The EM Algorithm

The EM algorithm is used for obtaining maximum likelihood estimates of parameters when some of the data is *missing*. More generally, however, the EM algorithm can also be applied when there is *latent*, i.e. unobserved, data which was never intended to be observed in the first place. In that case, we simply assume that the latent data is missing and proceed to apply the EM algorithm. The EM algorithm has many applications throughout statistics. It is often used for example, in machine learning and data mining applications, and in Bayesian statistics where it is often used to obtain the mode of the posterior marginal distributions of parameters.

1 The Classical EM Algorithm

We begin by assuming that the complete data-set consists of $\mathcal{Z} = (\mathcal{X}, \mathcal{Y})$ but that only $\mathcal{X}$ is observed. The complete-data log likelihood is then denoted by $l(\theta; \mathcal{X}, \mathcal{Y})$ where θ is the unknown parameter vector for which we wish to find the MLE.

E-Step: The E-step of the EM algorithm computes the expected value of $l(\theta; \mathcal{X}, \mathcal{Y})$ given the observed data, $\mathcal{X}$, and the current parameter estimate, θ_{old} say. In particular, we define

$$\begin{aligned} Q(\theta; \theta_{old}) &:= \mathsf{E}\left[l(\theta; \mathcal{X}, \mathcal{Y}) \mid \mathcal{X}, \theta_{old}\right] \\ &= \int l(\theta; \mathcal{X}, y)\, p(y \mid \mathcal{X}, \theta_{old})\, dy \end{aligned} \tag{1}$$

where $p(\cdot \mid \mathcal{X}, \theta_{old})$ is the conditional density of $\mathcal{Y}$ given the observed data, $\mathcal{X}$, and assuming $\theta = \theta_{old}$.

M-Step: The M-step consists of maximizing over θ the expectation computed in (1). That is, we set

$$\theta_{new} := \max_{\theta} Q(\theta; \theta_{old}).$$

We then set $\theta_{old} = \theta_{new}$.

The two steps are repeated as necessary until the sequence of θ_{new}'s converges. Indeed under very general circumstances convergence to a local maximum can be guaranteed and we explain why this is the case below. If it is suspected that the log-likelihood function has multiple local maximums then the EM algorithm should be run many times, using a different starting value of θ_{old} on each occasion. The ML estimate of θ is then taken to be the best of the set of local maximums obtained from the various runs of the EM algorithm.

Why Does the EM Algorithm Work?

We use $p(\cdot \mid \cdot)$ to denote a generic conditional PDF. Now observe that

$$\begin{aligned} l(\theta; \mathcal{X}) = \ln p(\mathcal{X} \mid \theta) &= \ln \int p(\mathcal{X}, y \mid \theta)\, dy \\ &= \ln \int \frac{p(\mathcal{X}, y \mid \theta)}{p(y \mid \mathcal{X}, \theta_{old})} p(y \mid \mathcal{X}, \theta_{old})\, dy \end{aligned}$$

$$
\begin{aligned}
&= \ln \mathsf{E}\left[\frac{p(\mathcal{X},\mathcal{Y}\mid\theta)}{p(\mathcal{Y}\mid\mathcal{X},\theta_{old})} \mid \mathcal{X},\theta_{old}\right] \\
&\geq \mathsf{E}\left[\ln\left(\frac{p(\mathcal{X},\mathcal{Y}\mid\theta)}{p(\mathcal{Y}\mid\mathcal{X},\theta_{old})}\right) \mid \mathcal{X},\theta_{old}\right] \qquad (2)\\
&= \mathsf{E}\left[\ln p(\mathcal{X},\mathcal{Y}\mid\theta)\mid\mathcal{X},\theta_{old}\right] \ - \ \mathsf{E}\left[\ln p(\mathcal{Y}\mid\mathcal{X},\theta_{old})\mid\mathcal{X},\theta_{old}\right] \\
&= Q(\theta;\theta_{old}) \ - \ \mathsf{E}\left[\ln p(\mathcal{Y}\mid\mathcal{X},\theta_{old})\mid\mathcal{X},\theta_{old}\right] \qquad (3)
\end{aligned}
$$

