Parameter Learning for Log-supermodular Distributions

Tatiana Shpakova and Francis Bach

INRIA - École Normale Supérieure

28th December 2016

Second Christmas Colloquium on Computer Vision

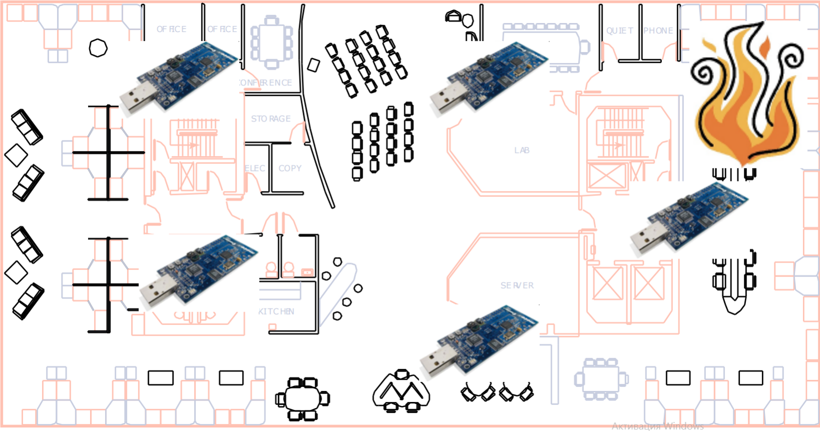
Slides credit Andreas Krause, Stefanie Jegelka

Outline

Document summarization

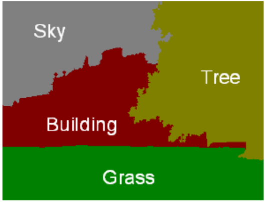
Goal: representative sentences selection

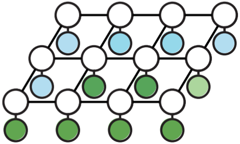
Sensor placement



Goal: place sensors to monitor temperature

MAP inference





$$\max_x p(x|z)$$

Goal: How find the MAP labeling in discrete graphical models efficiently?

Formalization

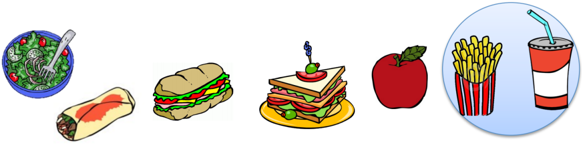
All these problems can be considered as an optimization of a set function $F(S)$, which is defined on subsets of some ground set V.

- General problem is **very hard**
- But structure can help!

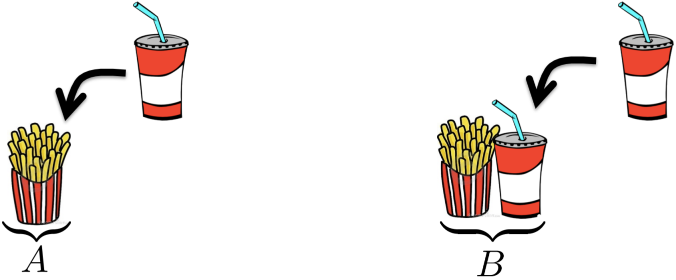
  If $F(S)$ is **submodular**, we have efficient optimization:

  - **Submodular Minimization** is computable in polynomial time.
  - Effective constant-factor approximation algorithms for **Submodular Maximization** exist.

Submodularity

$\mathcal{V} = \{$  $\}$

$\mathcal{V}$ - ground set. $F : 2^{\mathcal{V}} \to \mathbb{R}$ - set function.



$$A \subset B \Rightarrow F(A \cup s) - F(A) \geq F(B \cup s) - F(B)$$

diminishing marginal costs

From optimization to distributions

Instead of optimization, we take Bayesian approach:

$\min_x f(x) \Rightarrow P(x) = \frac{\exp(-f(x))}{\sum_{x \in D} \exp(-f(x))}$ - **log-supermodular distribution**

x lies in the power set D, e.g., the segmentation of an image.

