Variational Inference (contd)

CS772A: Probabilistic Machine Learning

Piyush Rai

Recap: Variational Inference (VI)

Variational distribution

Variational parameters

- Assuming $p(\boldsymbol{Z}|\boldsymbol{\mathcal{D}},\Theta)$ is intractable, VI approximates it by a distr $q(\boldsymbol{Z}|\phi)$ or $q_\phi(\boldsymbol{Z})$

KL minimization

$$\phi^* = \text{argmin}_\phi \, \text{KL}[q_\phi(\boldsymbol{Z})||p(\boldsymbol{Z}|\boldsymbol{\mathcal{D}},\Theta)]$$

ELBO maximization

$$\phi^* = \text{argmax}_\phi \mathbb{E}_{q_\phi(\boldsymbol{Z})}[\log p(\boldsymbol{\mathcal{D}}|\boldsymbol{Z},\Theta)] - \text{KL}[q_\phi(\boldsymbol{Z})||p(\boldsymbol{Z}|\Theta)]$$
$$= \text{argmax}_\phi \mathbb{E}_{q_\phi(\boldsymbol{Z})}\left[\log p(\boldsymbol{\mathcal{D}},\boldsymbol{Z}|\Theta) - \log q_\phi(\boldsymbol{Z})\right] = \text{argmax}_\phi \, \mathcal{L}(\phi,\Theta)$$

Can use gradient-based optimization to learn the parameters of the variational distribution

$$\phi_{t+1} = \phi_t + \eta_t \nabla_{\phi=\phi_t}\mathcal{L}(\phi,\Theta)$$

Case when Θ is also unknown will be discussed later

Mean-field assumption on the variational distribution

$$q(\boldsymbol{Z}|\phi) = \prod_{i=1}^{M} q(\boldsymbol{Z}_i|\phi_i)$$

$\mathbb{E}_{i\neq j}$ denotes expectations w.r.t. $\prod_{i\neq j} q(Z_i|\phi_i)$

$$q_j^*(\boldsymbol{Z}_j) = \frac{\exp(\mathbb{E}_{i\neq j}[\log p(\boldsymbol{\mathcal{D}},\boldsymbol{Z}|\Theta)])}{\int \exp(\mathbb{E}_{i\neq j}[\log p(\boldsymbol{\mathcal{D}},\boldsymbol{Z}|\Theta)]\, d\boldsymbol{Z}_j}$$

This, for simple enough model, when using mean-field VI, we can get optimal $\boldsymbol{q}$ "directly" without taking ELBO derivatives

Equivalent to writing $\log q_j^*(\boldsymbol{Z}_j) = \mathbb{E}_{i\neq j}[\log p(\boldsymbol{\mathcal{D}},\boldsymbol{Z}|\Theta)] + \text{const}$

"simple enough" means the cases where these expectations can be analytically computed

Example: Mean-field VI without ELBO Derivatives

No "latent variables" here. Data $\mathbf{X}$ is fully observed, and parameters μ, τ need to be estimated

- Consider data $\mathbf{X} = \{x_1, x_2, \dots, x_N\}$ from a one-dim Gaussian $\mathcal{N}(\mu, \tau^{-1})$
- Assume the following normal-gamma prior on μ and τ

Assume the hyperparameters $\mu_0, \lambda_0, a_0, b_0$ are known

$$p(\mu|\tau) = \mathcal{N}(\mu|\mu_0, (\lambda_0\tau)^{-1}) \quad p(\tau) = \text{Gamma}(\tau|a_0, b_0)$$

- Posterior is also normal-gamma due to the jointly conjugate prior
- Let's still try mean-field VI for this model

Note that we aren't even specifying the forms of these two distributions! We'll be able identify the forms in a few steps after working with the expectations

- With mean-field assumption on the variational posterior $q(\mu, \tau) = q_\mu(\mu) q_\tau(\tau)$

$$\begin{aligned} \log q_\mu^*(\mu) &= \mathbb{E}_{q_\tau}[\log p(\mathbf{X}, \mu, \tau)] + \text{const} \\ \log q_\tau^*(\tau) &= \mathbb{E}_{q_\mu}[\log p(\mathbf{X}, \mu, \tau)] + \text{const} \end{aligned}$$

- In this example, the log-joint $\log p(\mathbf{X}, \mu, \tau) = \log p(\mathbf{X}|\mu, \tau) + \log p(\mu|\tau) + \log p(\tau)$. Thus

$$\begin{aligned} \log q_\mu^*(\mu) &= \mathbb{E}_{q_\tau}[\log p(\mathbf{X}|\mu, \tau) + \log p(\mu|\tau)] + \text{const} \qquad \text{(only keeping terms that involve } \mu) \\ \log q_\tau^*(\tau) &= \mathbb{E}_{q_\mu}[\log p(\mathbf{X}|\mu, \tau) + \log p(\mu|\tau) + \log p(\tau)] + \text{const} \end{aligned}$$

