Log-Linear Approach to Discriminative Training

NIPS 2008 Workshop
"Speech and Language: Learning-based Methods and Systems"

Ralf Schlüter, Georg Heigold, and Hermann Ney

Lehrstuhl für Informatik 6
Human Language Technology and Pattern Recognition
Computer Science Department, RWTH Aachen University
D-52056 Aachen, Germany

Dec. 12, 2008

Outline

Introduction

Statistical Modeling from the Perspective of Automatic Speech Recognition

Modeling

- Gaussian mixture hidden Markov models (HMM): Standard in state-of-the-art automatic speech recognition (ASR) systems

Training Criteria

- Maximum likelihood (ML)
  - Former standard, initialization for further discriminative training
  - Pros: guaranteed convergence, low training complexity
  - Cons: limited generalization, mismatch conditions
- Discriminative training (DT)
  - E.g. maximum mutual information (MMI), minimum classification error (MCE), minimum phone/word error (MPE/MWE)
  - Pros: consider class confusability
  - Cons: high training complexity, optimization, overtraining

Parameter Optimization

- ML: expectation-maximization (EM): guaranteed convergence
- DT: usually gradient descent or similar approaches, convergence proofs often for non-finite step sizes

Introduction

Discriminative Training Criteria

- Usually consider class posteriors, e.g.:

$$\mathcal{F}_{\text{MMI}}(\theta) = \sum_n \log p_\theta(c_n|x_n)$$

$$\mathcal{F}_{\text{MPE}}(\theta) = \sum_n \sum_c p_\theta(c|x_n) A(c, c_n)$$

with observations x, classes c, training data (x_n, c_n) for $n = 1, \ldots N$, class accuracy $A(c, c_n)$ and the class posterior

$$p_\theta(c|x) = \frac{p_\theta(x|c) \cdot p_\theta(c)}{\sum_{c'} p_\theta(x|c') \cdot p_\theta(c')}$$

- Observation: posterior is explicitly normalized over word sequences.
- Local normalization of class cond. and prior cancels in posterior:
  - Therefore no effect in discriminative training and Bayes decision rule.

  $\Rightarrow$ Local normalization of class conditional and prior not needed!
- Explicit/global normalization: similar to log-linear models.

Introduction

Questions:

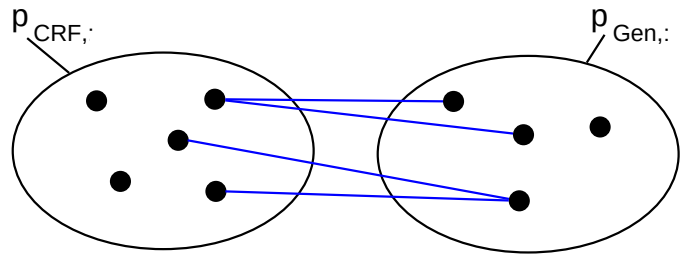
- Generalization of generative to log-linear models?
  - Transform Gaussian into log-linear model: straightforward.
  - Transform log-linear model into Gaussian: this work.
  - Transform log-linear model into *n*-gram language model: this work.
- Discriminative training of log-linear models for ASR?
  - MMI/MCE/MPE can be used for log-linear models.
  - Introduction of margin and regularization to MMI/MCE/MPE:
  - Strong relation to support vector machines (SVM): this work.
- Optimization of discriminative training criteria for log-linear models?
  - Due to equivalence similar to case of Gaussian mixture HMMs.
  - Finite step size with proven convergence: this work.

Outline

Generative vs. Discriminative Modeling

Posterior models

- discriminative posterior model: $p_{\mathrm{CRF},\Lambda}(c|x)$
- generative posterior model: $p_{\mathrm{Gen},\theta}(x,c) \overset{\text{Bayes}}{\rightarrow} p_{\mathrm{Gen},\theta}(c|x)$



- $p_{\mathrm{CRF},\cdot}$ and $p_{\mathrm{Gen},\cdot}$: (sub-)sets of posteriors

Equivalence:

- Generative and discriminative model with equal posterior: $\forall \Lambda\ \exists \theta : p_{\mathrm{CRF},\Lambda}(c|x) = p_{\mathrm{Gen},\theta}(c|x)\ \forall\ c,x$ (and vice versa)

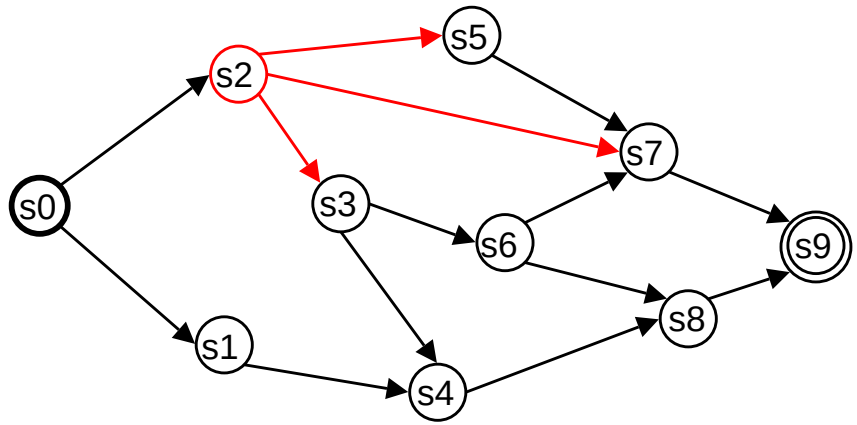
Remarks

- In case of equivalence posterior-based algorithms equally powerful.
- Examples: training (e.g. MMI/MCE/MPE) or decoding.
- Equivalence for non-parametric models: obvious.
- $p_{\mathrm{Gen},\cdot} \subset p_{\mathrm{CRF},\cdot}$: known in literature $\rightarrow$ not considered here.

