Variational Inference (Contd.)

Piyush Rai

Topics in Probabilistic Modeling and Inference (CS698X)

March 20, 2018

Variational Inference (VI)

- VI finds the $q(\mathbf{Z}|\phi)$ that is "closest" to the intractable target posterior $p(\mathbf{Z}|\mathbf{X})$
- Basically, we want q by solving $\arg\min_q \text{KL}(q||p)$
- Since q depends on the free variational parameteras ϕ, this amounts to solving

$$\phi^* = \arg\min_{\phi} \text{KL}[q_\phi(\mathbf{Z})||p(\mathbf{Z}|\mathbf{X})]$$

- Luckily, we never need to deal with $p(\mathbf{Z}|\mathbf{X})$ while solving the above optimization problem!
- Reason: Using the identity $\log p(\mathbf{X}) = \mathcal{L}(q) + \text{KL}(q||p)$ where $\mathcal{L}(q) = \int q(\mathbf{Z}) \log \frac{p(\mathbf{X},\mathbf{Z})}{q(\mathbf{Z})} d\mathbf{Z}$

$$\arg\min_q \text{KL}(q||p) = \arg\max_q \mathcal{L}(q)$$

- Thus VI finds the optimal q by maximizing the evidence lower bound (ELBO) $\mathcal{L}(q)$

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

- Solving $\arg\max_q \mathcal{L}(q)$ requires two things: Computing the expectations, taking derivatives

Mean-Field VI

- Assumes a partition of the latent variables $\mathbf{Z}$ into M groups $\mathbf{Z}_1, \ldots, \mathbf{Z}_M$
- Assumes that our approximation $q(\mathbf{Z})$ factorizes over these groups

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

- Greatly simplifies $\arg\max_q \mathcal{L}(q)$ with very simple update for each optimal factor

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

  where $\mathbb{E}_{i\neq j}$ denotes expectation w.r.t. current optimal q_i's of all other groups
- The above solution can also be written as

$$\boxed{q_j^*(\mathbf{Z}_j) = \frac{\exp(\mathbb{E}_{i\neq j}[\log p(\mathbf{X}, \mathbf{Z})])}{\int \exp(\mathbb{E}_{i\neq j}[\log p(\mathbf{X}, \mathbf{Z})])d\mathbf{Z}_j}}$$

- We can take a cyclic approach and update one q_j at a time, keeping all others fixed

Mean-Field VI: A Very Simple Example

- Consider data $\mathbf{X} = \{x_1, \ldots, x_N\}$ from a 1-D Gaussian $\mathcal{N}(x|\mu, \tau^{-1})$ with mean μ, precision τ
- Assume the following normal-gamma prior on μ and τ

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

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

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

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

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

Mean-Field VI: A Very Simple Example (Contd)

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

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

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

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

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

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

- Important: Updates of $q_\mu^*(\mu)$ and $q_\tau^*(\tau)$ depend on each-other (thus requires cyclic updates)

VI by explicitly taking ELBO's derivatives

- In mean-field VI, we could simply "read off" the solution for each q_j
- But VI in general is an optimization problem!
- A more general way of doing VI is to explicitly write down ELBO and optimize w.r.t. ϕ

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

- Let's do this for linear regression model (note: The "reading off" approach can be used too)

$$p(y_n|\boldsymbol{w}, \boldsymbol{x}_n) = \mathcal{N}(y_n|\boldsymbol{w}^\top \boldsymbol{x}_n, \beta^{-1}), \quad p(\boldsymbol{w}) = \mathcal{N}(\boldsymbol{w}|0, \lambda^{-1}\mathbf{I}), \quad p(\beta) = \text{Gamma}(a, b)$$

we'll assume λ to be known, so the unknowns are $\boldsymbol{w}, \beta$

VI by explicitly taking ELBO's derivatives

- Let's assume $q(\boldsymbol{w}, \beta) = q_w(\boldsymbol{w})q_\beta(\beta) = \mathcal{N}(\boldsymbol{w}|\mu_N, \Sigma_N)\text{Gamma}(\beta|a_N, b_N)$
- With this and denoting $\mathbf{Z} = (\boldsymbol{w}, \beta)$, the ELBO will be

$$\mathcal{L}(q) = \int q(\mathbf{Z}) \log p(\boldsymbol{y}|\mathbf{Z}, \mathbf{X}) d\mathbf{Z} + \int q(\mathbf{Z}) \log p(\mathbf{Z}) d\mathbf{Z} - \int q(\mathbf{Z}) \log q(\mathbf{Z}) d\mathbf{Z}$$

