Introduction to Variational Inference

Piyush Rai

Topics in Probabilistic Modeling and Inference (CS698X)

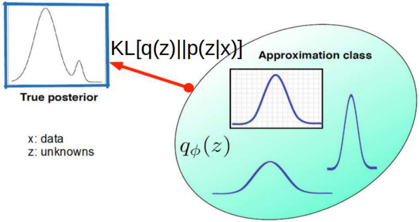
March 15, 2018

Variational Bayes (VB) or Variational Inference (VI)

- Consider a model with data $\mathbf{X}$ and unknowns $\mathbf{Z}$. Goal: Compute the posterior $p(\mathbf{Z}|\mathbf{X})$
- Assuming $p(\mathbf{Z}|\mathbf{X})$ is intractable, VB/VI approximates it using a distribution $q(\mathbf{Z}|\phi)$ or $q_\phi(\mathbf{Z})$
- VB/VI finds the $q(\mathbf{Z}|\phi)$ that is "closest" to $p(\mathbf{Z}|\mathbf{X})$ by finding the "optimal" value of ϕ

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

- This amounts of finding the best distribution from a class of distributions parametrized by ϕ



- VB/VI refers to the free parameters ϕ as variational parameters (w.r.t. which we optimize)
- **But wait!** If $p(\mathbf{Z}|\mathbf{X})$ itself is intractable, can we (easily) solve the above KL minimization problem?

Variational Bayes (VB) or Variational Inference (VI)

- The key identity central to VB/VI is the following (holds for any choice of q)

$$\ln p(\mathbf{X}) = \mathcal{L}(q) + \mathrm{KL}(q\|p)$$

$$\mathcal{L}(q) = \int q(\mathbf{Z}) \ln \left\{ \frac{p(\mathbf{X}, \mathbf{Z})}{q(\mathbf{Z})} \right\} \mathrm{d}\mathbf{Z}$$

$$\mathrm{KL}(q\|p) = -\int q(\mathbf{Z}) \ln \left\{ \frac{p(\mathbf{Z}|\mathbf{X})}{q(\mathbf{Z})} \right\} \mathrm{d}\mathbf{Z}$$

- Note that we saw something similar in EM. But, unlike EM, there's no "parameters" Θ vs latent variables $\mathbf{Z}$ distinction. We are treating everything as latent variables and collectively call it $\mathbf{Z}$
- Also, unlike EM, in VB/VI, we will infer the full posterior for all the latent variables
- Since $\log p(\mathbf{X})$ is a constant w.r.t. $\mathbf{Z}$, the following must hold: $\arg\min_q \mathrm{KL}(q||p) = \arg\max_q \mathcal{L}(q)$
- Good news: Unlike $\mathrm{KL}(q||p)$, $\mathcal{L}(q)$ does NOT depend on $p(\mathbf{Z}|\mathbf{X})$, so $\arg\max_q \mathcal{L}(q)$ is easier!
- Since $\mathrm{KL}(q||p) \geq 0$, $\ln p(\mathbf{X}) \geq \mathcal{L}(q)$. $\mathcal{L}(q)$ is also known as the **Evidence Lower Bound (ELBO)**
  - Reason for the name "ELBO": $\ln p(\mathbf{X})$ or $\ln p(\mathbf{X}|m)$ is the log-evidence of model m

Approximate Inference by Maximizing the ELBO

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

$$\mathcal{L}(q) = \int q(\mathbf{Z}) \ln \left\{ \frac{p(\mathbf{X}, \mathbf{Z})}{q(\mathbf{Z})} \right\} d\mathbf{Z}$$

- Since $q(\mathbf{Z})$ depends on ϕ, the ELBO is essentially a function of ϕ (and only ϕ)

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

- Makes sense: Maximizing $\mathcal{L}(q)$ will give a q that explains data well and is close to the prior
- Maximizing $\mathcal{L}(q)$ w.r.t. q can still be hard in general (note the expectation w.r.t. q)
- Some of the ways to make this problem easier
  1. Restricting the form of the q distribution, e.g., mean-field VB inference (today's discussion)
  2. Using Monte-Carlo approximation of the expectation/gradient of the ELBO (later)
- For conditionally/locally conjugate models VB/VI is particularly easy to derive

