Stochastic Variational Inference

Reza Babanezhad

rezababa@cs.ubc.ca

Outline

- VI
- Monte Carlo Gradient Approximation
- Stochastic Variational Inference(SVI)
- Bridging the GAP

VI

- VI can be used to approximate the posterior distribution
- Objective is minimizing the KL divergence between the approximate *q* and joint distribution *p*

$$\log p(x) = \mathbb{E}_{q_\theta(z|x)}[\log p(x,z) - \log q_\theta(z|x)] + D_{KL}(q_\theta(z|x)||p(z|x))$$
$$\geq \mathbb{E}_{q_\theta(z|x)}[\log p(x,z) - \log q_\theta(z|x)] = \mathcal{L}.$$

- To optimize ELBO, we can use the coordinate descent or ascent.
- Problems:
  - Computing the gradient of expectation
  - In each iteration, we need to go over all data.

Monte Carlo Gradient Approximation

- In ELBO some expectations cannot be computed in closed form.
- To solve it, let divide it to to part: $\underline{\mathcal{L}} = \mathbb{E}_q[f] + h(X, \Psi)$
  - h : closed form part.
  - f : its expectation does not have closed form
- The gradient: $\nabla_\psi \mathcal{L} = \nabla_\psi \mathbb{E}_q[f(\theta)] + \nabla_\psi h(X, \Psi)$
- The first term in RHS is intractable.
- Goal: finding a Monte Carlo approximation for intractable term.

Monte Carlo Gradient Approximation

$$\begin{aligned}\nabla_\psi \mathbb{E}_q[f(\theta)] &= \nabla_\psi \int_\theta f(\theta) q(\theta|\psi) d\theta \\ &= \int_\theta f(\theta) \nabla_\psi q(\theta|\psi) d\theta \\ &= \int_\theta f(\theta) q(\theta|\psi) \nabla_\psi \ln q(\theta|\psi) d\theta.\end{aligned}$$

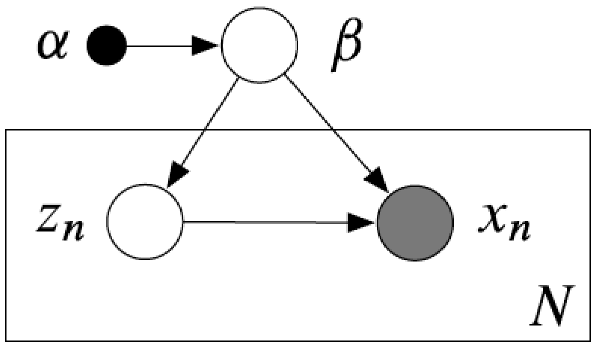
$$\nabla_\psi \mathbb{E}_q[f(\theta)] = \mathbb{E}_q[f(\theta) \nabla_\psi \ln q(\theta|\psi)]$$

$$\nabla_\psi \mathbb{E}_q[f(\theta)] \approx \frac{1}{S} \sum_{s=1}^{S} f(\theta^{(s)}) \nabla_\psi \ln q(\theta^{(s)}|\psi), \quad \theta^{(s)} \overset{iid}{\sim} q(\theta|\psi)$$

$$\psi^{(t+1)} = \psi^{(t)} + \rho_t \nabla_\psi h(X, \Psi^{(t)}) + \rho_t \zeta_t \quad \zeta_t = \nabla_{\psi_t} \mathbb{E}_q[f(\theta)]$$

SVI

- Model



$$z_n = z_{n,1:J}$$

$$p(x, z, \beta \mid \alpha) = p(\beta \mid \alpha) \prod_{n=1}^{N} p(x_n, z_n \mid \beta)$$

- Our goal: approximate the posterior

$$p(\beta, z \mid x)$$

- Locally independence

$$p(x_n, z_n \mid x_{-n}, z_{-n}, \beta, \alpha) = p(x_n, z_n \mid \beta, \alpha).$$

SVI

- Extra assumption
  - posterior is from exponential family

$$p(\beta \mid x, z, \alpha) = h(\beta) \exp\{\eta_g(x, z, \alpha)^\top t(\beta) - a_g(\eta_g(x, z, \alpha))\},$$

$$p(z_{nj} \mid x_n, z_{n,-j}, \beta) = h(z_{nj}) \exp\{\eta_\ell(x_n, z_{n,-j}, \beta)^\top t(z_{nj}) - a_\ell(\eta_\ell(x_n, z_{n,-j}, \beta))\}.$$

