Introduction

The ELBO Decomposition

Variational Inference

Variational Learning

COMPSCI 688: Probabilistic Graphical Models
Lecture 18: Variational Inference

Dan Sheldon
Manning College of Information and Computer Sciences
University of Massachusetts Amherst

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

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)

Want:

\[
p(z|x) = \frac{\tilde{p}(z|x)}{p(x)}
\]

Usually, have

\[
p(z|x) = p(z)p(x|z)
\]
Introduction

The ELBO Decomposition

Variational Inference

Variational Learning

Problem Setting

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

\[ p(z) = \frac{1}{Z} \hat{p}(z) \quad \text{(don’t know)} \]

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 \( \phi \)
2. Try to optimize

\[ \min_{\phi} KL(q_\phi(z) \| p(z|x)) \quad \text{("reverse KL")} \]

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

Why use VI?

- Can often get reasonable approximations faster than MCMC
- Gives a bound on \( p(x) \) (or “Z”), useful for learning (more later)
The ELBO Decomposition

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

**Big Idea: ELBO Decomposition**

This is the math trick that is at the heart of all VI methods:
\[
\log p(x) = \sum_z q_\phi(z) \log \frac{p(z, x)}{q_\phi(z)} + \sum_z q_\phi(z) \log \frac{q_\phi(z)}{p(z|x)}
\]

- ELBO: "Evidence Lower Bound" (will explain later)
- KL: what we want to minimize

**ELBO Significance**

\[
\text{ELBO} = \sum_z q_\phi(z) \log \frac{p(z, x)}{q_\phi(z)}
\]

- KL: what we want to minimize
- "evidence"

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 \( \log p(x) \geq \text{ELBO} \) ("Evidence lower bound")
5. Therefore, choosing \( \phi \) to maximize the ELBO is the same as choosing \( \phi \) to minimize the KL (since \( \log p(x) \) is constant with respect to \( \phi \))
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: \( \log p(x) \geq \text{ELBO}(q_\phi(z) \parallel p(x|z)) \), usually in a learning procedure for \( p_\theta(x) \) (details to come)
ELBO Intuition

\[ \text{ELBO} = \sum_z q_0(z) \log p(z, x) - \sum_z q_0(z) \log q_0(z) \]

- energy term encourages \( q_0(z) \) to be high where \( p(z|x) \) is high
- entropy term encourages \( q_0(z) \) to be spread out

Expectation Maximization (EM): VI + Learning

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

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

**Usual lower-bound derivation**

\[
\log p_\theta(x) = \log \sum_z q_\theta(z) \log \frac{p_\theta(z, x)}{q_\theta(z)} \\
\geq \sum_z q_\theta(z) \log \frac{p_\theta(z, x)}{q_\theta(z)} \\
= \sum_z q_\theta(z) \log p_\theta(z) - \sum_z q_\theta(z) \log q_\theta(z) - \sum_z q_\theta(z) \log q_\theta(z)
\]

(Jensen’s inequality) \( (\log \) is concave)
Variational EM

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

**Goal**: choose $\theta$ to maximize $\log p(\theta) = \log \sum_z p(\theta, x)$ given observed $x$.

Define

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

then jointly optimize $L(\phi, \theta)$ with respect to $\phi$ and $\theta$, e.g.:

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