VI Talk

roko

June 2023

1 Introduction

Suppose we believe something follows a latent variable model (LVM).

Idea: Bayesian Inference to the rescue!

Start with:

$$\text{Prior: } p_\theta(z)$$

$$\text{Likelihood: } p_\theta(x|z)$$

End with:

$$\text{Posterior: } \frac{p_\theta(x|z) * p_\theta(z)}{\int p_\theta(x|z) * p_\theta(z) dx}$$

Oh... this is bad (intractable).

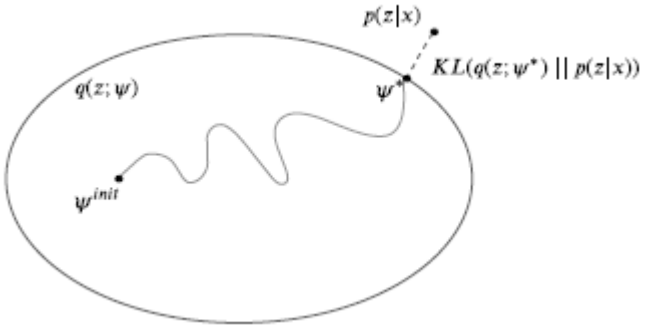

Figure 1: Caption

Approach: Use the best q distribution out of the distribution family $\mathcal{Q}$ such that:

$$q = \underset{q \in \mathcal{Q}}{\operatorname{argmin}} D_{\mathbb{KL}}\left(q(z) \| p_{\boldsymbol{\theta}}(z \mid x)\right)$$

Solution:

$$\begin{aligned}
\psi^* &= \underset{\psi}{\operatorname{argmin}} D_{\mathbb{KL}}\left(q_\psi(\boldsymbol{z}) \| p_{\boldsymbol{\theta}}(\boldsymbol{z} \mid \boldsymbol{x})\right) \\
&= \underset{\psi}{\operatorname{argmin}} \mathbb{E}_{q_\psi(\boldsymbol{z})}\left[\log q_{\boldsymbol{\psi}}(\boldsymbol{z}) - \log\left(\frac{p_{\boldsymbol{\theta}}(\boldsymbol{x} \mid \boldsymbol{z}) p_{\boldsymbol{\theta}}(\boldsymbol{z})}{p_{\boldsymbol{\theta}}(\boldsymbol{x})}\right)\right] \\
&= \underset{\psi}{\operatorname{argmin}} \underbrace{\mathbb{E}_{q_\psi(\boldsymbol{z})}\left[\log q_{\boldsymbol{\psi}}(\boldsymbol{z}) - \log p_{\boldsymbol{\theta}}(\boldsymbol{x} \mid \boldsymbol{z}) - \log p_{\boldsymbol{\theta}}(\boldsymbol{z})\right]}_{\mathcal{L}(\boldsymbol{\theta}, \boldsymbol{\psi} \mid \boldsymbol{x})} + \log p_{\boldsymbol{\theta}}(\boldsymbol{x}) \\
&= \underset{\psi}{\operatorname{argmin}} \mathbb{E}_{q_\psi(\boldsymbol{z})}\left[\log q_{\boldsymbol{\psi}}(\boldsymbol{z}) - \log p_{\boldsymbol{\theta}}(\boldsymbol{x}, \boldsymbol{z})\right] \\
&= \underset{\psi}{\operatorname{argmin}} \mathbb{E}_{q_\psi(\boldsymbol{z})}\left[-\log p_{\boldsymbol{\theta}}(\boldsymbol{x}, \boldsymbol{z}) + \log q_{\boldsymbol{\psi}}(\boldsymbol{z})\right]
\end{aligned}$$

Notice that the last term in line 3 is constant in terms of ψ so we can drop it. Note that because a KL divergence is always above 0, we have that (from line 3)

$$D_{\mathbb{KL}}\left(q_\psi(\boldsymbol{z}) \| p_{\boldsymbol{\theta}}(\boldsymbol{z} \mid \boldsymbol{x})\right) = \mathcal{L}(\boldsymbol{\theta}, \boldsymbol{\psi} \mid \boldsymbol{x}) + \log p_{\boldsymbol{\theta}}(\boldsymbol{x}) \geq 0$$

$$\log p_{\boldsymbol{\theta}}(\boldsymbol{x}) \geq -\mathcal{L}(\boldsymbol{\theta}, \boldsymbol{\psi} \mid \boldsymbol{x}) := ELBO$$

$$\textbf{ELBO} = E_{q_\phi(z)}\left[\log p_{\boldsymbol{\theta}}(\boldsymbol{x}, \boldsymbol{z}) - \log q_{\boldsymbol{\psi}}(\boldsymbol{z})\right]$$