  - h: base measure
  - t: sufficient statistics
  - η: natural parameter
  - a: partition function or log normalizer

SVI

- Conjugacy relation between the global variable and local variable

$$p(x_n, z_n|\beta) = h(x_n, z_n)\exp\{\beta^\top t(x_n, z_n) - a_\ell(\beta)\}.$$

- Prior of global variable is also exponential

$$p(\beta) = h(\beta)\exp\{\alpha^\top t(\beta) - a_g(\alpha)\}$$

- Posterior

$$p(z, \beta | x) = \frac{p(x, z, \beta)}{\int p(x, z, \beta) dz d\beta}.$$

SVI: Exp. Family

$$p(x|\lambda) = h(x)e^{\theta T(x) - A(\theta)}$$

- 2 main properties:

$$\mathbb{E}_p[T(x)] = \nabla_\lambda A(\theta)$$

$$\mathbb{E}_p[(T(x) - \mathbb{E}_p[T(x)])(T(x) - \mathbb{E}_p[T(x)])^T] = \nabla^2_\lambda A(\theta)$$

SVI

- Example of exp. family

Gaussian	$p(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\| x-\mu \|^2/(2\sigma^2)}$	$x \in \mathbb{R}$
Bernoulli	$p(x) = \alpha^x (1-\alpha)^{1-x}$	$x \in \{0,1\}$
Binomial	$p(x) = \binom{n}{x} \alpha^x (1-\alpha)^{n-x}$	$x \in \{0,1,2,\ldots,n\}$
Multinomial	$p(x) = \frac{n!}{x_1! x_2! \ldots x_n!} \prod_{i=1}^n \alpha_i^{x_i}$	$x_i \in \{0,1,2,\ldots,n\}, \sum_i x_i = n$
Exponential	$p(x) = \lambda e^{-\lambda x}$	$x \in \mathbb{R}^+$
Poisson	$p(x) = \frac{e^{-\lambda}}{x!} \lambda^x$	$x \in \{0,1,2,\ldots\}$
Dirichlet	$p(x) = \frac{\Gamma(\sum_i \alpha_i)}{\prod_i \Gamma(\alpha_i)} \prod_i x_i^{\alpha_i - 1}$	$x_i \in [0,1], \sum_i x_i = 1$

SVI

- Natural parameterization of Bernolli

$$
\begin{aligned}
p(x) &= \alpha^x (1-\alpha)^{1-x} \\
&= \exp\left[\log\left(\alpha^x (1-\alpha)^{1-x}\right)\right] \\
&= \exp\left[\, x \log \alpha + (1-x) \log (1-\alpha) \,\right] \\
&= \exp\left[x \log \frac{\alpha}{1-\alpha} + \log (1-\alpha) \right] \\
&= \exp\left[\, x\, \theta - \log\left(1+e^{\theta}\right) \right]
\end{aligned}
$$

$$
T(x) = x \qquad \theta = \log \frac{\alpha}{1-\alpha} \qquad A(\theta) = \log\left(1+e^{\theta}\right)
$$

SVI: ELBO

$$
\begin{aligned}
\log p(x) &= \log \int p(x, z, \beta) dz d\beta \\
&= \log \int p(x, z, \beta) \frac{q(z, \beta)}{q(z, \beta)} dz d\beta \\
&= \log \left(\mathbb{E}_q \left[\frac{p(x, z, \beta)}{q(z, \beta)} \right] \right) \\
&\geq \mathbb{E}_q[\log p(x, z, \beta)] - \mathbb{E}_q[\log q(z, \beta)] \\
&\triangleq \mathcal{L}(q).
\end{aligned}
$$

SVI: Mean Field VI

- Mean field variational family

$$q(z,\beta) = q(\beta \mid \lambda) \prod_{n=1}^{N} \prod_{j=1}^{J} q(z_{nj} \mid \phi_{nj}).$$

- Our approx. dist. is from exp. family

$$q(\beta \mid \lambda) = h(\beta) \exp\{\lambda^\top t(\beta) - a_g(\lambda)\},$$
$$q(z_{nj} \mid \phi_{nj}) = h(z_{nj}) \exp\{\phi_{nj}^\top t(z_{nj}) - a_\ell(\phi_{nj})\}.$$

