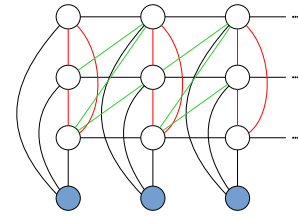


Variational Methods for Inference

based on a paper by Michael Jordan et al.

Patrick Pletscher
 ETH Zurich, Switzerland
 16th May 2006

The Need for Approximate Methods – FHMM



Inference

$$P(H|E) = \frac{P(H, E)}{P(E)}, \quad \text{complexity } \mathcal{O}(N^{M+1}T)$$

Overview

- 1 Motivation
- 2 Variational Methods
- 3 Discussion

Toy Example: $\ln(x)$

Idea of Variational Methods

Characterize a probability distribution as the solution of an optimization problem.

Intro: $\ln(x)$ variationally

Although no probability, still useful. Note $\ln(x)$ is a *concave* function.

$$\ln(x) = \min_{\lambda} \{\lambda x - \ln \lambda - 1\}$$

$\ln(x)$ now a linear function! Price: minimization has to be carried out for each x .

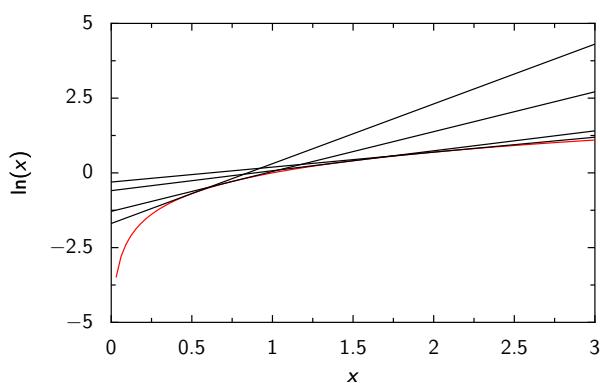
Upper bounds

For any given x , we have:

$$\ln(x) \leq \lambda x - \ln \lambda - 1,$$

for all λ .

Toy Example: $\ln(x)$



Convex Duality (1/2)

- 1 Transform function such that it becomes *convex or concave*. Transformation has to be *invertible*.
- 2 Calculate *conjugate function* (for concave function $f(\mathbf{x})$)

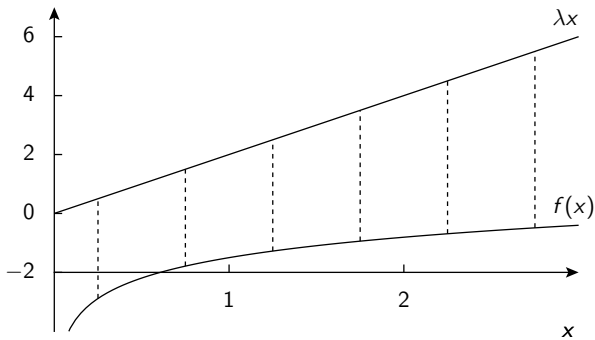
$$f(\mathbf{x}) = \min_{\lambda} \{\lambda^T \mathbf{x} - f^*(\lambda)\},$$

where

$$f^*(\lambda) = \min_{\mathbf{x}} \{\lambda^T \mathbf{x} - f(\mathbf{x})\}$$

- 3 Transform back.

Convex Duality (2/2)



Convex Duality and ln(x) Example

minimize:

$$\frac{d}{dx} \{ \lambda x - \ln(x) \} \stackrel{!}{=} 0,$$

we get

$$\lambda - \frac{1}{x} \stackrel{!}{=} 0 \rightarrow x = \frac{1}{\lambda}$$

Finally resubstitute:

$$f^*(\lambda) = \lambda \cdot \frac{1}{\lambda} + \ln \lambda = 1 + \ln \lambda$$

Which is the "magical" intercept of the ln example:

$$f(x) = \min_{\lambda} \{ \lambda x - \ln \lambda - 1 \}$$

Approximations using Convex Duality (1/2)

Basic idea

Simplify joint probability distribution by transforming the local probability functions. Usually only for "hard" nodes. Afterwards one can use *exact methods*.

This might look like this ...

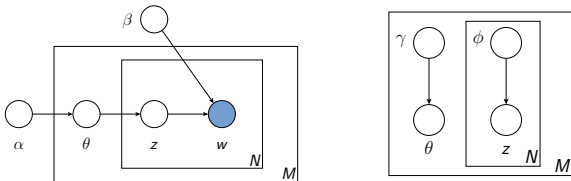


Figure: Replacing a difficult graphical model by a simpler one. Here for Latent Dirichlet Allocation.

Approximations using Convex Duality (2/2)

Joint Distribution

Product of upper bounds is an upper bound:

$$P(S) = \prod_i P(S_i | S_{\pi(i)}) \leq \prod_i P^U(S_i | S_{\pi(i)}, \lambda_i^U)$$

Marginalization

Upper bound for $P(E)$, the likelihood:

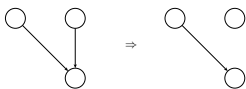
$$P(E) = \sum_{\{H\}} P(H, E) \leq \sum_{\{H\}} \prod_i P^U(S_i | S_{\pi(i)}, \lambda_i^U)$$

Sequential Approach

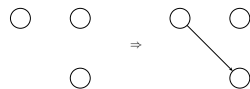
An unsupervised approach...

Algorithm transforms nodes, while needed. Backward-"elimination" popular as graph remains tractable.

Forward



Backward



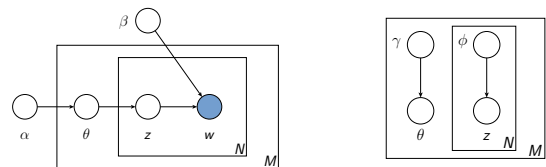
Discussion

- Flexible, out-of-the-box application,
- but: no "insider" knowledge is used.

Block Approach

A supervised approach...

Designate in advance which nodes are to be transformed.



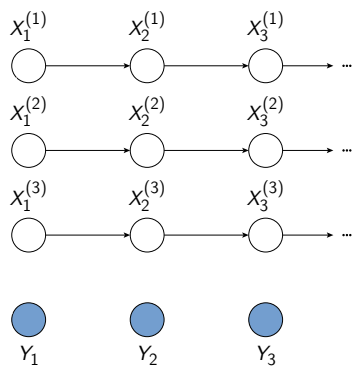
Minimize Kullback-Leibler Divergence

$$\lambda^* = \arg \min_{\lambda} D(Q(H|E, \lambda) || P(H|E)),$$

where

$$D(Q || P) := \sum_{\{S\}} Q(S) \ln \frac{Q(S)}{P(S)}$$

FHMM Variationally



Discussion: some pointers

Quite broad questions ...

- Does anybody know more about this new dependence, introduced by the optimization step?
- Any theoretical guarantees?
- Anybody already used variational methods? If so, for what? Experiences?

Junction Tree algorithm ...

- Translation from conditional probabilities to clique potentials?
- How do clique potentials change when we introduce the chords?