E6892 Bayesian Models for Machine Learning

Columbia University, Fall 2015

Lecture 6, 10/15/2015

Instructor: John Paisley

Variational inference review (simple notation)

- For this fast review, we compress this into a simple notation. We have data X generated from a model with parameters θ. These parameters are themselves generated from a prior distribution (and so called "variables" from now on), giving the hierarchical representation

$$X \sim p(X|\theta), \quad \theta \sim p(\theta) \tag{1}$$

- We want the posterior distribution

$$p(\theta|X) = \frac{p(X|\theta)p(\theta)}{p(X)}, \tag{2}$$

  but it's intractable. This can be because the prior and likelihood term are not conjugate. Just as common, the variables θ could actually be a large set of variables, the data X a complicated set of data, and the distributions $p(X|\theta)$ and $p(\theta)$ quite complicated in the dependency structure they induce on these variables.

- We approximate $p(\theta|X)$ with a distribution $q(\theta)$ and construct the equation

$$\underbrace{\ln p(X)}_{\text{constant}} = \underbrace{\int q(\theta) \ln \frac{p(X,\theta)}{q(\theta)} d\theta}_{\text{variational objective } \mathcal{L}} + \underbrace{\int q(\theta) \ln \frac{q(\theta)}{p(\theta|X)} d\theta}_{\text{KL-divergence} \geq 0} \tag{3}$$

- We've discussed how we can't actually compute the first and last terms in general, but that by maximizing $\mathcal{L}$ with respect to the parameters of $q(\theta)$, we are minimizing the KL-divergence between $q(\theta)$ and $p(\theta|X)$, which is like a distance measure between the two.

- Therefore, we can view $q(\theta)$ as an approximation of $p(\theta|X)$. The key thing is that we must be able to calculate

$$\mathcal{L} = \int q(\theta) \ln p(X,\theta) d\theta - \int q(\theta) \ln q(\theta) d\theta \tag{4}$$

  Fortunately we often can do this since we define the joint likelihood $p(X,\theta)$ and the distribution $q(\theta)$.

- In fact, the above oversimplifies the problem somewhat. In reality, we have a set of parameters $\theta = (\theta_1, \dots, \theta_m)$ which, along with their prior distributions, defines a joint likelihood on the data

$$p(X, \theta_1, \dots, \theta_m) \tag{5}$$

Notice that, as written, I haven't given you enough information to know how $p(X, \theta_1, \dots, \theta_m)$ should factorize into a product of conditional distributions. The following discussion doesn't change at all based on this factorization, so I will just keep using $p(X, \theta_1, \dots, \theta_m)$.

- Question: How should we pick $q(\theta_1, \dots, \theta_m)$?
Answer: Last week we discussed factorizing q using a "mean-field" assumption.

"Mean-field" assumption

- The "mean-field" assumption gets its name from physics, where these techniques were first developed in the context of a problem where this name makes sense. It constitutes the following steps:

1. Split θ into groups, usually (but not always) according to the "units" in which they are drawn from the prior. Here we've already assumed that to be $\theta_1, \dots, \theta_m$. Notice that we could have $\theta_1 \in \mathbb{R}^d$ and $\theta_2 \in \mathbb{R}$, so it's not correct to think of each θ_i as being in the same space.

2. Define $q_i(\theta_i|\psi_i)$ for variables θ_i. That is, q_i is the distribution family defined for θ_i and ψ_i is its parameters. Often you will just see "$q(\theta_i)$" for short, but for now we'll write it in this more complicated way to keep it more transparent.

3. Let $q(\theta_1, \dots, \theta_m) = \prod_{i=1}^m q_i(\theta_i|\psi_i)$ be the distribution we choose to approximate the posterior $p(\theta_1, \dots, \theta_m|X)$

- Using this q distribution, the variational objective is then computed as

$$\mathcal{L} = \int \Big(\prod_{i=1}^m q_i(\theta_i|\psi_i) \Big) \ln p(X, \theta_1, \dots, \theta_m) d\theta_1 \cdots d\theta_m - \sum_{i=1}^m \int q_i(\theta_i|\psi_i) \ln q_i(\theta_i|\psi_i) d\theta_i \tag{6}$$

- Assuming that we can calculate all of these integrals, the result is a function $\mathcal{L}(\psi_1, \dots, \psi_m)$ that we try to maximize over all ψ_i.

