DISCRIMINATIVE TRAINING BY ITERATIVE LINEAR PROGRAMMING OPTIMIZATION

Brian Mak, Benny Ng

Department of Computer Science and Engineering
The Hong Kong University of Science and Technology
Clear Water Bay, Hong Kong
{mak, bennyng}@cse.ust.hk

ABSTRACT

In this paper, we cast discriminative training problems into standard linear programming (LP) optimization. Besides being convex and having globally optimal solution(s), LP programs are well-studied with well-established solutions, and efficient LP solvers are freely available. In practice, however, one may not have complete knowledge of the feasible region since it is constructed from a limited number of competing hypotheses based on the current model — not the final model which, by definition, is not known *a priori* at the time of hypotheses generation. We investigate an iterative LP optimization algorithm in which an additional constraint on the parameters being optimized is further imposed. Our proposed method is evaluated on the estimation of global and state-dependent stream weights and biases of a multi-stream hidden Markov model system. Results show that the stream weights and biases found by our iterative LP optimization algorithm may give better recognition performance than the ones found by a brute-force grid search.

Index Terms: iterative linear programming, discriminative training, multi-stream HMM, parameter tying.

1. INTRODUCTION

Linear weighting functions are commonly found in automatic speech recognition (ASR). For instance, in a multi-stream hidden Markov model (HMM) (which is used in discrete HMM system [1, 2], multi-band ASR [3], and audio-visual ASR [4, 5]), the state log-likelihood is usually computed as a linear combination of the per-stream state log-likelihoods; in Viterbi search, the recognition score of a test utterance is a linearly weighted sum of the acoustic score and language score. It is known that the parameter of these linear functions cannot be determined by maximum-likelihood estimation as it will simply give all the weights to the most probable factor. As a consequence, discriminative training is commonly employed to estimate these linear functions in ASR by minimizing the classification errors (MCE), by maximizing the mutual information (MMI), or by maximizing the entropy (MAXENT), and so forth.

In this paper, we propose to improve the estimation of these linear functions by casting the respective problems into the standard linear programming (LP) optimization framework. Our proposed LP approach has the following advantages over other discriminative training schemes:

- Since linear programming problem is convex, the solution(s) is/are guaranteed to be globally optimal (though they may not be unique).
- Unlike some other discriminative training methods like MCE estimation that perform corrective training, LP utilizes all training data — both correct and incorrect data — to perform the joint optimization of the parameters in a linear function.
- There are fewer system parameters to tune. For instance, unlike gradient-based solutions, there is no learning rate to tune; also, unlike MCE training which introduces nonlinearity through the use of the sigmoid function, it does not need to tune the sigmoid function parameters.
- LP optimization is very well studied with established solutions, and efficient LP solvers are freely available.

Since LP solutions are supposed to be globally optimal, it requires complete knowledge of the feasible region (for searching the solution). However, as we will illustrate in the rest of the paper, the feasible region has to be constructed from the competing hypotheses in relevant ASR problems. Since the competing hypotheses are generated by the current model, and usually from N-best list or decoding lattice, they are not complete (due to pruning) and they are unlikely the same as those of the final model, which, by definition, cannot be known *a priori*. Hence, the globally optimal solution found using the competing hypotheses generated by the current model may not be really the true optimal solution. We devise an iterative LP optimization algorithm and impose an additional constraint on how much the parameters being optimized can be changed in each iteration.

To illustrate the idea of our novel iterative LP optimization, we will show how the stream weights and biases of a multi-stream HMM can be estimated using the new approach below by considering frame or word recognition correctness.

2. FORMULATION OF THE ESTIMATION OF STREAM WEIGHTS/BIASES AS AN LP PROBLEM

For simplicity, we will formulate the estimation of stream weights and biases of a multi-stream HMM by first considering the recognition correctness at the frame level, and then extend the formulation to the word level.

2.1. Based on Frame Recognition Correctness

Let us denote an observation vector at time t as $\mathbf{x}_t$ and the HMM state that generates it as y_t for $t = 1, \ldots, T$. Using the common HMM notation, the probability of $\mathbf{x}_t$ at state j is given by $b_j(\mathbf{x}_t)$. In

This work was supported by the Research Grants Council of the Hong Kong SAR under the grant numbers HKUST617406 and HKUST617507.

