Expectation Maximization Algorithms for Conditional Likelihoods

Jarkko Salojärvi JARKKO.SALOJARVI@HUT.FI
Kai Puolamäki KAI.PUOLAMAKI@HUT.FI
[1]Laboratory of Computer and Information Science, Helsinki University of Technology, P.O. Box 5400, FI-02015 HUT, Finland

Samuel Kaski[2],[1] SAMUEL.KASKI@HUT.FI
[2]Department of Computer Science, P.O. Box 68, FI-00014 University of Helsinki, Finland

Abstract

We introduce an expectation maximization-type (EM) algorithm for maximum likelihood optimization of conditional densities. It is applicable to hidden variable models where the distributions are from the exponential family. The algorithm can alternatively be viewed as automatic step size selection for gradient ascent, where the amount of computation is traded off to guarantees that each step increases the likelihood. The tradeoff makes the algorithm computationally more feasible than the earlier conditional EM. The method gives a theoretical basis for extended Baum Welch algorithms used in discriminative hidden Markov models in speech recognition, and compares favourably with the current best method in the experiments.

1. Introduction

We discuss optimizing generative models for conditional probability densities $p(c|x)$, that is, models that discriminate the values c of a dependent variable C, conditional on the values of an independent variable X. In practice, such models are currently optimized with algorithms effectively boiling down to gradient ascent. The problem common to gradient ascent-based algorithms is that the update step length needs to be selected empirically. Too optimistic step lengths may overshoot the local optimum resulting in a decrease of the objective function (and hence slower convergence), whereas too pessimistic step lengths lead to slow convergence. So-called second order gradient ascent algorithms take into account the curvature of the model and adjust the step length accordingly. The automation comes with a cost however, since inversion of the Hessian matrix of the model is required.

The most common algorithm used for optimizing hidden variable models is the expectation-maximization (EM) algorithm. It operates by constructing a global lower bound for the objective function, likelihood where the hidden variables have been marginalized out, with the aid of Jensen's inequality. The bound is tight at the current values of the model parameters, $\hat{\theta}$, and its gradient equals that of the objective function at $\hat{\theta}$. Since the lower bound is global, it is guaranteed that optimizing the bound always increases the value of the objective function. Plain EM could be applied in our case to modeling of the joint density of the data pairs (x, c).

Since the gradients of the objective function and its lower bound are equal at $\hat{\theta}$, the EM can be interpreted as a kind of gradient ascent algorithm with automatic selection of step length.

The plain EM algorithm is not applicable to optimizing conditional likelihoods, where the objective function is a rational function. It is possible to construct a lower bound by forming a global lower bound for the numerator and a global upper bound for the denominator. The approach was rigorously studied in (Jebara & Pentland, 2001; Jebara, 2001). Unfortunately the resulting formulas turned out to be very complicated which hinders their practical use, and obtaining even a conservative estimate of the bound is computationally demanding (Jebara, 2001). Moreover, the bounds allow only a very small step size which makes optimization slow and hence further increases computational demands.

An alternative family of algorithms, called extended

Appearing in *Proceedings of the 22nd International Conference on Machine Learning*, Bonn, Germany, 2005.

Baum Welch, EBW (Gopalakrishnan et al., 1991; Povey et al., 2003), is the current state-of-the-art in speech processing with hidden Markov models. During the last fifteen years, considerable experimental effort has been made in the field in order to find EM-type update rules that maximize the discriminative power of the models. Although there now exists a consensus on what the update formulas should look like, backed up with good heuristics for selecting optimal regularization, no solid theory that explains the formulas has been presented yet. The EBW can be easily extended to optimize mixture models, for example the Gaussian mixture model (Klautau, 2003), by assuming that instead of time series, the data consists of N samples of length one. This extension has been coined discriminative EM by Klautau (2003).

In this paper we introduce a discriminative EM-type algorithm that relies on an alternative derivation of the global lower bound for conditional probability densities, which is perhaps more intuitive than (Jebara & Pentland, 2001). The derivation suggests a practical algorithm that trades off the globality of the lower bound for computational efficiency and simpler update formulas. The resulting update formulas are very close to current extended Baum Welch formulas, for which they give a solid basis for choosing the currently partly heuristic step length. Note that none of the currently existing computationally feasible EM-type algorithms are guaranteed to always increase the likelihood. We validate our tradeoff in two different applications.

The computational complexity of the resulting discriminative EM is somewhat bigger than the ordinary EM ($\mathcal{O}(S^3T^2)$ vs. $\mathcal{O}(S^2T)$) algorithm. The algorithms can be used in the same way.

2. Background

In this Section, we will first discuss the exponential family distributions in order to introduce our notation, and to recall the basic characteristics of exponential families which will be needed in deriving the discriminative EM algorithm below.

2.1. Exponential family distributions

Exponential family distribution can always be written in the canonical form

$$p(x|\theta) = \exp\left(T(x)^T\theta - \log Z(\theta) - \log Y(x)\right), \quad (1)$$

where $T(x)$ are the (observed) sufficient statistics of x, θ the natural parameters, and $\log Z(\theta)$ is the convex normalization term (partition function). Table 1 gives representations of the Gaussian and multinomial distributions, the exponential family distributions applied in this paper.

