Variational Inference

Aahlad

New York University

apm470@nyu.edu

November 17, 2016

Overview

Variational Inference

Variants of VI

Variational Inference

Variants of VI

What is Variational?

Generally, variational means formulating the estimator as the solution to an optimization problem.

Example:

$$\min_{x;\|x\|=1} x^T A x?$$

Counterpart to MCMC.

Why use VI?

- Posterior is impossile/intractable to compute.

Why use VI?

- Posterior is impossile/intractable to compute.
- Posterior is hard to sample from.

Why use VI?

- Posterior is impossile/intractable to compute.
- Posterior is hard to sample from.
- "Variational inference is that thing you implement while waiting for your Gibbs sampler to converge." - Blei
  BBVI in 5 lines of python.

Why use VI?

- Posterior is impossile/intractable to compute.
- Posterior is hard to sample from.
- "Variational inference is that thing you implement while waiting for your Gibbs sampler to converge." - Blei
  BBVI in 5 lines of python.
- All optimization tricks can now be used.

Why use VI?

- Posterior is impossile/intractable to compute.
- Posterior is hard to sample from.
- "Variational inference is that thing you implement while waiting for your Gibbs sampler to converge." - Blei
  BBVI in 5 lines of python.
- All optimization tricks can now be used.

Best feature: We know when it converges.

The Variational Family

Search over a family specially designed to make our lives easier. But not too easy.

The fundamental problem of VI: It's parametric. Generally the actual distribution is not in this. Hence, the tradeoff.

Example

Take the generaitve mode for a mixture of gaussians. Draw centers from a gaussian. Draw cluster assignments from multinomial. Draw samples from corresponding clusters. We have

$$p(\mathbf{x}, \mathbf{z}|\tau^2, \pi, \sigma^2) = \prod_i p(x_i|\mu_{z_i}, \sigma^2)p(z_i|\pi) \prod_k p(\mu_k|\tau^2)$$

What is a natural variational family to consider? We want z, μ_k. So,

$$q(z|\nu) = \prod_i q(z_i|\nu_{z_i}) \prod_k q(\mu_k|\nu_{\mu_k})$$

What's the problem here?

Example

Take the generaitve mode for a mixture of gaussians. Draw centers from a gaussian. Draw cluster assignments from multinomial. Draw samples from corresponding clusters. We have

$$p(\mathbf{x}, \mathbf{z}|\tau^2, \pi, \sigma^2) = \prod_i p(x_i|\mu_{z_i}, \sigma^2)p(z_i|\pi) \prod_k p(\mu_k|\tau^2)$$

What is a natural variational family to consider? We want z, μ_k. So,

$$q(z|\nu) = \prod_i q(z_i|\nu_{z_i}) \prod_k q(\mu_k|\nu_{\mu_k})$$

What's the problem here? First product is a Gaussian, second is a multinomial. What happens in the graphical model?

The Variational Objective.

Start with a distribution family $q(z)$, with the variational parameters ν.

The Variational Objective.

Start with a distribution family $q(z)$, with the variational parameters ν.

We now have

$$\hat{\mathcal{L}} = \mathbb{E}_q \log p(x, z) - \mathbb{E}_q \log q(z)$$

The negative of this is the Evidence Lower Bound (ELBO), which is derived from the log-likelihood. Maximizing the ELBO minimizes the KL-divergence. Initialize carefully and then maximize using coordinate ascent. This is vanilla VI.

KL-divergence

Very interesting interpretation from information theory. Used to compare loss-less encoding schemes.

In our case we look at it as comparing candidates against the optimal distribution. Two ways: **Forward and Reverse KL.**

The Mean Field Approximation.

Pick distributions however you want. The hidden variables are indepdent given only their hyperparameters. Start with

$$q(z|\nu) = \Pi_i q(z_i|\nu_i)$$

Intuitively, the effect of independent variables could approach the effect of dependent variables and that's enough.

Why mean field?

Makes the optimization problem easiser. Works decently in practice where you expect weak interactions between hidden variables.

Plus, with mean field, when the conditionals $p(z_j|z_{-j}, x)$ are in some exponential family, then the optimal approximation q is in the same exponential family. Which is awesome why?

But, one can come up with examples where mean field approximation is horrible. The simplest is an MRF over two hidden nodes.

Example

Variational inference on LDA. With
$p(x, z, \theta|\alpha, \beta) = \prod_d p(\theta|\alpha) \prod_n (z_{nd}|\theta_d) p(w_{nd}|z_{nd}, \beta)$.

The math is slightly hairy. We want z, θ, the hidden variables.
What is the Variational family? Split ν into γ, ϕ,:

$$q(z, \theta|\gamma, \phi) =$$

Example

Variational inference on LDA. With
$p(x, z, \theta|\alpha, \beta) = \prod_d p(\theta|\alpha) \prod_n (z_{nd}|\theta_d) p(w_{nd}|z_{nd}, \beta)$.

The math is slightly hairy. We want z, θ, the hidden variables.
What is the Variational family? Split ν into γ, ϕ,:

$$q(z, \theta|\gamma, \phi) = \prod_d q(\theta_d|\gamma) \prod_n q(z_{nd}|\phi_{nd})$$

Throw out all dependence of $z_n|\theta_d$! Still works!

Nice things about Exponential family.

