# Convergence of DLLR Rapid Speaker Adaptation Algorithms

*Asela Gunawardana and William Byrne*

Center for Language and Speech Processing
Department of Electrical and Computer Engineering
The Johns Hopkins University,
3400 N. Charles St., Baltimore, MD 21218, USA
{zilla,byrne}@jhu.edu

## Abstract

Discounted Likelihood Linear Regression (DLLR) is a speaker adaptation technique for cases where there is insufficient data for MLLR adaptation. Here, we provide an alternative derivation of DLLR by using a censored EM formulation which postulates additional adaptation data which is hidden. This derivation shows that DLLR, if allowed to converge, provides maximum likelihood solutions. Thus the robustness of DLLR to small amounts of data is obtained by slowing down the convergence of the algorithm and by allowing termination of the algorithm before overtraining occurs. We then show that discounting the observed adaptation data by postulating additional hidden data can also be extended to MAP estimation of MLLR-type adaptation transformations.

## 1. Introduction

Discounted Likelihood Linear Regression (DLLR) [1] is a technique for robust speaker adaptation from very small amounts of data. In particular, it is a modification of the popular Maximum Likelihood Linear Regression (MLLR) technique [2, 3] that requires less adaptation data. DLLR is an application of the general discounted likelihood estimation scheme [4, 5] to the problem of estimating MLLR-type adaptation transformations. Discounted likelihood estimation is a modification of the Estimation Maximization (EM) algorithm [6, 7] for the case when there is little data.

Although it was initially motivated by an information geometric description of the EM algorithm as alternating minimization, discounted likelihood can also viewed as the use of the EM algorithm with an augmented hidden variable [1]. As such, the convergence properties of the EM algorithm [7] apply to discounted likelihood procedures, and they yield the same estimates as the standard EM algorithm (i.e., local maxima in likelihood) if allowed to run until convergence. However, the discounted likelihood procedures have slower convergence than the standard EM algorithm, and therefore allow stopping before overtraining takes place.

It was shown in previous work [1] that two efficient DLLR implementations could be obtained by making two different approximations. Because of these approximations, the resulting algorithms are no longer exact EM procedures. However, we show here that both procedures have the same fixed point sets as the EM algorithm, and therefore yield MLLR solutions if they converge. This underscores the importance of terminating the algorithm before overtraining occurs.

The paper is organized as follows. We start by reviewing the description of discounted likelihood as censored EM and deriving DLLR reestimation equations in section 2, showing that DLLR yields maximum likelihood solutions. We then show in section 3 how the efficient approximate DLLR implementations presented in previous work [1] also lead to maximum likelihood solutions. In section 4 we show how discounted likelihood estimation can be applied to MAP estimation of MLLR-type transformations, and we close with conclusions in section 5.

## 2. Discounted Likelihood and the EM Algorithm for Rapid Speaker Adaptation

For MLLR-type speaker adaptation, the HMM state emission densities are reparametrized as

$$q(o|s;\, T) = \mathcal{N}(o;\, T\nu_s, \Sigma_s)$$

where $\mathcal{N}$ is a normal density on the feature vector or observation $o$, $\nu_s = [\mu_s' \; 1]'$ is the extended speaker independent (SI) mean vector at state $s$, and $\Sigma_s$ is the covariance matrix at state $s$, so that the matrix $T$ defines an affine transformation of the SI means $\mu_s$. Note that, as discussed in [1], we assume without loss of generality that state emission densities are single Gaussians instead of mixtures of Gaussians. Also, for simplicity of presentation we will only treat the case where a single global transformation $T$ is used for all states, and note that all the results presented here extend easily to the case of multiple regression classes of states.

The speaker adaptation problem is to estimate $T$

based on adaptation data $(\hat{w}_1^{\hat{n}}, \hat{o}_1^{\hat{l}})$ consisting of a word sequence $\hat{w}_1^{\hat{n}}$ of length $\hat{n}$ and an observation sequence $\hat{o}_1^{\hat{l}}$ of length $\hat{l}$. In MLLR adaptation, the estimation is done to maximize the likelihood of the adaptation data, and since corresponding random state sequence $S_1^L$ is not observed, the EM algorithm is used.

