E6892 Bayesian Models for Machine Learning

Columbia University, Fall 2015

Lecture 5, 10/8/2015

Instructor: John Paisley

- Last week we talked about using the EM algorithm for MAP inference. Today, we will apply EM to a sequence of models to see where EM breaks down and variational inference (VI) takes over. The purpose is to discuss EM in more detail and also see how it closely relates to variational methods. Variational inference provides a general way to approximate the full posterior distribution of all model variables, and is a new technique to add to the arsenal, along with the Laplace approximation and MCMC methods (of which we only discussed Gibbs sampling).

The "Core Model"

- We are more interested in the general EM and VI inference techniques, but I think they're easier to differentiate through an easy modeling example. Therefore, we will build on the basic linear regression model discussed in an earlier lecture. Our development of this model will not be the main thing of interest. Instead, we will be more interested in what we can do with each version of this model in the context of EM and VI.

- Remember that we're given pairs $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^N$ where $x \in \mathbb{R}^d$ and $y \in \mathbb{R}$ and we model this data as

$$y_i \overset{ind}{\sim} \text{Normal}(x_i^T w, \alpha^{-1}), \qquad w \sim \text{Normal}(0, \lambda^{-1} I) \tag{1}$$

  There is no prior distribution on x, and so we are in the "discriminative" setting. This is in contrast to the Bayes classifier where we had priors on x and so were in the "generative" setting.

- Posterior calculation: Earlier, we saw how we could calculate the posterior distribution of this model in closed form,

$$p(w|x, y) = \frac{p(y|w, x)p(w)}{p(y|x)} = \text{Normal}(\mu, \Sigma) \tag{2}$$

  where

$$\Sigma = \Big(\lambda I + \alpha \sum_{i=1}^N x_i x_i^T\Big)^{-1}, \qquad \mu = \Sigma\Big(\sum_{i=1}^N \alpha y_i x_i\Big).$$

  (This time, we let the index-free x and y represent all x_i and y_i, respectively.)

- Therefore, for this simple model the inference problem is very easy. No iterative algorithms are needed and we can directly go to the full posterior of all model variables (in this case, only w) in one step. We next develop this model in a way where this is not possible.

- We notice that there are two model parameters that need to be set and will likely have a significant impact on $p(w|y, x)$. That is, even though we have a mathematically correct answer for this posterior distribution, it is only correct under the assumptions we make about the values of α and λ. (And as emphasized before, on the assumption of the model to begin with, which is a simplifying assumption about the data-generating processes underlying a real-world phenomenon.)

- Therefore, even though we have solved for $p(w|y, x)$, we can't claim that we've solved the problem of modeling (x, y) with a linear regression model—we probably can never claim that.

- Since α corresponds to the inverse noise variance, it arguably has the greater impact on our predictions, so let's address this one first:

    - The default solution in the wider machine learning community is to use "cross-validation." In short, that means we try a bunch of values for α and see which one is best according to some problem we're interested in.

    - In the Bayesian modeling world, the default solution is to put a prior distribution on α and try to learn it using posterior inference.

Core model (version 2.0)

- We expand the core model hierarchically by putting a prior distribution on α. Out of convenience we pick a conjugate gamma prior,

$$y_i \overset{ind}{\sim} \text{Normal}(x_i^T w, \alpha^{-1}), \qquad w \sim \text{Normal}(0, \lambda^{-1} I), \qquad \alpha \sim \text{Gamma}(a, b) \tag{3}$$

- As was the case with w, by adding a prior on α, we are committing ourselves to one of the following:

    1. Try to learn the full posterior of α and w (exactly or approximately)
    2. Learn a point estimate of the variables via MAP inference
    3. Integrate out (or marginalize) some variables to learn the others (e.g. with point estimate)

- Let's consider each of these:

    1. The full posterior distribution of α and w is

$$p(w, \alpha|y, x) = \frac{p(y|w, \alpha, x)p(w)p(\alpha)}{\int\int p(y|w, \alpha, x)p(w)p(\alpha)dwd\alpha} \tag{4}$$

    When we try to calculate the normalizing constant, we see that one integral can be performed (either over α or w), but this makes the other integral intractable. In this case, the interactions between w and α in the model are such that we can't find the full posterior.

