Statistical Models & Computing Methods

Lecture 9: Stochastic Variational Inference

Cheng Zhang

School of Mathematical Sciences, Peking University

December 03, 2019

- Mean-field VI can be slow when the data size is large.
- Moreover, the conditional conjugacy required by mean-field VI greatly reduces the general applicability of the method.
- Fortunately, as an optimization approach, VI allows us to easily combine it with various scalable optimization methods.
- In this lecture, we will introduce some of the recent advancements on scalable variational inference, both for mean-field VI and more general VI.
- We will also talk about alternative training objectives in VI besides KL divergence.

Mean-field VI Could Be Data-inefficient

- A generic class of models

$$p(\beta, z, x) = p(\beta) \prod_{i=1}^{n} p(z_i, x_i | \beta)$$

- The mean-field approximation

$$q(\beta, z) = q(\beta | \lambda) \prod_{i=1}^{n} q(z_i | \phi_i)$$

- Coordinate ascent could be data-inefficient

$$\lambda^* = \mathbb{E}_{q(z)}(\eta_g(x, z)), \quad \phi_i^* = \mathbb{E}_{q(\beta)}(\eta_\ell(x_i, \beta))$$

  - Requires local computation for each data points.
  - Aggregate these computation to update the global parameter.

Gradients of The ELBO

- Recall that the λ-ELBO (update to a constant) is

$$L(\lambda) = \nabla_\lambda A_g(\lambda)^\top \left(\alpha + \sum_{i=1}^{n} \mathbb{E}_{\phi_i}(T(z_i, x_i)) - \lambda \right) + A_g(\lambda)$$

- Differentiating this w.r.t. λ yields

$$\nabla_\lambda L(\lambda) = \nabla_\lambda^2 A_g(\lambda) \left(\alpha + \sum_{i=1}^{n} \mathbb{E}_{\phi_i}(T(z_i, x_i)) - \lambda \right)$$

- Similarly

$$\nabla_{\phi_i} L(\phi_i) = \nabla_{\phi_i}^2 A_\ell(\phi_i) \left(\mathbb{E}_\lambda(\eta_\ell(x_i, \beta)) - \phi_i \right)$$

- The gradient of f at λ, $\nabla_\lambda f(\lambda)$ points in the same direction as the solution to

$$\arg\max_{d\lambda} f(x + d\lambda), \quad s.t. \ \|d\lambda\|^2 \leq \epsilon^2$$

for sufficiently small ϵ.

- The gradient direction implicitly depends on the Euclidean distance, which might not capture the distance between the parameterized probability distribution $q(\beta|\lambda)$.
- We can use *natural gradient* instead, which points in the same direction as the solution to

$$\arg\max_{d\lambda} f(x + d\lambda), \quad s.t. \ \mathrm{D}_{\mathrm{KL}}^{\mathrm{sym}}(q(\beta|\lambda), q(\beta|\lambda + d\lambda)) \leq \epsilon$$

for sufficiently small ϵ, where $\mathrm{D}_{\mathrm{KL}}^{\mathrm{sym}}$ is the symmetrized KL divergence.

- We manage the symmetrized KL divergence constraint with a Riemannian metric $G(\lambda)$

$$\mathrm{D}_{\mathrm{KL}}^{\mathrm{sym}}(q(\beta|\lambda), q(\beta|\lambda + d\lambda)) \approx d\lambda^\top G(\lambda) d\lambda$$

as $d\lambda \to 0$. G is the **Fisher information** matrix of $q(\beta|\lambda)$

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

- The **natural gradient** (Amari, 1998)

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

- When $q(\beta|\lambda)$ is in the prescribed exponential family

$$G(\lambda) = \nabla_\lambda^2 A_g(\lambda)$$

- The natural gradient of the ELBO

$$\nabla_\lambda^{\text{nat}} L = \left(\alpha + \sum_{i=1}^{n} \mathbb{E}_{\phi_i}(T(z_i, x_i))\right) - \lambda$$
$$\nabla_{\phi_i}^{\text{nat}} L = \mathbb{E}_\lambda(\eta_\ell(x_i, \beta)) - \phi_i$$

  Classical coordinate ascent can be viewed as natural gradient descent with step size one

- Use the noisy natural gradient instead

$$\hat{\nabla}_\lambda^{\text{nat}} L(\lambda) = \alpha + n\mathbb{E}_{\phi_j}(T(z_j, x_j)) - \lambda, \quad j \sim \text{Uniform}(1, \ldots, n)$$

- This is a good noisy gradient
  - The expectation is the exact gradient (unbiased).
  - Depends merely on optimized local parameters (cheap).

Stochastic Variational Inference

Input: data $\mathbf{x}$, model $p(\beta, \mathbf{z}, \mathbf{x})$.

Initialize λ randomly. Set ρ_t appropriately.

repeat

Sample $j \sim \text{Unif}(1, \ldots, n)$.

Set local parameter $\phi \leftarrow \mathbb{E}_\lambda \left[\eta_\ell(\beta, x_j) \right]$.

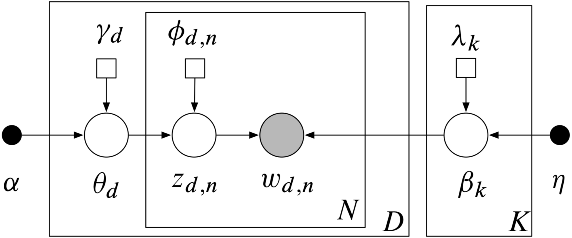
Set intermediate global parameter

$$\hat{\lambda} = \alpha + n\mathbb{E}_\phi [t(Z_j, x_j)].$$

Set global parameter

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

until *forever*



Classic Coordinate Ascent

$$\phi_{d,n,k} \propto \exp\left(\mathbb{E}(\log \theta_{d,k}) + \mathbb{E}(\log \beta_{k,w_{d,n}})\right)$$

$$\gamma_d = \alpha + \sum_{n=1}^{N} \phi_{d,n}, \quad \lambda_k = \eta + \sum_{d=1}^{D} \sum_{n=1}^{N} \phi_{d,n,k} w_{d,n}$$