Generative Models: Constraints/Model Structure

Examples:

- Gaussian distributions: variances must be positive
- Conditional probabilities (transition probabilities, language model):



  - many local normalization constraints, i.e., $\sum_{s} p(s|s') = 1 \ \forall s'$
  - log-linear/CRF: single global normalization constraint $\sum_{s_1^T} p(s_1^T) = 1$

- dependence (e.g. first order Markov model): $p(s_t|s_{t-1})$

Log-Linear Models: Invariance

Example log-linear model

$$p_\Lambda(c|x) \quad \propto \quad \exp(x^T \Lambda(c) x + \lambda(c)^T x + \alpha(c))$$

- $x \in \mathbb{R}^D$, $c = \{1, \ldots, C\}$, $\Lambda = \{\Lambda(c) \in \mathbb{R}^{D \times D}, \lambda(c) \in \mathbb{R}^D, \alpha(c) \in \mathbb{R}\}$

Definition

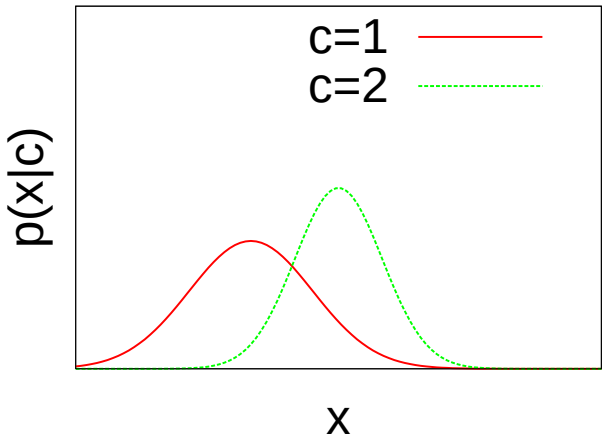
- invariance transformation f does not change posterior models, i.e.,

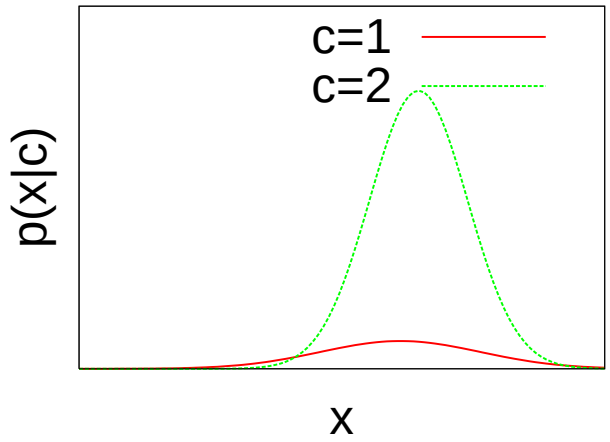
$$p_{\text{CRF}} \text{ invariant under } f \quad :\Leftrightarrow \quad p_{\text{CRF},\Lambda}(c|x) = p_{\text{CRF},f(\Lambda)}(c|x) \; \forall \Lambda, x, c$$

Example invariance transformations

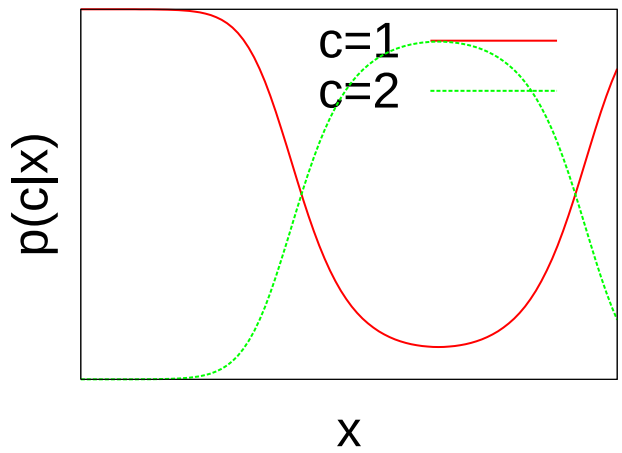
$$\begin{aligned}
\alpha(c) &\mapsto \alpha(c) + \alpha_0, \; \alpha_0 \in \mathbb{R} \\
\lambda(c) &\mapsto \lambda(c) + \lambda_0, \; \lambda_0 \in \mathbb{R}^D \\
\Lambda(c) &\mapsto \Lambda(c) + \Lambda_0, \; \Lambda_0 \in \mathbb{R}^{D \times D}
\end{aligned}$$

Example (Gaussian Mixture Model)





Ambiguous mixture weights, means, and variances...



...but identical posterior models

Outline

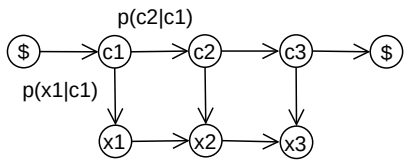
Equivalence Relations

Goal: Transform discriminative model into generative model with same posterior distribution.

Approach: Utilize invariances of log-linear models to fulfil constraints of generative model.