where (2) follows from Jensen's Inequality and since the $\ln$ function is concave. It is also clear (because the term inside the expectation becomes a constant) that the inequality in (2) becomes an equality if we take $\theta = \theta_{old}$. Letting $g(\theta \mid \theta_{old})$ denote the right-hand-side of (3), we therefore have

$$l(\theta;\mathcal{X}) \ \geq \ g(\theta\mid\theta_{old})$$

for all θ with equality when $\theta = \theta_{old}$. Therefore any value of θ that increases $g(\theta \mid \theta_{old})$ beyond $g(\theta_{old} \mid \theta_{old})$ must also increase $l(\theta;\mathcal{X})$ beyond $l(\theta_{old};\mathcal{X})$. The M-step finds such a θ by maximizing $Q(\theta;\theta_{old})$ over θ which is equivalent (why?) to maximizing $g(\theta \mid \theta_{old})$ over θ. It is also worth mentioning that in many applications the function $Q(\theta;\theta_{old})$ will be a convex function of θ and therefore easy to optimize.

---

2 Examples

Example 1 (Missing Data in a Multinomial Model)
Suppose $\mathbf{x} := (x_1, x_2, x_3, x_4)$ is a sample from a Mult(n, π_θ) distribution where

$$\pi_\theta \ = \ \left(\frac{1}{2}+\frac{1}{4}\theta,\ \frac{1}{4}(1-\theta),\ \frac{1}{4}(1-\theta),\ \frac{1}{4}\theta\right).$$

The likelihood, $L(\theta;\mathbf{x})$, is then given by

$$L(\theta;\mathbf{x}) = \frac{n!}{x_1!x_2!x_3!x_4!}\left(\frac{1}{2}+\frac{1}{4}\theta\right)^{x_1}\left(\frac{1}{4}(1-\theta)\right)^{x_2}\left(\frac{1}{4}(1-\theta)\right)^{x_3}\left(\frac{1}{4}\theta\right)^{x_4}$$

so that the log-likelihood $l(\theta;\mathbf{x})$ is

$$l(\theta;\mathbf{x}) \ = \ C \ + \ x_1\ln\left(\frac{1}{2}+\frac{1}{4}\theta\right) \ + \ (x_2+x_3)\ln(1-\theta) \ + \ x_4\ln(\theta)$$

where C is a constant that does not depend on θ. We could try to maximize $l(\theta;\mathbf{x})$ over θ directly using standard non-linear optimization algorithms. However, in this example we will perform the optimization instead using the EM algorithm. To do this we assume that the complete data is given by $\mathbf{y} := (y_1, y_2, y_3, y_4, y_5)$ and that $\mathbf{y}$ has a Mult(n, π^*_θ) distribution where

$$\pi^*_\theta \ = \ \left(\frac{1}{2},\ \frac{1}{4}\theta,\ \frac{1}{4}(1-\theta),\ \frac{1}{4}(1-\theta),\ \frac{1}{4}\theta\right).$$

However, instead of observing $\mathbf{y}$ we only observe $(y_1 + y_2, y_3, y_4, y_5)$, i.e, we only observe $\mathbf{x}$. We therefore take $\mathcal{X} = (y_1 + y_2, y_3, y_4, y_5)$ and take $\mathcal{Y} = y_2$. The log-likelihood of the complete data is then given by

$$l(\theta;\mathcal{X},\mathcal{Y}) \ = \ C \ + \ y_2\ln(\theta) \ + \ (y_3+y_4)\ln(1-\theta) \ + \ y_5\ln(\theta)$$

where again C is a constant containing all terms that do no depend on θ. It is also clear that the conditional density of $\mathcal{Y}$ satisfies

$$f(\mathcal{Y}\mid\mathcal{X},\theta) \ = \ \mathsf{Bin}\left(y_1+y_2,\ \frac{\theta/4}{1/2+\theta/4}\right).$$

We can now implement the E-step and M-step.