Stochastic Variational Inference in LDA

- Sample a document w_d uniform from the data set
- Estimate the local variational parameters using the current topics. For $n = 1, \ldots, N$

$$\phi_{d,n,k} \propto \exp\left(\mathbb{E}(\log \theta_{d,k}) + \mathbb{E}(\log \beta_{k,w_{d,n}})\right), \quad k = 1, \ldots, K$$

$$\gamma_d = \alpha + \sum_{n=1}^{N} \phi_{d,n}$$

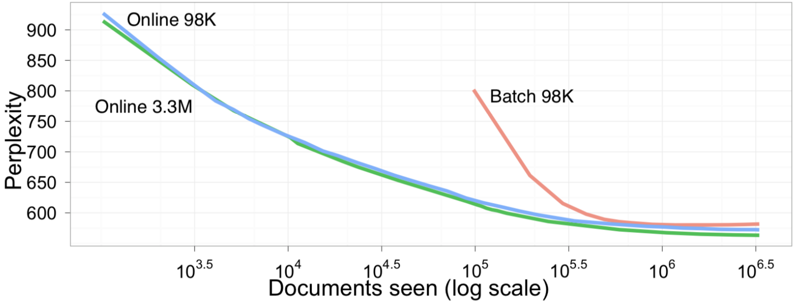
- Form the intermediate topics from those local parameters for noisy natural gradient

$$\hat{\lambda}_k = \eta + D \sum_{n=1}^{N} \phi_{d,n,k} w_{d,n}, \quad k = 1, \ldots, K$$

- Update topics using noisy natural gradient

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

Stochastic Variational Inference in LDA



Documents analyzed	**2048**	**4096**	**8192**	**12288**	**16384**	**32768**	**49152**	**65536**
Top eight words	systems	systems	service	service	service	business	business	business
	road	health	systems	systems	companies	service	service	industry
	made	communication	health	companies	systems	companies	companies	service
	service	service	companies	business	business	industry	industry	companies
	announced	billion	market	company	company	company	services	services
	national	language	communication	billion	industry	management	company	company
	west	care	company	health	market	systems	management	management
	language	road	billion	industry	billion	services	public	public

VI for General Models

- Mean-field VI works for **conjugate-exponential models**, where the local optimal has closed-form solution.
- For more general models, we may not have this conditional conjugacy
  - Nonlinear Time Series Models
  - Deep Latent Gaussian Models
  - Generalized Linear Models
  - Stochastic Volatility Models
  - Bayesian Neural Networks
  - Sigmoid Belief Network
- While we may derive a model specific bound for each of these models (Knowles and Minka, 2011; Paisley et al., 2012), it would be better if there is a solution that does not entail model specific work.

VI for Bayesian Logistic Regression

- The logistic regression model

$$y_i \sim \text{Bernoulli}(p_i),\ p_i = \frac{1}{1+\exp(-x_i^\top \beta)}. \quad \beta \sim \mathcal{N}(0, I_d)$$

- The mean-field approximation

$$q(\beta) = \prod_{j=1}^{d} \mathcal{N}(\beta_j | \mu_j, \sigma_j^2)$$

- The ELBO is

$$L(\mu, \sigma^2) = \mathbb{E}_q(\log p(\beta) + \log p(y|x, \beta) - \log q(\beta))$$

VI for Bayesian Logistic Regression

$$
\begin{aligned}
L(\mu, \sigma^2) &= \mathbb{E}_q(\log p(\beta) - \log q(\beta) + \log p(y|x, \beta)) \\
&= -\frac{1}{2}\sum_{j=1}^{d}(\mu_j^2 + \sigma_j^2) + \frac{1}{2}\sum_{j=1}^{d}\log\sigma_j^2 + \mathbb{E}_q \log p(y|x, \beta) + \text{Const} \\
&= \frac{1}{2}\sum_{j=1}^{d}(\log\sigma_j^2 - \mu_j^2 - \sigma_j^2) + Y^\top X\mu - \mathbb{E}_q(\log(1 + \exp(X\beta)))
\end{aligned}
$$

- We can not compute the expectation term
- This hides the objective dependence on the variational parameters, making it hard to directly optimize.

- Let $p(x, \theta)$ be the joint probability (i.e., the posterior up to a constant), and $q_\phi(\theta)$ be our variational approximation
- The ELBO is

$$L(\phi) = \mathbb{E}_q(\log p(x, \theta) - \log q_\phi(\theta))$$

- Instead of requiring a closed-form lower bound and differentiating afterwards, we can take derivatives directly
- As shown later, this leads to a stochastic optimization approach that handles massive data sets as well.

- Compute the gradient

$$
\begin{aligned}
\nabla_\phi L &= \nabla_\phi \mathbb{E}_q (\log p(x,\theta) - \log q_\phi(\theta)) \\
&= \int \nabla_\phi q_\phi(\theta)(\log p(x,\theta) - \log q_\phi(\theta)) \, d\theta \\
&\qquad\qquad - q_\phi(\theta) \nabla_\phi \log q_\phi(\theta) \, d\theta \\
&= \int q_\phi(\theta) \nabla_\phi \log q_\phi(\theta)(\log p(x,\theta) - \log q_\phi(\theta)) \\
&\qquad\qquad - q_\phi(\theta) \nabla_\phi \log q_\phi(\theta) \, d\theta \\
&= \mathbb{E}_q \left(\nabla_\phi \log q_\phi(\theta)(\log p(x,\theta) - \log q_\phi(\theta) - 1) \right)
\end{aligned}
$$

Using $\nabla_\phi \log q_\phi \theta = \frac{\nabla_\phi q_\phi(\theta)}{q_\phi(\theta)}$

- Recall that

$$\nabla_\phi L = \mathbb{E}_q \left(\nabla_\phi \log q_\phi(\theta) (\log p(x,\theta) - \log q_\phi(\theta) - 1) \right)$$

- Note that

$$\mathbb{E}_q \nabla_\phi \log q_\phi(\theta) = 0$$

- We can simplify the gradient as follows

$$\nabla_\phi L = \mathbb{E}_q \left(\nabla_\phi \log q_\phi(\theta) (\log p(x,\theta) - \log q_\phi(\theta)) \right)$$

- This is known as score function estimator or REINFORCE gradients (Williams, 1992; Ranganath et al., 2014; Minh et al., 2014)