Example: Mean-field VI without ELBO Derivatives

- Substituting $p(\mathbf{X}|\mu,\tau) = \prod_{n=1}^{N} p(x_n|\mu,\tau)$ and $p(\mu|\tau)$, we get

$$\begin{aligned}\log q_\mu^*(\mu) &= \mathbb{E}_{q_\tau}[\log p(\mathbf{X}|\mu,\tau) + \log p(\mu|\tau)] + \text{const} \\ &= -\frac{\mathbb{E}_{q_\tau}[\tau]}{2}\left\{\sum_{n=1}^{N}(x_n-\mu)^2 + \lambda_0(\mu-\mu_0)^2\right\} + \text{const}\end{aligned}$$

- (Verify) The above is log of a Gaussian. This $q_\mu^* = \mathcal{N}(\mu|\mu_N, \lambda_N^{-1})$ with

$$\mu_N = \frac{\lambda_0\mu_0 + N\bar{x}}{\lambda_0 + N} \quad \text{and} \quad \lambda_N = (\lambda_0 + N)\mathbb{E}_{q_\tau}[\tau]$$

This update depends on q_τ

- Proceeding in a similar way (verify), we can show that $q_\tau^* = \text{Gamma}(\tau|a_N, b_N)$

$$a_N = a_0 + \frac{N+1}{2} \quad \text{and} \quad b_N = b_0 + \frac{1}{2}\mathbb{E}_{q_\mu}\left[\sum_{n=1}^{N}(x_n-\mu)^2 + \lambda_0(\mu-\mu_0)^2\right]$$

This update depends on q_μ

- Note: Updates of q_μ^* and q_τ^* depend on each other (hence alternating updates needed)

Mean-Field VI for Locally Conjugate Models

- Since $\log q_j^*(\boldsymbol{Z}_j) = \mathbb{E}_{i\neq j}[\log p(\boldsymbol{X},\boldsymbol{Z})]$ + const $= \mathbb{E}_{i\neq j}\left[\log p(\boldsymbol{X},\boldsymbol{Z}_j,\boldsymbol{Z}_{-j})\right]$ + const

$$\log q_j^*(\mathbf{Z}_j) = \mathbb{E}_{i\neq j}[\log p(\mathbf{Z}_j|\mathbf{X},\mathbf{Z}_{-j})] + \text{const}$$

For any model

- Thus finding optimal $q_j^*(\boldsymbol{Z}_j)$ only requires expectations of params of CP $p(\boldsymbol{Z}_j|\boldsymbol{X},\boldsymbol{Z}_{-j})$
- For locally conjugate models, we know CP is easy and is an exp-fam distr of the form

$$p(\mathbf{Z}_j|\mathbf{X},\mathbf{Z}_{-j}) = h(\mathbf{Z}_j)\exp\left[\eta(\mathbf{X},\mathbf{Z}_{-j})^\top\mathbf{Z}_j - A(\eta(\mathbf{X},\mathbf{Z}_{-j}))\right]$$

- Using the above, we can rewrite the optimal variational distribution as follows

$$\log q_j^*(\mathbf{Z}_j) = \mathbb{E}_{i\neq j}\left[\log\left(h(\mathbf{Z}_j)\exp\left[\eta(\mathbf{X},\mathbf{Z}_{-j})^\top\mathbf{Z}_j - A(\eta(\mathbf{X},\mathbf{Z}_{-j}))\right]\right)\right] + \text{const}$$

$$\Longrightarrow q_j^*(\mathbf{Z}_j) \propto h(\mathbf{Z}_j)\exp\left[\mathbb{E}_{i\neq j}[\eta(\mathbf{X},\mathbf{Z}_{-j})]^\top\mathbf{Z}_j\right] \qquad \text{(verify)}$$

- Thus, with local conj, we just require expectation of nat. params. of CP of $\boldsymbol{Z}_j$

Variational EM

- In LVMs, latent vars $\boldsymbol{Z}$ and parameters $\boldsymbol{\Theta}$ both may be unknown. In such cases, we can use variational EM (VEM). Same as EM except VEM uses VI to approx. CP of $\boldsymbol{Z}$
- VEM alternates between the following two steps
  - Maximize the ELBO w.r.t. ϕ (gives the variational approximation $q(\boldsymbol{Z})$ of CP of $\boldsymbol{Z}$)

$$\phi^{(t)} = \text{argmax}_{\phi} \ \mathbb{E}_{q_\phi(\boldsymbol{Z})}\left[\log p\left(\mathcal{D}, \boldsymbol{Z} | \Theta^{(t-1)}\right) - \log q_\phi(\boldsymbol{Z})\right]$$

