STA2311: Advanced Computational Methods for Statistics I

Class 5: Variational Inference

Radu Craiu   Robert Zimmerman

University of Toronto

October 10, 2023

Section 1

Introduction

Variational Inference

- Variational inference provides a way to approximate complicated distributions by simpler ones (usually for the purposes of sampling)
  - Especially posterior distributions...
- For a given distribution of interest, the approximating distribution is chosen as the optimal one among a class of simpler ones
  - The meaning of "optimal" here will be discussed!
- Because one can then generate samples from the simpler distribution, variational inference is a popular alternative to MCMC, which we will learn about later in the course
- The topic gets its name from *variational calculus* (or the *calculus of variations*), with deals with optimizing functionals
- We mainly follow Bishop [2006] and Blei et al. [2017]

Optimizing Functionals

- A *functional* $S[\cdot]$ is a mapping from a function space $\mathcal{F}$ to a scalar field ($\mathbb{R}$, for our purposes)
- For example, the *differential entropy* $H[\cdot]$ can be viewed as a functional on the space of density functions, given by

$$H[f] = -\int \log(f(x)) \cdot f(x)\,\mathrm{d}x$$

- Since $S[f] \in \mathbb{R}$, in principle there usually exists at least one $f^* \in \mathcal{F}$ such that $S[f^*] \geq S[f]$ for all $f \in \mathcal{F}$
    - For example, among densities supported on (a, b), the Unif(a, b) density $f(x) = \frac{\mathbb{1}_{a<x<b}}{b-a}$ maximizes the differential entropy
- Techniques for determining such an f^* are the topic of variational calculus; these are broadly analogous to function optimization methods from basic calculus, but we will not go into details

Section 2

The Ingredients

Data and Latent Variables

- Let $\boldsymbol{X} = X_{1:m}$ represent our data and $\boldsymbol{Z} = Z_{1:m}$ represent auxiliary/latent variables (which may be parameters in the Bayesian setup)
- $\boldsymbol{x}$ and $\boldsymbol{z}$ are their observed counterparts
- Then the joint distribution of $(\boldsymbol{Z}, \boldsymbol{X})$ factorizes: $p(\boldsymbol{z}, \boldsymbol{x}) = p(\boldsymbol{z}) \cdot p(\boldsymbol{x} \mid \boldsymbol{z})$ so that the conditional distribution of $\boldsymbol{Z} \mid \boldsymbol{x}$ is

$$p(\boldsymbol{z} \mid \boldsymbol{x}) = \frac{p(\boldsymbol{z}) \cdot p(\boldsymbol{x} \mid \boldsymbol{z})}{\int p(\boldsymbol{z}) \cdot p(\boldsymbol{x} \mid \boldsymbol{z}) \,\mathrm{d}\boldsymbol{z}} \tag{1}$$

- We're interested in approximating $p(\boldsymbol{z} \mid \boldsymbol{x})$

The KL Divergence

- The *Kullback-Leibler (KL) divergence* is a measure of "distance" between distributions
- For mass functions p and q defined on a sample space $\mathcal{X}$, it is given by

$$\mathrm{KL}(p \,||\, q) = \sum_{x \in \mathcal{X}} p(x) \cdot \log\left(\frac{p(x)}{q(x)}\right)$$

- For density functions p and q defined on $\mathcal{X}$, it is given by

$$\mathrm{KL}(p \,||\, q) = \int_{\mathcal{X}} p(x) \cdot \log\left(\frac{p(x)}{q(x)}\right) \mathrm{d}x$$

- One can show that $\mathrm{KL}(p \,||\, q) \geq 0$ for any distributions p, q, with equality if and only if $p = q$
  - However, it is not a metric on the space of distributions on $\mathcal{X}$

Information Theory

- The KL divergence emerged from the field of information theory
- In statistics, p typically describes our observed data, and q represents a distribution which is hypothesized to have generated that data

  The KL divergence is then interpreted as the average difference of the number of bits required for encoding samples of p using a code optimized for q rather than one optimized for p.