The equation in blue is VERY important. It's important that we keep this equation in mind as we perform VI going forward. Keep in mind that the joint likelihood can ALWAYS be decomposed into the product of the likelihood and prior. It's also nice that you are maximizing a lower bound of the evidence. So higher ELBO implies a higher MINIMUM evidence.

$$\begin{aligned}
\mathrm{L}(\boldsymbol{\psi} \mid \boldsymbol{\theta}, \boldsymbol{x}) &= \mathbb{E}_{q_\psi(\boldsymbol{z})}\left[\log p_{\boldsymbol{\theta}}(\boldsymbol{x} \mid \boldsymbol{z}) + \log p_{\boldsymbol{\theta}}(\boldsymbol{z}) - \log q_{\boldsymbol{\psi}}(\boldsymbol{z})\right] \\
&= \mathbb{E}_{q_{\boldsymbol{\psi}}(\boldsymbol{z})}\left[\log p_{\boldsymbol{\theta}}(\boldsymbol{x} \mid \boldsymbol{z})\right] + \int q_\psi(\boldsymbol{z})(\log p_{\boldsymbol{\theta}}(\boldsymbol{z}) - \log q_{\boldsymbol{\psi}}(\boldsymbol{z}))dz \\
&= \mathbb{E}_{q_{\boldsymbol{\psi}}(\boldsymbol{z})}\left[\log p_{\boldsymbol{\theta}}(\boldsymbol{x} \mid \boldsymbol{z})\right] + \int q_\psi(\boldsymbol{z}) \frac{\log p_{\boldsymbol{\theta}}(\boldsymbol{z})}{\log q_{\boldsymbol{\psi}}(\boldsymbol{z})} dz \\
&= \mathbb{E}_{q_{\boldsymbol{\psi}}(\boldsymbol{z})}\left[\log p_{\boldsymbol{\theta}}(\boldsymbol{x} \mid \boldsymbol{z})\right] - \int q_\psi(\boldsymbol{z}) \frac{\log q_{\boldsymbol{\psi}}(\boldsymbol{z})}{\log p_{\boldsymbol{\theta}}(\boldsymbol{z})} dz \\
&= \mathbb{E}_{q_{\boldsymbol{\psi}}(\boldsymbol{z})}\left[\log p_{\boldsymbol{\theta}}(\boldsymbol{x} \mid \boldsymbol{z})\right] - D_{\mathbb{KL}}\left(q_{\boldsymbol{\psi}}(\boldsymbol{z}) \| p_{\boldsymbol{\theta}}(\boldsymbol{z})\right)
\end{aligned}$$

ELBO = expected log likelihood - KL from posterior to prior

2 Variational Posterior Forms

2.1 Fixed Form VI

Fixed-Form VI means that you pick a convenient functional form (i.e. MV Gaussian) for the posterior $q_\psi(z|x)$ and the ELBO is optimized with gradient based methods.

2.2 Free Form VI

Free-Form VI is VI that is optimized one parameter at a time in a coordinate ascent manner. This is typically done with the mean field assumption by assuming the posterior factorizes:

$$q_\phi(z|x) = \prod_{j=1}^{J} q_j(z_j)$$

3 Parameter Estimation

3.1 MLE for LVMs

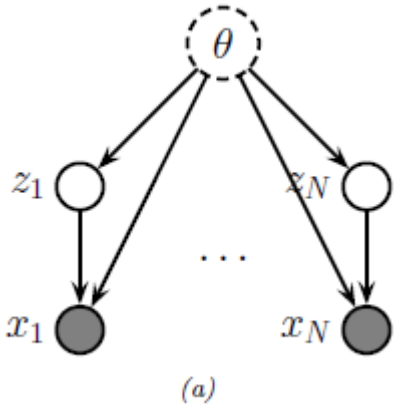

Figure 2: Caption

Suppose we have a latent variable model of the form.