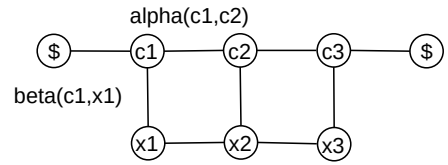
Example: simple tagging problem

$p_{\text{Gen}}\left(x_1^N, c_1^N\right)$



$$\underbrace{p(\$|c_N)\prod_{n=1}^{N} p(c_n|c_{n-1})}_{\text{bigram}}\ \underbrace{p(x_n|c_n)}_{\text{emission}}$$

$p_{\text{CRF}}\left(c_1^N|x_1^N\right)$



$$\underbrace{\exp(\alpha(c_N,\$))\prod_{n=1}^{N}\exp(\alpha(c_{n-1},c_n))}_{\text{bigram}}\ \underbrace{\exp(\beta(c_n,x_n))}_{\text{emission}}$$

Emission Probabilities

- Take advantage of invariances of log-linear models.
- Positionwise normalization of pseudo emission probabilities:

$$p(x|c) = \frac{\exp(\beta(c,x))}{Z(c)}$$

- Additional normalization constant $Z(c) = \sum_x \exp(\beta(c,x))$ can be included into bigram features:

$$\tilde{\alpha}(c',c) = \alpha(c',c) + \log Z(c)$$

- Log-linear model/CRF remains unchanged

$$\alpha(c',c) + \beta(c,x) = \big(\alpha(c',c) + \log Z(c)\big) + \big(\beta(c,x) - \log Z(c)\big)$$

Bigram Probabilities

Starting point:

- Solution for infinite sequences from [Jaynes 2003].
- Approach via transition matrix $Q_{c'c} = e^{\tilde{\alpha}(c',c)}$:

$$p(c|c') = \frac{Q_{c'c} v_c}{q v_{c'}}.$$

  with eigenvector v for largest eigenvalue q of transition matrix $Q_{c'c}$.
- Components of eigenvector pairwise cancel in class sequences: telescope product.

Generalization here:

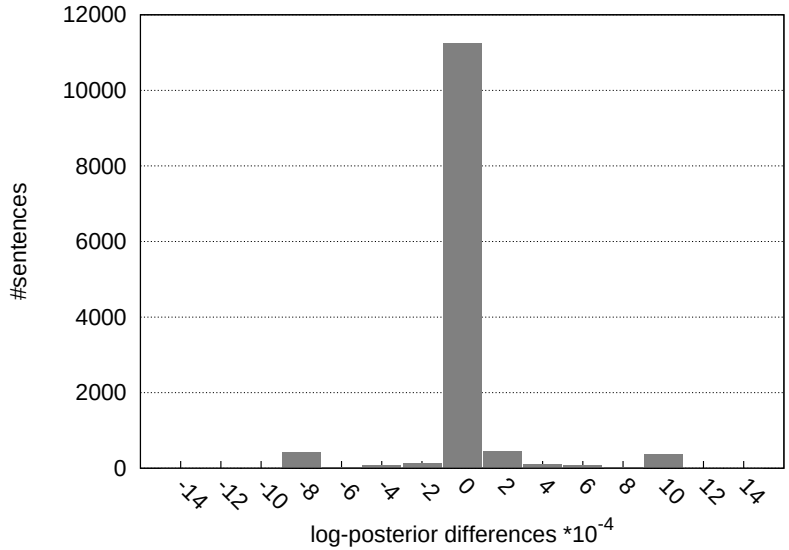
- Finite sequences: introduce sentence end symbol.
- Existence: from Perron-Frobenius theorem [Rao 1998, p. 467–475].
- Beyond bigrams: n-grams by generalization of class definition and correct handling of sentence boundaries.

Outline

Experimental Verification and Discussion

Experimental verification:

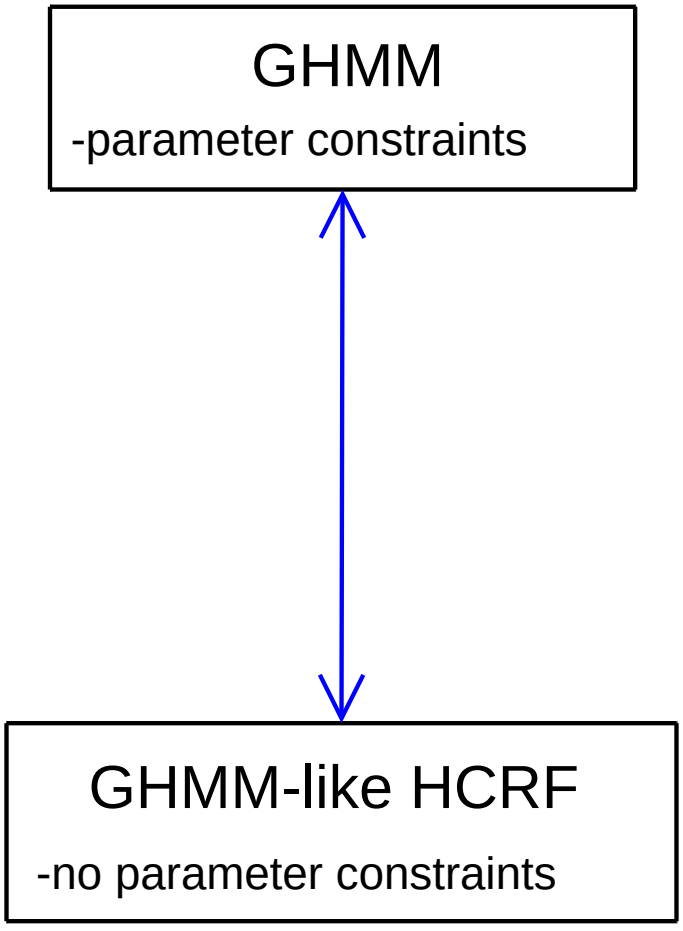
- Simple tagging problem, LUNA Media corpus
- zero difference = identical log-posteriors



Equivalence also holds for automatic speech recognition:

- Similar to tagging problem: propagation of normalization terms.
- Emission distribution: as before; positive definite covariance matrices could be fulfilled utilizing invariances of log-linear models. More details in [Heigold$^+$, Interspeech 2007].
- Transition probabilities: as before.
- Language model: generalization of case of transition probabilities.
- Further details in [Heigold$^+$, Interspeech 2008].

