Information Geometry and Maximum Likelihood Criteria

William Byrne
byrne@jhu.edu
Center for Language and Speech Processing and
Department of Electrical and Computer Engineering
The Johns Hopkins University
Barton Hall, 3400 N. Charles St., Baltimore, MD, 21218

1 Introduction

This paper presents a brief comparison of two information geometries as they are used to describe the EM algorithm used in maximum likelihood estimation from incomplete data. The Alternating Minimization framework based on the I-Geometry developed by Csiszár is presented first, followed by the *em*-algorithm of Amari. Following a comparison of these algorithms, a discussion of a variation in likelihood criterion is presented. The EM algorithm is usually formulated so as to improve the marginal likelihood criterion (as described in Section 2.1). Closely related algorithms also exist which are intended to maximize different likelihood criteria. The 1-Best criterion, for example, leads to the Viterbi training algorithm used in Hidden Markov Modeling. This criterion has an information geometric description that results from a minor modification of the marginal likelihood formulation.

The techniques discussed here are not given in rigorous detail, but rather at a level intended to allow comparison between them; the works cited in the bibliography should be consulted for complete and correct presentations of all methods discussed.

2 Likelihood Criteria for Incomplete Data Problems

The estimation task is to choose a distribution Q from a family $\mathcal{Q}$ to best describe a training set $T = \{y_i\}$ that consists of observations of a random process Y. Assuming that the observations are independent, the maximum likelihood estimation problem is then to find the distribution $Q \in \mathcal{Q}$ that maximizes $\prod_i Q_Y(y_i)$[1].

2.1 Marginal Likelihood Criterion

When the variable Y is not a sufficient statistic for the model Q, this is termed an incomplete data maximum likelihood problem. One formulation of this problem is that $Q \in \mathcal{Q}$ is a joint distribution on two random processes X and Y. If (y, x) is available as a single observation of Y and X, then the maximum likelihood training objective is

$$\max_{Q \in \mathcal{Q}} Q_{Y,X}(y, x) \tag{1}$$

However, if only y is available it is necessary to choose a likelihood criterion to assign likelihood to y as a function of the possible values X might assume. The marginal likelihood

$$Q_Y(y) = \sum_x Q_{Y,X}(y, x) \tag{2}$$

leads to the training objective

$$\max_{Q \in \mathcal{Q}} \sum_x Q_{Y,X}(y, x) \tag{3}$$

as in the EM algorithm [1]. No restrictions are placed on the values of X that might be associated with any observation y. The EM algorithm proceeds via the E-Step:

$$K(Q'', Q) = E_{Q_{Y,X}}[\log Q''_{Y,X}(y, x) | Y = y] \tag{4}$$

A new model Q' is then found by the M-Step:

$$Q'_{Y,X} = \arg \max_{Q'' \in \mathcal{Q}} K(Q'', Q) \tag{5}$$

This yields $Q'_Y(y) \geq Q_Y(y)$ [1].

[1] For simplicity, it is assumed here and elsewhere that all needed maxima can be attained.

Presented at the Conference on Information Sciences and Systems, Princeton, New Jersey, 1996

2.2 Alternating Minimization

It is well-known that the EM algorithm can be formulated in terms of the alternating minimization algorithm of Csiszár and Tusnády [2] and [3]. One formulation is as follows. An empirical distribution $\hat{P}(Y)$ is defined on the observed random variable Y so that it places all its mass on the observed data from the training set T

$$\hat{P}_Y(y) = \frac{1}{|T|} \sum_i \delta_{y_i}(y) \tag{6}$$

This distribution is used to define a family of distributions $\mathcal{D}$ on (Y, X) in terms of the following linear constraint

$$\mathcal{D} = \{P_{Y,X} : \hat{P}_Y(y) = \sum_x P_{Y,X}(y, x)\} \tag{7}$$

Note that for $P_{Y,X} \in \mathcal{D}$, $P_{Y,X} = P_{X|Y}\hat{P}_Y$.