Conditional/Local Conjugacy

- Consider a general model with data $\mathbf{X}$ and K unknowns $\mathbf{Z} = (\mathbf{Z}_1, \mathbf{Z}_2, \ldots, \mathbf{Z}_K)$
- Suppose the overall likelihood $p(\mathbf{X}|\mathbf{Z})$ and the joint prior $p(\mathbf{Z})$ aren't conjugate
- In this case, the posterior $p(\mathbf{Z}|\mathbf{X}) = \frac{p(\mathbf{X}|\mathbf{Z})p(\mathbf{Z})}{p(\mathbf{X})}$ wil be intractable
- However suppose, given $\mathbf{Z}_{-k}$, we can compute the following conditional posterior tractably

$$p(\mathbf{Z}_k|\mathbf{X}, \mathbf{Z}_{-k}) = \frac{p(\mathbf{X}|\mathbf{Z}_k, \mathbf{Z}_{-k})p(\mathbf{Z}_k)}{\int p(\mathbf{X}|\mathbf{Z}_k, \mathbf{Z}_{-k})p(\mathbf{Z}_k)d\mathbf{Z}_k} \propto p(\mathbf{X}|\mathbf{Z}_k, \mathbf{Z}_{-k})p(\mathbf{Z}_k)$$

  .. which would be possible if $p(\mathbf{X}|\mathbf{Z}_k, \mathbf{Z}_{-k})$ and $p(\mathbf{Z}_k)$ are conjugate to each other
- Such models are called "locally conjugate" models
- Important: In the above context, when considering the likelihood $p(\mathbf{X}|\mathbf{Z}_k, \mathbf{Z}_{-k})$
  - $\mathbf{X}$ actually refers to only that part of data $\mathbf{X}$ that depends on $\mathbf{Z}_k$
  - Local/conditional conjugacy makes it easy to do inference using VB (and using Gibbs sampling)

Mean-Field VB

- One of the simplest ways of doing VB
- In mean-field VB, we define a partition of the latent variables $\mathbf{Z}$ into M groups $\mathbf{Z}_1, \ldots, \mathbf{Z}_M$
- Assume our approximation $q(\mathbf{Z})$ factorizes over these groups

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

- As a short-hand, sometimes we write $q = \prod_{i=1}^{M} q_i$ where $q_i = q(\mathbf{Z}_i|\phi_i)$
- In mean-field VB, learning the optimal q reduces to learning the optimal $q_1, \ldots, q_M$
- The groups are usually chosen based on the model's structure, e.g., in Bayesian linear regression

$$q(\mathbf{Z}|\phi) = q(\boldsymbol{w}, \lambda, \beta|\phi) = q(\boldsymbol{w}|\phi_w)q(\lambda|\phi_\lambda)q(\beta|\phi_\beta)$$

- Note: Mean-field is quite a strong assumption (can destroy structure among latent variables)

Mean-Field VB

- With $q = \prod_{j=1}^{M} q_j$, what's each optimal q_j equal to when we do $\arg\max_q \mathcal{L}(q)$?
- Note that under this mean-field assumption, the ELBO simplifies to

$$\mathcal{L}(q) = \int q(\mathbf{Z}) \ln \left\{ \frac{p(\mathbf{X},\mathbf{Z})}{q(\mathbf{Z})} \right\} \mathrm{d}\mathbf{Z}$$

$$\begin{aligned}
\mathcal{L}(q) &= \int \prod_i q_i \left\{ \ln p(\mathbf{X},\mathbf{Z}) - \sum_i \ln q_i \right\} \mathrm{d}\mathbf{Z} \\
&= \int q_j \left\{ \int \ln p(\mathbf{X},\mathbf{Z}) \prod_{i \neq j} q_i \, \mathrm{d}\mathbf{Z}_i \right\} \mathrm{d}\mathbf{Z}_j - \int q_j \ln q_j \, \mathrm{d}\mathbf{Z}_j + \text{const} \\
&= \int q_j \ln \widetilde{p}(\mathbf{X},\mathbf{Z}_j) \, \mathrm{d}\mathbf{Z}_j - \int q_j \ln q_j \, \mathrm{d}\mathbf{Z}_j + \text{const}
\end{aligned}$$

where we have defined a new distribution $\widetilde{p}(\mathbf{X},\mathbf{Z}_j)$ by the relation

$$\ln \widetilde{p}(\mathbf{X},\mathbf{Z}_j) = \mathbb{E}_{i \neq j}[\ln p(\mathbf{X},\mathbf{Z})] + \text{const}.$$

- Here $\mathbb{E}_{i \neq j}$ denotes expectation w.r.t. the q distribution except component q_j

$$\mathbb{E}_{i \neq j}[\ln p(\mathbf{X},\mathbf{Z})] = \int \ln p(\mathbf{X},\mathbf{Z}) \prod_{i \neq j} q_i \, \mathrm{d}\mathbf{Z}_i$$