$$p\left(\mathcal{D}, \boldsymbol{z}_{1:N} \mid \boldsymbol{\theta}\right) = \prod_{n=1}^{N} p\left(\boldsymbol{z}_n \mid \boldsymbol{\theta}\right) p\left(\boldsymbol{x}_n \mid \boldsymbol{z}_n, \boldsymbol{\theta}\right)$$

How do you find the MLE of θ given the observable data? With unobservable latent variables, we can't take them as given in the likelihood function and need to marginalize them out.

$$\log p\left(\boldsymbol{x}_n \mid \boldsymbol{\theta}\right) = \log \left[\int p\left(\boldsymbol{x}_n \mid z_n, \boldsymbol{\theta}\right) p\left(\boldsymbol{z}_n \mid \boldsymbol{\theta}\right) dz_n \right]$$

This integral is intractable, but the ELBO is not!

$$\mathrm{L}\left(\boldsymbol{\theta}, \boldsymbol{\psi}_{1:N} \mid \mathcal{D}\right) = \sum_{n=1}^{N} \mathrm{L}\left(\boldsymbol{\theta}, \boldsymbol{\psi}_n \mid \boldsymbol{x}_n\right) \leq \log p(\mathcal{D} \mid \boldsymbol{\theta})$$

We can optimize this using VEM:

Algorithm 10.1: Amortized stochastic variational EM

Initialize $\boldsymbol{\theta}$, $\boldsymbol{\phi}$
repeat
  Sample $\boldsymbol{x}_n \sim p_{\mathcal{D}}$
  E step: $\boldsymbol{\phi} = \operatorname{argmax}_{\boldsymbol{\phi}} Ł(\boldsymbol{\theta}, \boldsymbol{\phi} | \boldsymbol{x}_n)$
  M step: $\boldsymbol{\theta} = \operatorname{argmax}_{\boldsymbol{\theta}} Ł(\boldsymbol{\theta}, \boldsymbol{\phi} | \boldsymbol{x}_n)$
until *converged*

Figure 3: Caption

Note that unlike the traditional EM algorithm, where the E-step involves taking an expectation for missing/unknown values, VEM's E-step is not an expectation. An EM algorithm might be a borderline misuse of the term. The important thing is that we update variational parameters ψ, and then use them for sampling purposes when learning model parameters θ.

3.2 Empirical Bayes

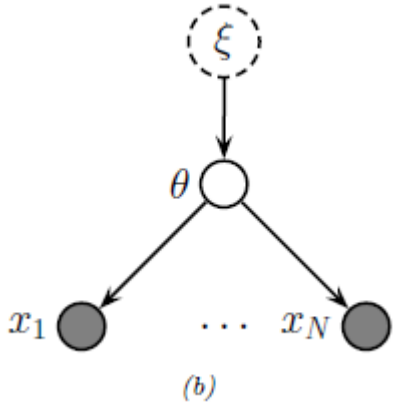

Figure 4: Caption

If you have a model with no local latents, we can try for an (empirical) Bayes approach. In this case, we assume the model has the following graphical decomposition

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

Obviously the goal is to compute the posterior:

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

Option 1: Use standard Bayesian inference. This assumes we "know" (more like have a reasonable guess at) the values of $\boldsymbol{\xi}$.
Option 2: Use empirical Bayes if hyperprior parameters are not known. Basically, maximize the hyperprior parameters s.t. the maximize the likelihood.

$$\hat{\xi} = \underset{\xi}{\operatorname{argmax}} \log p(\mathcal{D} \mid \xi)$$

We can use variational EM to compute this. The parameter to be estimated are ξ, the latent variables are the shared global parameters θ, and the observations are $\mathcal{D}$. We then get the lower bound

$$\log p(\mathcal{D} \mid \boldsymbol{\xi}) \geq L(\boldsymbol{\xi}, \boldsymbol{\psi} \mid \mathcal{D}) = \mathbb{E}_{q_\psi(\boldsymbol{\theta})} \left[\sum_{n=1}^{N} \log p\left(\boldsymbol{x}_n \mid \boldsymbol{\theta}\right) \right] - D_{\mathrm{KL}}\left(q_\psi(\boldsymbol{\theta}) \| p(\boldsymbol{\theta} \mid \boldsymbol{\xi})\right)$$

If ξ is fixed, then we only need to optimize the variational parameters and this puts in the case of Option 1, called Variational Bayes.

