Maximum Likelihood from Incomplete Data

Let $Q$ be a family of probability measures on $\{\mathcal{X}, \mathcal{F}\}$. Let $X_1, \ldots, X_n$ be i.i.d. random variables taking values in $\mathcal{X}$, possibly drawn according to some $Q \in Q$. Let $\hat{Q}$, given by

$$
\hat{Q} = \hat{Q}_{x_1, \ldots, x_n} = \arg \max_{Q \in \mathcal{Q}} \prod_{t=1}^{n} Q(x_t),
$$

be the measure in $Q$ which assigns the maximum likelihood to the sequence $x_1, \ldots, x_n$. We have seen that if we denote the empirical measure of the observations by

$$
\hat{P}(x) = \frac{1}{n} \sum_{t=1}^{n} \mathbf{1}(x_t = x), \quad x \in \mathcal{X},
$$

then

$$
\log \prod_{t=1}^{n} Q(x_t) = -H(\hat{P}) - D(\hat{P}||Q),
$$

and thus the ML estimate also satisfies

$$
\hat{Q} = \arg \min_{Q \in \mathcal{Q}} D(\hat{P}||Q).
$$

Thus the classical ML estimation problem is a problem of divergence minimization between a measure $\hat{P}$ and a family of measures $Q$.

A variation of the ML estimation problem which is often encountered is that of determining $\hat{Q}$ not directly from $x_1, \ldots, x_n$, but some function of these variables. The random variables $X_t$ may be

- intensities of pixels in an image, and we may be given quantized versions of the actual intensities, e.g., to eight bits of resolution (noisy data).
• the time-to-failure of a light bulb, and we may be given the actual lifetime if it is less than 1000 hours, or the information that the lifetime exceeds 1000 hours, at which time the test is terminated (censored or truncated data).

• the energy in multiple spectral bands of radiation measured on a remote sensing satellite, and we may be given the energies only in those bands for which sensors were on-board and functional at the time of measurement (missing data).

• the state-output pairs of a hidden Markov source, and we may be given only the observed outputs (hidden data).

Notice that while the first three examples have a real notion of incomplete data, the last is a typical case in which there really is no missing data – we simply hypothesize a missing part, either to better model that data or, sometimes, to replace an intractable model class on \( \mathcal{Y} \) with a more tractable one on \( \mathcal{X} \). These problems can be reformulated as follows.

Let \( \{ \mathcal{Y}, \tilde{\mathcal{F}} \} \) be the space of observations and let \( T: \mathcal{X} \to \mathcal{Y} \) be a measurable function. For a measure \( Q \) on \( \mathcal{X} \), define the induced measure \( Q^T \) on \( \mathcal{Y} \) by assigning to each event in \( \mathcal{Y} \) the probability of its inverse image (under \( T \)) in \( \mathcal{X} \). Let \( \mathcal{Q}^T \) denote the set of measures induced by \( Q \in \mathcal{Q} \). Then, the ML estimation problem, given the observations \( y_1, \ldots, y_n \), is to determine

\[
\hat{Q} = \hat{Q}_{y_1, \ldots, y_n} = \arg \max_{Q \in \mathcal{Q}} \prod_{t=1}^{n} Q^T(y_t).
\]

We shall see next, that this too is a divergence minimization problem. To this end, define

\[
P = \{ P : P^T = \hat{P} \},
\]

where \( P \) is a probability measure on \( \mathcal{X} \) and \( \hat{P} = \hat{P}_{y_1, \ldots, y_n} \) is the empirical measure induced on \( \mathcal{Y} \) by the observations. Next, note that

\[
D(P||Q) = \sum_{x \in \mathcal{X}} P(x) \log \frac{P(x)}{Q(x)} = \sum_{y \in \mathcal{Y}} \sum_{x: T(x) = y} P(x) \log \frac{P(x)}{Q(x)} = \sum_{y \in \mathcal{Y}} \sum_{x: T(x) = y} P^T(y) P(x|T(x) = y) \log \frac{P^T(y) P(x|T(x) = y)}{Q^T(y) Q(x|T(x) = y)} = \sum_{y \in \mathcal{Y}} P^T(y) \log \frac{P^T(y)}{Q^T(y)} + P^T(y) \sum_{x: T(x) = y} P(x|T(x) = y) \log \frac{P(x|T(x) = y)}{Q(x|T(x) = y)} = D(P^T||Q^T) + D(P^X|Y||Q^X|Y | P^T)
\]

where the last term denotes the divergence between the conditional distributions on \( \mathcal{X} \), given \( Y \). Now, if we fix a \( Q \in \mathcal{Q} \) and consider minimizing the abovementioned I-divergence over
all measures $P \in \mathcal{P}$, it is easy to see that the first term is the same for all $P \in \mathcal{P}$, and the second term, which is nonnegative, can be made to vanish by setting $P^X | Y = Q^X | Y$. Therefore

$$\min_{P \in \mathcal{P}} D(P || Q) = D(\hat{P} || Q^T), \quad Q \in \mathcal{Q}.$$ 

It follows then that the ML estimate $\hat{Q}$ based on the observations $y_1, \ldots, y_n$ satisfies

$$\hat{Q} = \arg\min_{Q \in \mathcal{Q}} D(\hat{P} || Q^T) = \arg\min_{Q \in \mathcal{Q}} \min_{P \in \mathcal{P}} D(P || Q).$$

We therefore must now minimize the I-divergence between two sets of measures $\mathcal{P}$ and $\mathcal{Q}$ and, in particular, find the $Q \in \mathcal{Q}$ which achieves this minimum.

This problem is a special instance of a more general problem of minimizing a distance-like function between two abstract sets of points.

- We present this general problem and a solution methodology commonly known as alternating minimization.

- We then show that the I-divergence problem is indeed a special case of this problem, and specialize the methodology to the case of I-divergence minimization, yielding the well known class of EM (or Expectation-Maximization) algorithms.

However, to give the reader a more concrete grasp of the nature of the problem and the mechanics of the solution, we first state the classical EM algorithm without proof. We also present a simple application of the EM algorithm to an incomplete data problem where the ML solution is directly computable as well (a rare occurrence indeed). The example illustrates, perhaps in an oversimplified manner, the relationship between the amount of missing data, and the rate of convergence of the EM algorithm. The EM algorithm is revisited after proofs of convergence of the more general alternation minimization problem.

**Classical Formulation of the EM Algorithm**

Let $X \in \mathcal{X}$ be a random vector generated by a probability density function

$$g_c(x; \Psi), \quad x \in \mathcal{X},$$

where $\Psi = (\Psi_1, \ldots, \Psi_d) \in \Omega$ is a vector of unknown parameters. Let $Y = T(X)$ be the observed random variable corresponding to $X$, where the transformation $T : \mathcal{X} \to \mathcal{Y}$ could
account for noisy, censored, or missing data, e.g., \(X = (Z, Y)\). Let the induced density on the space of \(Y\) be

\[
g(y; \Psi) = \int_{T(x) = y} g_c(x; \Psi) \, dx, \quad y \in \mathcal{Y}, \quad \Psi \in \Omega.
\]

Given this parametric model and observations \(Y = y\), we wish to determine the value of the parameter \(\Psi\) which maximizes the likelihood of the observed data

\[
L(\Psi) = g(y; \Psi).
\]

In many cases, however, this is a very difficult computational problem primarily because the induced likelihood function \(g\) is difficult to compute in closed form. For instance, \(Y\) may be obtained by passing \(X\) through a matched filter, a convolutional decoder, or a nonlinearity (e.g., thresholding), making it difficult to maximize \(g\) directly as a function of \(\Psi\). If the functional form of \(g_c\) is more tractable, this fact may be used to solve this problem under suitable conditions.

