

Variational Inference

David M. Blei

1 Set up

- As usual, we will assume that $x = x_{1:n}$ are observations and $z = z_{1:m}$ are hidden variables. We assume additional parameters α that are fixed.
- Note we are general—the hidden variables might include the “parameters,” e.g., in a traditional inference setting. (In that case, α are the hyperparameters.)
- We are interested in the **posterior distribution**,

$$p(z | x, \alpha) = \frac{p(z, x | \alpha)}{\int_z p(z, x | \alpha)}. \quad (1)$$

- As we saw earlier, the posterior links the data and a model. It is used in all downstream analyses, such as for the predictive distribution.
- (Note: The problem of computing the posterior is an instance of a more general problem that variational inference solves.)

2 Motivation

- We can't compute the posterior for many interesting models.
- Consider the Bayesian mixture of Gaussians,
 1. Draw $\mu_k \sim \mathcal{N}(0, \tau^2)$ for $k = 1 \dots K$.
 2. For $i = 1 \dots n$:
 - (a) Draw $z_i \sim \text{Mult}(\pi)$;

(b) Draw $x_i \sim \mathcal{N}(\mu_{z_i}, \sigma^2)$.

- Suppressing the fixed parameters, the posterior distribution is

$$p(\mu_{1:K}, z_{1:n} \mid x_{1:n}) = \frac{\prod_{k=1}^K p(\mu_k) \prod_{i=1}^n p(z_i) p(x_i \mid z_i, \mu_{1:K})}{\int_{\mu_{1:K}} \sum_{z_{1:n}} \prod_{k=1}^K p(\mu_k) \prod_{i=1}^n p(z_i) p(x_i \mid z_i, \mu_{1:K})}. \quad (2)$$

- The numerator is easy to compute for any configuration of the hidden variables. The problem is the denominator.
- Let's try to compute it. First, we can take advantage of the conditional independence of the z_i 's given the cluster centers,

$$p(x_{1:n}) = \int_{\mu_{1:K}} \prod_{k=1}^K p(\mu_k) \prod_{i=1}^n \sum_{z_i} p(z_i) p(x_i \mid z_i, \mu_{1:K}). \quad (3)$$

This leads to an integral that we can't (easily, anyway) compute.

- Alternatively, we can move the summation over the latent assignments to the outside,

$$p(x_{1:n}) = \int_{\mu_{1:K}} \prod_{k=1}^K p(\mu_k) \prod_{i=1}^n \sum_{z_i} p(z_i) p(x_i \mid z_i, \mu_{1:K}). \quad (4)$$

It turns out that we can compute each term in this summation. (This is an exercise.) However, there are K^n terms. This is intractable when n is reasonably large.

- This situation arises in most interesting models. This is why approximate posterior inference is one of the central problems in Bayesian statistics.

3 Main idea

- We return to the general $\{x, z\}$ notation.
- The main idea behind variational methods is to pick a family of distributions over the latent variables with its own **variational parameters**,

$$q(z_{1:m} \mid \nu). \quad (5)$$

- Then, find the setting of the parameters that makes q close to the posterior of interest.

- Use q with the fitted parameters as a proxy for the posterior, e.g., to form predictions about future data or to investigate the posterior distribution of the hidden variables.
- Typically, the true posterior is not in the variational family. (Draw the picture from Wainwright and Jordan, 2008.)

4 Kullback-Leibler Divergence

- We measure the closeness of the two distributions with Kullback-Leibler (KL) divergence.
- This comes from **information theory**, a field that has deep links to statistics and machine learning. (See the books “Information Theory and Statistics” by Kullback and “Information Theory, Inference, and Learning Algorithms” by MacKay.)

- The KL divergence for variational inference is

$$\text{KL}(q||p) = \mathbb{E}_q \left[\log \frac{q(Z)}{p(Z|x)} \right]. \quad (6)$$

- Intuitively, there are three cases
 - If q is high and p is high then we are happy.
 - If q is high and p is low then we pay a price.
 - If q is low then we don’t care (because of the expectation).
- (Draw a multi-modal posterior and consider various possibilities for single modes.)
- Note that we could try to reverse these arguments. In a way, that makes more intuitive sense. However, we choose q so that we can take expectations.
- That said, reversing the arguments leads to a different kind of variational inference than we are discussing. It is called “expectation propagation.” (In general, it’s more computationally expensive than the algorithms we will study.)