The alternating minimization algorithm proceeds by projecting back and forth between the $\mathcal{D}$ and $\mathcal{Q}$ families under the information divergence

$$P^*_{Y,X} = \arg\min_{P \in \mathcal{D}} D(P_{Y,X} \| Q_{Y,X}) \tag{8}$$

$$Q'_{Y,X} = \arg\min_{Q'' \in \mathcal{Q}} D(P^*_{Y,X} \| Q''_{Y,X}) \tag{9}$$

Equation 8 is referred to as the I-Projection of Q to $\mathcal{D}$ [4]. It is denoted $D(\mathcal{D}||Q)$ and can be thought of as the divergence "distance" from Q to $\mathcal{D}$. In the simple case where X, Y take a finite number of values, it can be found easily from the Log-Sum inequality [5] (see [6] for an example). In this case

$$P^*_{Y,X} = Q_{X|Y}\hat{P}_Y \tag{10}$$

$$D(\mathcal{D}||Q_{Y,X}) = D(\hat{P}_Y||Q_Y) \tag{11}$$

It is easily shown that

$$D(\hat{P}_Y||Q_Y) = -H(\hat{P}_Y) - \frac{1}{|T|}\sum_i \log Q_Y(y_i) \tag{12}$$

from which it follows that decreasing $D(\mathcal{D}||Q)$ increases $\sum_i \log Q_Y(y_i)$. Hence, $\mathcal{D}$ has been termed the family of desired distributions, because approaching the family in terms of the divergence improves the likelihood of the training data. Also, inserting Equation 10 into Equation 9 yields

$$Q'_{Y,X} = \arg\max_{Q'' \in \mathcal{Q}} \sum_y \hat{P}_Y(y) E_{Q_{X|Y}} \log Q''_{Y,X}(y, x) \tag{13}$$

By comparing this last result to Equations 4 and 5 as in [2], it is shown that the I-Projection does indeed lead to the same operations as the EM algorithm.

Repeated application of Equations 8 and 9 yields a sequence of models $\{Q^p_{Y,X}\}$ so that $D(\mathcal{D}||Q^p_{Y,X})$ decreases. This is shown graphically in Figure 1.

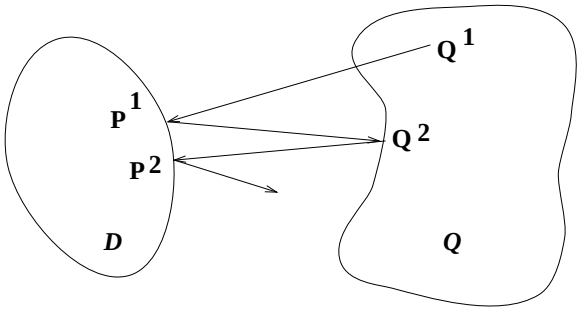

Figure 1: Alternating Minimization

2.3 The *em* algorithm

An information geometric descriptiion of the EM algorithm has been developed by Amari [7, 8] based on the geometry presented in [9]. The *em* algorithm is discussed in [10] and conditions for its equivalence to the EM algorithm have recently been given in [7]. A summary is presented here for comparison to the alternating minimization algorithm as presented above.

The set of model distributions $\mathcal{Q}$ is taken to be a subset of the family of exponential distributions parameterized by the n-dimensional parameter w

$$\mathcal{S} = \{S_{Y,X}(y, x; w) = e^{w \cdot g(y,x) + \psi_w}\} \tag{14}$$

Distributions in $\mathcal{S}$ are also parameterized by the expectation parameters η

$$\eta = E_{S_{Y,X}} g(Y, X) \tag{15}$$

Suppose the parameters of the models Q are restricted to be a function of another parameter u: $w = w(u)$. The function $w(u)$ is not allowed to vary freely over all possible values of w, so $\mathcal{Q} \subset \mathcal{S}$

$$\mathcal{Q} = \{Q_{Y,X}(y, x; w(u)) = e^{w(u) \cdot g(y,x) + \psi_u}\} \tag{16}$$

The observed data is introduced into the problem through the definition of the *observed data*

submanifold. For convenience, let the training data be a single observation $\hat{y}$. The observed data submanifold is defined as

$$\mathcal{N} = \{\eta : \eta = g(\hat{y}, x);\ \text{x varies over the range of X}\} \tag{17}$$

The coordinates $\eta \in \mathcal{N}$ define a submanifold of $\mathcal{S}$ different from the set of models $\mathcal{Q}$.