- The KL divergence shows up in many areas within statistics

Towards the ELBO

- First, we consider a family $\mathcal{Q}$ of approximate distributions of $\boldsymbol{Z}$
- Then, we find the member $q^* \in \mathcal{Q}$ that best approximates $p(\boldsymbol{Z} \mid \boldsymbol{X})$
- The "best" is defined in terms of the KL divergence:

$$q^*(\boldsymbol{z}) = \underset{q \in \mathcal{Q}}{\operatorname{argmin}} \, \mathrm{KL}\left(q(\cdot) \,||\, p(\cdot \mid \boldsymbol{x})\right) = \underset{q \in \mathcal{Q}}{\operatorname{argmin}} \int \log\left(\frac{q(\boldsymbol{z})}{p(\boldsymbol{z} \mid \boldsymbol{x})}\right) q(\boldsymbol{z}) \, \mathrm{d}\boldsymbol{z}$$

- We can recast this optimization problem more conveniently in terms of the evidence

The Evidence

- Another way to write (1) is

$$p(\boldsymbol{z} \mid \boldsymbol{x}) = \frac{p(\boldsymbol{z}, \boldsymbol{x})}{p(\boldsymbol{x})}$$

- Here $p(\boldsymbol{x}) = \int p(\boldsymbol{z}, \boldsymbol{x})\, \mathrm{d}\boldsymbol{z}$ is called the *evidence*, and is usually intractable

- Observe that for any q,

$$\begin{aligned} \mathrm{KL}\left(q(\cdot) \mid\mid p(\cdot \mid \boldsymbol{z})\right) &= \mathbb{E}_q[\log(q(\boldsymbol{Z}))] - \mathbb{E}_q[\log(p(\boldsymbol{Z} \mid \boldsymbol{x}))] \\ &= \mathbb{E}_q[\log(q(\boldsymbol{Z}))] - \mathbb{E}_q[\log(p(\boldsymbol{Z}, \boldsymbol{x}))] + \mathbb{E}_q[\log(p(\boldsymbol{x}))] \end{aligned}$$

- Since the rightmost term is constant in $\boldsymbol{Z}$, minimizing $\mathrm{KL}\left(q(\cdot) \mid\mid p(\cdot \mid \boldsymbol{x})\right)$ is equivalent to maximizing

$$\mathrm{ELBO}(q) := \mathbb{E}_q[\log(p(\boldsymbol{Z}, \boldsymbol{x}))] - \mathbb{E}_q[\log(q(\boldsymbol{Z}))]$$

The ELBO

- The quantity $\mathrm{ELBO}(q)$ is called the *evidence lower bound (ELBO)*
- The name comes from the fact that

$$\log(p(\boldsymbol{x})) = \mathrm{KL}\left(q(\cdot) \,||\, p(\cdot \mid \boldsymbol{x})\right) + \mathrm{ELBO}(q) \geq \mathrm{ELBO}(q),$$

  because the KL divergence is non-negative
- So the ELBO provides a lower bound on the (log) evidence
- Moreover, equality holds if and only if $q(\boldsymbol{z}) = p(\boldsymbol{z} \mid \boldsymbol{x})$
- But usually $p(\cdot \mid \boldsymbol{x}) \notin \mathcal{Q}$.

Section 3

Mean-Field Variational Inference

Choosing the Variational Family

- There are usually several choices of variational family to choose from
- We want the family to be rich enough to provide a reasonably good approximation to our target, but simple enough that its members satisfy the requirement of being easy to work with
- If the family contains the target itself, then the problem is trivial
- One choice is the set of densities from a given parametric family (such as Gaussian distributions)
  - Then the optimization problem reduces to finding the optimal parameters μ and σ^2, which is "easy"
- However, for complicated target distributions, it is preferable to optimize over a more flexible class

Choosing the Variational Family (Continued)

- The *mean-field* variational family is one in which the latent variables are independent
- That is, each has its own factor in the variational distribution: $q(\boldsymbol{z}) = \prod_{j=1}^{m} q_j(z_j)$
- Usually the posterior is not in the mean-field variational family because of dependencies between components of $\boldsymbol{Z}$
- However, this family allows us to use the *coordinate ascent* algorithm to find the optimal q
- We will discuss some extensions later

