Introduction

1. **Objectives**
   To improve the generalization of the existing ILP method in determining the values of decoding parameters — grammar factor and word insertion penalty.

2. **Motivation**
   - The solution found by the current ILP algorithm when the training data do not match well with the test data is significantly worse than under matched condition.
   - In modern machine learning, generalization is often achieved by increasing the margin of the classifier.

3. **Math Tools**:
   - Iterative linear programming(LP).
   - Large-margin training.

Decoding Parameters

- Given a sequence of T acoustic observations, $\mathbf{x}_1^T = \{\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_T\}$, find the corresponding N-word sequence, $\hat{\mathbf{w}}_1^N = \{\hat{\mathbf{w}}_1, \hat{\mathbf{w}}_2, \cdots, \hat{\mathbf{w}}_N\}$.

- In theory, using the MAP decision rule:

$$\hat{\mathbf{w}}_1^N = \underset{\mathbf{w}_1^N, N}{\operatorname{argmax}} \underbrace{\ln p(\mathbf{x}_1^T | \mathbf{w}_1^N)}_{\text{acoustic score}} + \underbrace{\ln p(\mathbf{w}_1^N)}_{\text{language score}} \quad .$$

- In practice, need adjustment because of the very different dynamic ranges of acoustic score and language score.

$$\hat{\mathbf{w}}_1^N = \underset{\mathbf{w}_1^N, N}{\operatorname{argmax}} \left\{ \ln p(\mathbf{x}_1^T | \mathbf{w}_1^N) + K_{gf} \ln p(\mathbf{w}_1^N) + K_{wip} N \right\}$$

where, K_{gf} = grammar factor; K_{wip} = word insertion penalty.

- Let $\hat{\mathbf{w}}_i$ = correct transcription of the ith training utterance $\mathbf{x}_i$
  $\mathbf{w}_{ij}$ = jth competing word sequence of $\mathbf{x}_i$ during decoding.
- Linear discriminants: $\boxed{\forall i, \forall j, \quad \ln p(\hat{\mathbf{w}}_i|\mathbf{x}_i) \geq \ln p(\mathbf{w}_{ij}|\mathbf{x}_i).}$

That is,

$$\begin{aligned}
\forall i, \forall j, \qquad & \ln p(\hat{\mathbf{w}}_i|\mathbf{x}_i) - \ln p(\mathbf{w}_{ij}|\mathbf{x}_i) \geq 0 \\
\Leftrightarrow \quad & [\ln p(\mathbf{x}_i|\hat{\mathbf{w}}_i) - \ln p(\mathbf{x}_i|\mathbf{w}_{ij})] + \\
& K_{gf}[\ln p(\hat{\mathbf{w}}_i) - \ln p(\mathbf{w}_{ij})] + K_{wip}(N_i - N_{ij}) \geq 0 \\
\Leftrightarrow \quad & u_{ij} + K_{gf} v_{ij} + K_{wip} z_{ij} \geq 0
\end{aligned}$$

where

$$\begin{aligned}
u_{ij} &= \ln p(\mathbf{x}_i|\hat{\mathbf{w}}_i) - \ln p(\mathbf{x}_i|\mathbf{w}_{ij}) \\
v_{ij} &= \ln p(\hat{\mathbf{w}}_i) - \ln p(\mathbf{w}_{ij}) \\
z_{ij} &= N_i - N_{ij}
\end{aligned}$$

LP Formulation

- To allow possible violations, relax the requirement by introducing slack variables, $\xi_{ij} \geq 0$, so that

$$u_{ij} + K_{gf} v_{ij} + K_{wip} z_{ij} + \xi_{ij} \geq 0 \ .$$

- The slack variables implements the hinge loss function: for correctly decoded utterances, $\xi_{ij} = 0$.
- ξ_{ij} is also an approximation of the utterance recognition error.
- LP form:

$$\min_{K_{gf}, K_{wip}} \sum_i \xi_{ij}$$

subject to the following constraints :

$$\begin{aligned} \forall i,\ \forall j, \quad u_{ij} + K_{gf} v_{ij} + K_{wip} z_{ij} + \xi_{ij} &\geq 0 \ , \\ \forall i,\ \forall j, \quad \xi_{ij} &\geq 0 \ , \\ K_{gf} &\geq 0 \ . \end{aligned}$$

Min-max Iterative Linear Programming (ILP)

Min-max Training Approach

- $\xi_{ij} \rightarrow \xi_i$: Tie the "errors" ξ_{ij} across all competitors (j's)
  $\Rightarrow$ minimize the maximum "error" for each training utterance (i.e. worst case or strongest competitor).

Iterative Linear Programming

- Incomplete knowledge of the feasible region. Two reasons:
  1. infinite competing word sequences for an utterance!
     - N-best solution only gives a subset of them.
  2. the feasible region is only an approximation since the N-best solution is not computed with the true decoding parameters $\Rightarrow$ non-optimal solution.
- Solution: don't move to the globally optimal solution of LP. Instead, constrain the change to the decoding parameters:

$$\begin{aligned} |K_{gf}(n+1) - K_{gf}(n)| &\leq \Delta K_{gf\,max} \\ |K_{wip}(n+1) - K_{wip}(n)| &\leq \Delta K_{wip\,max} \end{aligned}$$

and update the estimates in a number of LP iterations.

