A Convergence Analysis of Log-Linear Training and its Application to Speech Recognition

S. Wiesler, R.Schlüter, H. Ney

Human Language Technology and Pattern Recognition
RWTH Aachen - University of Technology, 52056 Aachen, Germany
{wiesler, schlueter, ney}@cs.rwth-aachen.de

***Abstract*—Log-linear models are a promising approach for speech recognition. Typically, log-linear models are trained according to a strictly convex criterion. Optimization algorithms are guaranteed to converge to the unique global optimum of the objective function from any initialization. For large-scale applications, considerations in the limit of infinite iterations are not sufficient. We show that log-linear training can be a highly ill-conditioned optimization problem, resulting in extremely slow convergence. Conversely, the optimization problem can be preconditioned by feature transformations. Making use of our convergence analysis, we improve our log-linear speech recognition system and achieve a strong reduction of its training time. In addition, we validate our analysis on a continuous handwriting recognition task.**

***Index Terms*—convergence analysis, log-linear models**

I. Introduction

Conventional speech recognition systems rely on hidden Markov models with Gaussian mixture models serving as models for the emission probabilities (GHMMs). The training of the acoustic model starts with a maximum likelihood training with the expectation maximization (EM) algorithm. In state-of-the-art systems, the acoustic model is further optimized according to discriminative criteria, e.g. the minimum phone error (MPE) or maximum mutual information (MMI) criterion. Recently, the interest in direct models, i.e. models for posterior probabilities, has greatly increased. In particular log-linear models are promising, because they fit into the probabilistic framework of hidden Markov models (HMMs). Log-linear models have been successfully applied to phoneme recognition [1], [2], [3]. Promising results have also been obtained on large vocabulary speech recognition tasks (LVCSR), e.g. in [4] or in our previous work [5]. Furthermore, log-linear models are of general interest, because they are widely used in natural language processing, e.g. [6], [7], and many other applications.

An important property of log-linear models is that their training according to the regularized MMI criterion is a strictly convex optimization problem. This property guarantees that apart from the global optimum no other local optima exist. Algorithms with guaranteed convergence, e.g. steepest descent and other more sophisticated gradient-based optimization algorithms, converge to the global optimum from any initialization.

For large-scale applications as speech recognition, considerations in the limit of infinite iterations of the optimization algorithm are not sufficient. With a restricted amount of computation time, the result of the optimization is only an approximation to the unique global optimum. The quality of the approximation depends on the initialization, on the choice of the optimization algorithm, and on the difficulty of the optimization problem.

A number of papers are concerned with the choice of optimization algorithms for log-linear training. Often, the quasi-Newton algorithm L-BFGS is considered as the best optimization algorithm for log-linear training [8], [7]. Good results have also been reported for Rprop [9], which we also used in our previous work.

The analysis of the optimization problem itself is quite limited. Experimentally, it has been observed that the use of correlated features slows down convergence, e.g. by Minka [10]. From optimization theory it is known that the convergence rate of gradient-based optimization algorithms can be described by the condition number of the Hessian matrix at the optimum, i.e. the ratio of its largest and smallest eigenvalue. The dependence on the condition number is very strong for steepest descent. For high condition numbers, steepest descent is useless in practice [11, Chapter 9.3]. It can be shown that more sophisticated gradient-based optimization algorithms as conjugate gradient and L-BFGS depend on the condition number as well [12, Chapter 5.1], [12, Chapter 9.1].

In this paper, we derive an estimate for the condition number of the optimization problem encountered in log-linear training. We show that in extreme cases, log-linear training can be highly ill-conditioned. Conversely, our analysis also shows that log-linear training can be accelerated by feature transformations. A more detailed analysis of the optimization problem and an experimental evaluation on a handwriting task has been given by us in another paper [13]. In this paper, we focus more on the experimental results and perform experiments on a speech recognition task. We compare the results to our previous work on log-linear acoustic models for LVCSR [5]. Furthermore, we compare the effect of the condition number on the behavior of different optimization algorithms.