  - Maximize the ELBO w.r.t. $\boldsymbol{\Theta}$ (gives us point estimate of $\boldsymbol{\Theta}$)

$$\Theta^{(t)} = \text{argmax}_{\Theta} \ \mathbb{E}_{q_{\phi^{(t)}}(\boldsymbol{Z})}\left[\log p(\mathcal{D}, \boldsymbol{Z}|\Theta) - \log q_{\phi^{(t)}}(\boldsymbol{Z})\right]$$

$$= \text{argmax}_{\Theta} \ \mathbb{E}_{q_{\phi^{(t)}}(\boldsymbol{Z})}[\log p(\mathcal{D}, \boldsymbol{Z}|\Theta)]$$

This looks very similar to the expected CLL with the CP replaced by its variational approximation

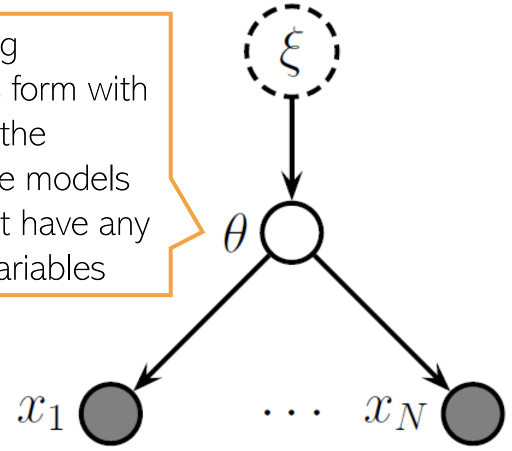
- Note: If we want posterior for $\boldsymbol{\Theta}$ as well, treat it similar to $\boldsymbol{Z}$ and apply variational approximation (instead of using VEM) if the posterior isn't tractable

VI for models without "latent variables"

Recall the Gaussian mean and variance estimation problem

- Suppose we have a "fully observed" case (no missing data/latent variables but just some unknown global parameters $\boldsymbol{\theta}$ and known hyperparams ξ)
- A simple example of the model is shown in the figure below

Even supervised learning problems may have this form with $\boldsymbol{\theta}$ being the weights of the generative/discriminative models and the models may not have any missing data or latent variables



$$p(\mathcal{D}, \boldsymbol{\theta}|\boldsymbol{\xi}) = p(\boldsymbol{\theta}|\boldsymbol{\xi}) \prod_{n=1}^{N} p(\boldsymbol{x}_n|\boldsymbol{\theta})$$

If this CP is intractable, we can use VI to approximate this

$$p(\boldsymbol{\theta}|\mathcal{D}, \boldsymbol{\xi}) = \frac{p(\mathcal{D}|\boldsymbol{\theta})p(\boldsymbol{\theta}|\boldsymbol{\xi})}{p(\mathcal{D}|\boldsymbol{\xi})}$$

- If ξ are also unknown then one way would be to alternate like Variational EM
  - Approximating the CP $p(\theta|\mathcal{D}, \xi)$ using VI
  - Using MLE-II to get point estimates of the hyperparameters ξ

Making VI Faster for LVMs: Stochastic VI (SVI)

- Many LVMs have local latent variables $\boldsymbol{Z} = \{\boldsymbol{z}_1, \boldsymbol{z}_2, \dots, \boldsymbol{z}_N\}$ and global params $\boldsymbol{\Theta}$
- VI updates of local and global variables depend on each other (similar to EM)
- This makes things slow (for VI and also for EM) especially when N is large
  - We must update $q(\boldsymbol{z}_n|\boldsymbol{\phi}_n)$, i.e., compute $\boldsymbol{\phi}_n$, for each latent variable before updating $\boldsymbol{\Theta}$
- Also need all the data $\boldsymbol{X} = \{\boldsymbol{x}_1, \boldsymbol{x}_2, \dots, \boldsymbol{x}_N\}$ in memory to do these updates
- Stochastic VI* is an efficient way using minibatches of data
- In each iteration, SVI takes a minibatch $\mathcal{B}$ of $|\mathcal{B}| \ll N$ data points, updates $q(\boldsymbol{z}_n|\boldsymbol{\phi}_n)$ examples in that minibatches and approximates the ELBO as follows

