A New Family of Extended Baum-Welch Update Rules

Dimitri Kanevsky, Daniel Povey,
Bhuvana Ramabhadran, Irina Rish
IBM T. J. Watson Research Center
Yorktown Heights, NY 10598
{kanevsky, dpovey}@us.ibm.com
{bhuvana, rish}@us.ibm.com

Tara N. Sainath
MIT Computer Science
Artificial Intelligence Laboratory
32 Vassar St. Cambridge, MA 02139
tsainath@mit.edu

Abstract

In this paper, we consider a generalization of the state-of-art discriminative method for optimizing the conditional likelihood in Hidden Markov Models (HMMs), called the Extended Baum-Welch (EBW) algorithm, that has had significant impact on the speech recognition community. We propose a generalized form of EBW update rules that can be associated with a weighted sum of updated and initial models, and demonstrate that using novel update rules can significantly speed up parameter estimation for Gaussian mixtures.

1 Introduction

Efficient methods for learning HMMs are essential for solving a wide range of natural language processing tasks, such as part-of-speech tagging, word segmentation, optical character recognition, as well as acoustic modeling in speech recognition, just to name a few applications. The EBW approach (Woodland, 2002), (Povey, 2007), (Kanevsky, 2004) is currently considered one of the most successful discriminative training techniques for estimating models parameters using HMM with Gaussian mixtures. EBW is an iterative algorithm for estimating HMM parameters that uses a specific set of update rules performed at each iteration. These rules involve special EBW parameters that control the amount of change in an objective function (e.g. the Maximum Mutual Information Estimation (MMIE) objective) at each iteration of the algorithm. Significant efforts in speech community has been devoted to learning what values of these control parameters lead to better estimation of parameters of Gaussian mixture in discriminative tasks. In this paper, we introduce a generalization of EBW update rules that leads to a novel family of EBW algorithms where EBW is included as a particular case.

We find that proper choice of the control parameters allows for faster training relative to previous methods. Recently EBW update rules have been used to derive a gradient steepness measurement to evaluate the quality of the model to match the distribution of the data (Sainath, 2007). In this paper we also derive a gradient steepness measurement for the family of EBW update rules that are applied to functions of Gaussian mixtures and demonstrate the growth property of these transformations. Namely, we show that the value of the objective function is non-decreasing under these updated rules.

The paper is structured as follows. In the next section we introduce a generalized family of EBW update rules. In Section 3, we reproduce explicit formulas to measure the gradient steepness, and show the relationship between our EBW family and a recently proposed "constrained line search" method of (Cong et al., 2007). Empirical results are presented in Section 4. Section 5 concludes the paper and discusses future work.

2 A New Family of EBW Update Rules

In this section we introduce a novel family of update rules for parameter estimation with diagonal Gaussian mixtures. Assume that data $y_i, i \in I = \{1, ..., n\}$ is drawn from a Gaussian mixture with each component of the mixture described by the parameters $\theta_j = (\tau_{1j}, \tau_{2j})$, where τ_{1j} is

the mean and τ_{2j} is the variance. Thus the probability of y_i given model θ_j is $z_{ij} = z_i(\theta_j) = \frac{1}{(2\pi)^{1/2}\tau_{2j}} e^{-(y_i-\tau_{1j})^2/2{\tau_{2j}}^2}$. Let $F(z) = F(\{z_{ij}\})$ be some objective function over $z = \{z_{ij}\}$, and let $c_{ij} = z_{ij}\frac{\delta}{\delta z_{ij}}F(z)$.

We will now define the following function that we will call the function *associated* with F:

$$Q(\theta'_j, \theta_j) = \sum_i z_i(\theta_j)\frac{\delta F(\{z_i(\theta_j)\})}{\delta z_i(\theta_j)} \log z_i(\theta'_j),$$

Optimizing this function will lead to closed-form update rules (that are generally not obtainable if optimizing F directly)[1]. Let $\{\tilde{\tau}_{rj}\}$ be solutions of $\frac{\delta Q(\theta'_j,\theta_j)}{\delta \tau'_{rj}} = 0$.