II. Model Definition and Training Criterion

Let $X \subset \mathbb{R}^D$ denote the observation space and $\mathcal{S} = \{1, \dots, S\}$ a finite set of classes. A *log-linear model* with parameters $\Lambda = (\lambda_{s,n})_{s,n} \in \mathbb{R}^{S\times N}$ is a model for class posterior probabilities of the form

$$p_\Lambda(s|x) = \frac{\exp\Big(\sum\limits_{n=1}^{N} \lambda_{s,n} f_n(x)\Big)}{\sum\limits_{\tilde{s}} \exp\Big(\sum\limits_{n=1}^{N} \lambda_{\tilde{s},n} f_n(x)\Big)}, \tag{1}$$

where the components of

$$f : X \to \mathbb{R}^N,\ x \mapsto (f_1(x), \ldots, f_N(x)) \tag{2}$$

are called *feature functions*.

The regularized MMI criterion is often regarded as the natural training criterion for log-linear models. In this work, we consider the MMI criterion with ℓ_2-regularization:

$$\mathcal{F} : \mathbb{R}^{S\times N} \to \mathbb{R}, \Lambda \mapsto -\frac{1}{T}\sum_{t=1}^{T} \ln p_\Lambda(s_t|x_t) + \frac{C}{2}\|\Lambda\|_2^2. \tag{3}$$

Here $(x_t, s_t)_{t=1,\ldots,T}$ is the training sample and $C > 0$ is the regularization constant. We refer to the minimization of $\mathcal{F}$ as *log-linear training*.

For the optimization of log-linear models, iterative optimization algorithms have to be employed, which require the evaluation of the gradient of the objective function. The first and second partial derivatives of the objective function are:

$$\frac{\partial \mathcal{F}}{\partial \lambda_{s,n}}(\Lambda) = \frac{1}{T}\sum_{t=1}^{T} \left(p_\Lambda(s|x_t) - \delta(s, s_n)\right) f_n(x_t) + \lambda_{s,n}, \tag{4}$$

and

$$\begin{aligned}\frac{\partial^2 \mathcal{F}}{\partial \lambda_{s,n}\partial \lambda_{\bar{s},\bar{n}}}(\Lambda) =& \frac{1}{T}\sum_{t=1}^{T} p_\Lambda(s|x_t)(\delta(s,\bar{s}) - p_\Lambda(\bar{s}|x_t)) \\ & \cdot f_n(x_t) f_{\bar{n}}(x_t) + \alpha\, \delta(s,\bar{s})\delta(n,\bar{n})\ ,\end{aligned} \tag{5}$$

where $1 \leq s, \bar{s} \leq S$ and $0 \leq n, \bar{n} \leq N$ and δ denotes the Kronecker delta. It can be shown that the Hessian matrix of $\mathcal{F}$ is positive semidefinite, and strictly positive definite for $\alpha > 0$. Thus, the optimization problem is convex, respectively strictly convex, see e.g. [14].

For speech recognition, the log-linear model can either be defined on frame level or on sequence level. In the latter case the log-linear model is a conditional random field [15]. For simplicity, we assume here, that the log-linear model is defined on frame-level, i.e. s_t denotes an HMM state and x_t an acoustic observation. The resulting estimates for the posterior probabilities can be used in HMM-based speech recognizers via the *hybrid approach*. The class-conditional probabilities required in HMMs are derived via Bayes rule:

$$p_\Lambda(x|s) = p_\Lambda(s|x)p(x)/p(s)\ . \tag{6}$$

The prior probabilities $p(s)$ can be estimated easily as relative frequencies, and $p(x)$ can be discarded in recognition without changing the maximizing word sequence.

There are numerous possibilities for the definition of appropriate feature functions for speech recognition. Widely used are polynomial feature functions. A polynomial feature of order k is a function

$$\phi : X \to \mathbb{R},\ x \mapsto x_{d_1} \cdot \ldots \cdot x_{d_k}, \tag{7}$$

where $1 \leq d_i \leq D$ for all $1 \leq i \leq k$. In our previous work [5], we applied in addition sparse posterior features:

$$\phi : X \to \mathbb{R}, x \mapsto p(l|x) = \frac{p(l)p(x|l)}{\sum_{l'} p(l')p(x|l')}, \tag{8}$$

where $(p(l))_{1\leq l\leq L}$ and $(p(x|l))_{1\leq l\leq L}$ are obtained by estimating a Gaussian mixture model (GMM) for the marginal probability $p(x)$.