- In the above, $\mathcal{L}(q) = -KL(q_j || \tilde{p}) + \text{const}$. Which q_j will maximize it? Answer: $q_j = \tilde{p}(\mathbf{X},\mathbf{Z}_j)$

Mean-Field VB

- The optimal $q_j^*(\mathbf{Z}_j)$ is therefore equal to $\tilde{p}(\mathbf{X}, \mathbf{Z}_j)$
- Since $\ln q_j^*(\mathbf{Z}_j) = \ln \tilde{p}(\mathbf{X}, \mathbf{Z}_j) = \mathbb{E}_{i \neq j}[\ln p(\mathbf{X}, \mathbf{Z})] + \text{const}$, we have

$$q_j^\star(\mathbf{Z}_j) = \frac{\exp\left(\mathbb{E}_{i \neq j}[\ln p(\mathbf{X}, \mathbf{Z})]\right)}{\int \exp\left(\mathbb{E}_{i \neq j}[\ln p(\mathbf{X}, \mathbf{Z})]\right) \, \mathrm{d}\mathbf{Z}_j}$$

- Note: Only need to compute the numerator. Denominator can usually be recognized by inspection
- Important: For estimating q_j, the required expectation depends on other $\{q_i\}_{i \neq j}$
- Thus we need to cycle through updating each q_j in turn (similar to co-ordinate ascent, alternating optimization, Gibbs sampling, etc.)
- Guaranteed to converge (to a local optima)
  - We are basically solving a sequence of concave maximization problems
  - Reason: $\mathcal{L}(q) = \int q_j \ln \tilde{p}(\mathbf{X}, \mathbf{Z}_j)\mathbf{Z}_j - \int q_j \ln q_j d\mathbf{Z}_j + \text{const}$ is concave w.r.t. each q_j

The Mean-Field VB Algorithm

- Also known as Co-ordinate Ascent Variational Inference (CAVI) Algorithm
- Input: Model $p(\mathbf{X}, \mathbf{Z})$, Data $\mathbf{X}$
- Output: A variational distribution $q(\mathbf{Z}) = \prod_{j=1}^{M} q_j(\mathbf{Z}_j)$
- Initialize: Variational distributions $q_j(\mathbf{Z}_j)$, $j = 1, \ldots, M$
- While the ELBO has not converged
  - For each $j = 1, \ldots, M$, set

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

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

Mean-Field VB by explicitly taking ELBO's derivatives

- Assume some form for each $q_i(\mathbf{Z}_i)$ (e.g., the same distribution as the prior $p(\mathbf{Z}_i)$)
  - Suppose the free variational parameters of $q_i(\mathbf{Z}_i)$ are ϕ_i (say mean and variance of the Gaussian)
- Now write down the **full ELBO** expression

$$
\begin{aligned}
\mathcal{L}(q) = \mathcal{L}(\phi_1, \ldots, \phi_M) &= \mathbb{E}_q[\log p(\mathbf{X}, \mathbf{Z})] - \mathbb{E}_q[\log q(\mathbf{Z})] \\
&= \int q(\mathbf{Z}) \log p(\mathbf{X}, \mathbf{Z}) d\mathbf{Z} - \int q(\mathbf{Z}) \log q(\mathbf{Z}) d\mathbf{Z} \\
&= \int q(\mathbf{Z}) \log p(\mathbf{X}|\mathbf{Z}) d\mathbf{Z} + \int q(\mathbf{Z}) \log p(\mathbf{Z}) d\mathbf{Z} - \int q(\mathbf{Z}) \log q(\mathbf{Z}) d\mathbf{Z}
\end{aligned}
$$

- Note: The above may get further simplified due to independence structures, e.g.,
  - i.i.d. observations simplify $\log p(\mathbf{X}|\mathbf{Z})$; conditionally independent priors simplify $\log p(\mathbf{Z})$
  - The mean-field assumption simplifies $q(\mathbf{Z})$ as $q(\mathbf{Z}) = \prod_{i=1}^{M} q_i(\mathbf{Z}_i)$
  - Note that the last term reduces to sum of entropies of q_i's (which usually has known forms)
- Now take partial derivatives of $\mathcal{L}(\phi_1, \ldots, \phi_M)$ w.r.t. $\phi_1, \ldots, \phi_M$ to find their optimal values
  - Note: Each ϕ_j usually depends on the other ϕ_i's ($i \neq j$). Co-ordinate ascent/descent like procedure