E-Step: Recalling that $Q(\theta;\theta_{old}) := \mathsf{E}\left[l(\theta;\mathcal{X},\mathcal{Y}) \mid \mathcal{X},\theta_{old}\right]$, we have

$$\begin{aligned} Q(\theta;\theta_{old}) &:= C \;+\; \mathsf{E}\left[y_2 \ln(\theta) \mid \mathcal{X},\theta_{old}\right] \;+\; (y_3+y_4)\ln(1-\theta) \;+\; y_5\ln(\theta) \\ &= C \;+\; (y_1+y_2)p_{old}\ln(\theta) \;+\; (y_3+y_4)\ln(1-\theta) \;+\; y_5\ln(\theta) \end{aligned}$$

where

$$p_{old} := \frac{\theta_{old}/4}{1/2+\theta_{old}/4}. \tag{4}$$

M-Step: We now maximize $Q(\theta;\theta_{old})$ to find θ_{new}. Taking the derivative we obtain

$$\frac{dQ}{d\theta} \;=\; \frac{(y_1+y_2)}{\theta}\,p_{old} \;-\; \frac{(y_3+y_4)}{1-\theta} \;+\; \frac{y_5}{\theta}$$

which is zero when we take $\theta \;=\; \theta_{new}$ where

$$\theta_{new} \;:=\; \frac{y_5 + p_{old}(y_1+y_2)}{y_3+y_4+y_5+p_{old}(y_1+y_2)} \tag{5}$$

Equations (4) and (5) now define the EM iteration which begins with some (judiciously) chosen value of θ_{old}. ∎

Example 2 (A Simple Normal-Mixture Model)
An extremely common application of the EM algorithm is to estimate the MLE of normal mixture models. This is often used, for example, in clustering algorithms. Suppose for example that $\mathcal{X} = (X_1,\ldots,X_n)$ are IID random variables each with PDF

$$f_x(x) \;=\; \sum_{j=1}^{m} p_j\,\frac{e^{-(x-\mu_j)^2/2\sigma_j^2}}{\sqrt{2\pi\sigma_j^2}}$$

where $p_j \geq 0$ for all j and where $\sum p_j = 1$. The parameters in this model are the p_j's, the μ_j's and the σ_j's. Instead of trying to finding the maximum likelihood estimates of these parameters directly via numerical optimization, we can use the EM algorithm. We do this by assuming the presence of an additional random variable, Y say, where $P(Y=j) = p_j$ for $j = 1,\ldots,m$. The realized value of Y then determines which of the m normal distributions generates the corresponding value of X. There are n such random variables, $(Y_1,\ldots,Y_n) := \mathcal{Y}$. Note that

$$f_{x|y}(x_i \mid y_i = j,\theta) \;=\; \frac{1}{\sqrt{2\pi\sigma_j^2}}e^{-(x_i-\mu_j)^2/2\sigma_j^2} \tag{6}$$

where $\theta := (p_1,\ldots,p_m,\mu_1,\ldots,\mu_m,\sigma_1,\ldots,\sigma_m)$ is the unknown parameter vector and that the likelihood is given by

$$L(\theta;\mathcal{X},\mathcal{Y}) = \prod_{i=1}^{n} p_{y_i}\frac{1}{\sqrt{2\pi\sigma_{y_i}^2}}e^{-(x_i-\mu_{y_i})^2/2\sigma_{y_i}^2}.$$

The EM algorithm starts with an initial guess, θ_{old}, and then iterates the E-step and M-step as described below until convergence.

E-Step: We need to compute $Q(\theta;\theta_{old}) := \mathsf{E}\left[l(\theta;\mathcal{X},\mathcal{Y}) \mid \mathcal{X},\theta_{old}\right]$. Indeed it is straightforward to show that

$$Q(\theta;\theta_{old}) \;=\; \sum_{i=1}^{n}\sum_{j=1}^{m} P(Y_i=j \mid x_i,\theta_{old})\;\ln\left(f_{x|y}(x_i \mid y_i=j,\theta)\,P(Y_i=j \mid \theta)\right). \tag{7}$$