Monte Carlo Estimate

$$\nabla_\phi L = \mathbb{E}_q\left(\nabla_\phi \log q_\phi(\theta)(\log p(x,\theta) - \log q_\phi(\theta))\right)$$

- Unbiased stochastic gradients via **Monte Carlo!**

$$\frac{1}{S}\sum_{s=1}^{S} \nabla_\phi \log q_\phi(\theta_s)(\log p(x,\theta_s) - \log q_\phi(\theta_s)), \quad \theta_s \sim q_\phi(\theta)$$

- The requirements for inference
  - Sampling from $q_\phi(\theta)$
  - Evaluating $\nabla_\phi \log q_\phi(\theta)$
  - Evaluating $\log p(x,\theta)$ and $\log q_\phi(\theta)$
- This is called **Black Box Variational Inference** (BBVI): no model specific work! (Ranganath et al., 2014)

Basci BBVI

Algorithm 1: Basic Black Box Variational Inference

Input : Model $\log p(\mathbf{x}, \mathbf{z})$,
Variational approximation $q(\mathbf{z}; \boldsymbol{\nu})$

Output : Variational Parameters: $\boldsymbol{\nu}$

while *not converged* **do**
- $\mathbf{z}[s] \sim q$ **// Draw S samples from q**
- $\rho = t$-th value of a Robbins Monro sequence
- $\boldsymbol{\nu} = \boldsymbol{\nu} + \rho \frac{1}{S} \sum_{s=1}^{S} \nabla_{\boldsymbol{\nu}} \log q(\mathbf{z}[s]; \boldsymbol{\nu})(\log p(\mathbf{x}, \mathbf{z}[s]) - \log q(\mathbf{z}[s]; \boldsymbol{\nu}))$
- $t = t + 1$

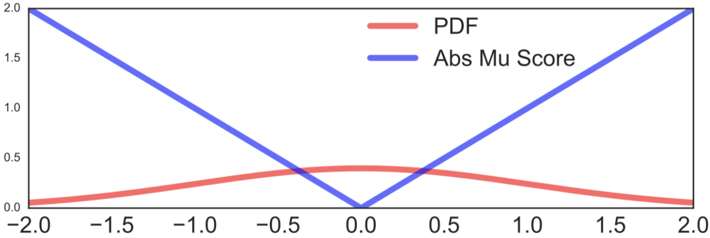
end

Ranganath et al., 2014

Basic BBVI Doesn't Work

Variance of the gradient can be a problem

$$\mathrm{Var}_{q_\phi(\theta)} = \mathbb{E}_q\left((\nabla_\phi \log q_\phi(\theta)(\log p(x,\theta) - \log q_\phi(\theta)) - \nabla_\phi L)^2\right)$$

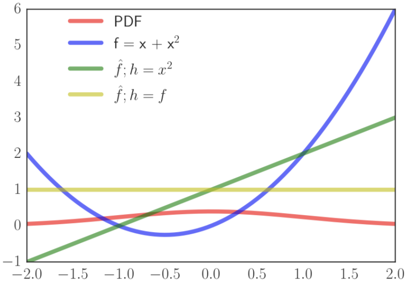


Adapted from Blei, Ranganath and Mohamed

- magnitude of $\log p(x,\theta) - \log q_\phi(\theta)$ varies widely
- rare values sampling
- too much variance to be useful

- To make BBVI work in practice, we need methods to reduce the variance of naive Monte Carlo estimates
- **Control Variates**. To reduce the variance of Monte Carlo estimates of $\mathbb{E}(f(x))$, we replace f with $\hat{f}$ such that $\mathbb{E}(\hat{f}(x)) = \mathbb{E}(f(x))$. A general class

$$\hat{f}(x) = f(x) - a(h(x) - \mathbb{E}h(x))$$



- a can be chosen to minimize the variance.
- h is a function of our choice. Good h have high correlation with the original function f.

$$\hat{f}(x) = f(x) - a(h(x) - \mathbb{E}h(x))$$

- For variational inference, we need h functions with known q expectation
- A commonly used one is $h(\theta) = \nabla_\phi \log q_\phi(\theta)$, where

$$\mathbb{E}_q(\nabla_\phi \log q_\phi(\theta)) = 0, \quad \forall q$$

- The variance of $\hat{f}$ is

$$\mathrm{Var}(\hat{f}) = \mathrm{Var}(f) + a^2\mathrm{Var}(h) - 2a\mathrm{Cov}(f, h)$$

and the optimal scaling is $a^* = \mathrm{Cov}(f, h)/\mathrm{Var}(h)$. In practice this can be estimated using the empirical variance and covariance on the samples

Baseline

- When $h(\theta) = \nabla_\phi \log q_\phi(\theta)$, the control variate gradient is

$$\nabla_\phi L = \mathbb{E}_q \left(\nabla_\phi \log q_\phi(\theta)(\log p(x, \theta) - \log q_\phi(\theta) - a)\right)$$

  and a is called a **baseline**.
- Baselines can be constant, or input-dependent $a(x)$.
- While we can estimate the baseline using the samples as before, people often use a *model-agnostic* baseline to *centre the learning signal* (Minh and Gregor, 2014)

$$\rho = \arg\min_\rho \mathbb{E}_q(\ell(x, \theta, \phi) - a_\rho(x))^2$$

  where the learning signal is

$$\ell(x, \theta, \phi) = \log p(x, \theta) - \log q_\phi(\theta)$$

- We can use Rao-Blackwellization to reduce the variance by integrating out some random variables.
- Consider the mean-field variational family

$$q(\theta) = \prod_{i=1}^{d} q_i(\theta_i|\phi_i)$$

- Let $q_{(i)}$ be the distribution of variables that depend on the ith variable (i.e., the Markov blanket of θ_i and θ_i), and let $p_i(x, \theta_{(i)})$ be the terms in the joint probability that depend on those variables.

$$\nabla_{\phi_i} L = \mathbb{E}_{q_{(i)}} \left(\nabla_{\phi_i} \log q_i(\theta_i|\phi_i)(\log p_i(x, \theta_{(i)}) - \log q_i(\theta_i|\phi_i)) \right)$$

- This can be combined with control variates.