III. Convergence Analysis of Log-Linear Model Training

In [13] we showed that for the unregularized objective function, the condition number of the Hessian matrix of the objective function can be approximated by the condition number of the uncentered covariance matrix

$$X = \frac{1}{T}\sum_{t=1}^{T} f(x_t)f(x_t)^T. \tag{9}$$

The idea for this derivation is to approximate the Hessian at the unknown optimum $\Lambda^\star$ by the Hessian at $\Lambda = 0$. The Hessian at zero has a Kronecker product structure (see e.g. [16]), which allows for the analytic derivation of the eigenvalues.

For large regularization constants, the optimization behavior of the regularized criterion is dominated by the regularization term. Since the regularization term is a well-conditioned quadratic function, the convergence behavior is strongly accelerated in this case. However, for large-scale problems as speech recognition, the regularization constant is typically very small, because the problem of overfitting is less severe with large amounts of training data. In this case, the regularized training criterion behaves similar to the unregularized one and is determined by the eigenvalue distribution of X.

The dependence of the convergence behavior on the properties of X is in accordance to experimental observations. Other researchers have noted before, that the use of correlated features leads to slower convergence [7]. Minka [10] noted that convergence slows down when adding a constant to the features, because this “introduces correlation, in the sense that” X “has significant off-diagonals.”. How can we verify these findings formally? The following theorem concerns the case of uncorrelated features. The proof is an application of Weyl’s inequalities (see [17, Theorem 4.3.7]).

Theorem 1. *Suppose the features $f_n(x), 1 \leq n \leq N$, are uncorrelated with respect to the empirical distribution. Let μ_n and σ_n^2 denote the empirical mean and variance of $f_i(x)$ for $1 \leq i \leq N$. Without loss of generality, we assume that the features are ordered such that $\sigma_1^2 \leq \ldots \leq \sigma_N^2$. Then the condition number of $X = \frac{1}{T}\sum_{t=1}^{T} f(x_t)f(x_t)^T$ is bounded by*

$$\frac{\max\{\sigma_N^2 + \mu_N^2, \sigma_1^2 + \|\mu\|_2^2\}}{\min\{\sigma_1^2 + \mu_1^2, \sigma_2^2\}} \leq \kappa(X) \leq \frac{\sigma_N^2 + \|\mu\|_2^2}{\sigma_1^2}\ . \tag{10}$$

Proof of Theorem 1: Since the features are uncorrelated, we have

$$X = \mathrm{diag}(\sigma_1^2, \dots, \sigma_\mathrm{N}^2) + \mu\mu^T \overset{\mathrm{def}}{=} A + B \ . \tag{11}$$

A lower bound on the condition number is obtained by using the fact that the diagonal elements $X_{1,1} = \sigma_1^2 + \mu_1^2$ and $X_{N,N} = \sigma_N^2 + \mu_N^2$ of X are upper bounds for the smallest eigenvalue and lower bounds for the largest eigenvalue (see [17, p181]).

A tighter lower bound and an upper bound to the condition number are obtained by the application of Weyl's inequalities. Let $\lambda_j(M)$ denote the j-th eigenvalue in ascending order of a Hermitian $N \times N$-matrix M. Weyl's inequalities state that for all Hermitian $N \times N$-matrices A, B and all j, k:

$$\lambda_{j+k-N}(A+B) \quad \leq \lambda_j(A) + \lambda_k(B) \ , \tag{12}$$

$$\lambda_{j+k-1}(A+B) \quad \geq \lambda_j(A) + \lambda_k(B) \ . \tag{13}$$

The eigenvalues of A are the diagonal elements $\lambda_j(A) = \sigma_j^2$. B is a rank-one matrix with the eigenvalues $\lambda_N(B) = \|\mu\|_2^2$ and $\lambda_j(B) = 0$ for $1 \leq j \leq N-1$. The bounds for $\kappa(X)$ follow with the application of (13) and (12) to the smallest and largest eigenvalue. For instance, the upper bound on the condition number follows from the application of (12) with $j = k = N$ to the largest eigenvalue and (13) with $j = k = 1$ to the lowest eigenvalue. The proof of the lower bound is analogous. ■

Theorem 1 shows that even for uncorrelated features, the matrix X can be ill-conditioned, only because the features have a non-zero mean and the variances have a wide range. In particular, the norm $\|\mu\|_2^2$ can get very large for high-dimensional features. Conversely, Theorem 1 shows that convergence of log-linear training can be accelerated by feature transformations. Centering the features and normalizing their variances are only simple preprocessing steps and result in a preconditioning of the optimization problem.