- The likelihood function, if the observations \(X\) were given, is denoted by

\[
L_c(\Psi) = g_c(X; \Psi).
\]

Since \(X\) is not available to compute \(\Psi\) from this form of \(g_c\), we “estimate” \(X\) based on the observed \(Y\). In particular we use the conditional distribution of \(X\) to compute the average log-likelihood

\[
Q(\Psi; \Psi^{(0)}) = E_{\Psi^{(0)}}[\log L_c(\Psi) \mid Y = y],
\]

for some nominal value \(\Psi^{(0)}\) of the parameters. This is sometimes easier than computing the induced measure \(g\).

- We then maximize this as a function of \(\Psi\), i.e. we replace \(\Psi^{(0)}\) by

\[
\Psi^{(1)} = \arg \max_{\Psi \in \Omega} Q(\Psi; \Psi^{(0)}).
\]

This maximization, again, is sometimes easier to perform than the maximization of \(g\) as a function of \(\Psi\).

These two steps may then be repeated with \(\Psi^{(1)}\) in the role of \(\Psi^{(0)}\), etc. This general framework for computing the maximizer of \(L(\Psi)\) is called the Expectation-Maximization (EM) algorithm, after the nature of the two alternating operation involved in each iteration. It is perhaps more appropriate to call it the EM technique because the framework does not address the computational issues in either step.
We shall see shortly that this technique indeed yields reasonable values of $\Psi$ – the asymptotic values are local maximizers of the likelihood function $L(\Psi)$ under fairly mild conditions. In particular, we shall see that if we start with $\Psi^{(k)}$, and compute $\Psi^{(k+1)}$ using the two steps described above then, under most circumstances, $L(\Psi^{(k+1)}) \geq L(\Psi^{(k)})$. Therefore, if the likelihood function is bounded above, it must converge to a local maximum. This fact also provides us with a sanity check for this technique – if we chance upon the true ML estimate $\hat{\Psi}$, then $\Psi^{(k)} = \hat{\Psi}$ implies $\Psi^{(k+1)} = \hat{\Psi}$.

An Example of Censored Data

Let $W_1, \ldots, W_n$ be i.i.d. random variables taking values in $X = \mathbb{R}$, with common density

$$f(w; \mu) = \frac{1}{\mu} \exp \left\{ -\frac{w}{\mu} \right\} 1(w > 0), \quad w \in \mathbb{R},$$

for some parameter $\mu \in \Omega = \mathbb{R}^+$. These, for instance, may represent the time to failure of a batch of devices, and the manufacturer of the devices may be interested in estimating the mean lifetime, $\mu$. If the mean lifetime is large, however, the test must be conducted until a period corresponding to the largest lifetime in the batch has elapsed, so as to “observe” the $W_i$’s. This may be unacceptable for various reasons. Typically, the test is conducted until a suitable fraction of the samples have failed, and the remaining are marked as having a lifetime longer than the test period. The test may further be conducted separately for each sample, and the time limit to censorship may vary from sample to sample. Let

$$Y_i = (y_i, \delta_i), \quad i = 1, \ldots, n,$$

be the censored observations, where

$$y_i = \min \{W_i, c_i\}, \quad \text{and}$$

$$\delta_i = \begin{cases} 0 & \text{if } W_i \leq c_i, \\ 1 & \text{if } W_i > c_i, \end{cases}$$

is the censorship information about the $i$-th observation. The problem is to find the ML estimate of $\mu$ given $\{Y_i, i = 1, \ldots, n\}$. We may assume without loss of generality that $\delta_1 = \cdots = \delta_r = 0$ and $\delta_{r+1} = \cdots = \delta_n = 1$ for some known $r$, with $0 < r \leq n$. Also assume without loss of generality that each observed $y_i$ is nonnegative.

Now, the cumulative distribution function of the time to failure, given $\mu$, is

$$F(w; \mu) = \text{Prob}(W \leq w) = \left( 1 - \exp \left\{ -\frac{w}{\mu} \right\} \right) 1(w > 0), \quad w \in \mathbb{R}.$$\(^{\text{1}}\)

\(^{\text{1}}\)The case of $r = 0$ is not considered, as it is not very interesting.
Therefore, if we evaluate the density for the observed \( Y_i \)'s, and take logarithms, we get

\[
\log L(Y_1, \ldots, Y_n; \mu) = \sum_{i=1}^{n} \log L(Y_i; \mu)
\]

\[
= \sum_{i=1}^{r} \log \frac{1}{\mu} \exp \left\{ -\frac{y_i}{\mu} \right\} + \sum_{i=r+1}^{n} \log \exp \left\{ -\frac{y_i}{\mu} \right\}
\]

\[
= -r \log \mu - \frac{1}{\mu} \sum_{i=1}^{n} y_i.
\]

It is then easy to see that the value of \( \mu \) that maximizes \( L(\mu) \) is

\[
\hat{\mu}_{ML} = \frac{1}{r} \sum_{i=1}^{n} y_i.
\]

We shall now employ the EM algorithm to solve this problem, and compare the EM iterates with the true ML estimate.

1. Recall that the exponential distribution is “memoryless” – for the censored observations \( W_{r+1}, \ldots, W_n \), the residual life time \( W_i - c_i \) is also exponentially distributed. If we denote by \( L_c(\mu) \) the likelihood of the complete observations, then the E-step is

\[
Q(\mu; \mu^{(k)}) = E_{\mu^{(k)}} [\log L_c(\mu) | Y_1, \ldots, Y_n]\n\]

\[
= \sum_{i=1}^{r} \log \frac{1}{\mu} \exp \left\{ -\frac{y_i}{\mu} \right\} + \sum_{i=r+1}^{n} E_{\mu^{(k)}} \left[ \frac{W_i}{\mu} \right] W_i > y_i
\]

\[
= -r \log \mu - \frac{1}{\mu} \sum_{i=1}^{r} y_i - \sum_{i=r+1}^{n} E_{\mu^{(k)}} \left[ \frac{W_i}{\mu} \right] W_i > y_i - (n - r) \log \mu
\]

\[
= -n \log \mu - \frac{1}{\mu} \sum_{i=1}^{r} y_i - \frac{1}{\mu} \sum_{i=r+1}^{n} y_i + \mu^{(k)}. \]

2. The M-step is to compute the value of \( \mu \) which maximizes

\[
Q(\mu, \mu^{(k)} = -n \log \mu - \frac{1}{\mu} \left[ (n - r)\mu^{(k)} + \sum_{i=1}^{n} y_i \right].
\]

Differentiating \( Q \) with respect to \( \mu \) and setting the derivative to zero, we obtain the maximizer to be

\[
\mu^{(k+1)} = \frac{1}{n} \left[ (n - r)\mu^{(k)} + \sum_{i=1}^{n} y_i \right].
\]

To compare this recursive EM solution with the true ML solution, rearrange the EM recursion as

\[
\mu^{(k+1)} - \hat{\mu}_{ML} = \frac{1}{n} \left[ (n - r)\mu^{(k)} + \sum_{i=1}^{n} y_i \right] - \frac{1}{r} \sum_{i=1}^{n} y_i
\]
\[
\begin{align*}
  &= \frac{n-r}{n} \mu^{(k)} + \frac{r-n}{nr} \sum_{i=1}^{n} y_i \\
  &= \left(1 - \frac{r}{n}\right)(\mu^{(k)} - \hat{\mu}_{ML}).
\end{align*}
\]

The following observations therefore apply to the EM iteration.

- The EM estimates \( \mu^{(k)} \) converge monotonically to the true ML estimate \( \hat{\mu}_{ML} \), as \( k \to \infty \). Thus the EM algorithm indeed finds asymptotically the global maximizer of the likelihood function. This, of-course, is possible because in this simple case the likelihood function has a unique extremum at the ML solution.