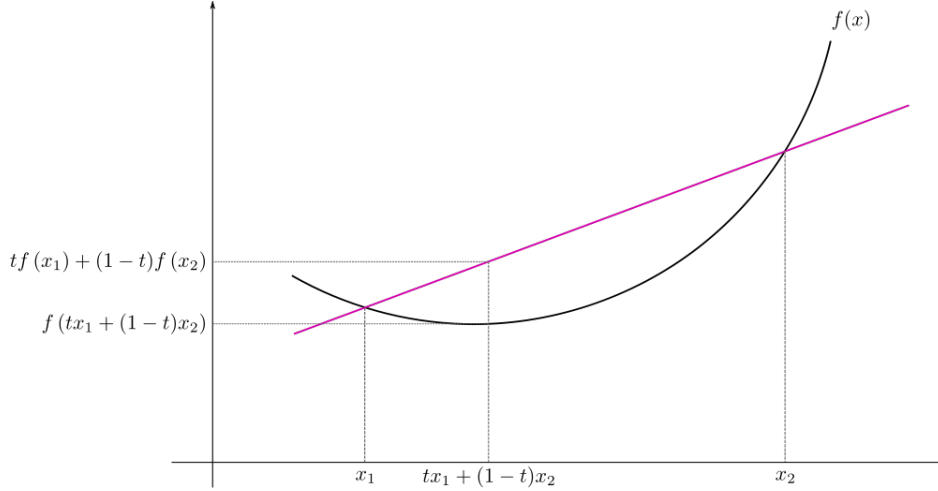
5 The evidence lower bound

- We actually can’t minimize the KL divergence exactly, but we can minimize a function that is equal to it up to a constant. This is the **evidence lower bound** (ELBO).

- Recall Jensen's inequality as applied to probability distributions. When f is concave,

$$f(\mathbb{E}[X]) \geq \mathbb{E}[f(X)]. \quad (7)$$

- If you haven't seen Jensen's inequality, spend 15 minutes to learn about it.



(This figure is from Wikipedia.)

- We use Jensen's inequality on the log probability of the observations,

$$\log p(x) = \log \int_z p(x, z) \quad (8)$$

$$= \log \int_z p(x, z) \frac{q(z)}{q(z)} \quad (9)$$

$$= \log \left(\mathbb{E}_q \left[\frac{p(x, Z)}{q(Z)} \right] \right) \quad (10)$$

$$\geq \mathbb{E}_q[\log p(x, Z)] - \mathbb{E}_q[\log q(Z)]. \quad (11)$$

This is the ELBO. (Note: This is the same bound used in deriving the expectation-maximization algorithm.)

- We choose a family of variational distributions (i.e., a parameterization of a distribution of the latent variables) such that the expectations are computable.
- Then, we maximize the ELBO to find the parameters that gives as tight a bound as possible on the marginal probability of x .
- Note that the second term is the entropy, another quantity from information theory.

- What does this have to do with the KL divergence to the posterior?

– First, note that

$$p(z | x) = \frac{p(z, x)}{p(x)}. \quad (12)$$

– Now use this in the KL divergence,

$$\text{KL}(q(z) || p(z | x)) = \mathbb{E}_q \left[\log \frac{q(Z)}{p(Z | x)} \right] \quad (13)$$

$$= \mathbb{E}_q[\log q(Z)] - \mathbb{E}_q[\log p(Z | x)] \quad (14)$$

$$= \mathbb{E}_q[\log q(Z)] - \mathbb{E}_q[\log p(Z, x)] + \log p(x) \quad (15)$$

$$= -(\mathbb{E}_q[\log p(Z, x)] - \mathbb{E}_q[\log q(Z)]) + \log p(x) \quad (16)$$

This is the negative ELBO plus the log marginal probability of x .