- Entropy term:

$$-\mathbb{E}_q[\log q(z,\beta)] = -\mathbb{E}_\lambda[\log q(\beta)] - \sum_{n=1}^{N} \sum_{j=1}^{J} \mathbb{E}_{\phi_{nj}}[\log q(z_{nj})]$$

  - $\mathbb{E}_{\phi_{nj}}[\cdot]$ an $\mathbb{E}_\lambda[\cdot]$ denote expectation w.r. $q(z_{nj} \mid \phi_{nj})$ an $q(\beta \mid \lambda)$

SVI: coordinate ascent inference

- Updating one variational parameter while holding others fixed.
- Elbo for global parameter

$$\mathcal{L}(\lambda) = \mathbb{E}_q[\log p(\beta \,|\, x, z)] - \mathbb{E}_q[\log q(\beta)] + \text{const.}$$

$$\mathcal{L}(\lambda) = \mathbb{E}_q[\eta_g(x, z, \alpha)]^\top \nabla_\lambda a_g(\lambda) - \lambda^\top \nabla_\lambda a_g(\lambda) + a_g(\lambda) + \text{const.}$$

- Recall that: $\mathbb{E}_q[t(\beta)] = \nabla_\lambda a_g(\lambda)$
- $\mathbb{E}_q[a_g(\eta_g(x, z, \alpha))]$ does not depend on λ
- Gradient of elbo w.r.t. λ $\quad \nabla_\lambda \mathcal{L} = \nabla_\lambda^2 a_g(\lambda)(\mathbb{E}_q[\eta_g(x, z, \alpha)] - \lambda).$
- Set it to 0: $\quad \lambda = \mathbb{E}_q[\eta_g(x, z, \alpha)].$

SVI: coordinate ascent inference

- Similarly for local parameters

$$\nabla_{\phi_{nj}} \mathcal{L} = \nabla^2_{\phi_{nj}} a_\ell(\phi_{nj}) (\mathbb{E}_q[\eta_\ell(x_n, z_{n,-j}, \beta)] - \phi_{nj}).$$

$$\phi_{nj} = \mathbb{E}_q[\eta_\ell(x_n, z_{n,-j}, \beta)]$$

```
Initialize λ^(0) randomly.
repeat
    for each local variational parameter φ_nj do
        Update φ_nj, φ_nj^(t) = E_{q^(t-1)}[η_{ℓ,j}(x_n, z_{n,-j}, β)].
    end for
    Update the global variational parameters, λ^(t) = E_{q^(t)}[η_g(z_{1:N}, x_{1:N})].
until the ELBO converges
```

We need to go over all data before updating λ

SVI: Natural Gradient

- Classical gradient method

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

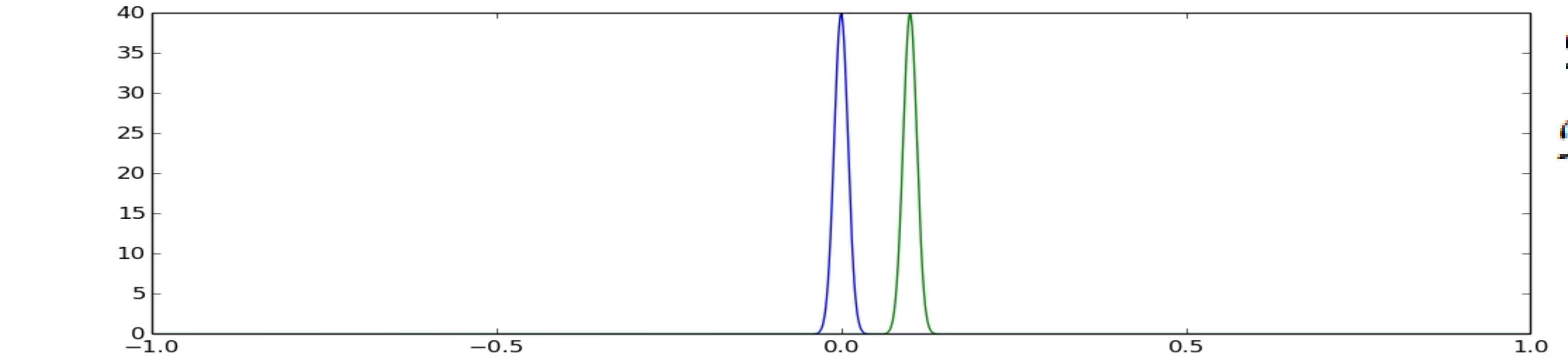
- Equal formulation

$$\arg\max_{d\lambda} f(\lambda + d\lambda) \quad \text{subject to } ||d\lambda||^2 < \varepsilon^2$$

- ε needs to be small enough.