(This is not always the case, e.g., in the first homework.) To approximate this posterior, we could use Laplace or MCMC sampling (Gibbs sampling will work in this case). However, since we're more interested in introducing new inference techniques than in solving specific models now, let's consider other options.

2. We could also simply do MAP over w and α,

$$\begin{aligned} w, \alpha &= \arg\max_{w,\alpha} \ln p(y, w, \alpha|x) \\ &= \arg\max_{w,\alpha} \ln p(y|w, \alpha, x) + \ln p(w) + \ln p(\alpha) \end{aligned} \tag{5}$$

We can take the derivative of this with respect to w or α and solve for each variable in closed form (but not both at once). This would lead to a "coordinate ascent" algorithm, where we iterate between maximizing w.r.t. w or α holding the other one fixed.

However, this in some sense is unsatisfying because we're giving up all measures of uncertainty and learning a point estimate of all variables. In any case, I argue that it should be unsatisfying in light of option #3.

3. Another option is to integrate out α. That is, to calculate the marginal likelihood

$$p(y, w|x) = \int p(y, w, \alpha|x) d\alpha \tag{6}$$

We have a few options after doing this. The first one to try is to do posterior inference for $p(w|y, x)$ using Bayes rule. However, notice that a different model is implied here. When we write $p(w|y, x) \propto p(y|w, x)p(w)$, we see that the term $p(y|w, x)$ isn't a Gaussian anymore since it's found by integrating out α. That is, $p(y|w, x) = \int p(y, \alpha|w, x) d\alpha$, which gives a student-t distribution, not a Gaussian like in the core model.[1] Therefore, the prior and likelihood aren't conjugate anymore.

Another option would be to maximize $p(y, w|x)$ over w using MAP. However, using the student-t marginal $p(y|w, x)$, we see that $\nabla_w \ln p(y|w, x)p(w) = 0$ does not have a closed form solution for w. Therefore, we would have to use gradient methods if we wanted to directly maximize $p(y, w|x)$ over w. However, this setup is reminiscent of EM...

- Let's now focus on option #3. We want to maximize $p(y, w|x) = \int p(y, w, \alpha|x) d\alpha$ over w. Doing so directly working with $\ln p(y, w|x)$ requires gradient methods, which is something we might want to avoid. (This is not a blanket statement and optimization researchers would probably disagree. But if we have an option that gives closed form updates, we would rather use that.)

- Remember from our discussion on the EM algorithm that we want to maximize a joint likelihood, e.g., $p(y, w|x)$, but we want to avoid gradient methods. To this end, we need to find a hidden variable, e.g., α, such that $p(y, w|x) = \int p(y, w, \alpha|x) d\alpha$. Our initial discussion started with defining the model $p(y, w|x)$ and the tricky part was finding the α. In this case, the latent variable is actually part of the original model definition and we are then finding a point estimate of the marginal distribution of that model.

[1] In that model, α was a parameter so we didn't bother to write $p(y|w, x, \alpha)$, but that's technically what $p(y|w, x)$ corresponds to in the core model, whereas now $p(y|w, x)$ corresponds to the above integral over α.

1. The first direction was hard: Start with a desired marginal distribution and introduce a variable that gives that marginal distribution after integrating it out.

2. This new direction is easy: Start with the expanded model as a definition and then try to maximize over the marginal. We care about the distributions in the larger model, not the distribution that comes from marginalizing. "Whatever $p(y, w|x) = \int p(y, w, \alpha|x)d\alpha$ ends up being is what it ends up being." We're just going to maximize it, possibly without even knowing what it is, since EM doesn't require us to ever solve this integral in order to optimize the marginal over w.

EM for the core model (version 2.0)

- Therefore, we're going to find a point estimate of w to maximize the marginal distribution $p(y, w|x) = \int p(y, w, \alpha|x)d\alpha$ of the second version of the core model. We'll do so by treating α as the extra variable in an EM setting. The EM equation in this case is

$$\ln p(y, w|x) = \underbrace{\int q(\alpha) \ln \frac{p(y, w, \alpha|x)}{q(\alpha)} d\alpha}_{\mathcal{L}(w)} + \underbrace{\int q(\alpha) \ln \frac{q(\alpha)}{p(\alpha|y, w, x)} d\alpha}_{\mathrm{KL}(q\|p)} \tag{7}$$

E-Step

Can we find $p(\alpha|y, w, x)$?