4 VI in Practice

4.1 Stochastic VI

Now that we have seen the ways with which we can leverage variational inference to find an approximating distribution in theory, how does it work in practice?

After all, we have assumed thus far that we have access to all N samples and that it's computationally feasible to use all of them (spoiler alert, it's not). Luckily,

$$\mathrm{L}(\boldsymbol{\psi}, \boldsymbol{\theta} \mid \boldsymbol{x}) = \mathrm{L}(\boldsymbol{\psi}_{1:N}, \boldsymbol{\theta} \mid \boldsymbol{x_n}) = \sum_{n=1}^{N} \mathrm{L}(\boldsymbol{\psi}_n, \boldsymbol{\theta} \mid \boldsymbol{x}) = \frac{N}{B} \sum_{x_n \in B} L(\boldsymbol{\psi}_{x_n}, \boldsymbol{\theta} | \boldsymbol{x_n})$$

Similar to how we do not need to process all of the data to make an update in gradient descent, we don't need all of the samples to make an update in maximizing the ELBO.

4.2 Amortized VI

First of all, what does it mean for something to be amortized?

amortized: optimized in a shared way (usually via parameter sharing)

This can be useful because we can optimize several parameters at once. By allowing our ψ_n values to be predicted all at once as outputs of a neural network (the inference network in our case), we can avoid having to optimize them separately.

$$q(\boldsymbol{z_n}|\psi_n) = q(\boldsymbol{z_n}|f_\phi^{\text{inf}}(\boldsymbol{x_n})) = q_\phi(\boldsymbol{z_n}|\boldsymbol{x_n})$$

So, if we created the inference net for amortized inference and combine it with Stochastic VI, we get an ELBO formula of...

$$\mathrm{L}(\boldsymbol{\psi}, \boldsymbol{\theta} \mid \mathcal{D}) = \frac{N}{B} \sum_{x_n \in B} \mathbb{E}_{q_\phi(\boldsymbol{z_n}|\boldsymbol{x_n})} \left[\log p_{\boldsymbol{\theta}}(\boldsymbol{x_n}, \boldsymbol{z_n}) - \log q_{\boldsymbol{\phi}}(\boldsymbol{z_n}|\boldsymbol{x_n})\right]$$

4.2.1 Semi-amortized VI

Amortized inference is no free lunch though (what is though really)...

There exists a tradeoff between joint optimization and optimizing each variational parameter in isolation: accuracy vs. computational expense. The gap

in accuracy is called the amortization gap. A possible solution is to use the inference net as a warm start for the initial guess of each variational parameter before individually learning optimal values.

5 VI and the Reparameterization Trick

How exactly do we optimize over distributions exactly? In this section we tackle the ability to take gradients over *seemingly* random variables.

The key is to rewrite the random variable $\mathbf{z} \sim q_\phi(\boldsymbol{z}|\boldsymbol{x})$ as a differentiable AND invertible transformation g in terms of another random variable $\epsilon \sim p(\epsilon)$ which does not depend on ϕ.
If we can create such a reparameterization, we can create move the gradient under the expectation and create a tractable gradient update.

$$\nabla_\phi \mathbb{E}_{q_\phi(z|x)}[f(z)] = \nabla_\phi \mathbb{E}_{p(\epsilon)}[f(z)] = \mathbb{E}_{p(\epsilon)}\left[\nabla_\phi f(z)\right]$$

5.1 Example: Isotropic Gaussian

$$\epsilon \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$$
$$z = \boldsymbol{\mu} + \boldsymbol{\sigma} \odot \boldsymbol{\epsilon}$$
$$(\boldsymbol{\mu}, \log \boldsymbol{\sigma}) = f_\phi^{\text{inf}}(\mathbf{x})$$

Recall that right now, the ELBO formula for a single sample is:

$$\log p_{\boldsymbol{\theta}}(\boldsymbol{x_n}, \boldsymbol{z_n}) - \log q_{\boldsymbol{\phi}}(\boldsymbol{z_n}|\boldsymbol{x_n})$$

