Variational Inference via Stochastic Backpropagation

Kai Fan

February 27, 2016
Preliminaries

Stochastic Backpropagation

Variational Auto-Encoding

Related Work

Summary
Outline

Preliminaries

Stochastic Backpropagation

Variational Auto-Encoding

Related Work

Summary
Bayesian inference on latent variable model

- $\mathbf{y}$ observed data
- $\mathbf{x}$ latent variable
- $p_\theta(\mathbf{x}, \mathbf{y})$ probabilistic model
Bayesian inference on latent variable model

- $y$ observed data
- $x$ latent variable
- $p_\theta(x, y)$ probabilistic model

- Purpose: we are (very) interested in inferring a posterior distribution $p_\theta(x|y)$
  - Enables learning parameters in latent variable models
  - Deep learning

$P(x|y) = \frac{P(x, y)}{P(y)}$ is most often intractable.
Bayesian inference on latent variable model

- $\mathbf{y}$ observed data
- $\mathbf{x}$ latent variable
- $p_\theta(\mathbf{x}, \mathbf{y})$ probabilistic model

- Purpose: we are (very) interested in inferring a posterior distribution $p_\theta(\mathbf{x}|\mathbf{y})$
  - Enables learning parameters in latent variable models
  - Deep learning

- Difficulty: $p(\mathbf{x}|\mathbf{y}) = \frac{p(\mathbf{x}, \mathbf{y})}{p(\mathbf{y})}$ is most often intractable.
Non-variational approx. inference methods

- Point estimate of $p_\theta(x|y)$ (MAP)
  - Fast
  - Overfitting

- Markov Chain Monte Carlo (MCMC)
  - Asymptotically unbiased
  - Expensive, slow to assess convergence
Variational Inference

- Introduce variational distribution $q_\phi(x)$ or $q_\phi(x|y)$ of true posterior. 
  $\phi$ variational parameters

- Objective: minimize w.r.t. the KL-divergence

$$D_{KL}(q_\phi(x|y) \| p_\theta(x|y))$$

- $q_\phi(x|y) = p_\theta(x|y)$ achieves 0 KL divergence.
Lower Bound

- From marginal log-likelihood to lower bound,
\[
\log p_\theta(y) = \mathbb{E}_q \left[ \log \frac{p_\theta(y, x)}{q_\phi(x|y)} \right] + D_{KL}(q_\phi(x|y) || p_\theta(x|y)) \\
\geq \mathbb{E}_q[\log p_\theta(y, x) - \log q_\phi(x|y)] \\
\triangleq \mathcal{L}
\]

- Objective: maximize w.r.t the Lower Bound

- Non-gradient-based optimization technique: Mean-Field VB with fixed-point equations
  - Efficiency
  - Intractable / not applicable in many cases
Outline

Preliminaries

Stochastic Backpropagation

Variational Auto-Encoding

Related Work

Summary
Reparameterized Gradient Estimator

- Consider a general form of lower bound $\mathcal{L} = \mathbb{E}_{q(\mathbf{x}|\mathbf{y})}[f(\mathbf{y}, \mathbf{x})]$.

- Monte Carlo Gradient Approximation at Iteration $t$
  - sample $\epsilon^t$ from some base distribution $p(\epsilon)$
  - transformation $\mathbf{x}^t = g(\epsilon^t)$, s.t. $\mathbf{x}^t \sim q(\mathbf{x}|\mathbf{y})$
  - compute $\nabla \phi f(\mathbf{y}, \mathbf{x}^t)$ to approximate $\nabla \phi \mathcal{L}$

- Reparameterization has to exist. E.g. Gaussian, Laplace, Student t’s, etc.
Reparameterization Trick

Original form

\[ f(z; \phi, x) \sim \sim q(z; \phi, x) \]

Reparameterised form

\[ f(z; \phi, x) = g(\phi, x, \epsilon) \]

\[ \frac{\partial f}{\partial \epsilon} = \frac{\partial L}{\partial \phi_i} \]

Backprop
Gaussian Backpropagation

- $\mathbf{x} \sim \mathcal{N}(\mu, \mathbf{C})$, we have following identities.

\[
\nabla_{\mu_i} \mathbb{E}_{\mathcal{N}(\mathbf{x}|\mu, \mathbf{C})}[f(\mathbf{x})] = \mathbb{E}_{\mathcal{N}(\mathbf{x}|\mu, \mathbf{C})}[\nabla_{z_i} f(\mathbf{x})]
\]

\[
\nabla_{C_{ij}} \mathbb{E}_{\mathcal{N}(\mathbf{x}|\mu, \mathbf{C})}[f(\mathbf{x})] = \frac{1}{2} \mathbb{E}_{\mathcal{N}(\mathbf{x}|\mu, \mathbf{C})}[\nabla^2_{z_i z_j} f(\mathbf{x})],
\]

\[
\nabla^2_{C_{i,j}, C_{k,l}} \mathbb{E}_{\mathcal{N}(\mathbf{x}|\mu, \mathbf{C})}[f(\mathbf{x})] = \frac{1}{4} \mathbb{E}_{\mathcal{N}(\mathbf{x}|\mu, \mathbf{C})}[\nabla^4_{z_i z_j z_k z_l} f(\mathbf{x})],
\]

\[
\nabla^2_{\mu_i, C_{k,l}} \mathbb{E}_{\mathcal{N}(\mathbf{x}|\mu, \mathbf{C})}[f(\mathbf{x})] = \frac{1}{2} \mathbb{E}_{\mathcal{N}(\mathbf{x}|\mu, \mathbf{C})} \left[\nabla^3_{z_i z_k z_l} f(\mathbf{x})\right].
\]

