MCMC and Variational Inference: Bridging the gap

Kolesov Alexander

Moscow Institute of Physics and Technology

01.12.2021

Introduction

- Bayes's Theorem:

$$p(\phi|x) = \frac{p(x|\phi)p(\phi)}{\int p(x|\phi)p(\phi)d\phi}$$

- Intractable denominator:

$$p(\phi|x) = \frac{p(x|\phi)p(\phi)}{Z(\phi)}$$

- Idea:
  - Ratio of probabilities
  - direction of gradient

Markov chains

- Definition of Homogeneous Markov's chain

$$p(x_1, ..., x_n) = p(x_n|x_{n-1})...p(x_2|x_1)p(x_0)$$

- Some intuition from looking for eigen vectors in Hilbert space
- Theorem of stationary distribution (Fixed Point Equation)

$$\int q(x^{'}|x)p(x)dx = p(x^{'})$$

- Simplest way to satisfy FPE - Detailed balance

$$q(x^{'}|x)p(x) = q(x|x^{'})p(x^{'})$$

- Aperiodic, irreducible , Geometric ergodicity

Metropolis - Hastings scheme

Algorithm

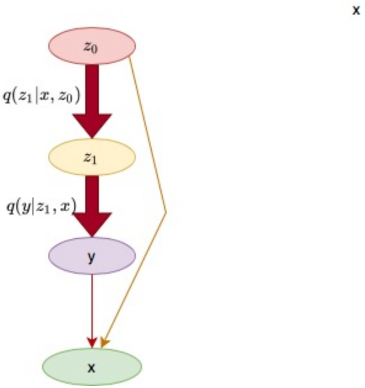
- $x_0 \sim \pi_0$
- $x^{'} \sim q(x^{'}|x)$
- $Acc = min(1, \frac{p(x^{'})q(x|x^{'})}{p(x)q(x^{'}|x)})$
- $\alpha \sim U[0,1]$
- if $\alpha < Acc$ accept, otherwise reject

- Problems:
  - Curse of Dimensionality
  - Accept-reject stochastic problem
  - Poor mode exploration
  - Proposal Distribution
  - Slow convergence
  - Small steps
  - Geometric of space
  - low ESS
  - Detailed Balance

Variational Inference

Usual deducing of Variational inference starts from MLE problem:

- $\log p(x|\phi) = \log p(x|\phi) \to \max_\phi$
-
- $\log p_\phi(x) \int q_\psi(z|x) dz = \int q_\psi(z|x) \log p_\phi(x) dz$
-
- $\int q_\psi(z|x) \log \frac{p_\phi(x,z) q_\psi(z|x)}{p_\phi(z|x) q_\psi(z|x)} dz = ELBO + KL$
-
- $ELBO = \int q_\psi(z|x) \log \frac{p_\phi(x,z)}{q_\psi(z|x)} dz =$
  $\int q_\psi(z|x) \log p_\phi(x,z) dz - \int q_\psi(z|x) \log q_\psi(z|x) dz$
-
- $\mathcal{L} = \mathbb{E}_{q_\psi(z|x)} \log p_\phi(x,z) - \mathbb{E}_{q_\psi(z|x)} \log q_\psi(z|x)$



- Vanilla VAE - Poor Prior
- $q(z_0|x)$: latent var from $\mathbb{R}^d$
- $q(z_1|z_0, x)$: $\mathbb{R}^D \to \mathbb{R}^d$
- $p(z|x)$ is necessary
- y $\sim p(z|x)$: true posterior
- $q(z|x) = q(z_0|x) \prod_{t=1}^{T} q(z_t|z_{t-1}, x)$
- z_{t-1} better, than $z_0 \sim \mathcal{N}(0, I)$
- $y = [z_0, .., z_{t-1}]$

Auxiliary variables

Lets expand previous notations for new auxiliary variables

- $\mathcal{L} = \mathbb{E}_{q_\psi(z|x)} \log p_\phi(x, z) - \mathbb{E}_{q_\psi(z|x)} \log q_\psi(z|x)$
-
- $p(x, z) \rightarrow p(x, y, z_t) = p(x, z_t) p(y|x, z_t)$
-
- $q(z|x) \rightarrow q(y, z_t|x)$
-
- $\mathcal{L}_{aux} = \mathbb{E}_{q(y, z_t|x)}[\log p(x, z_t) r(y|x, z_t) - \log q(y, z_t|x)]$

Formalisations

Let me tell more about these distributions

- $r(y|x, z_t) = \prod_{s=1}^{T} r_s(z_{s-1}|x, z_t)$ - Reversed kernel
-
- $L_{aux} = \mathbb{E}_{q(y,z_t|x)}[\log p(x, z_t) - \log q_t(z_0, ..., z_t|x) + \log r_t(z_0, .., z_{t-1}|x, z_t)]$
-
- $L_{aux} = \mathbb{E}_{q(y,z_t|x)}[\log \frac{p(x,z_t)}{q(z_0|x)} + \sum_{t=1}^{T} \log \frac{r_t(z_{t-1}|x,z_t)}{q_t(z_t|x,z_{t-1})}]$
-
- q_t is proposal of MCMC, r_t is like posterior

Optimizing the Lower Bound

Algorithm 1 MCMC lower bound estimate

Require: Model with joint distribution $p(x, z)$ and a desired but intractable posterior $p(z|x)$
Require: Number of iterations T
Require: Transition operator(s) $q_t(z_t|x, z_{t-1})$
Require: Inverse model(s) $r_t(z_{t-1}|x, z_t)$
  Draw an initial random variable $z_0 \sim q(z_0|x)$
  Initialize the lower bound estimate as
  $L = \log p(x, z_0) - \log q(z_0|x)$
  for $t = 1 : T$ **do**
    Perform random transition $z_t \sim q_t(z_t|x, z_{t-1})$
    Calculate the ratio $\alpha_t = \frac{p(x,z_t)r_t(z_{t-1}|x,z_t)}{p(x,z_{t-1})q_t(z_t|x,z_{t-1})}$
    Update the lower bound $L = L + \log[\alpha_t]$
  end for
  return the unbiased lower bound estimate L

Key Insight

Algorithm

Algorithm 2 Markov Chain Variational Inference (MCVI)

Require: Forward Markov model $q_\theta(z)$ and backward Markov model $r_\theta(z_0, \ldots, z_{t-1}|z_T)$
Require: Parameters θ
Require: Stochastic estimate $L(\theta)$ of the variational lower bound $\mathcal{L}_{\text{aux}}(\theta)$ from Algorithm 1
  while not converged **do**
    Obtain unbiased stochastic estimate $\hat{g}$ with $E_q[\hat{g}] = \nabla_\theta \mathcal{L}_{\text{aux}}(\theta)$ by differentiating $L(\theta)$
    Update the parameters θ using gradient $\hat{g}$ and a stochastic optimization algorithm
  end while
  return final optimized variational parameters θ

Auxiliary Variables MCMC

- Curse of Dimensionality
- Continious or discrete
- One can recover p(x) as

$$p(x) = \int \tilde{q}_\phi(x|a)\hat{q}(a)da$$

- $q(z|x) = q(z_0|x)\prod_{t=1}^{T} q(z_t|z_{t-1}, x)$
- $\hat{q}_\lambda(a)$ might be learned