Bi/BE/CS 183 2021-2022
Instructor: Lior Pachter
TAs: Tara Chari, Zitong (Jerry) Wang, Meichen Fang

Problem Set 4 Solution

Problem 1

Given observed data X, unobserved/latent variable Z and model parameter θ, the Expectation-maximization (EM) algorithm attempts to maximize the following Q-function

$$\begin{aligned} Q(\theta \mid \theta_t) &= E_{Z|X,\theta_t}[\log p(X, Z; \theta)] \\ &= \int p(Z|X; \theta_t) \log p(X, Z; \theta). \end{aligned} \tag{1}$$

We will show that improving $Q(\theta|\theta_t)$ will indirectly improve the log-likelihood $\log p(X|\theta)$, and that $\log p(X|\theta)$ is guaranteed to be non-decreasing with each step of the algorithm.

(a) Show that the log-likelihood of the observed data $\log p(X; \theta)$ can be expressed in term of the Q-function,

$$\log p(X|\theta) = Q(\theta|\theta_t) - R(\theta|\theta_t),$$

where $R(\theta|\theta_t) = E_{Z|X,\theta_t}[\log p(Z|X; \theta)]$

A: Note that $p(X|\theta)p(Z|X; \theta) = p(X, Z|\theta)$.

Thus, the log-likelihood of the observed data $\log p(X|\theta)$ can be expressed as follows:

$$\begin{aligned} \log p(X|\theta) &= \int p(Z|X; \theta_t) \log p(X|\theta) dZ \\ &= \int p(Z|X; \theta_t) \log \frac{p(X, Z|\theta)}{p(Z|X; \theta)} dZ \\ &= \int p(Z|X; \theta_t) \log p(X, Z|\theta) dZ - \int p(Z|X; \theta_t) \log p(Z|X; \theta) dZ \\ &= Q(\theta|\theta_t) - R(\theta|\theta_t) \end{aligned}$$

(b) Show that for any θ, the following inequality holds,

$$R(\theta|\theta_t) \leq R(\theta_t|\theta_t)$$

Hint: use the fact about the logarithm of a positive number x: $log(x) \leq x-1$ with equality if and only if x = 1.

A: We will show that $R(\theta_t|\theta_t) - R(\theta|\theta_t)$ is always non-negative. In fact, it is the Kullback-Leibler distance between these two probability distribution $p(Z|X; \theta_t)$ and $p(Z|X; \theta)$

$$\begin{aligned} R(\theta_t|\theta_t) - R(\theta|\theta_t) &= \int -p(Z|X; \theta_t) \log \frac{p(Z|X; \theta)}{p(Z|X; \theta_t)} dZ \\ &\geq \int p(Z|X; \theta_t)(1 - \frac{p(Z|X; \theta)}{p(Z|X; \theta_t)}) dZ \\ &= 1 - \int p(Z|X; \theta) dZ \\ &= 1 - 1 = 0 \end{aligned}$$

(c) Denoting the estimated parameter at two consecutive step of the EM algorithm as θ_t and θ_{t+1}, show that $\log p(X|\theta_{t+1}) - \log p(X|\theta_t) \geq Q(\theta_{t+1}|\theta_t) - Q(\theta_t|\theta_t)$. This shows that by improving the Q-function, the EM algorithm causes $\log p(X;\theta)$ to improve by at least as much.

A: In the E-step, we calculate Q-function based on $p(Z|X;\theta_t)$. Then, in the M-step, we optimize Q-function over parameters and obtain new parameters θ_{t+1}. Note that $p(Z|X;\theta_t)$ doesn't change in the M-step.

Therefore,

$$\begin{aligned}\log p(X|\theta_{t+1}) - \log p(X|\theta_t) &= (Q(\theta_{t+1}|\theta_t) - R(\theta_{t+1}|\theta_t)) - (Q(\theta_t|\theta_t) - R(\theta_t|\theta_t)) \\ &= Q(\theta_{t+1}|\theta_t) - Q(\theta_t|\theta_t) + R(\theta_t|\theta_t) - R(\theta_{t+1}|\theta_t)\end{aligned}$$

From part (b) we know that $R(\theta_t|\theta_t) - R(\theta_{t+1}|\theta_t) \geq 0$. Therefore, $\log p(X|\theta_{t+1}) - \log p(X|\theta_t) \geq Q(\theta_{t+1}|\theta_t) - Q(\theta_t|\theta_t)$.