The *em* algorithm is defined in terms of the m-straight line that joins two distributions in $\mathcal{S}$ with parameters η and η':

$$\eta_t = t\,\eta + (1-t)\,\eta' \qquad 0 \le t \le 1 \tag{18}$$

and the e-straight line that joins two distributions in $\mathcal{S}$ with parameters w and w'

$$w_t = t\,w + (1-t)\,w' \qquad 0 \le t \le 1 \tag{19}$$

Given a model Q with parameters w and η, the e-projection to $\mathcal{N}$ is found first. This distribution is specified by w^* and η^* and is found such that the e-straight line between w and w^* is orthogonal to $\mathcal{N}$. With this e-projection fixed, a new model Q' with parameters w' and η' is specified by finding the m-straight line between η^* and η' that is orthogonal to $\mathcal{Q}$. In [7], it is shown that repeated application yields a sequence of models Q^p that approaches a local minimum in $D(\mathcal{N}||Q)$ presented graphically in Figure 2 (after Amari [7])

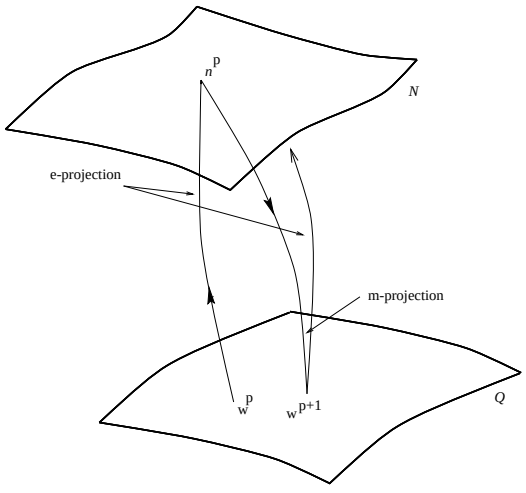

Figure 2: The *em* algorithm

This does not lead directly to improvement in the marginal likelihood. According to Theorem 3 of [7], the *em* and EM algorithms are equivalent if and only if $E_Q[g(Y,X)|Y]$ is linear in Y. This implies that improvement under the marginal likelihood criterion is not assured unless this condition is met. The relative merits of this are discussed in [7]. Two significant modeling assumptions are that the $\mathcal{Q}$ and $\mathcal{N}$ manifolds are flat, in that if two distributions belong to $\mathcal{Q}$, then the e-straight line joining them must also belong to $\mathcal{Q}$. Similarly, if two distributions belong to $\mathcal{N}$, the m-straight line joining them must also belong to $\mathcal{N}$. Manifolds with these properties are termed e-flat and m-flat, respectively.

2.4 Summary and Comparison

The Alternating Minimization procedure as described in Section 2.2 has the following properties:

- the desired distributions are defined using
  - the likelihood criterion
  - the observed training data
- the I-Projection directly yields the marginal likelihood of the training data
  - can be found easily for some problems through the log-sum inequality
- directly increases the marginal likelihood
- leads to the same operations as the EM algorithm

The *em* algorithm as summarized in Section 2.3 has the following properties:

- the observed data submanifold is defined through
  - the observed training data
  - the minimal sufficient statistics of the model family
- model and observed data submanifolds are flat
- equivalence to the EM algorithm and improvement under the marginal likelihood criterion depends on the model

3 Variations in Likelihood Criterion

The previous section discussed training algorithms under the marginal likelihood criterion. Another widely used likelihood criterion for missing data problems that is easily described is

$$\max_x Q_{Y,X}(y,x) \tag{20}$$

This assumes that only the single, best value of X should be associated with y. This is called here the 1-Best criterion. In this case, the E-Step as given in Equation 4 can be replaced by

$$\begin{array}{c} K_1(Q'',Q) = \log Q''_{Y,X}(y,x_v) \\ \text{where } x_v = \arg\max_x Q_{Y,X}(y,x) \end{array} \tag{21}$$

The M-Step is unchanged. A new model is found as

$$Q'_{Y,X} = \arg\max_{Q''\in\mathcal{Q}} \log Q''_{Y,X}(y,x_v) \tag{22}$$

This leads to

$$\max_x Q'(y,x) \geq \max_x Q(y,x) \tag{23}$$

When Q is an HMM, this algorithm is known as "Viterbi-style" training, or as the Segmental K-Means [11] algorithm. Both names are appropriate. The Viterbi algorithm is used to find the needed value x_v, while the M-Step used to implement Equation 22 resembles K-Means clustering.

3.1 Alternating Minimization Description of the 1-Best Criterion

In [12] it is shown that the Viterbi training algorithm also has a formulation in the alternating minimization framework. The set of desired distributions $\mathcal{D}$ defined in Equation 7 is modified by adding a size constraint on the support of $P_{X|Y}$:

$$\mathcal{D}_1 = \mathcal{D} \cap \{P_{Y,X} : |\{x : P_{X|Y}(x|y) > 0\}| = 1\} \tag{24}$$