Figure: Examples of $x \in \{0, 1\}^{200 \times 200}$

Example: binary pairwise Markov random fields (MRFs)

$$P(x_1, \ldots, x_n) = \frac{1}{Z} \prod_{i,j} \phi_{i,j}(X_i, X_j)$$

Proposed approach

$$P(x) = \frac{\exp(-f(x))}{\sum_{x \in D} \exp(-f(x))} = \frac{\exp(-f(x))}{Z(f)}.$$

Via log-supermodular model we can:

- learn parameters
- do inference
- quantify uncertainty about the solutions of optimization problem

Difficulty: Normalization constant $Z(f)$ is intractable when $|D|$ is huge.

Proposed approach

$$P(x) = \frac{\exp(-f(x))}{\sum_{x \in D} \exp(-f(x))} = \frac{\exp(-f(x))}{Z(f)}.$$

Via log-supermodular model we can:

- learn parameters
- do inference
- quantify uncertainty about the solutions of optimization problem

Difficulty: Normalization constant $Z(f)$ is intractable when $|D|$ is huge.

Solution: We can approximate the normalizer!

Upper bounds of partition function

- **State-of-the-art**

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

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

  The result was obtained by J. Djolonga and A. Krause.

Upper bounds of partition function

- **State-of-the-art**

  $A_{L-field} = \min\limits_{s \in B(f)} \sum\limits_{d=1}^{D} \log\left(1 + e^{-s_d}\right)$, where $B(f)$ is a base polyhedron of $f(x)$, i.e.

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

  The result was obtained by J. Djolonga and A. Krause.

- We took an upper bound of the normalizer for an abstract function $f(x)$ and investigate its properties under the assumption of submodularity. The bound for general function $f(x)$ was obtained by T. Hazan and T. Jaakkola: $A_{logistic} = \mathbb{E}_z\left[\max\limits_{y \in \{0,1\}^D} z^T y - f(y)\right]$, where z is a random vector consisting of independent logistic distributed random variables.

Upper bounds of partition function

- **State-of-the-art**

  $A_{L-field} = \min\limits_{s \in B(f)} \sum\limits_{d=1}^{D} \log\left(1 + e^{-s_d}\right)$, where $B(f)$ is a base polyhedron of $f(x)$, i.e.

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

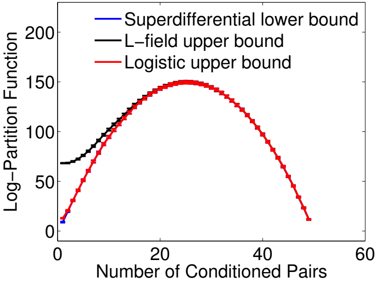
  The result was obtained by J. Djolonga and A. Krause.

- We took an upper bound of the normalizer for an abstract function $f(x)$ and investigate its properties under the assumption of submodularity. The bound for general function $f(x)$ was obtained by T. Hazan and T. Jaakkola: $A_{logistic} = \mathbb{E}_z\left[\max\limits_{y \in \{0,1\}^D} z^T y - f(y)\right]$, where z is a random vector consisting of independent logistic distributed random variables.

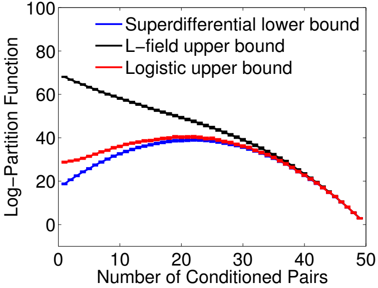
- **We proved** the following inequality:

  $$A_{logistic} \leq A_{L-field}$$

Example



(a) Mean bounds, c = 1



(b) Mean bounds, c = 3

We consider 2 Gaussian clusters and sample $n = 50$ points from each cluster. Graphcut function is used as submodular function $f(x)$.
Conditional distributions are considered:
one for each $k = 1, \ldots, n$, on the events that at least k points from the first cluster lie on the one side of the cut and at least k points from the second cluster lie on the other side of the cut.

Learning

We introduce parameters governing the distribution.
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.

