19 The Information Geometry of ML and EM

In the previous section, we introduced the EM algorithm as an iterative procedure for obtaining maximum likelihood (ML) and maximum a posteriori (MAP) estimates. Our derivation was both straightforward and fairly general, handling both discrete- and continuous-valued observations. Here we develop an information geometric view of the EM algorithm. In particular, it is possible to view the EM algorithm as corresponding to an alternating projections algorithm on the probability simplex. Specifically, the EM algorithm is equivalent to an alternating divergence minimization procedure, as we now develop.

19.1 Maximum Likelihood as a Reverse I-Projection

For convenience, we focus on the case of independent, identically distributed (i.i.d.) discrete data and a nonrandom parameter. In particular, our observations (the "incomplete data") are the i.i.d. discrete-valued sequence

$$\mathbf{y} = \begin{bmatrix} y_1 & y_2 & \dots & y_N \end{bmatrix}^{\mathrm{T}}. \tag{1}$$

As usual, we let $\mathcal{Y}$ denote the alphabet for the observations, and let $p_y(\cdot; x)$ be our model for each y_n, so we can write the distribution for $\mathbf{y}$ as

$$p_{\mathbf{y}}(\mathbf{y}; x) = \prod_{n=1}^{N} p_y(y_n; x), \tag{2}$$

Our goal is to produce the ML estimate of x based on (2), or equivalently based on the corresponding log-likelihood

$$\frac{1}{N} \log p_{\mathbf{y}}(\mathbf{y}; x) = \frac{1}{N} \sum_{n=1}^{N} \log p_y(y_n; x). \tag{3}$$

With such i.i.d. models (2), the ordering of values in the observed sequence evidently carries no information about the unknown parameter. Thus, only the relative frequency of occurrence of symbols in the alphabet $\mathcal{Y}$ is important. Phrased differently, a sufficient statistic for our model is the *empirical distribution* or *type* of the data, defined via

$$\hat{p}_y(b; \mathbf{y}) = \frac{1}{N} \sum_{n=1}^{N} \mathbb{1}_b(y_n), \quad b \in \mathcal{Y}. \tag{4}$$

Evidently, $\hat{p}_y(b; \mathbf{y})$ in (4) is the fraction of times the symbol b appears in the sequence $\mathbf{y}$.

Without loss of generality, we let $\mathcal{Y} = \{1, 2, \ldots, M\}$, and observe that this sufficient statistic can be equivalently written as the M-dimensional vector

$$\mathbf{t} = \begin{bmatrix} \hat{p}_y(1;\mathbf{y}) & \hat{p}_y(2;\mathbf{y}) & \ldots & \hat{p}_y(M;\mathbf{y}) \end{bmatrix}^{\mathrm{T}}, \tag{5}$$

where we note that the last element of the vector is technically redundant, so could be dropped, but we retain it for convenience.

As a sufficient statistic, $\mathbf{t}$ is as good as the original data for the purpose of inference, so

$$\hat{x}_{\mathrm{ML}}(\mathbf{y}) = \arg\max_a p_{\mathbf{y}}(\mathbf{y}; a) = \arg\max_a p_{\mathbf{t}}(\mathbf{t}; a). \tag{6}$$

The following general identity involving empirical distributions will be useful in our development.

Fact 1. *For any deterministic sequence* $\mathbf{v} = [v_1, \ldots, v_N]$ *whose elements are drawn from alphabet* $\mathcal{V}$*, and any function* $f(\cdot)$ *whose domain is* $\mathcal{V}$*, we have*

$$\frac{1}{N}\sum_{n=1}^{N} f(v_n) = \sum_{v\in\mathcal{V}} f(v)\, \hat{p}(v;\mathbf{v}) \tag{7}$$

where

$$\hat{p}(v;\mathbf{v}) = \frac{1}{N}\sum_{n=1}^{N} \mathbb{1}_{v_n}(v), \quad v \in \mathcal{V}$$

is the empirical distribution of the sequence $\mathbf{v}$*.*

Proof. It suffices to note

$$\frac{1}{N}\sum_{n=1}^{N} f(v_n) = \frac{1}{N}\sum_{n=1}^{N}\left(\sum_{v\in\mathcal{V}} \mathbb{1}_{v_n}(v) f(v)\right) = \sum_{v\in\mathcal{V}} f(v)\left(\frac{1}{N}\sum_{n=1}^{N}\mathbb{1}_{v_n}(v)\right),$$

where the first equality is due to the sifting property of the Kronecker function, and where the second equality results from interchanging the order of summations. □

Now we have the following geometric view of ML estimation in this i.i.d. case.