Example (direct method)

- We will refer to the process of explicitly defining each q_i and calculating $\mathcal{L}$ as the "direct method." We show an example of this from the model discussed last week.

- We have data $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^N$ with $x \in \mathbb{R}^d$ and $y \in \mathbb{R}$ and the model

$$y_i \sim \text{Normal}(x_i^T w, \alpha^{-1}), \quad w \sim \text{Normal}(0, \lambda^{-1} I), \quad \alpha \sim \text{Gamma}(a, b) \tag{7}$$

- Based on the definition of this model, the joint likelihood factorizes as

$$p(y, w, \alpha|x) = p(\alpha)p(w)\prod_{i=1}^{N} p(y_i|x_i, w, \alpha) \tag{8}$$

- To approximate the full posterior, we pick the distribution

$$q(w, \alpha) = q(\alpha)q(w) = \text{Gamma}(\alpha|a', b')\text{Normal}(w|\mu', \Sigma') \tag{9}$$

- We the calculate the variational objective function

$$\begin{aligned}\mathcal{L}(a', b', \mu', \Sigma') &= \int q(\alpha)\ln p(\alpha)d\alpha + \int q(w)\ln p(w)dw \\ &+ \sum_{i=1}^{N}\int\int q(\alpha)q(w)\ln p(y_i|x_i, w, \alpha)dwd\alpha \\ &- \int q(\alpha)\ln q(\alpha)d\alpha - \int q(w)\ln q(w)dw \end{aligned} \tag{10}$$

- When we write out the distributions involved, we see that this will require us to calculate $\mathbb{E}[\alpha]$, $\mathbb{E}[\ln\alpha]$, $\mathbb{E}[w]$ and $\mathbb{E}[ww^T]$. This can be looked up when necessary. If we were to actually solve these integrals, we would get something that looks like below:

$$\begin{aligned}\mathcal{L}(a', b', \mu', \Sigma') &= (a-1)(\psi(a') - \ln b') - b\frac{a'}{b'} + \text{constant} \\ &-\frac{\lambda}{2}(\mu'^T\mu' + \text{tr}(\Sigma')) + \text{constant} \\ &+\frac{N}{2}(\psi(a') - \ln b') - \sum_{i=1}^{N}\frac{1}{2}\frac{a'}{b'}\Big((y_i - x_i^T\mu')^2 + x_i^T\Sigma' x_i\Big) + \text{constant} \\ &+a' - \ln b' + \ln\Gamma(a') + (1-a')\psi(a') \\ &+\frac{1}{2}\ln|\Sigma'| + \text{constant}\end{aligned} \tag{11}$$

The only reason this is being written out here is to give an idea of how complicated the variational objective function can look (remember how simple the model is). It's just here to make things 100% concrete, but you don't have to actually parse this if you don't want to. Of course in practice it might be unavoidable that you have to calculate at this level of detail.

- In this function, $\psi(\cdot)$ is a special function called the digamma function, and notice that the priors parameters are included without the $'$.

- Generally speaking, when optimizing $\mathcal{L}$ we want to update *all* parameters of a single q_i distribution at a time. Therefore, we want to update the pair (a', b') together and (μ', Σ') together.

- We'll discuss a faster method later, but the fail-safe method is to directly calculate $\mathcal{L}(a', b', \mu', \Sigma')$ and then solve "two equations, two unknowns" problem $\partial\mathcal{L}/\partial a' = 0$ and $\partial\mathcal{L}/\partial b' = 0$ for (a', b') and similarly for (μ', Σ').