Learning

We introduce parameters governing the distribution.
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.
Here some results are listed:

- Firstly, we tried to learn using $A_{L-field}$. Maximizing loglikelihood we obtain a linear function with constant coefficient:

$$\max_{\alpha \in \mathbb{R}_+^K} \sum_{k=1}^{K} \alpha_k \left[f_k\left(\sum_{n=1}^{N} x_n \right) - \sum_{n=1}^{N} f_k(x_n) \right].$$

Learning

We introduce parameters governing the distribution.
Family of submodular functions has the form:

$$f(x) = \sum_{k=1}^{K} \alpha_k f_k(x) - t^T x$$

and $\alpha \in \mathbb{R}_+^K$, $t \in \mathbb{R}^D$, $f_1, \dots, f_K$ are submodular base functions.
Here some results are listed:

- Firstly, we tried to learn using $A_{L-field}$. Maximizing loglikelihood we obtain a linear function with constant coefficient:

$$\max_{\alpha \in \mathbb{R}_+^K} \sum_{k=1}^{K} \alpha_k \left[f_k\Big(\sum_{n=1}^{N} x_n\Big) - \sum_{n=1}^{N} f_k(x_n) \right].$$

- **We show** how to learn using $A_{logistic}$ on the next slide.

Learning with $A_{logistic}$ via MaxLogLikelihood

We consider the following optimization problem:

$$\max_{t\in\mathbb{R}^D,\alpha\in\mathbb{R}_+^K} -\sum_{n=1}^{N}\sum_{k=1}^{K}\Big(\alpha_k f_k(x_n)\Big) + t^T\sum_{n=1}^{N}x_n - N\cdot A_{logistic}(\alpha, t),$$

where $A_{logistic} = \mathbb{E}_z\left[\max_{y\in\{0,1\}^D} z^T y - f(y)\right]$.

Learning with $A_{logistic}$ via MaxLogLikelihood

We consider the following optimization problem:

$$\max_{t\in\mathbb{R}^D,\alpha\in\mathbb{R}_+^K} -\sum_{n=1}^{N}\sum_{k=1}^{K}\Big(\alpha_k f_k(x_n)\Big) + t^T\sum_{n=1}^{N} x_n - N\cdot A_{logistic}(\alpha, t),$$

where $A_{logistic} = \mathbb{E}_z\Big[\max_{y\in\{0,1\}^D} z^T y - f(y)\Big]$.

1. Subgradient Descent

   In this case an empirical version of the $A_{logistic}$ bound is used:

$$A_{logistic} \approx \frac{1}{M}\sum_{m=1}^{M}\max_{y^m\in\{0,1\}^D}(z^m)^T y^m - f(y^m).$$

Learning with $A_{logistic}$ via MaxLogLikelihood

We consider the following optimization problem:

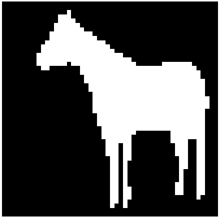
$$\max_{t\in\mathbb{R}^D,\alpha\in\mathbb{R}^K_+} -\sum_{n=1}^{N}\sum_{k=1}^{K}\Big(\alpha_k f_k(x_n)\Big) + t^T\sum_{n=1}^{N} x_n - N\cdot A_{logistic}(\alpha, t),$$

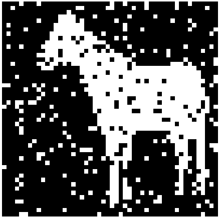
where $A_{logistic} = \mathbb{E}_z\left[\max_{y\in\{0,1\}^D} z^T y - f(y)\right]$.