$$\begin{aligned} p(\alpha|y, w, x) &\propto \prod_{i=1}^{N} p(y_i|\alpha, w, x_i)p(\alpha) && (8) \\ &\propto \underbrace{\alpha^{\frac{N}{2}} \mathrm{e}^{-\frac{\alpha}{2}\sum_{i=1}^{N}(y_i - x_i^T w)^2}}_{\text{product of normal likelihoods}} \times \underbrace{\alpha^{a-1}\mathrm{e}^{-b\alpha}}_{\text{gamma prior}} \\ &= \mathrm{Gamma}\Big(a + \frac{N}{2}, b + \frac{1}{2}\sum_{i=1}^{N}(y_i - x_i^T w)^2\Big) && (9) \end{aligned}$$

So we can set $q_t(\alpha) = p(\alpha|y, w_{t-1}, x)$ at iteration t. We then calculate the expectation,

$$\begin{aligned} \mathcal{L}_t(w) &= \mathbb{E}_q[\ln p(y, \alpha|w, x)p(w)] - \mathbb{E}_q[\ln q(\alpha)] && (10) \\ &= -\frac{\mathbb{E}_{q_t}[\alpha]}{2}\sum_{i=1}^{N}(y_i - x_i^T w)^2 - \frac{\lambda}{2}w^T w + \text{constant w.r.t. } w \end{aligned}$$

M-Step

We can find q and complete the E-step by calculating an analytic function $\mathcal{L}(w)$. Next, we need to see if we can maximize $\mathcal{L}(w)$ in closed form. If not, then we're likely no better off than we were with $\ln p(y, w|x)$. Differentiating $\nabla_w \mathcal{L}_t(w)$ and setting to zero, we find that

$$w_t = \Big(\lambda I + \mathbb{E}_{q_t}[\alpha]\sum_{i=1}^{N} x_i x_i^T\Big)^{-1}\Big(\sum_{i=1}^{N}\mathbb{E}_{q_t}[\alpha]y_i x_i\Big) \tag{11}$$

The expectation of α is of a gamma random variable with the parameters from the E-step, so

$$\mathbb{E}_{q_t}[\alpha] = \frac{a + \frac{N}{2}}{b + \frac{1}{2}\sum_{i=1}^{N}(y_i - x_i^T w_{t-1})^2} \tag{12}$$

- Notice that if we plug this expression for $\mathbb{E}_{q_t}[\alpha]$ directly into the update of w_t, the EM algorithm is giving us a way of iteratively updating w using the previous value.

- Also notice that, while this is a "pure" EM problem along the lines of what we discussed last time, the interpretation here feels slightly different:

    1. Last week, the latent variable we introduced was just a stepping stone to get the point estimate we wanted. We didn't motivate that new variable or its q distribution as being something interesting itself.

    2. In this model, the latent variable α has a clear interpretation as relating to the observation noise. This was originally an important model parameter that we decided to put a prior distribution on. Therefore, we have a clear picture of what the "introduced" variable α means here, since it was introduced at the point of the model formulation and not at the point of inference. Therefore, we can think of $q(\alpha) = p(\alpha|y, w, x)$ as being interesting in its own right since it's the conditional posterior distribution of an important model variable.

    3. Therefore, using EM for this problem we make a compromise: We can't get the full posterior $p(w, \alpha|y, x)$, but we don't want to make point estimates for these two variables either, so we settle for a point estimate of w and a *conditional* posterior of α.

    4. Also notice that *we could have done the reverse*: We could have learned a point estimate of α and a conditional posterior $q(w)$ distribution. In this case, we would be doing MAP for $p(y, \alpha|x) = \int p(y, w, \alpha|x)dw$. For this specific model, the EM algorithm would work perfectly well since $q(w) = p(w|y, x, \alpha)$ is a multivariate Gaussian, $\mathcal{L}(\alpha)$ is analytic and maximized in closed form. However, w is clearly not just a stepping stone to learn α! In fact, we could interpret EM here as allowing us to calculate the posterior $p(w|y, x)$ of the original "core model" while additionally helping us *set* the parameter α.

Core model (version 3.0)

- If we were doing EM for version 2.0 of the core model, we would probably want to maximize point-wise over α and learn $q(w)$ as discussed in point #4 above. However, this discussion is directed primarily towards connecting EM with VI and so the development in this next version would not work out the way I want it to.