- The convergence is geometric, with rate \( \frac{r-x}{n} \). Thus the smaller amount of complete observations \( r \), the slower is the convergence to \( \hat{\mu}_{ML} \). This is generally true about EM techniques – the rate of convergence decreases with increasing incompleteness of the observed data.

**General Alternating Minimization Techniques**

Let \( \mathcal{P} \) and \( \mathcal{Q} \) represent two abstract sets. Throughout this section, the symbol \( P \) with any superscripts or subscripts shall represent members of \( \mathcal{P} \) and similarly \( Q \) of \( \mathcal{Q} \). Let \( d: \mathcal{P} \times \mathcal{Q} \to \mathbb{R} \cup \{+\infty\} \) be an extended real valued function. We shall use the notation

\[
P \xrightarrow{1} Q^* \quad \text{iff} \quad d(P, Q^*) = \min_{Q \in \mathcal{Q}} d(P, Q) \\
Q \xrightarrow{2} P^* \quad \text{iff} \quad d(P^*, Q) = \min_{P \in \mathcal{P}} d(P, Q),
\]

with the understanding that the “projection” is undefined whenever the minimum is not achieved. The superscript 1 (or 2) on the arrows may be used to remember that the first (or second) argument of \( d(\cdot, \cdot) \) is being held fixed in the minimization, as also to distinguish these symbol from the symbol \( \to \) for limits. Next, for \( A \subseteq \mathcal{P} \) and \( B \subseteq \mathcal{Q} \), define

\[
d(A, B) = \inf_{P \subseteq A, Q \subseteq B} d(P, Q)
\]

with the usual convention that the infimum is \(+\infty\) if either \( A \) or \( B \) is void. We are interested in techniques to find the minimum “distance” between the sets \( \mathcal{P} \) and \( \mathcal{Q} \) and in the points which achieve this minimum distance. More specifically, we shall study iterative techniques which find points \( P_n \in \mathcal{P} \), and \( Q_n \in \mathcal{Q} \), \( n = 0, 1, \ldots \), such that

\[
d(P_n, Q_n) \to d(\mathcal{P}, \mathcal{Q}).
\]

\( ^2 \)Adapted from “Information Geometry and Alternating Minimization,” by I. Csiszár and G. Tusnády (1984)
Additionally, we shall study conditions under which the iterates satisfy $P_n \to P^*$, and $Q_n \to Q^*$ such that
\[ d(P^*, Q^*) = d(P, Q), \]
where the convergence of the sequences is under suitable topologies on $\mathcal{P}$ and $\mathcal{Q}$.

The behavior of iterative techniques often depends on the starting point of the iteration, and this is true for the class of problems we shall consider as well. At the risk of jumping ahead of ourselves, define, for a given sequence $\{Q_n\}_{n=0}^{\infty}$ of points from $\mathcal{Q}$, a subset of points from $\mathcal{P}$ which are at a finite distance from the sequence, \textit{i.e.},
\[ \mathcal{P}_0 = \{P \in \mathcal{P} : d(P, Q_n) < \infty \text{ for some } n\}, \]
and observe that $\mathcal{P}_0$ depends on the entire $Q_n$ sequence. The reason we omit this dependence in our notation is that we shall later concentrate on iterative schemes which take the form
\[ P_0 \xrightarrow{1} Q_0 \xrightarrow{3} P_1 \xrightarrow{1} Q_1 \xrightarrow{3} P_2 \xrightarrow{1} \ldots, \]
and $\mathcal{P}_0$ shall be considered as defined with respect to this sequence of $Q_n$'s. Since the starting point $P_0$ determines the entire sequence of iterates, $\mathcal{P}_0$ shall then depend only on $P_0$, making our notation less ambiguous. But until such specifics are considered, $\mathcal{P}_0$ is defined as above for a given sequence of $Q_n$'s.

Our first result is that under very mild conditions on the sets $\mathcal{P}$ and $\mathcal{Q}$ and the function $d$, alternating minimization schemes indeed converge to the minimum "distance" between $\mathcal{P}$ and $\mathcal{Q}$. We begin by stating a technical lemma for sequences of real numbers which we shall use in the proof of the next theorem.

\textbf{Lemma 1: } Let $a_n > -\infty$, $b_n > -\infty$, $n = 0, 1, \ldots$, be a sequence of extended real numbers and $c$ a finite number such that

1. $c + b_{n-1} \geq b_n + a_n$, \quad $n = 1, 2, \ldots$,
2. $\limsup_{n \to \infty} b_n > -\infty$, and
3. $b_{n_0} < +\infty$ for some $n_0$.

Then
\[ \liminf_{n \to \infty} a_n \leq c. \tag{1} \]

If, in addition to the three conditions above, it holds that
\[ \cdot \sum_{n=1}^{\infty} (c - a_n)^+ < +\infty, \]

then we get
\[ \sum_{n=n_0+1}^{\infty} |c - a_n| < +\infty \] which immediately implies that \( \lim_{n \to \infty} a_n = c \). \hfill (2)

**Proof:** If \( \sum_{n=1}^{\infty} (c - a_n)^+ = +\infty \), it follows that \( a_n < c \) infinitely often, and \( \liminf_{n \to \infty} a_n \leq c \). Hence it suffices to prove the second part of the assertion. Now, note that if \( b_{n_0} \) is finite then, since
\[ c + b_{n_0} \geq b_{n_0+1} + a_{n_0+1}, \]
it follows that \( b_{n_0+1} \) and \( a_{n_0+1} \) are finite. By induction, therefore, \( a_n \) and \( b_n \) are finite for all \( n > n_0 \). Thus we can rearrange the inequality in the statement of the lemma to write
\[ a_n - c \leq b_{n-1} - b_n, \quad n = n_0 + 1, n_0 + 2, \ldots, \] (3)
and hence, for \( N > n_0 \) it holds that
\[ \sum_{n=n_0+1}^{N} a_n - c \leq b_{n_0} - b_N, \]
\[ \sum_{n=n_0+1}^{N} (a_n - c)^+ - \sum_{n=n_0+1}^{N} (c - a_n)^+ \leq b_{n_0} - b_N. \]
By assumption, the quantity on the left side has a well defined limit as \( N \to \infty \). Therefore it suffices to show that this limit is less than \( +\infty \), because once we do so, it follows that
\[ \sum_{n=n_0+1}^{\infty} |c - a_n| = \sum_{n=n_0+1}^{\infty} (c - a_n)^+ + \lim_{N \to \infty} \sum_{n=n_0+1}^{N} (a_n - c)^+ < +\infty. \]
But since the inequality (3) holds for every \( N > n_0 \) we get
\[ \lim_{N \to \infty} \sum_{n=n_0+1}^{N} a_n - c \leq \liminf_{N \to \infty} b_{n_0} - b_N \]
\[ = b_{n_0} - \limsup_{N \to \infty} b_N, \]
which is finite by assumption. This completes the proof of the lemma. \( \square \)

**Remark:** A qualitative interpretation of the lemma above is the following. If \( a_n \) is larger than \( c \) for some \( n \), but \( c + b_{n-1} \geq b_n + a_n \) due to a particularly negative \( b_n \), then for \( n + 1 \), \( b_n \) goes over to the left side of the inequality, forcing \( a_{n+1} \) to be dominated by \( c \), unless \( b_{n+1} \) too is heavily negative. However this cannot go on ad infinitum unless \( b_n \to -\infty \), which is ruled out by assumption! Thus, eventually, \( a_{n+m} \leq c \) for some \( m \). Infinitely long runs of \( a_n > c \) are therefore not possible, and there must be a subsequence of \( a_n \) which is dominated by \( c \). The finiteness of at least one \( b_{n_0} \) is needed because if all \( b_n \)'s are \( +\infty \), nothing interesting can be inferred about \( a_n \) from the inequality \( c + b_{n-1} \geq b_n + a_n \).
Theorem 1: For an arbitrary sequence \( \{P_n\}_{n=0}^{\infty} \) from \( P \) and \( \{Q_n\}_{n=0}^{\infty} \) from \( Q \), if it holds that
\[
d(P, Q) + d(P, Q_{n-1}) \geq d(P, Q_n) + d(P_n, Q_n), \quad n = 1, 2, \ldots,
\]
either for
1. every \( P \in P_0 \), or
2. some \( P \in P_0 \) such that \( d(P, Q) = d(P_0, Q) \),
then
\[
\lim_{n \to \infty} d(P_n, Q_n) = d(P_0, Q).
\]
Under the first hypothesis the sequence \( d(P_n, Q_n) \) decreases to \( d(P_0, Q) \) monotonically while under second hypothesis
\[
\sum_{n=n_1}^{\infty} (d(P_n, Q_n) - d(P_0, Q)) < +\infty,
\]
for some index \( n_1 \).