- Notice that $\log p(x)$ does not depend on q . So, as a function of the variational distribution, minimizing the KL divergence is the same as maximizing the ELBO.
- And, the difference between the ELBO and the KL divergence is the log normalizer—which is what the ELBO bounds.

6 Mean field variational inference

- In mean field variational inference, we assume that the variational family **factorizes**,

$$q(z_1, \dots, z_m) = \prod_{j=1}^m q(z_j). \quad (17)$$

Each variable is independent. (We are suppressing the parameters ν_j .)

- This is more general than it initially appears—the hidden variables can be grouped and the distribution of each group factorizes.
- Typically, this family does not contain the true posterior because the hidden variables are dependent.
 - E.g., in the Gaussian mixture model all of the cluster assignments z_i are dependent on each other and the cluster locations $\mu_{1:K}$ given the data $x_{1:n}$.
 - These dependencies are often what makes the posterior difficult to work with.

– (Again, look at the picture from Wainwright and Jordan.)

- We now turn to optimizing the ELBO for this factorized distribution.
- We will use **coordinate ascent inference**, iteratively optimizing each variational distribution holding the others fixed.
- We emphasize that this is not the only possible optimization algorithm. Later, we'll see one based on the natural gradient.
- First, recall the chain rule and use it to decompose the joint,

$$p(z_{1:m}, x_{1:n}) = p(x_{1:n}) \prod_{j=1}^m p(z_j \mid z_{1:(j-1)}, x_{1:n}) \quad (18)$$

Notice that the z variables can occur in any order in this chain. The indexing from 1 to m is arbitrary. (This will be important later.)

- Second, decompose the entropy of the variational distribution,

$$\mathbb{E}[\log q(z_{1:m})] = \sum_{j=1}^m \mathbb{E}_j[\log q(z_j)], \quad (19)$$

where \mathbb{E}_j denotes an expectation with respect to $q(z_j)$.

- Third, with these two facts, decompose the the ELBO,

$$\mathcal{L} = \log p(x_{1:n}) + \sum_{j=1}^m \mathbb{E}[\log p(z_j \mid z_{1:(j-1)}, x_{1:n})] - \mathbb{E}_j[\log q(z_j)]. \quad (20)$$

- Consider the ELBO as a function of $q(z_k)$.
 - Employ the chain rule with the variable z_k as the last variable in the list.
 - This leads to the objective function

$$\mathcal{L} = \mathbb{E}[\log p(z_k \mid z_{-k}, x)] - \mathbb{E}_j[\log q(z_k)] + \text{const}. \quad (21)$$

- Write this objective as a function of $q(z_k)$,

$$\mathcal{L}_k = \int q(z_k) \mathbb{E}_{-k}[\log p(z_k \mid z_{-k}, x)] dz_k - \int q(z_k) \log q(z_k) dz_k. \quad (22)$$

- Take the derivative with respect to $q(z_k)$

$$\frac{d\mathcal{L}_j}{dq(z_k)} = \mathbb{E}_{-k}[\log p(z_k | z_{-k}, x)] - \log q(z_k) - 1 = 0 \quad (23)$$

- This (and Lagrange multipliers) leads to the coordinate ascent update for $q(z_k)$

$$q^*(z_k) \propto \exp\{\mathbb{E}_{-k}[\log p(z_k | Z_{-k}, x)]\} \quad (24)$$

- But the denominator of the posterior does not depend on z_j , so

$$q^*(z_k) \propto \exp\{\mathbb{E}_{-k}[\log p(z_k, Z_{-k}, x)]\} \quad (25)$$

- Either of these perspectives might be helpful in deriving variational inference algorithms.

- The coordinate ascent algorithm is to iteratively update each $q(z_k)$. The ELBO converges to a *local minimum*. Use the resulting q is as a proxy for the true posterior.

- Notice

- The RHS only depends on $q(z_j)$ for $j \neq k$ (because of factorization).
- This determines the form of the optimal $q(z_k)$. We didn't specify the form in advance, only the factorization.
- Depending on that form, the optimal $q(z_k)$ might not be easy to work with. However, for many models it is. (Stay tuned.)

- There is a strong relationship between this algorithm and Gibbs sampling.