The Learning Algorithm

Step 0. Set the iteration index $n = 0$, and determine

- initial values of $K_{gf}(0)$ and $K_{wip}(0)$.
- $\Delta K_{gf\,max}$ and $\Delta K_{wip\,max}$.
- maximum number of iterations n_{max}.
- convergence measure θ.

Step 1. N-best decoding for each training utterance using the current decoding parameters, $K_{gf}(n)$ and $K_{wip}(n)$.

Step 2. Compute acoustic score difference u_{ij}, language score difference v_{ij}, and the number of words difference z_{ij}.

Step 3. Construct the LP cost and constraints and add:

$$|K_{gf}(n+1) - K_{gf}(n)| \leq \Delta K_{gf\,max} \tag{1}$$

$$|K_{wip}(n+1) - K_{wip}(n)| \leq \Delta K_{wip\,max} . \tag{2}$$

Step 4. Solve the LP problem of Step 3.

Step 5. If the relative change of $\sqrt{K_{gf}(n)^2 + K_{wip}(n)^2} \leq \theta$, or n_{max} is reached, stop.

Step 6. Set $n = n + 1$, and go to Step 1.

Large Margin Training

- It is a regularization method that avoids overfitting of the training data.
- For each utterance, the recognition score of the correct word sequence is required to be greater than any of its competing word sequences by a positive margin $M \geq 0$.
- We modify the constraints as follows:

$$\forall i,\ \forall j, \quad u_{ij} + K_{gf} v_{ij} + K_{wip} z_{ij} + \xi_{ij} \geq M \ .$$

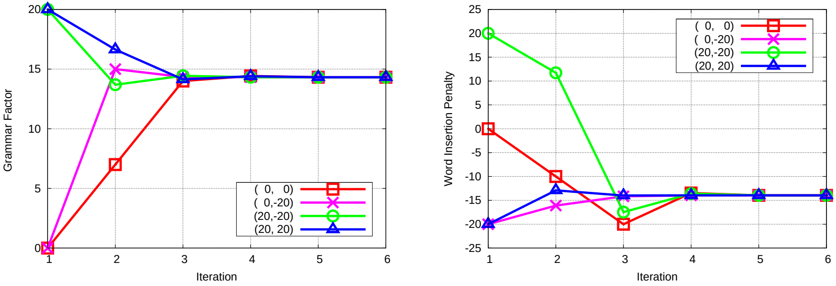
- When the margin M is greater than a certain value, all the constraints are satisfied. Therefore, the estimated parameters converge with increasing margin.
- M may be determined by cross validation using extra development data. But in practice, a very big M suffices.

Evaluation on WSJ0

Item	WS0
# train utt.	7138
# test utt.	330
# dev. utt.	442
acoustic unit	cross-word triphones
# acoustic models	15,449
# HMM states	3,132 (tied)
# Gaussians/state	16
language model	bigram
LM perplexity	111 (test), 121 (dev)
baseline word acc. by grid search	93.16%

- acoustic feature: standard 39-dimensional MFCC vectors .
- competing hypotheses: found by N-best decoding with $N = 20$.
- LP solver: the Mosek optimization software.
- $\Delta K_{gf\,max} = 7$, $\Delta K_{wip\,max} = 10$.
- maximum number of iterations $= 10$.
- convergence threshold $\theta = 10^{-4}$.

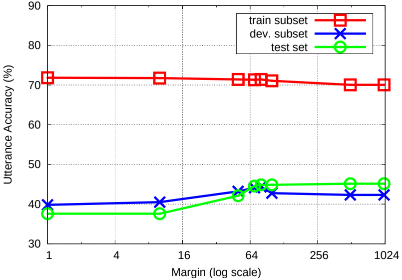
- M=80



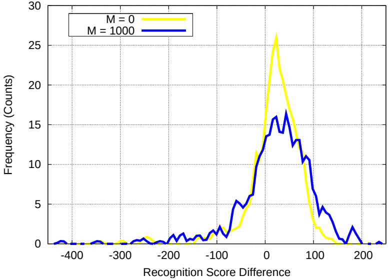
(a) Grammar Factor

(b) Insertion Penalty

Effect of Large Margin



(a) Utterance Accuracy vs M



(b) Distribution of Score Diff.

Comparison Among Different Approaches

Table: Recognition performance on the standard Nov'92 test set(330).

Training Set (#Utt.)	Method	Word (Utt.) Acc. (%)
test set (330)	grid search	93.16 (44.55)
dev subset (442)	grid search	92.92 (44.55)
dev subset (442)	ILP (M=0)	92.53 (42.42)
train subset (1175)	ILP (M=0)	91.72 (37.58)
dev subset (442)	LMILP ($M=\infty$)	92.86 (45.76)
train subset (1175)	LMILP ($M=\infty$)	93.03 (45.15)

Summary & Conclusions

- The result obtained by LMILP(93.03%), where no additional data were used, is close to the the upper bound(93.16%) which was achieved by a grid search on the test data (cheating experiment).
- The algorithm shows good convergence within 5-7 iterations.
- The results seem to be independent of the initial values of the parameters.
- Recommended strategy:
  1. first run the LMILP algorithm to determine a good estimate of the decoding parameters.
  2. fine-tune the estimates from LMILP with a grid search using a fine grid.