Optimize this approximate ELBO w.r.t. $\boldsymbol{\Theta}$
(note: this is an unbiased estimate*)

$$\tilde{\mathcal{L}}(\phi, \Theta) = \frac{N}{B}\sum_{x_i \in \mathcal{B}} \mathbb{E}_{q(\boldsymbol{z}_i|\phi_i)}[\log p(\boldsymbol{x}_i|\boldsymbol{z}_i, \Theta)] - \mathrm{KL}[q_\phi(\boldsymbol{z}_i)||p(\boldsymbol{z}_i|\Theta)]$$

*Stochastic Variational Inference (Hoffman et al, 2013)

Making VI Faster for LVMs: Amortized VI

- Instead of computing the optimal ϕ_n for each $q(z_n|\phi_n)$, learn a function to do so

$$q(z_n|\phi_n) \approx q(z_n|\hat{\phi}_n) \quad \text{where} \quad \hat{\phi}_n = \text{NN}_\phi(x_n)$$

- Function is usually a neural network with weights ϕ
  - Usually referred to as "inference network" or "recognition model"
- Amortization: We are shifting the cost of finding ϕ_n for each data point to finding the weights ϕ of the neural network shared by all data points
- Can also combine amortized VI with stochastic VI
  - Each iteration only uses a minibatch to optimize NN weights ϕ and global params Θ
- ELBO expression remains the same but $q(z_n|\phi_n)$ is replaced by $q(z_n|\text{NN}_\phi(x_n))$
- Amortized VI quality can be poor but it is fast and can give a quick solution
  - We can refine this solution other methods (e.g., using sampling; will see later)
  - This refinement based approach is called "semi-amortized VI"

VI using ELBO's gradients

- For simple locally conjugate models, VI updates are usually easy
  - Sometimes, can find the optimal $\boldsymbol{q}$ even without taking the ELBO's gradients
- For complex models, we have to use the more general gradient-based approach
- Consider the setting when we have latent variables $\boldsymbol{Z}$ and parameters $\boldsymbol{\Theta}$
- The ELBO's gradient w.r.t. $\boldsymbol{\Theta}$

$$\nabla_\Theta \mathcal{L}(\phi, \Theta) = \nabla_\Theta \, \mathbb{E}_{q_\phi(\boldsymbol{Z})}[\log p(\boldsymbol{\mathcal{D}}, \boldsymbol{Z}|\Theta) - \log q_\phi(\boldsymbol{Z})]$$

$$= \mathbb{E}_{q_\phi(\boldsymbol{Z})}\left[\nabla_\Theta \left\{\log p(\boldsymbol{\mathcal{D}}, \boldsymbol{Z}|\Theta) - \log q_\phi(\boldsymbol{Z})\right\}\right]$$

Gradient can go inside expectation since $q(Z)$ doesn't depend on Θ

Monte-Carlo approximation using samples of $q_\phi(\boldsymbol{Z})$ is straightforward here

- The ELBO's gradient w.r.t. $\boldsymbol{\phi}$

$$\nabla_\phi \mathcal{L}(\phi, \Theta) = \nabla_\phi \, \mathbb{E}_{q_\phi(\boldsymbol{Z})}[\log p(\boldsymbol{\mathcal{D}}, \boldsymbol{Z}|\Theta) - \log q_\phi(\boldsymbol{Z})]$$

$$\neq \mathbb{E}_{q_\phi(\boldsymbol{Z})}\left[\nabla_\phi \left\{\log p(\boldsymbol{\mathcal{D}}, \boldsymbol{Z}|\Theta) - \log q_\phi(\boldsymbol{Z})\right\}\right]$$

Gradient can't go inside expectation since $q(Z)$ depends on ϕ

Monte-Carlo approximation using samples of $q_\phi(\boldsymbol{Z})$ is NOT as straightforward

Black-Box Variational Inference (BBVI)

- Black-box Var. Inference* (BBVI) approximates ELBO derivatives using Monte-Carlo
- Uses the following identity for the ELBO's derivative

$$\begin{aligned} \nabla_\phi \mathcal{L}(q) &= \nabla_\phi \mathbb{E}_q[\log p(\mathbf{X}, \mathbf{Z}) - \log q(\mathbf{Z}|\phi)] \\ &= \mathbb{E}_q[\nabla_\phi \log q(\mathbf{Z}|\phi)(\log p(\mathbf{X}, \mathbf{Z}) - \log q(\mathbf{Z}|\phi))] \quad \text{(proof on next slide)} \end{aligned}$$