a multi-stream HMM with K streams and N states, the state pdf is represented by a factored pdf as follows:

$$b_j(\mathbf{x}_t) = c_j \prod_{k=1}^{K} b_j^{(k)}(\mathbf{x}_t^{(k)})^{w_j^{(k)}} \ , \tag{1}$$

or equivalently in the log domain as

$$\log b_j(\mathbf{x}_t) = \log c_j + \sum_{k=1}^{K} w_j^{(k)} \log b_j^{(k)}(\mathbf{x}_t^{(k)}) \tag{2}$$

where $\mathbf{x}_t^{(k)}$ is the feature vector of the kth stream; $b_j^{(k)}(\mathbf{x}_t^{(k)})$ is the state observation probability of $\mathbf{x}_t^{(k)}$ in the kth stream; $w_j^{(k)}$ is the weight of the kth stream in state j with the constraint that $\sum_{k=1}^{K} w_j^{(k)} = 1$; c_j is the normalization factor to make the right-hand side of Eqn. (1) a true probability density function. Theoretically, we should have

$$\int_{\mathbf{x}_t^1} \int_{\mathbf{x}_t^2} \cdots \int_{\mathbf{x}_t^K} c_j \prod_{k=1}^{K} b_j^{(k)}(\mathbf{x}_t^{(k)})^{w_j^{(k)}} d\mathbf{x}_t^1 d\mathbf{x}_t^2 \cdots d\mathbf{x}_t^K = 1 \ .$$

However, that will render the problem intractable. In this paper, we will not pursue this requirement, and we will treat $\log c_j$ as a bias for each state and $\log b_j(\mathbf{x}_t)$ should be treated more as a likelihood score than a strict probability term.

Furthermore, with the following variable substitutions[1]:

$$\begin{aligned} \mathbf{w}_j &= [w_j^{(1)}, w_j^{(2)}, \cdots, w_j^{(K)}]' \ , \\ \mathbf{z}_{jt} &= [\log b_j^{(1)}(\mathbf{x}_t^{(1)}), \log b_j^{(2)}(\mathbf{x}_t^{(2)}), \cdots, \log b_j^{(K)}(\mathbf{x}_t^{(K)})]' \ , \\ v_j &= \log c_j \ , \end{aligned}$$

Eqn. (2) can be expressed in vector form as

$$\log b_j(\mathbf{x}_t) = \mathbf{w}_j' \mathbf{z}_{jt} + v_j \ . \tag{3}$$

2.1.1. The Basic Requirement

For each training frame $\mathbf{x}_t = [\mathbf{x}_t'^{(1)}, \ldots, \mathbf{x}_t'^{(K)}]'$ belonging to the truth state y_t, we would like its probability computed by the truth state to be greater than its probability computed by any other competing state. That is,

$$\begin{aligned} \forall j \neq y_t \qquad & \log b_{y_t}(\mathbf{x}_t) - \log b_j(\mathbf{x}_t) \geq 0 \\ \Rightarrow \quad & (\mathbf{w}_{y_t}' \mathbf{z}_{y_t t} - \mathbf{w}_j' \mathbf{z}_{jt}) + (v_{y_t} - v_j) \geq 0 \ . \end{aligned}$$

To allow possible "noise" in the training data, we may relax the requirement by introducing *slack variables* $\xi_{tj} \geq 0$, and require

$$(\mathbf{w}_{y_t}' \mathbf{z}_{y_t t} - \mathbf{w}_j' \mathbf{z}_{jt}) + (v_{y_t} - v_j) + \xi_{tj} \geq 0 \ . \tag{4}$$

The slack variables basically implements the hinge loss function so that their values for correctly recognized frames are zero, and their values for incorrectly recognized frames are positive. From another point of view, the slack variables are a measure of frame recognition errors.

[1] In this paper, vector quantities are written in bold.

2.1.2. LP Form

From Eqn. (4), we may formulate the estimation of the stream weight vector $\mathbf{w}_j$ as a standard LP problem as follows:

$$\min_{\mathbf{w}_j, v_j} \sum_t \sum_{j \neq y_t} \xi_{tj} \tag{5}$$

such that

$$\forall t, \ \forall j \neq y_t, \ (\mathbf{w}_{y_t}' \mathbf{z}_{y_t t} - \mathbf{w}_j' \mathbf{z}_{jt}) + (v_{y_t} - v_j) + \xi_{tj} \geq 0 \ , \tag{6}$$

$$\forall t, \ \forall j \neq y_t, \ \xi_{tj} \geq 0 \ , \tag{7}$$

$$\forall j, \ \sum_{k=1}^{K} w_j^{(k)} = \text{constant} \ , \tag{8}$$

$$\forall j, \ \forall k, \ w_j^{(k)} \geq 0 \ . \tag{9}$$

Note that

- although we formulate the problem with state-dependent weights $\mathbf{w}_j$, state-dependent biases v_j, and frame-and-state-dependent error ξ_{tj}, one may tie these parameters to provide various degrees of smoothing for his problem at hand.
- this is not corrective training; all training frames are used. The LP solver will find optimal weights that will increase the number of correctly recognized frames, and reduce the log likelihood difference between the correct state and competing states for the incorrectly recognized frames.
- in the basic setting, the sum of weights is set to 1; in the LP formulation, we only require the sum to be a constant.