Note that $f_{x|y}(x_i \mid y_i = j, \theta)$ is given by (6) and that $P(Y_i = j \mid \theta_{old}) = p_{j,old}$. Finally, since

$$P(Y_i = j \mid x_i, \theta_{old}) \;=\; \frac{P(Y_i = j,\ X_i = x_i \mid \theta_{old})}{P(X_i = x_i \mid \theta_{old})} \;=\; \frac{f_{x|y}(x_i \mid y_i = j, \theta_{old})\ P(Y_i = j \mid \theta_{old})}{\sum_{k=1}^{m} f_{x|y}(x_i \mid y_i = k, \theta_{old})\ P(Y_i = k \mid \theta_{old})}$$

it is clear that we can compute (7) analytically.

M-Step: We can now maximize $Q(\theta; \theta_{old})$ by setting the vector of partial derivatives, $\partial Q / \partial \theta$, equal to 0 and solving for θ_{new}. After some algebra, we obtain

$$\mu_{j,new} \;=\; \frac{\sum_{i=1}^{n} x_i P(Y_i = j \mid x_i, \theta_{old})}{\sum_{i=1}^{n} P(Y_i = j \mid x_i, \theta_{old})} \tag{8}$$

$$\sigma^2_{j,new} \;=\; \frac{\sum_{i=1}^{n} (x_i - \mu_j)^2 P(Y_i = j \mid x_i, \theta_{old})}{\sum_{i=1}^{n} P(Y_i = j \mid x_i, \theta_{old})} \tag{9}$$

$$p_{j,new} \;=\; \frac{1}{n} \sum_{i=1}^{n} P(Y_i = j \mid x_i, \theta_{old}). \tag{10}$$

Given an initial estimate, θ_{old}, the EM algorithm cycles through (8) to (10) repeatedly, setting $\theta_{old} = \theta_{new}$ after each cycle, until the estimates converge. ∎

---

3 A More General Version of the EM Algorithm

The EM algorithm is often stated more generally using the language of information theory. In this section[1] we will describe this more general formulation and relate it back to EM algorithm as described in Section 1. As before the goal is to maximize the likelihood function, $L(\theta; \mathcal{X})$, which is given by[2]

$$L(\theta; \mathcal{X}) = p(\mathcal{X} \mid \theta) = \int_y p(\mathcal{X}, y \mid \theta)\, dy. \tag{11}$$

The implicit assumption underlying the EM algorithm is that it is difficult to optimize $p(\mathcal{X} \mid \theta)$ with respect to θ but that it is much easier to optimize $p(\mathcal{X}, \mathcal{Y} \mid \theta)$. We first introduce an arbitrary distribution, $q(\mathcal{Y})$, over the latent variables, $\mathcal{Y}$, and note that we can decompose the log-likelihood, $l(\theta; \mathcal{X})$, into two terms (the first of which is sometimes called the "energy" term) according to

$$l(\theta; \mathcal{X}) := \ln p(\mathcal{X} \mid \theta) \;=\; \underbrace{\mathcal{L}(q, \theta)}_{\text{"energy"}} \;+\; \mathsf{KL}(q \,||\, p_{\mathcal{Y}|\mathcal{X}}) \tag{12}$$

where $\mathcal{L}(q, \theta)$ and $\mathsf{KL}(q \,||\, p_{\mathcal{Y}|\mathcal{X}})$ are the likelihood and Kullback-Leibler (KL) divergence[3] which are given by

$$\begin{aligned} \mathcal{L}(q, \theta) \;&=\; \int_{\mathcal{Y}} q(\mathcal{Y}) \ln \left(\frac{p(\mathcal{X}, \mathcal{Y} \mid \theta)}{q(\mathcal{Y})} \right) \\ \mathsf{KL}(q \,||\, p_{\mathcal{Y}|\mathcal{X}}) \;&=\; -\int_{\mathcal{Y}} q(\mathcal{Y}) \ln \left(\frac{p(\mathcal{Y} \mid \mathcal{X}, \theta)}{q(\mathcal{Y})} \right). \end{aligned} \tag{13}$$

It is well-known (see the Appendix) that the KL divergence satisfies $\mathsf{KL}(q \,||\, p_{\mathcal{Y}|\mathcal{X}}) \geq 0$ and equals 0 if and only if $q(\mathcal{Y}) = p_{\mathcal{Y}|\mathcal{X}}$. It therefore follows that $\mathcal{L}(q, \theta) \leq l(\theta; \mathcal{X})$ for all distributions, $q(\cdot)$. We can now use the decomposition of (12) to define the EM algorithm. We begin with an initial parameter estimate, θ_{old}.