Fact 2. *Let the set of models be*

$$\mathcal{P} = \{p_y(\cdot; x),\ x \in \mathcal{X}\}. \tag{8}$$

Then the ML estimate can be written in the form

$$\hat{x}_{\mathrm{ML}}(\mathbf{y}) = \arg\min_{p\in\mathcal{P}} D(\hat{p}_y(\cdot;\mathbf{y}) \| p) \tag{9}$$

$$= \arg\min_a D(\hat{p}_y(\cdot;\mathbf{y}) \| p_y(\cdot; a)), \tag{10}$$

which is termed a reverse I-projection.[1]

[1] Although we don't develop the concept here, such reverse I-projections also obey their own version of "Pythagoras' theorem."

Proof. To verify (10), it suffices to expand (3) according to

$$\begin{aligned}\frac{1}{N}\log p_{\mathbf{y}}(\mathbf{y};x) &= \frac{1}{N}\sum_{n=1}^{N}\log p_{y}(y_n;x)\\ &= \sum_{b\in\mathcal{Y}}\hat{p}_{y}(b;\mathbf{y})\log p_{y}(b;x) \qquad (11)\\ &= -\sum_{b\in\mathcal{Y}}\hat{p}_{y}(b;\mathbf{y})\log\frac{\hat{p}_{y}(b;\mathbf{y})}{p_{y}(b;x)} + \sum_{b\in\mathcal{Y}}\hat{p}_{y}(b;\mathbf{y})\log\hat{p}_{y}(b;\mathbf{y})\\ &= -D(\hat{p}_{y}(\cdot;\mathbf{y})\|p_{y}(\cdot;x)) - H(\hat{p}(\cdot;\mathbf{y})), \qquad (12)\end{aligned}$$

where to obtain (11) we have used Fact 1, and where we note that the second term in (12) does not depend on x. □

19.2 Estimate-Maximize as Alternating Projections

First, recall that the EM algorithm comprises the following steps:

E-step Compute $U(x,\hat{x}^{(n-1)})$ where

$$U(x,x') = \sum_{\mathbf{z}\in\mathcal{Z}^N} p_{\mathbf{z}|\mathbf{y}}(\mathbf{z}|\mathbf{y};x')\log p_{\mathbf{z}}(\mathbf{z};x) \qquad (13)$$

M-step Compute

$$\hat{x}^{(n)} = \arg\max_{x} U(x,\hat{x}^{(n-1)}). \qquad (14)$$

In our case of interest, in addition to our observed ("incomplete") data (1), our "complete data"

$$\mathsf{z} = \begin{bmatrix}\mathsf{z}_1 & \mathsf{z}_2 & \dots & \mathsf{z}_N\end{bmatrix}^{\mathrm{T}},$$

is also i.i.d. and discrete-valued, with elements are drawn from the alphabet $\mathcal{Z}$ and distributed according to $p_{\mathsf{z}}(\cdot;x)$. Furthermore, the observations are obtained from the complete data via $y_n = g(z_n)$, where $g(\cdot)$ is a deterministic, many-to-one (i.e., noninvertible) mapping from $\mathcal{Z}$ to $\mathcal{Y}$.

In this case, (13) specializes to

$$U(x,x') = \sum_{n=1}^{N}\sum_{c\in\mathcal{Z}} p_{z|y}(c|y_n;x')\,\log p_{z}(c;x), \qquad (15)$$

where

$$p_{\mathbf{z}|\mathbf{y}}(\mathbf{z}|\mathbf{y};x) = \prod_{n=1}^{N} p_{z|y}(z_n|y_n;x). \qquad (16)$$

To evaluate (15), we use

$$p_{\mathsf{z}|\mathsf{y}}(z|y;x) = \frac{p_{\mathsf{z},\mathsf{y}}(z,y;x)}{p_{\mathsf{y}}(y;x)} = \frac{p_{\mathsf{y}|\mathsf{z}}(y|z)\, p_{\mathsf{z}}(z;x)}{p_{\mathsf{y}}(y;x)} = \frac{p_{\mathsf{z}}(z;x)}{p_{\mathsf{y}}(y;x)}\, \mathbb{1}_{z\in\mathcal{G}(y)} \tag{17}$$

with

$$\mathcal{G}(y) = \{z\colon g(z) = y\}, \tag{18}$$

since

$$p_{\mathsf{y}|\mathsf{z}}(y|z) = \mathbb{1}_{z\in\mathcal{G}(y)}. \tag{19}$$

In addition, we have, again using (19)

$$p_{\mathsf{y}}(y;x) = \sum_z p_{\mathsf{z},\mathsf{y}}(z,y;x) = \sum_z p_{\mathsf{y}|\mathsf{z}}(y|z)\, p_{\mathsf{z}}(z;x) = \sum_{z\in\mathcal{G}(y)} p_{\mathsf{z}}(z;x). \tag{20}$$