- In Gibbs sampling we sample from the conditional.
- In coordinate ascent variational inference, we iteratively set each factor to

$$\text{distribution of } z_k \propto \exp\{\mathbb{E}[\log(\text{conditional})]\}. \quad (26)$$

- Easy example: Multinomial conditionals

- Suppose the conditional is multinomial

$$p(z_j | z_{-j}, x_{1:n}) := \pi(z_{-j}, x_{1:n}) \quad (27)$$

- Then the optimal $q(z_j)$ is also a multinomial,

$$q^*(z_j) \propto \exp\{\mathbb{E}[\log \pi(z_{-j}, x)]\} \quad (28)$$

7 Exponential family conditionals

- Suppose each conditional is in the exponential family

$$p(z_j | z_{-j}, x) = h(z_j) \exp\{\eta(z_{-j}, x)^\top t(z_j) - a(\eta(z_{-j}, x))\} \quad (29)$$

- This describes *a lot* of complicated models
 - Bayesian mixtures of exponential families with conjugate priors
 - Switching Kalman filters
 - Hierarchical HMMs
 - Mixed-membership models of exponential families
 - Factorial mixtures/HMMs of exponential families
 - Bayesian linear regression
- Notice that any model containing conjugate pairs and multinomials has this property.
- Mean field variational inference is straightforward

- Compute the log of the conditional

$$\log p(z_j | z_{-j}, x) = \log h(z_j) + \eta(z_{-j}, x)^\top t(z_j) - a(\eta(z_{-j}, x)) \quad (30)$$

- Compute the expectation with respect to $q(z_{-j})$

$$\mathbb{E}[\log p(z_j | z_{-j}, x)] = \log h(z_j) + \mathbb{E}[\eta(z_{-j}, x)]^\top t(z_j) - \mathbb{E}[a(\eta(z_{-j}, x))] \quad (31)$$

- Noting that the last term does not depend on q_j , this means that

$$q^*(z_j) \propto h(z_j) \exp\{\mathbb{E}[\eta(z_{-j}, x)]^\top t(z_j)\} \quad (32)$$

and the normalizing constant is $a(\mathbb{E}[\eta(z_{-j}, x)])$.

- So, the optimal $q(z_j)$ is in the same exponential family as the conditional.
- Coordinate ascent algorithm
 - Give each hidden variable a variational parameter ν_j , and put each one in the same exponential family as its model conditional,

$$q(z_{1:m} | \nu) = \prod_{j=1}^m q(z_j | \nu_j) \quad (33)$$

- The coordinate ascent algorithm iteratively sets each natural variational parameter ν_j equal to the expectation of the natural conditional parameter for variable z_j given all the other variables and the observations,

$$\nu_j^* = \mathbb{E}[\eta(z_{-j}, x)]. \quad (34)$$

8 Example: Bayesian mixtures of Gaussians

- Let's go back to the Bayesian mixture of Gaussians. For simplicity, assume that the data generating variance is one.

- The latent variables are cluster assignments z_i and cluster means μ_k .

- The mean field family is

$$q(\mu_{1:K}, z_{1:n}) = \prod_{k=1}^K q(\mu_k | \tilde{\mu}_k, \tilde{\sigma}_k^2) \prod_{i=1}^n q(z_i | \phi_i), \quad (35)$$

where $(\tilde{\mu}_k, \tilde{\sigma}_k)$ are Gaussian parameters and ϕ_i are multinomial parameters (i.e., positive K -vectors that sum to one.)

- (Draw the graphical model and draw the graphical model with the mean-field family.)

- We compute the update for $q(z_i)$.

- Recall that

$$q^*(z_i) \propto \exp\{\mathbb{E}_{-i}[\log p(\mu_{1:K}, z_i, z_{-i}, x_{1:n})]\}. \quad (36)$$

- Because z_i is a multinomial, this has to be one too.