Table 1. Canonical representation of some exponential family distributions.

	Gaussian	Multinomial
$T(x)$	$\begin{pmatrix} x^2 & x \end{pmatrix}$	n_k
θ	$\begin{pmatrix} -\frac{1}{2\sigma^2} & \frac{\mu}{\sigma^2} \end{pmatrix}$	$\log \pi_k$
$\log Z$	$-\frac{\mu^2}{2\sigma^2} - \log\sigma$	$\log\left(\sum \pi_k\right)$
$\log Y$	$-\frac{1}{2}\log 2\pi$	$\log N! - \sum_k \log n_k!$

Two key definitions needed here are the dual parameter μ and covariance matrix Σ (Buntine, 2002),[1]

$$\mu = \langle T(x)\rangle_{p(x|\theta)} = \frac{\partial \log Z}{\partial \theta}, \quad (2)$$

$$\Sigma = \left\langle (T(x)-\mu)\,(T(x)-\mu)^T\right\rangle_{p(x|\theta)} = \frac{\partial^2 \log Z}{\partial\theta\partial\theta} = \frac{\partial\mu}{\partial\theta}. \quad (3)$$

It is always possible to find a θ^* corresponding to the sufficient statistics $T(x)$ by solving a mapping $\frac{\partial}{\partial\theta}\log Z(\theta)\big|_{\theta\ \theta^*} = \mu(\theta^*) = T(x)$. This leads to an alternative way of writing exponential models (see Efron (1978), for example) by $p(x|\theta) = \frac{1}{Z}e^{-B(\theta,\theta^*)}$, where $B(\theta,\theta^*)$ is the Bregman divergence

$$B(\theta,\theta^*) = \log Z(\theta) - \log Z(\theta^*) - \mu(\theta^*)\,(\theta-\theta^*) \quad . \quad (4)$$

In this respect the Bregman divergence is the natural distance measure for the selected exponential family.

For all exponential family models, the Bregman divergence is always non-negative due to convexity of the log-partition function. The second derivative of Bregman divergence is the Fisher information, Σ. Notice also that $B(\hat\theta,\hat\theta) = \frac{\partial}{\partial\theta}\,B(\theta,\hat\theta)\Big|_{\theta=\hat\theta} = 0$.

Due to the convexity of the log-partition function $\log Z(\hat\theta)$, it is always possible to form a log-linear upper bound for exponential family distributions, having the form

$$\begin{aligned} &T(x)^T\theta - \log Z(\theta) - \log Y(x) \le \\ &(T(x)-\mu(\hat\theta))^T\theta - \log Z(\hat\theta) + \mu(\hat\theta)\hat\theta - \log Y(x). \end{aligned} \quad (5)$$

2.2. Joint likelihood hidden variable models

In (marginalized) joint likelihood hidden variable models we optimize

$$\log p(y|\theta) = \log \sum_s p(y,s|\theta) \quad , \quad (6)$$

[1] For compactness of our formulas we denote $\langle T(x)\rangle_{p(x|\theta)} = E_{p(x|\theta)}\{T(x)\}$

where y is the data, s the values of hidden variables, and θ the model parameters. In the following we assume that $p(y, s|\theta)$ is of the exponential family.

The objective function of the EM algorithm is a global lower bound for $\log p(y|\theta)$, obtained with the aid of Jensen's inequality (see Buntine (2002)):

$$\begin{aligned} \log p(y|\theta) &= \log \sum_s p(y, s|\theta) \geq \mathcal{F}(\theta) \\ &= \log p(y|\theta) - KL(q(s)||p(s|y,\theta)) \quad (7) \\ &= \langle \log p(y, s|\theta) \rangle_{q(s)} + H(q(s)) \ , \quad (8) \end{aligned}$$

where $KL(||)$ denotes the Kullback-Leibler divergence and $H()$ the entropy. The globality of the lower bound is easy to see from Eq. (7), since $KL(||)$ is always ≥ 0.

The EM algorithm operates by iteratively minimizing the Kullback-Leibler divergence of equation (7) with respect to the distribution over the hidden variable, $q(s)$, at $\theta = \hat{\theta}$, and then optimizing the likelihood (8) with respect to θ, keeping $q(s)$ fixed. In an EM algorithm there is always a distribution $q(s)$ such that the bound is tight (equality),[2] making the bound tangential to the likelihood at $\theta = \hat{\theta}$ (Buntine, 2002).

