- HW4 due Wed
- Quiz n due Fri

COMPSCI 688: Probabilistic Graphical Models

Lecture 18: Variational Inference

Dan Sheldon

Manning College of Information and Computer Sciences
University of Massachusetts Amherst

Partially based on materials by Benjamin M. Marlin (marlin@cs.umass.edu) and Justin Domke (domke@cs.umass.edu)

Introduction

Models
BNs, MRFs
learning + inference

Approx. inference
MCMC
Variational inference < learning / models: VAEs

Additional
- GPs?
- normalizing flows

Variational Inference (VI) Overview

- Variational inference is an approximate inference approach (alternative to MCMC)
- Variational inference is at the core of a large family of techniques, **all of which start with the same mathematical idea**
  - mean-field and structured VI
  - black-box VI √
  - expectation maximization (EM) √
  - variational EM √
  - variational Bayes
  - variational auto-encoders √
  - loopy belief propagation and advanced message-passing algorithms

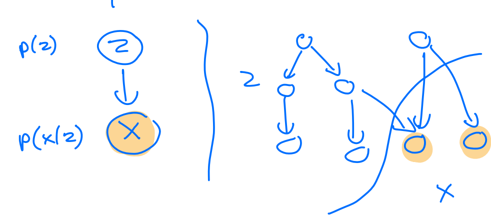
Problem Setting

$\tilde{p}(z)$ - can evaluate at any z

Assume we have an unnormalized probability model over z. Two examples:

1. Bayesian model $p(z|x)$ for latent z, observed x, unknown $p(x)$
2. Unnormalized model $p(z) = \frac{1}{Z}\tilde{p}(z)$ with unknown Z (e.g., loopy MRF)

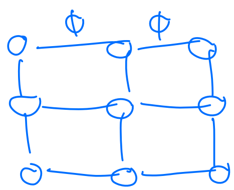
1. Bayesian



"hard" Want: $p(z|x) = \frac{p(z,x)}{p(x)}$ — $\tilde{p}(z)$, "easy", "hard"

usually $= \frac{p(z)p(x|z)}{p(x)}$

2. Unnormalized model

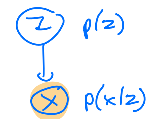


"hard"

$$p(z) = \frac{1}{Z}\tilde{p}(z) = \frac{1}{Z}\prod_{c\in\mathcal{C}}\phi_c(z_c)$$

"hard" "easy"

Problem Setting



For concreteness, henceforth we'll assume the Bayesian model setting:

- $p(z, x) = p(z)p(x|z)$ easy to compute
- We observe x, but not z
- We want to approximate

$$p(z|x) = \frac{p(z,x)}{p(x)}$$

unnormalized density $\tilde{p}(z)$

but don't know the normalization constant $p(x)$

General Strategy

- Gaussian
- diagonal covariance Gaussian
- dist over z w/ independent components

1. Let $q_\phi(z)$ be a "simple" distribution from some family with parameters ϕ
2. Try to optimize

"posterior"

$$\min_\phi \mathsf{KL}(q_\phi(z) \,\|\, p(z|x)) \qquad \text{("reverse KL")}$$

Then use $q_\phi(z)$ in place of $p(z|x)$

Why use VI?

- Can often get reasonable approximations faster than MCMC
  - get "whole" (approx.) dist instead of samples
- Gives a bound on $p(x)$ (or "Z"), useful for learning (more later)

Downside: "bias" due to simplifying assumption

The ELBO Decomposition

Big Idea: ELBO Decomposition

$z \rightarrow x$

$\tilde{p}(z) = p(z, x)$

This is the math trick that is at the heart of all VI methods:

$$\underbrace{\log p(x)}_{\text{const., “}\log Z\text{”}} = \underbrace{\sum_z q_\phi(z) \log \frac{p(z,x)}{q_\phi(z)}}_{\text{ELBO}(q_\phi(z) \,\|\, p(z,x))} + \underbrace{\sum_z q_\phi(z) \log \frac{q_\phi(z)}{p(z|x)}}_{\text{KL}(q_\phi(z) \,\|\, p(z|x))}$$

unnormalized

- ELBO: “**E**vidence **L**ower **BO**und” (will explain later)
- KL: what we want to minimize

Derivation

Claim:

$$\log p(x) = \sum_z q_\phi(z) \log \frac{p(z,x)}{q_\phi(z)} + \sum_z q_\phi(z) \log \frac{q_\phi(z)}{p(z|x)}$$