- We've integrated out α, but what about λ? We can again update the model by adding a conveniently chosen prior here,

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

- Again, we can try to learn this model using Laplace, or MCMC sampling (Gibbs sampling works here), or point estimates of w, α and λ, or point estimates of some variables and marginalization over others...

- Let's again look into maximizing w over the the marginal distribution

$$p(y, w|x) = \int \int p(y, w, \alpha, \lambda|x) d\alpha d\lambda \tag{13}$$

EM for the core model (version 3.0)

- We now have two latent variables, but again we can write

$$\ln p(y, w|x) = \int\int q(\alpha, \lambda) \ln \frac{p(y, w, \alpha, \lambda|x)}{q(\alpha, \lambda)} d\alpha d\lambda + \int\int q(\alpha, \lambda) \ln \frac{q(\alpha, \lambda)}{p(\alpha, \lambda|y, w, x)} d\alpha d\lambda \tag{14}$$

- Side comment: Notice that the left hand side again contains $\ln p(y, w|x)$, but the right hand side is different. Does this mean that EM for versions 2.0 and 3.0 are equivalent? The answer is emphatically *no*. This is simply because

$$\int \underbrace{p(y, w, \alpha|x)}_{\text{model \#2}} d\alpha \;\neq\; \int\int \underbrace{p(y, w, \alpha, \lambda|x)}_{\text{model \#3}} d\alpha d\lambda \tag{15}$$

  Even though both of these can be written as $p(y, w|x)$, they are of *different models*. There is always an assumed model underlying a joint likelihood and simply writing $p(y, w|x)$ is not enough information to know what it is.

- Therefore, the EM algorithm we can derive for doing MAP of w in this model is over a different objective function than the previous model: If we were to actually calculate the marginal distributions $p(y, w|x)$ for version 2.0 and 3.0, we would see that they are different functions.

E-Step

According to the rules of EM, we need to set $q(\alpha, \lambda) = p(\alpha, \lambda|y, w, x)$. Using Bayes rule,

$$\begin{aligned} p(\alpha, \lambda|y, w, x) &= \frac{p(y|w, \alpha, \lambda)p(\alpha)p(w|\lambda)p(\lambda)}{\int\int p(y|w, \alpha, \lambda)p(\alpha)p(w|\lambda)p(\lambda)d\alpha d\lambda} \\ &= \underbrace{\frac{p(y, \alpha|w, \lambda)}{\int p(y, \alpha|w, x)d\alpha}}_{= p(\alpha|y,w,x)} \cdot \underbrace{\frac{p(w, \lambda)}{\int p(w, \lambda)d\lambda}}_{= p(\lambda|w)} \end{aligned} \tag{16}$$

The conditional posterior of α is found exactly the same way, and the conditional posterior of λ is found in a similar way,

$$p(\alpha|y, w, x) = \text{Gamma}\Big(a + \frac{N}{2}, b + \frac{1}{2}\sum_{i=1}^{N}(y_i - x_i^T w)^2\Big), \quad p(\lambda|w) = \text{Gamma}\Big(e + \frac{d}{2}, f + \frac{1}{2}w^T w\Big)$$

M-Step

Using the same exact method as for the previous version, we can compute $\mathcal{L}(w)$ and maximize over w to find that

$$w_t = \Big(\mathbb{E}_{q_t}[\lambda]I + \mathbb{E}_{q_t}[\alpha]\sum_{i=1}^{N} x_i x_i^T\Big)^{-1}\Big(\sum_{i=1}^{N}\mathbb{E}_{q_t}[\alpha]y_i x_i\Big) \tag{17}$$

The difference is that we have expectations over both α and λ since these are the variables being marginalized.

- We notice that marginalizing two variables works out because we can still calculate their full conditional posteriors. In this case, the posterior has the form

$$q(\alpha, \lambda) = p(\alpha, \lambda | y, w, x) = p(\alpha | y, w, x) p(\lambda | w) \tag{18}$$

and so we say the posterior "factorizes." Therefore, q factorizes as well

$$q(\alpha, \lambda) = q(\alpha) q(\lambda) \tag{19}$$

where $q(\alpha) = p(\alpha | y, w, x)$ and $q(\lambda) = p(\lambda | w)$. When we calculate expectations over both q distributions, we can focus only on the relevant ones. That is,

$$\mathbb{E}_q[\alpha] = \int \int \alpha q(\alpha) q(\lambda) d\alpha \lambda = \int \alpha q(\alpha) d\alpha \tag{20}$$

Technically, during the E-step we are performing the first integral. However, because α and λ are *conditionally independent* given w, we simply integrate out λ to give the expectation of α restricted to its conditional posterior. *This is not the rule.* It is a by-product of the fact that $p(\alpha, \lambda | y, w, x)$ factorizes the way it does. In a way, this factorization of the (conditional) posterior makes life easier—something to keep in mind when we discuss variational inference next.

