Parameter Learning for Log-supermodular Distributions

Tatiana Shpakova
INRIA - École Normale Supérieure
Paris, France
Email: tatiana.shpakova@inria.fr

Francis Bach
INRIA - École Normale Supérieure
Paris, France
Email: francis.bach@inria.fr

***Abstract*—We consider log-supermodular models on binary variables, which are probabilistic models with negative log-densities which are submodular. These models provide probabilistic interpretations of common combinatorial optimization tasks such as image segmentation. In this paper, we focus primarily on parameter estimation in the models from known upper-bounds on the intractable log-partition function. We show that the bound based on separable optimization on the base polytope of the submodular function is always inferior to a bound based on "perturb-and-MAP " ideas. Then, to learn parameters, given that our approximation of the log-partition function is an expectation (over our own randomization), we use a stochastic subgradient technique to maximize a lower-bound on the log-likelihood. This can also be extended to conditional maximum likelihood. We illustrate our new results in a set of experiments in binary image denoising, where we highlight the flexibility of a probabilistic model to learn with missing data.**

I. Introduction

Submodular functions provide efficient and flexible tools for learning on discrete data. Several common combinatorial optimization tasks, such as clustering, image segmentation, or document summarization, can be achieved by the minimization or the maximization of a submodular function [1], [4], [8]. The key benefit of submodularity is the ability to model notions of diminishing returns, and the availability of exact minimization algorithms and approximate maximization algorithms with precise approximation guarantees [7].

In practice, it is not always straightforward to define an appropriate submodular function for a problem at hand. Given fully-labeled data, e.g., images and their foreground/background segmentations in image segmentation, structured-output prediction methods such as the structured-SVM may be used [10]. However, it is common (a) to have missing data, and (b) to embed submodular function minimization within a larger model. These are two situations well tackled by *probabilistic modelling*. This work has been published in the proceedings of the 2016 NIPS conference [9].

II. Log-supermodular models

We consider submodular functions on the vertices of the hypercube $\{0,1\}^D$. For any two vertices of the hypercube, $x, y \in \{0,1\}^D$, a function $f : \{0,1\}^D \to \mathbb{R}$ is submodular if $f(x) + f(y) \geqslant f(\min\{x,y\}) + f(\max\{x,y\})$, where the min and max operations are taken component-wise. We will only use the fact that f can be efficiently minimized on $\{0,1\}^D$, by a variety of generic algorithms, or by more efficient dedicated ones for subclasses such as graph-cuts. Log-supermodular models are introduced in [3] to model probability distributions on a hypercube, $x \in \{0,1\}^D$, and are defined as

$$p(x) = \frac{1}{Z(f)} \exp(-f(x)),$$

where $f : \{0,1\}^D \to \mathbb{R}$ is a submodular function such that $f(0) = 0$ and the partition function is $Z(f) = \sum_{x \in \{0,1\}^D} \exp(-f(x))$. It is more convenient to deal with the convex log-partition function $A(f) = \log Z(f) = \log \sum_{x \in \{0,1\}^D} \exp(-f(x))$. In general, the calculation of the partition function $Z(f)$ or the log-partition function $A(f)$ is intractable, as it includes simple binary Markov random fields—the exact calculation is known to be $\#P$-hard [6].

III. Upper-bounds on the log-partition function

Two upper bounds of log-partition function are presented:

- [3]: $A_{\text{L-field}}(f) = \min_{s \in B(f)} \sum_{d=1}^{D} \log\left(1 + e^{-s_d}\right)$, where $B(f)$ is a base polyhedron of $f(x)$, i.e.

  $$B(f) = \{s \in \mathbb{R}^D | s(\mathbb{1}) = f(\mathbb{1}), \forall x \in \{0,1\}^D : s(x) \leq f(x)\}.$$

- [5]: $A_{\text{Logistic}}(f) = \mathbb{E}_l\Big[\max_{y \in \{0,1\}^D} l^T y - f(y)\Big]$, where l is a logistic distributed random vector.
- We proved the inequality: $\boxed{\boldsymbol{A_{\text{Logistic}}(f) \leq A_{\text{L-field}}(f)}}$