Discussion



Possible differences in practice:

- Numerical issues
  (e.g. inversion of covariance matrix).
- Spurious local optima.
- Different optimization criteria used
  (e.g. ML vs. MMI, regularization).
- Parameters are kept fixed.
- Unsuitable parameter tying.

Outline

Machine Learning and ASR

Machine Learning:

- Support Vector Machines (SVM)/large margin classifiers
- Probably Approximately Correct (PAC) bound
- amount of training data: typically $<$100,000 observations

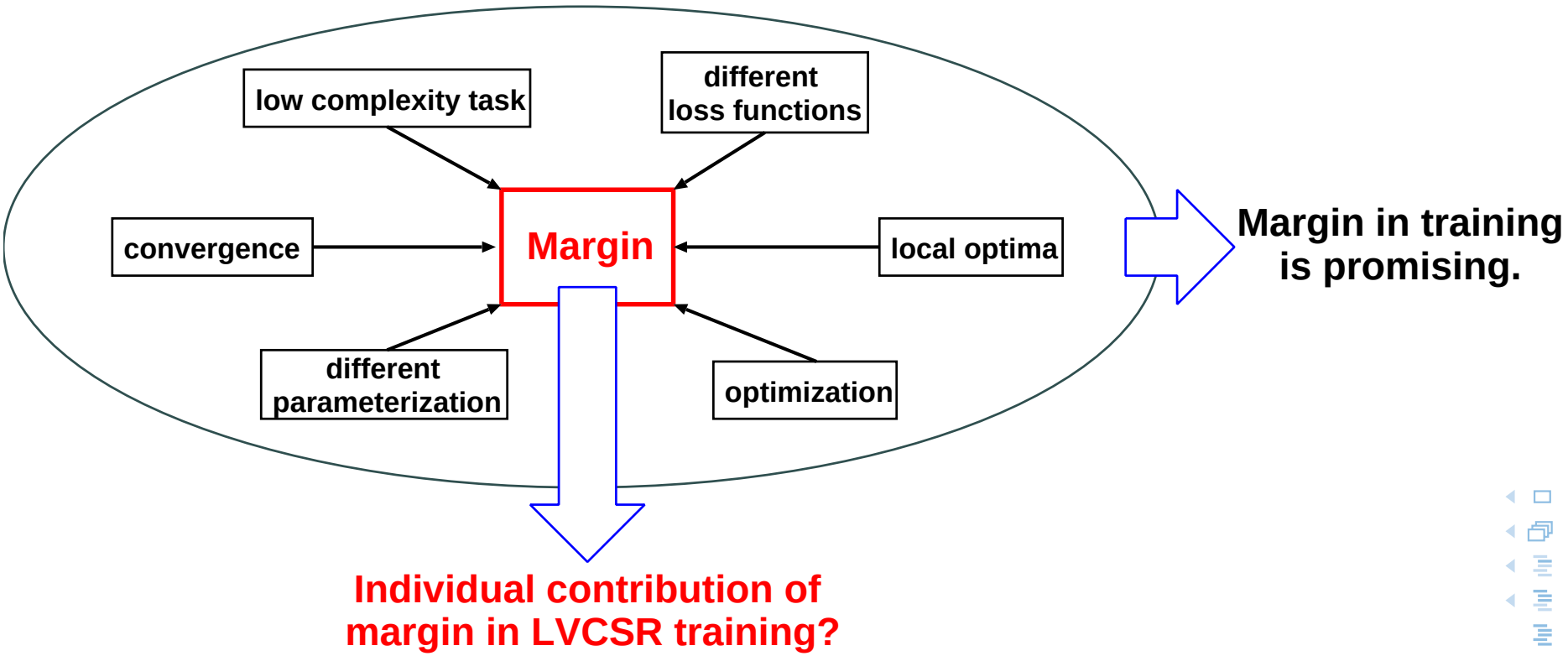
Automatic Speech Recognition (ASR):

- probabilistic criteria: e.g.
  - Maximum Likelihood (ML)
  - Maximum Mutual Information (MMI)
- error-based criteria: e.g.
  - Minimum Phone Error (MPE)
  - Minimum Classification Error (MCE)
- all without margin, (some of them) with regularization
- amount of training data:
  typically $\gtrsim$100h acoustic data, i.e. $\gtrsim$36,000,000 observations.

Margin & Discriminative Training in ASR

Current state:

- Usual criteria in ASR (ML/MMI/MPE) do not include margin term.
- Recently, margin-based criteria are investigated.
- Missing: investigation of the contribution of the margin alone.



Margin & Discriminative Training in ASR

Goals:

- Investigate potential of margin term in state-of-the-art large vocabulary speech recognition systems (LVCSR).
- Consistent evaluation of performance of margin term.

Study effect of margin w/o further modifying

- loss function,
- optimization algorithm,
- model parameterization, etc.

Approach:

- Modify MMI/MPE to incorporate margin.
- Relationship of conventional ASR training criteria and SVMs.
- Experimental evaluation for LVCSR.