Proof: If \( P_0 = \phi \), then \( d(P_n, Q_n) = d(P_0, Q) = +\infty \) for every \( n \) and the assertion of the theorem is trivially true. If (4) holds for some \( P \in P_0 \), then the first of the three conditions of Lemma 1 holds with
\[
a_n = d(P_n, Q_n), \quad b_n = d(P, Q_n), \quad \text{and} \quad c = d(P, Q).
\]
Since \( P \in P_0 \), it must be true that \( d(P, Q_{n_0}) < \infty \) for some \( n_0 \), satisfying the third condition of the lemma, and it follows that \( c < \infty \) as well. Next, if we set \( n = n_0 + 1 \) in (4), it follows that \( c > -\infty \). Finally, \( b_n \geq c \) for every \( n \), and the second condition of the lemma is also satisfied.

1. Under our first hypothesis, the first assertion of Lemma 1 then implies that
\[
\liminf_{n \to \infty} d(P_n, Q_n) \leq d(P, Q), \quad P \in P_0.
\]
If \( d(P_n, Q_{n-1}) < \infty \) for some \( n \), then setting \( P = P_{n-1} \) in (4) gives
\[
d(P_{n-1}, Q_{n-1}) \geq d(P_n, Q_n).
\]
If \( d(P_{n-1}, Q_{n-1}) = \infty \), then this inequality is trivially true. Therefore, \( d(P_n, Q_n) \) is a nonincreasing sequence. Finally, by definition,
\[
d(P_n, Q_n) \geq d(P_0, Q), \quad n = 1, 2, \ldots,
\]
and combining this with the two preceding inequalities yields the assertion of the theorem under the first hypothesis.
2. Under our second hypothesis, the definition of $d(P_0, Q)$ implies that

$$d(P_n, Q_n) \geq d(P_0, Q) = d(P, Q), \quad n = 0, 1, \ldots,$$

which then provides that

$$\sum_{n=0}^{\infty} (c - a_n)^+ = 0 < +\infty.$$

The assertion of the theorem then follows from the second part of Lemma 1.

**Definition:** We shall say that the *five points property* holds for $P \in \mathcal{P}$ if

$$Q_0 \xrightarrow{\hat{\Delta}} P_1 \xrightarrow{\iota} Q_1 \quad \Rightarrow \quad d(P, Q) + d(P, Q_0) \geq d(P, Q_1) + d(P_1, Q_1), \quad Q \in \mathcal{Q}. \quad (5)$$

The five points property is inspired by the condition (4) in Theorem 1, and is in some sense the least requirement for an alternating minimization procedure to find the minimum “distance” between $\mathcal{P}$ and $\mathcal{Q}$. It, of course, must be checked in a particular problem of interest if the property (5) holds for any or every $P \in \mathcal{P}_0$.

We next define two other properties which imply the five points property. These are sometimes easier to check, as will be the case with the $\mathbf{I}$-divergence in the role of $d$. Let $\delta : \mathcal{P} \times \mathcal{P} \rightarrow \mathbb{R}^+$ be any function such that $\delta(P, P) = 0$ for every $P \in \mathcal{P}$.

**Definitions:** We shall say that the *three points property* holds for $P \in \mathcal{P}$ if

$$Q_0 \xrightarrow{\hat{\Delta}} P_1 \quad \Rightarrow \quad d(P, Q_0) \geq \delta(P, P_1) + d(P_1, Q_0). \quad (6)$$

We shall say that the *four points property* holds for $P \in \mathcal{P}$ if

$$P_1 \xrightarrow{\iota} Q_1 \quad \Rightarrow \quad d(P, Q) + \delta(P, P_1) \geq d(P, Q_1), \quad Q \in \mathcal{Q}. \quad (7)$$

Now, if $Q_0 \xrightarrow{\hat{\Delta}} P_1 \xrightarrow{\iota} Q_1$, and the three and four points properties hold for some $P \in \mathcal{P}$, then, adding the inequalities implied by (6) and (7) we get

$$d(P, Q_0) + d(P, Q) + \delta(P, P_1) \geq \delta(P, P_1) + d(P_1, Q_0) + d(P, Q_1), \quad Q \in \mathcal{Q}.$$

If $d(P, Q_0) < \infty$, then the three points property implies that $\delta(P, P_1) < \infty$ as well. In this case, we can subtract $\delta(P, P_1)$ from either side of the inequality above, observe that $d(P_1, Q_0) \geq d(P_1, Q_1)$ by construction of $Q_1$, and conclude that (6) and (7) imply the five point property. If $d(P, Q_0) = \infty$, (5) holds trivially.
**Theorem 2:** Let $P_0 \in \mathcal{P}$ be given. If $\{P_n\}_{n=0}^{\infty}$ and $\{Q_n\}_{n=0}^{\infty}$ are sequences obtained by alternating minimization of $d(P, Q)$ as $P_0 \xrightarrow{\sim} Q_0 \xrightarrow{\sim} P_1 \xrightarrow{\sim} Q_1 \xrightarrow{\sim} \ldots$, and if either

1. every $P \in \mathcal{P}_0$, or

2. some $P \in \mathcal{P}_0$ with $d(P, Q) = d(\mathcal{P}_0, Q)$

has the five points property (5), then

$$\lim_{n \to \infty} d(P_n, Q_n) = d(\mathcal{P}_0, Q).$$

Further, if the three and four points properties hold for some $P \in \mathcal{P}_0$ with $d(P, Q) = d(\mathcal{P}_0, Q)$, then (the second condition above is true, and) in addition to the convergence of $d$, we also have for this $P$ that

$$\delta(P, P_{n+1}) \leq \delta(P, P_n), \quad n = 0, 1, \ldots.$$

**Proof:** The first assertion of the theorem follows easily from the observation that the five points property implies the condition (4) and Theorem 1 applies. Next, if the three and four points properties hold for $P$ then, for every $Q \in \mathcal{Q}$, we get

$$Q_n \xrightarrow{\sim} P_{n+1} \quad \Rightarrow \quad \delta(P, P_{n+1}) + d(P_{n+1}, Q_n) \leq d(P, Q_n), \quad \text{ and }$$

$$P_n \xrightarrow{\sim} Q_n \quad \Rightarrow \quad d(P, Q_n) \leq d(P, Q) + \delta(P, P_n), \quad \text{ and thus}$$

$$\delta(P, P_{n+1}) + d(P_{n+1}, Q_n) \leq \delta(P, P_n) + d(P, Q).$$

Therefore, if $d(P, Q) = d(\mathcal{P}_0, Q)$, then $d(P_{n+1}, Q_n) \geq d(P, Q)$ and the second assertion of the theorem follows.

