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

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

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)

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

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

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

General Strategy

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

$$\min_\phi \mathsf{D}(q_\phi(z) \,\|\, p(z|x))$$

   where D is some "distance". Then use $q_\phi(z)$ in place of $p(z|x)$

Why use VI?

- Can often get reasonable approximations faster than MCMC
- Gives a bound on $p(x)$ (or "Z"), useful for learning (more later)

KL Minimization and ELBO

Idea: "Distance" Minimization

We want $q_\phi(z) \approx p(z|x)$.

Idea: define a "distance" $D(q_\phi(z) \| p(z|x))$ and choose ϕ to minimize it.

KL Divergence

A widely used "distance" between distributions is the Kullback-Leibler divergence:

$$\mathsf{KL}(q||p) = \int q(\mathbf{z}) \log\left(\frac{q(\mathbf{z})}{p(\mathbf{z})}\right) d\mathbf{z}$$

It is a *divergence* because it only satisfies some properties of a distance metric. It satisfies:

- $\mathsf{KL}(q||p) \geq 0$ for all q and p
- $\mathsf{KL}(q||p) = 0$ if and only if $q = p$

It does **not** satisfy:

- $\mathsf{KL}(q||p) = \mathsf{KL}(p||q)$ for all q, p
- $\mathsf{KL}(q||p) \leq \mathsf{KL}(q||s) + \mathsf{KL}(s||p)$ for all q, p, s

Big Idea: ELBO Decomposition

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

$$\log p(x) = \underbrace{\sum_z q_\phi(z) \log \frac{p(z,x)}{q_\phi(z)}}_{\mathsf{ELBO}\left(q_\phi(z) \,\|\, p(z,x)\right)} + \underbrace{\sum_z q_\phi(z) \log \frac{q_\phi(z)}{p(z|x)}}_{\mathsf{KL}\left(q_\phi(z) \,\|\, p(z|x)\right)}$$

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

Derivation

Claim:

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

Proof. Start with RHS and simplify:

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

ELBO Significance

$$\log p(x) = \underbrace{\sum_z q_\phi(z) \log \frac{p(z,x)}{q_\phi(z)}}_{\text{ELBO}\left(q_\phi(z) \,\|\, p(z,x)\right)} + \underbrace{\sum_z q_\phi(z) \log \frac{q_\phi(z)}{p(z|x)}}_{\text{KL}\left(q_\phi(z) \,\|\, p(z|x)\right)}$$

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:

   sample $z^{(1)}, \ldots, z^{(N)} \sim q_\phi(z)$, 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

Variational Inference

Uses of VI

There are two different uses of VI

1. Approximate a posterior distribution: $p(z|x) \approx q_\phi(z)$
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)

Basic VI Algorithm

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

Many choices for

- Approximating family q_ϕ
- How to estimate ELBO
- How to do optimization

ELBO Intuition

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

- 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

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

$$\begin{aligned}\log p_\theta(x) &= \log \sum_z p_\theta(x, z) \\ &= \log \sum_z q(z) \frac{p_\theta(x, z)}{q(z)} \\ &\geq \sum_z q(z) \log \frac{p_\theta(x, z)}{q(z)} \\ &= \text{ELBO}\end{aligned}$$

(Jensen's inequality)

EM Algorithm

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

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

Variational EM

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.
Define

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

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

- (Stochastic) gradient ascent
- Alternating (partial) optimization steps