Deriving the Coordinate Ascent Algorithm

- For any j, let $\boldsymbol{Z}_{-j} = (Z_1, \ldots, Z_{j-1}, Z_{j+1}, \ldots, Z_m)$ and $q_{-j} = \prod_{i \neq j}^{m} q_i$
- Under the mean-field assumption, the ELBO depends on q_j through

$$\text{ELBO}\,(q_j) = \int q_j(Z_j) \log(\tilde{p}(X, Z_j))\,\mathrm{d}Z_j - \int \log(q_j(Z_j)) q_j(Z_j)\,\mathrm{d}Z_j + const$$

where

$$\log(\tilde{p}(X, Z_j)) = \mathbb{E}_{\boldsymbol{Z}_{-j}}[\log(p(X, \boldsymbol{Z}))]$$

- Note that the ELBO (q_j) is just the negative KL divergence between q_j and $\tilde{p}(X, Z_j)$ so we know it is minimized when $q_j = \tilde{p}(X, Z_j)$

The Optimal Solution

- This implies that the optimal q_j satisfies

$$\log(q_j(z_j)) = \mathbb{E}_{q_{-j}}[\log(p(z_j, \boldsymbol{Z}_{-j}, \boldsymbol{x}))] + c_j, \quad 1 \leq j \leq m, \tag{2}$$

  for an appropriate constant c_j (used for normalization)
- This is optimal, but not quite explicit because the expectation involved is taken with respect to q_{-j}, which is a product of the other mean-field factors
- This suggests an iterative algorithm in which we first initialize $q_1, \ldots, q_m$, and then repeatedly update them one at a time using (2)

The Algorithm

- Given data $\boldsymbol{x}$ and a joint distribution $p(\boldsymbol{z}, \boldsymbol{x})$, the *mean-field variational inference* algorithm is

  1. Initialize $q_j^{(0)}(z_j)$ for $1 \leq j \leq m$
  2. For $t \geq 0$:
     - for $1 \leq j \leq m$, compute

$$q_j^{(t+1)}(z_j) \propto \exp\Big(\mathbb{E}_{q_{-j}^{(t)}}[\log(p(z_j, \boldsymbol{Z}_{-j}, \boldsymbol{x}))]\Big),$$

where $q_{-j}^{(t)} = \prod_{i=1}^{j-1} q_i^{(t+1)} \cdot \prod_{i=j+1}^{m} q_i^{(t)}$, with edge cases are treated in the obvious manner

- It can be shown that this algorithm is guaranteed to converge

Caveats

- In order to use the algorithm, we need to evaluate $\exp\Big(\mathbb{E}_{q_{-j}}[\log(p(z_j, \boldsymbol{Z}_{-j}, \boldsymbol{x}))]\Big)$ and the normalizing constant

$$\int \exp\Big(\mathbb{E}_{q_{-j}}[\log(p(z_j, \boldsymbol{Z}_{-j}, \boldsymbol{x}))]\Big)\,\mathrm{d}z_j$$

- These can be extremely challenging to compute for all but the simplest toy models
- There is no guarantee that the expectation and/or the normalizing constant exists in closed form
  - Especially in Bayesian models

A Toy Example

- To get a feel for how the algorithm works, consider finding a mean-field approximation to a bivariate normal distribution:

$$p(\boldsymbol{z} \mid \boldsymbol{x}) = p(\boldsymbol{z}) = \frac{1}{\sqrt{2\pi|\boldsymbol{\Sigma}|}}\exp\Big(-(\boldsymbol{z}-\boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{z}-\boldsymbol{\mu})/2\Big), \quad \boldsymbol{z} \in \mathbb{R}^2$$

- This target involves no “data” $\boldsymbol{x}$, but that’s okay
- The parameters in $p(\boldsymbol{z})$ are the mean $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$, but it easier to work in terms of the precision matrix $\boldsymbol{\Lambda} := \boldsymbol{\Sigma}^{-1}$ and transform back later