Given this modified set of desired distributions, it is necessary to find the I-Projection of a model $Q_{Y,X}$ onto this set. The P^* which satisfies

$$P^*_{Y,X} = \arg\min_{P\in D_1} D(P_{Y,X}\|Q_{Y,X}) \tag{25}$$

is found as

$$P^*_{Y,X}(y,x) = \begin{cases} \hat{P}_Y(y) & x = \arg\max Q_{Y,X}(y,x) \\ 0 & \text{otherwise} \end{cases} \tag{26}$$

$$\begin{array}{c} D(P^*_{Y,X}||Q_{Y,X}) = \\ -H(\hat{P}_Y) - \frac{1}{|T|}\sum_{y_i} \max_x Q_{Y,X}(y,x) \end{array} \tag{27}$$

As in the case of the marginal likelihood, the I-Projection here yields the likelihood of the training data under the criterion that defines the training algorithm. Finding $Q'_{Y,X}$ according to Equation 9 then yields $\max_x Q'_{Y,X}(y) \geq \max_x Q_{Y,X}(y)$. This is a special case of adding a more general constraint to $\mathcal{D}$

$$\mathcal{D}_N = \mathcal{D} \cap \{P_{Y,X} : |\{x : P_{X|Y}(x|y) > 0\}| \leq N\} \tag{28}$$

which leads to N-Best training algorithms and extends the formulation presented in [13] to other commonly used likelihood criteria used in training HMMs.

3.2 Boltzmann Machine learning under the 1-Best criterion

Boltzmann machines [14] are binary valued stochastic neural networks which may have both visible and hidden units. Information geometric discussions of Boltzmann Machine learning are presented in [10, 7, 6, 15]. The network state is denoted $Z = [Y(1),,Y(m),X(1),\ldots,X(n)]'$, $Z(j) \in \{0,1\}$. The vector Y is the visible units and X is the hidden units. The network behavior is such that, when free running, the network achieves a stationary distribution

$$B_Z(z;W) = c_W e^{-z'Wz} \tag{29}$$

parameterized by the matrix of network connectivities W.

In a network with no hidden units, maximum likelihood estimation of the parameters W is completely specified by the training data through the values of $p(j,k) = E_{\hat{P}_Y} Z(j)Z(k)$ which are defined for all pairs of network units. These empirical observations of the desired network behavior are used by whatever algorithm is chosen to implement the maximum likelihood solution. For example, gradient methods [16] can be used and

a description of an iterative proportional fitting solution is given in [6].

In a network with hidden units, however, the behavior of all units is not determined by the training data. It is therefore necessary to estimate values of $p(j,k)$ where they are not specified by the training data. In training under the marginal likelihood criterion, the unspecified components are estimated by "clamping" [14] which can be thought of as a stochastic approximation to the E-Step. It produces estimates $\bar{p}(j,k)$ as $E_{\hat{P}_Y} E_{B_{Z|Y}}[Z(j)Z(k);W]$ by observing the behavior of the Boltzmann machine itself. In [6] these are shown to be $E_{P_Z^*} Z(j)Z(k)$, where P_Z^* is the I-Projection of $B_Z(z;W)$ onto $\mathcal{D}$. These estimates $\bar{p}(j,k)$ are then used by whatever algorithm is chosen to implement the M-Step.

It is possible to describe a Boltzmann machine learning algorithm under the 1-Best criterion. As with HMM training under this criterion, the M-Step is not changed. However, the values of $p(i,j)$ not determined by the training data are found from the I-Projection of $B_Z(z;W)$ onto $\mathcal{D}_1$

$$P^*_{Y,X}(x,y) = \begin{cases} \hat{P}_Y(y) & x = \arg\max B_Z([y,x];W) \\ 0 & \text{otherwise} \end{cases} \tag{30}$$

Unlike in HMM training, however, dynamic programming algorithms are not in general available to perform this computation. While it is therefore difficult to find exactly the best value of the hidden vector for a given visible vector, an approximate value can be found by running the Boltzmann machine deterministically as a Hopfield network [17]. For each training vector y_i, this approach can be used to find a vector of hidden units x_i which is at a local maximum of $B_Z([y_i,x];W)$. Taken together, these vectors form a "temporary" training set $\{z_i\}$, where $z_i = [y_i, x_i]$ for $y_i \in T$. The I-Projection $B_Z(z;W)$ onto $\mathcal{D}_1$ can then be approximated by

$$P^*_Z(z) = \begin{cases} \hat{P}_Y(y_i) & z = [y_i,x_i],\ y_i \in T \\ 0 & \text{otherwise} \end{cases} \tag{31}$$

This distribution is then used to compute the estimate $p_1(j,k) = E_{P_Z^*} Z(j)Z(k)$ as

$$p_1(j,k) = \frac{1}{|T|} \sum_i z_i(j) z_i(k) \tag{32}$$

The resulting training procedure is similar to that used in usual Boltzmann machine learning, except that clamping is replaced by the deterministic dynamics of the Hopfield network.