Analyzing the general case of correlated and unnormalized features is more difficult. The idea of the following theorem is regarding the off-diagonals as a perturbation of the diagonal matrix. This case can be analyzed with Geršgorin's circle theorem [17, Theorem 6.1.1], which states that all eigenvalues lie in circles around the diagonal entries of the matrix.

Theorem 2. *Let μ_n and σ_n^2 denote the empirical mean and variance of $f_i(x)$ for $1 \leq i \leq N$ and assume that $\sigma_1^2 \leq \ldots \leq \sigma_N^2$. Let*

$$R_i = \sum_{j, j \neq i} |\mathrm{Cov}\,(f_j(x), f_i(x))\,| \tag{14}$$

denote the radius of the i-th Geršgorin circle. Then, the largest and smallest eigenvalues of $X = \frac{1}{T}\sum_{t=1}^{T} f(x_t) f(x_t)^T$ are bounded by

$$\sigma_1^2 - R_1 \leq \lambda_1(X) \leq \min\{\sigma_1^2 + \mu_1^2, \sigma_N^2 + R_N\} \ , \tag{15}$$

and

$$\begin{aligned} \max\{\sigma_N^2 + \mu_N^2, \sigma_1^2 - R_1 + \|\mu\|_2^2\} \\ \leq \lambda_N(X) \leq \sigma_N^2 + R_N + \|\mu\|_2^2 \ . \end{aligned} \tag{16}$$

TABLE I

Corpus and baseline model statistics for the IAM handwriting database using a recognition lexicon with 50k entries.

	Train	Dev	Test	LM data
Words	53,884	8,717	25,472	3,363,402
Characters	219,749	31,724	96,637	13,871,031
Writers	283	57	162	-
Out-of-vocabulary words (%)	-	3.94	3.42	-
WER baseline model (%)	-	32.8	39.4	-

The proof of Theorem 2 is a direct generalization of Theorem 1. In contrast to Theorem 1, only the bounds for the eigenvalues of A obtained by Geršgorin's theorem are known instead of the exact eigenvalues. For strongly correlated features corresponding to large values of R_i, the eigenvalues can be distributed almost arbitrarily according to the bounds (15) and (16). For weakly correlated features, the bounds are tighter. In particular, for normalized features and $R_1 < 1$, Theorem 2 implies:

$$1 \leq \kappa(X) \leq \frac{1 + R_N}{1 - R_1}. \tag{17}$$

This shows that the best conditioning of the optimization problem is obtained for decorrelated and normalized features.

IV. Experimental Results

In this section, we validate our theoretical result on two recognition tasks. The first one is the continuous handwriting recognition task IAM. The second is the Wall Street Journal (WSJ) task for English read speech.

A. Continuous Handwriting Recognition

The IAM database [18] is a continuous handwriting recognition task with open vocabulary. This task is of interest for us, because it requires the same techniques as an LVCSR system. The main difference of the handwriting and the speech recognition system is the feature extraction. We can use exactly the same software for speech recognition as for handwriting recognition. The corpus has a predefined subdivision into training, development and testing folds, see Table I. The amount of training data is quite small in comparison to a speech recognition system, roughly corresponding to seven hours of speech. In comparison to other handwriting recognition tasks it is considered as a large dataset.