Outline

Support Vector Machines (SVM) for ASR

SVM formulation w/o constraints/slack variables:

$$SVM^{(\mathcal{L})}(\lambda) = \frac{1}{2}\|\lambda\|^2 + \frac{J}{N}\sum_{n=1}^{N}\mathcal{L}(c_n; d_{n1}, \ldots, d_{nC})$$

- "Distance" $d_{nc} = \lambda^\top (f(x_n, c_n) - f(x_n, c))$, feature fcts. $f(x, c)$.
- loss function $\mathcal{L}$, e.g. hinge loss, or margin error
- Include margin (and regularization) into MMI/MPE.

MMI/MPE in log-linear form (recall: covers Gaussian mixture HMMs):

- Modified MMI:

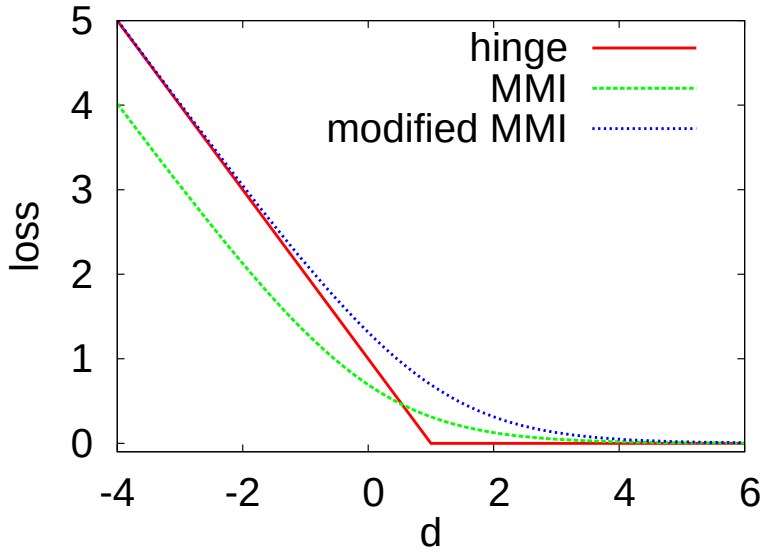
$$\mathcal{F}_\gamma^{(\mathrm{MMI})}(\lambda) = \frac{1}{2}\|\lambda\|^2 - \frac{J}{N}\sum_{n=1}^{N}\frac{1}{\gamma}\log\left(\frac{\exp(\gamma\,(\lambda^\top f(x_n, c_n) - 1))}{\sum_c \exp(\gamma\,(\lambda^\top f(x_n, c) - \delta(c, c_n)))}\right)$$

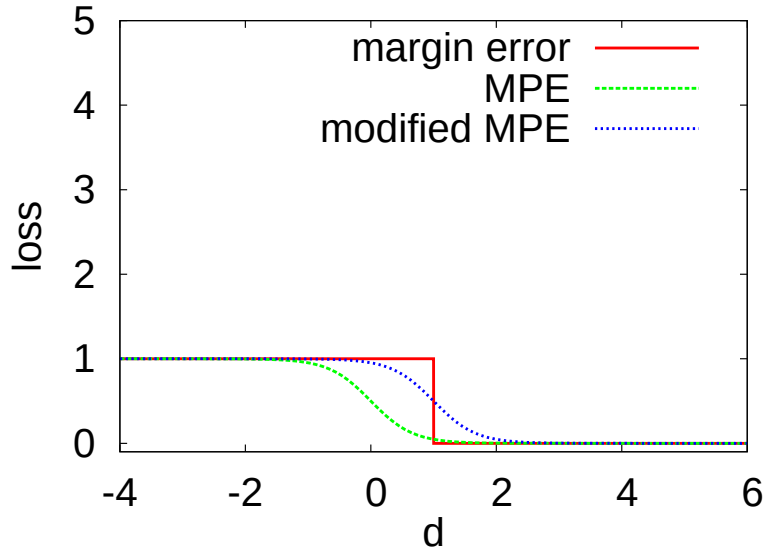
- Modified MPE:

$$\mathcal{F}_\gamma^{(\mathrm{MPE})}(\lambda) = \frac{1}{2}\|\lambda\|^2 + \frac{J}{N}\sum_{n=1}^{N}\sum_c E[c|c_n]\frac{\exp(\gamma\,(\lambda^\top f(x_n, c) - \delta(c, c_n)))}{\sum_{c'} \exp(\gamma\,(\lambda^\top f(x_n, c') - \delta(c', c_n)))}$$

- Approximation level γ: control smoothness of loss function
- $E[c|c_n]$: e.g. phoneme error, i.e., 1-0 loss generalized to strings.

Different loss functions (two-class case):





Asymptotic behaviour:

$$\mathcal{F}_{\gamma}^{(\mathrm{MMI})}(\lambda) \overset{\gamma\to\infty}{\to} SVM^{(\mathrm{hinge})}(\lambda)$$

$$\mathcal{F}_{\gamma}^{(\mathrm{MPE})}(\lambda) \overset{\gamma\to\infty}{\to} SVM^{(\mathrm{error})}(\lambda)$$

Potential shortcomings of hinge loss:

- Mismatch of loss in training and testing (relation?).
- PAC bound for hinge loss (and not recognition error).
- Hinge loss is not bounded, i.e., single observation can dominate objective function.

Specific Issues in ASR

Important differences to "simple" classification:

- Sequences instead of simple observations.
- Class c: HMM state sequence.
- Loss function: phoneme error.
- Margin: proportional to (approximate) number of phonemes.
- I-smoothing: replace regularization $\|\lambda\|^2$with $\|\lambda - \lambda_0\|^2$ for some reasonable λ_0.
- relationship of HCRFs with n^{th} order features, and Gaussian HMMs.

Justification of (some) heuristics in ASR?

