting schemes. In *UAI*, 2012.

[14] Michael I. Jordan, Zoubin Ghahramani, Tommi S. Jaakkola, and Lawrence K. Saul. An introduction to variational methods for graphical models. *Mach. Learn.*, 1999.

[15] Greg Linden, Brent Smith, and Jeremy York. Amazon.com recommendations: Item-to-item collaborative filtering. *IEEE Internet Computing*, 2003.

[16] Nathan Linial, Yishay Mansour, and Noam Nisan. Constant depth circuits, fourier transform, and learnability. *J. ACM*, 40(3), 1993.

[17] Yishay Mansour. Learning Boolean functions via the Fourier transform. *advances in neural computation and learning*, 0:1–28, 1994.

[18] Robert Mateescu, Kalev Kask, Vibhav Gogate, and Rina Dechter. Join-graph propagation algorithms. *J. Artif. Intell. Res. (JAIR)*, 37, 2010.

[19] Ryan O'Donnell. Some topics in analysis of boolean functions. *Proceedings of the fourtieth annual ACM symposium on Theory of computing - STOC 08*, page 569, 2008.

[20] Yosihiko Ogata and Masaharu Tanemura. Estimation of interaction potentials of spatial point patterns through the maximum likelihood procedure. *Annals of the Institute of Statistical Mathematics*, 1981.

[21] Matthew Richardson and Pedro Domingos. Markov logic networks. *Mach. Learn.*, 2006.

[22] L Chris G Rogers. Evaluating first-passage probabilities for spectrally one-sided lévy processes. *Journal of Applied Probability*, pages 1173–1180, 2000.

[23] David Sontag, Talya Meltzer, Amir Globerson, Tommi Jaakkola, and Yair Weiss. Tightening lp relaxations for map using message passing. In *UAI*, pages 503–510, 2008.

[24] Martin J. Wainwright, Tommi S. Jaakkola, and Alan S. Willsky. Tree-reweighted belief propagation algorithms and approximate ML estimation by pseudo-moment matching. In *Proceedings of the Ninth International Workshop on Artificial Intelligence and Statistics*, 2003.