We used the maximum likelihood GHMM system of [19] as our baseline system. In the feature extraction of this system, only elementary preprocessing steps (deslanting and size normalization) are used, which are commonly employed in image recognition. An image slice was extracted at every position. Seven features in a sliding window were concatenated and projected to a thirty dimensional vector by a principal component analysis (PCA). The 78 characters were modeled by context-independent five-state left-to-right HMMs, resulting in 390 distinct states plus one state for the whitespace model. The emission probabilities are modeled by GMMs with a pooled diagonal covariance matrix and trained with the

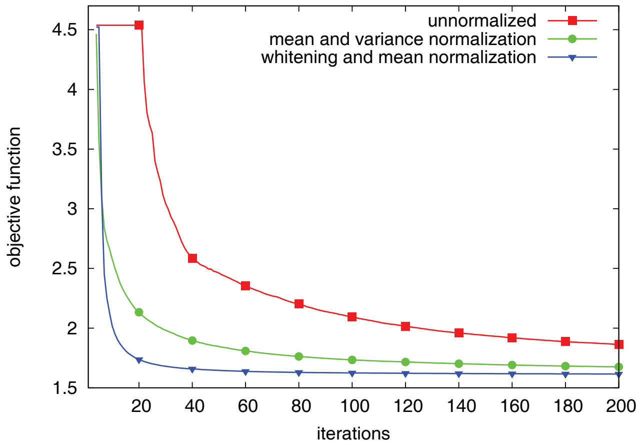

Fig. 1. Plot of the objective function versus iteration number for experiments on the IAM database with second-order features, zero initialization, and different preprocessings. Note that the objective function for the unnormalized features remains constant for the first iterations, because L-BFGS performs a backtracking.

TABLE II
RESULTS ON THE IAM DATABASE FOR TRAINING WITH L-BFGS WITH POLYNOMIAL FEATURE SPACES OF DEGREE $m \in \{1,2,3\}$ AND DIFFERENT INITIALIZATIONS AND PREPROCESSINGS ("-" CORRESPONDS TO NO PREPROCESSING, "M" TO MEAN NORMALIZATION, "V" TO VARIANCE NORMALIZATION, AND "D" TO DECORRELATION).

m	Preprocessing	Initialization	WER/dev (%)	WER/test (%)
1	-	zero / random	49.9 / 68.3	60.1 / 75.5
1	mv	zero / random	49.7 / 48.9	58.9 / 58.5
2	-	zero / random	32.4 / >100.0	40.2 / >100.0
2	mv	zero / random	30.2 / 34.4	38.5 / 41.3
2	mv	1st order	26.8	33.1
2	dmv	zero / random	25.1 / 25.9	31.6 / 32.3
3	mv	2nd order	23.0	27.4

EM algorithm with a splitting procedure. Best results were obtained with 25k mixture components in total. A Kneser-Ney smoothed trigram language model has been trained on the text data which is commonly used for this corpus. A recognition lexicon with 50k entries has been chosen on the same data. The baseline system achieves a WER of 32.8% on the development corpus and 39.4% on the test corpus. A more detailed description of the baseline system can be found in [19].

We generated a state alignment with our baseline system, and then trained the log-linear models on the resulting training sample $(x_t, s_t)_{t=1,\dots,T}$. We used the L-BFGS algorithm for training, because, it is considered as the best optimization algorithm for log-linear training in literature. We set the history length of L-BFGS to ten, which is a standard value given in literature [8], [7]. For comparison with our previous work [5], we also performed trainings with the improved Rprop algorithm proposed by [20]. For all configurations, we stopped training after 200 training iterations. At this point, the change in the objective function is small for all models.

For the log-linear models, we used polynomial features of degree one ($N = 30$), two ($N = 495$) and three ($N = 5455$). In preliminary experiments, we obtained almost no improvements by regularization. The reason for this is that with our choice of features, the frame-classification error on the training data ranges from forty to sixty percent and benefits from regularization can only be expected, when the training error is small. Therefore, we only report the results without regularization.

The results on the IAM database are summarized in Table II. Our theoretical analysis predicted the convergence behavior very well. The first-order features are already decorrelated, but without mean and variance normalization, the convergence is slower, resulting in a worse WER on development and test set. The difference is moderate, when the parameters are initialized with zero, corresponding to a uniform distribution. In a next experiment, we initialized all parameters randomly with plus or minus one. This results in a huge degradation for the unnormalized features and – with exactly the same random initialization – has only a minor impact when normalized features are used. The differences are even larger for the second-order experiments. This can be expected, since mean and variance take on more extreme values when the features are squared. Furthermore, projecting the observations to second-order polynomials introduces correlation among the features. For the zero initialization, the improvement from mean and variance normalization is only moderate in WER, although convergence is already strongly accelerated (see Figure 1). For the unnormalized features and random initialization, the optimization did not lead to a usable model for recognition at all. Fastest convergence and best results are obtained after decorrelation by means of a principal component analysis (PCA) and mean and variance normalization of the features. In addition, the influence of the initialization is the smallest in this case. Because of the high dimension of the third-order features, the estimation of the decorrelation matrix itself is already computationally very expensive. Therefore, we only performed a mean and variance normalization of the third-order features, but initialized the models incrementally from first to second to third-order features. In this manner, we obtain our best result of 27.4% WER, which is a drastic improvement over the maximum likelihood GHMM baseline system (39.4% WER).