SVI: Natural Gradient

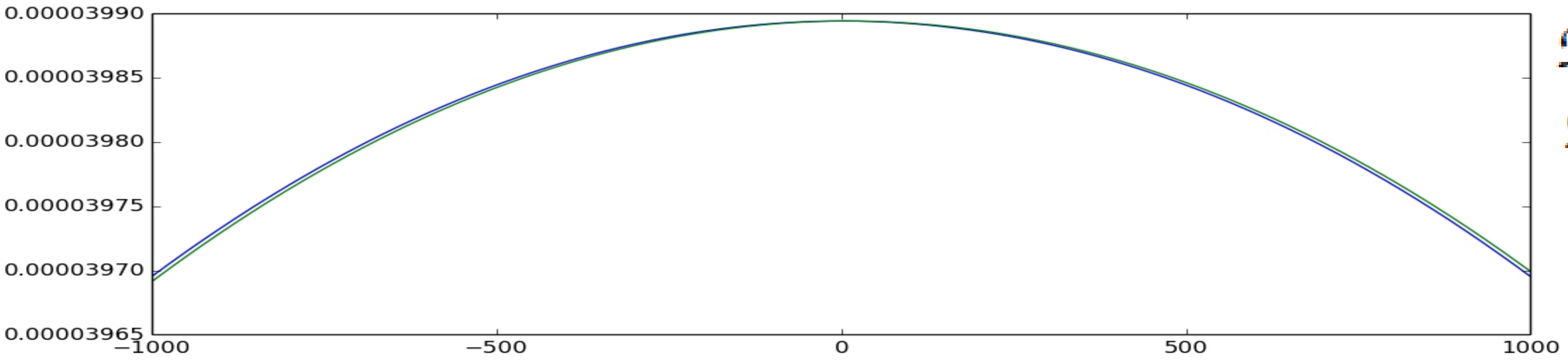
- Which of these two distributions are more different?



$\mathcal{N}(0, 0.01)$

$\mathcal{N}(0.1, 0.01)$

Euclidian Distance = 0.1



$\mathcal{N}(0, 10000)$

$\mathcal{N}(10, 10000)$

Euclidian Distance = 10

SVI: Natural Gradient

- Natural Measure of dissimilarity between probability measures:

$$D_{KL}^{\text{sym}}(\lambda, \lambda') = \mathbb{E}_{\lambda}\left[\log \frac{q(\beta \mid \lambda)}{q(\beta \mid \lambda')}\right] + \mathbb{E}_{\lambda'}\left[\log \frac{q(\beta \mid \lambda')}{q(\beta \mid \lambda)}\right]$$

$$\arg\max_{d\lambda} f(\lambda + d\lambda) \quad \text{subject to } D_{KL}^{\text{sym}}(\lambda, \lambda + d\lambda) < \varepsilon.$$

- Riemannian Metric :

$$d\lambda^T G(\lambda) d\lambda = D_{KL}^{\text{sym}}(\lambda, \lambda + d\lambda),$$

- Natural Gradient

$$\hat{\nabla}_{\lambda} f(\lambda) \triangleq G(\lambda)^{-1} \nabla_{\lambda} f(\lambda),$$

SVI: Natural Gradient

- Here, G is Fisher information matrix

$$G(\lambda) = \mathbb{E}_\lambda \left[(\nabla_\lambda \log q(\beta \mid \lambda))(\nabla_\lambda \log q(\beta \mid \lambda))^\top \right]$$

- We need to find G for exponential family.

SVI: Natural Gradient

$$\log q(\beta|\lambda + d\lambda) = O(d\lambda^2) + \log q(\beta|\lambda) + d\lambda^\top \nabla_\lambda \log q(\beta|\lambda),$$
$$q(\beta|\lambda + d\lambda) = O(d\lambda^2) + q(\beta|\lambda) + q(\beta|\lambda) d\lambda^\top \nabla_\lambda \log q(\beta|\lambda),$$