2.2.1. Connection to gradient ascent

We will next show a simple connection between the EM algorithm and gradient ascent. If we assume that all the probabilities within the model are expressed using exponential family distributions, the derivative of the model is

$$\frac{\partial}{\partial \theta} \log \sum_s p(y, s|\theta) = \sum_s p(s|y,\theta)\{T_s(x) - \mu(\theta)\}, \quad (9)$$

by using Eq. (2), and denoting sufficient statistics associated with the current assignment s of hidden variables by $T_s(x)$. It can be shown that the EM update is obtained when we compute the distribution $p(s|y, \hat{\theta})$ (so called Expectation phase), and set the derivative to zero,

$$\mu(\theta) = \frac{\sum_s p(s|y,\hat{\theta}) T_s(x)}{\sum_s p(s|y,\hat{\theta})}. \quad (10)$$

Gradient ascent, on the other hand, operates by iterating

$$\mu(\theta) = \mu(\hat{\theta}) + \Gamma^{-1} \frac{\partial}{\partial \theta} \log \sum_s p(y, s|\theta),$$

where Γ^{-1} is a small value. Inserting Eq.(9) and evaluating the derivative at $\hat{\theta}$, we may solve for $\mu(\theta)$, resulting in

$$\mu(\theta) = \frac{\Gamma\mu(\hat{\theta}) + \sum_s p(s|y,\hat{\theta}) T_s(x)}{\Gamma + \sum_s p(s|y,\hat{\theta})}. \quad (11)$$

Gradient ascent thus gives us a version of the EM where the update step length is regularized with Γ.

2.3. Conditional likelihood hidden variable models

In a *discriminative model* the set of observations Y is divided into two classes, $Y = C \cup X$, where C are the variables over which we want to discriminate and X are the remaining observations. The likelihood of the discriminative model is

$$\log p(c|x,\theta) = \log p(y|\theta) - \log p(x|\theta) \quad . \quad (12)$$

Discriminative models are usually optimized using gradient descent methods.

Extended Baum-Welch. In speech processing an *extended Baum-Welch* algorithm has given the best results so far. The algorithm was first presented by Gopalakrishnan et al. (1991) for multinomial observation distributions, and extended by Normandin (1991) to Gaussian distributions. The algorithm can be interpreted to lower bound both $\log p(y|\theta)$ and $\log p(x|\theta)$ using Jensen's inequality. Using Eq. (7), the result is

$$\begin{aligned} \log p(y|\theta) &- KL(q_C(s)||p(s|y,\theta)) + \\ &- \log p(x|\theta) + KL(q_F(s)||p(s|x,\theta)) \quad , \end{aligned}$$

where $q_C(s), q_F(s)$ denote the hidden variable distributions in cases where c is known ("clamped"), or marginalized out ("free"), respectively. Since the last term is positive, globalness of the lower bound cannot be guaranteed. Therefore, some regularization is needed in the update formulas, resulting in a functional form similar to Eq. (11). Gopalakrishnan et al. (1991) present a formula for computing a regularization value Γ which is large enough such that convergence can be guaranteed. However, the resulting value is so large that even in the original publication approximations needed to be made for practical implementation of EBW. As noted in Section 2.2.1, a large regularization coefficient reduces the EBW to a gradient ascent-type optimization algorithm which is known to converge to a local optimum. Since the original publication, choosing the proper amount of regularization has been under considerable debate. See Woodland and Povey (2002) for the most recent best heuristics.

[2] If the distribution family $q(s)$ is not rich enough to include $p(s|y,\hat{\theta})$, i.e. the Kullback-Leibler divergence of equation (7) cannot generally be made to vanish, we have a variational algorithm, where $q(s)$ is the variational approximation. EM algorithm is thus a special case of the variational method.

Conditional Expectation Maximization. To construct an objective function for EM that maximizes the conditional likelihood (12), we need a global lower bound.

In (Jebara & Pentland, 1999; Jebara & Pentland, 2001) this is achieved by using a lower bound $\mathcal{F}(\theta)$ for $\log p(y|\theta)$, as in the EBW, and an upper bound $\mathcal{G}(\theta)$ for $\log p(x|\theta)$. The EM algorithm for the discriminative model then follows straightforwardly from the lower bound given by the difference,

$$\log p(c|x,\theta) \geq \mathcal{F}(\theta) - \mathcal{G}(\theta) \quad . \tag{13}$$

The problem of discriminative training thus reduces to finding a global upper bound for log-likelihood.

In conditional EM, Jebara and Pentland (1999) bound the $\log p(x|\theta)$ by the function itself (plus a constant), $p(x|\theta)+1$, and thus achieve a global lower bound. However, the resulting update rules are complicated, which hinders their practical use.

The update rules would be far simpler if a form of Jensen's inequality could be used to derive an upper bound, since instead of $\log \sum \exp(T(x)\theta + \log Z(\theta) + \dots)$ we would then have functions of the type $\sum \log \exp(T(x)\theta + \log Z(\theta) + \dots)$, where log exp cancel each other. The functional form of the upper bound would then be similar to the lower bound obtained using the ordinary Jensen inequality.

Jebara and Pentland (2001) solve the problem by taking a trial function which has the same functional form as the lower bound:

$$\sum_s q(s)\left[T(y)\Theta - \log Z(\Theta)\right] - \log Y(y),$$

and solve its coefficients $q(s)$, $T(y)$, $\log Y(y)$ so that (i) the bound is tight at $\hat{\theta}$ (thus getting $\log Y(y)$), and (ii) has the same derivative as the log-likelihood at $\hat{\theta}$ (getting $T(y)$).