The Reparameterization Trick

- Another commonly used variance reduction technique is **the reparameterization trick** (Kingma et al., 2014; Rezende et al., 2014)
- The Reparameterization

$$\theta = g_\phi(\epsilon), \;\; \epsilon \sim q_\epsilon(\epsilon) \quad \Longrightarrow \quad \theta \sim q_\phi(\theta)$$

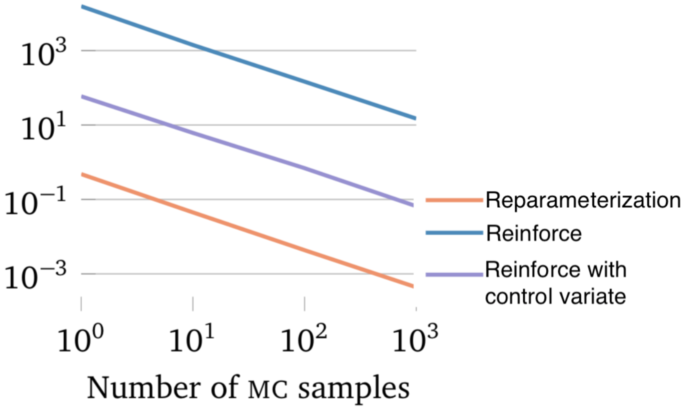
- Example:

$$\theta = \epsilon\sigma + \mu, \;\; \epsilon \sim \mathcal{N}(0,1) \quad \Longleftrightarrow \quad \theta \sim \mathcal{N}(\mu, \sigma^2)$$

- Compute the gradient via the reparameterization trick

$$\begin{aligned}\nabla_\phi L &= \nabla_\phi \mathbb{E}_{q_\phi(\theta)}(\log p(x,\theta) - \log q_\phi(\theta)) \\ &= \nabla_\phi \mathbb{E}_{q_\epsilon(\epsilon)}(\log p(x, g_\phi(\epsilon)) - \log q_\phi(g_\phi(\epsilon))) \\ &= \mathbb{E}_{q_\epsilon(\epsilon)} \nabla_\phi (\log p(x, g_\phi(\epsilon)) - \log q_\phi(g_\phi(\epsilon)))\end{aligned}$$

Variance Comparison



Kucukelbir et al., 2016

Score Function

- Differentiates the density $\nabla_\phi q_\phi(\theta)$
- Works for general models, including both discrete and continuous models.
- Works for large class of variational approximations
- May suffer from large variance

Reparameterization

- Differentiates the function $\nabla_\phi(\log p(x,\theta) - \log q_\phi(\theta))$
- Requires differentiable models
- Requires variational approximation to have form $\theta = g_\phi(\epsilon)$
- Better behaved variance in general

- Scale up previous stochastic variational inference methods to large data set via **data subsampling**.
- Replace the log joint distribution with unbiased stochastic estimates

$$\log p(x, \theta) \simeq \log p(\theta) + \frac{n}{m} \sum_{i=1}^{m} \log p(x_{t_i} | \theta), \quad m \ll n$$

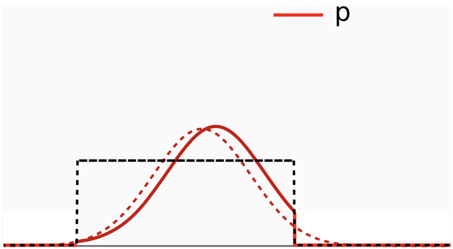
- Example: score function estimator

$$\hat{\nabla}_\phi L = \frac{1}{S} \sum_{s=1}^{S} \nabla_\phi \log q_\phi(\theta_s) \Bigg(\log p(\theta_s) + \frac{n}{m} \sum_{i=1}^{m} \log p(x_{t_i} | \theta_s) \\ - \log q_\phi(\theta_s) \Bigg), \quad \theta_s \sim q_\phi(\theta)$$

Summary on Stochastic VI

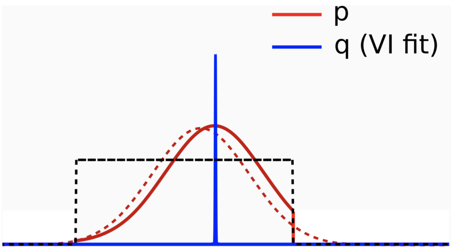
- When the data size is large, we can use **stochastic optimization** to scale up VI.
- For conditional exponential models, we can use noisy natural gradient.
- For general models, naive stochastic gradient estimators may have large variance, variance reduction techniques are often required.
  - Score function estimator (for both discrete and continuous latent variable)
  - The reparameterization trick (for continuous variable, and requires reparameterizable variational family)
- We can also combine score function estimators with the reparameterization trick for more general and robust stochastic gradient estimators (Ruiz et al., 2016)

- So far, we have only used the KL divergence as a distance measure in VI.
- Other than the KL divergence, there are many alternative statistical distance measures between distributions that admit a variety of statistical properties.
- In this lecture, we will introduce several alternative divergence measures to KL, and discuss their statistical properties, with applications in VI.



- VI does not work well for non-smooth potentials
- This is largely due to the zero-avoiding behaviour
  - The area where $p(\theta)$ is close to zero has very negative $\log p$, so does the variational distribution q distribution when trained to minimize the KL.
- In this truncated normal example, VI will fit a delta function!

Potential Problems with The KL Divergence



- VI does not work well for non-smooth potentials
- This is largely due to the zero-avoiding behaviour
  - The area where $p(\theta)$ is close to zero has very negative $\log p$, so does the variational distribution q distribution when trained to minimize the KL.
- In this truncated normal example, VI will fit a delta function!