$$\begin{aligned}
D_{KL}^{\text{sym}}(\lambda, \lambda + d\lambda) &= \int_\beta (q(\beta|\lambda + d\lambda) - q(\beta|\lambda))(\log q(\beta|\lambda + d\lambda) - \log q(\beta|\lambda)) d\beta \\
&= O(d\lambda^3) + \int_\beta q(\beta|\lambda)(d\lambda^\top \nabla_\lambda \log q(\beta|\lambda))^2 d\beta \\
&= O(d\lambda^3) + \mathbb{E}_q[(d\lambda^\top \nabla_\lambda \log q(\beta|\lambda))^2] = O(d\lambda^3) + d\lambda^\top G(\lambda) d\lambda.
\end{aligned}$$

$$\begin{aligned}
G(\lambda) &= \mathbb{E}_\lambda\left[(\nabla_\lambda \log p(\beta \,|\, \lambda))(\nabla_\lambda \log p(\beta \,|\, \lambda))^\top\right] \\
&= \mathbb{E}_\lambda\left[(t(\beta) - \mathbb{E}_\lambda[t(\beta)])(t(\beta) - \mathbb{E}_\lambda[t(\beta)])^\top\right] \\
&= \nabla_\lambda^2 a_g(\lambda).
\end{aligned}$$

SVI: Natural Gradient

- Using natural gradient for variational parameters

$$\hat{\nabla}_{\lambda} \mathcal{L} = \mathbb{E}_{\phi}[\eta_g(x, z, \alpha)] - \lambda.$$

$$\hat{\nabla}_{\phi_{nj}} \mathcal{L} = \mathbb{E}_{\lambda, \phi_{n,-j}}[\eta_{\ell}(x_n, z_{n,-j}, \beta)] - \phi_{nj}.$$

SVI: Stochastic elbo

- Elbo for λ

Elbo for

$$\mathcal{L}(\lambda) = \mathbb{E}_q[\log p(\beta)] - \mathbb{E}_q[\log q(\beta)] + \sum_{n=1}^{N} \max_{\phi_n} (\mathbb{E}_q[\log p(x_n, z_n \mid \beta)] - \mathbb{E}_q[\log q(z_n)]).$$

- Stochastic Elbo for λ

Stochastic Elbo for

$$\mathcal{L}_I(\lambda) \triangleq \mathbb{E}_q[\log p(\beta)] - \mathbb{E}_q[\log q(\beta)] + N \max_{\phi_I} (\mathbb{E}_q[\log p(x_I, z_I \mid \beta) - \mathbb{E}_q[\log q(z_I)]]).$$

SVI: Stochastic Natural Gradient

- Natural Gradient and update

$$\hat{\nabla}\mathcal{L}_i = \mathbb{E}_q\left[\eta_g\left(x_i^{(N)}, z_i^{(N)}, \alpha\right)\right] - \lambda,$$

$$\eta_g\left(x_i^{(N)}, z_i^{(N)}, \alpha\right) = \alpha + N \cdot (t(x_n, z_n), 1). \quad \hat{\nabla}_\lambda \mathcal{L}_i = \alpha + N \cdot \left(\mathbb{E}_{\phi_i(\lambda)}[t(x_i, z_i)], 1\right) - \lambda,$$

$$\hat{\lambda}_t \triangleq \alpha + N\mathbb{E}_{\phi_i(\lambda)}[(t(x_i, z_i), 1)].$$

$$\begin{aligned} \lambda^{(t)} &= \lambda^{(t-1)} + \rho_t\left(\hat{\lambda}_t - \lambda^{(t-1)}\right) \\ &= (1-\rho_t)\lambda^{(t-1)} + \rho_t\hat{\lambda}_t. \end{aligned}$$

SVI algorithm

Initialize $\lambda^{(0)}$ randomly.

Set the step-size schedule ρ_t appropriately.

repeat

  Sample a data point x_i uniformly from the data set.

  Compute its local variational parameter,

$$\phi = \mathbb{E}_{\lambda^{(t-1)}}[\eta_g(x_i^{(N)}, z_i^{(N)})].$$

  Compute intermediate global parameters as though x_i is replicated N times,

$$\hat{\lambda} = \mathbb{E}_{\phi}[\eta_g(x_i^{(N)}, z_i^{(N)})].$$

  Update the current estimate of the global variational parameters,

$$\lambda^{(t)} = (1 - \rho_t)\lambda^{(t-1)} + \rho_t \hat{\lambda}.$$

until forever

Bridging the GAP

- What if taking samples from posterior approximate is not easy?

- Basic Idea:
  - Use monte carlo method to generate samples from posterior approximate.

Thank you!