(d) Show that for any step t of the EM algorithm, the following holds

$$\log p(X|\theta_{t+1}) \geq \log p(X|\theta_t).$$

In other words, the log-likelihood is non-decreasing with every step of the EM algorithm.

A: Since in the M-step we optimize Q-function over parameters, $Q(\theta_{t+1}|\theta_t) - Q(\theta_t|\theta_t) \geq 0$. Therefore, $\log p(X|\theta_{t+1}) \geq \log p(X|\theta_t)$ follows from part (c).

Problem 2

The EM procedure is often used to estimate relative abundances of transcripts as sequencing reads may align to multiple transcript sequences. To resolve these ambiguities the EM algorithm can be applied as shown below (Fig. 1). Given N reads and K transcripts, the reads are aligned to the transcript sequences to determine which transcript(s) they originate from. We can summarize this alignment with a matrix $\mathbf{Y}$ defined by $y_{k,n} = 1$ if read n aligns to transcript k and 0 otherwise. $\mathbf{Y}$ is denoted as the *compatibility matrix*. For Fig. 1, with 5 reads and 3 transcripts, the compatibility matrix is

$$\begin{array}{c} \\ \texttt{red} \\ \texttt{green} \\ \texttt{blue} \end{array} \begin{array}{c} \begin{array}{ccccc} a & b & c & d & e \end{array} \\ \begin{pmatrix} 1 & 0 & 1 & 1 & 1 \\ 1 & 1 & 0 & 0 & 1 \\ 1 & 1 & 1 & 0 & 0 \end{pmatrix} \end{array}.$$

For example $y[red, b] = 0$ as read b (denoted by a black bar in Fig. 1) does not overlap with any part of transcript red (read bars in Fig. 1). From these reads and their overlapping alignments, we want to estimate the relative abundances, denoted as α, for each transcript, $(\alpha_1, ...\alpha_k)$.

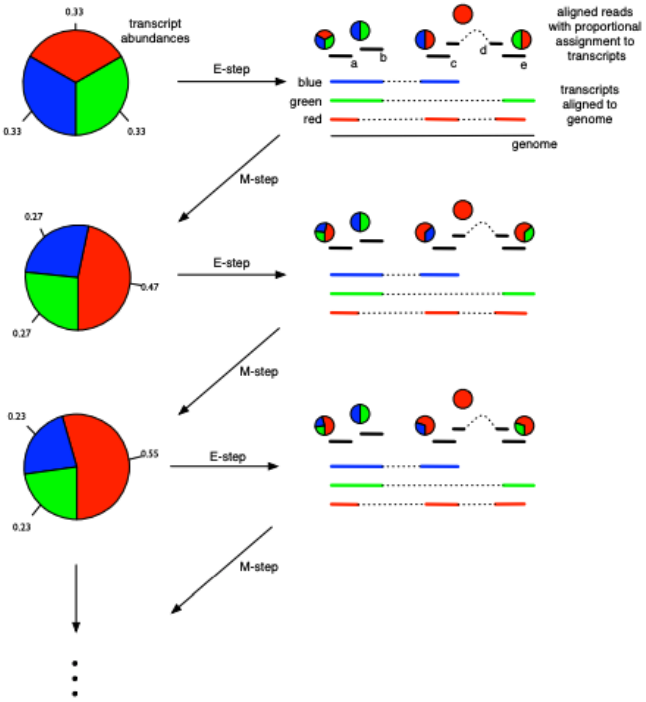

Figure 1: Example of EM algorithm steps for 3 transcripts and 5 reads. From [1].

For Fig. 1 , we are attempting to estimate the parameters $\alpha_{red}, \alpha_{green}, \alpha_{red}$, and given $\mathbf{Y}$ the likelihood to maximize is then

$$\begin{aligned} \mathcal{L}(\alpha) &= \prod_{n=1}^{N} (\sum_{k=1}^{K} y_{k,n}\alpha_k) \\ &= (\alpha_{red} + \alpha_{blue})(\alpha_{red} + \alpha_{green})(\alpha_{blue} + \alpha_{green})\alpha_{red}, \end{aligned}$$

subject to $\alpha_{red} + \alpha_{blue} + \alpha_{green} = 1$.

In the EM algorithm, we iteratively approximate the read counts given some α estimate (Fig. 1, E step), then we recalculate the α parameters given the (proportional) read counts (Fig. 1, M step). Denote the α in step t by $\alpha^{(t)}$. We will let the latent variable Z denote whether a read comes from transcript k.