Defined by canonical parameter η and sufficient statistic $t(x)$:

$$p(x|\eta) = \exp\{\langle \eta(x), t(x) \rangle - a(\eta)\}$$

$$\begin{aligned} \frac{\partial a(\eta)}{\partial \eta} &= \frac{1}{\exp\{a(\eta)\}} \frac{\partial \int \exp\{a(\eta)\}}{\partial \eta} \\ &= \frac{\int t(x) \exp\{\langle \eta(x), t(x) \rangle\} dx}{\int \exp\{\langle \eta(x), t(x) \rangle\} dx} \\ &= \mathbb{E}_p [t(X)] \end{aligned} \tag{1}$$

This will be used in expectations for LDA for the dirichlet dist.

Expectations for LDA.

Supressing hyper-parameters, we need $\mathbb{E}_q[p(x, z, \theta)]$. Note this is for a single document. We want to sample a θ and n topics.

$$\begin{aligned}
\mathbb{E}_q[\log p(\theta|\alpha)] &= \sum_t (\alpha_t - 1)\mathbb{E}_q[\log \theta_t] - \log \mathbf{B}(\alpha) \\
&= \sum_t (\alpha_t - 1)\left(\psi(\gamma_t) - \psi(\sum_d \gamma_t)\right) + \Gamma, \alpha \textit{ stuff} \\
\mathbb{E}_q[\log p(z|\theta)] &= \sum_n \sum_t \phi_{nt}\left(\psi(\gamma_t) - \psi(\sum_t \gamma_t)\right) \\
\mathbb{E}_q[\log p(w|z, \beta)] &= \sum_n \sum_t \sum_v \phi_{nt} w_{nv} \log \beta_{tv}
\end{aligned} \tag{2}$$

w_{nv} is an indicator function.

Expectations for LDA.

Supressing hyper-parameters, we need $\mathbb{E}_q[p(x, z, \theta)]$. Note this is for a single document. We want to sample a θ and n topics.

$$\begin{aligned}
\mathbb{E}_q[\log p(\theta|\alpha)] &= \sum_t (\alpha_t - 1)\mathbb{E}_q[\log \theta_t] - \log \mathbf{B}(\alpha) \\
&= \sum_t (\alpha_t - 1)\left(\psi(\gamma_t) - \psi(\sum_d \gamma_t)\right) + \Gamma, \alpha \textit{ stuff} \\
\mathbb{E}_q[\log p(z|\theta)] &= \sum_n \sum_t \phi_{nt}\left(\psi(\gamma_t) - \psi(\sum_t \gamma_t)\right) \\
\mathbb{E}_q[\log p(w|z, \beta)] &= \sum_n \sum_t \sum_v \phi_{nt} w_{nv} \log \beta_{tv}
\end{aligned} \tag{2}$$

w_{nv} is an indicator function. And two entropy terms which aren't as cool.

Expectations for LDA.

Pretty intuitive right? Now coordinate ascent. After all the calculus, one step would do the following.

$$\phi_{nt} \propto \beta_{tv} \exp\left(\psi(\gamma_t) - \psi(\sum_t \gamma_t)\right)$$

$$\gamma_t = \alpha_t + \sum_n \phi_{nt}$$

Notice that the coupling between all the hyperparameters.

Variational Inference contd.

We assumed we know the parameters of the true model.

Variational Inference contd.

We assumed we know the parameters of the true model.
Enter Variational EM: Add another maximization step:

$$\theta_{new} = \arg\max_{\theta} \hat{\mathcal{L}}(\nu_{new}, \theta)$$

In the LDA case, $\theta = (\alpha, \beta)$ and the objective is the sum of ELBO for each document.

Variational EM

The idea is really simple. Model both the actual posterior and the approximate posterior.

A natural generalization of the EM algorithm. The necessity is obvious if you take a look at the EM steps.

EM vs. Variational EM

EM is a special case of VI. Very easy to see once the equations are written down. Look at Prof. Bruna's slides.

In EM, we get pretty close to the actual distribution. But we need the assumption that the actual posterior is tractable.

In contrast, Variational Inference has the ability to approximate arbitrary distributions. The downside is that we impose additional independence on the approximate distribution. This could lead to extremely bad approximations.

Structured Mean Field

A natural extension to mean field in networks. Works much better in cases where the interactions captured are actually strong.

Trade-off: Much computation per update. But number of updates still remain about the same.

Generally, pick substructures that we can do exact inference over.

Variants of VI

Variants of VI

What can we improve?

What are the bottlenecks?

Stochastic VI

Can we improve the optimization? With larger data, per step computation increases. Can we learn something about global structure and improve updates? YES!

1. Natural Gradients. Moving in prob. space vs. decrease in KL.
2. + Stochastic Optimization. Paralellizing
3. + Update Global parameters. Feedback to global structure.

Guaranteed to converge to local optimum.

Black-Box VI

Don't want to do the math? "By the time you derive VI, the sampling algorithm will coverge."

Instead of giving update rules, make the program estimate the derivative. Estimate

$$\nabla_\nu \mathbb{E}_q(z|\nu)\left[\log p(x,z)\right] \sim \frac{1}{n}\sum_{i=1}^{n} \log p(x,z_i)\nabla_\nu q(z_i|\nu)$$

There's a lot of detail I left out(don't understand).

Thanks!

Asketh your questions.