Inserting $\log Y(y)$ and $T(y)$, and regrouping the variables results in

$$\begin{aligned}\log p(x|\hat{\theta}) + \sum_s p(s|x,\hat{\theta})(T(x) - \mu(\hat{\theta}))(\theta - \hat{\theta}) \\ + q(s)B(\theta,\hat{\theta}) \quad ,\end{aligned} \tag{14}$$

where $T(x)$ and $\mu(\hat{\theta})$ are the observed and expected sufficient statistics, respectively. The term $B(\theta,\hat{\theta})$ is the Bregman divergence between θ and $\hat{\theta}$. Jebara and Pentland (2001) proceed by mapping the Bregman distances to a parable (since every convex function has a diffeomorphic mapping to a parable), and solve the remaining values, $q(s)$. It turns out that the mapping need not be solved explicitly. However, the mapping affects the resulting update rules by restricting the allowed values for $q(s)$, making optimisation difficult. Furthermore, although guaranteed, the speed of convergence and especially the computational demands reported in (Jebara, 2001) leave room for improvement.

3. Discriminative Expectation Maximization

We will next derive a global upper bound for $p(x|\theta)$ by inspecting the first and second derivatives of the objective function, that is, the log-likelihood, and its Jensen lower bound.

3.1. Functional form

We begin from the same functional form as Jebara and Pentland (2001), shown in Equation (14). The choice can be justified by its more flexible form than the normal Jensen lower bound. Namely, if we choose $q(s) = p(s|x,\hat{\theta})$, we get the familiar result:

$$\begin{aligned}\sum_s p(s|x,\hat{\theta})\left[T(x)^T(\theta - \hat{\theta}) - \log Z(\theta) + \log Z(\hat{\theta})\right] + \\ + \log p(x|\hat{\theta}) \quad ,\end{aligned} \tag{15}$$

which is the Jensen lower bound expressed in terms of Bregman divergences. This also motivates the selection of the trial function (14).

The new result in this paper follows from realizing that $q(s)$ does not necessarily have to be a distribution. Furthermore, $q(s)$ affects neither the value nor the derivative of the trial function at $\hat{\theta}$ (since $B(\theta,\hat{\theta})$ is zero in both cases). The term $q(s)B(\theta,\hat{\theta})$ determines the curvature of the trial function, however. We will next show that by choosing the curvature properly we can get new bounds for the objective function.

3.1.1. Curvature considerations

The second derivative (i.e., curvature) of log likelihood is

$$\frac{\partial^2 \log p(x|\theta)}{\partial\theta\partial\theta} = C(\theta,\mu) - \langle\Sigma\rangle_{p(s|\theta)} \quad . \tag{16}$$

Here C is the *correlation matrix* over hidden states,

$$\begin{aligned}C(\theta,\mu) = \sum_s \big(\langle(T_s(x)-\mu)(T_s(x)-\mu)^T\rangle_{p(s|x,\theta)} \\ -\langle(T_s(x)-\mu)\rangle_{p(s|x,\theta)}\langle(T_s(x)-\mu)^T\rangle_{p(s|x,\theta)}\big) \ ,\end{aligned} \tag{17}$$

where we denote by $T_s(x)$ the sufficient statistics associated with the assignment s of the hidden variables. The $\langle\Sigma\rangle_{p(s|\theta)}$ is the Fisher information matrix.

Since both $C(\theta, \mu)$ and Fisher information are positive-semidefinite, the resulting curvature may be positive or negative, but it is always upper bounded by $C(\theta, \mu)$ and lower bounded by the (negative) Fisher information matrix.

The curvature of the Jensen lower bound (15) is defined by the second derivative of the Bregman divergence, the (negative) Fisher information $-\langle\Sigma\rangle_{p(s|x,\hat{\theta})}$. The lower bound is thus concave, which makes the optimisation problem easy to solve, and its curvature is defined only by the model and $q(s)$, not data.

The sufficient condition for the globality of an upper bound is that it must have a larger curvature than the objective function (Boyd & Vandenberghe, 2004). Hence a valid upper bound should have a larger curvature than the correlation matrix $C(\theta, \mu)$.

3.2. Upper bound

We can now find the upper bound by using a trial function $\mathcal{G}(\theta)$, having a similar functional form as (14):

$$\log p(x|\theta) \leq \mathcal{G}(\theta) = \log p(x|\hat{\theta}) + \\ + \sum_s p(s|x,\hat{\theta})(T(x) - \mu(\hat{\theta}))(\theta - \hat{\theta}) + \Lambda D(\theta, \hat{\theta}) \ , \quad (18)$$

where Λ is an appropriate constant and $D(\theta, \hat{\theta})$ is a distance function. We may choose D quite freely, as long as it is a convex function (which is equivalent to having a diffeomorphic transformation to a parable). The constant Λ is chosen to be large enough, such that the curvature of ΛD bounds the curvature of C from above. Then Eq. (18) always remains an upper bound. The sufficient conditions for D are:

1. $D(\hat{\theta}, \hat{\theta}) = 0$,
2. $\frac{\partial}{\partial\theta} D(\theta, \hat{\theta})\Big|_{\theta=\hat{\theta}} = 0$,
3. $\frac{\partial^2}{\partial\theta^2} D(\theta, \hat{\theta}) > 0$.