E-Step: The E-step maximizes the lower bound, $\mathcal{L}(q, \theta_{old})$, with respect to $q(\cdot)$ while keeping θ_{old} fixed. In principle this is a variational problem since we are optimizing a functional, but the solution is easily found.

First note that $l(\theta_{old}; \mathcal{X})$ does not depend on $q(\cdot)$. It then follows from (12) (with $\theta = \theta_{old}$) that maximizing $\mathcal{L}(q, \theta_{old})$ is equivalent to minimizing $\mathsf{KL}(q \,||\, p_{\mathcal{Y}|\mathcal{X}})$. Since this latter term is always non-negative we see that

[1] The material in this section is drawn from *Pattern Recognition and Machine Learning* (2006) by Chris Bishop.
[2] If $\mathcal{Y}$ is discrete then we replace the integral in (11) with a summation.
[3] The KL divergence is also often called the *relative entropy*.

$\mathcal{L}(q, \theta_{old})$ is optimized when $\mathsf{KL}(q \,||\, p_{\mathcal{Y}|\mathcal{X}}) = 0$ which, by our earlier observation, is the case when we take $q(\mathcal{Y}) = p(\mathcal{Y} \,|\, \mathcal{X}, \theta_{old})$. At this point we see that the lower bound, $\mathcal{L}(q, \theta_{old})$, will now equal the current value of the log-likelihood, $l(\theta_{old}; \mathcal{X})$.

M-Step: In the M-step we keep $q(\mathcal{Y})$ fixed and maximize $\mathcal{L}(q, \theta)$ over θ to obtain θ_{new}. This will therefore cause the lower bound to increase (if it is not already at a maximum) which in turn means that the log-likelihood must also increase. Moreover, at this new value θ_{new} it will no longer be the case that $\mathsf{KL}(q \,||\, p_{\mathcal{Y}|\mathcal{X}}) = 0$ and so by (12) the increase in the log-likelihood will be greater than the increase in the lower bound.

Comparing the General EM Algorithm with the Classical EM Algorithm

It is instructive to compare the E-step and M-step of the general EM algorithm with the corresponding steps of Section 1. To do this, first substitute $q(\mathcal{Y}) = p(\mathcal{Y} \,|\, \mathcal{X}, \theta_{old})$ into (13) to obtain

$$\mathcal{L}(q, \theta) \;=\; Q(\theta; \theta_{old}) \;+\; \text{constant} \tag{14}$$

where $Q(\theta; \theta_{old})$ is the expected complete-date log-likelihood as defined in (1) where the expectation is taken assuming $\theta = \theta_{old}$. The M-step of the general EM algorithm is therefore identical to the M-step of Section 1 since the constant term in (14) does not depend on θ.

The E-step in the general EM algorithm takes $q(\mathcal{Y}) = p(\mathcal{Y} \,|\, \mathcal{X}, \theta_{old})$ which, at first glance, appears to be different to the E-step in Section 1. But there is no practical difference: the E-step in Section 1 simply uses $p(\mathcal{Y} \,|\, \mathcal{X}, \theta_{old})$ to compute $Q(\theta; \theta_{old})$ and, while not explicitly stated, the general E-step must also do this since it is required for the M-step.