Picking and then solving $q_i(\theta_i|\psi_i)$

- We return to the general problem where we have data X and model variables $\theta_1, \dots, \theta_m$. After defining how $q(\theta_1, \dots, \theta_m)$ factorizes, the two key problems are

  1. Picking the distribution family of $q_i(\theta_i|\psi_i)$ (e.g., Gaussian, gamma, etc.)
  2. Finding a way to update the parameters ψ_i to find the best q_i in its family

- Previously, we solved #1 arbitrarily and #2 through brute force calculation of $\mathcal{L}$. We have two natural questions that arise at this point:

  Q1: Is there a way to pick a *best* family for q_i?

  Q2: Is there a faster way to solve for ψ_i that bypasses calculating $\mathcal{L}$?

  The answer to both of these questions is "yes" (at least in principle) and we can get the solutions to both questions at once! We will now show how to do this.

- General setup: Exactly as before, we have a joint likelihood $p(X, \theta_1, \dots, \theta_m)$ and we approximate the posterior distribution $p(\theta_1, \dots, \theta_m|X) \approx \prod_{i=1}^m q_i(\theta_i|\psi_i)$.

- Let's now focus on the q distribution for a specific variable, say $q_i(\theta_i|\psi_i)$. Our goal is to think about the variational objective function only in the context of this q_i distribution and see what it tells us,

$$\begin{aligned}\mathcal{L} &= \int q_i(\theta_i|\psi_i) \underbrace{\left[\int \Big(\prod_{j\neq i} q_j(\theta_j|\psi_j)\Big) \ln p(X, \theta_1, \dots, \theta_m) d\theta_{j\neq i}\right]}_{= \mathbb{E}_{q_{j\neq i}}[\ln p(X,\theta_1,\dots,\theta_m)]} d\theta_i && (12)\\ &\quad - \int q_i(\theta_i|\psi_i) \ln q_i(\theta_i|\psi_i) d\theta_i - \sum_{j\neq i} \int q_j(\theta_j|\psi_j) \ln q_j(\theta_j|\psi_j) d\theta_j && (13)\end{aligned}$$

- As indicated, the term $\mathbb{E}_{q_{j\neq i}}[\ln p(X, \theta_1, \dots, \theta_m)]$ is the expectation over all θ_j for $j \neq i$. Therefore, the result of this expectation is a function of θ_i and of ψ_j for all $j \neq i$.

- We haven't defined any other q_j yet either, but we'll see that this doesn't matter for choosing the family of q_i.

- The next step is to manipulate how we write the variational objective function. First,

$$\begin{aligned}\mathcal{L} &= \int q_i(\theta_i|\psi_i) \mathbb{E}_{q_{j\neq i}}[\ln p(X, \theta_1, \dots, \theta_m)] d\theta_i - \int q_i(\theta_i|\psi_i) \ln q_i(\theta_i|\psi_i) d\theta_i + \text{const. w.r.t. } \theta_i \\ &= \int q_i(\theta_i|\psi_i) \ln \frac{e^{\mathbb{E}_{q_{j\neq i}}[\ln p(X,\theta_1,\dots,\theta_m)]}}{q_i(\theta_i|\psi_i)} d\theta_i + \text{const. w.r.t. } \theta_i && (14)\end{aligned}$$

- Next we add and subtract $\ln Z$ and define

$$Z = \int e^{\mathbb{E}_{q_{j\neq i}}[\ln p(X,\theta_1,\dots,\theta_m)]} d\theta_i \qquad (15)$$

This gives

$$\mathcal{L} = \int q_i(\theta_i|\psi_i) \ln \frac{\frac{1}{Z}\mathrm{e}^{\mathbb{E}_{q_{j\neq i}}[\ln p(X,\theta_1,\ldots,\theta_m)]}}{q_i(\theta_i|\psi_i)} d\theta_i + \ln Z + \text{const. w.r.t. } \theta_i \tag{16}$$

