Black Box Variational Inference

Rajesh Ranganath, Sean Gerrish, David M. Blei

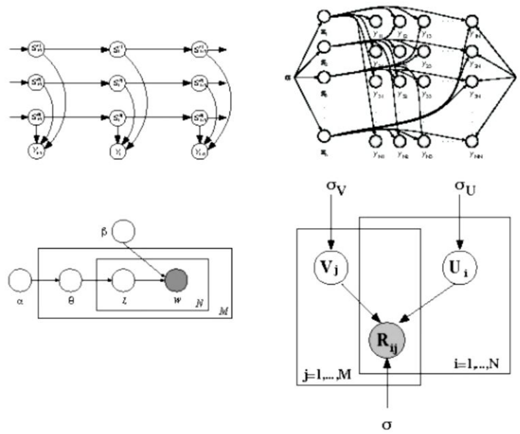
Presented by Dingcheng Hu

Problem

- Deriving a variational algorithm generally requires significant **model-specific** analysis
- These efforts hinder us from quickly developing and exploring a variety of models for a problem at hand.

Goal

Provide a **simple** way to approximate posteriors **quickly** in a **broad** class of latent variable models



Variational Inference

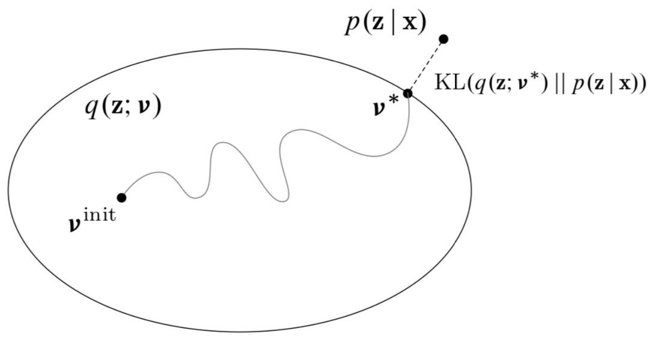
- A probabilistic model is a joint distribution of hidden variable z and observed variables x $p(\mathbf{z}, \mathbf{x})$
- Inference about the unknowns is through the posterior, the conditional distribution of the hidden variables given the observations

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

- For most interesting models, the denominator is not tractable. We appeal to approximate the posterior.

Variational Inference

- Variational inference solves posterior computation with optimization
- Define a variational family of distributions over the latent variables $q(\mathbf{z}; \boldsymbol{v})$
- Update the parameter to fit the variational distribution to be close (in KL) to the exact posterior



Variational Inference

- Optimization is over a family q
- Find q that minimizes KL(*q*||*p(z|x)*)
- But KL is intractable; VI optimizes the evidence lower bound (ELBO) instead:

$$\log p(x) \geq \mathbb{E}_{q_\lambda(z)}[\log p(x, z) - \log q(z)]$$

- **The first term** rewards variational distributions that place high mass on configurations of the latent variable that also explain the observations
- **The second term** encourages variational distribution to be diffuse
- Minimizing the KL is the same as Maximizing the ELBO

Problem

- ELBO:

$$\mathcal{L}(\lambda) = \mathbb{E}_{q_\lambda(z)}[\log p(x, z) - \log q(z)]$$

- Variational inference algorithms are normally derived by computing expectations and gradients
- Expectations often have no closed-form representation
- Computing the required expectations becomes intractable
- Require various expectation for each new model
- How do we get around this?

We want

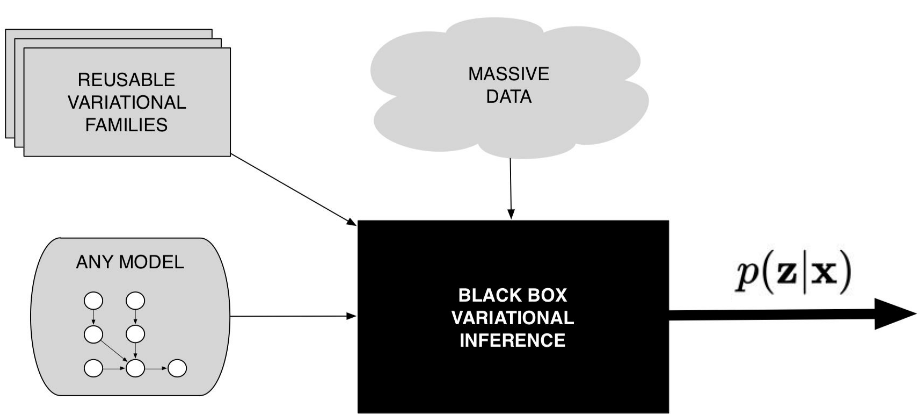
Black Box Variational Inference

Why do we want Black Box Variational Inference

- Easily use variational inference with **any model**
- Minimal mathematical work beyond specifying the model $p(\mathbf{z}, \mathbf{x})$