- Thus ELBO gradient can be written solely in terms of expec. of gradient of $\log q(\boldsymbol{Z}|\phi)$
  - Required gradients don't depend on the model; only on chosen var. distribution (hence "black-box")
- Given S samples $\{Z_s\}_{s=1}^S$ from $q(\boldsymbol{Z}|\phi)$, we can get (noisy) gradient as follows

$$\nabla_\phi \mathcal{L}(q) \approx \frac{1}{S} \sum_{s=1}^{S} \nabla_\phi \log q(\mathbf{Z}_s|\phi)(\log p(\mathbf{X}, \mathbf{Z}_s) - \log q(\mathbf{Z}_s|\phi))$$

- Above is also called the "score function" based gradient (also REINFORCE method)

Gradient of a log-likelihood or log-probability function w.r.t. its params is called score function; hence the name

*Black Box Variational Inference - Ranganath et al (2014)

Proof of BBVI Identity

- The ELBO gradient can be written as

$$
\begin{aligned}
\nabla_\phi \mathcal{L}(q) &= \nabla_\phi \int (\log p(\mathbf{X},\mathbf{Z}) - \log q(\mathbf{Z}|\phi))q(\mathbf{Z}|\phi)d\mathbf{Z} \\
&= \int \nabla_\phi[(\log p(\mathbf{X},\mathbf{Z}) - \log q(\mathbf{Z}|\phi))q(\mathbf{Z}|\phi)]d\mathbf{Z} \qquad (\nabla \text{ and } \int \text{ interchangeable; dominated convergence theorem}) \\
&= \int \nabla_\phi[(\log p(\mathbf{X},\mathbf{Z}) - \log q(\mathbf{Z}|\phi))]q(\mathbf{Z}|\phi) + \nabla_\phi q(\mathbf{Z}|\phi)[(\log p(\mathbf{X},\mathbf{Z}) - \log q(\mathbf{Z}|\phi))]d\mathbf{Z} \\
&= \mathbb{E}_q[-\nabla_\phi \log q(\mathbf{Z}|\phi)] + \int \nabla_\phi q(\mathbf{Z}|\phi)[(\log p(\mathbf{X},\mathbf{Z}) - \log q(\mathbf{Z}|\phi))]d\mathbf{Z}
\end{aligned}
$$

- Note that $\mathbb{E}_q[\nabla_\phi \log q(\mathbf{Z}|\phi)] = \mathbb{E}_q\left[\frac{\nabla_\phi q(\mathbf{Z}|\phi)}{q(\mathbf{Z}|\phi)}\right] = \int \nabla_\phi q(\mathbf{Z}|\phi)d\mathbf{Z} = \nabla_\phi \int q(\mathbf{Z}|\phi)d\mathbf{Z} = \nabla_\phi 1 = 0$
- Also note that $\nabla_\phi q(\mathbf{Z}|\phi) = \nabla_\phi[\log q(\mathbf{Z}|\phi)]q(\mathbf{Z}|\phi)$, using which

$$
\begin{aligned}
\int \nabla_\phi q(\mathbf{Z}|\phi)[(\log p(\mathbf{X},\mathbf{Z}) - \log q(\mathbf{Z}|\phi))]d\mathbf{Z} &= \int \nabla_\phi \log q(\mathbf{Z}|\phi)[(\log p(\mathbf{X},\mathbf{Z}) - \log q(\mathbf{Z}|\phi))]q(\mathbf{Z}|\phi)d\mathbf{Z} \\
&= \mathbb{E}_q[\nabla_\phi \log q(\mathbf{Z}|\phi)(\log p(\mathbf{X},\mathbf{Z}) - \log q(\mathbf{Z}|\phi))]
\end{aligned}
$$

- Therefore $\nabla_\phi \mathcal{L}(q) = \mathbb{E}_q[\nabla_\phi \log q(\mathbf{Z}|\phi)(\log p(\mathbf{X},\mathbf{Z}) - \log q(\mathbf{Z}|\phi))]$

Benefits of BBVI

- Recall that BBVI approximates the ELBO gradients by the Monte Carlo expectations

$$\nabla_\phi \mathcal{L}(q) \approx \frac{1}{S} \sum_{s=1}^{S} \nabla_\phi \log q(\mathbf{Z}_s|\phi)(\log p(\mathbf{X}, \mathbf{Z}_s) - \log q(\mathbf{Z}_s|\phi))$$