Suppose that the adaptation data $(\hat{w}_1^{\hat{n}}, \hat{o}_1^{\hat{l}})$, which is an observation of the joint random sequences $(W_1^N, O_1^L)$ is insufficient for robust MLLR estimation. Suppose also that $n$ observations of $(W_1^N, O_1^L)$ would have been sufficient for robust MLLR estimation. We will denote these i.i.d. pairs of random sequences by $(\underline{W}^{(1)}, \underline{O}^{(1)}), \cdots, (\underline{W}^{(n)}, \underline{O}^{(n)})$, where we have omitted the random lengths of the sequences to simplify the notation. We will now assume that all $n$ sequences were recorded, but that all but the first were censored or hidden. That is, we will treat the adaptation data $(\hat{w}_1^{\hat{n}}, \hat{o}_1^{\hat{l}})$ as an observation $(\underline{\hat{w}}^{(1)}, \underline{\hat{o}}^{(1)})$ of $(\underline{W}^{(1)}, \underline{O}^{(1)})$. Then, the hidden variables are $\underline{S}^{(1)}$ and $(\underline{W}^{(2)}, \underline{S}^{(2)}, \underline{O}^{(2)}), \cdots, (\underline{W}^{(n)}, \underline{S}^{(n)}, \underline{O}^{(n)})$.

The transformation $T$ is now the parameter of the joint distribution on $(\underline{W}^{(1)}, \underline{S}^{(1)}, \underline{O}^{(1)}), \cdots, (\underline{W}^{(n)}, \underline{S}^{(n)}, \underline{O}^{(n)})$. Iterative EM reestimation of $T^{(p+1)}$ from $T^{(p)}$ based on the observation $(\underline{\hat{w}}^{(1)}, \underline{\hat{o}}^{(1)})$ of $(\underline{W}^{(1)}, \underline{O}^{(1)})$ is done by choosing $T^{(p+1)}$ to maximize the EM auxiliary function

$$\Phi(T|T^{(p)}) = \mathbf{E}\Big[\log q(\underline{W}^{(1)}, \underline{S}^{(1)}, \underline{O}^{(1)}, \cdots, \underline{W}^{(n)}, \underline{S}^{(n)}, \underline{O}^{(n)}; T) \;\Big|\; \underline{\hat{w}}^{(1)}, \underline{\hat{o}}^{(1)}; T^{(p)}\Big]$$

over $T$.

Since the triplets of sequences $(\underline{W}^{(1)}, \underline{S}^{(1)}, \underline{O}^{(1)})$ through $(\underline{W}^{(n)}, \underline{S}^{(n)}, \underline{O}^{(n)})$ are in fact i.i.d. versions of $(W_1^N, S_1^L, O_1^L)$, we get

$$\Phi(T|T^{(p)}) = \mathbf{E}\Big[\log q(W_1^N, S_1^L, O_1^L; T) \;\Big|\; \hat{w}_1^{\hat{n}}, \hat{o}_1^{\hat{l}}; T^{(p)}\Big] + (k-1)\mathbf{E}\Big[\log q(W_1^N, S_1^L, O_1^L; T);\; T^{(p)}\Big].$$

Thus, $T^{(p+1)}$ is chosen to maximize

$$\lambda\mathbf{E}\Big[\log q(W_1^N, S_1^L, O_1^L; T) \;\Big|\; \hat{w}_1^{\hat{n}}, \hat{o}_1^{\hat{l}}; T^{(p)}\Big] + (1-\lambda)\mathbf{E}\Big[\log q(W_1^N, S_1^L, O_1^L; T);\; T^{(p)}\Big]$$

over $T$, where the discounting factor $\lambda = \frac{1}{k}$ is the ratio between the amount of data available and the amount of data necessary for robust estimation.