- As with the previous version, the output of this EM algorithm is a point estimate of w and conditional posteriors on the other variables. We've again made a compromise between the desired full posterior and a point estimate of everything.

- Before discussing how variational methods allow us to approximate the full posterior of all variables, it's worth quickly pointing out another way to approximate the full posterior distribution based on what we've done. This can further highlight how there is no single solution to approximate posterior inference (e.g., Laplace, MCMC, VI, etc.)

    - The output of this EM algorithm are conditional posterior distributions on α and λ and a MAP estimate of w. If we wanted to get some approximation of the posterior distribution of w, notice that we could do a Laplace approximation on $\ln p(y, w|x)$ at the MAP estimate. We now have this! Therefore, we only need to calculate the Hessian $\nabla^2_w \ln p(y, w|x)$ and evaluate it at w_{MAP} in order to get a Gaussian approximation of the posterior of w.

    - In this case, we will be approximating the posterior distribution

$$p(\alpha, \lambda, w | y, x) \approx q(\alpha) q(\lambda) q(w) \tag{21}$$

where

$$\begin{aligned} q(\alpha) = p(\alpha | y, x, w_{\text{MAP}}), \quad q(\lambda) = p(\lambda | w_{\text{MAP}}), \\ q(w) = \text{Normal}(w_{\text{MAP}}, (-\nabla^2_w \ln p(y, w_{\text{MAP}}, |x))^{-1}) \end{aligned} \tag{22}$$

Notice that to do this we will need to actually calculate $p(y, w|x)$, while we didn't need to in order to optimize it over w using EM.

    - This approach of approximating the full posterior with a factorized distribution over its variables will appear again as a major component of variational inference. In the above approach there doesn't appear to be a single, unified objective function that we optimize in order to get these q. We simply combine two techniques: Two of the q are found with EM and the last one with Laplace, and w_{MAP} is what connects them. Variational inference learns this factorized q using a single objective function that is closely related to EM.

From EM to variational inference

- Let's try to be clever. We were able to get conditional posteriors on α and λ by integrating them out and running EM. Can we add w to this? That is, what does it mean to do EM for the marginal distribution

$$p(y|x) = \int\int\int p(y, w, \alpha, \lambda) d\alpha d\lambda dw \tag{23}$$

  First, we notice that there aren't any free parameters in the marginal distribution on the left side. Still, this is a marginal distribution of *something* (the data), and so we can write

$$\begin{aligned} \ln p(y|x) &= \int\int\int q(\alpha, \lambda, w) \ln \frac{p(y, w, \alpha, \lambda|x)}{q(\alpha, \lambda, w)} d\alpha d\lambda dw \; + \\ & \int\int\int q(\alpha, \lambda, w) \ln \frac{q(\alpha, \lambda, w)}{p(\alpha, \lambda, w|y, x)} d\alpha d\lambda dw \end{aligned} \tag{24}$$

- However, this scenario feels different. There is nothing to optimize over the left hand side, so there is no M-step that can be performed. As for the E-step, we need to set

$$q(\alpha, \lambda, w) = p(\alpha, \lambda, w|y, x) \tag{25}$$

- Finally we can see the big problem. Trying to work through EM in this instance requires us to calculate the full posterior distribution of all model variables. This was the problem to begin with and so, even though there is no M-step to perform, we can't get to that point anyway because we can't complete the E-step. Before moving on, I think the following digression is worthwhile.

- Digression: A side comment relating EM to the original core model.