The likelihood is well defined in our model. In order to evaluate the term $q_\phi(\mathbf{z_n}|\mathbf{x_n})$, we need to leverage the change of variables formula:

$$q_\phi(z|x) = p_\epsilon(g^{-1}(z)) * \left|\det\left(\frac{\partial}{\partial z} g^{-1}(z)\right)\right| = p_\epsilon((z-\mu)/\sigma) * \left|\det\left(\frac{\partial}{\partial z}(z-\mu)./\sigma\right)\right| = p_\epsilon(\epsilon)\left|\det\left(\frac{\partial}{\partial z}\epsilon\right)\right|$$
$$= p_\epsilon(\epsilon)\left|\det\left(\frac{\partial \epsilon}{\partial z}\right)\right| = p_\epsilon(\epsilon)\left|\det\left(\frac{\partial \epsilon}{\partial z}\right)\right| = p_\epsilon(\epsilon)\frac{1}{\left|\det\left(\frac{\partial z}{\partial \epsilon}\right)\right|}$$

Taking the logarithm of both sides gives up

$$\log q_\phi(z|x) = \log p_\epsilon(\epsilon) - \log\left(\left|\det\left(\frac{\partial z}{\partial \epsilon}\right)\right|\right)$$

where the partial derviative is the Jacobian.

5.2 Example: Full Rank Gaussian

$$\epsilon \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$$
$$\Sigma = LL^T \text{(Cholesky decomposition)}$$
$$z = \boldsymbol{\mu} + \mathbf{L}\boldsymbol{\epsilon}$$
$$(\boldsymbol{\mu}, \log \boldsymbol{\sigma}, \mathbf{L}') = f_\phi^{\text{inf}}(\mathbf{x})$$

5.3 Example: Low Rank Gaussian

$$\epsilon_1, \epsilon_2 \overset{i.i.d.}{\sim} \mathcal{N}(\mathbf{0}, \mathbf{I})$$
$$z = \boldsymbol{\mu} + \mathbf{B}\boldsymbol{\epsilon_1} + \mathbf{C}\boldsymbol{\epsilon_2}$$
$$B \in \mathbb{R}^{dxf}$$
$$C \in \mathbb{R}^{dxd} \text{ (Diagonal)}$$
$$\Sigma = BB^T + C^2$$
$$(\mathbf{B}, \text{diag}(\mathbf{C}), \boldsymbol{\mu}) = f_\phi^{\text{inf}}(\mathbf{x})$$

It may not be immediately clear why this paradigm makes sense until you break it down to properties of covariance.

$$Cov(\mathbf{z}) = Cov(\boldsymbol{\mu} + \mathbf{B}\boldsymbol{\epsilon_1} + \mathbf{C}\boldsymbol{\epsilon_2}) = Cov(\mathbf{B}\boldsymbol{\epsilon_1} + \mathbf{C}\boldsymbol{\epsilon_2}) = \mathbf{B}Cov(\boldsymbol{\epsilon_1})\mathbf{B}^T + \mathbf{C}Cov(\boldsymbol{\epsilon_2})\mathbf{C}^T$$
$$= \mathbf{B}I\mathbf{B}^T + \mathbf{C}I\mathbf{C}^T = \mathbf{B}\mathbf{B}^T + \mathbf{C}\mathbf{C}^T = \mathbf{B}\mathbf{B}^T + \mathbf{C}^2$$

This type of setup is useful if a full Gaussian rank requires learning TOO many parameters. You go from learning d (diagonals) + d(d+1)/2 (off diagonals) in the full case to fd (B) + d (diag C) + d (μ) = (f+2)d in the low rank case.

6 ADVI

Notice that VI seems like a tool that has free range and for the most part does not have many restrictions (barring valid transformations in the reparameterization trick).

But variances need to be positive, and probabilities need to be in $[0, 1]$. So, in order to use gradient based methods to learn certain parameters of our variational distributions, we need to be careful not to output parameter values that would violate constraints. ENTER ADVI!

Approach: Map the constrained support space to the real line (in the dimension of the parameter ofc) and use THAT space for gradient algorithms.