2.2. Based on Word Recognition Correctness

The LP formulation in Eqn. (5) can easily be extended to consider recognition accuracy at the word level. For the ith instance X_{mi} of the word, $X_m, m = 1, \ldots, M$, where M is the vocabulary size, we would like to have its probability given by the HMM λ_m of the word X_m greater than that of all its competing hypotheses. That is, if the function $\mathcal{T}(\cdot)$ maps an instance of a word in an utterance to its time span, $\bar{\lambda_m}$ represents the models in a competing hypothesis, $\bar{y_t}$ represents the state in the competing hypothesis at time t, and we ignore the contribution of likelihoods due to the transition probabilities, we have,

$$\begin{aligned} \forall i, \forall m \qquad & \log P(X_{mi}|\lambda_m) - \log P(X_{mi}|\bar{\lambda_m}) \geq 0 \\ \Rightarrow \quad & \textstyle\sum_{t \in \mathcal{T}(X_{mi})} \left[(\mathbf{w}_{y_t}' \mathbf{z}_{y_t t} - \mathbf{w}_{\bar{y_t}}' \mathbf{z}_{\bar{y_t} t}) + (v_{y_t} - v_{\bar{y_t}}) \right] \geq 0 \ . \end{aligned}$$

Thus, if we assume that the slack variable is tied at the word level, then the corresponding LP problem is,

$$\min_{\mathbf{w}_j, v_j} \sum_i \sum_m \xi_m \tag{10}$$

such that

$$\forall i, \forall m, \forall \bar{m}, \quad \sum_{t \in \mathcal{T}(X_{mi})} \left[(\mathbf{w}_{y_t}' \mathbf{z}_{y_t t} - \mathbf{w}_{\bar{y_t}}' \mathbf{z}_{\bar{y_t} t}) + (v_{y_t} - v_{\bar{y_t}}) \right] + \xi_m \geq 0 \ , \tag{11}$$

$$\forall m, \ \xi_m \geq 0 \ , \tag{12}$$

$$\forall j, \ \sum_{k=1}^{K} w_j^{(k)} = \text{constant} \ , \tag{13}$$

$$\forall j, \ \forall k, \ w_j^{(k)} \geq 0 \ . \tag{14}$$

2.2.1. Iterative LP Optimization

There is one problem with the formulation as described in Eqns. (10 – 14). LP optimization gives a globally optimal solution in the feasible region. However, in speech recognition, we can only generate the competing hypotheses — hence the feasible region — using the current models. Once we change the stream weights and biases according to the LP solution, there will be a new set of models, and they may give a different feasible region as they will probably generate a different set of hypotheses. Thus, unless we have complete knowledge of the feasible region, the globally optimal solution given by LP optimization is only correct with respect to the feasible region created by the current set of competing hypotheses. Other discriminative training methods such as MCE have the same problem, but the problem is not as serious as ours since those methods do not find a globally optimal solution in an iteration but try to approach the local optimum slowly in an iterative algorithm. Here, we investigate an iterative LP optimization approach as follows: The LP optimization will run iteratively. In each iteration, the competing hypotheses are generated by the current model, and LP optimization is performed with an additional constraint to control the amount of change in $\mathbf{w}$ as follows:

$$\forall j, \forall k, \quad \Delta w_j^{(k)} \leq \Delta w_{max}. \tag{15}$$

Then, a new model is obtained with the new $\mathbf{w}$, which then is used to generate a new set of competing hypotheses, and the algorithm is repeated. By carefully controlling how much the parameters being optimized (Δw_{max} here) can change in each iteration, it is hoped that the locally optimal solution in each LP iteration will converge to the globally optimal solution.

Table 1. Word accuracy of the baseline model and the model with the "best" global stream weights found by grid search.

CDHMM	Word Accuracy
4-stream, global weights, $\forall j, \forall k, w_j^{(k)} = 1$	91.43%
4-stream, global weights found by grid search	**92.23%**

Table 2. Effect of tying stream weights when the LP is formulated in terms of frame recognition correctness, and solved using 3,600 training frames.