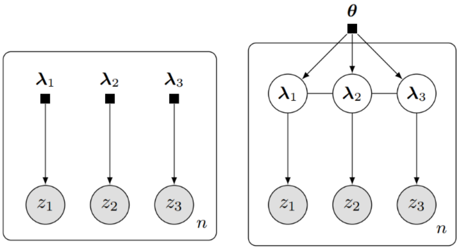
- Further using $p(\boldsymbol{y}|\mathbf{Z}, \mathbf{X}) = p(\boldsymbol{y}|\mathbf{X}, \boldsymbol{w}, \beta)$ and $p(\mathbf{Z}) = p(\boldsymbol{w}, \beta) = p(\boldsymbol{w})p(\beta)$, we get

$$\begin{aligned}\mathcal{L}(\mu_N, \Sigma_N, a_N, b_N) &= \sum_{n=1}^{N} \int \int q(\boldsymbol{w})q(\beta) \log p(y_n|\boldsymbol{w}, \boldsymbol{x}_n, \beta) d\boldsymbol{w} d\beta \\ &+ \int q(\boldsymbol{w}) \log p(\boldsymbol{w}) d\boldsymbol{w} + \int q(\beta) \log p(\beta) d\beta \\ &- \int q(\boldsymbol{w}) \log q(\boldsymbol{w}) d\boldsymbol{w} - \int q(\beta) \log q(\beta) d\beta\end{aligned}$$

- Can now take gradients w.r.t. each of $a_N, b_N, \mu_N, \Sigma_N$ and estimate these in an alternating fashion
- This will give us $q(\boldsymbol{w}, \alpha) = \text{Normal}(\boldsymbol{w}|\mu_N, \Sigma_N)\text{Gamma}(\alpha|a_N, b_N)$

Beyond Mean-Field

- Many approaches that relax the mean-field assumption
- A recent approach is based on "hierarchical variational model" (Ranganath et al, ICML 2016)
  - Same structure as traditional mean-field but variational parameters "tied" via a shared prior



$$q_{\mathrm{MF}}(\mathbf{z};\boldsymbol{\lambda}) = \prod_{i=1}^{d} q(z_i;\boldsymbol{\lambda}_i)$$

Traditional Fully-Factorized Mean-Field

$$q_{\mathrm{HVM}}(\mathbf{z};\boldsymbol{\theta}) = \int q(\boldsymbol{\lambda};\boldsymbol{\theta}) \prod_i q(z_i \,|\, \boldsymbol{\lambda}_i)\, \mathrm{d}\boldsymbol{\lambda}$$

Hierarchical Variariational Mean-Field

VI with non-conjugate models (Some approaches)

Some Model-Specific Tricks

- ELBO $\mathcal{L}(q) = \mathbb{E}_q[\log p(\mathbf{X}, \mathbf{Z})] - \mathbb{E}_q[\log q(\mathbf{Z})]$ requires computing expectations w.r.t. var. dist. q
- The ELBO and its derivatives can be difficult to compute for non-conjugate models
- A common approach is to replace each difficult terms by a tight lower bound. Some examples:
- Assuming $q(a,b) = \prod_i q(a_i)q(b_i)$, the expectation below can be replaced by a lower bound

$$\mathbb{E}_q\left[\log \sum_i a_i b_i\right] = \mathbb{E}_q\left[\log \sum_i p_i \frac{a_i b_i}{p_i}\right] \geq \mathbb{E}_q\left[\underbrace{\sum_i p_i \log \frac{a_i b_i}{p_i}}_{\text{via Jensen's inequality}}\right] = \sum_i p_i \mathbb{E}_q[\log a_i + \log b_i] - \sum_i p_i \log p_i$$

  where p_i is a variable (depends on a_i and b_i) that we need to optimize. Expectations above easy to compute
- For models with logistic likelihood, we use the following (trick by Jaakkola and Jordan, 2000)

$$-\mathbb{E}_q[\log(1+\exp(-y_n \boldsymbol{w}^\top \boldsymbol{x}_n))] \geq \log \sigma(\xi_n) + \mathbb{E}_q\left[\frac{1}{2}(y_n \boldsymbol{w}^\top \boldsymbol{x}_n - \xi_n) - \lambda(\xi_n)(\boldsymbol{w}^\top \boldsymbol{x}_n \boldsymbol{x}_n^\top \boldsymbol{w} - \xi_n^2)\right]$$

  where ξ_n is a variable (depends on $\boldsymbol{w}$) that we need to optimize. Expectations above easy to compute