- Unbiased estimator of $\nabla^k \mathbb{E}[f]$, $k = 1, 2$

- Higher order derivatives need to calculated w.r.t. $f$
Reparameterized Gaussian Backpropagation

- $x \sim \mathcal{N}(\mu, RR^\top)$, thus $x = \mu + R\epsilon$ where $\epsilon \sim \mathcal{N}(0, I)$

- New identities

\[
\nabla_R \mathbb{E}_{\mathcal{N}(\mu, C)}[f(x)] = \mathbb{E}_{\epsilon \sim \mathcal{N}(0, I_{dz})}[\epsilon g^\top]
\]

\[
\nabla^2_{\mu, R} \mathbb{E}_{\mathcal{N}(\mu, C)}[f(x)] = \mathbb{E}_{\epsilon \sim \mathcal{N}(0, I_{dz})}[\epsilon^\top \otimes H]
\]

\[
\nabla^2_R \mathbb{E}_{\mathcal{N}(\mu, C)}[f(x)] = \mathbb{E}_{\epsilon \sim \mathcal{N}(0, I_{dz})}[(\epsilon \epsilon^T) \otimes H]
\]

where $\otimes$ is Kronecker product, and gradient $g$, Hessian $H$ are evaluated at $\mu + R\epsilon$ in terms of $f(x)$.

- Still easy to obtain unbiased estimator

- Hessian-vector multiplication due to the fact that

\[(A^\top \otimes B) \text{vec}(V) = \text{vec}(AVB)\]
Bayesian Logreg

- Prior $\mathcal{N}(0, \Lambda)$ where $\Lambda$ is diagonal

- Variational distribution $q(\beta|\mu, D)$ where $D$ is diagonal for simplicity.
Outline

Preliminaries

Stochastic Backpropagation

Variational Auto-Encoding

Related Work

Summary
Model Formulation

- Gaussian latent variable, prior $p(x) = \mathcal{N}(0, I)$

- Generative model $p_\theta(y|x)$, characterize a non-linear transformation, e.g. MLP

- Recognition model $q_\phi(x|y) = \mathcal{N}(\mu, D)$, where $\phi = [\mu, D] = \text{MLP}(y; W, b)$ and denote $\psi = (W, b)$

Objective Function: $L = \log p(y|x) + \log p(x) - \log q(x|y)$

- $\log p(y|x)$ reconstruction error
- $\log p(x) - \log q(x|y)$ regularization
- Unlike VEM, $(\theta, \psi)$ is optimized simultaneously, by gradient based algorithm.
Unrolled VAE

Input: $x$

Output: $y$

$\mu$, $\Sigma$

$N(\mu, \Sigma)$

$W_5, b_5$

$W_4, b_4$

$W_2, b_2$, $W_3, b_3$

$h_d$

$h_e$

Input: $x$

Hidden decoder layer

Gaussian latent layer

Hidden encoder layer

$W_1, b_1$

(W_5, b_5), (W_6, b_6)$ if $x$ is continuous.
Back to Backpropagation

▶ Fast Gradient computation

\[ \nabla_{\psi_1} \mathbb{E}_\mathcal{N}(\mu, c)[f(x)] = \mathbb{E}_{\epsilon \sim \mathcal{N}(0, I)} \left[ g^\top \frac{\partial (\mu + R\epsilon)}{\partial \psi_1} \right] \]

\[ \nabla^2_{\psi_1 \psi_2} \mathbb{E}_\mathcal{N}(\mu, c)[f(x)] = \mathbb{E}_{\epsilon \sim \mathcal{N}(0, I)} \left[ \frac{\partial (\mu + R\epsilon)}{\partial \psi_1}^\top H \frac{\partial (\mu + R\epsilon)}{\partial \psi_2} + g^\top \frac{\partial^2 (\mu + R\epsilon)}{\partial \psi_1 \partial \psi_2} \right] \]

▶ \(O(d_z^2)\) algorithmic complexity for both 1st and 2nd derivative.
Back to Backpropagation

- For any $F$, $H_\psi v = \lim_{\gamma \to 0} \frac{\nabla F(\psi + \gamma v) - \nabla F(\psi)}{\gamma}$

$$H_\psi v = \left. \frac{\partial}{\partial \gamma} \nabla F(\psi + \gamma v) \right|_{\gamma=0}$$

$$= \frac{\partial}{\partial \gamma} \mathbb{E} \mathcal{N}(0,1) \left[ g^T \left. \frac{\partial (\mu(\psi) + R(\psi) \epsilon)}{\partial \psi} \right|_{\psi \leftarrow \psi + \gamma v} \right] \bigg|_{\gamma=0}$$

$$= \mathbb{E} \mathcal{N}(0,1) \left[ \left. \frac{\partial}{\partial \gamma} \left( g^T \left. \frac{\partial (\mu(\psi) + R(\psi) \epsilon)}{\partial \psi} \right|_{\psi \leftarrow \psi + \gamma v} \right) \right] \bigg|_{\gamma=0}$$

- PCG only requires $H_\psi v$ to solve linear system $Hp = -g$.