In the E-step, we first calculate the posterior distribution:

$$\begin{aligned} p(Z_n = k|Y_n; \alpha^{(t)}) &= \frac{p(Z_n = k, Y_n|\alpha^{(t)})}{\sum_k p(Z_n = k, Y_n|\alpha^{(t)})} \\ &= \frac{y_{k,n}\alpha_k^{(t)}}{\sum_{k=1}^K y_{k,n}\alpha_k^{(t)}} \end{aligned} \tag{2}$$

Then the Q-function is

$$\begin{aligned} Q &= E_{Z|Y,\alpha^{(t)}}[\log p(Y, Z; \theta)] \\ &= \sum_{n=1}^N E_{Z_n|Y_n,\alpha^{(t)}}[\log p(Y_n, Z_n; \alpha)] \\ &= \sum_{n=1}^N \sum_{k=1}^K p(Z_n = k|Y_n; \alpha^{(t)}) \log p(Z_n = k, Y_n; \alpha) \\ &= \sum_{n=1}^N \sum_{k=1}^K \frac{y_{k,n}\alpha_k^{(t)}}{\sum_{l=1}^K y_{ln}\alpha_l^{(t)}} \log(y_{k,n}\alpha_k) \end{aligned}$$

In the M-step, α is optimized. Using a Lagrange multiplier and maximizing the following quantity

$$\sum_{n=1}^N \sum_{k=1}^K \frac{y_{k,n}\alpha_k^{(t)}}{\sum_{l=1}^K y_{l,n}\alpha_l^{(t)}} \log(y_{k,n}\alpha_k) - \lambda(\sum_k \alpha_k - 1),$$

gives

$$\sum_{n=1}^N \frac{y_{k,n}\alpha_k^{(t)}}{\sum_{l=1}^K y_{l,n}\alpha_l^{(t)}} \frac{1}{\alpha_k} - \lambda = 0,$$

and

$$\alpha_k^{(t+1)} = \frac{1}{N} \sum_{n=1}^N \frac{y_{k,n}\alpha_k^{(t)}}{\sum_{l=1}^K y_{l,n}\alpha_l^{(t)}} \quad . \tag{3}$$

(a) (5 points) Identifiability in this model means that different parameter values (transcript abundances) result in different (unique) probability distributions on the read counts. Equivalently, identifiability is satisfied if the compatibility matrix is full rank. Determine if this system is identifiable, from its compatibility matrix.

A: Using Gaussian elimination:

$\begin{pmatrix} 1 & 0 & 1 & 1 & 1 \\ 1 & 1 & 0 & 0 & 1 \\ 1 & 1 & 1 & 0 & 0 \end{pmatrix} \to \begin{pmatrix} 1 & 0 & 1 & 1 & 1 \\ 0 & 1 & -1 & -1 & 0 \\ 0 & 1 & 0 & -1 & -1 \end{pmatrix} \to \begin{pmatrix} 1 & 0 & 1 & 1 & 1 \\ 0 & 1 & -1 & -1 & 0 \\ 0 & 0 & 1 & 0 & -1 \end{pmatrix}$ Since it has full rank, this system is identifiable.

For (b) - (d) fill out the Problem 2 notebook here to estimate the transcript abundances α.

Your edited version of the notebook *must be submitted* for this problem. Reminder to check that your notebook runs all the way through with the the `Runtime` → `Restart` and `Runtime` → `Run All` commands.

Problem 2 solution here

Problem 3

In this problem you will develop code for running the EM algorithm to fit a Gaussian Mixture Model (GMM). You will learn the mixture weights for a set of (multivariate) Gaussian distributions, which describe the input, single-cell data. This is a common approach to determine clusters within a dataset. Access the Problem 3 notebook here.

Briefly, a GMM is defined as

$$f_{GMM}(\mathbf{x}) = \sum_{j=1}^{k} \phi_j f(\mathbf{x}; \boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j) \tag{4}$$

subject to $\sum_{j=1}^{k} \phi_j = 1$.

ϕ denotes the weights for each Gaussian pdf f, and together the GMM is defined as the weighted sum of these Gaussians each with parameters $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$. From the EM procedure you will fit the parameters $\boldsymbol{\phi}, \boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ given a single-cell dataset.

Your edited version of the notebook *must be submitted* for this problem. Reminder to check that your notebook runs all the way through with the the `Runtime` → `Restart` and `Runtime` → `Run All` commands.

Problem 3 solution here

References

1. Pachter, L. Models for transcript quantification from RNA-Seq. arXiv: `1104.3889 [q-bio.GN]` (Apr. 2011).