**Example:** The three and four points properties can be easily verified in the case when $\mathcal{P}$ and $\mathcal{Q}$ are closed convex subsets of $\mathbb{R}^2$, $d(P, Q) = ||P - Q||^2$, and $\delta(P, P') = ||P - P'||^2$. The three points property holds due to the convexity of $\mathcal{P}$, and the four property due to the convexity of $\mathcal{Q}$. Therefore the five points property holds for every $P \in \mathcal{P}$. Further, $d$ is finite valued and thus $\mathcal{P}_0 = \mathcal{P}$. Thus the alternating minimization procedure converges to the minimum squared-distance between the sets in this case.

**Alternating Minimization of I-Divergence**

We verify next that when $\mathcal{P}$ and $\mathcal{Q}$ are convex sets of finite measures, and $d$ is the I-divergence, the three and four points properties are satisfied for every $P \in \mathcal{P}$ with $\delta$ also
being an I-divergence. This will then let us conclude that the alternating minimization of 
\( D(P||Q) \) indeed solves the ML problem in the case of incomplete data. Recall that the two steps in that case are

- to project \( Q_n \in \mathcal{Q} \) on the set \( \mathcal{P} \), which is accomplished by computing and setting 
  \[ P_{n+1} = Q_n^{X|Y \hat{P}}, \]

- to project \( P_{n+1} \in \mathcal{P} \) on the set \( \mathcal{Q} \), which is the complete data ML estimation problem, 
  with \( P_{n+1} \) in the role of \( \hat{P} \) – the empirical distribution of the complete observations.

Therefore, a technique for solving the complete data problem can be used to solve the incomplete data problem. This is the general class of EM techniques, where the first of the steps above is called the *Expectation* computation, and the second the *Maximization* step.

**Lemma 2:** Let \( \mathcal{P} \) be a convex set of finite measures and let \( Q_0 \) be another finite measure, none identically zero, on \( \{X, \mathcal{F}\} \). Then

\[ Q_0 \rightarrow P_1 \quad \Rightarrow \quad D(P||P_1) + D(P_1||Q_0) \leq D(P||Q_0), \]

for every \( P \in \mathcal{P} \), where

\[ D(P||Q) = \begin{cases} 
\int \log \frac{dP}{dQ} dP + Q(X) - P(X) & \text{if } P << Q, \\
+\infty & \text{otherwise}. 
\end{cases} \]

**Proof:** This is the Pythagorean inequality for \( P_1 \), the I-projection of \( Q_0 \) on \( \mathcal{P} \).

**Lemma 3:** Let \( \mathcal{Q} \) be a convex set of finite measures and let \( P_1 \) be another finite measure, none identically zero, on \( \{X, \mathcal{F}\} \). Then

\[ P_1 \rightarrow Q_1 \quad \Rightarrow \quad D(P||Q_1) \leq D(P||P_1) + D(P||Q), \]

for every measure \( P \) on \( \{X, \mathcal{F}\} \) and every \( Q \in \mathcal{Q} \).

**Proof:** For a fixed \( Q \in \mathcal{Q} \), define

\[ Q_\alpha = \alpha Q_1 + (1 - \alpha) Q, \]

and note that \( Q_\alpha \in \mathcal{Q} \) for every \( 0 \leq \alpha \leq 1 \). We define

\[ g(\alpha) = D(P_1||Q_\alpha), \quad 0 < \alpha \leq 1, \]
and note that \(g(\alpha)\) attains its minimum at \(\alpha = 1\). We also denote by \(\overline{Q}\) and \(\overline{Q}_1\) the absolutely continuous component, with respect to \(P_1\), of \(Q\) and \(Q_1\) respectively, and write

\[
q = \frac{d\overline{Q}}{dP_1}, \quad q_1 = \frac{d\overline{Q}_1}{dP_1}.
\]

But \(D(P_1||Q_1) = D(P_1||Q) < +\infty\). Therefore \(P_1 << Q_1\), and consequently \(q_1 > 0, P_1\)-a.e. Therefore \(P_1 << Q_\alpha\) for every \(0 < \alpha \leq 1\), and since the density of the absolutely continuous part of \(Q_\alpha\) w.r.t. \(P_1\) is

\[
q_\alpha = \alpha q_1 + (1 - \alpha)q,
\]

we use the minimization property of \(\alpha = 1\) to write

\[
\frac{g(1) - g(\alpha)}{1 - \alpha} \leq 0,
\]

\[
\frac{Q_1(\mathcal{X}) - Q_\alpha(\mathcal{X})}{1 - \alpha} + \int \frac{1}{1 - \alpha} (- \log q_1 + \log q_\alpha) \, dP_1 \leq 0,
\]

\[
Q(\mathcal{X}) - Q_1(\mathcal{X}) + \int \left( \frac{\log q_1 - \log q_\alpha}{1 - \alpha} \right) \, dP_1 \geq 0.
\]

The integrand, however, is the difference quotient of the function \(\log q_\alpha\), which is concave in \(\alpha\) (because \(q_\alpha\) is linear in \(\alpha\)). Therefore the integrand converges monotonically, as \(\alpha \to 1\), to

\[
\left. \frac{d}{d\alpha} \log q_\alpha \right|_{\alpha = 1} = 1 - \frac{q}{q_1}.
\]

By the monotone convergence theorem, we therefore get

\[
Q(\mathcal{X}) - Q_1(\mathcal{X}) + \int \left( 1 - \frac{q}{q_1} \right) \, dP_1 \geq 0. \tag{8}
\]

Now, for a given \(Q \in Q\), the assertion of the lemma need be proved only for \(P\) such that \(P << P_1\), and \(P << Q\) (otherwise \(D(P||P_1)\), resp. \(D(P||Q)\), is \(+\infty\)). For such \(P\), it then follows that \(P << Q_1\). Further, if we denote

\[
p = \frac{dP}{dP_1}, \quad \text{then} \quad pqq_1 > 0, \quad P\text{-a.e.}
\]

and therefore

\[
\frac{dP}{dQ_1} = \frac{p}{q_1}, \quad \frac{dP}{dQ} = \frac{p}{q}.
\]

Therefore

\[
D(P||P_1) + D(P||Q) - D(P||Q_1)
\]

\[
= P_1(\mathcal{X}) + Q(\mathcal{X}) - Q_1(\mathcal{X}) - P(\mathcal{X}) + \int \left( \log p + \log \frac{p}{q} - \log \frac{p}{q_1} \right) \, dP
\]

14
\[ P_1(\mathcal{X}) + Q(\mathcal{X}) - Q_1(\mathcal{X}) - P(\mathcal{X}) + \int -\log \frac{q}{pq_1} dP \]
\[ \geq P_1(\mathcal{X}) + Q(\mathcal{X}) - Q_1(\mathcal{X}) - P(\mathcal{X}) + \int \left(1 - \frac{q}{pq_1}\right) dP \]
\[ = P_1(\mathcal{X}) + Q(\mathcal{X}) - Q_1(\mathcal{X}) - \int \frac{q}{q_1} dP_1 \]
\[ = Q(\mathcal{X}) - Q_1(\mathcal{X}) - \int \left(1 - \frac{q}{q_1}\right) dP_1, \]

where we have used the inequality \(-\log t \geq 1 - t\). By virtue of (8), the quantity above is nonnegative, and the assertion of the lemma follows. ■

From the last two lemmas, it follows that the alternating minimization of I-divergence between two convex sets \(\mathcal{P}\) and \(\mathcal{Q}\) converges. This is formally stated below.

