Intro to Variational Inference

CS698X: Topics in Probabilistic Modeling and Inference

Piyush Rai

Variational Bayes (VB) or Variational Inference (VI)

- Consider a model with data $\boldsymbol{X}$ and unknowns $\boldsymbol{Z}$. Goal: Compute the posterior $p(\boldsymbol{Z}|\boldsymbol{X})$
- $\boldsymbol{Z}$ denotes <u>all unknowns</u> (params, latent vars, hyperparams of likelihood, prior, etc)

Defines a class of distributions parametrized by ϕ

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

Often called variational parameters

- We find the best approx. distr by finding ϕ s.t. its <u>distance</u> from $p(\boldsymbol{Z}|\boldsymbol{X})$ is minimized

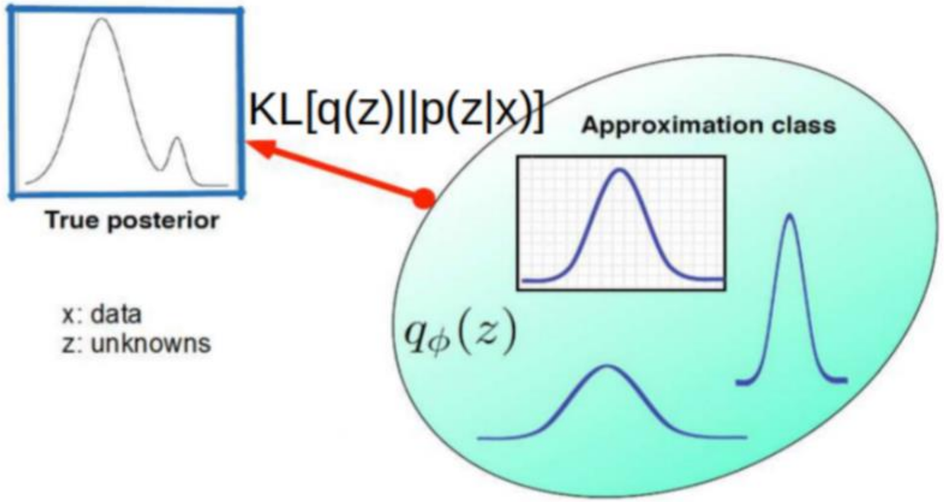
But since we don't know $p(\boldsymbol{Z}|\boldsymbol{X})$, can we easily solve this optimization problem?

Often, we will simply write it as $\text{argmin}_q \ \text{KL}[q||p_z]$

Other measures have also been used such as reverse KL ($\text{KL}[p||q]$), and various other divergence functions defined for distributions

VI turns inference into optimization

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



- Note: The name "variational" comes from Physics
  - Optimizing functions of distributions (KL is a func of distr)

Variational Bayes (VB) or Variational Inference (VI)

- VB/VI is based on following identity for the log marg-lik (log evidence) of a model $\boldsymbol{m}$

Similar as the identify we had in case of EM, which was defined for log of the ILL

$$\log p(\boldsymbol{X}|m) = \mathcal{L}(q) + \mathrm{KL}(q||p_z)$$

Unlike EM, we don't have any distinction b/w latent var and parameters (all unknowns are being called $\boldsymbol{Z}$ here)

$$\mathcal{L}(q) = \int q(\boldsymbol{Z})\log\left\{\frac{p(\boldsymbol{X},\boldsymbol{Z})}{q(\boldsymbol{Z})}\right\}d\boldsymbol{Z} \qquad \mathrm{KL}(q||p_z) = -\int q(\boldsymbol{Z})\log\left\{\frac{p(\boldsymbol{Z}|\boldsymbol{X})}{q(\boldsymbol{Z})}\right\}d\boldsymbol{Z}$$

- Since the log evidence $\log p(\boldsymbol{X}|m)$ is constant w.r.t $\boldsymbol{Z}$, we must have

$$\mathrm{argmin}_q \ \mathrm{KL}[q||p_z] = \mathrm{argmax}_q \ \mathcal{L}(q)$$

- Also note that since $\mathrm{KL}[q||p_z] \geq 0$, we must have $\log p(\boldsymbol{X}|m) \geq \mathcal{L}(q)$

- Therefore, $\mathcal{L}(q)$ is also known as Evidence Lower Bound (ELBO)
  - VB/VI finds the best $q(Z)$ by maximizing the ELBO w.r.t. q