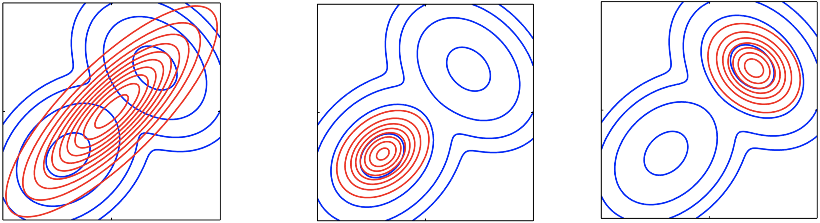
Beyond The KL Divergence

- Recall that the KL divergence from q to p is

$$D_{\mathrm{KL}}(q\|p) = \mathbb{E}_q \log \frac{q(x)}{p(x)} = \int q(x) \log \frac{q(x)}{p(x)}\, dx$$

- An alternative: the reverse KL divergence

$$D_{\mathrm{KL}}^{\mathrm{Rev}}(p\|q) = \mathbb{E}_p \log \frac{p(x)}{q(x)} = \int p(x) \log \frac{p(x)}{q(x)}\, dx$$



Reverse KL **KL**

The f-Divergence

- The f-divergence from q to p is defined as

$$D_f(q\|p) = \int p(x) f\left(\frac{q(x)}{p(x)}\right)\, dx$$

where f is a convex function such that $f(1) = 0$.

- The f-divergence defines a family of valid divergences

$$\begin{aligned} D_f(q\|p) &= \int p(x) f\left(\frac{q(x)}{p(x)}\right)\, dx \\ &\geq f\left(\int p(x)\frac{q(x)}{p(x)}\, dx\right) = f(1) = 0 \end{aligned}$$

and

$$D_f(q\|p) = 0 \Rightarrow q(x) = p(x) \text{ a.s.}$$

Many common divergences are special cases of f-divergence, with different choices of f.

- KL divergence. $f(t) = t \log t$
- reverse KL divergence. $f(t) = -\log t$
- Hellinger distance. $f(t) = \frac{1}{2}(\sqrt{t} - 1)^2$

$$H^2(p, q) = \frac{1}{2} \int (\sqrt{q(x)} - \sqrt{p(x)})^2 dx = \frac{1}{2} \int p(x) \left(\sqrt{\frac{q(x)}{p(x)}} - 1 \right)^2 dx$$

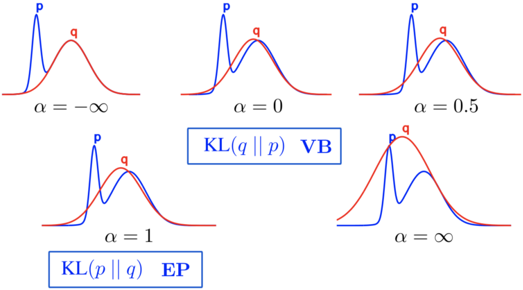
- Total variation distance. $f(t) = \frac{1}{2}|t - 1|$

$$d_{\mathrm{TV}}(p, q) = \frac{1}{2} \int |p(x) - q(x)| dx = \frac{1}{2} \int p(x) \left| \frac{q(x)}{p(x)} - 1 \right| dx$$

Amari's α-Divergence

When $f(t) = \frac{t^\alpha - t}{\alpha(\alpha-1)}$, we have the Amari's α-divergence (Amari, 1985; Zhu and Rohwer, 1995)

$$D_\alpha(p\|q) = \frac{1}{\alpha(1-\alpha)}\left(1 - \int p(\theta)^\alpha q(\theta)^{1-\alpha}\, d\theta\right)$$



$$D_{\mathrm{KL}}(q\|p) = \lim_{\alpha\to 0} D_\alpha(p\|q)$$

$$D_{\mathrm{KL}}(p\|q) = \lim_{\alpha\to 1} D_\alpha(p\|q)$$

Adapted from Hernández-Lobato et al.

$$D_\alpha(q\|p) = \frac{1}{\alpha - 1} \log \int q(\theta)^\alpha p(\theta)^{1-\alpha}\, d\theta$$

- Some special cases of Rényi's α-divergence
  - $D_1(q\|p) := \lim_{\alpha\to 1} D_\alpha(q\|p) = D_{\mathrm{KL}}(q\|p)$
  - $D_0(q\|p) = -\log \int_{q(\theta)>0} p(\theta)d\theta = 0$ *iff* $supp(p) \subset supp(q)$.
  - $D_{+\infty}(q\|p) = \log \max_\theta \frac{q(\theta)}{p(\theta)}$
  - $D_{\frac{1}{2}}(q\|p) = -2\log\left(1 - \mathrm{Hel}^2(q\|p)\right)$
- Importance properties
  - Rényi divergence is non-decreasing in α

$$D_{\alpha_1}(q\|p) \geq D_{\alpha_2}(q\|p), \qquad \text{if } \alpha_1 \geq \alpha_2$$

  - Skew symmetry: $D_{1-\alpha}(q\|p) = \frac{1-\alpha}{\alpha} D_\alpha(p\|q)$

- Consider approximating the exact posterior $p(\theta|x)$ by minimizing Rényi's α-divergence $D_\alpha(q(\theta)\|p(\theta|x))$ for some selected $\alpha > 0$
- Using $p(\theta|x) = p(\theta, x)/p(x)$, we have

$$\begin{aligned} D_\alpha(q(\theta)\|p(\theta|x)) &= \frac{1}{\alpha - 1}\log\int q(\theta)^\alpha p(\theta|x)^{1-\alpha}\, d\theta \\ &= \log p(x) - \frac{1}{1-\alpha}\log\int q(\theta)^\alpha p(\theta, x)^{1-\alpha}\, d\theta \\ &= \log p(x) - \frac{1}{1-\alpha}\log \mathbb{E}_q\left(\frac{p(\theta, x)}{q(\theta)}\right)^{1-\alpha} \end{aligned}$$

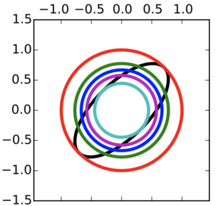
- The Rényi lower bound (Li and Turner, 2016)

$$L_\alpha(q) \triangleq \frac{1}{1-\alpha}\log \mathbb{E}_q\left(\frac{p(\theta, x)}{q(\theta)}\right)^{1-\alpha}$$

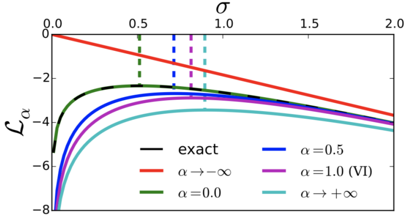
- **Theorem**(Li and Turner 2016). The Rényi lower bound is continuous and non-increasing on $\alpha \in [0,1] \cup \{|L_\alpha| < +\infty\}$. Especially for all $0 < \alpha < 1$

$$L_{\mathrm{VI}}(q) = \lim_{\alpha \to 1} L_\alpha(q) \le L_\alpha(q) \le L_0(q)$$

$L_0(q) = \log p(x)$ iff $supp(p(\theta|x)) \subset supp(q(\theta))$.