**Theorem 3:** Let \(\mathcal{P}\) and \(\mathcal{Q}\) be convex sets of measures on \(\mathcal{X}, \mathcal{F}\) and \(\{P_n\}_{n=0}^{\infty}, \{Q_n\}_{n=0}^{\infty}\) be sequences obtained by alternating minimization of \(D(P||Q)\), starting from some \(P_0 \in \mathcal{P}\). Then

\[ \lim_{n \to \infty} D(P_n||Q_n) = D(P_0||Q). \]

Further, if \(\mathcal{X}\) is finite and \(\mathcal{P}\) and \(\mathcal{Q}\) are closed in the topology of pointwise convergence, then

\[ P_n \to P^*, \quad \text{such that} \quad D(P^*||Q) = D(P_0||Q). \]

**Proof:** The first assertion follows from Lemmas 2 and 3, which show that the three and four points properties hold for every \(P \in \mathcal{P}\), and from Theorem 2. Next, if \(\mathcal{X}\) is finite, then \(D(P||Q) < \infty\) iff \(P \ll Q\). Therefore, if \(\mathcal{P}\) is closed, then so is \(\mathcal{P}_0\). This, however, means that \(P_n\) has a convergent subsequence, say \(P_{N_k} \to P^*\). Similarly, thus, there is a further subsequence, say \(k_i\), for which \(Q_{N_{k_i}} \to Q^*\). Since \(D(P||Q)\) is lower semicontinuous in \(P, Q\), we therefore have

\[ D(P^*||Q^*) \leq \lim_{i \to \infty} D(P_{N_{k_i}}||Q_{N_{k_i}}) = D(P_0||Q), \]

where the strict inequality is impossible. Therefore

\[ D(P^*||Q) = D(P_0||Q). \]

By the second assertion of Theorem 2, therefore, \(D(P^*||P_n)\) is monotonically decreasing in \(n\) and the limit

\[ \lim_{n \to \infty} D(P^*||P_n) = D^*, \]

is well defined. However, \(P_{N_k} \to P^*\). Therefore \(D^*\) must necessarily equal zero and this, in turn, implies that \(P_n \to P^*\). ■
Maximum Likelihood from Incomplete Data (Revisited)

We are now in a position to formally state the technique for solving the maximum likelihood problem in the case of incomplete data provided a technique for solving the problem in the case of complete observations is known.

**Theorem 4(i):** Given a measure $\tilde{P}$ on $\{\mathcal{Y}, \mathcal{F}\}$, and a convex set $Q$ of measures on $\{\mathcal{X}, \mathcal{F}\}$, consider the set of measures on $\{\mathcal{X}, \mathcal{F}\}$ given by

$$\mathcal{P} = \{P : P^T = \tilde{P}\}.$$  

Starting from some $Q_0 \in Q$, let $Q_0 \xrightarrow{\mathcal{P}} P_1 \xrightarrow{\mathcal{Q}} Q_1 \xrightarrow{\mathcal{P}} P_2 \xrightarrow{\mathcal{Q}} \ldots$, be obtained by alternating minimization of $D(P||Q)$, $P \in \mathcal{P}$, $Q \in Q$, where the step $Q_{n-1} \xrightarrow{\mathcal{Q}} P_n$ is given by

$$\frac{dP_n}{dQ_{n-1}}(x) = \frac{d\tilde{P}}{dQ_{n-1}^T}(T(x)), \quad n = 1, 2, \ldots, \quad x \in \mathcal{X}.$$  

Then,

$$\lim_{n \to \infty} D(\tilde{P}||Q_n^T) = D(\tilde{P}||Q^T), \quad \iff \quad D(P_0||Q) = D(P||Q).$$

**Proof:** By construction of $P_n$

$$D(P_n||Q_{n-1}) = D(\tilde{P}||Q_{n-1}^T),$$

and the assertion of the theorem then follows from Theorem 3. \[\blacksquare\]

**Decomposition of Finite Mixtures by Alternating Minimization**

The alternating minimization technique is also useful for ML estimation of mixture coefficients (weights) from observed data when the mixture components are known. There is no “missing” data in this case – the modeler hypothesizes that the observations are i.i.d., and the source distribution is a mixture of some known components. But one can cast this as an incomplete data problem by viewing each mixture component as an individual i.i.d. source, and each observation as a “random sampling” from this collection of sources, where the information about the source from which an observation derives is missing. The mixture weight of a component is then the probability with which the corresponding source is selected for generating an observation.
This alternate view satisfies the underlying assumption of Theorem 4(i), namely, that given the complete observations, the ML estimation problem is “easy” to solve. Indeed, the ML estimate of the probability with which an individual source is selected is simply the relative frequency with which the source is selected! This is formally stated below.

**Lemma 4:** Let \( \{\mathcal{X}, \mathcal{F}\} = \{\mathcal{Z}, 2^\mathcal{Z}\} \times \{\mathcal{Y}, \tilde{\mathcal{F}}\} \), where \( 2^\mathcal{Z} \) is the collection of all subsets of \( \mathcal{Z} = \{1, \ldots, k\} \). Let \( \mu_1, \ldots, \mu_k \) be measures on \( \{\mathcal{Y}, \tilde{\mathcal{F}}\} \), and let \( \mathcal{R} \) be the set of all measures on \( \{\mathcal{X}, \mathcal{F}\} \) of the form

\[
R(\{i\}, A) = c_i \mu_i(A), \quad i \in \mathcal{Z}, \quad A \in \tilde{\mathcal{F}},
\]

for some \((c_1, \ldots, c_k) = \zeta \in \mathbb{P}^k\). For any given measure \( S \) on \( \{\mathcal{X}, \mathcal{F}\} \), with \( D(S||\mathcal{R}) < +\infty \), the I-projection \( S \rightharpoonup R^* \) on \( \mathcal{R} \) is uniquely determined by the choice of

\[
c_i^* = \frac{S(\{i\} \times \mathcal{Y})}{S(\mathcal{Z} \times \mathcal{Y})} = \frac{S(i)}{S(\mathcal{X})}.
\]

**Proof:** For \( R \in \mathcal{R} \), we have

\[
D(S||R) = \int_{\mathcal{X}} \log \left( \frac{dS}{dR} \right) dS = \sum_{i=1}^{k} S(i) \log \frac{S(i)}{c_i} + \sum_{i=1}^{k} S(i) \int_{\mathcal{Y}} \log \left( \frac{d\nu_i}{d\mu_i} \right) d\nu_i,
\]

where \( \nu_i, i \in \mathcal{Z} \), are measures on \( \{\mathcal{Y}, \tilde{\mathcal{F}}\} \) defined by

\[
\nu_i(A) = \frac{S(\{i\} \times A)}{S(i)}, \quad A \in \tilde{\mathcal{F}}.
\]

Since the only free variable for minimizing \( D(S||R) \) is \( \zeta \), the assertion follows. \( \blacksquare \)

**Theorem 5:** Let \( \tilde{P} \), and \( \mu_1, \ldots, \mu_k \) be measures on \( \{\mathcal{Y}, \tilde{\mathcal{F}}\} \), and let \( \mathcal{Q} \) be the set of all measures of the form

\[
\mathcal{Q}_\zeta = \sum_{i=1}^{k} c_i \mu_i, \quad \zeta \in \mathbb{P}^k.
\]

Suppose that \( D(\tilde{P}||\mathcal{Q}) < +\infty \). Starting from some vector \( \zeta^0 = (c_1^0, \ldots, c_k^0) \) with strictly positive components, let \( \zeta^n \) be defined as

\[
c_i^n = c_i^{n-1} \int \frac{d\mu_i}{d\mathcal{Q}_{n-1}} \frac{d\tilde{P}}{\tilde{P}(\mathcal{Y})}, \quad i = 1, \ldots, k, \quad n = 1, 2, \ldots.
\]

Then, \( \lim_{n \to \infty} \zeta^n = \zeta^* \) exists and

\[
\lim_{n \to \infty} D(\tilde{P}||\mathcal{Q}_{\zeta^n}) = D(\tilde{P}||\mathcal{Q}_\zeta^*) = D(\tilde{P}||\mathcal{Q}).
\]