IV. Parameter learning via maximum likelihood(MLL)

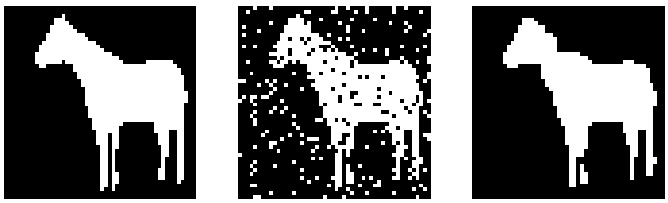
We introduce parameters governing the distribution. Set $\mathcal{F}$ is a considered family of submodular functions has the form $f(x) = \sum_{k=1}^{K} \alpha_k f_k(x) - t^T x$ and $\alpha \in \mathbb{R}_+^K$, $t \in \mathbb{R}^D$, $f_1, \ldots, f_K$ are submodular base functions. Our goal is to estimate the parameters from the sample $x_1, \ldots, x_N$ using MLL approach. Firstly, we prove that if we replace $A(f)$ by $A_{\text{L-field}}(f)$, we obtain a degenerate trivial solution. We were able to learn nonzero parameters using $A_{\text{Logistic}}(f)$ and a descent algorithm. Our second contribution is the use of stochastic gradient descent, not on the data as usually done, but on our own internal randomization.

A. Extension to conditional maximum likelihood

We consider a joint model over two binary vectors $x, z \in \mathbb{R}^D$, as follows $p(x, z|\alpha, t, \pi) = p(x|\alpha, t)p(z|x, \pi) = \exp(-f(x) - A(f)) \prod_{d=1}^{D} \pi_d^{\delta(z_d \neq x_d)} (1 - \pi_d)^{\delta(z_d = x_d)}$, which corresponds to sampling x from a log-supermodular model and considering z that switches the values of x with probability π_d for each d, that is, a noisy observation of x. Thus, if we observe both z and x, we can consider a conditional maximization of the log-likelihood (still a convex optimization problem), which we do in our experiments for supervised image denoising(Fig. 1), where we assume we know both noisy and original images at training time. Stochastic gradient on the logistic samples can then be used. As base submodular function horizontal and vertical cuts are used. The principal feature of a probabilistic approach is ability to deal with missing labels. While supervised learning can be achieved by other techniques such as structured-output-SVMs [10], [11], [12], our approach also applies when we do not observe the original image, which we consider in our experiments.

Acknowledgment

We acknowledge support the European Union's H2020 Framework Programme (H2020-MSCA-ITN-2014) under grant agreement n°642685 MacSeNet.



(a) original image (b) noisy image (c) denoised image

Fig. 1: Supervised denoising of a horse image from the Weizmann horse database [2].

REFERENCES

[1] F. Bach. Learning with submodular functions: a convex optimization perspective.*Foundations and Trends in Machine Learning*, 6(2-3):145 373, 2013.

[2] E. Borenstein, E. Sharon, and S. Ullman. Combining Top-down and Bottom-up Segmentation. In *Proc. ECCV*, 2004.

[3] J. Djolonga and A. Krause. From MAP to Marginals: Variational Inference in Bayesian Submodular Models. In *Adv. NIPS*, 2014.

[4] D. Golovin and A. Krause. Adaptive Submodularity: Theory and Applications in Active Learning and Stochastic Optimization. *Journal of Artificial Intelligence Research*, 42:427486,2011.

[5] T. Hazan and T. Jaakkola. On the Partition Function and Random Maximum A-Posteriori Perturbations. In *Proc. ICML*, 2012.

[6] M. Jerrum and A. Sinclair. Polynomial-time approximation algorithms for the Ising model. *SIAM Journal on Computing*, 22(5):10871116, 1993.

[7] Andreas Krause and Daniel Golovin. Submodular function maximization. In *Tractability:Practical Approaches to Hard Problems.* Cambridge University Press, February 2014.

[8] H. Lin and J. Bilmes. A class of submodular functions for document summarization. In *Proc. NAACL/HLT*, 2011.

[9] T. Shpakova and F. Bach. Parameter Learning for Log-supermodular Distributions. In *Adv. NIPS*, 2016.

[10] M. Szummer, P. Kohli, and D. Hoiem. Learning CRFs using graph cuts. In *Proc. ECCV*, 2008.

[11] B. Taskar, C. Guestrin, and D. Koller. Max-margin Markov networks. 2003.

[12] I. Tsochantaridis, Thomas Joachims, T., Y. Altun, and Y. Singer. Large margin methods for structured and interdependent output variables. *Journal of Machine Learning Research*, 6:14531484, 2005.