- What is $\frac{1}{Z}\mathrm{e}^{\mathbb{E}_{q_{j\neq i}}[\ln p(X,\theta_1,\ldots,\theta_m)]}$? Notice that we can think of it as a probability distribution on θ_i. As we will now see, this is very relevant, since our goal is to pick $q_i(\theta_i|\psi_i)$ such that, if we only focus on this term and ignore all other q_j, we maximize

$$\int q_i(\theta_i|\psi_i) \ln \frac{\frac{1}{Z}\mathrm{e}^{\mathbb{E}_{q_{j\neq i}}[\ln p(X,\theta_1,\ldots,\theta_m)]}}{q_i(\theta_i|\psi_i)} d\theta_i = -\mathrm{KL}\big(q_i \| \tfrac{1}{Z}\mathrm{e}^{\mathbb{E}_{q_{j\neq i}}[\ln p(X,\theta_1,\ldots,\theta_m)]}\big) \tag{17}$$

- Remember that $KL \geq 0$ and so $-\mathrm{KL} \leq 0$. We know when KL is minimized, so we equivalently know when the *negative* KL is maximized, that is, we now know we should set

$$q_i(\theta_i|\psi_i) = \frac{1}{Z}\mathrm{e}^{\mathbb{E}_{q_{j\neq i}}[\ln p(X,\theta_1,\ldots,\theta_m)]} \tag{18}$$

- Notice a few things about this:

  1. This gives the optimal family for $q_i(\theta_i|\psi_i)$. It doesn't matter what the other q_j are or their ψ_j parameters.

  2. This also gives the optimal parameters ψ_i of $q_i(\theta_i|\psi_i)$ for given settings of the other q_j. If we know all other q_j except for q_i, there's nothing unknown in this optimal $q_i(\theta_i|\psi_i)$. Put another way, we haven't just written out a distribution family above, we've written out a specific distribution including all its parameters.

  3. If we don't know what the family of the other q_j are (e.g., Gaussian, gamma, etc.) we still can say the *family* of $q_i(\theta_i|\psi_i)$. That is, the above equation will allow us to say something like "q_i should be defined to be a gamma distribution."

  4. Since our choice of i was arbitrary, we therefore know what *every* q_i should be set to by looking at the *form* of $\exp\{\mathbb{E}_{q_{j\neq i}}[\ln p(X,\theta_1,\ldots,\theta_m)]\}$ for each i. Therefore, this is the first step in variational inference: To iterate between each θ_i and decide what q_i distribution should be defined for it.

  5. We can write this abstractly, but if the integral is intractable (in the numerator or the denominator), then this does us no good. Does this work in general? The answer is "yes," for a large class of models (to be discussed in a later lecture).

- In words, this says that to find the optimal q distribution for θ_i:

  1. Take the log of the joint likelihood

  2. Take the expectation of all variables using their respective q distribution except for $q_i(\theta_i|\psi_i)$

  3. Exponentiate the result and normalize

General variational inference algorithm

- Given data X and a joint likelihood $p(X, \theta_1, \dots, \theta_m)$
  - For iteration $t = 1, 2, \dots$
    1. For model variable index $i = 1, \dots, m$ set
    $$q_i(\theta_i|\psi_i) = \frac{\mathrm{e}^{\mathbb{E}_{q_{j \neq i}}[\ln p(X,\theta_1,\dots,\theta_m)]}}{\int \mathrm{e}^{\mathbb{E}_{q_{j \neq i}}[\ln p(X,\theta_1,\dots,\theta_m)]} d\theta_i}$$
    2. Evaluate the variational objective function using the updated q
    $$\mathcal{L}_t = \mathbb{E}_q[\ln p(X, \theta_1, \dots, \theta_m)] - \sum_{i=1}^{m} \mathbb{E}_{q_i}[\ln q_i(\theta_i|\psi_i)]$$
    3. If the marginal increase in $\mathcal{L}_t$ compared with $\mathcal{L}_{t-1}$ is “small,” terminate, otherwise continue to the next iteration.

- As with Gibbs sampling, we always use the most recent parameters for all $q_j(\theta_j|\psi_j)$ when updating q_i. Let’s look at a concrete example.

