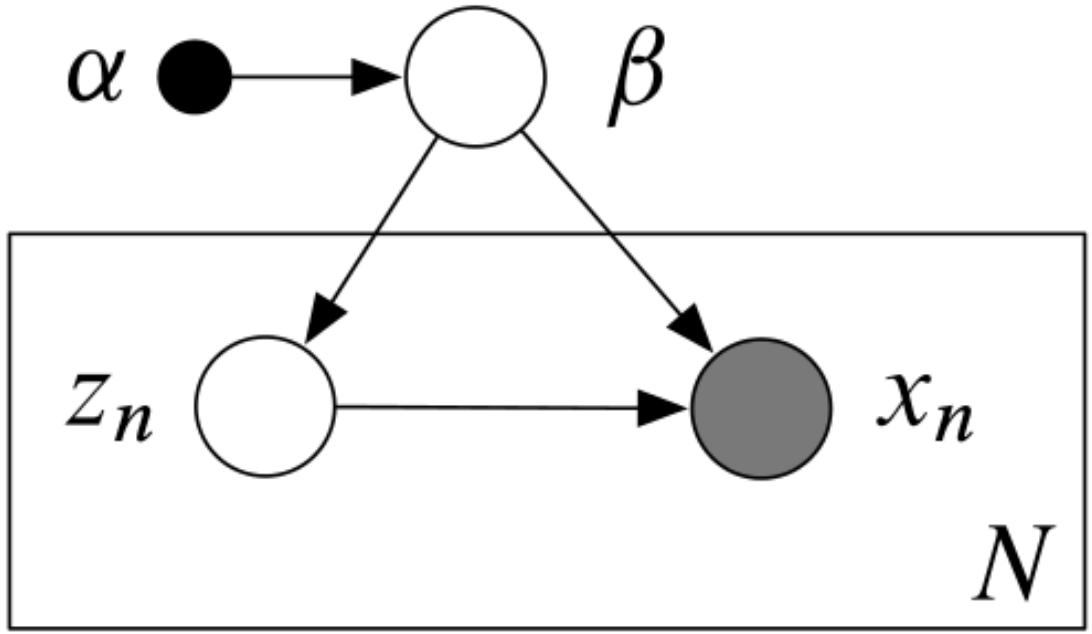
Stochastic Variational Inference

Michael Zhang

September 25, 2017

Introduction to Variational Inference

- Suppose we have a model with local variables, Z, and global variables β, similar to graphical model shown below (i.e. a Gaussian mixture model)



Introduction to Variational Inference

- Posterior distribution, $P(Z, \beta|X)$, is difficult to deal with, so approximate with simpler distribution $Q(Z, \beta)$.
- Criterion for optimal choice of Q? Kullback-Leibler divergence.
- Basic idea of Variational Inference? Minimize KL divergence (equivalently optimize evidence lower bound).
- Evidence Lower Bound:

$$\text{ELBO}(q) = E_Q\left[\log Q(Z, \beta)\right] - E_Q\left[\log P(X, Z, \beta)\right]$$

Introduction to Variational Inference

- We approximate the posterior from the following family of distributions:

$$Q(Z, \beta) = Q(\beta|\lambda) \prod_{n=1}^{N} \prod_{k=1}^{K} Q(z_{nk}|\phi_{nk})$$

  (called mean field assumption).
- λ and ϕ are the parameters of the variational approximation (called variational parameters).
- This task is easier with exponential family distributions on full conditionals, $P(z_n|-), P(\beta|-)$.

Exponential Family

- Recall exponential family distribution is written as:

$$P(X|\theta) = h(X) \exp\left\{\sum_i \eta_i(\theta) T_i(X) - A(\theta)\right\}$$

- $\eta(\theta)$ is called the natural parameter.
- If complete conditional is in exponential family and variational approximation is in same family then the optimal variational parameter is $E_Q[\eta(\cdot)]$, for natural parameter of complete conditional

Natural Gradients

- In typical gradient descent, we move with step size ρ in the direction of the gradient:

$$\lambda^{(t)} = \lambda^{(t-1)} + \rho \nabla f(\lambda)$$

- This is not necessarily the best method for optimizing parameters of probability distributions. Instead, moving according to "natural gradient" is better.
- Natural gradient for exponential families is

$$\hat{\nabla} f(\lambda) = E_Q[\eta(\cdot)] - \lambda$$

- Gradient update becomes $\lambda^{(t)} = (1-\rho)\lambda^{(t-1)} + \rho E_Q[\eta(\cdot)]$

Stochastic Variational Inference

- We can perform inference by updating ϕ_n for all $n = 1, \ldots, N$ and then updating λ.
- However, $\lambda^{(t)}$ depends on every value of $\phi^{(t)}$.
- This is inefficient. So we take noisy (but unbiased) estimates of gradients (hence "stochastic" part of name) by subsampling data and treating it like the full data set.