- Enables applying VI for a wide variety of probabilistic models
- Can also work with small minibatches of data rather than full data
- BBVI has very few requirements
  - Should be able to sample from $q(\boldsymbol{Z}|\boldsymbol{\phi})$ (usually sampling routines exists!)
  - Should be able to compute $\nabla_{\boldsymbol{\phi}} \log q(\boldsymbol{Z}|\boldsymbol{\phi})$ (automatic differentiation methods exist!)
  - Should be able to evaluate $\log p(\boldsymbol{X}, \boldsymbol{Z})$ and $\log q(\boldsymbol{Z}|\boldsymbol{\phi})$ for any value of $\boldsymbol{Z}$
- Some tricks needed to control the variance in the Monte Carlo estimate of the ELBO gradient (if interested in the details, please refer to the BBVI paper)

Reparametrization Trick

- Another Monte-Carlo approx. of ELBO grad (with often lower var than BBVI gradient)
- Suppose we want to compute ELBO's gradient $\nabla_\phi \mathbb{E}_{q_\phi(\mathbf{Z})}[\log p(\mathbf{X}, \mathbf{Z}) - \log q_\phi(\mathbf{Z})]$
- Assume a deterministic transformation g

$$\boldsymbol{Z} = g(\epsilon, \phi) \quad \text{where} \quad \epsilon \sim p(\epsilon)$$

Assumed to not depend on ϕ

- With this reparametrization, and using LOTUS rule, the ELBO's gradient would be

$$\nabla_\phi \mathbb{E}_{p(\epsilon)}[\log p(\mathbf{X}, g(\epsilon, \phi)) - \log q_\phi(g(\epsilon, \phi))] = \mathbb{E}_{p(\epsilon)} \nabla_\phi [\log p(\mathbf{X}, g(\epsilon, \phi)) - \log q_\phi(g(\epsilon, \phi))]$$

- Given S i.i.d. random samples $\{\epsilon_s\}_{s=1}^S$ from $p(\epsilon)$, we can get a Monte-Carlo approx.

$$\nabla_\phi \mathbb{E}_{q_\phi(\mathbf{Z})}[\log p(\mathbf{X}, \mathbf{Z}) - \log q_\phi(\mathbf{Z})] \approx \frac{1}{S} \sum_{s=1}^{S} [\nabla_\phi \log p(\mathbf{X}, g(\epsilon_s, \phi)) - \nabla_\phi \log q_\phi(g(\epsilon_s, \phi))]$$

- Such gradients are called pathwise gradients* (since we took a "path" from ϵ to $\boldsymbol{Z}$)

*Autoencoding Variational Bayes - Kingma and Welling (2013), Stochastic Backpropagation and Approximate Inference in Deep Generative Models- Rezende et al (2014)

Reparametrization Trick: An Example

- Suppose our variational distribution is $q(\boldsymbol{w}|\boldsymbol{\phi}) = \mathcal{N}(\boldsymbol{w}|\boldsymbol{\mu}, \boldsymbol{\Sigma})$, so $\boldsymbol{\phi} = \{\boldsymbol{\mu}, \boldsymbol{\Sigma}\}$
- Suppose our ELBO has a difficult expectation term $\mathbb{E}_q[f(\boldsymbol{w})]$

Or $\boldsymbol{\phi} = \{\boldsymbol{\mu}, \mathbf{L}\}$ where $\mathbf{L} = \text{chol}(\Sigma)$

- However, note that we need ELBO gradient, not ELBO itself. Let's use the trick
- Reparametrize $\boldsymbol{w}$ as $\boldsymbol{w} = \boldsymbol{\mu} + \mathbf{L}\mathbf{v}$ where $\mathbf{v} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I})$

Note that we will still have $q(\boldsymbol{w}|\boldsymbol{\phi}) = \mathcal{N}(\boldsymbol{w}|\boldsymbol{\mu}, \boldsymbol{\Sigma})$

$$\nabla_{\mu,\mathbf{L}}\mathbb{E}_{\mathcal{N}(\boldsymbol{w}|\mu,\Sigma)}[f(\boldsymbol{w})] = \nabla_{\mu,\mathbf{L}}\mathbb{E}_{\mathcal{N}(\boldsymbol{v}|0,\mathbf{I})}[f(\mu + \mathbf{L}\boldsymbol{v})] = \mathbb{E}_{\mathcal{N}(\boldsymbol{v}|0,\mathbf{I})}[\nabla_{\mu,\mathbf{L}} f(\mu + \mathbf{L}\boldsymbol{v})]$$

- The above is now straightforward
  - Easily take derivatives of $f(\boldsymbol{w})$ w.r.t. variational params $\boldsymbol{\mu}, \mathbf{L}$
  - Replace exp. by Monte-Carlo averaging using samples of $\mathbf{v}$ from $\mathcal{N}(\mathbf{0}, \boldsymbol{I})$