Weight Tying	Word Accuracy
state-dependent	90.86%
phoneme-dependent	91.14%
global	**92.19%**

3. EXPERIMENTAL EVALUATION

Discriminative training by the proposed (iterative) linear programming (LP) optimization approach was evaluated in the estimation of stream weights and biases for a 4-stream continuous density HMM (CDHMM) system. Monophone HMMs were used so that when the LP approach was formulated in terms of frame recognition correctness, the number of competing states, being 143, was small enough that complete knowledge of the feasible region could be constructed, and only *one iteration* of the LP optimization was needed to obtain the globally optimal solution. On the other hand, when the LP approach was formulated in terms of word recognition correctness, incomplete knowledge of the feasible region, which was constructed from N-best competing hypotheses, required iterative LP optimization with an additional constraint on the maximum amount of change in weights and biases.

3.1. Baseline Systems

The speaker-independent (SI) training set of the Resource Management Corpus (RM1) was used for training the SI model. It consists of 3990 utterances from 109 speakers. Evaluation was done on the 300 utterances in the SI Feb'91 test set using the standard word-pair grammar with a perplexity of 60. All model training and decoding was performed using the HTK software.

The conventional 39-dimensional MFCC vectors were extracted at every 10ms over a window of 25ms. Each MFCC vector was split into 4 streams: static MFCCs, delta MFCCs, delta-delta MFCCs, and energies respectively. The SI model consists of 47 monophones plus the silence and short pause. Each of them was modeled as a 4-stream CDHMM which is strictly left-to-right and has three states with 10 Gaussian mixture components per state. The models were trained by fixing all stream weights $w_j^{(k)}$ and biases to 1.

We also tried to locate the optimal global stream weights through an extensive grid search in the numerical region of 0.7 – 1.5 for all the stream weights simultaneously. Table 1 shows the word recognition accuracies of the 4-stream baseline HMMs with global stream weights set to 1, or found by brute-force grid search. It can be seen that better stream weights are found by the grid search, which reduce the word error rate (WER) of the baseline system by 9.33%.

3.2. Experiment 1: LP Optimization with Complete Knowledge of the Feasible Region Based on Frame Recognition Correctness

Thirty-six seconds or 3,600 frames of speech were randomly selected from the training set so that the amount of training frames for each truth state[2] was the same. The truth state of a frame was obtained from forced alignments of the training utterances using the baseline 4-stream 10-mixture CDHMMs with uniform stream weights and bias of 1. The log likelihood vectors $\mathbf{z}_{jt}$ were then computed from each frame accordingly for all 144 possible state. Thus, each truth state likelihood had 143 competing state likelihoods at each frame.

The sum of weights at any state $\sum_{k=1}^{K} w_j^{(k)}$ was set to 4. Various tying schemes of the weights and biases at the global, phoneme, and state level were considered. For the slack variables, we also tried no tying at all, or tying at the frame, phoneme, and state level[3]. Finally, the LP problem was solved by the Mosek software [6] using the interior-point method.

It is found that global stream biases and frame-dependent slack variables give the best results regardless how stream weights are tied. Table 2 presents the results when the stream weights were tied in various levels with global stream biases and frame-dependent slack variables. It can be seen that global stream weights give the best accuracy of 92.19%. We further ran experiment for this best setting with more training frames, and confirmed that the recognition performance had already converged, though a slightly better performance of 92.23% — which is the same as the result obtained through computationally extensive grid search — could be obtained with 7,200 or more training frames.

[2]There are totally $3 \times 47 + 3 = 144$ states in the HMMs.

[3]The frame-level tying of the slack variables means that for a training frame $\mathbf{x}_t$, its log likelihood from the truth state has to be better than *all* competing states — both the nearby and the farthest competing states — by the same amount.

Table 3. Estimation of state-dependent stream weights when the LP is formulated in terms of word recognition correctness using (one iteration of) LP without any constraint on $\Delta\mathbf{w}$.

Size of N-best List	Word Accuracy
10	89.98%
25	90.66%
50	90.50%
250	90.86%
800	90.90%

3.3. Experiment 2: Iterative LP with Incomplete Knowledge of the Feasible Region Based on Word Recognition Correctness

The LP formulation of Experiment 1 is based on frame recognition correctness. It does not match with the common WER performance measure in ASR, though usually we expect increasing frame accuracy to give non-decreasing word accuracy. Here, we repeated the LP optimization but minimized the word recognition errors to estimate the $144 * 4 = 576$ state-dependent stream weights; stream biases were assumed global. In the following experiments, the initial HMMs used the stream weights found by Experiment 1 to compute the competing hypotheses.