Example (optimal method)

- We return to the original example

$$y_i \sim \text{Normal}(x_i^T w, \alpha^{-1}), \quad w \sim \text{Normal}(0, \lambda^{-1} I), \quad \alpha \sim \text{Gamma}(a, b) \tag{19}$$

- We want to approximate the posterior $p(w, \alpha|y, x)$ with $q(\alpha, w)$ using variational inference and a mean-field assumption. Therefore, the first step is to choose the factorization of q. Again we choose
$$p(\alpha, w|y, x) \approx q(\alpha, w) \equiv q(\alpha)q(w)$$

- Next, we want to learn what distribution family we should set $q(\alpha)$ and $q(w)$ to be. For example, should $q(\alpha)$ be Gaussian? Should it be gamma? Poisson? etc.

- $\boldsymbol{q(\alpha)}$: We know from the general approach that we can find $q(\alpha)$ as follows:

$$\begin{aligned} q(\alpha) &\propto \exp\left\{\mathbb{E}_{q(w)}[\ln p(y|x,\alpha,w) + \ln p(\alpha) + \ln p(w)]\right\} \\ &\propto \exp\left\{\mathbb{E}_{q(w)}[\ln p(y|x,\alpha,w)] + \ln p(\alpha)\right\} \end{aligned} \tag{20}$$

Notice that we can remove any terms not involving α, therefore $\ln p(w)$ is removed. Also, the expectation is only over w, therefore the expectation doesn’t impact $\ln p(\alpha)$. Then,

$$\begin{aligned} q(\alpha) &\propto \exp\left\{ \textstyle\sum_{i=1}^{N} \mathbb{E}_{q(w)}[\ln p(y_i|x_i,\alpha,w)]\right\} p(\alpha) \\ &\propto \Big[\prod_{i=1}^{N} \alpha^{\frac{1}{2}} \mathrm{e}^{-\frac{\alpha}{2}\mathbb{E}_{q(w)}[(y_i - x_i^T w)^2]}\Big] \alpha^{a-1} \mathrm{e}^{-b\alpha} \end{aligned} \tag{21}$$

- Since we haven't defined $q(w)$ yet, we can't take this expectation. However, notice that we have enough information to be able to say that

$$q(\alpha) = \text{Gamma}(\alpha|a', b'), \quad a' = a + \frac{N}{2}, \quad b' = b + \frac{1}{2}\sum_{i=1}^{N} \mathbb{E}_{q(w)}[(y_i - x_i^T w)^2] \tag{22}$$

- $\boldsymbol{q(w)}$: Next, we perform similar operations to find the optimal $q(w)$,

$$\begin{aligned} q(w) &\propto \exp\left\{\mathbb{E}_{q(\alpha)}[\ln p(y|x,\alpha,w) + \ln p(\alpha) + \ln p(w)]\right\} \\ &\propto \exp\left\{\mathbb{E}_{q(\alpha)}[\ln p(y|x,\alpha,w)] + \ln p(w)\right\} \end{aligned} \tag{23}$$

Again we can remove any terms not involving w, and so $\ln p(\alpha)$ is removed. Also, since the expectation is only over α, we don't have an expectation for the term $\ln p(w)$. Again, we continue:

$$\begin{aligned} q(w) &\propto \exp\left\{\textstyle\sum_{i=1}^N \mathbb{E}_{q(\alpha)}[\ln p(y_i|x_i,\alpha,w)]\right\}p(w) \\ &\propto \Big[\prod_{i=1}^N \mathrm{e}^{\frac{1}{2}\mathbb{E}[\ln\alpha]}\mathrm{e}^{-(\mathbb{E}_{q(\alpha)}[\alpha]/2)(y_i - x_i^T w)^2}\Big]\mathrm{e}^{-\frac{\lambda}{2}w^T w} \end{aligned} \tag{24}$$

First notice that we can simply ignore $\mathrm{e}^{\frac{1}{2}\mathbb{E}[\ln\alpha]}$ because it will be canceled out when we compute the normalizing constant.