A Toy Example (Continued)

- The first step is to compute

$$
\begin{aligned}
q_1(z_1) &\propto \exp(\mathbb{E}_{q_2}[\log(p(z_1, Z_2))]) \\
&= \exp\left(\mathbb{E}_{q_2}\left[-\frac{1}{2}(z_1 - \mu_1)^2\Lambda_{11} - (z_1 - \mu_1)\Lambda_{12}(Z_2 - \mu_2)\right]\right) \\
&= \exp\left(-\frac{1}{2}z_1^2\Lambda_{11} + z_1(\mu_1\Lambda_{11} - \Lambda_{12}(\mathbb{E}_{q_2}[Z_2] - \mu_2))\right)
\end{aligned}
$$

- This is the kernel of a normal distribution!

- Working out the mean and variance (e.g., by completing the square) gives $q_1(z_1) = \phi(z_1 \mid m_1, \Lambda_{11}^{-1})$ where

$$
m_1 = \mu_1 - \frac{\Lambda_{12}}{\Lambda_{11}}(\mathbb{E}_{q_2}[Z_2] - \mu_2) \tag{3}
$$

    - Here $\phi(z \mid \mu, \sigma^2)$ is the $\mathcal{N}(\mu, \sigma^2)$ pdf

A Toy Example (Continued)

- A similar calculation (or a symmetry argument) yields $q_2(z_2) = \phi(z_2 \mid m_2, \Lambda_{22}^{-1})$ where

$$m_2 = \mu_2 - \frac{\Lambda_{12}}{\Lambda_{22}}(\mathbb{E}_{q_1}[Z_1] - \mu_1) \tag{4}$$

- In fact, since $\mathbb{E}_{q_1}[Z_1] = m_1$ and $\mathbb{E}_{q_2}[Z_2] = m_2$, we can plug these into (3) and (4) to get a linear system which is easy to solve
- That is, the optimal mean field approximation here has an explicit solution
- Since this is rarely the case, we will practice solving the system iteratively instead

A Toy Example (Continued)

```r
norm <- function(x) {sqrt(sum(x^2))}

mu <- c(-3, 3)
Sigma <- matrix(c(1,0.5,0.5,3), nrow=2, ncol=2, byrow=T)
Lambda <- solve(Sigma)

m1.old <- NaN; m2.old <- NaN
m1 <- 0; m2 <- 0

pars.old <- c(m1.old, m2.old)
pars <- c(m1, m2)

while(is.nan(m1.old) || norm(pars.old - pars) > 10e-6) {
  m1.old <- m1
  m2.old <- m2
  pars.old <- c(m1.old, m2.old)

  m1 <- mu[1] - Lambda[1,1]^(-1)*Lambda[1,2]*(m2.old - mu[2])
  m2 <- mu[2] - Lambda[2,2]^(-1)*Lambda[2,1]*(m1.old - mu[1])
  pars <- c(m1, m2)
}
```

Section 4

Local Methods

The Local Approach

- The mean-field approach seeks an optimal approximation to the entire posterior $p(\boldsymbol{z} \mid \boldsymbol{x})$
- Instead, we might settle on optimizing the distribution of a certain component z_i or a group of components $\boldsymbol{z}'$ within the full model
- In the context of variational inference, "optimizing" means "getting as close to the ELBO as possible"
- Combining such bounds then provides a bound on the target $p(\boldsymbol{z} \mid \boldsymbol{x})$ that is still easier to work with
- Bishop [2006] calls these approaches *local variational methods*

Variational Parameters

- The idea is to introduce a free parameter ξ into the function we wish to optimize, and then select — perhaps iteratively — the ξ that brings us as close to optimality as possible
  - We call ξ a *variational parameter*
- For example, to obtain a linear lower bound on the function $f(x) = e^{-x}$, we can take a first-order Taylor expansion around any ξ to get
$$f(\xi) + f'(\xi) \cdot (x - \xi) = e^{-\xi} - e^{-\xi} \cdot (x - \xi)$$
- To keep track of the variational parameter, we denote the linear function above as $y(x, \xi)$
- Then $y(x', \xi) \leq f(x')$ for all x', and the bound is optimal (i.e., as tight as possible) when $\xi = x'$
- In fact $f(x) = \sup_\xi y(x, \xi)$