The key contribution of this paper is introduction of the following novel iterative rules for updating the current model parameters τ_{rj} to their next values $\bar{\tau}_{rj}$ (and a subsequent analysis of their properties):

$$\bar{\tau}_{rj}(\alpha_{rj}) = \alpha_{rj}\tilde{\tau}_{rj} + (1-\alpha_{rj})\tau_{rj} + f_r(\alpha_{rj}) \quad (1)$$

Note that the above rules generalize the ones considered previously in (Cong et al., 2007) (see Section 3.1 for more details) via adding the term $f_r(\alpha_{rj}) = o(\alpha_{rj})$ (recall that $o(\epsilon)$ means that $\lim_{\epsilon\to 0} o(\epsilon)/\epsilon \to 0$). It can be shown that our new update rules (1) also include as a particular case the following EBW rules (Woodland, 2002)

$$\hat{\tau}_{1j} = \tau_{1j}(C) = \frac{\sum_{i\in I} c_{ij}y_i + C\tau_{1j}}{\sum_{i\in I} c_{ij} + C} \quad (2)$$

$$\hat{\tau}^2_{2j} = \tau_{2j}(C)^2 = \frac{\sum_{i\in I} c_{ij}y_i^2 + C(\tau^2_{1j} + {\tau_{2j}}^2)}{\sum_{i\in I} c_{ij} + C} - \hat{\tau}^2_{1j} \quad (3)$$

Indeed, assuming $\sum_i c_{ij} \neq 0$ and $\alpha_{rj} = \frac{\sum_i c_{ij}}{C}$ we have

$$\tau_{rj}(C) = \bar{\tau}_{rj}(\alpha_{rj}) \quad (4)$$

Here $|f_r(\alpha_{rj})| < d/C^2$ for sufficiently large C and for some constant d. To show this inequality one needs to observe that $\tilde{\tau}_{1j} = \frac{\sum_i c_{ij}y_i}{\sum_i c_{ij}}$ and $\hat{\tau}_{1j} = \tau_{1j}(0)\tilde{\alpha}_{1j} + \tau_{1j}(1-\tilde{\alpha}_{1j})$ where $\tilde{\alpha}_{1j} = \tilde{\alpha}_{1j}(C) = \frac{\sum_i c_{ij}}{\sum_i c_{ij}+C}$ and $\tau_{1j}(0)$ is defined as in (2) for $C = 0$. This implies statement (4) for $r = 1$ (i.e. for the mean parameter). Statement (4) for $r = 2$ (i.e., for the variance) follows from the fact that $\tilde{\tau}^2_{2j} = \frac{\sum_i c_{ij}(y_i-\mu_j)^2}{\sum_i c_{ij}}$ and from the linearized equations for variances in (18) in (Kanevsky, 2004).

[1]Note that when the objective F is the log-likelihood function (e.g., standard MLE estimation in HMM, i.e. the Baum-Welch method), then Q coincides with the auxiliary function.

3 Gradient Steepness Measurements

Using the linearization technique (Kanevsky, 2004) it was proved that transformations (2, 3) are growth transformations (i.e., cannot decrease the objective function) for large C if the function F obeys certain smoothness constraints. In what follows, we formulate a somewhat more general result.

Proposition 1 *Let $F(\{z_{ij}\})$, $i = 1...m$, be differentiable at τ_{1j}, τ_{2j} and $\frac{\delta F(\{z_{ij}\})}{\delta z_{ij}}$ exist at z_{ij}. Let $\hat{z}_{ij} = \frac{1}{(2\pi)^{1/2}\tau_{2j}(D_j)} e^{-(y_i-\tau_{1j}(C_j))^2/2\tau_{2j}(D_j)^2}$. Let $\hat{\tau}_{rj} \neq \tau_{rj}$ for some $r \in \{1,2\}$. Then for sufficiently large C_j and D_j we get $F(\{\hat{z}_{ij}\}) - F(\{z_{ij}\}) = \sum_j(\alpha_j T_{1j} + \beta_j T_{2j}) + \sum_j(o(\alpha_j) + o(\beta_j))$ where $\alpha_j = 1/C_j, \beta_j = 1/D_j$ and where*

$$T_{1j} = \frac{[\sum_i c_{ij}(y_i - \tau_{1j})]^2}{\tau^2_{2j}} > 0 \quad (5)$$

$$T_{2j} = \frac{\{\sum_i c_{ij}[(y_i - \tau_{1j})^2 - \tau^2_{2j}]\}^2}{2\tau^4_{2j}} > 0 \quad (6)$$

In other words, $F(\{\hat{z}_{ij}\})$ grows proportionally to $\sum_j \alpha_j T_{1j} + \sum_j \beta_j T_{2j}$ for sufficiently small $\alpha_j, \beta_j > 0$.

The proof is similar to Theorem 1 (Kanevsky, 2004), which assumed $C = D$. A similar gradient steepness result was proved for multidimensional multivariate Gaussian mixtures in (Kanevsky, 2005). Gradient steepness measurements T_{1j}, T_{2j} are always non-negative (sums of squares). This guarantees the growth property, i.e. that $F(\{\hat{z}_{ij}\}) \geq F(\{z_{ij}\})$ for sufficiently large C and D.

3.1 Relationship to "constrained line search"

(Cong et al., 2007) provides an optimization method known as the "Constrained Line Search" (CLS) . In what follows we will show that CLS could be considered as a member of a family of EBW transformations (1). In order to demonstrate this we represent the EBW family updates of model

parameters (1) in the following three steps.