Mean-Field VB for Exponential Family

Mean-Field VB updates are easy to identify/derive if likelihoods/priors are in exponential family

- Let's assume some model with data $\mathbf{X} = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N\}$, local latent variables $\mathbf{Z} = \{\boldsymbol{z}_1, \ldots, \boldsymbol{z}_N\}$
- For many models, the joint distribution of $\mathbf{X}$ and $\mathbf{Z}$ is a product of exp. family distributions

$$p(\mathbf{X}, \mathbf{Z}|\boldsymbol{\eta}) = \prod_{n=1}^{N} h(\mathbf{x}_n, \mathbf{z}_n) g(\boldsymbol{\eta}) \exp\left\{\boldsymbol{\eta}^{\mathrm{T}} \mathbf{u}(\mathbf{x}_n, \mathbf{z}_n)\right\}$$

  where $\boldsymbol{\eta}$ denotes the natural parameters (global unknowns) and $\boldsymbol{u}$ is the sufficient statistics func

- Suppose the conjugate prior for the parameters $\boldsymbol{\eta}$ is

$$p(\boldsymbol{\eta}|\nu_0, \mathbf{v}_0) = f(\nu_0, \boldsymbol{\chi}_0) g(\boldsymbol{\eta})^{\nu_0} \exp\left\{\nu_o \boldsymbol{\eta}^{\mathrm{T}} \boldsymbol{\chi}_0\right\}$$

- Note: This prior is equivalent to having ν_0 pseudo-observations, with sufficient statistics $\boldsymbol{u} = \chi_0$
- We are interested in the posterior over both $\mathbf{Z}$ and $\boldsymbol{\eta}$. Usually intractable. Let's use Mean-Field VB.

Mean-Field VB for Exponential Family

- The variational posterior for $\mathbf{Z}$ can be written (only keeping terms that depend on $\mathbf{Z}$)

$$
\begin{aligned}
\ln q^{\star}(\mathbf{Z}) &= \mathbb{E}_{\boldsymbol{\eta}}[\ln p(\mathbf{X}, \mathbf{Z}|\boldsymbol{\eta})] + \text{const} \\
&= \sum_{n=1}^{N} \left\{ \ln h(\mathbf{x}_n, \mathbf{z}_n) + \mathbb{E}[\boldsymbol{\eta}^{\mathrm{T}}]\mathbf{u}(\mathbf{x}_n, \mathbf{z}_n) \right\} + \text{const}
\end{aligned}
$$

- Exponentiating again, we get the following form for each $q^*(\boldsymbol{z}_n)$

$$
q^{\star}(\mathbf{z}_n) = h(\mathbf{x}_n, \mathbf{z}_n) g\left(\mathbb{E}[\boldsymbol{\eta}]\right) \exp\left\{ \mathbb{E}[\boldsymbol{\eta}^{\mathrm{T}}]\mathbf{u}(\mathbf{x}_n, \mathbf{z}_n) \right\}
$$

- Likewise, the variational posterior for the parameters $\boldsymbol{\eta}$ (only keeping terms that depend on $\boldsymbol{\eta}$)

$$
\begin{aligned}
\ln q^{\star}(\boldsymbol{\eta}) &= \ln p(\boldsymbol{\eta}|\nu_0, \boldsymbol{\chi}_0) + \mathbb{E}_{\mathbf{Z}}[\ln p(\mathbf{X}, \mathbf{Z}|\boldsymbol{\eta})] + \text{const} \\
&= \nu_0 \ln g(\boldsymbol{\eta}) + \boldsymbol{\eta}^{\mathrm{T}} \boldsymbol{\chi}_0 + \sum_{n=1}^{N} \left\{ \ln g(\boldsymbol{\eta}) + \boldsymbol{\eta}^{\mathrm{T}} \mathbb{E}_{\mathbf{z}_n}[\mathbf{u}(\mathbf{x}_n, \mathbf{z}_n)] \right\} + \text{const}
\end{aligned}
$$

- Again, exponentiating gives the following variational distribution

$$
\begin{aligned}
q^{\star}(\boldsymbol{\eta}) &= f(\nu_N, \boldsymbol{\chi}_N) g(\boldsymbol{\eta})^{\nu_N} \exp\left\{ \boldsymbol{\eta}^{\mathrm{T}} \boldsymbol{\chi}_N \right\} \\
\nu_N &= \nu_0 + N \\
\boldsymbol{\chi}_N &= \boldsymbol{\chi}_0 + \sum_{n=1}^{N} \mathbb{E}_{\mathbf{z}_n}[\mathbf{u}(\mathbf{x}_n, \mathbf{z}_n)]
\end{aligned}
$$