- We already saw this type of proportionality before when we calculated the posterior of w in the Bayesian linear regression problem. As a result, we know that

$$q(w) = \text{Normal}(w|\mu', \Sigma') \tag{25}$$

$$\Sigma' = \Big(\lambda I + \mathbb{E}_{q(\alpha)}[\alpha]\sum_{i=1}^N x_i x_i^T\Big)^{-1}, \qquad \mu' = \Sigma'\Big(\mathbb{E}_{q(\alpha)}[\alpha]\sum_{i=1}^N y_i x_i\Big)$$

- Notice that by cycling through each parameter and learning its q distribution, we also learn what the expectations are when learning other q distributions. That is, when we found that $q(\alpha)$ was a gamma distribution, we weren't able to say what the expectation with respect to $q(w)$ was because we didn't know what distribution to use for $q(w)$. Now we know it's Gaussian and so we can retroactively solve for the expectation.

- Similarly, because we first found that $q(\alpha)$ was a gamma distribution, we know what the expectation is that we should use when we update $q(w)$. This is the general pattern. We first find what the distributions are for each q_i, and after we have them all, we will know what all the expectations will be.

- For example, for this problem

$$\mathbb{E}_{q(\alpha)}[\alpha] = a'/b' \tag{26}$$

$$\mathbb{E}_{q(w)}[(y_i - x_i^T w)^2] = (y_i - x_i^T\mu')^2 + x_i^T\Sigma' x_i \tag{27}$$

- Let's look at the final variational inference algorithm for this problem.

VI algorithm for Bayesian linear regression with unknown noise precision

Inputs: Data and definitions $q(\alpha) = \text{Gamma}(\alpha|a', b')$ and $q(w) = \text{Normal}(w|\mu', \Sigma')$

Output: Values for a', b', μ' and Σ'

1. Initialize a'_0, b'_0, μ'_0 and Σ'_0 in some way
2. For iteration $t = 1, \dots, T$
   - Update $q(\alpha)$ by setting

$$\begin{aligned} a'_t &= a + \frac{N}{2} \\ b'_t &= b + \frac{1}{2}\sum_{i=1}^{N} (y_i - x_i^T \mu'_{t-1})^2 + x_i^T \Sigma'_{t-1} x_i \end{aligned}$$

   - Update $q(w)$ by setting

$$\begin{aligned} \Sigma'_t &= \Big(\lambda I + \frac{a'_t}{b'_t}\sum_{i=1}^{N} x_i x_i^T\Big)^{-1} \\ \mu'_t &= \Sigma'_t\Big(\frac{a'_t}{b'_t}\sum_{i=1}^{N} y_i x_i\Big) \end{aligned}$$

   - Evaluate $\mathcal{L}(a'_t, b'_t, \mu'_t, \Sigma'_t)$ to assess convergence (i.e., decide T).

- Notice that this is exactly the solution we would have found if we solved the system of equations

$$\frac{\partial \mathcal{L}(a', b', \mu'_{t-1}, \Sigma'_{t-1})}{\partial a'} = 0, \quad \frac{\partial \mathcal{L}(a', b', \mu'_{t-1}, \Sigma'_{t-1})}{\partial b'} = 0$$

  to find a'_t and b'_t. We also would find this update for μ'_t and Σ'_t by solving the system

$$\nabla_{\mu'} \mathcal{L}(a'_t, b'_t, \mu', \Sigma') = 0, \quad \nabla_{\Sigma'} \mathcal{L}(a'_t, b'_t, \mu', \Sigma') = 0$$

- You can also see the different between VI for this model and EM for maximizing $\ln p(y, w|x)$ using α as the marginalized variable, which we discussed earlier. In EM, the update for the point estimate is $w_t = \mu'_t$ using μ'_t from above. When we wanted to update $q(\alpha)$, we again had a gamma distribution, but because there was no distribution on w, we simply plugged in w_t where μ'_t appears and remove Σ'_t (again, because there is no uncertainty about w.