  This is purely a digression. At this point we've laid enough groundwork that we can do something potentially useful very fast. We saw with the core model,

$$y_i \overset{ind}{\sim} \text{Normal}(x_i^T w, \alpha^{-1}), \qquad w \sim \text{Normal}(0, \lambda^{-1} I)$$

  that we could calculate the posterior in closed form and that it was a Gaussian. For this model, we can also write the marginal likelihood $p(y|x) = \int p(y, w|x) dw$ and set up an EM-like equality,

$$\ln p(y|x) = \int q(w) \ln \frac{p(y, w|x)}{q(w)} dw + \int q(w) \ln \frac{q(w)}{p(w|y, x)} dw$$

  Again there is nothing to optimize over on the left hand side. However, we can solve for $q(w)$ in this case because the model is simple enough that we know $p(w|y, x)$. Therefore,

$$p(y|x) = \exp\Big\{ \int p(w|y, x) \ln \frac{p(y, w|x)}{p(w|y, x)} dw \Big\}$$

  Actually we didn't need EM to make this statement (Bayes rule is enough), but given the context of our discussion at this point, I thought I would mention it.

- Back to the main discussion: We have nothing to optimize in the marginal distribution $p(y|x)$ and we can't update $q(\alpha, \lambda, w)$ as EM tells us we must because we don't know the full posterior

distribution of these model variables (which was the problem to begin with). The question is whether the equation:

$$\ln p(y|x) = \int q(\alpha, \lambda, w) \ln \frac{p(y, w, \alpha, \lambda|x)}{q(\alpha, \lambda, w)} d\alpha d\lambda dw + \int q(\alpha, \lambda, w) \ln \frac{q(\alpha, \lambda, w)}{p(\alpha, \lambda, w|y, x)} d\alpha d\lambda dw$$

tells us anything interesting. The answer is yes, and it will lead to a new approximate posterior inference technique called variational inference (VI). I'll refer to this as the VI equation below.

- Before we discuss the three terms in this equation, remember how we set up the EM problem:
  - When we were arriving at this equation we simply said that q is *some* distribution on the model variable we wanted to integrate out in the marginal likelihood. That is, this equation is true for every q we can define, so long as it is defined on the correct space corresponding to what values the variables can take.
  - Then, because we wanted to maximize the LHS (back when it had something we could optimize over, that is), EM told us our *one option* for setting q in order to be able to modify our point estimate such that we can guarantee that we're monotonically increasing the marginal likelihood.

- However, in the VI equation scenario there is nothing to maximize on the LHS. Therefore, this strict requirement made by EM on what we can set q to equal seems irrelevant. To make the transition from EM to VI, we still find this equality to be useful. However, we shift our perspective: With EM we were focused on the LHS and on making sure we were always improving it via the RHS. With VI we are more interested in what is going on with the q distributions, and whether the freedom we have to pick q can be useful.

- In light of this new "VI perspective," let's break down the three terms in the VI equation, keeping in mind that we are not going to be so strict about the distribution $q(\alpha, \lambda, w)$ that we choose:
  - $\ln p(y|x)$: This is the marginal likelihood of the data given the model. We don't know what this is, and in fact it's because of this that we're discussing any of these things to being with. That's because the target posterior is

$$p(\alpha, \lambda, w|y, x) = \frac{p(y, \alpha, \lambda, w|x)}{p(y|x)} \tag{26}$$

    and since we can't solve the integral in the denominator to find $p(y|x)$, we have to use an inference algorithm. Therefore, we can't write out an analytic expression for $\ln p(y|x)$. However, we do know one crucial fact: $\ln p(y|x)$ *is constant*. That is, given the model definition and the data, there are no other degrees of freedom for $p(y|x)$, and so whatever value it takes, it's something that will never change no matter what we do on the RHS of the VI equation.
  - Next, consider the far right term. This is $\mathrm{KL}(q\|p)$ where q is some distribution we're going to pick. This is the KL-divergence between our chosen q and the full posterior. Again, we don't actually know what this is, so we can't give a number for KL in this case. However, we again know two crucial facts about the KL-divergence: *It is always non-negative and only equals zero when* $q = p$. Therefore, this term is a similarity measure between our chosen q distribution and the true, full posterior that we want.

- Finally there is the middle term. Fortunately, this is something we can calculate!—that is, provided that we pick a distribution for $q(\alpha, \lambda, w)$ that helps us toward this end. This is because we have defined the model, so we have defined how to write the function $p(y, w, \alpha, \lambda|x)$. We just need to calculate the integrals.