From Fact 2, we seek x such that $p_{\mathsf{y}}(\cdot;x)$ minimizes

$$D(\hat{p}_{\mathsf{y}}(\cdot;\mathbf{y})\|p_{\mathsf{y}}(\cdot;x)),$$

which by assumption is a difficult optimization. By contrast, by assumption, with our choice of complete data $\mathbf{z}$, finding x such that $p_{\mathsf{z}}(\cdot;x)$ minimizes

$$D(\hat{p}_{\mathsf{z}}(\cdot;\mathbf{z})\|p_{\mathsf{z}}(\cdot;x)) \tag{21}$$

is a comparatively easy optimization.

Of course, since the realization $\mathbf{z}$ of our complete data is fictitious, we have some flexibility in our choice of $\hat{p}_{\mathsf{z}}(\cdot;\mathbf{z})$ in (21). In particular, the set of possible empirical distributions over $\mathcal{Z}$ is

$$\hat{\mathcal{P}}^{\mathcal{Z}}(\mathbf{y}) \triangleq \Big\{\hat{p}_{\mathsf{z}}(\cdot)\colon \sum_{c\in\mathcal{G}(b)} \hat{p}_{\mathsf{z}}(c) = \hat{p}_{\mathsf{y}}(b;\mathbf{y}) \text{ for all } b\in\mathcal{Y}\Big\}. \tag{22}$$

In fact, it turns out that a particular element of (22) will be of interest, viz.,

$$\hat{p}^*_{\mathsf{z}}(\cdot;x) = \operatorname*{arg\,min}_{\hat{p}_{\mathsf{z}}\in\hat{\mathcal{P}}^{\mathcal{Z}}(\mathbf{y})} D(\hat{p}_{\mathsf{z}}(\cdot)\|p_{\mathsf{z}}(\cdot;x)) \tag{23}$$

for a given value of x.

The relevance of (23) follows from the following claim, whose proof we postpone for the moment

Claim 1. *The empirical distribution $\hat{p}^*_{\mathsf{z}}(\cdot;x)$ in (23) can be expressed in the form*

$$\hat{p}^*_{\mathsf{z}}(\cdot;x) = \frac{p_{\mathsf{z}}(z;x)\,\hat{p}_{\mathsf{y}}(g(z))}{p_{\mathsf{y}}(g(z);x)}. \tag{24}$$

As a consequence of (24), we have that for corresponding values of x, the conditional empirical distribution associated with (23) matches with the actual conditional distribution, i.e.,

$$\hat{p}^*_{\mathsf{z}|\mathsf{y}}(z|g(z);x) = \frac{\hat{p}^*_{\mathsf{z},\mathsf{y}}(z,g(z))}{\hat{p}_{\mathsf{y}}(g(z))} = \frac{\hat{p}^*_{\mathsf{z}}(z;x)}{\hat{p}_{\mathsf{y}}(g(z))} = \frac{p_{\mathsf{z}}(z;x)}{p_{\mathsf{y}}(g(z);x)} = p_{\mathsf{z}|\mathsf{y}}(z|g(z);x).$$