VB/VI = Maximizing the ELBO

- Notation: $\boldsymbol{q(Z)}$, $\boldsymbol{q(Z|\phi)}$, $\boldsymbol{q_\phi(Z)}$, all refer to the same thing (the approx. distr.)
- VB/VI finds an approximating distribution $\boldsymbol{q(Z)}$ that maximizes the ELBO

$$\mathcal{L}(q) = \int q(\boldsymbol{Z}) \log \left[\frac{p(\boldsymbol{X},\boldsymbol{Z})}{q(\boldsymbol{Z})}\right] d\boldsymbol{Z}$$

- Since $\boldsymbol{q(Z)}$ depends on $\boldsymbol{\phi}$, the ELBO is essentially a function of $\boldsymbol{\phi}$

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

- Thus maximizing the ELBO will give an approximating distr. $\boldsymbol{q(Z)}$ which
  - Explains the data well, i.e., gives it large probability (large $\mathbb{E}_q[\log p(\boldsymbol{X}|\boldsymbol{Z})]$)
  - Is close to the prior $\boldsymbol{p(Z)}$, i.e. is simple/regularized (small $\mathrm{KL}[\log q(\boldsymbol{Z})||p(\boldsymbol{Z})]$)

Maximizing the ELBO

- The goal is to maximize the ELBO

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

$$= \mathbb{E}_q[\log p(\boldsymbol{X}|\boldsymbol{Z})] - \text{KL}[\log q(\boldsymbol{Z})||p(\boldsymbol{Z})]$$

- This may still be hard because
  - ELBO requires expectations to computed which may be intractable
  - Maximizing the ELBO will require computing gradients which may not always be easy

E.g., part of the ELBO may have terms that are not differentiable

- Some of the ways to make this problem easier
  - Restricting the form of our approximation $q(\boldsymbol{Z})$, e.g., mean-field VB (today's discussion)
  - Using Monte-Carlo approximation of the expectation/gradient of the ELBO (later)
- For locally conjugate models, ELBO maximization is easy
  - Closed form updates for $q(\boldsymbol{Z})$

Mean-Field VI

The name "mean-field" comes from statistical physics literature

- One of the simplest ways for doing VB/VI
- Assumes unknowns $\boldsymbol{Z}$ can be partitioned into $\boldsymbol{M}$ groups $\boldsymbol{Z}_1, \boldsymbol{Z}_2, \ldots, \boldsymbol{Z}_M$, s.t.,

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

As a shorthand, often written as $q = \prod_{i=1}^{M} q_i$ where $q_i = q(Z_i|\phi_i)$

- Learning the optimal $\boldsymbol{q}$ reduces to learning the optimal $\boldsymbol{q_1, q_2, \ldots, q_M}$
- Groups usually chosen based on model's structure, e.g., in Bayesian linear regression

$$p(\boldsymbol{w}, \beta, \lambda|X, y) \approx q(Z|\phi) = q(\boldsymbol{w}, \beta, \lambda|\phi) = q(\boldsymbol{w}|\phi_w)p\left(\beta\middle|\phi_\beta\right)p(\lambda|\phi_\lambda)$$

- Mean-field is a very restrictive assumption. Ignores the correlations among unknowns
  - Less restrictive versions also exist, such as structured mean-field (factorization is still there but only among groups of unknowns)

Deriving Mean-Field VI Updates

- With $q = \prod_{i=1}^{M} q_i$, what's the optimal q_i is when we do $\text{argmax}_q \, \mathcal{L}(q)$?
- Note that under this mean-field assumption, the ELBO simplifies to

$$\mathcal{L}(q) = \int q(\boldsymbol{Z}) \log \left[\frac{p(\boldsymbol{X}, \boldsymbol{Z})}{q(\boldsymbol{Z})} \right] d\boldsymbol{Z} = \int \prod_i q_i \left[\log p(X, Z) - \sum_i \log q_i \right] d\boldsymbol{Z}$$

- Suppose we wish to find the optimal q_j given all other q_i's $(i \neq j)$ as fixed, then

$$\mathcal{L}(q) = \int q_j \left[\int \log p(X, Z) \prod_{i \neq j} q_i \, Z_i \right] Z_j - \int q_j \log q_j Z_j + \text{const w.r.t. } q_j$$

$$= \int q_j \log \hat{p}(X, Z_j) \, Z_j - \int q_j \log q_j Z_j$$

$$= -\text{KL}(q_j || \hat{p})$$

$\log \hat{p}(X, Z_j) = \mathbb{E}_{i \neq j}[\log p(\boldsymbol{X}, \boldsymbol{Z})]$ + const