Qualitatively, we observed the same effect of the condition number and the initialization in the experiments with Rprop. With Rprop, we obtained 31.4% WER with decorrelated second-order features and zero initialization, and 36.5% WER with normalized features and random initialization. That means, for a well-conditioned problem and a reasonable initialization, the quality of the final model is almost independent of the choice of the optimization algorithm. For the worse-conditioned problem the RProp-result is better than the L-BFGS-result. This is surprising, because L-BFGS uses an approximation to the full Hessian matrix. We suspect that the simple diagonal scaling of Rprop is more reliable for high-

TABLE III

Results on the WSJ corpus for monophone log-linear systems with polynomial features of degree $m \in \{1, 2, 3\}$ with different preprocessings (see Table II) and for the triphone systems with additional sparse posterior features.

m	sparse feat. dim.	HMM states	Preprocessing	WER/test (%)
1	-	130	-	21.7
1	-	130	mv	21.2
2	-	130	-	9.7
2	-	130	mv	9.3
2	-	130	dmv	8.8
3	-	130	-	8.6
3	-	130	mv	7.0
2	9216	1500	-	3.6
2	9216	1500	dmv	3.3

dimensional problems than the Hessian approximation used in L-BFGS.

Our log-linear system outperforms other systems based on HMMs with comparable preprocessing. Bertolami and Bunke [21] obtain 32.9% WER by a ROVER-combination of different HMM systems. Especially interesting is the comparison to the results obtained by methods which are commonly used in speech recognition. Dreuw et al. [19] obtain 31.6% WER with GHMMs with lattice-based MMI training and 30.0% with MPE training. Their system is further improved with an additional discriminative adaptation method (29.0% WER). The system of Graves [22], which has a completely different architecture based on recurrent neural networks, outperforms our system with 25.9% WER. The best published result of 21.2% WER on the IAM database is by España-Boquera et al. [23], who use several specialized neural networks for preprocessing.

B. Wall Street Journal

In order to validate our theoretical analysis on a speech recognition task, we conducted experiments on the Wall Street Journal corpus (WSJ0). The Wall Street Journal corpus consists of 15h training data and half an hour of evaluation data. Since the official corpus does not provide a development set, we extracted 410 sentences from the North American Business (NAB) task and used it as a development set.

The experimental setup is the same as in our previous work [5]. We trained a standard GHMM system according to the maximum likelihood criterion with the EM algorithm, and a log-linear system in the same manner as the handwriting recognition system on IAM. In both systems, MFCC features with vocal tract length normalization and a voicedness feature are used. Acoustic context is incorporated by using a sliding window of nine frames. The dimension of the resulting feature vector is reduced to 33 by means of a linear discriminant analysis. 1500 generalized triphones are modeled, which are obtained by a hierarchical clustering. The GMM has a pooled, diagonal covariance matrix and a total of 223k mixture components. The GHMM system achieves a WER of 3.6% WER with a trigram language model on the evaluation set. For the log-linear system, we used polynomial features and $9 \cdot 2^{10}$ sparse posterior features, which are derived as posterior probabilities of a GMM with 2^{10} components and context expansion by nine frames. The WER of the log-linear system on the evaluation set is 3.6% as well.

Considering our previous work, our results on the IAM database are surprising. In this work, we were not aware of the possible ill-conditioning of log-linear training. In our experiments, we observed that after 50 to 100 iterations with Rprop, the objective function only decreased very slowly. Therefore, we stopped the optimization at the point where the decrease in the objective function became very small. We neither applied normalizations to the polynomial features nor to the sparse posterior features. In addition, we initialized all parameters randomly, which had a very negative effect in the experiments on IAM. Nevertheless, all our experiments behaved reasonable and we could obtain competitive results on this task.