General Idea

- Avoid computing model specific expectations and gradients
- Construct noisy gradients by sampling the variational family



Stochastic Optimization

Stochastic optimization maximizes a function using noisy gradients of that function. Formally:

- Objective: f
- Parameters: x
- A noisy gradient H; $E[H] = \nabla f$
- Step size: ρ_t;
- Update: $x_{t+1} \leftarrow x_t + \rho_t H(x_t)$

Converges to a local optimum when:

$$\begin{aligned} \textstyle\sum_{t=1}^{\infty} \rho_t &= \infty \\ \textstyle\sum_{t=1}^{\infty} \rho_t^2 &< \infty \end{aligned}$$

A Noisy Gradient of the ELBO

- The ELBO: $\mathcal{L}(\lambda) = \mathbb{E}_{q(\lambda)}[\log p(x, z) - \log q(z)]$
- Gradient of the ELBO:

$$\frac{\partial \mathcal{L}}{\partial \lambda} = \mathbb{E}_q[\nabla_\lambda \log q(z|\lambda)(\log p(x, z) - \log q(z|\lambda))]$$

- How do we construct a noisy gradient of the ELBO?
- Using samples from q
- Noisy Gradient:

$$\frac{1}{S} \sum_{s=1}^{S} \nabla_\lambda \log q(z_s|\lambda)(\log p(x, z_s) - \log q(z_s|\lambda)),$$

$$\text{where } z_s \sim q(z|\lambda)$$

A Noisy Gradient of the ELBO

The noisy gradient:

$$\frac{1}{S}\sum_{s=1}^{S}\nabla_\lambda \log q(z_s|\lambda)(\log p(x, z_s) - \log q(z_s|\lambda)),$$

$$\text{where } z_s \sim q(z|\lambda)$$

To compute the noisy gradient of the ELBO we need

- Sampling from $q(z)$
- Evaluating $\nabla_\lambda \log q(z_s|\lambda)$
- Evaluating $\log p(x, z)$ and $\log q(z)$

A Noisy Gradient of the ELBO

The noisy gradient:

$$\frac{1}{S}\sum_{s=1}^{S}\nabla_\lambda \log q(z_s|\lambda)(\log p(x, z_s) - \log q(z_s|\lambda)),$$

$$\text{where } z_s \sim q(z|\lambda)$$

Fully black box

- Evaluating $\log p(x, z)$ is akin to defining the model
- The computations about q can be shared across models

Algorithm 1 Black Box Variational Inference

Input: data x, joint distribution p, variational family q.
Initialize $\lambda_{1:n}$ randomly, $t = 1$.
repeat
  // **Draw** S **samples from** q
  for $s = 1$ **to** S **do**
    $z[s] \sim q$
  end for
  $\rho =$ t-th value of a Robbins Monro sequence
  $\frac{\hat{\partial \mathcal{L}}}{\partial \lambda} = \frac{1}{S} \sum_{s=1}^{S} \nabla_\lambda \log q(z[s] | \lambda)(\log p(x, z[s]) - \log q(z[s] | \lambda))$
  $\lambda = \lambda + \rho \frac{\hat{\partial \mathcal{L}}}{\partial \lambda}$
  $t = t + 1$
until change of held out predictive likelihood is less than ϵ

Variance Control

- Theoretically this is guaranteed to converge to local optima.
- But unfortunately the variance of the estimator can be very high
- We have to control the variance of the gradient
- Two techniques
    - Rao-Blackwellization
    - Control variates