(a) Approximated posterior.



(b) Hyper-parameter optimisation.

Monte Carlo Estimation

- Monte Carlo estimation of the Rényi lower bound

$$\hat{L}_{\alpha,K}(q) = \frac{1}{1-\alpha} \log \frac{1}{K} \sum_{i=1}^{K} \left(\frac{p(\theta_i, x)}{q(\theta_i)} \right)^{1-\alpha}, \quad \theta_i \sim q(\theta)$$

- Unlike traditional VI, here the Monte Carlo estimate is **biased**. Fortunately, the bias can be characterized by the following theorem
- **Theorem**(Li and Turner, 2016). $\mathbb{E}_{\{\theta_i\}_{i=1}^K}(\hat{L}_{\alpha,K}(q))$ as a function of α and K is
  - non-decreasing in K for fixed $\alpha \leq 1$, and converges to $L_\alpha(q)$ as $K \to +\infty$ if $supp(p(\theta|x)) \subset supp(q(\theta))$.
  - continuous and non-increasing in α on $[0, 1] \cup \{|L_\alpha| < +\infty\}$

Multiple Sample ELBO

- When $\alpha = 0$, the Monte Carlo estimate reduces to the multiple sample lower bound (Burda et al., 2015)

$$\hat{L}_K(q) = \log\left(\frac{1}{K}\sum_{i=1}^{K}\frac{p(x,\theta_i)}{q(\theta_i)}\right), \quad \theta_i \sim q(\theta)$$

- This recovers the standard ELBO when $K = 1$.
- Using more samples improves the tightness of the bound (Burda et al., 2015)

$$\log p(x) \geq \mathbb{E}(\hat{L}_{K+1}(q)) \geq \mathbb{E}(\hat{L}_K(q))$$

Moreover, if $p(x,\theta)/q(\theta)$ is bounded, then

$$\mathbb{E}(\hat{L}_K(q)) \to \log p(x), \quad \text{as } K \to +\infty$$

Lower Bound Maximization

Using the reparameterization trick

$$\theta \sim q_\phi(\theta) \Leftrightarrow \theta = g_\phi(\epsilon),\ \epsilon \sim q_\epsilon(\epsilon)$$

$$\nabla_\phi \hat{L}_{\alpha,K}(q_\phi) = \sum_{i=1}^{K} \left(\hat{w}_{\alpha,i} \nabla_\phi \log \frac{p(g_\phi(\epsilon_i), x)}{q_\phi(g_\phi(\epsilon_i))} \right), \quad \epsilon_i \sim q_\epsilon(\epsilon)$$

where

$$\hat{w}_{\alpha,i} \propto \left(\frac{p(g_\phi(\epsilon_i), x)}{q_\phi(g_\phi(\epsilon_i))} \right)^{1-\alpha},$$

the normalized importance weight with finite samples. This is a biased estimate of $\nabla_\phi L_\alpha(q_\phi)$ (except $\alpha = 1$).

- $\alpha = 1$: Standard VI with the reparamterization trick
- $\alpha = 0$: Importance weighted VI (Burda et al., 2015)

- Full batch training for maximizing the Rényi lower bound could be very inefficient for large datasets
- Stochastic optimization is non-trivial since the Rényi lower bound can not be represented as an expectation on a datapoint-wise loss, except for $\alpha = 1$.
- Two possible methods:
  - derive the fixed point iteration on the whole dataset, then use the minibatch data to approximately compute it (Li et al., 2015)
  - approximate the bound using the minibatch data, then derive the gradient on this approximate objective (Hernández-Lobato et al., 2016)

Remark: the two methods are equivalent when $\alpha = 1$ (standard VI).

Minibatch Training: Energy Approximation

- Suppose the true likelihood is

$$p(x|\theta) = \prod_{n=1}^{N} p(x_n|\theta)$$

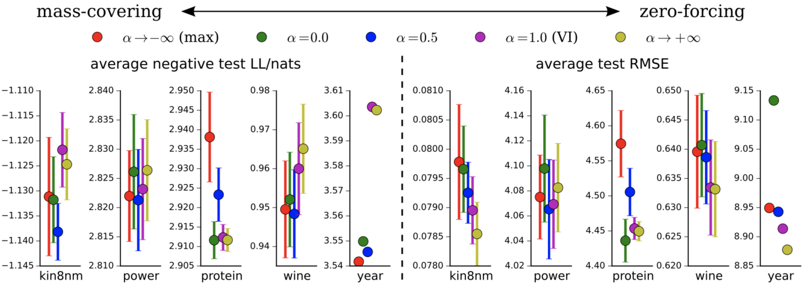
- Approximate the likelihood as

$$p(x|\theta) \approx \left(\prod_{n \in \mathcal{S}} p(x_n|\theta) \right)^{\frac{N}{|\mathcal{S}|}} \triangleq \bar{f}_{\mathcal{S}}(\theta)^N$$

- Use this approximation for the energy function

$$\tilde{L}_\alpha(q, \mathcal{S}) = \frac{1}{1-\alpha} \log \mathbb{E}_q \left(\frac{p_0(\theta)\bar{f}_{\mathcal{S}}(\theta)^N}{q(\theta)} \right)^{1-\alpha}$$

Example: Bayesian Neural Network



Adapted from Li and Turner, 2016

- The optimal α may vary for different data sets.
- Large α improves the predictive error, while small α provides better test log-likelihood.
- $\alpha = 0.5$ seems to produce overall good results for both test LL and RMSE.