1. Gradient for Model Changes: From (1) one can see that a direction along which models are updated is the following: $\lim_{\alpha_{rj}\to 0} \frac{\bar{\tau}_{rj}(\alpha_{rj})-\tau_{rj}}{\alpha_{rj}} = \tilde{\tau}_{rj} - \tau_{rj}$ This coincides with the direction of the line search that is defined in (Cong et al., 2007).

2. Finding a step along the search curve: If $\sum_j c_{ij} \neq 0$ then one can represent (5, 6) as

$$T'_{1j} = \frac{(\bar{\tau}_{1j} - \tau_{1j})^2}{\tau_{2j}^2},\ T'_{2j} = \frac{(\bar{\tau}_{2j}^2 - \tau_{2j}^2)^2}{2\sigma_j^4}$$

and $F(\{\hat{z}_{ij}\}) - F(\{z_{ij}\}) = \sum_j \alpha'_j T'_{1j} + \sum_j \beta'_j T'_{2j} + \sum_j (o(\alpha'_j) + o(\beta'_j))$ where $\alpha'_j = (\sum_i c_{ij})^2/C_j, \beta'_j = (\sum_i c_{ij})^2/D_j$. One can show (Cong et al., 2007) that the above equations approximate mean and variance "components" of the KL-divergence between two Gaussians (for updated and initial models). Therefore the gradient steepness measure can be used to evaluate the closeness of updated models to initial models. These metrics were used in (Sainath, 2007) for various speech tasks. They also can be used to avoid overfitting in EBW training if one chooses C_j inversely proportionally to gradient steepness metrics.

3. Finding a step direction on the search curve: Let us connect models $\{\tau_{1j}, \tau_{2j}^2\}$ and $\{\tilde{\tau}_{1j}, \tilde{\tau}_{2j}^2\}$ with a curve segment (which generalizes the straight line segment used by (Cong et al., 2007)). Then the following cases for location of an updated model (at which F increases its value) can be considered: 1) If $\sum_i c_{ij} > 0$ then a step $\alpha_j > 0$ and $\{\bar{\tau}_{1j}(\alpha_j), \bar{\tau}_{2j}^2(\alpha_j)\}$ lies on a segment that connects $\{\tau_{1j}, \tau_{2j}^2\}$ and $\{\tilde{\tau}_{1j}, \tilde{\tau}_{2j}^2\}$. 2) If $\sum_i c_{ij} < 0$ then a step $\alpha_j < 0$ and $\{\bar{\tau}_{1j}(\alpha_j), \bar{\tau}_{2j}^2(\alpha_j)\}$ lies outside of the segment that connects $\{\tau_{1j}, \tau_{2j}^2\}$ and $\{\tilde{\tau}_{1j}, \tilde{\tau}_{2j}^2\}$. These cases correspond to cases in (Cong et al., 2007) where a sign of a step along a gradient was chosen depending on whether F has the minimum or the maximum at $\{\tilde{\tau}_{1j}, \tilde{\tau}_{2j}^2\}$. The above process is illustrated in Fig. 1.

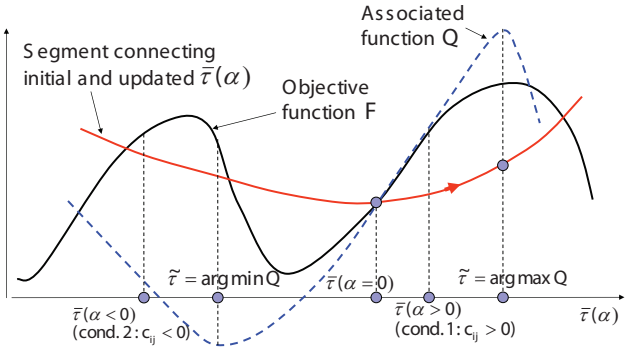

Figure 1: Illustration of the new update rules.

4 Experiments

In this section we report results on a speaker independent English broadcast news system. The discriminative baseline for training is done as in (Woodland, 2002). The acoustic model is trained on 450 hours of speech comprising the 1996 and 1997 English Broadcast News Speech collections and the English broadcast audio from TDT-4. Lightly-supervised training was performed on the TDT-4 audio because only closed captions were available. The model has 6000 quinphone context dependent states and 250K Gaussians. We test on the rt04 test set as defined for the English portion of the EARS program. More details are given in (Povey, 2008) which uses the same testing setup.

We performed experiments testing various members in an EBW family for transformations where we varied f_r and ratio of α_{1j}/α_{2j} in (1). Specifically, we investigated the following conditions.