- Our goal is to pick a $q(\alpha, \lambda, w)$ that is close to the posterior distribution $p(\alpha, \lambda, w|y, x)$. We saw previously how the Laplace approximation tries to do this by letting q be a multivariate Gaussian. However, we didn't really have an objective function we were trying to optimize there. Rather we were solving a second order Taylor approximation problem.

- A natural measure of closeness between $q(\alpha, \lambda, w)$ and $p(\alpha, \lambda, w|y, x)$ is the KL-divergence. Since this measure is not symmetric (i.e., $\text{KL}(q\|p) \neq \text{KL}(p\|q)$), we have two options. Variational inference optimizes $\text{KL}(q\|p)$ using the VI equation, which we equivalently write as

$$\ln p(y|x) = \underbrace{\mathbb{E}_q[\ln p(y, w, \alpha, \lambda|x)] - \mathbb{E}_q[\ln q(\alpha, \lambda, w)]}_{\equiv\ \mathcal{L}\ \leftarrow\ \text{"variational objective function"}} \ +\ \text{KL}(q\|p(\alpha, \lambda, w|y, x))$$

- How can we optimize $\text{KL}(q\|p)$ using this equation? After all, we don't know two of the three terms, including the KL term we're now setting out to optimize. The key insight is that we don't need to know these terms: The LHS is constant and the KL divergence is non-negative. Therefore, $\mathcal{L} + \text{KL}$ must add up to the same number for every possible q that we define. By finding a q that maximizes $\mathcal{L}$, we are equivalently finding a q that minimizes the KL divergence between it and the target posterior distribution because $\text{KL} \geq 0$.

- Therefore, as mentioned, VI is purely interested in the q distribution. And in this case we have a good reason to interpret this q distribution as an approximation to the full posterior distribution.

Variational inference for the core model (version 3.0)

- This leads us to a variational inference algorithm for the last model we have been discussing. This entails the following steps:

  1. We need to define a q distribution *family* for α, λ and w. Compare this with EM in which we were told what to set q equal to.

  2. We then need to construct the *variational objective function*

  $$\mathcal{L} = \mathbb{E}_q[\ln p(y, w, \alpha, \lambda|x)] - \mathbb{E}_q[\ln q(\alpha, \lambda, w)] \tag{27}$$

  After doing this, w, α and λ will be gone since they're integrated out. All that will remain are the parameters of q.

  3. We define the distribution family, meaning we don't actually say what it's parameters are. How do we know what to set these too? Our goal is to maximize $\mathcal{L}$ over these parameters because the VI equation tells us that doing so will minimize the KL-divergence between q and the target posterior. Therefore, as its name implies, we treat $\mathcal{L}$ as an objective function to be maximizes over the parameters of q.

- In the EM algorithm, $\mathcal{L}$ was a function over a model variable—the one in the marginal likelihood we were doing MAP inference for. The parameters of the q distribution were always known because we set them to the conditional posterior distribution. Now, it's the parameters of q that are unknown and $\mathcal{L}$ is a function of these parameters. Therefore, even though it's the same $\mathcal{L}$ in EM and VI, $\mathcal{L}$ is a function of different things in these two methods. This is another subtle difference between VI and EM that really makes them two distinct inference techniques.

- And so it remains to define what $q(\alpha, \lambda, w)$ actually is. We need to pick a recognizable distribution so we can carry out the integral and optimize over its parameters. In all cases this entails some sort of simplification. By far the most common is the "mean-field" assumption (a physics term where these ideas originate from). In this approach we write

$$q(\alpha, \lambda, w) = q(\alpha)q(\lambda)q(w) \tag{28}$$

  and pick individual distributions for each variable. Why did we split across these three? Basically because these were the three variable "units" of the prior (notice that w is a vector, so we don't necessarily split across every variable, although we certainly could define $q(w) = \prod_j q(w_j)$).

- In the next lecture we will see how to find the optimal distribution family of each q. However, we are free to define them however we want and the algorithm will still work, and so we define them to be in the same family as the prior. (These also happen to be the optimal choices.)

$$q(\alpha) = \text{Gamma}(a', b'), \qquad q(\lambda) = \text{Gamma}(e', f'), \qquad q(w) = \text{Normal}(\mu', \Sigma') \tag{29}$$

  We just don't know what a', b', e', f', μ' and Σ' are. To find this we first calculate $\mathcal{L}$.

- Before doing that, why do we pick this q distribution and what is it doing?