$p(z|x) = \frac{p(z,x)}{p(x)}$

Proof. Start with RHS and simplify:

$$\text{RHS} = \sum_z q_\phi(z) \left[\log p(z,x) - \cancel{\log q_\phi(z)} + \cancel{\log q_\phi(z)} - \log p(z|x) \right]$$

$$= \sum_z q_\phi(z) \left[\cancel{\log p(z,x)} - \cancel{\log p(z,x)} + \log p(x) \right]$$

$$= \sum_z q_\phi(z) \log p(x)$$

$$= \log p(x) \sum_z q_\phi(z)$$

$$= \log p(x)$$

ELBO Significance

$\boxed{\text{VI: max ELBO}(q_\phi \| p) \text{ wrt } \phi}$

$\tilde{p}(z)$, unnormalized

$$\underbrace{\log p(x)}_{\text{“evidence”}\ 20} = \underbrace{\sum_z q_\phi(z) \log \frac{p(z,x)}{q_\phi(z)}}_{8,\ \text{ELBO}(q_\phi(z) \,\|\, p(z,x))} + \underbrace{\sum_z q_\phi(z) \log \frac{q_\phi(z)}{p(z|x)}}_{12,\ \text{KL}(q_\phi(z) \,\|\, p(z|x))}$$

target “hard” (normalized)

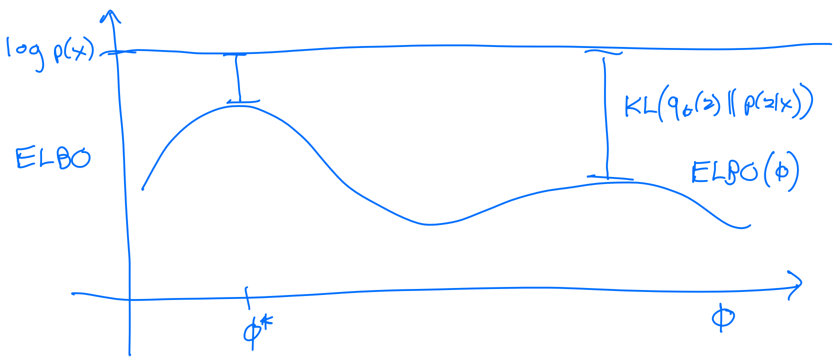
1. KL is “hard”: can't evaluate the *normalized* distribution $p(z|x)$
2. ELBO is “easy”(ish). Uses *unnormalized* distribution $p(z,x)$. Can often evaluate or approximate it, e.g., by Monte Carlo:

$$\text{sample } z^{(1)}, \ldots, z^{(N)} \sim q_\phi(z), \text{then compute } \frac{1}{N} \sum_{i=1}^N \log \frac{p(z^{(i)}, x)}{q_\phi(z^{(i)})}$$

3. KL is non-negative
4. Therefore $\log p(x) \geq$ ELBO (“Evidence lower bound”)
5. Therefore, choosing ϕ to maximize the ELBO **is the same** as choosing ϕ to minimize the KL (since $\log p(x)$ is constant with respect to ϕ)

ELBO Interpretation: Picture

$\phi \mapsto q_\phi$ Fixed $p(z,x)$



Variational Inference

Uses of VI

There are two different uses of VI

1. Approximate a posterior distribution: $\underbrace{p(z|x)}_{\text{target}} \approx \underbrace{q_\phi(z)}_{\text{simple}}$
2. Bound the log-likelihood: $\log p_\theta(x) \geq \text{ELBO}(q_\phi(z) \,\|\, p_\theta(z,x))$, usually in a learning procedure for $p_\theta(x)$ (details to come)

MLE: find θ to max $\log p_\theta(x)$

Basic VI Algorithm

1. **Input**: $\overset{\text{target}}{p(z,x)}$ and fixed x Want: $p(z|x)$
2. Choose some approximating family $q_\phi(z)$
3. Maximize ELBO$(q_\phi(z) \,\|\, p(x,z))$ wrt ϕ
4. Use $q_\phi(z)$ as a proxy for $p(z|x)$

Many choices for

- Target $p(z,x)$
- Approximating family q_ϕ
- How to estimate ELBO
- How to do optimization