- The log joint distribution is

$$\begin{aligned} \log p(\mu_{1:K}, z_i, z_{-i}, x_{1:n}) = \\ \log p(\mu_{1:k}) + \left(\sum_{j \neq i} \log p(z_j) + \log p(x_j | z_j) \right) + \log p(z_i) + \log p(x_i | z_i). \end{aligned} \quad (37)$$

- Restricting to terms that are a function of z_i ,

$$q^*(z_i) \propto \exp\{\log \pi_{z_i} + \mathbb{E}[\log p(x_i | \mu_{z_i})]\}. \quad (38)$$

- Let's compute the expectation,

$$\mathbb{E}[\log p(x_i | \mu_i)] = -(1/2) \log 2\pi - x_i^2/2 + x_i \mathbb{E}[\mu_{z_i}] - \mathbb{E}[\mu_{z_i}^2]/2. \quad (39)$$

- We will see that $q(\mu_i)$ is Gaussian, so these expectations are easy to compute.
- Thus the coordinate update for $q(z_i)$ is

$$q^*(z_i = k) \propto \exp\{\log \pi_k + x_i E[\mu_k] - E[\mu_k^2]/2\}. \quad (40)$$

- Now we turn to the update for $q(\mu_k)$.

- Here, we are going to use our reasoning around the exponential family and conditional distributions.
- What is the conditional distribution of μ_k given $x_{1:n}$ and $z_{1:n}$?
- Intuitively, this is the posterior Gaussian mean with the data being the observations that were assigned (in $z_{1:n}$) to the k th cluster.
- Let's put the prior and posterior, which are Gaussians, in their canonical form. The parameters are

$$\hat{\lambda}_1 = \lambda_1 + \sum_{i=1}^n z_i^k x_i \quad (41)$$

$$\hat{\lambda}_2 = \lambda_2 + \sum_{i=1}^n z_i^k. \quad (42)$$

- Note that z_i^k is the indicator of whether the i th data point is assigned to the k th cluster. (This is because z_i is an indicator vector.)
- See how we sum the data in cluster k with $\sum_{i=1}^n z_i^k x_i$ and how $\sum_{i=1}^n z_i^k$ counts the number of data in cluster k .
- So, the optimal variational family is going to be a Gaussian with natural parameters

$$\tilde{\lambda}_1 = \lambda_1 + \sum_{i=1}^n E[z_i^k] x_i \quad (43)$$

$$\tilde{\lambda}_2 = \lambda_2 + \sum_{i=1}^n E[z_i^k] \quad (44)$$

- Finally, because z_i^k is an indicator, its expectation is its probability, i.e., $q(z_i = k)$.

- It's convenient to specify the Gaussian prior in its mean parameterization, and we need the expectations of the variational posterior for the updates on z_i .

- The mapping from natural parameters to mean parameters is

$$E[X] = \eta_1 / \eta_2 \quad (45)$$

$$\text{Var}(X) = 1 / \eta_2 \quad (46)$$

(Note: this is an alternative parameterization of the Gaussian, appropriate for the conjugate prior of the unit-variance likelihood. See the exponential family lecture.)

- So, the variational posterior mean and variance of the cluster component k is

$$E[\mu_k] = \frac{\lambda_1 + \sum_{i=1}^n E[z_i^k] x_i}{\lambda_2 + \sum_{i=1}^n E[z_i^k]} \quad (47)$$

$$\text{Var}(\mu_k) = 1 / (\lambda_2 + \sum_{i=1}^n E[z_i^k]) \quad (48)$$

- We'd rather specify a prior mean and variance.

- For the Gaussian conjugate prior, we map

$$\eta = \langle \mu/\sigma^2, 1/\sigma^2 \rangle. \quad (49)$$

- This gives the variational update in mean parameter form,

$$\mathbb{E}[\mu_k] = \frac{\mu_0/\sigma_0^2 + \sum_{i=1}^n \mathbb{E}[z_i^k]x_i}{1/\sigma_0^2 + \sum_{i=1}^n \mathbb{E}[z_i^k]} \quad (50)$$

$$\text{Var}(\mu_k) = 1/(1/\sigma_0^2 + \sum_{i=1}^n \mathbb{E}[z_i^k]). \quad (51)$$

These are the usual Bayesian updates with the data weighted by its variational probability of being assigned to cluster k .