Upper bound for Gaussian distribution. For simplicity, let us first consider the Gaussian distribution. We select the Bregman divergence of the Gaussian partition function as the distance function, i.e., $D(\theta, \hat{\theta}) = \frac{1}{2}(\theta - \hat{\theta})^2$. The curvature of the trial function is then constant,

$$\frac{\partial^2 \mathcal{G}(\theta)}{\partial\theta\partial\theta} = \Lambda \mathbf{1} \quad . \quad (19)$$

Upper bound for the multinomial distribution. Multinomial distribution is more complicated than the Gaussian, since the curvature of the Bregman divergence is not constant. For example in the binomial case the curvature will be $\propto \mu(1-\mu)$, being zero at $\mu = \{0, 1\}$, and hence cannot provide an upper bound for $C(\theta, \mu)$. Moreover, using the same distance function as in the Gaussian case would result in unnecessarily complicated update formulas, since the mapping from natural parameters to dual parameters is not the identity function (but $\mu = e^\theta$ instead), as it is for Gaussians.

For these reasons, we use the function

$$D(\mu, \hat{\mu}) = \frac{1}{2}\left(\sqrt{\frac{\mu}{\hat{\mu}}} - \sqrt{\frac{\hat{\mu}}{\mu}}\right)^2 \quad .$$

The distance function fulfills the required conditions for D, upper bounds the true Bregman divergence for the multinomial distribution, and is symmetric around $\hat{\mu}$. When this distance function is used, the update rules result from solving a second order polynomial.

3.2.1. Choosing Λ

After choosing the distance function, we will have to find a proper value for the Λ. The distance functions D we have used have the curvature of 1 at $\hat{\theta}$. It is therefore the task of the constant Λ to provide the upper bound for C. This can be fulfilled by requiring that Λ is no less than the largest eigenvalue of $C(\theta, \mu)$.

The value of $C(\theta, \mu)$ in (17) depends on the parameters θ. Since we assumed that $p(s, x|\theta)$ is within the exponential family, we can form an upper bound $C(\theta, \hat{\mu})$, which is valid for all θ, by constructing *a log-linear upper bound* at $\hat{\mu}$ for each distribution in the likelihood $p(s, x|\theta)$ (i.e., we use the inequality (5) to construct the upper bound).

A global upper bound would then result from finding the worst-case hidden variable assignment. To be able to calculate a practical upper limit for the correlation we approximate $p(s|x, \theta)$ by $p(s|x, \hat{\theta})$ in the following. This is the only approximation we make. As a result, we will end up computing the upper bound for C, at the current parameter values $\hat{\mu}$, with the current hidden variable distribution $p(s|x, \hat{\theta})$.

Since the largest eigenvalue of $C(\theta, \mu)$ is bounded from above by the trace of the matrix, we obtain a (conservative) upper bound by setting

$$\Lambda = \mathrm{Tr}\ C \quad . \quad (20)$$

The upper bound parameter Λ can be computed separately for all block-diagonal elements of C.

About the approximation. It is possible to find a global worst-case estimate for the curvature by considering all possible distributions for $p(s|x,\theta)$ (an upper bound can be found by quadratic programming in $\mathcal{O}(ST^4)$). This would result in an alternative derivation of the bound derived by Jebara and Pentland (2001). By switching to a local approximation $p(s|x,\hat{\theta})$, we trade off the globality of the bound to a computationally more feasible solution. This tradeoff is justified in practice by the extended Baum Welch, since for the case of Gaussian distributions we arrive at the same update formulas. Furthermore, our approach gives a justification to the currently state-of-the-art heuristics used for optiming discriminative HMMs (Woodland & Povey, 2002). By using the approximation, the computational complexity in the case of HMMs is $\mathcal{O}(S^3T^2)$, where S is the number of hidden states and T the length of the time sequence. The complexity is thus of the order of ST larger than in the ordinary EM algorithm (Rabiner, 1989), but still manageable.

4. Discriminative Hidden Markov Model

A hidden Markov model (HMM) is defined formally as a 5-tuple $(s_{1:T},Y,\Pi,A,B)$, where $s_{1:T} = \{s_1,\ldots,s_T\}$ is a finite set of N states over a time sequence given by $t = 1,\ldots,T$, $Y = \{y_1,\ldots,y_T\}$ are the observations, $\Pi = \{\pi_1,\ldots,\pi_N\}$ are the initial state probabilities, $A = \{a_{ij}\}_{i,j\in\{1,\ldots,N\}}$ are the state transition probabilities, and $B = \{b_i(Y)\}$ are the emission probabilities. We use $\theta = \{\Pi, A, B\}$ to denote all model parameters.

The HMM is an exponential family mixture model where the path over the states $s_{1:T}$ has the role of a hidden variable in Eq. (6) of the EM algorithm. See (Rabiner, 1989) for the use of HMMs for joint likelihood maximization.

The log likelihood of a *discriminative* HMM is

$$\mathcal{L}_D(\theta) = \log p(c|x_{1:T},\theta) = \log p(y_{1:T}|\theta) - \log p(x_{1:T}|\theta) \quad . \tag{21}$$

The EM algorithm for discriminative HMM proceeds as follows: First we find the path probabilities corresponding to $\log p(y_{1:T}|\theta)$ ("clamped model") and $\log p(x_{1:T}|\theta)$ ("free model") using the forward-backward algorithm, resulting in the expressions for the clamped and free state probabilities, $\gamma_{C,t}(i)$ and $\gamma_{F,t}(i)$, and the transition probabilities, $\xi_{C,t}(i,j)$ and $\xi_{F,t}(i,j)$, respectively.