Maximizing this auxiliary function using calculus shows after some manipulation that the $(p+1)$st iteration of DLLR estimation can be done in two steps [1]. In the first step, which can be thought of as the E step of the EM algorithm, the sufficient statistics $f_s^{(p+1)}$ and $g_s^{(p+1)}$ at iteration $(p+1)$ given by

$$f_s^{(p+1)} = \lambda \sum_{\tau=1}^{\hat{l}} \gamma_s^{(p)}(\tau) + (1-\lambda)\tilde{\gamma}_s \tag{1}$$

and

$$g_s^{(p+1)} = \lambda \sum_{\tau=1}^{\hat{l}} \gamma_s^{(p)}(\tau)\hat{o}_\tau + (1-\lambda)\tilde{\gamma}_s T^{(p)}\nu_s \tag{2}$$

are calculated for every state. Here, $\gamma_s^{(p)}(\tau)$ is the conditional state occupancy of state $s$ at time $\tau$ given the adaptation data, evaluated under the adapted model at iteration $(p)$. On the other hand, $\tilde{\gamma}_s$ is the total (over all time) marginal occupancy of state $s$, without conditioning on the adaptation data, and is given by

$$\tilde{\gamma}_s = \sum_{w_1^n} q(w_1^n) \sum_{s_1^l} q(s_1^l|w_1^n)\#_s(s_1^l), \tag{3}$$

where $q(w_1^n)$ is the languages model probability of $w_1^n$, $q(s_1^l|w_1^n)$ is the probability of $s_1^l$ given $w_1^l$, and $\#_s(s_1^l)$ is the number of times the state $s$ occurs in the state sequences $s_1^l$.

In the second step, which is analogous to the M step of the EM algorithm, a new transformation $T^{(p+1)}$ is chosen to satisfy

$$\sum_s \Sigma_s^{-1} f_s^{(p+1)} T^{(p+1)} \nu_s \nu_s' = \sum_s \Sigma_s^{-1} g_s^{(p+1)} \nu_s'. \tag{4}$$

It is easy to see that at $\lambda = 1$, this DLLR reestimation equation reduces to the usual MLLR estimation algorithm [2].

Since this DLLR reestimation equation was derived as an instance of the EM algorithm, the usual convergence results for the EM algorithm [7] are guaranteed for DLLR. Thus, DLLR gives local maxima of the likelihood of the adaptation data $(\hat{w}_1^{\hat{n}}, \hat{o}_1^{\hat{l}})$ if allowed to run to convergence, as does MLLR. Thus, as was seen in previous work [1], it is important to terminate DLLR procedures before overtraining occurs.

## 3. Efficient DLLR Implementations

As described above, DLLR estimation requires the computation of the marginal state occupancies $\tilde{\gamma}_s$ of equation (3). Obviously, direct computation of these occupancies is impossible, as they involve summations over all word sequences and all state sequences of all lengths. In previous work [1], two approximate approaches for getting around this difficulty, namely moment interpolation and sampling speaker independent data, were presented. Here, we review the two approaches, and then

show that these approximate algorithms both have MLLR fixed points. Thus, both approximations will be shown to share the maximum likelihood property of the exact algorithm discussed above, and will therefore need careful consideration of stopping criteria in practice.

### 3.1. Moment Interpolation

In the first approximate DLLR implementation, the sufficient statistics (moments) $f_s^{(p+1)}$ and $g_s^{(p+1)}$ at iteration $(p+1)$ are computed as follows:

$$f_s^{(p+1)} = \lambda \sum_{\tau=1}^{\hat{l}} \gamma_s^{(p)}(\tau) + (1-\lambda) f_s^{(p)} \tag{5}$$

and

$$g_s^{(p+1)} = \lambda \sum_{\tau=1}^{\hat{l}} \gamma_s^{(p)}(\tau)\hat{o}_\tau + (1-\lambda) g_s^{(p)}. \tag{6}$$

It can be shown [1] that these sufficient statistics would yield exact discounted likelihood estimates if the model densities were regular exponential families (linear in the parameter $T$) instead of curved exponential families (nonlinear in $T$). However, the curvature of the model is small when $T$ is approximately identity, and we therefore treat moment interpolation as an approximate DLLR implementation.

At a fixed point of the moment interpolation algorithm, we have $f_s^{(p)}$, $g_s^{(p)}$ and $T^{(p)}$ such that an iteration of the algorithm gives

$$\begin{aligned} f_s^{(p+1)} &= f_s^{(p+1)} \\ g_s^{(p+1)} &= g_s^{(p+1)} \\ T^{(p+1)} &= T^{(p)}. \end{aligned}$$

Substituting the first two equations into equations (5) and (6) gives

$$f_s^{(p+1)} = \sum_{\tau=1}^{\hat{l}} \gamma_s^{(p)}(\tau)$$

and

$$g_s^{(p+1)} = \sum_{\tau=1}^{\hat{l}} \gamma_s^{(p)}(\tau)\hat{o}_\tau$$

which when substituted into equation (4) yields the usual MLLR reestimation equation [2]. Thus, fixed points of the moment interpolating DLLR implementation are MLLR solutions.