Heuristic in ASR	SVM
scaling of probabilities	approximation level γ
i-smoothing	(refined) regularization
weak language model ("weak prior")	margin

Outline

Effect of Margin on a Simple ASR Task

- SieTill: German digit recognition task (11 whole word HMMs).
- Train and test data: 5.5h/43k runnings words/2M obs. each.
- Standard cepstral acoustic observations with temporal context.
- Log-linear HMMs: in case of $0^{\text{th}} + 1^{\text{st}}$ order features equivalent to Gaussian HMMs with globally pooled variances.
- Maximum mutual information (MMI) training criterion: no margin.
- Modified MMI training criterion: includes margin.
- Margin set to approximate word accuracy.

Features	Dns/Mix	Margin	WER [%]
$0^{\text{th}} + 1^{\text{st}}$	16	no	1.88
		yes	1.72
	64	no	1.77
		yes	1.59
$0^{\text{th}} + 1^{\text{st}} + 2^{\text{nd}} + 3^{\text{rd}}$	1	no	1.68
		yes	1.53

Effect of Margin on a Large Vocabulary ASR Task

- European Parliament plenary speeches (EPPS) English task.
- Spontaneous speech, large vocabulary (52k words).
- Mixture HMM with globally pooled variances, ∼1M Gaussians.
- Training: 92h/661k running words/33M observations.
- Testing (Eval' 07): 2.9h/27k running words/1M observations.
- Minimum phone error (MPE) training criterion: no margin ('none').
- Modified MPE training criterion: margin set to
  - approximate word accuracy ('word'), or
  - approximate phoneme accuracy ('phoneme').
- Also: study effect of choice of language model in training.

Training LM	Margin	WER [%]
unigram	none	11.5
	word	11.3
	phoneme	11.3
bigram	none	11.6
	word	11.3
	phoneme	11.3

Comparison of the Effect of the Margin

- Additional large scale task: GALE Chinese Broadcasts, 60k voc.

Task	Voc.	Training Data [h]	Training Data words	Training Data obs.	Training Criterion	Margin	WER [%]
SieTill	11	5.5	43k	2M	MMI	no	1.68
					mod. MMI	yes	1.53
EPPS English	52k	92	661k	33M	MPE	no	11.5
					mod. MPE	yes	11.3
GALE Chinese	60k	230	2,200k	83M	MPE	no	20.6
					mod. MPE	yes	20.3
		1,500	15,500k	540M	MPE	no	16.5
					mod. MPE	yes	16.3

- Digit recognition (practically no training errors): margin helps.
- Large vocabulary/many training errors: margin has little effect.

Outline

Motivation

Models:

- Conditional Random Fields (CRFs)
  - Discriminative, direct, graphical models, log-linear.
- Hidden CRFs (HCRFs)
  - Hidden variables, e.g. HMM: hidden state sequence ("alignment").

Criteria/objective functions:

- probabilistic: distribution estimation (e.g. Maximum Entropy), or
- error based: (smoothed) error minimization (e.g. MPE).

Optimization:

- gradient based, e.g. RProp, or
- via auxiliary function, e.g. Generalized Iterative Scaling (GIS) for optimization of MMI using CRF

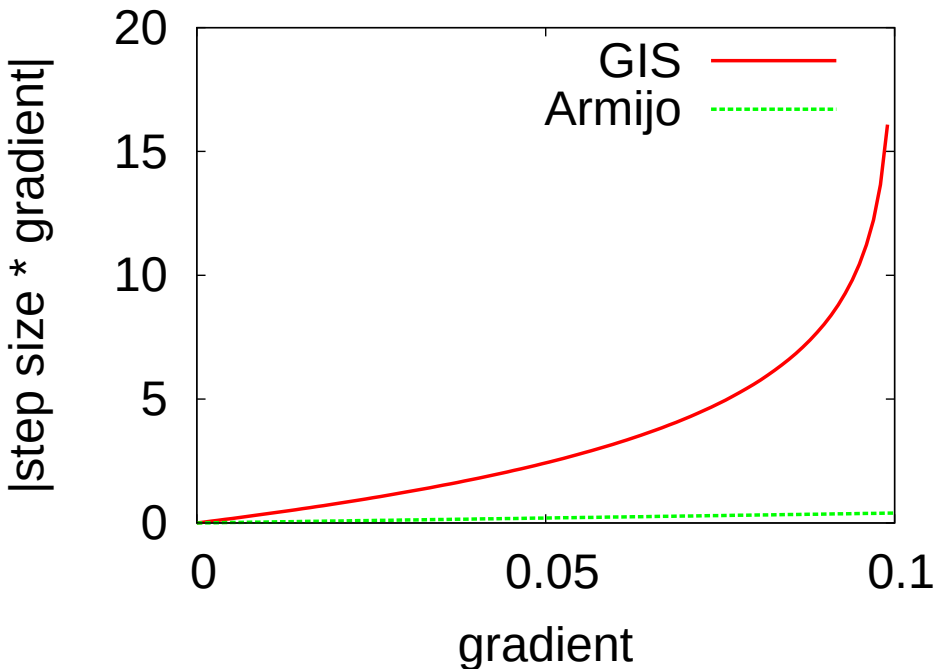
Limitations of GIS:

- HCRFs not covered,
- optimizes MMI criterion only.

Motivation

Why GIS-like algorithm?

- guaranteed increase of objective function in each iteration.
- convergence to critical point.
- parameter free (e.g. no tuning of step sizes!).



Why not existing algorithms?