Maximization of the (global) lower bound with respect to the model parameters θ then results in update rules; see Woodland and Povey (2002) for EBW updates and appendix A for discriminative EM updates.

4.1. Choosing Λ for HMMs

Upper bounding the correlation matrix for HMMs is not trivial. The largest eigenvalue of $C(\theta,\mu)$ is bounded from above by the trace of the matrix. To compute the trace we need to define the probability $A_{t,\tau}(i) = p(s_\tau = i|s_t = i, x_{1:T},\hat{\theta})$. This probability can be expressed using a matrix product of the quantities $\gamma_t(i) = p(s_t = i|x_{1:T},\hat{\theta})$ and $\xi_t(i,j) = p(s_{t+1} = j|s_t = i, x_{1:T},\hat{\theta})$, given by the forward-backward algorithm. We obtain $A_{t,\tau}(i) = (\xi_t \times \xi_{t+1} \times \ldots \times \xi_{\tau-1})_{ii}$, where $t < \tau$, matrix multiplication is denoted by $\times$, and $A_{t,t}(i) = 1$ and $A_{\tau,t}(i) = A_{t,\tau}(i)$. We further define $B_{t,\tau}(i) = A_{t,\tau}(i) - \gamma_\tau(i)$ and express the trace as

$$\mathrm{Tr}\, C = \sum_{i,t,\tau} \gamma_t(i) B_{t,\tau(i)} (T_i(x_t) - \mu_i)(T_i(x_\tau) - \mu_i) \quad .$$

The trace can be computed easily and efficiently in a loop of size $\mathcal{O}(S^3T^2)$, with $\mathcal{O}(T^2)$ matrix multiplications. See appendix A for the resulting update formulas.

The step length of the discriminative EM is determined mainly by Λ. It is therefore useful to know its order of magnitude. Assume that our data has a typical correlation time, T_C. That is, the correlation is weak, $B_{t,\tau}(i) \approx 0$, if $|\tau - t| > T_C$, and strong otherwise, $B_{t,\tau}(i) \approx 1$, where $|\tau - t| < T_C$. A typical value of the trace is thus of the order $\mathrm{Tr}\, C \propto TT_C S^{-1} E\left((T(x) - \mu)^2\right)$, where we have used $\gamma_t(i) \approx S^{-1}$. The trace should therefore scale linearly with respect to the length T of the time series.

5. Experiments

We compared convergence properties of the EBW and our algorithm using the log likelihood of the test data set as a measure of performance, i.e. the *perplexity*

$$\text{perp.} = e^{-\frac{\mathcal{L}}{N}} \text{ , where } \mathcal{L} = \sum_{i\ 1}^{N} \log P(C = c(i) \mid x^i_{1:T},\theta) \text{ ,}$$

where we denote by $c(i)$ the class associated with the time series $x^i_{1:T}$, and θ denotes the parameters of the model under evaluation. N is the size of the test set.

Experiments were carried out with two different data sets. The first was the homo sapiens splice sites data set[3], consisting of nucleic acid sequences of introns and

[3] Available at `http://www.sci.unisannio.it/docenti/rampone/`.

exons. A subset of 100 introns and 100 exons with length of less than 135 base pairs was selected and split into an equal-sized training and test data set.

The second data set consisted of eye movement measurements. The task here was to predict the known relevance of a text associated with the measured eye movement data.[4] The features computed for the HMMs were the same as the example features computed by the competition organizers (Salojärvi et al., 2005). A split into training and validation data sets is provided in the competition.

For the Extended Baum-Welch update formulas we used the heuristics given by Woodland and Povey (2002). Since the objective functions of both algorithms are the same, and since the interest here lies only in comparing the convergence of the algorithms, we did not tune the HMMs for either of the tasks. Each class was modelled simply with two hidden states, resulting in a 4-state HMM for the exon data, and a 6-state HMM for the eye movement data.

The perplexity through iterations is plotted in Figure 1. For the exon data the EBW overshoots the minimum at an early stage. The algorithm then sets to a different minimum which is different from the minimum found by the discriminative EM. With the eye movement data, the EBW overshoots the minimum at a later stage whereas the discriminative EM converges nicely.

In both cases the EBW with heuristics tuned for speech recognition exhibits oscillatory behavior. It therefore seems that the data sets require different heuristics. Our approach makes such tuning unnecessary.