Intuition

- Variance reduction methods work by replace the function whose expectation is being approximated by Monte Carlo with another function that has the same expectation but smaller variance
- To estimate $\mathbf{E}_q[f]$ via Monte Carlo
- We compute the empirical average of $\hat{f}$

$$\mathrm{E}_q[f] = E_q[\hat{f}\,]$$

$$\mathbf{Var}_q[f] > \mathbf{Var}_q[\hat{f}\,]$$

Technique: Rao-Blackwellization

- Replace a random variable with its conditional expectation
- Try to estimate $\mathbb{E}[J(X,Y)]$ with Monte Carlo
- Compute $\mathbb{E}[J(X,Y)|X]$ then estimate via Monte Carlo
- Proof: Define $\hat{J}(X) = \mathrm{E}[J(X,Y)|X]$
- We have $\mathrm{E}[\hat{J}(X)] = \mathrm{E}[J(X,Y)]$
- $\mathrm{Var}(\hat{J}(X)) = \mathrm{Var}(J(X,Y)) - \mathrm{E}\left[(J(X,Y) - \hat{J}(X))^2\right]$

Rao-Blackwellize the ith component in a Black box manner

$$\frac{1}{S}\sum_{s=1}^{S} \nabla_{\lambda_i} \log q_i(z_s|\lambda_i)(\log p_i(x, z_s) - \log q_i(z_s|\lambda_i)),$$

$$\text{where } z_s \sim q_{(i)}(z|\lambda).$$

Technique: Control Variates

- Replace with $\hat{f}(z) \triangleq f(z) - a(h(z) - E[h(z)])$
- a is a scalar chosen to minimize the variance
- h is a function of our choice
- $E_q[\hat{f}] = \mathrm{E}_q[f]$
- $\mathrm{Var}(\hat{f}) = \mathrm{Var}(f) + a^2\,\mathrm{Var}(h) - 2a\,\mathrm{Cov}(f, h)$
- Good h have high correlation with f

But in **black box** variational inference

- Set h as $\nabla_{\lambda_i} \log q_i(z_s|\lambda_i)$
- Simply because $\mathbb{E}_q[\nabla_{\lambda_i} \log q_i(z_s|\lambda_i)] = 0$ for any q

Maintains black box nature of the algorithm

Variance Reduction

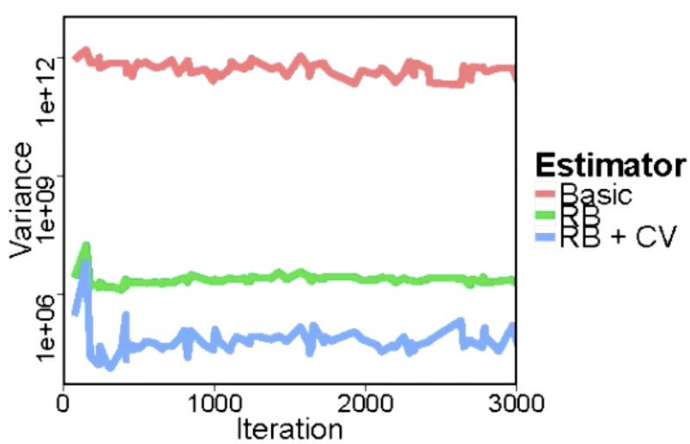


Figure: Variance comparison between our three estimators. The variance is reduced by several orders of magnitude.

Extension

- Set step size using AdaGrad
- Scalability by subsampling observations
  - Get a monte carlo estimate of $\log p(x, z)$

Experiment

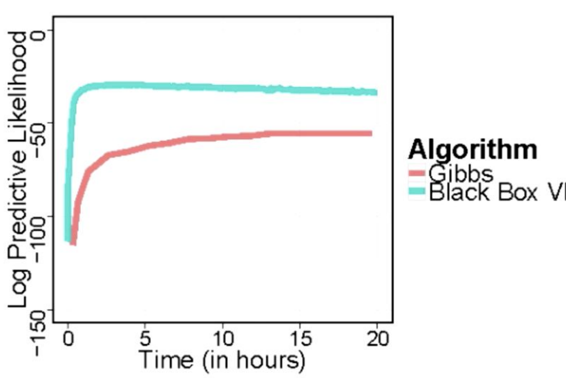


Figure: Comparison between Metropolis-Hastings within Gibbs and Black Box Variational Inference on a Gamma-Normal-TS model.

Summary

Black box variational inference only needs

- Log joint of the model $p(\mathbf{z}, \mathbf{x})$
- Computations of the variational approximation that can be shared across models