In order to further investigate this different behavior, we repeated a number of illustrative experiments from our previous work with varied optimization. We used L-BFGS for training and initialized all parameters with zero. We stopped training after 100 iterations. First, we performed experiments with polynomial features and different feature preprocessings on the monophone system. The results are summarized in Table III. The effect of the feature transformations is weaker than on the handwriting task, but well observable for all polynomial degrees. For first-order features, the WER is improved from 21.7% to 21.2% WER, and for second-order features from 9.7% to 9.3% (mean and variance normalization) and 8.8% WER (decorrelation and mean and variance normalization). As expected, the effect is strongest for third-order features. Here, the WER is improved from 8.6% to 7.0% by mean and variance normalization and initialization with the second-order model.

The explanation for the better convergence behavior on WSJ is simple. In our feature extraction, we used mean and variance normalization of the MFCC features on segment level (but not for the voicedness feature), which is a common technique in speech recognition for improving generalization of the models. The normalization of the basic features partly carries over to the higher-order polynomial features. For this reason, unintentionally, the condition of the optimization problem was strongly improved. Nevertheless, additional gains are obtained by the application of feature transformations.

For our final triphone system, we used sparse posterior features in addition to the polynomial features. According to our convergence analysis, the use of sparse posterior features is not only advantageous because of the lower costs per iteration, but also because they result in a better convergence behavior. First, the features are only weakly correlated. Second, because of their definition as posterior probabilities of clusters with similar prior probabilities, their means are close to zero and their variances are similar. Therefore, convergence is fast, when sparse posterior features are used. Further accelerating convergence by shifting the mean or decorrelation is not possible without loosing the sparsity of the features. Therefore, we applied the feature transformations only to the polynomial features. Surprisingly, for the setups with sparse features, Rprop converged much faster than L-BFGS. We obtained our best result of 3.3% WER on the evaluation set after only 20 Rprop iterations. This is a slight improvement over the log-linear system without decorrelation of the second-order features after 75 Rprop iterations and the GHMM baseline system (3.6% WER both) as well. More importantly, the reduced training time allows for the application of the log-linear training on larger datasets.

V. Discussion

In this paper, we presented a convergence analysis for the optimization of the parameters of log-linear models. We showed that the convergence behavior of log-linear training depends on the mean and variance of the features and the correlation among the features. In extreme cases, the optimization problem can be highly ill-posed. Conversely, our analysis shows that log-linear training can be preconditioned by feature transformations.

We verified our findings on a continuous handwriting recognition task and a large vocabulary continuous speech recognition task. We found that the theoretical analysis is in accordance to the experimental observations. The effect of the feature normalizations was very strong on the handwriting recognition task IAM, where the improvement of the condition number of the optimization problem was essential for obtaining a competitive result. On the speech recognition task WSJ, the effect was less pronounced, because common normalization techniques in speech recognition already alleviate the ill-conditioning of the optimization problem. Furthermore, the use of sparse posterior features on the WSJ setup was beneficial for the convergence behavior. Still, we could improve our log-linear speech recognition system on WSJ and strongly reduce training time. The log-linear speech recognition system now slightly outperforms the conventional maximum likelihood GHMM system on WSJ. In contrast to other methods for accelerating log-linear training, e.g. feature selection, the acceleration of training is achieved by a transformation to an equivalent optimization problem and therefore does not require any approximation or heuristic. In future work, we want to evaluate log-linear models on larger tasks, which is now tractable with the accelerated training.

We found that for experiments with sparse features, Rprop converges much faster than L-BFGS, which is commonly regarded as the best optimization algorithm for log-linear training. This may be due to the much higher feature dimension when sparse features are used. A detailed analysis of the convergence properties of Rprop and a comparison to numerical optimization algorithms with a better theoretical foundation as L-BFGS would therefore be valuable.

Acknowledgment — We would like to thank Alexander Richard for his support. This work was partly realized as part of the Quaero Programme, funded by OSEO, French State agency for innovation.

References

[1] A. Gunawardana, M. Mahajan, A. Acero, and J. C. Platt, "Hidden conditional random fields for phone classification," in *Proc. Interspeech*, 2005, pp. 1117–1120.