Let $T : \Theta \to \mathbb{R}^D$ be a bijective mapping that maps from the constrained to the unconstrained space. Because this transformation is invertible, we can always get back our variational parameter value (u): $u = T^{-1}(\theta)$

$$p(\boldsymbol{u}) = p\left(T^{-1}(\boldsymbol{u})\right) \left|\det\left(\mathbf{J}_{T^{-1}}(\boldsymbol{u})\right)\right|$$

where $\mathbf{J}_{T^{-1}}$ is the Jacobian of the inverse mapping $\boldsymbol{u} \to \boldsymbol{\theta}$. Hence the ELBO becomes

$$\mathrm{L}(\boldsymbol{\psi}) = E_{\boldsymbol{u}\sim q_\psi(\boldsymbol{u})}\left[\log p\left(\mathcal{D} \mid T^{-1}(\boldsymbol{u})\right) + \log p\left(T^{-1}(\boldsymbol{u})\right) + \log\left|\det\left(\mathbf{J}_{T^{-1}}(\boldsymbol{u})\right)\right|\right] + \mathbb{H}(\boldsymbol{\psi})$$

In the equation above, the first term is $\log p(x|z)$, the second and thrid terms make up $\log p(z)$ and the last term is the entropy: $E_{\boldsymbol{u}\sim q_\psi(\boldsymbol{u})}(-\log q_\phi(u))$. Example transformations include the natural logarithm to map from positives to reals for variances.

Example transformations include the inverse hyperbolic tangent to map from probabilities $[0, 1]$ to reals (may need to use epsilon to prevent overflow errors). $\tanh^{-1}(2x - 1)$ to be exact.

7 CAVI

So far, we discussed ways to make VI work with gradient descent. However, this need not be the only way we can optimize variational parameters. Recall the mean field approximation and free-form VI.

$$q_\phi(z|x) = \prod_{j=1}^{J} q_{\psi_j}(z_j) \overset{def}{=} \prod_{j=1}^{J} q_j(z_j)$$

The overall idea is summarized in the algorithm below.

Algorithm 10.4: Coordinate ascent variational inference (CAVI).

Initialize $q_j(\boldsymbol{z}_j)$ for $j = 1 : J$
foreach $t = 1 : T$ **do**
  foreach $j = 1 : J$ **do**
    Compute $g_j(\boldsymbol{z}_j) = \mathbb{E}_{\boldsymbol{z}_{\mathrm{mb}_t}}\left[\log \tilde{p}(\boldsymbol{z}_i, \boldsymbol{z}_{\mathrm{mb}_t})\right]$
    Compute $q_j(\boldsymbol{z}_j) \propto \exp(g_j(\boldsymbol{z}_j))$

Figure 5: Caption

7.1 Derivation

10.3.1 Derivation of CAVI algorithm

In this section, we derive the coordinate ascent variational inference (CAVI) procedure.

To derive the update equations, we initially assume there are just 3 discrete latent variables, to simplify notation. In this case the ELBO is given by

$$Ł(q_1, q_2, q_3) = \sum_{z_1}\sum_{z_2}\sum_{z_3} q_1(z_1)q_2(z_2)q_3(z_3)\log\tilde{p}(z_1, z_2, z_3) + \sum_{j=1}^{3}\mathbb{H}(q_j) \tag{10.77}$$

where we define $\tilde{p}(\boldsymbol{z}) = p_{\boldsymbol{\theta}}(\boldsymbol{z}, \boldsymbol{x})$ for brevity. We will optimize this wrt each q_i, one at a time, keeping the others fixed.

Let us look at the objective for q_3:

$$Ł_3(q_3) = \sum_{z_3} q_3(z_3)\left[\sum_{z_1}\sum_{z_2} q_1(z_1)q_2(z_2)\log\tilde{p}(z_1, z_2, z_3)\right] + \mathbb{H}(q_3) + \text{const} \tag{10.78}$$

$$= \sum_{z_3} q_3(z_3)\left[g_3(z_3) - \log q_3(z_3)\right] + \text{const} \tag{10.79}$$

where

$$g_3(z_3) \triangleq \sum_{z_1}\sum_{z_2} q_1(z_1)q_2(z_2)\log\tilde{p}(z_1, z_2, z_3) = \mathbb{E}_{\boldsymbol{z}_{-3}}\left[\log\tilde{p}(z_1, z_2, z_3)\right] \tag{10.80}$$