17
Further,
\[ D(\tilde{P}||\tilde{Q}_n) - D(\tilde{P}||\tilde{Q}) \leq \max_{1 \leq i \leq k} \log \frac{c_i^{n+1}}{c_i^n}. \]

**Proof:** Let \( \{X, \mathcal{F}\} = \{Z, 2^Z\} \times \{Y, \tilde{\mathcal{F}}\} \), where \( Z = \{1, \ldots, k\} \), and let \( T : X \to Y \) be the coordinate map from \( X \) to \( Y \). Let \( Q \) be the set of measures on \( X \) of the form
\[ Q_\mathcal{G}(\{i\} \times A) = c_i \mu_i(A), \quad i \in Z, \quad A \in \tilde{\mathcal{F}}, \quad \text{where } \mathcal{G} \in \mathcal{P}^k, \]
and note that \( Q_\mathcal{G}^T \) equals \( \mathcal{G} \) defined above. Let \( P \) be the set of all measures on \( X \) with \( Y \)-marginals equal to \( \tilde{P} \). The \( E \)-step \( Q_{\mathcal{F}^{n-1}} \rightarrow P_n \) is given, as in Theorem 4(i), by
\[ \frac{dP_n}{dQ_{\mathcal{F}^{n-1}}}(i, y) = \frac{d\tilde{P}}{d\tilde{Q}_{\mathcal{F}^{n-1}}}(y), \quad i \in Z, \quad y \in Y, \]
and the \( M \)-step \( P_n \rightarrow Q_{\mathcal{G}^n} \) is determined as in Lemma 4 by
\[ c_i^n = \frac{P_n(\{i\} \times Y)}{P(Y)}, \quad i \in Z. \]
Combining the two steps above, we obtain
\[ c_i^n = \int_Y \frac{dP_n}{P(Y)} \frac{1}{P(Y)} \int_Y \frac{d\tilde{P}}{d\tilde{Q}_{\mathcal{F}^{n-1}}} dQ_{\mathcal{F}^{n-1}} = \int_Y \frac{c_i^{n-1} d\mu_i}{P(Y)} \frac{d\tilde{P}}{d\tilde{Q}_{\mathcal{F}^{n-1}}}, \]
which is the recursion for \( \mathcal{G}^n \) in the statement of the theorem. Thus the convergence of \( D(\tilde{P}||\tilde{Q}_n) \) follows from Theorem 4(i) provided we show that \( D(P_0||Q) = D(P||Q) \). This, in turn, follows if we show that \( P_0 = P \). We shall show next that for every \( P \in \mathcal{P} \), \( D(P||Q_{\mathcal{G}}) < \infty \). For \( Q_{\mathcal{G}} \in \mathcal{Q} \),
\[ Q_{\mathcal{G}} \ll Q_{\mathcal{G}^n} \quad \text{iff} \quad c_i = 0 \quad \Rightarrow \quad c_i^n = 0. \]
But for all such \( Q_{\mathcal{G}} \), we can then write
\[ \frac{dQ_{\mathcal{G}}}{dQ_{\mathcal{G}^n}}(i, y) = \frac{c_i}{c_i^n}. \]
This implies that
\[ D(P||Q_{\mathcal{G}^n}) = \int \log \frac{dP}{dQ_{\mathcal{G}}} dQ_{\mathcal{G}} dP = D(P||Q_{\mathcal{G}}) + \sum_{i=1}^k P(i) \log \frac{c_i}{c_i^n}. \] (9)
Now, by assumption, \( D(\tilde{P}||\tilde{Q}) < +\infty \). Therefore, for every \( P \in \mathcal{P} \), \( D(P||Q_{\mathcal{G}}) < +\infty \) for some \( Q_{\mathcal{G}} \in \mathcal{Q} \). If we let this \( Q_{\mathcal{G}} \) be the measure in (9) above, and note that \( c_0 > 0 \), then it follows that \( D(P||Q_{\mathcal{G}}) < +\infty \), as claimed.
To show the convergence of the sequence $\mathcal{Q}^n$, note that $D(\tilde{P}||\sum_i c_i \mu_i)$ is lower semicontinuous in $\mathcal{Q}$ on $\mathbb{P}^k$ (it is continuous except possibly at the boundary of the simplex). Since $\mathbb{P}^k$ is compact, consider a convergent subsequence $\mathcal{Q}_{N_k}^n \rightarrow \mathcal{Q}^*$, and let $P^* \in \mathcal{P}$ be constructed according to the E-step $Q_{\mathcal{V}^*} \rightarrow P^*$. As in the proofs of Theorems 3 and 4(i),

$$D(\mathcal{P}||\mathcal{Q}) = D(P^*||\mathcal{Q}_{\mathcal{V}^*}) = D(\tilde{P}||\tilde{Q}_{\mathcal{V}^*}) = D(\tilde{P}||\tilde{Q}),$$

where the first equality comes from the convergence of $D(P^n||\mathcal{Q}_{\mathcal{V}^n})$, the second from the construction of $P^*$ and $\mathcal{P}$, and the third from lower semicontinuity. Thus $Q_{\mathcal{V}^*}$ also satisfies the condition $P^* \rightarrow Q_{\mathcal{V}^*}$, i.e.,

$$c_i^* = \frac{P^*(i)}{\tilde{P}(\mathcal{V})}.$$ 

Therefore, for every $n$, the partition inequality for I-divergence gives

$$D(P^*||P_n) \geq \sum_{i=1}^k P^*(i) \log \frac{P^*(i)}{P_n(i)} = \tilde{P}(\mathcal{V}) \sum_{i=1}^k c_i^* \log \frac{c_i}{c_i^*} = \tilde{P}(\mathcal{V})D(\mathcal{Q}^*||\mathcal{Q}^n).$$

Further, by the three points property and (9),

$$D(P^*||P_n) + D(P_n||Q_{\mathcal{V}^n-1}) \leq D(P^*||Q_{\mathcal{V}^n-1}) = D(P^*||Q_{\mathcal{V}^n}) + \sum_{i=1}^k P^*(i) \log \frac{c_i^*}{c_i^{n-1}} = D(P^*||Q_{\mathcal{V}^n}) + \tilde{P}(\mathcal{V})D(\mathcal{Q}^*||\mathcal{Q}^{n-1}),$$

and thus

$$\tilde{P}(\mathcal{V})D(\mathcal{Q}^*||\mathcal{Q}^n) + D(P_n||Q_{\mathcal{V}^n-1}) \leq D(P^*||Q_{\mathcal{V}^n}) + \tilde{P}(\mathcal{V})D(\mathcal{Q}^*||\mathcal{Q}^{n-1}).$$

Combining this with the minimality of $D(P^*||Q_{\mathcal{V}^n})$ shows that $D(\mathcal{Q}^*||\mathcal{Q}^n)$ is a decreasing sequence which is bounded below, and invoking $\mathcal{Q}_{N_k}^n \rightarrow \mathcal{Q}^*$ shows that $D(\mathcal{Q}^*||\mathcal{Q}^n) \rightarrow 0$.