[2] Y. Hifny and S. Renals, "Speech recognition using augmented conditional random fields," *IEEE Trans. Audio Speech Lang. Process.*, vol. 17, no. 2, pp. 354–365, 2009.

[3] E. Fosler Lusier and J. Morris, "Crandem systems: Conditional random field acoustic models for hidden Markov models," in *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing, ICASSP*, 2008, pp. 4049–4052.

[4] G. Zweig and P. Nguyen, "A segmental crf approach to large vocabulary continuous speech recognition," in *IEEE Automatic Speech Recognition and Understanding Workshop*. IEEE, 2009, pp. 152–157.

[5] S. Wiesler, M. Nußbaum, G. Heigold, R. Schlüter, and H. Ney, "Investigations on features for log-linear acoustic models in continuous speech recognition," in *IEEE Automatic Speech Recognition and Understanding Workshop*, 2009.

[6] A. McCallum, D. Freitag, and F. Pereira, "Maximum entropy markov models for information extraction and segmentation," in *Proceedings of the 17th International Conference on Machine Learning*, 2000, pp. 591–598.

[7] F. Sha and F. Pereira, "Shallow parsing with conditional random fields," in *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology*, 2003, pp. 134–141.

[8] R. Malouf, "A comparison of algorithms for maximum entropy parameter estimation," in *Proceedings of the Sixth Conference on Natural Language Learning*, 2002, pp. 49–55.

[9] M. Mahajan, A. Gunawardana, and A. Acero, "Training algorithms for hidden conditional random fields," in *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing, ICASSP*, vol. 1, 2006, pp. 273–276.

[10] T. Minka, "Algorithms for maximum-likelihood logistic regression," Carnegie Mellon University, Tech. Rep., 2001.

[11] S. Boyd and L. Vandenberghe, *Convex Optimization*. Cambridge University Press, 2004.

[12] J. Nocedal and S. Wright, *Numerical Optimization*. Springer, 1999.

[13] S. Wiesler, P. Dreuw, and H. Ney, "A convergence analysis of log-linear training," in *submitted to Advances in Neural Information Processing Systems*, 2011.

[14] C. Sutton and A. McCallum, "An introduction to conditional random fields for relational learning," in *Introduction to Statistical Relational Learning*, L. Getoor and B. Taskar, Eds. MIT Press, 2007.

[15] J. Lafferty, A. McCallum, and F. Pereira, "Conditional random fields: Probabilistic models for segmenting and labeling sequence data," in *Proceedings of the 18th International Conference on Machine Learning*, 2001, pp. 282–289.

[16] R. Horn and C. Johnson, *Topics in Matrix Analysis*. Cambridge University Press, 1994.

[17] ——, *Matrix Analysis*. Cambridge University Press, 2005.

[18] U. Marti and H. Bunke, "The IAM-Database: An English Sentence Database for Offline Handwriting Recognition," *Int. J. Doc. Anal. Recogn.*, vol. 5, no. 1, pp. 39–46, 2002.

[19] P. Dreuw, G. Heigold, and H. Ney, "Confidence- and Margin-Based MMI/MPE Discriminative Training for Off-Line Handwriting Recognition," *Int. J. Doc. Anal. Recogn.*, pp. 1–16, 2011.

[20] C. Igel and M. Hsken, "Empirical evaluation of the improved rprop learning algorithm," *Neurocomputing*, vol. 50, p. 2003, 2003.

[21] R. Bertolami and H. Bunke, "HMM-based Ensamble Methods for Offline Handwritten Text Line Recognition," *Pattern Recogn.*, vol. 41, pp. 3452–3460, 2008.

[22] A. Graves, M. Liwicki, S. Fernandez, R. Bertolami, H. Bunke, and J. Schmidhuber, "A Novel Connectionist System for Unconstrained Handwriting Recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 31, no. 5, pp. 855–868, May 2009.

[23] S. España-Boquera, M. Castro-Bleda, J. Gorbe-Moya, and F. Zamora-Martinez, "Improving Offline Handwritten Text Recognition with Hybrid HMM/ANN Models," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 33, no. 4, pp. 767 –779, april 2011.