### 3.2. DLLR by Sampling SI Data

In the second approximate DLLR implementation, the marginal state occupancies $\tilde{\gamma}_s$ are approximated directly by using the SI training database. The approach here is to divide the SI training data into a large number of 'adaptation sets,' compute the SI state occupancy of each state on each of these sets, and use the average occupancy $\hat{\gamma}_s$ of each state over all 'adaptation sets' as an approximation to the true marginal occupancy $\tilde{\gamma}_s$. The justification for this is that this can be viewed as computing the marginal occupancy after replacing the marginal $q(w_1^n, o_1^l; T)$ of the model by the empirical distribution over the training data [1]. This is a good approximation because the 45 hour SI training set yields on the order of 32,000 sample adaptation sets over which to compute this empirical distribution, and because the maximum likelihood training of the SI models attempts to fit this empirical distribution as well as possible.

At a fixed point of the algorithm, we have

$$T^{(p+1)} = T^{(p)}.$$

Using $\hat{\gamma}_s$ instead of $\tilde{\gamma}_s$ in equations (1) and (2) and then substituting into equation (4) yields

$$\begin{aligned} \lambda \sum_s \sum_{\tau=1}^{\hat{l}} \Sigma_s^{-1} \gamma_s^{(p)}(\tau) T^{(p+1)} \nu_s \nu_s' + \\ (1-\lambda) \sum_s \Sigma_s^{-1} \hat{\gamma}_s T^{(p+1)} \nu_s \nu_s' = \\ \lambda \sum_s \sum_{\tau=1}^{\hat{l}} \Sigma_s^{-1} \gamma_s^{(p)}(\tau) \hat{o}_\tau \nu_s' + \\ (1-\lambda) \sum_s \Sigma_s^{-1} \hat{\gamma}_s T^{(p)} \nu_s \nu_s'. \end{aligned}$$

Since $T^{(p+1)} = T^{(p)}$, this reduces to

$$\begin{aligned} \sum_s \sum_{\tau=1}^{\hat{l}} \Sigma_s^{-1} \gamma_s^{(p)}(\tau) T^{(p+1)} \nu_s \nu_s' = \\ \sum_s \sum_{\tau=1}^{\hat{l}} \Sigma_s^{-1} \gamma_s^{(p)}(\tau) \hat{o}_\tau \nu_s', \end{aligned}$$

which is the usual MLLR update equation. Thus, fixed points of the SI data sampling approximate DLLR procedure are MLLR solutions.

## 4. Combination with Bayesian Methods

Recently, there has been some interest in using Bayesian methods to estimate MLLR-type adaptation transformations [8, 9] as well as in combining MAP adaptation of unconstrained HMM parameters with transformation based methods such as MLLR and MAPLR [10]. We point out here that discounted likelihood estimation can also be combined with Bayesian methods. In general, such Bayesian methods introduce a penalty term to the usual EM auxiliary function $\Phi(\theta|\theta^{(p)})$, where $\theta$ are the

HMM parameters to be estimated (unconstrained means $\mu$ or transformations $T$, for example. Then, the MAP auxiliary function is given by

$$\Psi(\theta|\theta^{(p)}) = \Phi(\theta|\theta^{(p)}) + \log \pi(\theta)$$

where $\theta$ is a prior density on the parameters. If discounted estimation is being done, the hidden variable is augmented with $n-1$ censored observations as described in section 2, and $\Phi(\theta|\theta^{(p)})$ has the usual conditional expectation corresponding to the real observation and an unconditional expectation corresponding to the $k-1$ 'censored' observations. Then, sufficient statistics are computer according to equations (1) and (2) (or using the appropriate approximations), as with maximum likelihood DLLR estimation. What is changed by the presence of the prior is the M step, which reestimate the parameters based on the sufficient statistics.