1. Subgradient Descent
   In this case an empirical version of the $A_{logistic}$ bound is used:
   $$A_{logistic} \approx \frac{1}{M}\sum_{m=1}^{M}\max_{y^m\in\{0,1\}^D}(z^m)^T y^m - f(y^m).$$
2. Stochastic Gradient Descent
   On each iteration of gradient method we sample only one logistic vector z:
   $$A^h_{logistic} \approx \max_{y\in\{0,1\}^D} z^T y - f(y).$$

Supervised denoising

We consider the train sample of 100 binary images and the test sample of 100 binary images. We add some noise by flipping pixels values independently with the probability π.


a) original image


b) noised image


c) denoised image

noise π	max-marginals	mean-marginals	SVM-Struct
1%	0.4%	0.4%	0.6%
5%	1.1%	1.1%	1.5%
10 %	2.1%	2.0%	2.8%
20 %	4.2%	4.1%	6.0%

Unsupervised denoising

As training sample we consider only noisy images $z_1, \ldots, z_N$. We know a prior distribution of true images and the conditional distribution of noise:

$$p(x) = \frac{\exp(-f(x, \alpha, t))}{Z(\alpha, t)}, \quad p(z^i|x^i) = \begin{cases} x^i, \textbf{with } p, \\ \tilde{x}^i, \textbf{with } 1-p. \end{cases}$$

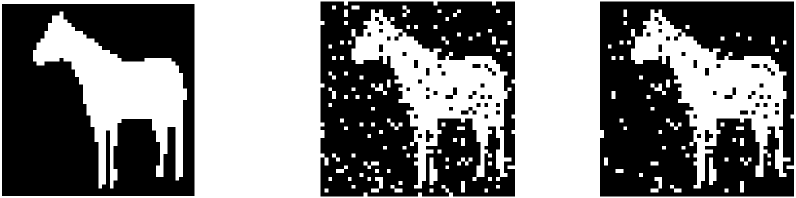
Let's consider the marginal loglikelihood of the observed data:

$$L(\alpha, t, z1, \ldots, z_n) = \sum_{n=1}^{N} \log p(z_n|\alpha, t) = \sum_{n=1}^{N} \log \sum_{x_n} p(x_n, z_n|\alpha, t) =$$

$$\sum_{n=1}^{N} \log \sum_{x_n} p(x_n)p(z_n|x_n) = \sum_{n=1}^{N} \log \sum_{x_n} e^{-f(x_n,\alpha,t)} p(z_n|x_n) - N \log Z(\alpha, t)$$

Experiments. Unsupervised case

We consider $N = 100$ train data (only noise images). After the learning procedure, we will be able to denoise train and test images.

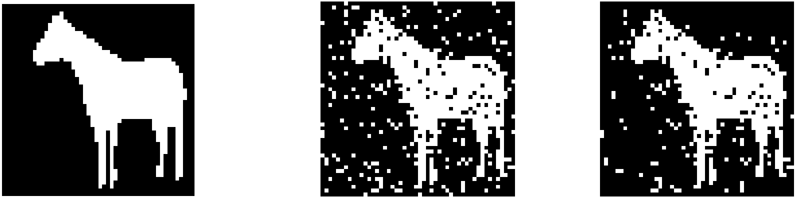


(a) original image (b) noisy image (c) denoised image

Figure: Denoising of a horse image. Unsupervised case.

Experiments. Unsupervised case

We consider $N = 100$ train data (only noise images). After the learning procedure, we will be able to denoise train and test images.



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

Figure: Denoising of a horse image. Unsupervised case.

	π is fixed		π is not fixed	
π	max-marg	mean-marg	max-marg	mean-marg
1%	0.5%	0.5%	1.0%	1.0%
5%	0.9%	1.0%	3.5%	3.6%
10%	1.9%	2.1%	6.8%	7.0%
20%	5.3%	6.0%	20.0%	20.0%

Our contribution

- We show that the **logistic bound** formally dominates a state-of-the-art bound [1].
- We demonstrate an impossibility of parameter learning via the existing state-of-the-art bound [1].
- We propose an automatic way to learn parameters using the **logistic bound**.
- We propose to use a stochastic subgradient technique over our own randomization during learning phase.
- We illustrate our new results on a set of experiments in binary image denoising (supervised and unsupervised problems).

This work has been accepted for **NIPS 2016**!

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

Future work

Exploration of larger-scale applications in computer vision:

- Foreground / Background segmentation (current work)
- Semantic multilabeled segmentation
- Interactive segmentation

Happy New Year!

Thank you!