Often even one or very few samples suffice

$$\nabla_{\mu}\mathbb{E}_{\mathcal{N}(\boldsymbol{w}|\mu,\Sigma)}[f(\boldsymbol{w})] = \mathbb{E}_{\mathcal{N}(\boldsymbol{v}|0,\mathbf{I})}[\nabla_{\mu} f(\mu + \mathbf{L}\boldsymbol{v})] \approx \nabla_{\mu} f(\mu + \mathbf{L}\boldsymbol{v}_s)$$

$$\nabla_{\mathbf{L}}\mathbb{E}_{\mathcal{N}(\boldsymbol{w}|\mu,\Sigma)}[f(\boldsymbol{w})] = \mathbb{E}_{\mathcal{N}(\boldsymbol{v}|0,\mathbf{I})}[\nabla_{\mathbf{L}} f(\mu + \mathbf{L}\boldsymbol{v})] \approx \nabla_{\mathbf{L}} f(\mu + \mathbf{L}\boldsymbol{v}_s)$$

$\frac{\partial f}{\partial \boldsymbol{w}}\frac{\partial \boldsymbol{w}}{\partial \boldsymbol{\mu}}$

Chain Rule

$\frac{\partial f}{\partial \boldsymbol{w}}\frac{\partial \boldsymbol{w}}{\partial \mathbf{L}}$

- Std. reparam. trick assumes differentiability (recent work on removing this req).

*Autoencoding Variational Bayes - Kingma and Welling (2013), Stochastic Backpropagation and Approximate Inference in Deep Generative Models- Rezende et al (2014)

Reparametrization Trick: Some Comments

- Standard Reparametrization Trick assumes the model to be differentiable

$$\nabla_\phi \mathbb{E}_{q_\phi(\mathbf{Z})}[\log p(\mathbf{X},\mathbf{Z}) - \log q_\phi(\mathbf{Z})] = \mathbb{E}_{p(\epsilon)}[\nabla_\phi \log p(\mathbf{X}, g(\epsilon,\phi)) - \nabla_\phi \log q_\phi(g(\epsilon,\phi))]$$

- In contrast, BBVI (score function gradients) only required $\boldsymbol{q(Z)}$ to be differentiable
- Thus rep. trick often isn't applicable, e.g., when $\boldsymbol{Z}$ is discrete (e.g., binary /categorical)
  - Recent work on continuous relaxation† of discrete variables† (e.g., Gumbel Softmax for categorical)
- The transformation function $\boldsymbol{g}$ may be difficult to find for general distributions
  - Recent work on generalized reparametrizations*
- Also, the transformation function $\boldsymbol{g}$ needs to be invertible (difficult/expensive)
  - Recent work on implicit reparametrized gradients#
- Assumes that we can directly draw samples from $p(\epsilon)$. If not, then rep. trick isn't valid@

†Categorical Reparameterization with Gumbel-Softmax (Jang et al, 2017), ∗ The Generalized Reparameterization Gradient (Ruiz et al, 2016), # Implicit Reparameterization Gradients (Figurnov et al, 2018), @ Reparameterization Gradients through Acceptance-Rejection Sampling Algorithms (Naesseth et al, 2016)

Automatic Differentiation Variational Inference

- Suppose $\boldsymbol{Z}$ is $\boldsymbol{D}$-dim r.v. with constraints (e.g., non-negativity) and distribution $\boldsymbol{q(Z|\phi)}$
- Assume a transformation $\boldsymbol{T}$ such that $\boldsymbol{u} = \boldsymbol{T(Z)}$ s.t. $\boldsymbol{u} \in \mathbb{R}^D$ (unconstrained) then

$$q(\boldsymbol{u}) = q(\boldsymbol{Z}) \left| \det\left(\frac{\partial \boldsymbol{Z}}{\partial \boldsymbol{u}}\right) \right|$$

- Recall the original ELBO expression $\mathcal{L}(\boldsymbol{\phi}) = \mathbb{E}_{q(\boldsymbol{Z}|\phi)}\left[\log \frac{p(\boldsymbol{D},\boldsymbol{Z})}{q(\boldsymbol{Z}|\phi)}\right]$
- Assuming $\boldsymbol{q(u|\psi)} = \mathcal{N}(\boldsymbol{u}|\mu, \Sigma)$ and using $\boldsymbol{Z} = \boldsymbol{T^{-1}(u)}$, the ELBO becomes

$$\mathcal{L}(\psi) = \mathbb{E}_{q(\boldsymbol{u}|\psi)}\left[\log \frac{p(\boldsymbol{D}, T^{-1}(\boldsymbol{u}))|\det(\partial \boldsymbol{Z}/\partial \boldsymbol{u})|}{q(\boldsymbol{u}|\psi)}\right]$$

$$= \mathbb{E}_{q(\boldsymbol{u}|\psi)}[\log p(\boldsymbol{D}, T^{-1}(\boldsymbol{u})) + \log|\det(\partial \boldsymbol{Z}/\partial \boldsymbol{u})|] + \mathrm{H}(q(\boldsymbol{u}|\psi))$$