- Comment: Sometimes researchers will define a q "distribution" that is a point mass – that is, they will let $q(w) = \delta_{w'}$ for example. For this distribution, $P(w = w') = 1$. When we "integrate" using this $q(w)$ to calculate $\mathcal{L}$, we find that we simply replace w with w' and do a point estimate of w' (and ignore the technical issue that the entropy of $\delta_{w'}$ is $-\infty$). This is sometimes done when the integral using any other $q(w)$ is intractable, which is not the case here.

VI algorithm for probit regression

- Recall the setup (slightly modified): We have data $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^N$ where $x \in \mathbb{R}^d$ and $y \in \{-1, +1\}$ (previously $\{0, 1\}$, but notice that it's no difference below). Including a hidden variable ϕ_i for each (x_i, y_i) pair, the probit regression model is

$$y_i = \text{sign}(\phi_i), \quad \phi_i \sim \text{Normal}(x_i^T w, \sigma^2), \quad w \sim \text{Normal}(0, \lambda^{-1} I) \tag{28}$$

- Previously, we derived an EM algorithm for maximizing the marginal distribution $p(y, w|x)$ under this model. Let's now look at a variational inference algorithm.

- The unknowns this time are w and the vector $\phi = (\phi_1, \ldots, \phi_N)$. We set up the variational inference equation

$$\ln p(y|x) = \int q(w, \phi) \ln \frac{p(y, w, \phi|x)}{q(w, \phi)} dw d\phi + \int q(w, \phi) \ln \frac{q(w, \phi)}{p(w, \phi|y, x)} dw d\phi \tag{29}$$

- From this setup, we can see that we're designing $q(w, \phi)$ to approximate the full posterior of both w and the extra variables ϕ together. We pick the factorization

$$q(w, \phi) = q(w) \prod_{i=1}^N q(\phi_i) \tag{30}$$

  The first problem using our "optimal approach" is to find out what these distributions should be. First, notice that the joint likelihood factorizes as

$$p(y, w, \phi|x) = p(w) \prod_{i=1}^N p(y_i|\phi_i) p(\phi_i|w, x_i) \tag{31}$$

  The "distribution" $p(y_i|\phi_i)$ has no randomness in it: $p(y_i|\phi_i) = \mathbb{1}\{y_i = \text{sign}(\phi_i)\}$.

- $\boldsymbol{q(\phi_i)}$ **:** Following the rules, we take the log of $p(y, w, \phi|x)$ and the expectation with respect to $q(w)$ and $q(\phi_j)$ for $j \neq i$, and then exponentiate. As a result, we can write that

$$\begin{aligned} q(\phi_i) \;\; \propto \;\; & \exp\Big\{\mathbb{E}_{q(w)}[\ln p(w)] + \sum_{j \neq i} \mathbb{E}_{q(\phi_j)}[\ln p(y_j|\phi_j) + \ln p(\phi_j|w, x_j)]\Big\} \times \\ & \exp\Big\{ \ln p(y_i|\phi_i) + \mathbb{E}_{q(w)}[\ln p(\phi_i|w, x_i)\Big\} \end{aligned} \tag{32}$$

  Notice that the first line doesn't contain anything having to do with ϕ_i, therefore it's a constant as far as the proportionality is concerned and can be ignored. As a result,

$$\begin{aligned} q(\phi_i) & \propto \mathbb{1}\{y_i = \text{sign}(\phi_i)\} \exp\{-\tfrac{1}{2\sigma^2}\mathbb{E}_{q(w)}[(\phi_i - x_i^T w)^2]\} \\ & \propto \mathbb{1}\{y_i = \text{sign}(\phi_i)\} \exp\{-\tfrac{1}{2\sigma^2}[(\phi_i - x_i^T \mathbb{E}_{q(w)}[w])^2 + x_i^T \mathbb{E}_{q(w)}[ww^T]x_i]\} \\ & \propto \mathbb{1}\{y_i = \text{sign}(\phi_i)\} \exp\{-\tfrac{1}{2\sigma^2}(\phi_i - x_i^T \mathbb{E}_{q(w)}[w])^2\} \exp\{-\tfrac{1}{2\sigma^2} x_i^T \mathbb{E}_{q(w)}[ww^T]x_i\} \\ & \propto \mathbb{1}\{y_i = \text{sign}(\phi_i)\} \exp\{-\tfrac{1}{2\sigma^2}(\phi_i - x_i^T \mathbb{E}_{q(w)}[w])^2\} \end{aligned} \tag{33}$$