  - We pick it, like how we pick many things, purely out of convenience. It's easy to define a distribution on a variable that is in the same family as the prior. By way of comparison, if we wanted to pick a 2-dimensional distribution $q(\alpha, \lambda)$ that wasn't factorizable, what could we pick? (What multivariate non-negative distributions are there readily at hand? The multivariate Gaussian isn't one of them since it is on all $\mathbb{R}^d$, not $\mathbb{R}^d_+$.) Likewise, how can we define a distribution, e.g., $q(\alpha, w)$? What distributions are there that we can easily write down where one dimension is in $\mathbb{R}_+$ and the others in $\mathbb{R}^d$? It's just easier to pick distributions individually for each variable.

  - The other consideration is that we need to be able to calculate $\mathcal{L}$. Picking complicated distributions on large sets of variables may produce an expectation (i.e., an integral) that is not solvable and then we're stuck with an objective we can't even write out, let alone easily optimize.

  - By making this factorization, we are assuming that all variables are independent *in the posterior*. In the core model version 2.0, we saw that α and λ are *conditionally* independent given w, which let us write $q(\alpha, \lambda) = q(\alpha)q(\lambda)$. In that case that was true, but now this is no longer true. That is, for every possible choice of the individual q distributions, we will always have

$$q(\alpha)q(\lambda)q(w) \neq p(\alpha, \lambda, w|y, x) \tag{30}$$

    Therefore, the KL-divergence will always be > 0.

- So now that we've defined q, we need to calculate $\mathcal{L}$. Next week we will see how we can update q in some situations without actually going through $\mathcal{L}$. That is, there is sometimes a trick that can be done where we can go straight to the parameter updates without calculating $\mathcal{L}$.

- However, the default is to calculate $\mathcal{L}$. Unfortunately this is not a pleasant task, but it's the most failsafe approach. The variational objective is

$$\begin{aligned}
\mathcal{L} &= \mathbb{E}_q[\ln p(y, w, \alpha, \lambda|x)] - \mathbb{E}_q[\ln q(\alpha, \lambda, w)] \\
&= -\frac{\mathbb{E}_q[\alpha]}{2}\sum_{i=1}^{N}\mathbb{E}_q[(y_i - x_i^T w)^2] + \frac{1}{2}\mathbb{E}_q[\ln \alpha] - \frac{\mathbb{E}_q[\lambda]}{2}\mathbb{E}_q[w^T w] + \frac{d}{2}\mathbb{E}_q[\ln \lambda] \\
&\quad + (a-1)\mathbb{E}_q[\ln \alpha] - b\mathbb{E}_q[\alpha] + (e-1)\mathbb{E}_q[\ln \lambda] - f\mathbb{E}_q[\lambda] + \text{constant} \\
&\quad - \mathbb{E}_q[\ln q(\alpha)] - \mathbb{E}_q[\ln q(\lambda)] - \mathbb{E}_q[\ln q(w)] \qquad (31)
\end{aligned}$$

- Notice that the last line is the sum of the entropies of each individual q distribution, which is a result of the factorization. At the very least we need to pick q so that we can calculate its entropy.

- The first two lines contain the expected log joint likelihood. Notice that a convenient result of the factorization is that q assumes all variables are independent. Therefore we have simplifications such as

$$\mathbb{E}_q\Big[\frac{\alpha}{2}\sum_{i=1}^{N}(y_i - x_i^T w)^2\Big] = \frac{\mathbb{E}_q[\alpha]}{2}\sum_{i=1}^{N}\mathbb{E}_q[(y_i - x_i^T w)^2] \qquad (32)$$

  And the expectations use only the part of q relevant to the variable being integrated over. This makes calculating $\mathcal{L}$ much easier.

- We will stop here for now. The function $\mathcal{L}$ will be nasty looking. However, after calculating it we can take derivatives and optimize over a', b', e', f', μ' and Σ'. It's not obvious at first sight, but after setting the respective derivatives to zero and solving, the final algorithm will actually be very simple and intuitively satisfying (in my opinion). We'll discuss this later.

- For now, the important take-home message is that by maximizing $\mathcal{L}$ over these six parameters, we are finding a *point estimate* of $q(\alpha)q(\lambda)q(w)$ such that it is an approximation of $p(\alpha, \lambda, w|y, x)$. That is, we get a point estimate of a probability distribution that approximates the posterior.