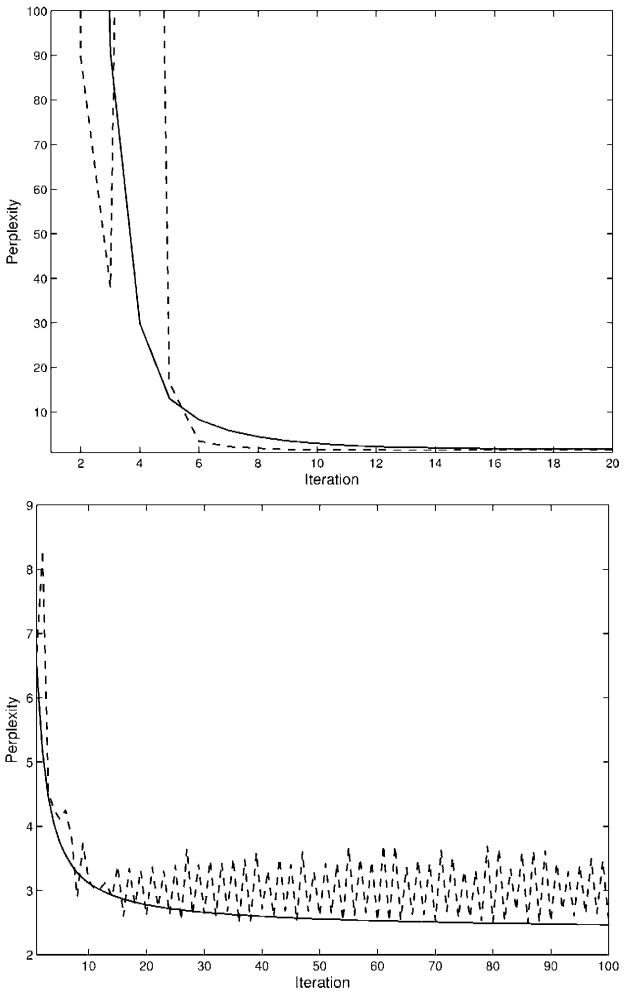

Figure 1. Perplexity of the validation data set. Top: Exon data. Bottom: Eye movement data Solid line: discriminative EM, dashed line: extended Baum-Welch.

6. Discussion

We introduced an EM-type algorithm that can be used to maximize conditional likelihoods of hidden variable models. As an example, we derived update formulas for discriminative hidden Markov models, and used them in two applications. Our results improve upon the earlier work by providing a practical solution with a considerably smaller time complexity. Our work additionally gives a sound theoretical basis for the extended Baum Welch update rules used widely in speech recognition. We validated our method by observing the convergence of hidden Markov models using two publicly available data sets.

[4]The data is available for competititors at `http://www.cis.hut.fi/eyechallenge2005/`

Acknowledgments

This work was supported by the Academy of Finland, decision no. 202209, and by the IST Programme of the European Community, under the PASCAL Network of Excellence, IST-2002-506778.[5]

References

Boyd, S., & Vandenberghe, L. (2004). *Convex optimization.* Cambridge, UK: Cambridge University Press.

Buntine, W. (2002). Variational extensions to EM and multinomial PCA. *Proceedings of the 13th European Conference on Machine Learning* (pp. 23–34). Springer-Verlag.

[5]This publication represents the authors' views. Access rights are restricted due to other commitments.

Efron, B. (1978). The geometry of exponential families. *The Annals of Statistics, 6*, 362–376.

Gopalakrishnan, P., Kanevsky, D., Nádas, A., & Nahamoo, D. (1991). An inequality for rational functions with applications to some statistical estimation problems. *IEEE Transactions on Information Theory, 37*, 107–113.

Jebara, T. (2001). *Discriminative, generative and imitative learning.* Doctoral dissertation, Media laboratory, MIT.

Jebara, T., & Pentland, A. (1999). Maximum conditional likelihood via bound maximization and the CEM algorithm. *Advances in Neural Information Processing Systems 11* (pp. 494 500). Cambridge,MA: The MIT Press.

Jebara, T., & Pentland, A. (2001). On reversing Jensen's inequality. *Advances in Neural Information Processing Systems 13.* Cambridge, MA: MIT Press.

Klautau, A. (2003). Discriminative Gaussian mixture models: A comparison with kernel classifiers. *Proceedings of the Twentieth International Conference on Machine Learning (ICML-2003)* (pp. 353–360). AAAI Press.

Normandin, Y. (1991). *Hidden Markov models, maximum mutual information estimation and the speech recognition problem.* Doctoral dissertation, McGill University.

Povey, D., Woodland, P., & Gales, M. (2003). Discriminative MAP for acoustic model adaptation. *IEEE International Conference on Acoustics, Speech, and Signal Processing, 2003. Proceedings. (ICASSP'03)* (pp. 312–315).

Rabiner, L. R. (1989). A tutorial on hidden Markov models and selected applications in speech recognition. *Proceedings of the IEEE, 77*, 257–286.

Salojärvi, J., Puolamäki, K., Simola, J., Kovanen, L., Kojo, I., & Kaski, S. (2005). *Inferring relevance from eye movements: Feature extraction* (Technical Report A82). Helsinki University of Technology, Publications in Computer and Information Science. http://www.cis.hut.fi/eyechallenge2005/.

Woodland, P. C., & Povey, D. (2002). Large scale discriminative training of hidden Markov models for speech recognition. *Computer Speech and Language, 16*, 25–47.