ELBO Intuition

$= \sum_z q_\phi(z) \log \frac{p(z,x)}{q_\phi(z)}$

$$\text{ELBO} = \underbrace{\sum_z q_\phi(z) \log p(z,x)}_{\text{energy}} \boxed{- \underbrace{\sum_z q_\phi(z) \log q_\phi(z)}_{\text{entropy}}}$$

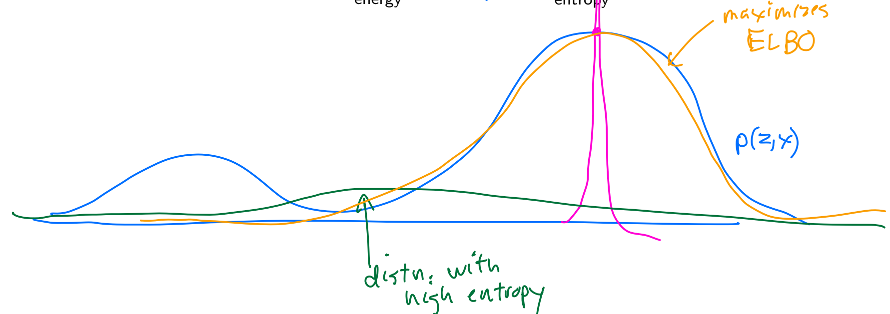
energy + entropy

- energy term encourages $q_\phi(z)$ to be high where $p(z|x)$ is high
- entropy term encourages $q_\phi(z)$ to be spread out

ELBO Intuition

$$\text{ELBO} = \underbrace{\sum_z q_\phi(z) \log p(z,x)}_{\text{energy}} \underbrace{- \sum_z q_\phi(z) \log q_\phi(z)}_{\text{entropy}}$$

+



Variational Learning

instead

MLE: find θ to max $\log p_\theta(\text{data})$

$z \to x$

$\max_\theta \log p_\theta(x)$

$\max \log \sum_z p_\theta(z,x)$

$\max_\theta \text{ELBO}(\theta) \le \log p_\theta(x)$

Expectation Maximization (EM): VI + Learning

EM is a classical algorithm for maximum-likelihood learning with latent variables

Goal: choose θ to maximize $\log p_\theta(x) = \log \sum_z p_\theta(z,x)$ given observed x

Usual lower-bound derivation

$\log p_\theta(x) = \log \sum_z p_\theta(z,x)$

$= \log \underbrace{\sum_z}_{\text{"f"}} \underbrace{q_\phi(z) \cdot}_{\text{"E"}} \frac{p_\theta(z,x)}{q_\phi(z)}$

$\ge \sum_z q_\phi(z) \log \frac{p_\theta(z,x)}{q_\phi(z)}$

$= \text{ELBO}$

(Jensen's inequality)

Learn θ

EM Algorithm. Init $\theta \mapsto p_\theta(z,x)$

- Set $q(z) = p_\theta(z|x)$ (maximize ELBO wrt q)
- Maximize $\underbrace{\sum_z q(z) \log \frac{p_\theta(x,z)}{q(z)}}_{\text{ELBO}}$ wrt θ
- Repeat

Gives local maximum of $\log p_\theta(x)$ wrt θ

Variational EM

$\max_q \mathrm{ELBO}(q(z) \,\|\, p(z,x)) \iff \min_q \mathrm{KL}(q(z) \,\|\, p(z|x))$

- It is not always possible or practical to compute $p_\theta(z|x)$ exactly in EM.
- Variational EM is an extension where the ELBO is maximized jointly with respect the the parameters ϕ of the approximating distribution and parameters θ of the model ("simultaneous inference and learning")

Goal: choose θ to maximize $\log p_\theta(x) = \log \sum_z p_\theta(z, x)$ given observed x.
~~fixed~~ $p_\theta(z,x)$
Define

$$\mathcal{L}(\phi, \theta) = \mathrm{ELBO}(q_\phi(z) \,\|\, p_\theta(z, x)) = \sum_z q_\phi(z) \log \frac{p_\theta(z, x)}{q_\phi(z)} \leq \log p_\theta(x)$$

then jointly optimize $\mathcal{L}(\phi, \theta)$ with respect to ϕ and θ, e.g.:

- (Stochastic) gradient ascent
- Alternating (partial) optimization steps $\longrightarrow$ Traditional EM
  1. max fully wrt $q \Rightarrow q(z) = p(z|x)$
  2. max wrt θ