In the case of discounted MAPLR, if we assume a matrix-variate normal form for the prior [8], we have

$$\pi(T) \propto |R|^{-(p+1)/2}|S|^{-p/2} \cdot \exp\left\{-\frac{1}{2}\text{tr}(T-M)'R^{-1}(T-M)S^{-1}\right\}$$

where $M \in \mathbb{R}^{(d+1)\times d}$ is the hyper-mean of the transformation matrix, and $R \in \mathbb{R}^{d\times d}$ and $S \in \mathbb{R}^{(d+1)\times(d+1)}$ are symmetric positive definite hyper-covariance matrices. The resulting discounted MAPLR reestimation equation is

$$\sum_s \Sigma_s^{-1} f_s^{(p+1)} T^{(p+1)} \nu_s \nu_s' + RUS = \sum_s \Sigma_s^{-1} g_s^{(p+1)} \nu_s' + RMS,$$

where $U \in \mathbb{R}^{(d+1)\times d}$ has all entries one. Note that while discounting adds a relaxation term to the sufficient statistics to slow estimation down, the prior adds a regularization term to the reestimation equation, and penalizes estimates that deviate from the prior. Thus, while one approach simply slows down the convergence of the algorithm to allow stopping before overtraining occurs (regardless of the training criterion), the other approach discourages the algorithm from giving solutions that deviate significantly from what is expected based on whatever prior knowledge is used in estimating the prior. As shown above, the two approaches can be combined, to give slow, controlled convergence to MAP solutions.

## 5. Conclusions

In sections 2 and 3, it was shown that discounted likelihood estimation changes the weight that is put on the observed data, in that the available adaptation set is treated as one out of many possible adaptation sets. This change is independent of the training criterion, which is still maximum likelihood. Thus the DLLR procedure converges to MLLR solutions. Therefore, it is important to note that the benefit of DLLR procedures comes from a slowing down of MLLR, which allows termination before overtraining occurs, rather than by avoiding overtraining altogether. It was shown in section 4 that discounting of the data can also be performed under a MAP criterion, to yield a combined estimation procedure. In this case, discounting leads to relaxation in moment space, while the prior leads to regularization in parameter space, which shows that DLLR is complementary to MAPLR.

## 6. References


[1] A. Gunawardana and W. Byrne, "Discounted likelihood linear regression for rapid speaker adaptation," *Comp. Spch. & Lang.*, vol. 15, pp. 15–38, Jan. 2001.

[2] C. J. Leggetter and P. C. Woodland, "Maximum likelihood linear regression for speaker adaptation of continuous density hidden Markov models," *Comp. Spch. & Lang.*, vol. 9, pp. 171–185, Apr. 1995.

[3] V. V. Digalakis, D. Rtischev, and L. G. Neumeyer, "Speaker adaptation using constrained estimation of Gaussian mixtures," *IEEE Trans. Spch. & Aud. Proc.*, vol. 3, pp. 357–366, Sept. 1995.

[4] W. J. Byrne, "Generalization and maximum likelihood from small data sets," in *IEEE-SP Wkshp. Neur. Net. Sig. Proc.*, 1993.

[5] A. J. R. Gunawardana, *The Information Geometry of EM Variants for Speech and Image Processing*. PhD thesis, The Johns Hopkins University, 2001.

[6] A. P. Dempster, A. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data," *J. Roy. Stat. Soc., Ser. B*, vol. 39, no. 1, pp. 1–38, 1977.

[7] C. F. J. Wu, "On the convergence properties of the EM algorithm," *Ann. Stat.*, vol. 11, no. 1, pp. 95–103, 1983.

[8] O. Siohan, C. Chesta, and C.-H. Lee, "Hidden Markov model adaptation using maximum a posteriori linear regression," in *Robust Methods for Speech Recognitino in Adverse Conditions*, (Tampere, Finland), 199.

[9] W. Chou, "Maximum a posterior linear regression with elliptically symmetric matrix variate priors," in *Eurospeech*, vol. 1, pp. 1–4, 1999.

[10] O. Siohan, C. Chesta, and C.-H. Lee, "Join maximum a posteriori estimation of transformation and hidden Markov model parameters," in *ICASSP*, IEEE, 2000.