A. Update formulas for discriminative HMM

Priors

$$\begin{aligned}
\pi(i) &= \frac{f}{2g} + \frac{1}{|2g|}\sqrt{f^2 + 2\lambda\hat{\pi}(i)g} \\
f &= \sum_n \gamma^n_{C,1}(i) - \gamma^n_{F,1}(i)(1-\hat{\pi}(i)) \\
g &= N + \frac{\Lambda}{2\hat{\pi}(i)} - \Gamma \\
\Lambda &= N - \sum_{i,n} \left(\gamma^n_{F,1}(i)\right)^2 \ .
\end{aligned}$$

Select Γ such that $\sum_i \pi(i) = 1$. Here i is an index over hidden states and n over sequences. A lower bound for Γ is given by requiring that the discriminant $f^2 + 2\lambda\hat{\pi}(i)g \geq 0$.

Gaussian observation densities

$$\mu^G_{b_i}(x) = \frac{\sum_{n,t} \eta^n_t T(x^n_t) + \hat{\mu}_{b_i}\left(\Lambda + \sum_t \gamma^n_{F,t}(i)\right)}{\Lambda + \sum_{n,t} \gamma^n_{C,t}(i)} \ ,$$

$$\Lambda - \sum_{n,t,\tau} \gamma^n_{F,t}(i) B^n_{t,\tau}(i) \left(T(x^n_t) - \hat{\mu}_{b_i}\right)\left(T(x^n_\tau) - \hat{\mu}_{b_i}\right) \ ,$$

$$B^n_{t,\tau}(i) = A^n_{t,\tau}(i,i) - \gamma^n_{F,\tau}(i) \ , \ \eta^n_t = \gamma^n_{C,t}(i) - \gamma^n_{F,t}(i) \ .$$

Multinomial observation densities

$$\begin{aligned}
\mu^M_{b_{ij}}(x) &= \frac{f}{2g} + \frac{1}{|2g|}\sqrt{f^2 + 2\Lambda\hat{\mu}_{b_{ij}}g} \\
f &= \sum_t \left(\gamma^n_{C,t}(i) - \gamma^n_{F,t}(i)\right)\delta(x^n_t, j) + \hat{\mu}_{b_{ij}} \sum_t \gamma^n_{F,t}(i) \\
g &= \sum_t \gamma^n_{C,t}(i) + \frac{\Lambda}{2\hat{\mu}_{b_{ij}}} - \Gamma \\
\Lambda &= \sum_j \sum_{t,\tau} \gamma^n_{F,t}(i) B^n_{t,\tau}(i) \omega^n_{tij}\omega^n_{\tau ij} \ , \\
B^n_{t,\tau}(i) &= \Lambda^n_{t,\tau}(i,i) - \gamma^n_{F,\tau}(i) \ , \ \omega^n_{tij} = \delta(x^n_t, j) - \hat{\mu}_{b_{ij}} \ .
\end{aligned}$$

Select Γ such that $\sum_j \mu_{b_{ij}} = 1$.

Transitions

$$\begin{aligned}
a_{ij} &= \frac{f}{2g} + \frac{1}{|2g|}\sqrt{f^2 + 2\Lambda\hat{a}_{ij}g} \\
f &= \sum_{n,t} \xi^n_{C,t}(i,j) - \xi^n_{F,t}(i,j) + \hat{a}_{ij} \sum_{n,t} \gamma^n_{F,t}(i) \\
g &= \sum_{n,t} \gamma^n_{C,t}(i) + \frac{\Lambda}{2\hat{a}_{ij}} - \Gamma \\
\Lambda &= \sum_{n,j} \sum_t \sum_{\tau > t} 2\xi_{F,t}(i,j) A_{t+1,\tau}(j,i) \xi_{F,\tau}(i|j) \\
&\quad -2\hat{a}_{ij}\xi_{F,t}(i,j) A_{t+1,\tau}(j,i) \\
&\quad -2\hat{a}_{ij}\gamma^n_{F,t}(i) A_{t,\tau}(i,i)\xi_{F,\tau}(i|j) + 2\gamma^n_{F,t}(i) A_{t,\tau}(i,i)\hat{a}^2_{ij} \\
&\quad + \sum_t \xi_{F,t}(i,j)(1-2\hat{a}_{ij}) + \gamma^n_{F,t}(i)(1+\hat{a}^2_{ij}) \\
&\quad + \sum_j \left(\sum_{n,t} \xi_{F,t}(i,j) - \hat{a}_{ij}\gamma^n_{F,t}(i)\right)^2 .
\end{aligned}$$

Select Γ such that $\sum_j a_{ij} = 1$. Here t and τ are indexes over the length of the time sequence n.

B. Extended Baum-Welch vs. discriminative EM

By selecting $\Gamma = \Lambda + \sum_t \gamma_{F,t}(i)$, we get the extended Baum-Welch update rule

$$\mu^{(Gauss)}_{b_i}(x) = \frac{\sum_t \left(\gamma_{C,t}(i) - \gamma_{F,t}(i)\right) T(x(t)) + \hat{\mu}_{b_i}\Gamma}{\sum_t \gamma_{C,t}(i) - \gamma_{F,t}(i) + \Gamma} .$$

The heuristic by Woodland and Povey (2002) for selecting Γ is to choose $\max\left(\sum_t \gamma_{F,t}(i), 2B\right)$, where B is the minimum value such that the updated variance is positive.