- In standard VI, we often minimize $D_{\mathrm{KL}}(q\|p)$. Sometimes, we can also minimize $D_{\mathrm{KL}}(p\|q)$ (can be viewed as MLE).

$$q^* = \arg\min_q D_{\mathrm{KL}}(p\|q) = \arg\max_q \mathbb{E}_p \log q(\theta)$$

- Assume q is from the exponential family

$$q(\theta|\eta) = h(\theta)\exp\left(\eta^\top T(\theta) - A(\eta)\right)$$

- The optimal η^* satisfies

$$\begin{aligned}\eta^* &= \arg\max_\eta \mathbb{E}_p \log q(\theta|\eta)\\ &= \arg\max_\eta \left(\eta^\top \mathbb{E}_p\left(T(\theta)\right) - A(\eta)\right) + \mathrm{Const}\end{aligned}$$

- Differentiate with respect to η

$$\mathbb{E}_p\left(T(\theta)\right) = \nabla_\eta A(\eta^*)$$

- Note that $q(\theta|\eta)$ is a valid distribution $\forall\eta$

$$\begin{aligned} 0 &= \nabla_\eta \int h(\theta) \exp\left(\eta^\top T(\theta) - A(\eta)\right)\, d\theta \\ &= \int q(\theta|\eta)\left(T(\theta) - \nabla_\eta A(\eta)\right)\, d\theta \\ &= \mathbb{E}_q\left(T(\theta)\right) - \nabla_\eta A(\eta) \end{aligned}$$

- The KL divergence is minimized if the expected sufficient statistics are the same

$$\mathbb{E}_q\left(T(\theta)\right) = \mathbb{E}_p\left(T(\theta)\right)$$

Expectation Propagation

- An approximate inference method proposed by Minka 2001.
- Suitable for approximating product forms. For example, with iid observations, the posterior takes the following form

$$p(\theta|x) \propto p(\theta) \prod_{i=1}^{n} p(x_i|\theta) = \prod_{i=0}^{n} f_i(\theta)$$

- We use an approximation

$$q(\theta) \propto \prod_{i=0}^{n} \tilde{f}_i(\theta)$$

One common choice for $\tilde{f}_i$ is the exponential family

$$\tilde{f}_i(\theta) = h(\theta) \exp\left(\eta_i^\top T(\theta) - A(\eta_i)\right)$$

- Iteratively refinement of the terms $\tilde{f}_i(\theta)$

- Take out term approximation i

$$q^{\setminus i}(\theta) \propto \prod_{j \neq i} \tilde{f}_j(\theta)$$

- Put back in term i

$$\hat{p}(\theta) \propto f_i(\theta) \prod_{j \neq i} \tilde{f}_j(\theta)$$

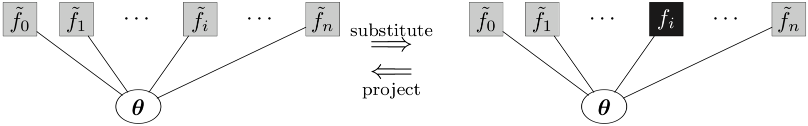
- Match moments. Find q such that

$$\mathbb{E}_q(T(\theta)) = \mathbb{E}_{\hat{p}}(T(\theta))$$

- Update the new term approximation

$$\tilde{f}_i^{\text{new}}(\theta) \propto \frac{q(\theta)}{q^{\setminus i}(\theta)}$$

How Does EP Work?

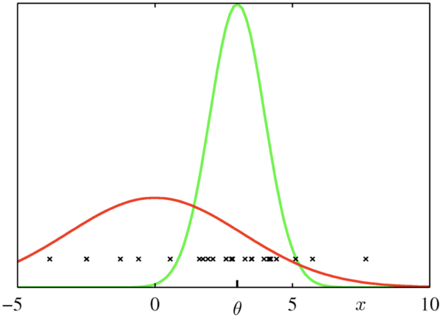


- Minimize the KL divergence from $\hat{p}$ to q

$$D_{\mathrm{KL}}(\hat{p}\|q) = \mathbb{E}_{\hat{p}} \log\left(\frac{\hat{p}(\theta)}{q(\theta)}\right)$$

- Equivalent to moment matching when q is in the exponential family.

Example: The Clutter Problem



- **Goal**: fit a multivariate Gaussian into data in the presence of background clutter (also Gaussian)

$$p(x|\theta) = (1-w)\mathcal{N}(x|\theta, I) + w\mathcal{N}(x|0, aI)$$

- The prior is Gaussian: $p(\theta) = \mathcal{N}(\theta|0, bI)$.

- The joint distribution

$$p(\theta, x) = p(\theta) \prod_{i=1}^{n} p(x_i|\theta)$$

  is a mixture of 2^n Gaussians, intractable for large n.
- We approximate it using a spherical Gaussian

$$q(\theta) = \mathcal{N}(\theta|m, vI)$$

- This is an exponential family with
  - sufficient statistics $T(\theta) = (\theta, \theta^\top \theta)$
  - natural parameters $\eta = (v^{-1}m, -\frac{1}{2}v^{-1})$
  - normalizing constant $Z(\eta) = (2\pi v)^{d/2} \exp\left(\frac{m^\top m}{2v}\right)$

- For the clutter problem, we have

$$
\begin{aligned}
f_0(\theta) &= p(\theta) \\
f_i(\theta) &= p(x_i|\theta),\ i = 1, \ldots, n
\end{aligned}
$$

- The approxmation is of the form

$$
\begin{aligned}
\tilde{f}_0(\theta) &= f_0(\theta) = p(\theta) \\
\tilde{f}_i(\theta) &= s_i \exp(\eta_i^\top T(\theta)),\ i = 1, \ldots, n \\
q(\theta) &\propto \prod_{i=0}^{n} \tilde{f}_i(\theta) = s\mathcal{N}(\theta; \eta)
\end{aligned}
$$

- Initialize $\eta_i = (0, 0)$ for $i = 1, \ldots, n$

- With natural parameters, taking out term approximation i is trivial.

$$q^{\setminus i}(\theta) \propto \frac{q(\theta)}{\tilde{f}_i(\theta)} \propto \mathcal{N}(\theta; \eta^{\setminus i})$$

where

$$\eta^{\setminus i} = \eta - \eta_i$$

- Now we put back in term i

$$\begin{aligned}
\hat{p}(\theta) &\propto ((1-w)\mathcal{N}(x_i|\theta, I) + w\mathcal{N}(x_i|0, aI))\,\mathcal{N}(\theta; \eta^{\setminus i}) \\
&= (1-w)\frac{Z(\eta^+)}{Z(\eta^{x_i})Z(\eta^{\setminus i})}\mathcal{N}(\theta; \eta^+) + w\mathcal{N}(x_i|0, aI)\mathcal{N}(\theta; \eta^{\setminus i}) \\
&\propto r\mathcal{N}(\theta; \eta^+) + (1-r)\mathcal{N}(\theta; \eta^{\setminus i})
\end{aligned}$$

where $\eta^+ = \eta^{\setminus i} + \eta^{x_i}, \quad \eta^{x_i} = (x_i, -\frac{1}{2})$.