- For $K$ iteration of PCG, relative tolerance $\epsilon < \exp(-2K/\sqrt{c})$, where $c$ is matrix conditioner. Thus, $c$ can be nearly as large as $O(K^2)$.

- Complexity for each iteration: $O(Kdd_z^2)$ v.s. $O(dd_z^2)$
Theoretical Perspective

- If $f$ is an $L$-Lipschitz differentiable function and $\epsilon \sim \mathcal{N}(0, I_d)$, then $\mathbb{E}[(f(\epsilon) - \mathbb{E}[f(\epsilon)])^2] \leq \frac{L^2 \pi^2}{4}$.

- $\mathbb{P}\left(\left| \frac{1}{M} \sum_{m=1}^{M} f(\epsilon_m) - \mathbb{E}[f(\epsilon)] \right| \geq t \right) \leq 2e^{-\frac{2Mt^2}{\pi^2 L^2}}$.

- In most application, $M = 1$ is used as MC integration.
VAE Experiments

Manifold of Generative Model by setting $d_z = 2$
VAE Experiments

Lower Bound

Frey Face

Olivetti Face

Lower Bound

Ada train
Ada test
L-BFGS-SGVI train
L-BFGS-SGVI test
HFSGVI train
HFSGVI test
Outline

Preliminaries

Stochastic Backpropagation

Variational Auto-Encoding

Related Work

Summary
Semi-supervised VAE (NIPS 2014)

Generative Model: \( p(y) = \text{Cat}(y|\pi) \); \( p(x) = \mathcal{N}(0, I) \);
\( p(y|x, x) = \text{MLP} \)

Recognition Model: \( q(x|y, y) = \mathcal{N}(\mu_\phi(y, y), D_\phi(y)) \) and
\( q(y|y) = \text{Cat}(y|\pi(y)) \), parameter function is also MLP.
Neural Variational Inference (ICML 2014)

Sigmoid Belief Networks

- Generative Model: $h_L \rightarrow h_{L-1} \ldots \rightarrow h_1 \rightarrow y$
- Recognition Model: reverse the arrow direction
- Learning signal or control variate for variance reduction
  borrowing idea from RL

$$\nabla_\phi L = \mathbb{E}_q[(\log p_\theta(x, z) - \log q_\phi(z|x)) \times \nabla_\phi \log q_\phi(z|x)]$$

$$= \mathbb{E}_q[(\log p_\theta(x, z) - \log q_\phi(z|x) - C_\xi(x)) \times \nabla_\phi \log q_\phi(z|x)]$$

- $(\theta, \phi, \xi)$ joint learning
Dynamic Modeling

- DRAW (Dynamic VAE with LSTM, ICML 2015, reviewed before)
- DSBN (NIPS 2015), generative model is similar to HMM
Dynamic Modeling, ctd

hGCHMM, my model

(g) NVI  (h) Doctor  (i) GibbsEM
Bayesian Dark Knowledge (NIPS 2015)

- **Teacher Model:** deep neural networks $T(y|x, \theta)$, prior $p(\theta|\lambda)$
- **Student Model:** deep neural networks $S(y|x, \omega)$, prior $p(\omega|\gamma)$
- **Two step training** (or distilled SGLD, term they used in paper)
  - Mini-batch data $(X, Y)$ with size $B$
  - SGLD update $\theta$
    \[
    \Delta \theta_{t+1} = \frac{\eta_t}{2} \left( \nabla_\theta \log p(\theta|\lambda) + \frac{N}{B} \sum_{x_i \in X} \nabla_\theta \log p(y_i|x_i, \theta) \right) + \mathcal{N}(0, \eta_t)
    \]
  - SGD update $\omega$ with noisy data $\tilde{X}$ only; $\tilde{y}_i$ is obtained by feeding $\tilde{x}_i$ to current teacher model
    \[
    \Delta \omega_{t+1} = \rho_t \left( \frac{1}{B} \sum_{\tilde{x}_i \in \tilde{X}} \nabla_\omega \log p(\tilde{y}_i|\tilde{x}_i, \omega) + \gamma \omega_t \right)
    \]
Outline

Preliminaries

Stochastic Backpropagation

Variational Auto-Encoding

Related Work

Summary
Summary

- Minimize the difference between Generative model and recognition model
- Variational inference framework