- The ELBO is the sum of two terms,

$$\left(\sum_{k=1}^K \mathbb{E}[\log p(\mu_k)] + H(q(\mu_k)) \right) + \left(\sum_{i=1}^n \mathbb{E}[\log p(z_i)] + \mathbb{E}[\log p(x_i | z_i, \mu_{1:K})] + H(q(z_i)) \right).$$

- The expectations in these terms are the following.

- The expected log prior over mixture locations is

$$\mathbb{E}[\log p(\mu_k)] = -(1/2) \log 2\pi\sigma_0^2 - \mathbb{E}[\mu_k^2]/2\sigma_0^2 + \mathbb{E}[\mu_k]\mu_0/\sigma_0^2 - \mu_0^2/2\sigma_0^2, \quad (52)$$

where $\mathbb{E}[\mu_k] = \tilde{\mu}_k$ and $\mathbb{E}[\mu_k^2] = \tilde{\sigma}_k^2 + \tilde{\mu}_k^2$.

- The expected log prior over mixture assignments is not random,

$$\mathbb{E}[\log p(z_i)] = \log(1/K) \quad (53)$$

- The entropy of each variational location posterior is

$$H(q(\mu_k)) = (1/2) \log 2\pi\tilde{\sigma}_k^2 + 1/2. \quad (54)$$

If you haven't seen this, work it out at home by computing $-\mathbb{E}[\log q(\mu_k)]$.

- The entropy of each variational assignment posterior is

$$H(q(z_i)) = - \sum_{k=1}^K \phi_{ij} \log \phi_{ij} \quad (55)$$

- Now we can describe the coordinate ascent algorithm.

- We are given data $x_{1:n}$, hyperparameters μ_0 and σ_0^2 , and a number of groups K .

- The variational distributions are
 - * n variational multinomials $q(z_i)$
 - * K variational Gaussians $q(\mu_k | \tilde{\mu}_k, \tilde{\sigma}_k^2)$.
- Repeat until the ELBO converges:
 1. For each data point x_i
 - * Update the variational multinomial $q(z_i)$ from Equation 40.
 2. For each cluster $k = 1 \dots K$
 - * Update the mean and variance from Equation 50 and Equation 51.
- We can obtain a posterior decomposition of the data.
 - Points are assigned to $\arg \max_k \phi_{i,k}$.
 - Cluster means are estimated as $\tilde{\mu}_k$.
- We can approximate the predictive distribution with a mixture of Gaussians, each at the expected cluster mean. This is

$$p(x_{\text{new}} | x_{1:n}) \approx \frac{1}{K} \sum_{k=1}^K p(x_{\text{new}} | \tilde{\mu}_k), \quad (56)$$

where $p(x | \tilde{\mu}_k)$ is a Gaussian with mean $\tilde{\mu}_k$ and unit variance.

9 Multivariate mixtures of Gaussians

- We adjust the algorithm (slightly) when the data are multivariate. Assume the observations $x_{1:n}$ are p -dimensional and, thus, so are the mixture locations $\mu_{1:K}$.
- The multinomial update on Z_i is

$$q^*(z_i = k) \propto \exp\{\log \pi_k + x_i \mathbb{E}[\mu_k] - \mathbb{E}[\mu_k^\top \mu_k]/2\}. \quad (57)$$

- The expected log prior over mixture locations is

$$\mathbb{E}[\log p(\mu_k)] = -(p/2) \log 2\pi\sigma_0^2 - \mathbb{E}[\mu_k^\top \mu_k]/2\sigma_0^2 + \mathbb{E}[\mu_k]^\top \mu_0/\sigma_0^2 - \mu_0^\top \mu_0/2\sigma_0^2, \quad (58)$$

where $\mathbb{E}[\mu_k] = \tilde{\mu}_k$ and $\mathbb{E}[\mu_k^\top \mu_k] = p\tilde{\sigma}_k^2 + \tilde{\mu}_k^\top \tilde{\mu}_k$.

- The entropy of the Gaussian is

$$\mathbb{H}(q(\mu_k)) = (p/2) \log 2\pi\tilde{\sigma}_k^2 + p/2. \quad (59)$$