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

- Thus $q_j^* = \text{argmax}_{q_j} \mathcal{L}(q) = \text{argmin}_{q_j} \text{KL}(q_j || \hat{p}) = \hat{p}(X, Z_j)$

Deriving Mean-Field VI Updates

- So we saw that the optimal $\boldsymbol{q_j}$ when doing mean-field VI is

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

- Note: Can often just compute the numerator and recognize denominator by inspection
- Important: For locally conjugate models, $q_j^*(\boldsymbol{Z}_j)$ will have the same form as prior $p(Z_j)$
  - Only the distribution parameters will be different
- Important: For estimating $\boldsymbol{q_j}$ the required expectation depends on other $\{q_i\}_{i\neq j}$
  - Thus we use an alternating update scheme (ALT-OPT, Gibbs sampling, etc)
- Guaranteed to converge (to a local optima)
  - We are basically solving a sequence of concave maximization problems
  - Reason: $\mathcal{L}(q) = \int q_j \log \hat{p}(X, Z_j)\, Z_j \; - \; \int q_j \log q_j Z_j$ is concave in $\boldsymbol{q_j}$

The Mean-Field VI Algorithm

- Also known as Co-ordinate Ascent Variational Inference (CAVI) Algorithm
- Input: Model in form of priors and likelihood, or joint $p(\boldsymbol{X}, \boldsymbol{Z})$, Data $\boldsymbol{X}$
- Output: A variational distribution $q(Z) = \prod_{j=1}^{M} q_j(\boldsymbol{Z}_j)$
- Initialize: Variational distributions $q_j(\boldsymbol{Z}_j)$, $j = 1,2, \dots M$
- While the ELBO has not converged
  - For each $j = 1,2, \dots M$, set

$$q_j(\boldsymbol{Z}_j) \propto \exp(\mathbb{E}_{i \neq j}[\log p(\boldsymbol{X}, \boldsymbol{Z})])$$

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

Mean-Field VI: A Simple Example

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

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

- Posterior is also normal-gamma due to the jointly conjugate prior
- Let's try mean-field VI nevertheless to illustrate the idea
- 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

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

Mean-Field VI: A Simple Example

- 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 $q_\mu^* = \mathcal{N}(\mu|\mu_N,\lambda_N)$ 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: A Closer Look

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

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

For any model

- Thus opt variational distr $q_j^*(\boldsymbol{Z}_j)$ basically requires expectations of CP $p(\boldsymbol{Z}_j|\boldsymbol{X},\boldsymbol{Z}_{-j})$
- For locally conjugate models, CP can be easily found and is usually an exp-fam distr

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

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

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

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

For locally conjugate model

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

VI by Computing ELBO Gradients

Modern VI methods (e.g., those used in Bayesian deep learning) use this idea (more later)

- Can also do VI by computing ELBO's gradient and doing gradient ascent/descent
- Gradient based approach is broadly applicable, not just for mean-field VI
    1. Assume $q(\boldsymbol{Z})$ to be from some family of distributions with variational parameters $\boldsymbol{\phi}$
    2. Write down the full ELBO expression (will give us a function of var parameters $\boldsymbol{\phi}$)

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

    3. Compute ELBO gradients, i.e., $\nabla_{\boldsymbol{\phi}} \mathcal{L}(\boldsymbol{\phi})$ and use gradient methods to find optimal $\boldsymbol{\phi}$
- Step 2 may be simplified due to the problem structure or the form of $q(\boldsymbol{Z})$
    - i.i.d. observations simplify $\log p(\boldsymbol{X}|\boldsymbol{Z})$; conditionally independent priors simplify $\log p(\boldsymbol{Z})$
    - Locally-conjugate models
    - The mean-field assumption simplifies $q(\boldsymbol{Z})$ as $q = \prod_{i=1}^{M} q_i$
        - Moreover, the last term reduces to sum of entropies of q_i's (which usually has known forms)

Coming Up Next

- VI for latent variable models with local and global unknowns
- Some properties of VI
- VI for non-conjugate models