The last inequality also provides the final (rate of convergence) result. By rearrangement,

$$D(P_n||Q_{\mathcal{V}^n-1}) - D(P^*||Q_{\mathcal{V}^n}) \leq D(\mathcal{Q}^*||\mathcal{Q}^{n-1}) - D(\mathcal{Q}^*||\mathcal{Q}^n)$$

$$D(\tilde{P}||\tilde{Q}_{\mathcal{V}^n-1}) - D(\tilde{P}||\tilde{Q}) = \sum_{i=1}^k P^*(i) \log \frac{c_i^n}{c_i^{n-1}},$$

which is bounded above by the largest logarithmic term.

\[ \blacksquare \]

**Revisiting the EM Procedure: An Analytical View**

The commonly presented analysis of the EM procedure does not take the information geometric view that has been presented for the more general alternating minimization procedure. What follows is perhaps more in keeping with the usual presentation.
Given the observations $\mathbf{y} = \mathbf{y}$, let

$$k(x|y; \Psi) = \frac{g_c(x; \Psi)}{g(y; \Psi)}, \quad x \in \mathcal{X}, \quad y \in \mathcal{Y},$$

denote the conditional density of the complete data. Then,

$$\log L(\Psi) = \log g(y; \Psi) = \log g_c(x; \Psi) - \log k(X|y; \Psi)$$

with probability one under any measure (depending on $\Psi$) for which the random variable $\log g_c(X; \Psi) = \log L_c(\Psi)$ is finite a.s. Further,

$$\log L(\Psi) = E_{\Psi(k)}[\log L_c(\Psi)|Y = y] - E_{\Psi(0)}[\log k(X|y; \Psi)|Y = y]$$

$$= Q(\Psi; \Psi^{(k)}) - H(\Psi; \Psi^{(k)}),$$

provided the conditional expectation of $\log L_c(\Psi)$ is also finite under the measure given by the density $g_c(x; \Psi^{(k)})$, where

$$H(\Psi; \Psi^{(k)}) = E_{\Psi(k)}[\log k(X|y; \Psi)|Y = y].$$

Now,

$$H(\Psi; \Psi^{(k)}) - H(\Psi^{(k)}; \Psi^{(k)}) = E_{\Psi(k)} \left[ \log \frac{k(X|y; \Psi)}{k(X|y; \Psi^{(k)})}\bigg| Y = y \right]$$

$$\leq \log E_{\Psi(k)} \left[ \frac{k(X|y; \Psi)}{k(X|y; \Psi^{(k)})}\bigg| Y = y \right]$$

$$= \log \int_{x: T(x) = y} \frac{g_c(x; \Psi)}{g_c(x; \Psi^{(k)})} k(x|y; \Psi^{(k)}) \, dx$$

Therefore, if $\Psi^{(k+1)}$ is chosen to maximize $Q(\Psi; \Psi^{(k)})$ as a function of $\Psi$, then

$$L(\Psi^{(k+1)}) - L(\Psi^{(k)})$$

$$= Q(\Psi^{(k+1)}; \Psi^{(k)}) - Q(\Psi^{(k)}; \Psi^{(k)}) - \left( H(\Psi^{(k+1)}; \Psi^{(k)}) - H(\Psi^{(k)}; \Psi^{(k)}) \right)$$

$$\geq Q(\Psi^{(k+1)}; \Psi^{(k)}) - Q(\Psi^{(k)}; \Psi^{(k)})$$

$$\geq 0.$$

Therefore, each iteration of the EM algorithm results in an updated parameter value for which the likelihood of the observed $\mathbf{y}$ is no less than that under the previous value of the parameter. If the improvement in the $Q$-function during the M-step is strictly positive, then the improvement in the likelihood of the data is strictly positive as well. Therefore, if the likelihood function is bounded above, then $L(\Psi^{(k)})$ converges as $k \to \infty$. 

20
Observe that to get an increase in the likelihood from step to step, all we need to do is choose $\Psi^{(k+1)}$ in such a manner that $Q(\Psi^{(k+1)}; \Psi^{(k)}) \geq Q(\Psi^{(k)}; \Psi^{(k)})$. Techniques which relax the maximization operation in the EM algorithm in this manner are called Generalized EM (GEM) algorithms or techniques.

The abstract formulation of the problem we have so far examined leaves two important issues unanswered even after this illustration of the connection between the $Q$-function improvement by the EM algorithm and the improvement in the likelihood function:

- Does $L(\Psi^{(k)})$ converge to a local maximum (or to a local minimum or a point of inflection) of $L$? Under what conditions does it converge to the global maximum?

- When does the convergence of the likelihood value $L(\Psi^{(k)})$ guarantee the convergence of the parameters $\Psi^{(k)}$?

Stronger assumptions about the structure of the likelihood function and the parameter space $\Omega$ must be made to answer these questions.

**The Csiszár-Tusnády Conditions**

Observe that the two steps of the EM algorithm are precisely the two steps in the alternating minimization procedure in Theorem 4(i) of the Csiszár-Tusnády (1984) paper discussed earlier. Therefore, if $g_c$ and $\Omega$ are such that for some fixed $\sigma$-finite measure $\mu$ the family of measures given by

$$Q = \{Q : Q \text{ has density } g_c(\cdot ; \Psi) \text{ w.r.t. } \mu \text{ for some } \Psi \in \Omega\},$$

is convex then it follows easily, from the abovementioned theorem, that the EM algorithm converges to the global maximizer of the likelihood, and that the iterates also converge to the corresponding ML estimate.

**Other Conditions for Convergence to Local Extrema$^3$**

Weaker convergence results can be obtained under different and slightly weaker conditions for the EM algorithm. In particular, if we assume that

- the parameters are real valued, i.e., $\Omega \subset \mathbb{R}^d$,

---

$^3$Extensive literature exists on the convergence properties of the EM algorithm for various applications. For a good starting point, see *The EM Algorithm and Extensions*, by G. J. McLachlan and T. Krishnan (1997). This section is adapted from Chapter 3 of this book.
for every $\Psi^{(0)} \in \Omega$ such that $\log L(\Psi^{(0)}) > -\infty$, the set of possible GEM iterates given by

$$\Omega_{\Psi^{(0)}} = \left\{ \Psi \in \Omega : L(\Psi) \geq L(\Psi^{(0)}) \right\},$$

is compact, and

- the likelihood function $L(\Psi)$ for the observed $y$ is continuous on $\Omega$ and differentiable in the interior of $\Omega$,

then the following statements can be made about the behavior of the GEM or the EM algorithms.

**Theorem 3.1:** Let $\{\Psi^{(k)}$, $k = 0, 1, \ldots\}$, be an instance of a GEM algorithm generated by $\Psi^{(k+1)} \in \{\Psi : Q(\Psi; \Psi^{(k)}) \geq Q(\Psi^{(k)}; \Psi^{(k)})\}$. Let $S$ denote the set of stationary points of the algorithm in the interior of $\Omega$. Suppose that

- $L(\Psi^{(k+1)}) > L(\Psi^{(k)})$ whenever $\Psi^{(k)} \notin S$, and

- the point-to-set mapping $\Omega_{\Psi^{(k)}}$ is closed over the complement of $S$.

Then, all limit points of $\Psi^{(k)}$ are stationary points of $L(\Psi)$, and for some $\Psi^* \in S$ we get that

$$\lim_{k \to \infty} L(\Psi^{(k)}) = L(\Psi^*).$$

**Remark:** We say that a point-to-set map $f : \mathcal{X} \rightarrow 2^Y$ is closed if

$$x_n \rightarrow x^*, \quad \text{and} \quad y_n \in f(x_n) \text{ such that } y_n \rightarrow y^* \implies y^* \in f(x^*).$$

**Theorem 3.2:** Let $\{\Psi^{(k)}$, $k = 0, 1, \ldots\}$, be an instance of a EM algorithm generated by $\Psi^{(k+1)} = \arg \max_{\Psi \in \Omega} Q(\Psi; \Psi^{(k)})$. Let $S$ denote the set of stationary points of the algorithm in the interior of $\Omega$. Suppose that

- the function $Q(\Psi; \Phi)$ is continuous in both $\Psi$ and $\Phi$.

Then, all limit points of $\Psi^{(k)}$ are stationary points of $L(\Psi)$, and for some $\Psi^* \in S$ we get that

$$\lim_{k \to \infty} L(\Psi^{(k)}) = L(\Psi^*),$$

where the convergence is monotone.