where $\boldsymbol{z}_{-3} = (z_1, z_2)$ is all variables except z_3. Here $g_3(z_3)$ can be interpreted as an expected negative energy (log probability). We can convert this into an unnormalized probability distribution by defining

$$\tilde{f}_3(z_3) = \exp(g_3(z_3)) \tag{10.81}$$

which we can normalize to get

$$f_3(z_3) = \frac{\tilde{f}_3(z_3)}{\sum_{z_3'}\tilde{f}_3(z_3')} \propto \exp(g_3(z_3)) \tag{10.82}$$

Since $g_3(z_3) \propto \log f_3(z_3)$ we get

$$Ł_3(q_3) = \sum_{z_3} q_3(z_3)\left[\log f_3(z_3) - \log q_3(z_3)\right] + \text{const} = -D_{\mathrm{KL}}\left(q_3 \parallel f_3\right) + \text{const} \tag{10.83}$$

Since $D_{\mathrm{KL}}\left(q_3 \parallel f_3\right)$ achieves its minimal value of 0 when $q_3(z_3) = f_3(z_3)$ for all z_3, we see that $q_3^*(z_3) = f_3(z_3)$.

Now suppose that the joint distribution is defined by a Markov chain, where $z_1 \to z_2 \to z_3$, so $z_1 \perp z_3|z_2$. Hence $\log\tilde{p}(z_1, z_2, z_3) = \log\tilde{p}(z_2, z_3|z_1) + \log\tilde{p}(z_1)$, where the latter term is independent of $q_3(z_3)$. Thus the ELBO simplifies to

$$Ł_3(q_3) = \sum_{z_3} q_3(z_3)\left[\sum_{z_2} q_2(z_2)\log\tilde{p}(z_2, z_3)\right] + \mathbb{H}(q_3) + \text{const} \tag{10.84}$$

$$= \sum_{z_3} q_3(z_3)\left[\log f_3(z_3) - \log q_3(z_3)\right] + \text{const} \tag{10.85}$$

where

$$f_3(z_3) \propto \exp\left[\sum_{z_2} q_2(z_2)\log\tilde{p}(z_2, z_3)\right] = \exp\left[\mathbb{E}_{\boldsymbol{z}_{\mathrm{mb}_3}}\left[\log\tilde{p}(z_2, z_3)\right]\right] \tag{10.86}$$

8 Variational Bayes

Recall the model with shared latent variable θ.

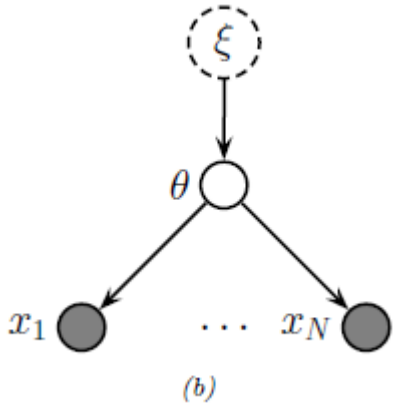

Figure 7: Caption

We discussed the possibility that ξ is fixed. In that case, it cannot be updated and it is independent of θ and can be removed from the graph.

Our new goal is to learn the posterior:

$$p(\boldsymbol{\theta} \mid \mathcal{D}) = \frac{p(\mathcal{D} \mid \boldsymbol{\theta}) p(\boldsymbol{\theta})}{p(\mathcal{D})} \propto p(\boldsymbol{\theta}) \prod_{n=1}^{N} p(\mathcal{D}_n \mid \boldsymbol{\theta})$$

The ELBO takes on the familiar form:

$$L(\phi_\theta|\mathcal{D}) = \mathbb{E}_{q(\theta|\phi_\theta)} \left[p(\boldsymbol{\theta}) \prod_{n=1}^{N} p(\mathcal{D}_n \mid \boldsymbol{\theta}) - \sum_{j=1}^{J} log(q(\theta_j|\phi_{\theta_j})) \right]$$

Now that the ELBO is expressed after taking the mean field approximation into account, we can leverage CAVI to learn the variational parameters!