3.1 The Case of Independent and Identically Distributed Observations

When the data set consists of N IID observations, $\mathcal{X} = \{x_n\}$ with corresponding latent or unobserved variables, $\mathcal{Y} = \{y_n\}$, then we can simplify the calculation of $p(\mathcal{Y} \,|\, \mathcal{X}, \theta_{old})$. In particular we obtain

$$p(\mathcal{Y} \,|\, \mathcal{X}, \theta_{old}) \;=\; \frac{p(\mathcal{X}, \mathcal{Y} \,|\, \theta_{old})}{\sum_{\mathcal{Y}} p(\mathcal{X}, \mathcal{Y} \,|\, \theta_{old})} \;=\; \frac{\prod_{n=1}^{N} p(x_n, y_n \,|\, \theta_{old})}{\sum_{\mathcal{Y}} \prod_{n=1}^{N} p(x_n, y_n \,|\, \theta_{old})} \;=\; \frac{\prod_{n=1}^{N} p(x_n, y_n \,|\, \theta_{old})}{\prod_{n=1}^{N} p(x_n \,|\, \theta_{old})} \;=\; \prod_{n=1}^{N} p(y_n \,|\, x_n, \theta_{old})$$

so that the posterior distribution also factorizes which makes its calculation much easier. Note that the E-step of Example 2 is consistent with this observation.

3.2 Bayesian Applications

The EM algorithm can also be used to compute the mode of the posterior distribution, $p(\theta \,|\, \mathcal{X})$, in a Bayesian setting where we are given a prior, $p(\theta)$, on the unknown parameter (vector), θ. To see this, first note that we can write $p(\theta \,|\, \mathcal{X}) = p(\mathcal{X} \,|\, \theta) p(\theta) / p(\mathcal{X})$ which upon taking logs yields

$$\ln p(\theta \,|\, \mathcal{X}) = \ln p(\mathcal{X} \,|\, \theta) + \ln p(\theta) - \ln p(\mathcal{X}). \tag{15}$$

If we now use (12) to substitute for $\ln p(\mathcal{X} \,|\, \theta)$ on the right-hand-side of (15) we obtain

$$\ln p(\theta \,|\, \mathcal{X}) \;=\; \mathcal{L}(q, \theta) \;+\; \mathsf{KL}(q \,||\, p_{\mathcal{Y}|\mathcal{X}}) + \ln p(\theta) - \ln p(\mathcal{X}). \tag{16}$$

We can now find the posterior mode of $\ln p(\theta \,|\, \mathcal{X})$ using a version of the EM algorithm. The E-step is exactly the same as before since the final two terms on the right-hand-side of (16) do not depend on $q(\cdot)$. The M-step, where we keep $q(\cdot)$ fixed and optimize over θ, must be modified however to include the $\ln p(\theta)$ term.

There are also related methods that can be used to estimate the variance-covariance matrix, $\mathbf{\Sigma}$, of θ. In this case it is quite common to approximate the posterior distribution of θ with a Gaussian distribution centered at the mode and with variance-covariance matrix, $\mathbf{\Sigma}$. This is called a *Laplacian approximation* and it is a simple but commonly used framework for deterministic inference. It only works well, however, when the posterior is

unimodal with contours that are approximately elliptical. It is also worth pointing out, however, that it is often straightforward to compute the mode of the posterior and determine a suitable $\mathbf{\Sigma}$ for the Gaussian approximation so that the Laplacian approximation need not rely on the EM algorithm.

We also note in passing that the decomposition in (12) also forms the basis of another commonly used method of deterministic inference called *variational Bayes*. The goal with variational Bayes is to select $q(\cdot)$ from some parametric family of distributions, $\mathcal{Q}$, to approximate $p(\mathcal{Y} \,|\, \mathcal{X})$. The dependence on θ is omitted since we are now in a Bayesian setting and θ can be subsumed into the latent or hidden variables, $\mathcal{Y}$. In choosing $q(\cdot)$ we seek to maximize the lower bound, $\mathcal{L}(q)$, or equivalently by (12), to minimize $\mathsf{KL}(q \,||\, p_{\mathcal{Y}|\mathcal{X}})$. A common choice of $\mathcal{Q}$ is the set of distributions under which the latent variables are independent.