Stochastic Variational Inference

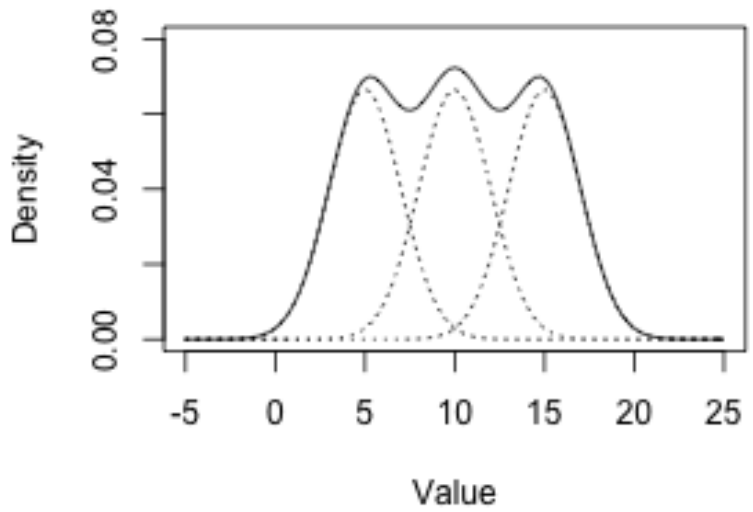
- Algorithm is as follows
  1. Select size M minibatch of data, $x_n, n = 1, \ldots, M$.
  2. For $n = 1, \ldots, M$, set $\phi_n = E_{Q(\lambda)}\left[\eta(\cdot)\right]$
  3. For $n = 1, \ldots, M$, compute $\hat{\lambda}_n = E_{Q(\phi)}\left[\eta(\cdot)\right]$ with estimate of gradient
  4. Update $\lambda^{(t)} = (1 - \rho_t)\lambda^{(t-1)} + \frac{\rho_t}{M}\sum_{n=1}^{M}\hat{\lambda}_n$

Example: Gaussian Mixture Model

- Suppose we have data from the following 1D mixture model with K components:

$$x_n \sim N(z_n^T \mu, \sigma^2),\ \mu_k \sim N(\mu_0, \tau^2),\ z_n \sim \text{Multinomial}\left(\frac{1}{K}, \ldots, \frac{1}{K}\right)$$

- Local variables are Z and global variables are μ_k.



Example: Gaussian Mixture Model

- Assume variational approximations:

$$Q(z_n|\phi_n) \sim \text{Multinomial}(\phi_{n,1}, \ldots, \phi_{n,k})$$
$$Q(\mu_k|m_k, s_k^2) \sim N(m_k, s_k^2)$$

- We need to fit variational parameters ϕ, m, s^2.

Example: Gaussian Mixture Model

- Obtain full conditional distribution of each parameter:

$$P(z_{n,k}|-) \propto \exp\left\{-\frac{\mu_k^2}{2\sigma^2} + \frac{\mu_k x_n}{\sigma^2}\right\}$$

- If $z \sim \text{Multinomial}(p_1, \ldots, p_K)$, natural parameter is $\eta(p_k) = \log p_k$.
- Thus, $E_{Q(\lambda)}\left[\eta\left(x_n^{(N)}, z_n^{(N)}\right)\right] = \frac{-m_k^2 - s^2}{2\sigma^2} + \frac{m_k x_n}{\sigma^2}$
- Reparameterize into multinomial form: $\phi_n \propto \exp\left\{E_{Q(\lambda)}\left[\eta\left(\cdot\right)\right]\right\}$

Example: Gaussian Mixture Model

- Raise likelihood of $P(x_n|z_n, \mu)$ to N/M-th power to obtain estimate of full likelihood
- Full conditional of $\hat{P}(\mu_k|-)$ is $N(\mu_x, \sigma_x^2)$

$$\sigma_x^2 = \left(\frac{1}{\tau^2} + \frac{N\sum_{n=1}^{M} z_{n,k}}{M\sigma^2}\right)^{-1}, \mu_x = \sigma_x^2\left(\frac{\mu_0}{\tau^2} + \frac{N\sum_{n=1}^{M} z_{n,k} \cdot x_n}{M\sigma^2}\right)$$

- Natural parameter for $N(\mu, \sigma^2)$ is $\left[\frac{\mu}{\sigma^2}, -\frac{1}{2\sigma^2}\right]$
- Variational update is:

$$\hat{m}_k = \left(\frac{\mu_0}{\tau^2} + \frac{N\sum_{n=1}^{M} \phi_k \cdot x_n}{M\sigma^2}\right), \hat{s}_k = -\frac{1}{2}\left(\frac{1}{\tau^2} + \frac{N\sum_{n=1}^{M} \phi_k}{M\sigma^2}\right)$$