Example: Bayesian Logistic Regression

- Consider logistic regression: we have independent observations $Y_1, \ldots, Y_n$ and covariates $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n \in \mathbb{R}^p$ with $Y_i \mid \boldsymbol{x}_i \sim \text{Bernoulli}(\sigma(\beta^\top \boldsymbol{x}_i))$, where $\sigma(x) = (1 + e^{-x})^{-1}$
- We adopt a Bayesian model and impose a $\mathcal{N}_p(\boldsymbol{m}_0, \boldsymbol{S}_0)$ prior on β
  - This is a canonical prior for Bayesian logistic regression
- We seek a local variational approximation to the posterior $p(\beta \mid \boldsymbol{y})$ by finding a lower bound on the evidence, and then maximizing it

Example: Bayesian Logistic Regression (Continued)

- Our prior is $p(\beta) \propto \exp\left(-\frac{1}{2}(\beta - \boldsymbol{m}_0)^\top \boldsymbol{S}_0^{-1}(\beta - \boldsymbol{m}_0)\right)$
- The likelihood for a single observation is

$$p(y_i \mid \beta) = \sigma(\beta^\top \boldsymbol{x}_i)^{y_i} \cdot (1 - \sigma(\beta^\top \boldsymbol{x}_i))^{1-y_i} = \cdots = e^{\beta^\top \boldsymbol{x}_i y_i} \cdot \sigma(-\beta^\top \boldsymbol{x}_i)$$

- The evidence is therefore given by

$$\begin{aligned} p(\beta) &= \int p(\beta) \cdot p(\boldsymbol{y} \mid \beta) \, \mathrm{d}\beta \\ &= \int p(\beta) \cdot \left(\prod_{i=1}^{n} p(y_i \mid \beta) \right) \mathrm{d}\beta \end{aligned}$$

- The plan is to lower bound the integrand by the kernel of a distribution that's easy to work with

Example: Bayesian Logistic Regression (Continued)

- To do this, we use a lower bound on the expit function $\sigma(x)$:

$$\sigma(x) \geq \sigma(\xi) \cdot \exp\left(\frac{(x-\xi)}{2} - \lambda(\xi) \cdot (x^2 - \xi^2)\right), \quad x \in (-\xi, \xi)$$

where $\lambda(\xi) = \frac{1}{2\xi}(\sigma(\xi) - \frac{1}{2})$

  - This bound is derived using some mild convex analysis (see p.495 of Bishop [2006] for details)

- We allow each $p(y_i \mid \beta)$ to get its own variational parameter ξ_i
- Thus

$$p(y_i \mid \beta) \geq \sigma(\xi_i) \cdot \exp\left(\frac{(-\boldsymbol{\beta}^\top \boldsymbol{x}_i - \xi_i)}{2} - \lambda(\xi_i) \cdot ([-\boldsymbol{\beta}^\top \boldsymbol{x}_i]^2 - \xi_i^2)\right)$$

Example: Bayesian Logistic Regression (Continued)

- This gives us

$$p(\beta \mid \boldsymbol{y}) \geq \exp\Bigg(-\frac{1}{2}(\beta - \boldsymbol{m}_0)^\top \boldsymbol{S}_0^{-1}(\beta - \boldsymbol{m}_0) + \sum_{i=1}^{n}\Big(\beta^\top \boldsymbol{x}_i(y_i - 1/2) - \lambda(\xi_n) \cdot \beta^\top \boldsymbol{x}_i \boldsymbol{x}_i^\top \beta\Big) + c\Bigg) \tag{5}$$

  where $c = \sum_{i=1}^{n}\Big(\log\Big(\sigma(\beta^\top \boldsymbol{x}_i) - \lambda(\xi_i) \cdot \xi_i^2\Big)\Big)$ is constant with respect to β