- Again note that updates of $q(\mathbf{Z})$ and $q(\boldsymbol{\eta})$ are coupled

VB and Expectation Maximization (EM)

- VB can be seen as a generalization of the EM algorithm
- Unlike EM, in VB there is no distinction between parameters Θ and latent variables $\mathbf{Z}$
- VB treats all unknowns of the model as latent variables and calls them $\mathbf{Z}$
- Since there is no notion of "parameters", VB is like EM without the "M step"
- VB can be used within an EM algorithm if the E step is intractable
    - This is known as Variational EM algorithm

Some Properties of VB

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

$$\text{KL}(q||p) = -\int q(\mathbf{Z}) \ln \left\{ \frac{p(\mathbf{Z})}{q(\mathbf{Z})} \right\} \mathrm{d}\mathbf{Z}$$

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 VB

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

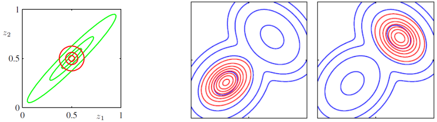


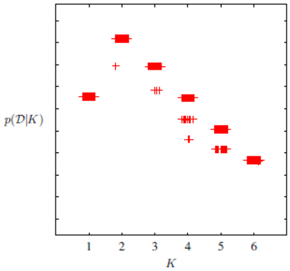
Figure: (Left) Zero-Forcing Property of VB, (Right) For multi-modal posterior, VB locks onto one of the models

Note: Some other inference methods, e.g., Expectation Propagation (EP) can avoid this behavior

ELBO for Model Selection

- ELBO can also be used for model selection
- We can compute ELBO for each model and then choose the one with largest value of ELBO
- An Example: The ELBO plot for a Gaussian Mixture Model with different K values

Plot of the variational lower bound $\mathcal{L}$ versus the number K of components in the Gaussian mixture model, for the Old Faithful data, showing a distinct peak at $K = 2$ components. For each value of K, the model is trained from 100 different random starts, and the results shown as '+' symbols plotted with small random horizontal perturbations so that they can be distinguished. Note that some solutions find suboptimal local maxima, but that this happens infrequently.



- Note that unlike likelihood, ELBO doesn't monotonically increase with K (penalizes large K)

Figure courtesy: PRML (Bishop, 2006)

Some Comments

- VB is very widely used for approximate inference in probabilistic models
- In general, VB requires two steps: Writing down ELBO $\mathcal{L}(q)$ and optimizing it w.r.t. q
- Mean-field assumption and exponential family distributions make it easy to derive VB
- In other cases, we usually need to approximate the ELBO and its derivatives
  - E.g., Using Monte-Carlo approximations of ELBO and its derivatives (more on this later)

Appendix

Mean-Field VB: An Example

- Suppose we have N obs. $\mathcal{D} = \{x_1, \ldots, x_N\}$ from a 1-D Gaussian $\mathcal{N}(\mu, \tau)$

$$p(\mathcal{D}|\mu,\tau) = \left(\frac{\tau}{2\pi}\right)^{N/2} \exp\left\{-\frac{\tau}{2}\sum_{n=1}^{N}(x_n-\mu)^2\right\}$$

- Assume the following priors on the mean μ and precision τ

$$\begin{aligned} p(\mu|\tau) &= \mathcal{N}\left(\mu|\mu_0, (\lambda_0\tau)^{-1}\right) \\ p(\tau) &= \mathrm{Gam}(\tau|a_0, b_0) \end{aligned}$$

- Goal: Infer the posterior distribution $p(\mu, \tau|\mathcal{D})$
- The prior $p(\mu, \tau)$ is jointly conjugate to the likelihood so we can easily get closed form posterior
- Actually no need to do VB for this model but we will nevertheless do it as a simple example of VB
- Our mean-field VB approximation will assume the following factorized distribution

$$q(\mu, \tau) = q_\mu(\mu)q_\tau(\tau)$$

Mean-Field VB: An Simple Example

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

$$q_j^\star(\mathbf{Z}_j) = \frac{\exp\left(\mathbb{E}_{i\neq j}[\ln p(\mathbf{X}, \mathbf{Z})]\right)}{\int \exp\left(\mathbb{E}_{i\neq j}[\ln p(\mathbf{X}, \mathbf{Z})]\right) \mathrm{d}\mathbf{Z}_j}$$