3.4. Experiment 2.1: Single LP Iteration with Different Sizes of N-best Lists

We first investigated the conjecture that the feasible region constructed by an N-best list generated by a given model is incomplete. Only one LP iteration was run as in Experiment 1 with no further constraint on $\Delta\mathbf{w}$. The results with different values of N are shown in Table 3. It is observed that the global solution found by one iteration of LP optimization gives worse results than the baseline system. Increasing the number of competing hypotheses helps slightly but the performance is still unsatisfactory.

3.5. Experiment 2.2: Iterative LP Optimization

Experiment 2.1 was repeated using the iterative LP algorithm described in Section 2.2.1 with 50-best competing hypotheses, and constraining the change in all stream weights to be less than Δw_{max}. The results with varied values of Δw_{max} are plotted in Fig. 1. It is found that the iterative LP algorithm effectively improves the estimation of stream weights. A smaller value of Δw_{max} gives better convergence performance in a few iterations. The best state-dependent stream weights were obtained with $\Delta w_{max} = 0.01$, giving a word accuracy of 92.79% which is better than the result obtained with global stream weights found by grid search.

4. CONCLUSIONS

We investigate the use of standard linear programming (LP) optimization for discriminative training. We analyze the problem of incomplete knowledge of the feasible region that can be constructed from competing hypotheses in practical ASR system, and propose an iterative LP optimization algorithm. It is empirically found that the proposed LP approach is effective in estimating the stream weights for a multi-stream HMM system either with (1) complete knowledge of the feasible region when it is formulated on frame recognition correctness in a single iteration, or (2) incomplete knowledge of the feasible region when it is formulated on word recognition correctness in several iterations with further constraint on the change in weights. Weights estimated by (1) give the same performance as the "optimal" global weights found by extensive (and computationally expensive) grid search, whereas weights estimated by (2) give even better performance.

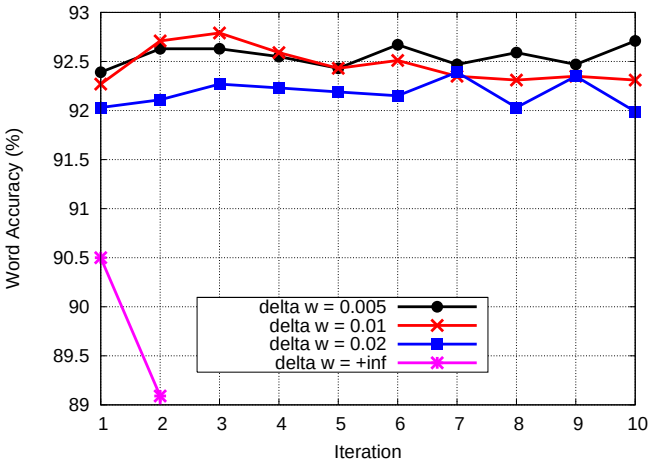

Fig. 1. Effect of $\Delta\mathbf{w}$ on iterative LP optimization.

It is worth noting that in [7], state-dependent stream weights perform worse than global stream weights when they are trained by maximum-entropy estimation. In light of the result, our findings are encouraging. Future work will incorporate the estimation of state biases into the iterative LP framework.

We would like to emphasize that the RM speech recognition task used in this preliminary study is not the intended application of the new method. It is only used as an example to show that our new algorithm is effective in estimating stream weights, and should be able to improve the performance of any multi-stream classifiers, such as the product HMM in audio-visual ASR.

5. REFERENCES

[1] K. F. Lee, "Context-dependent phonetic hidden Markov models for speaker-independent continuous speech recognition," *IEEE Trans. on ASSP*, vol. 38, no. 4, pp. 599–609, April 1990.

[2] Brian Mak, S. K. Au Yeung, Y. P. Lai, and M. Siu, "High-density discrete HMM with the use of scalar quantization indexing," in *Proc. of Eurospeech*, Lisbon, Portugal, Sept 2005.

[3] H. Bourlard and S. Dupont, "A new ASR approach based on independent processing and recombination of partial frequency bands," in *Proc. of ICSLP*, October 1996.

[4] G. Potamianos and H. P. Graf, "Discriminative training of HMM stream exponents for audio-visual speech recognition," in *Proc. of ICASSP*, 1998, pp. 3733–3736.

[5] S.Tamura, K. Iwano, and S. Furui, "A stream weight optimization method for audio-visual speech recognition using multi-stream HMMs," in *Proc. of ICASSP*, 2004, vol. 1, pp. 857–860.

[6] http://www.mosek.com.

[7] G. Gravier, S. Axelrod, G. Potamianos, and C. Neti, "Maximum entropy and MCE based HMM stream weight estimation for audio-visual ASR," in *Proc. of ICASSP*, 2002, vol. 1, pp. 853–856.