Appendix: Kullback-Leibler Divergence

Let P and Q be two probability distributions such that if $P(\mathbf{x}) = 0$ then $Q(\mathbf{x}) = 0$. The Kullback-Leibler (KL) divergence or *relative entropy* of Q from P is defined to be

$$\mathsf{KL}(P \,||\, Q) \;=\; \int_{\mathbf{x}} P(\mathbf{x}) \ln\left(\frac{P(\mathbf{x})}{Q(\mathbf{x})}\right) \tag{17}$$

with the understanding[4] that $0 \log 0 = 0$. The KL divergence is a fundamental concept in *information theory* and machine learning. One can imagine P representing some true but unknown distribution that we approximate with Q and that $\mathsf{KL}(P \,||\, Q)$ measures the "distance" between P and Q. This interpretation is valid because we will see below that $\mathsf{KL}(P \,||\, Q) \geq 0$ with equality if and only if $P = Q$. Note however, that the KL divergence is not a true measure of distance since it is asymmetric in that $\mathsf{KL}(P \,||\, Q) \neq \mathsf{KL}(Q \,||\, P)$ and does not satisfy the triangle inequality.

In order to see that $\mathsf{KL}(P \,||\, Q) \geq 0$ we first recall that a function $f(\cdot)$ is *convex* on $\mathbb{R}$ if

$$f(\alpha x + (1-\alpha)y) \leq \alpha f(x) + (1-\alpha) f(y) \quad \text{for all } \alpha \in [0,1].$$

We also recall Jensen's inequality:

Jensen's Inequality: Let $f(\cdot)$ be a convex function on $\mathbb{R}$ and suppose $\mathrm{E}[X] < \infty$ and $\mathrm{E}[f(X)] < \infty$. Then $f(\mathrm{E}[X]) \leq \mathrm{E}[f(X)]$.

Noting that $-\ln(x)$ is a convex function we have

$$\begin{aligned}
\mathsf{KL}(P \,||\, Q) &= -\int_{\mathbf{x}} P(\mathbf{x}) \ln\left(\frac{Q(\mathbf{x})}{P(\mathbf{x})}\right) \\
&\geq -\ln\left(\int_{\mathbf{x}} P(\mathbf{x}) \frac{Q(\mathbf{x})}{P(\mathbf{x})}\right) \qquad \text{by Jensen's inequality} \\
&= 0.
\end{aligned}$$

Moreover it is clear from (17) that $\mathsf{KL}(P \,||\, Q) = 0$ if $P = Q$. In fact because $-\ln(x)$ is strictly convex it is easy to see that $\mathsf{KL}(P \,||\, Q) = 0$ only if $P = Q$.

Example 3 (An Optimization "Trick" that's Worth Remembering)
Suppose $\mathbf{c}$ is a non-negative vector in $\mathbb{R}^n$, i.e. $\mathbf{c} \in \mathbb{R}^n_+$, and we wish to maximize $\sum_{i=1}^n c_i \ln(q_i)$ over *probability mass functions*, $\mathbf{q} = \{q_1, \ldots, q_n\}$. Let $\mathbf{p} = \{p_1, \ldots, p_n\}$ where $p_i := c_i / \sum_j c_j$ so that $\mathbf{p}$ is a (known) pmf. We then have:

$$\max_{\mathbf{q}} \sum_{i=1}^n c_i \ln(q_i) \;=\; \Big(\sum_{i=1}^n c_i\Big) \max_{\mathbf{q}} \Big\{ \sum_{i=1}^n p_i \ln(q_i) \Big\}$$

[4] This is consistent with the fact that $\lim_{x \to 0} x \log x = 0$.

$$
\begin{aligned}
&= \Big(\sum_{i=1}^{n} c_i\Big) \max_{\mathbf{q}} \Big\{ \sum_{i=1}^{n} p_i \ln(p_i) - \sum_{i=1}^{n} p_i \ln\left(\frac{p_i}{q_i}\right) \Big\}, \\
&= \Big(\sum_{i=1}^{n} c_i\Big) \left(\sum_{i=1}^{n} p_i \ln(p_i) - \min_{\mathbf{q}} \mathsf{KL}(\mathbf{p} \,||\, \mathbf{q}) \right)
\end{aligned}
$$

from which it follows (why?) that the optimal $\mathbf{q}^*$ satisfies $\mathbf{q}^* = \mathbf{p}$. ∎