- The optimal variational distribution $q_\mu^*(\mu)$ for the Gaussian's mean (verify):

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

Completing the square over μ we see that $q_\mu(\mu)$ is a Gaussian $\mathcal{N}\left(\mu|\mu_N, \lambda_N^{-1}\right)$ with mean and precision given by

$$\begin{aligned}
\mu_N &= \frac{\lambda_0\mu_0 + N\overline{x}}{\lambda_0 + N} \\
\lambda_N &= (\lambda_0 + N)\mathbb{E}[\tau].
\end{aligned}$$

Mean-Field VB: A Simple Example

- Assuming shape-rate parameterization of gamma prior on the precision τ, the optimal variational distribution $q_\tau^*(\tau)$ for the Gaussian's precision (verify):

$$
\begin{aligned}
\ln q_\tau^\star(\tau) &= \mathbb{E}_\mu\left[\ln p(\mathcal{D}|\mu,\tau) + \ln p(\mu|\tau)\right] + \ln p(\tau) + \text{const} \\
&= (a_0 - 1)\ln\tau - b_0\tau + \frac{N}{2}\ln\tau + \frac{1}{2}\ln\tau \\
&\quad -\frac{\tau}{2}\mathbb{E}_\mu\left[\sum_{n=1}^{N}(x_n-\mu)^2 + \lambda_0(\mu-\mu_0)^2\right] + \text{const}
\end{aligned}
$$

and hence $q_\tau(\tau)$ is a gamma distribution $\mathrm{Gam}(\tau|a_N, b_N)$ with parameters

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

Mean-Field VB: Another Example (using ELBO Derivatives)

- Consider a Bayesian linear regression model with unknown noise variance α^{-1} (assume λ known)

$$y_i \sim \mathrm{Normal}(x_i^T w, \alpha^{-1}), \quad w \sim \mathrm{Normal}(0, \lambda^{-1} I), \quad \alpha \sim \mathrm{Gamma}(a, b)$$

$$p(y, w, \alpha | x) = p(\alpha) p(w) \prod_{i=1}^{N} p(y_i | x_i, w, \alpha)$$

$$q(w, \alpha) = q(\alpha) q(w) = \mathrm{Gamma}(\alpha | a', b') \mathrm{Normal}(w | \mu', \Sigma')$$

- The ELBO $\mathcal{L}(q) = \int q(\mathbf{Z}) \log p(\mathbf{X}, \mathbf{Z}) d\mathbf{Z} - \int q(\mathbf{Z}) \log q(\mathbf{Z}) d\mathbf{Z}$ is

$$\begin{aligned}
\mathcal{L}(a', b', \mu', \Sigma') &= \int q(\alpha) \ln p(\alpha) d\alpha + \int q(w) \ln p(w) dw \\
&+ \sum_{i=1}^{N} \int \int q(\alpha) q(w) \ln p(y_i | x_i, w, \alpha) dw d\alpha \\
&- \int q(\alpha) \ln q(\alpha) d\alpha - \int q(w) \ln q(w) dw
\end{aligned}$$

Mean-Field VB: Another Example (using ELBO Derivatives)

- ELBO is now a function of the variational parameters a', b', μ', Σ'

$$
\begin{aligned}
\mathcal{L}(a', b', \mu', \Sigma') = \; & (a-1)(\psi(a') - \ln b') - b\frac{a'}{b'} + \text{constant} \\
& -\frac{\lambda}{2}(\mu'^T\mu' + \text{tr}(\Sigma')) + \text{constant} \\
& +\frac{N}{2}(\psi(a') - \ln b') - \sum_{i=1}^{N} \frac{1}{2}\frac{a'}{b'}\left((y_i - x_i^T\mu')^2 + x_i^T\Sigma' x_i\right) + \text{constant} \\
& +a' - \ln b' + \ln\Gamma(a') + (1-a')\psi(a') \\
& +\frac{1}{2}\ln|\Sigma'| + \text{constant}
\end{aligned}
$$

- ψ is the digamma function (derivative of log of gamma function)
- Can now take gradients w.r.t. each of a', b', μ', Σ' and estimate these in an alternating fashion
- This will give us $q(\boldsymbol{w}, \alpha) = \text{Normal}(\boldsymbol{w}|\mu', \Sigma')\text{Gamma}(\alpha|a', b')$