This is a truncated normal on the half of $\mathbb{R}$ defined by y_i. Therefore, the optimal $q(\phi_i)$ is

$$q(\phi_i) = \text{TN}_{y_i}(\mu'_{\phi_i}, \sigma^2), \qquad \mu'_{\phi_i} = x_i^T \mathbb{E}_{q(w)}[w] \tag{34}$$

- $\boldsymbol{q(w)}$ **:** We use the same approach to find the optimal q distribution of w,

$$\begin{aligned} q(w) &\propto \exp\Big\{\ln p(w) + \sum_{i=1}^{N} \mathbb{E}_{q(\phi_i)}[\ln p(y_j|\phi_j) + \ln p(\phi_j|w, x_j)]\Big\} \\ &\propto \exp\Big\{\ln p(w) + \sum_{i=1}^{N} \mathbb{E}_{q(\phi_i)}[\ln p(\phi_j|w, x_j)]\Big\} \\ &\propto \mathrm{e}^{-\frac{\lambda}{2}w^T w} \prod_{i=1}^{N} \mathrm{e}^{-\frac{1}{2\sigma^2}(\mathbb{E}_{q(\phi_i)}[\phi_i] - x_i^T w)^2} \end{aligned} \tag{35}$$

This is exactly the Bayesian linear regression problem we have discussed. The difference is that because we don't observe ϕ_i, we end up using the expectation of this latent variable using its $q(\phi_i)$ distribution. As a result,

$$q(w) = \text{Normal}(\mu', \Sigma') \tag{36}$$

$$\Sigma' = \Big(\lambda I + \frac{1}{\sigma^2}\sum_{i=1}^{N} x_i x_i^T\Big)^{-1}, \qquad \mu' = \Sigma'\Big(\frac{1}{\sigma^2}\sum_{i=1}^{N} \mathbb{E}_{q(\phi_i)}[\phi_i] x_i\Big)$$

- Now that we've defined the q distributions, we can write out the relevant expectations,

$$\mathbb{E}_q[\phi_i] = \begin{cases} x_i^T \mu'_{\phi_i} + \sigma \times \dfrac{\Phi'(-x_i^T \mu'_{\phi_i}/\sigma)}{1 - \Phi(-x_i^T \mu'_{\phi_i}/\sigma)} & \text{if } y_i = +1 \\ \\ x_i^T \mu'_{\phi_i} + \sigma \times \dfrac{-\Phi'(-x_i^T \mu'_{\phi_i}/\sigma)}{\Phi(-x_i^T \mu'_{\phi_i}/\sigma)} & \text{if } y_i = -1 \end{cases}$$

$$\mathbb{E}_{q(w)}[w] = \mu'$$

- As before, Φ is the CDF of a standard normal and Φ' is its PDF.

- This gives the following two steps that can be iterated,

  - At iteration t

    1. Update μ'_{ϕ_i} for each i using the μ' from iteration $t-1$ as in Equation (34)
    2. Update μ' and Σ' using all μ'_{ϕ_i} just updated in Step 1 as in Equation (36)

  - Assess convergence by evaluating the variational objective $\mathcal{L}(\mu', \Sigma', \mu'_{\phi_1}, \dots, \mu'_{\phi_N})$ using the values of these variational parameters from iteration t.

- As a (painful) exercise, you can try deriving this variational inference algorithm using the "direct method" of first calculating $\mathcal{L}$ and then taking derivatives, setting to zero and solving. This will make it very easy to appreciate how much easier this optimal approach is. Also, if we weren't able to find the optimal q distributions first, and instead defined some other distribution for $q(\phi_i)$, the direct method most likely would lead to a dead end.