- simple auxiliary function [Armijo 1966]: e.g. HCRFs/ GHMMs, overly pessimistic.
- reverse Jensen inequality: vanishing second derivative, linear auxiliary function.
- Generalized EM, [Saul 2002]: indirect optimization, alternates between mixture weights and Gaussian parameters.
- equivalence of log-linear HMMs ($\subset$HCRFs) and GHMMs: convergence speed sensitive to ambiguous parameters.

Outline

Objective Function

So far (CRF+MMI):

$$\mathcal{F}(\Lambda) = \sum_n \log \left(\frac{\exp\left(\sum_i \lambda_i f_i(x_n, c_n)\right)}{\sum_c \exp\left(\sum_i \lambda_i f_i(x_n, c)\right)} \right)$$

Generalized objective function (cf. rational form):

$$\mathcal{F}^{(\text{hidden})}(\Lambda) = \sum_n \log \left(\frac{\sum_c q_n(c) \exp\left(\sum_i \lambda_i f_i(x_n, c)\right)}{\sum_c p_n(c) \exp\left(\sum_i \lambda_i f_i(x_n, c)\right)} \right)$$

- non-negative numerator/denominator weights $q_n(c)$ and $p_n(c)$
- additional problem: (weighted) sum in numerator
- in general: no longer convex

Examples:

- log-linear mixtures/log-linear HMMs (LHMMs)/context priors
- MPE for HCRFs

Auxiliary Functions

Goal: auxiliary function $\mathcal{A}(\Lambda|\Lambda')$ (in the strong sense) of $\mathcal{F}(\Lambda)$ at Λ'

$$:\Leftrightarrow \quad \mathcal{F}(\Lambda) - \mathcal{F}(\Lambda') \geq \mathcal{A}(\Lambda|\Lambda') \quad \text{with equality for } \Lambda = \Lambda'$$

Consequence: convergence with $\mathcal{A}(\Lambda|\Lambda') > 0 \Rightarrow \mathcal{F}(\Lambda) > \mathcal{F}(\Lambda')$

Approach:

1. Decomposition of objective function:
   $\mathcal{F}^{(\text{hidden})}(\Lambda) = \mathcal{F}^{(\text{num})}(\Lambda) - \mathcal{F}^{(\text{den})}(\Lambda)$
2. Find auxiliary functions of (well-known) subproblems:
   - EM gives an auxiliary function for $\mathcal{F}^{(\text{num})}$
   - GIS gives an auxiliary function for $\mathcal{F}^{(\text{den})}$
     ($f(x,c) \geq 0$, feature count $F \equiv \sum_i f_i(x_n, c)$)
3. Combination of subproblems (transitivity):
   $\mathcal{A}^{(hidden)} = \mathcal{A}^{(EM)} + \mathcal{A}^{(GIS)}$
4. Update rules by maximization of $\mathcal{A}^{(hidden)}$:
   - similar to GIS
   - same efficient algorithms (e.g. accumulation statistics)