- Optimize $\mathcal{L}(\boldsymbol{\psi})$ w.r.t. $\boldsymbol{\psi}$ to get $\boldsymbol{q(u|\psi)}$ as a Gaussian and use distribution transformation equation to get $\boldsymbol{q(Z|\phi)}$

Structured Variational Inference

- Here "structured" may refer to anything that makes VI approx. more expressive, e.g.,
  - Removing the independence assumption of mean-field VI
  - In general, learning more complex forms for the variational approximation family $q(\boldsymbol{Z}|\phi)$
- To remove the mean-field assumption in VI, various approaches exist
  - Structured mean-field (Saul et al, 1996)
  - Hierarchical VI (Ranganath et al, 2016): Variational params $\phi_1, \phi_2, \ldots, \phi_M$ "tied" via a shared prior

$$q(\boldsymbol{z}_1, \ldots, \boldsymbol{z}_M|\theta) = \int \left[\prod_{m=1}^{M} q(\boldsymbol{z}_m|\phi_m) \right] p(\phi|\theta) d\phi$$

- Recent work on learning more expressive variational approx. for general VI
  - Boosting or mixture of simpler distributions, e.g., $q(\boldsymbol{Z}) = \sum_{c=1}^{C} \rho_c q_c(\boldsymbol{Z})$

    Even simple unimodal components will give a multimodal $q(\boldsymbol{Z})$
  - Normalizing flows*: Turn a simple var. distr. into a complex one via series of invertible transfor.

$$\mathbf{z}_K = f_K \circ \cdots \circ f_1(\mathbf{z}_0), \quad \mathbf{z}_0 \sim q_0(\mathbf{z}_0),$$

A simple unimodal variational distribution (e.g.. $\mathcal{N}(0, I)$)

$$\mathbf{z}_K \sim q_K(\mathbf{z}_K) = q_0(\mathbf{z}_0) \prod_{k=1}^{K} \left| \det \frac{\partial f_k}{\partial \mathbf{z}_{k-1}} \right|^{-1}$$

A much more complex (e.g., multimodal) variational distribution obtained via the flow idea

*Variational Inference with Normalizing Flows (Rezende and Mohamed, 2015)

Other Divergence Measures

- VI minimizes $KL(q||p)$ but other divergences can be minimized as well
  - Recall that VI with minimization of $KL(q||p)$ leads to underestimated variances
- A general form of divergence is Renyi's α-divergence defined as

$$D_\alpha^R(p(\boldsymbol{Z})||q(\boldsymbol{Z})) = \frac{1}{\alpha - 1} \log \int p(\boldsymbol{Z})^\alpha q(\boldsymbol{Z})^{1-\alpha} d\boldsymbol{Z}$$

- $KL(p||q)$ is a special case with $\alpha \rightarrow 1$ (can verify using L'Hopital rule of taking limits)
- An even more general form of divergence is f-Divergence

$$D_f(p(\boldsymbol{Z})||q(\boldsymbol{Z})) = \int q(\boldsymbol{Z}) f\left(\frac{p(\boldsymbol{Z})}{q(\boldsymbol{Z})}\right) d\boldsymbol{Z}$$

- Many recent variational inference algorithms are based on minimizing such divergences

Variational Inference: Some Comments

- Many probabilistic models nowadays rely on VI to do approx. inference
- Even mean-field with locally-conjugacy used in lots of models
  - This + SVI gives excellent scalability as well on large datasets
- Progress in various areas has made VI very popular and widely applicable
  - Stochastic Optimization (e.g., SGD)
  - Automatic Differentiation
  - Monte-Carlo gradient of ELBO

We covered many of the threads being explored in recent work but a lot of work still being done in this area

- Note: Most of these ideas apply also to Variational EM
- Many VI and advanced VI algos are implemented in probabilistic prog. packages (e.g., Tensorflow Probability, PyTorch, etc), making VI easy even for complex models
- Still a very active area of research, especially for doing VI in complex models
  - Models with discrete latent variables
  - Reducing the variance in Monte-Carlo estimate of ELBO gradients
  - More expressive variational distribution for better approximation