Reparametrization Trick

- Instead of model-specific tricks, recent work on VI has focused on general methods for computing ELBO and its derivatives for non-conjugate models
- A Monte-Carlo sampling based approximation of the derivatives is via the Reparametrization Trick
- Example: Suppose our var. dist. is $q(\boldsymbol{w}) = \mathcal{N}(\boldsymbol{w}|\mu, \Sigma)$. Suppose the ELBO has a term $\mathbb{E}_q[f(\boldsymbol{w})]$
- We can reparametrize $\boldsymbol{w}$ as $\boldsymbol{w} = \mu + \mathbf{L}\boldsymbol{v}$ where $\mathbf{L} = \text{chol}(\Sigma)$ and $\boldsymbol{v} \sim \mathcal{N}(0, \mathbf{I})$, and write

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

- It is now easy to take derivatives w.r.t. variational parameters using Monte Carlo sampling, e.g.,

$$\begin{aligned}
\nabla_\mu \mathbb{E}_{\mathcal{N}(\boldsymbol{w}|\mu,\Sigma)}[f(\boldsymbol{w})] &= \mathbb{E}_{\mathcal{N}(\boldsymbol{v}|0,\mathbf{I})}[\nabla_\mu f(\mu + \mathbf{L}\boldsymbol{v})] \approx \nabla_\mu f(\mu + \mathbf{L}\boldsymbol{v}_s) \quad \text{where } \boldsymbol{v}_s \sim \mathcal{N}(\boldsymbol{v}|0, \mathbf{I}) \\
\nabla_{\mathbf{L}} \mathbb{E}_{\mathcal{N}(\boldsymbol{w}|\mu,\Sigma)}[f(\boldsymbol{w})] &\approx \nabla_{\mathbf{L}} f(\mu + \mathbf{L}\boldsymbol{v}_s) = \nabla_{\boldsymbol{w}}[f(\boldsymbol{w})]\boldsymbol{v}_s^\top
\end{aligned}$$

.. the above just requires being able to take derivatives of $f(\boldsymbol{w})$ w.r.t. $\boldsymbol{w}$

* Autoencoding Variational Bayes - Kingma and Welling (2013)

Black-box Variational Inference (BBVI)

- Black-box Variational Inference (BBVI) also approximate ELBO derivatives using Monte-Carlo
- BBVI 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 expectation of gradient of $\log q(\mathbf{Z}|\phi)$
  - Required gradients don't depend on the model. Only on the chosen variational distribution
  - That's why this approach is called "black-box"
- Given S samples $\{\mathbf{Z}_s\}_{s=1}^S$ from $q(\mathbf{Z}|\phi)$, we can get (noisy) gradient $\nabla_\phi \mathcal{L}(q)$ as follows

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

* 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 VB inference for a wide variety of probabilistic models
- Very few requirements
  - Should be able to sample from $q(\mathbf{Z}|\phi)$
  - Should be able to compute $\nabla_\phi \log q(\mathbf{Z}|\phi)$ (automatic differentiation methods exist!)
  - Should be able to evaluate $p(\mathbf{X}, \mathbf{Z})$ and $\log q(\mathbf{Z}|\phi)$
- 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)

Some Properties of VI

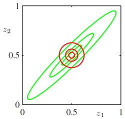
Recall that VI is equivalent to finding q by minimizing $\text{KL}(q||p)$

$$\text{KL}(q||p) = \int q(\mathbf{Z}) \log \frac{q(\mathbf{Z})}{p(\mathbf{Z})} d\mathbf{Z}$$

If the true posterior p is very small in some region then, to minimize $\text{KL}(q||p)$, the approx. dist. q will also have to be very small (otherwise KL will be very large)

This has two key consequences for VI

- Underestimates the variances of the true posterior
- For multimodal posteriors, VI locks onto one of the modes



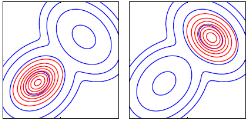


Figure: (Left) Zero-Forcing Property of VI, (Right) For multi-modal posterior, VB locks onto one of the models

Note: **Expectation Propagation (EP)** which minimizes $\text{KL}(p||q)$ instead can avoid this behavior