- Now we match the sufficient statistics of the Gaussian mixture

$$\hat{p}(\theta) = r\mathcal{N}(\theta; \eta^{+}) + (1-r)\mathcal{N}(\theta; \eta^{\setminus i})$$

From $\mathbb{E}_q(T(\theta)) = \mathbb{E}_{\hat{p}}(T(\theta))$, we have

$$m = rm^{+} + (1-r)m^{\setminus i}$$
$$v + m^\top m = r\left(v^{+} + (m^{+})^\top m^{+}\right) + (1-r)\left(v^{\setminus i} + (m^{\setminus i})^\top m^{\setminus i}\right)$$

- Similarly, the update of $\tilde{f}_i$ is trivial

$$\tilde{f}_i(\theta) \propto \frac{q(\theta)}{q^{\setminus i}(\theta)} \propto \mathcal{N}(\theta; \eta_i)$$

where

$$\eta_i = \eta - \eta^{\setminus i}$$

Marginal Likelihood by EP

- We can use EP to evaluate the marginal likelihood $p(x)$
- To do this, we include a scale on $\tilde{f}_i(\theta)$

$$\tilde{f}_i(\theta) = Z_i \frac{q^*(\theta)}{q^{\backslash i}(\theta)}$$

where $q^*(\theta)$ is a normalized version of $q(\theta)$ and

$$Z_i = \int q^{\backslash i}(\theta) f_i(\theta) \, d\theta$$

- Use the normalizing constant of $q(x)$ to approximate $p(x)$

$$p(x) \approx \int \prod_{i=0}^{n} \tilde{f}_i(\theta) \, d\theta$$

- For the clutter problem

$$s_i \exp(\eta_i^\top T(\theta)) = \tilde{f}_i(\theta) = Z_i \frac{q^*(\theta)}{q^{\setminus i}(\theta)}$$

implies

$$s_i = Z_i \frac{Z(\eta^{\setminus i})}{Z(\eta)}$$

$$Z_i = (1-w)\frac{Z(\eta^+)}{Z(\eta^{x_i})Z(\eta^{\setminus i})} + w\mathcal{N}(x_i|0, aI)$$

- The marginal likelihood estimate is

$$p(x) \approx \int \prod_{i=0}^{n} \tilde{f}_i(\theta)\, d\theta = \frac{Z(\eta)}{Z(\eta_0)} \prod_{i=1}^{n} s_i$$

Summary

- Other than the standard KL divergence, there are many alternative distance measures for VI (e.g., f-divergence, Rényi α-divergence).
- The Rényi α-divergences allow tractable lower bound and promote different learning behaviors through the choice of α (from mode-covering to model-seeking as α goes from $-\infty$ to ∞), which can be adapted to specific learning tasks.
- We also introduced another approximate inference method, expectation propagation (EP), that uses the reversed KL. More recent development on EP (Li et al., 2015, Hernández-Lobato et al., 2016).
- Many other options including variational upper bounds, adaptive variational bounds, etc.

- S. Amari. Natural gradient works efficiently in learning. Neural computation, 10(2):251–276, 1998.
- Hoffman, M., Blei, D., Wang, C., and Paisley, J. (2013). Stochastic variational inference. Journal of Machine Learning Research, 14:1303–1347.
- D. Knowles and T. Minka. Non-conjugate variational message passing for multinomial and binary regres- sion. In Advances in Neural Information Processing Systems, 2011.
- J. Paisley, D. Blei, and M. Jordan. Variational Bayesian inference with stochastic search. International Conference in Machine Learning, 2012.

- Williams, R. J. (1992). Simple statistical gradient-following algorithms for connectionist reinforcement learning. In Machine Learning, pages 229–256.
- R. Ranganath, S. Gerrish, and D. Blei. Black box variational inference. In Artificial Intelligence and Statistics, 2014.
- Rezende, D. J., Mohamed, S., and Wierstra, D. (2014). Stochastic backpropagation and approximate inference in deep generative models. In International Conference on Machine Learning, pages 1278–1286.
- D. P. Kingma and M. Welling. Auto-encoding variational Bayes. In International Conference on Learning Representations, 2014.

- A. Mnih and K. Gregor. Neural variational inference and learning in belief networks. Advances in Neural Information Processing Systems, 2014.
- F. R. Ruiz, M. Titsias, and D. Blei. The generalized reparameterization gradient. Advances in Neural Information Processing Systems, 2016.
- Amari, Shun-ichi. Differential-Geometrical Methods in Statistic. Springer, New York, 1985.
- Zhu, Huaiyu and Rohwer, Richard. Information geometric measurements of generalisation. Technical report, Technical Report NCRG/4350. Aston University., 1995.

- Y. Li and R. E. Turner. Rényi Divergence Variational Inference. NIPS, pages 1073–1081, 2016.
- Y. Burda, R. Grosse, and R. Salakhutdinov. Importance weighted autoencoders. International Conference on Learning Representations (ICLR), 2016.
- Y. Li, J. M. Hernández-Lobato, and R. E. Turner. Stochastic expectation propagation. In Advances in Neural Information Processing Systems (NIPS), 2015.
- J. M. Hernández-Lobato, Y. Li, M. Rowland, D. Hernández-Lobato, T. Bui, and R. E. Turner. Black-box α-divergence minimization. In Proceedings of The 33rd International Conference on Machine Learning (ICML), 2016.

- A. B. Dieng, D. Tran, R. Ranganath, J. Paisley and D. M. Blei. Variational Inference via χ Upper Bound Minimization. Advances in Neural Information Processing Systems, 2017.
- D. Wang, H. Liu and Q. Liu. Variational Inference with Tail-adaptive f-Divergence. Advances in Neural Information Processing Systems, 2018.