- The RHS is the kernel of a normal distribution with covariance matrix

$$\boldsymbol{S}_n = \left(\boldsymbol{S}_0^{-1} + 2\sum_{i=1}^{n}\lambda(\xi_i) \cdot \boldsymbol{x}_i \boldsymbol{x}_i^\top\right)^{-1}$$

  and mean

$$\boldsymbol{m}_n = \boldsymbol{S}_n\left(\boldsymbol{S}_0^{-1}\boldsymbol{m}_0 + \sum_{i=1}^{n}(y_i - 1/2)\boldsymbol{x}_i\right)$$

Example: Bayesian Logistic Regression (Continued)

- So we have a family of normal approximations to the posterior: one for each $\boldsymbol{\xi} = (\xi_1, \ldots, \xi_n)$
- The next step is to determine the optimal $\boldsymbol{\xi}$
- To do this, we let

$$\mathcal{L}(\boldsymbol{\xi}) = \log\left(\int h(\boldsymbol{\beta}, \boldsymbol{\xi}) \, \mathrm{d}\boldsymbol{\beta} \right)$$

  where $h(\boldsymbol{\beta}, \boldsymbol{\xi})$ is the RHS of (5)
- We have that $\log(p(\boldsymbol{y})) \geq \mathcal{L}(\boldsymbol{\xi})$ for any $\boldsymbol{\xi}$

Example: Bayesian Logistic Regression (Continued)

- Since $h(\beta, \boldsymbol{\xi})$ involves the exponential of a quadratic form in β, $\int h(\beta, \boldsymbol{\xi})\,\mathrm{d}\beta$ can be evaluated in closed form, which gives

$$\mathcal{L}(\boldsymbol{\xi}) = \frac{1}{2}\left(\log(|\boldsymbol{S}_n|) + \boldsymbol{m}_n^\top \boldsymbol{S}_n^{-1}\boldsymbol{m}_n\right) + \sum_{i=1}^{n}\left(\log(\sigma(\xi_i)) - \xi_i/2 + \lambda(\xi_i)\cdot\xi_i^2\right) + c'$$

where $c' = -\frac{1}{2}\left(\log(|\boldsymbol{S}_0|) + \boldsymbol{m}_0^\top \boldsymbol{S}_0^{-1}\boldsymbol{m}_0\right)$

- Differentiating with respect to ξ_i and doing the (tedious) algebra yields the optimal values

$$\xi_i = \sqrt{\boldsymbol{x}_i^\top(\boldsymbol{S}_n + \boldsymbol{m}_n\boldsymbol{m}_n^\top)\boldsymbol{x}_i}$$

- This can also be derived by viewing β as a latent variable in $\log(\int h(\beta, \boldsymbol{\xi})\,\mathrm{d}\beta)$ and working out an EM algorithm
  - See p.501 of Bishop [2006] for details

Example: Bayesian Logistic Regression (Continued)

```r
set.seed(2311)
expit <- function(x) {1/(1+exp(-x))}
logit <- function(p) {log(p/(1-p))}
norm <- function(x) {sqrt(sum(x^2))}

n <- 1000

X1 <- rnorm(n=n)
X2 <- rbinom(n=n, size=1, prob=0.2)
X3 <- rpois(n=n, lambda=0.7)
X <- cbind(1, X1, X2, X3)

y <- rbinom(n=n, size=1, prob=expit(0.4 + 0.7*X1 + 3*X2 - X3))

S0 <- (1/4)*diag(4)
m0 <- rep(0, times=4)
```

Example: Bayesian Logistic Regression (Continued)

```
xi <- rep(1, times=n)
xi.old <- rep(10, times=n)

lambda <- function(xi) {(1/(2*xi))*(expit(xi) - 1/2)}

Sn <- S0
mn <- m0

while (norm(xi - xi.old) > 10e-6) {
  xi.old <- xi
  xi <- sqrt(apply(X, 1, function(x) t(x)%*%(Sn + mn%*%t(mn))%*%x))

  Sn <- solve( solve(S0) + 2*Reduce('+', lapply(1:n,
            function(j) {lambda(xi.old[j])*X[j,]%*%t(X[j,])})) )
  mn <- Sn %*% ( solve(S0)%*%m0 + colSums((y-1/2)*X) )
}
```