4 Discussion

The Boltzmann machine training procedure presented above may or may not have practical value. It is presented to show the ease with which a likelihood criterion developed for one model architecture can be extended to a very different model architecture when both models share an information geometric description.

The 1-Best criterion is mentioned in the context of the two information geometries to provide an additional point of comparison between the geometries. Unlike the marginal likelihood, the 1-Best criterion does not appear to have a straightforward formulation in the *em* algorithm. The difficulty is that of formulating an observed data submanifold that incorporates the likelihood criterion and yet is still an m-flat subset of $\mathcal{S}$.

Other likelihood criteria can also be formulated within information geoemtric settings. Although they are beyond the scope of this paper, penalized likelihood criteria intended to prevent overtraining and avoid local maxima in the EM algorithm [18, 19, 20] also have interesting geometric descriptions.

References

[1] A.P.Dempster, N.M.Laird, and D.B.Rubin. Maximum likelihood from incomplete data. *Journal of the Royal Statistical Society*, 39(1):1–38, 1977.

[2] I. Csiszár and G. Tusnády. Information geometry and alternating minimization procedures. *Statistics and Decisions, Supplementary Issue Number 1*, pages 205–237, 1984.

[3] I. Csiszár. ENEE 728F. Information Theory and Statistics Class Notes, Dept. Electrical Engineering, University of Maryland, Spring 1990.

[4] I. Csiszár. I-divergence geometry of probability distributions and minimization problems. *The Annals of Probability*, 3(1):146–158, 1975.

[5] I. Csiszár and J. Körner. *Information theory : coding theorems for discrete memoryless systems*. Academic Press, Orlando, 1981.

[6] W. J. Byrne. Alternating Minimization and Boltzmann Machine learning. *I.E.E.E. Transactions on Neural Networks*, 3(4):612–620, 1992.

[7] S.-I. Amari. Information geometry of the EM and em algorithms for neural networks. *Neural Networks*, 8(9):1379–1408, 1995.

[8] S.-I. Amari. Information geometry of the EM and em algorithms for neural networks. Technical report, Dept. of Mathematical Engineering and Information Physics, University of Tokyo, Bunkyo-ku, Tokyo 113, Japan, 1994.

[9] S.-I. Amari. *Differential-Geometrical methods in statistics*. Springer-Verlag, New York, 1985.

[10] S.-I. Amari, K. Kurata, and H. Nagaoka. Information geometry of Boltzmann Machines. *IEEE Transactions on Neural Networks*, pages 260–271, March 1992.

[11] B.-H. Juang and L. Rabiner. The Segmental K-Means algorithm for estimating parameters of Hidden Markov Models. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 38(9):1639–1641, September 1990.

[12] W. J. Byrne. *Encoding and representing phonemic sequences using nonlinear networks*. PhD thesis, University of Maryland, College Park, 1993.

[13] Y. Ephraim and L. R. Rabiner. On the relations between modeling approaches for information sources. *IEEE Transactions on Information Theory*, 36(2):372–380, March 1990.

[14] G. H. D. Ackley and T. Sejnowski. A learning algorithm for Boltzmann Machines. *Cognitive Science*, 9:147–169, 1985.

[15] N. Anderson and D. Titterington. Boltzmann Machines. In F. Kelly, editor, *Probability, Statistics, and Optimization*. Wiley, 1994.

[16] G. E. Hinton, T. J. Sejnowski, and D. H. Ackley. Boltzmann Machines: Constraint satisfaction networks that learn. Technical Report CMU-CS-84-119, Carnegie-Mellon University, Pittsburgh, PA 15213, May 1984.

[17] J.J.Hopfield. Neural networks and physical systems with emergent collective computational abilities. *Proceedings of the National. Acadamy of Sciences USA*, 79:2554–2558, 1982.

[18] W. Byrne. Generalization and maximum likelihood from small data sets. In *Proceedings of IEEE-SP Workshop on Neural Networks for Signal Processing*, 1993.

[19] N. Ueda and R.Nakano. Deterministic annealing variant of the EM algorithm. In *Advances in Neural Information Processings Systems*, 1994.

[20] M. I. Miller and D. L. Snyder. The role of likelihood and entropy in incomplete-data problems: Applications to estimating point-process intensities and Toeplitz constrained covariances. *Proceedings of the I.E.E.E.*, 75(7):892–907, July 1985.