1. *Linearized update of means:*

$$\hat{\tau}_{1j} = \tau_{1j}(C_j) = \tau_{1j} + \frac{\sum_{i\in I} c_{ij}(y_i - \tau_{1j})}{C_j}$$

$$\hat{\tau}_{2j}^2 = \tau_{2j}(D_j)^2 = \frac{\sum_{i\in I} c_{ij} y_i^2 + D_j(\tau_{1j}^2 + {\tau_{2j}}^2)}{\sum_{i\in I} c_{ij} + D_j} - \tau_{1j}(D)^2 \quad (7)$$

2. *Ratio of control parameters:* $D_j/C_j = 1.5$

3. *Low value of control parameters:* C_j for each Gaussian prototype is chosen to keep variance positive, e.g. starting from low $C_j = 1$ and multiplying C_j by 1.1 until variance (7) becomes positive.

It was observed that any one of above conditions alone do not provide improvement in a decoding accuracy when decoding was done on a PBS subset in the rt04 test set using Boosted MMIE setting as described in (Povey, 2008); Table 1 shows WER

1st cond	No		Yes	
2nd cond / 3rd cond	No	Yes	No	Yes
No	15.3	15.3	15.4	15.4
Yes	16.4	16.2	14.8	14.6

Table 1: Combinations of three conditions: word error rate (WER) on PBS subset in rt04, 1st iteration of update.

Iteration	Test set Training method			
	rt04 bas.I	rt04 comb.I	rt04 bas.II	rt04 comb.II
0	20.5%		20.5%	
1	19.9%	18.9%	18.7%	17.8%
2	19.5%	**18.7%**	17.8%	**17.3%**
3	19.1%		17.4%	
4	**18.8%**		**17.3%**	

Table 2: Word error rate on the test set rt04 (4:00 hours).

with all combinations of the above conditions. The best result occurs when all three conditions are combined.

Table 2 describes experiments on test sets in which columns that are labeled as *test rt04/bas.x* contain results for baseline MMI training for 4 iterations (starting from a ML baseline). Columns labeled as *test rt04/xxx.I* represent setting for the baseline MMI (for backoff) as described in (Woodland, 2002) and columns labeled as *test rt04/xxx.II* represent the setting with boosted MMI that was introduced recently (Povey, 2008). Columns labeled as *rt04/comb.I* and *rt04/comb.II* represent two subsequent iterations with modified EBW (combined 3 conditions shown in table 1) on standard and boosted MMI. These results show that the modified EBW in 2 iterations allows to achieve the same decoding result as 4 iterations with the baseline methods and therefore is significantly faster. We reproduced only two subsequent iterations of modified EBW here since application of a third iteration of modified EBW leads to degradation of the accuracy. This is because using low C and D in (7) leads to overfitting. In order to avoid overfitting in a consequent iteration, one needs to increase C,D at each iteration. The preliminary experiments provide evidence that one can control the size of C and D at each iteration by measuring gradient steepness (see Section 3) or relative changes in likelihood.

5 Conclusion and future work

In the paper we considered a family of transformations that can be associated with weighted sums of updated and initial models. We showed that this family of transformations has the same gradients as EBW transformations and therefore provide estimates that converges to local maximum. We demonstrated that considering different members in this EBW family allows leads to faster discriminative training. We also demonstrated that CLS updates of model parameters in diagonal Gaussian mixtures (Cong et al., 2007) can be considered as members of an EBW family of transformations. We plan to continue to study EBW based training in which EBW control parameters are correlated to gradient steepness along "mean and variance directions." We also plan to extend results of this paper for multivariate multidimensional Gaussian mixture densities.

References

Cong Liu Peng Liu Hui Jiang Soong, F. Ren-Hua Wang. 2007. *Constrained Line Search Optimization for Discriminative Training in Speech Recognition*, Proc. ICASSP, 2007.

D. Kanevsky. 2004. *Extended Baum Transformations for General Functions*. Proc. ICASSP.

D. Kanevsky. 2005. *Extended Baum Transformations for General Functions, II*. tech. Rep. RC23645(W0506-120), Human Language technologies, IBM.

Daniel Povey, Brian Kingsbury. 2007. *Evaluation of Proposed Modifications to MPE for Large Scale Discriminative Training*. ICASSP'07.

Daniel Povey, Dimitri Kanevsky, Brian Kingsbury, Bhuvana Ramabhadran, George Saon and Karthik Visweswariah. 2008. *Boosted MMI for model and feature-space discriminative training*. ICASSP'08.

Tara N. Sainath, Dimitri Kanevsky, Bhuvana Ramabhadran. 2007. *Gradient Steepness Metrics Using Extended Baum-Welch Transformations for Universal Pattern Recognition Tasks*. Proc. ASRU, 2007

P.C. Woodland and D. Povey. 2002. *Large Scale Discriminative Training of hidden Markov modles for Speech Recognition*. Computer Speech and Language, pp. 25-47, Vol. 16, No. 1, January 2002.