Section 5

Connections

Connection to EM

- Suppose we move back to the frequentist realm
- $\boldsymbol{X}$ is our data, and $\boldsymbol{Z}$ is a set of latent variables, and now $\boldsymbol{\theta}$ is a parameter in a parametric model for $\boldsymbol{X}$ that we seek to estimate
- In Class 3, we learned how the EM algorithm increases the likelihood in $\boldsymbol{\theta}$
- In fact, we can view the EM algorithm as a special case of variational inference
- Write the ELBO as

$$\text{ELBO}(q, \boldsymbol{\theta}) = \mathbb{E}_q[\log(p(\boldsymbol{Z}, \boldsymbol{X}; \boldsymbol{\theta}))] - \mathbb{E}_q[\log(q(\boldsymbol{Z}))] \tag{6}$$

The E-Step

- Recall that in the E-step of the EM algorithm, we compute $Q(\boldsymbol{\theta} \mid \boldsymbol{\theta}^{(t)})$, the expected complete-data log-likelihood $\mathbb{E}[\log(p(\boldsymbol{Z}, \boldsymbol{X}; \boldsymbol{\theta}))]$ where $\boldsymbol{Z} \sim p(\cdot \mid \boldsymbol{X}, \boldsymbol{\theta}^{(t)})$ and $\boldsymbol{\theta}^{(t)}$ is our current parameter estimate
- But we know that the ELBO (6) is maximized when $q(\boldsymbol{Z}) = p(\cdot \mid \boldsymbol{X}, \boldsymbol{\theta}^{(t)})$
- So computing $Q(\boldsymbol{\theta} \mid \boldsymbol{\theta}^{(t)})$ is the same as computing $\text{ELBO}(q^{(t)}, \boldsymbol{\theta})$, where $q^{(t)} = \underset{q}{\operatorname{argmax}}\, \text{ELBO}(q, \boldsymbol{\theta})$

The M-Step

- In the M-step of the EM algorithm, we choose $\boldsymbol{\theta}^{(t+1)}$ by maximizing $Q(\boldsymbol{\theta} \mid \boldsymbol{\theta}^{(t)})$ with respect to $\boldsymbol{\theta}$
- From the previous slide, we see that this is the same as $\boldsymbol{\theta}^{(t+1)} = \underset{\boldsymbol{\theta}}{\operatorname{argmax}} \, \mathsf{ELBO}(q^{(t)}, \boldsymbol{\theta})$
- Alternatively, note that maximizing $Q(\boldsymbol{\theta} \mid \boldsymbol{\theta}^{(t)})$ means setting $\boldsymbol{\theta}^{(t+1)} = \underset{\boldsymbol{\theta}}{\operatorname{argmax}} \, \mathbb{E}[\log(p(\boldsymbol{Z}, \boldsymbol{X}; \boldsymbol{\theta}))]$ where again $\boldsymbol{Z} \sim p(\cdot \mid \boldsymbol{X}, \boldsymbol{\theta}^{(t)})$
- And

$$
\begin{aligned}
\boldsymbol{\theta}^{(t+1)} &= \underset{\boldsymbol{\theta}}{\operatorname{argmax}} \, \left(\mathbb{E}[\log(p(\boldsymbol{Z}, \boldsymbol{X}; \boldsymbol{\theta}))] - \mathbb{E}[\log\Big(p(\boldsymbol{Z} \mid \boldsymbol{X}, \boldsymbol{\theta}^{(t)})\Big)]\right) \\
&= \underset{\boldsymbol{\theta}}{\operatorname{argmax}} \, \mathsf{ELBO}(q^{(t)}, \boldsymbol{\theta})
\end{aligned}
$$

References I

Christopher M. Bishop. *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Springer-Verlag, Berlin, Heidelberg, 2006. ISBN 0387310738.

David M Blei, Alp Kucukelbir, and Jon D McAuliffe. Variational inference: A review for statisticians. *Journal of the American statistical Association*, 112(518):859–877, 2017.