Hence, evaluating (15), we have

$$\begin{aligned}
\frac{1}{N}U(x,x') &= \frac{1}{N}\sum_{n=1}^{N}\sum_{z\in\mathcal{G}(y)} p_{\mathsf{z}|\mathsf{y}}(z|y_n;x')\log p_{\mathsf{z}}(z;x) \\
&= \sum_{y}\hat{p}_{\mathsf{y}}(y)\sum_{z\in\mathcal{G}(y)} p_{\mathsf{z}|\mathsf{y}}(z|y;x')\log p_{\mathsf{z}}(z;x) && (25)\\
&= \sum_{y}\hat{p}_{\mathsf{y}}(y)\sum_{z\in\mathcal{G}(y)} \hat{p}^*_{\mathsf{z}|\mathsf{y}}(z|y;x')\log p_{\mathsf{z}}(z;x) && (26)\\
&= \sum_{y}\sum_{z\in\mathcal{G}(y)} \hat{p}^*_{z}(z;x)')\log p_{\mathsf{z}}(z;x) && (27)\\
&= \sum_{z} \hat{p}^*_{\mathsf{z}}(z;x')\log p_{\mathsf{z}}(z;x) \\
&= -D(\hat{p}^*_{\mathsf{z}}(\cdot;x')\|p_{\mathsf{z}}(\cdot;x)) - H(\hat{p}^*_{\mathsf{z}}(z;x')), && (28)
\end{aligned}$$

where to obtain (25) we have used Fact 1, and where to obtain (26) and (27) we have used (19.2). Thus,

$$\arg\max_x U(x,x') = \arg\min_x D(\hat{p}^*_{\mathsf{z}}(\cdot;x')\|p_{\mathsf{z}}(\cdot;x)) \tag{29}$$

Hence, combining (23) and (29), we can rewrite the EM algorithm in the equivalent form:

E-step Compute

$$\hat{p}^*_{\mathsf{z}}(\cdot;\hat{x}^{(n-1)}) = \arg\min_{\hat{p}_{\mathsf{z}}\in\hat{\mathcal{P}}^{\mathsf{z}}(\mathbf{y})} D(\hat{p}_{\mathsf{z}}(\cdot)\|p_{\mathsf{z}}(\cdot;\hat{x}^{(n-1)})).$$

M-step Compute

$$\hat{x}^{(n)} = \arg\min_x D(\hat{p}^*_{\mathsf{z}}(\cdot;\hat{x}^{(n-1)})\|p_{\mathsf{z}}(\cdot;x)).$$

From this characterization, it follows that if $\hat{\mathcal{P}}^{\mathsf{z}}(\mathbf{y})$ and the set of models $\mathcal{P}$ [cf. (8)] are convex, then the EM algorithm will converge to the the ML estimate. However, it is straightforward to verify that $\hat{\mathcal{P}}^{\mathsf{z}}(\mathbf{y})$ is always convex, so only $\mathcal{P}$ must be tested for the problem of interest.

It remains only to show Claim 1, which is an immediate consequence of the decision form of the data processing inequality.

Lemma 1 (Data Processing Inequality—Decision Form). [2] *Let $\mathcal{Y}$ and $\mathcal{Z}$ be two alphabets, let $g\colon \mathcal{Z} \mapsto \mathcal{Y}$ be an arbitrary mapping, and let $p_{\mathsf{y}}, q_{\mathsf{y}} \in \mathcal{P}^{\mathcal{Y}}$ be the (respective) distributions induced by arbitrary distributions $p_{\mathsf{z}}, q_{\mathsf{z}} \in \mathcal{P}^{\mathcal{Z}}$ via the mapping g. Then*

$$D(p_{\mathsf{z}} \| q_{\mathsf{z}}) \geq D(p_{\mathsf{y}} \| q_{\mathsf{y}}), \tag{30}$$

with equality if and only if

$$\frac{p_{\mathsf{z}}(z)}{q_{\mathsf{z}}(z)} = \frac{p_{\mathsf{y}}(g(z))}{q_{\mathsf{y}}(g(z))}. \tag{31}$$

Applying Lemma 1 to the argument of the minimization in (23), we obtain, for any $\hat{p}_{\mathsf{z}}(\cdot) \in \hat{\mathcal{P}}_{\mathsf{z}}(\mathbf{y})$,

$$D(\hat{p}_{\mathsf{z}}(\cdot) \| p_{\mathsf{z}}(\cdot; x)) \geq D(\hat{p}_{\mathsf{y}}(\cdot; \mathbf{y}) \| p_{\mathsf{y}}(\cdot; x)). \tag{32}$$

But the right-hand side of (32) does not depend on the particular choice of $\hat{p}_{\mathsf{z}}$. Hence, the minimum in (23) occurs when (32) is satisfied with equality, which from specializing Lemma 1 is when (24) is satisfied.

We leave the proof of Lemma 1, which can be developed using the log-sum inequality, as a homework exercise.

19.3 Further reading

For an alternative derivation, see, for example, the tutorial "Information Theory and Statistics" by Csiszar and Shields.

[2] A more general form of this Lemma involves randomized mappings, but the present version is sufficient for our needs.