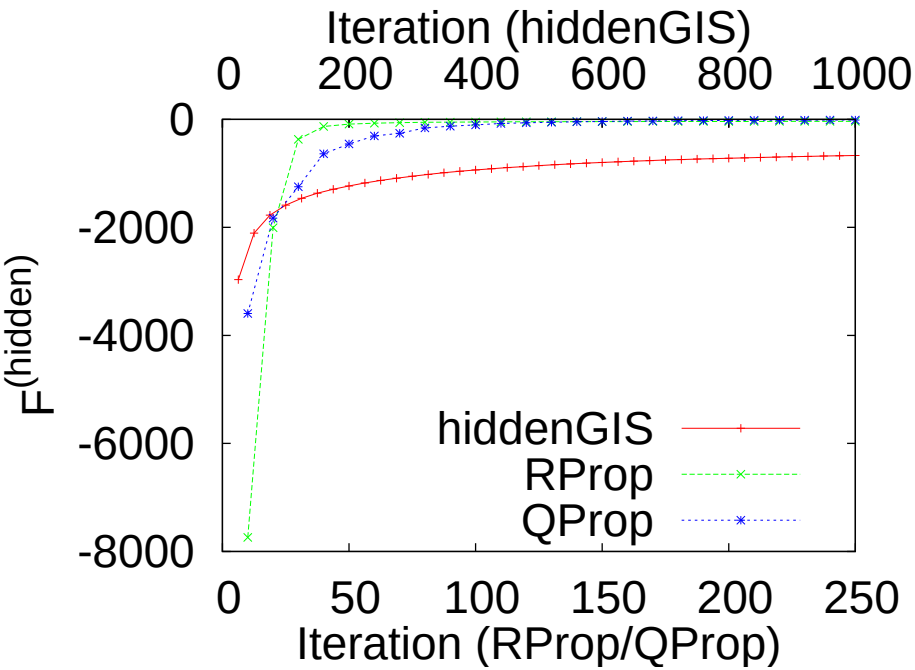
Outline

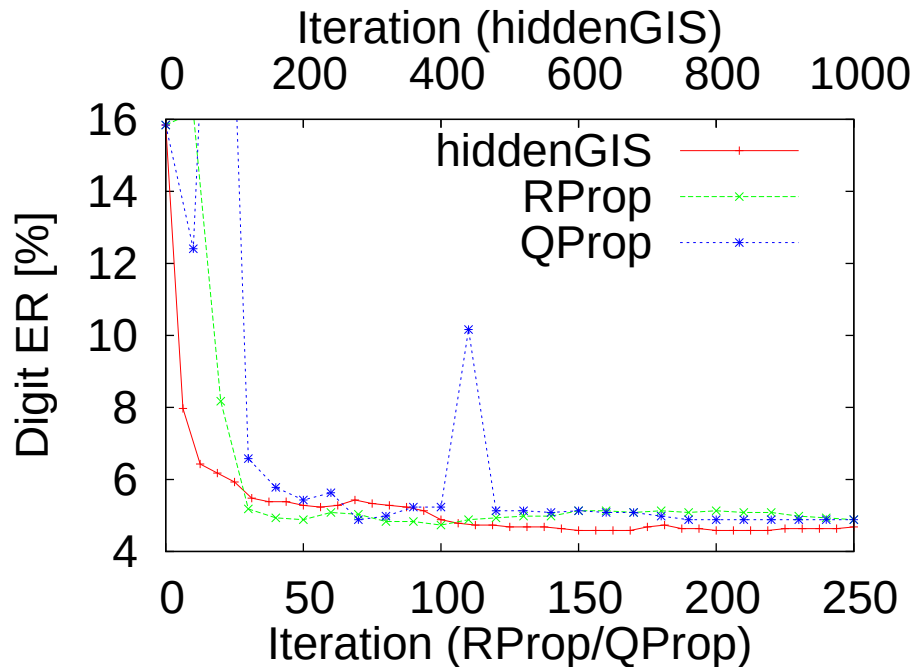
Experiments: Handwritten Digit Classification

- USPS Handwritten digits (US postal codes).
- Large amount of image variability.
- Separate training (7,000 images) and test (2,000 images) set.
- Gray-scale features + Sobel based derivatives.
- Log-linear mixture models, 16 Gaussians/mixture.
- Regularization based on Gaussian prior.
- Advanced modeling: 2-3% WER.

Experiments: Handwritten Digit Classification

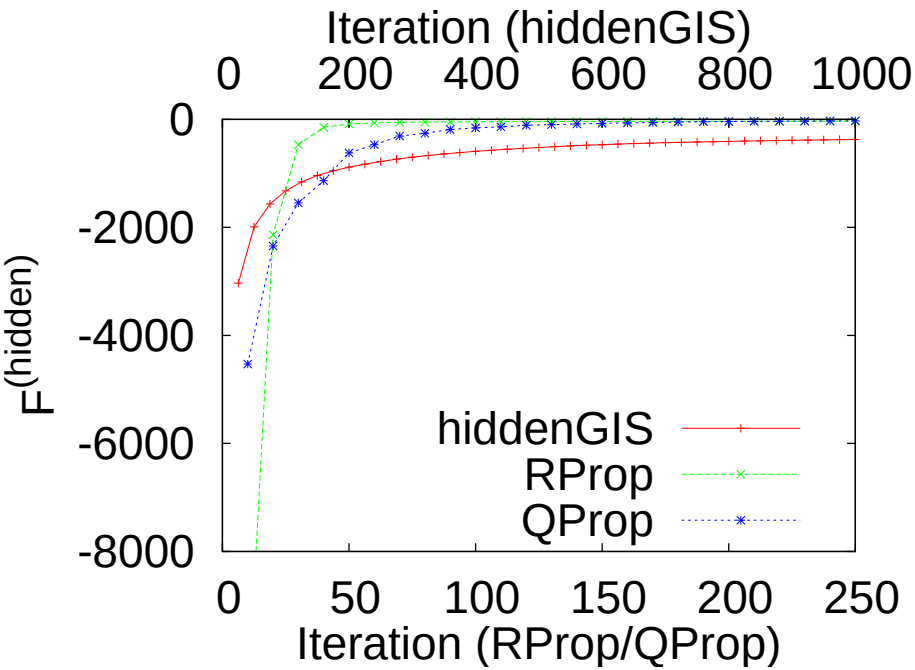
USPS: Initialization from Gaussian mixture model (GMM)

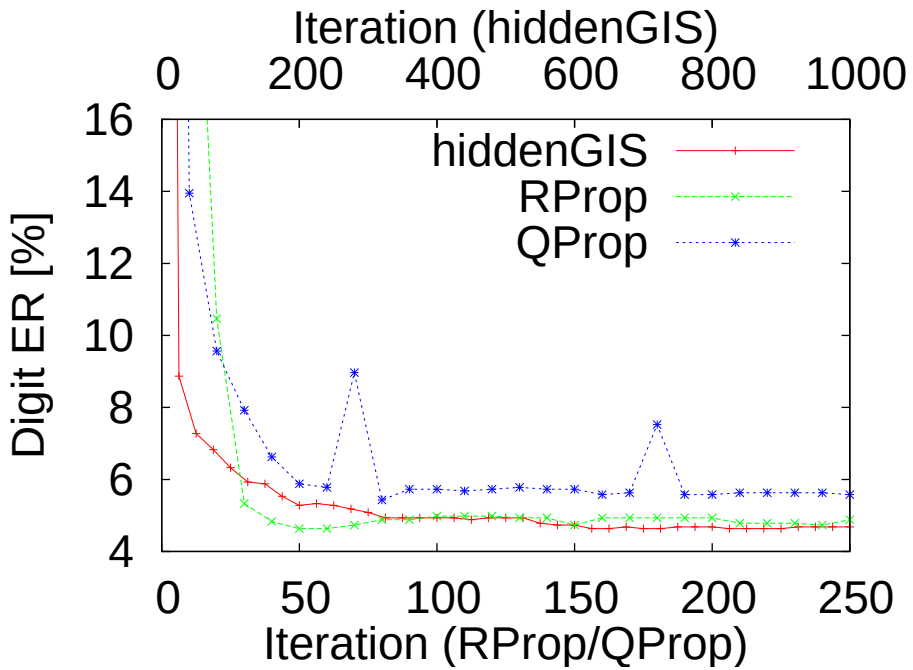




Experiments: Handwritten Digit Classification

USPS: Random initialization

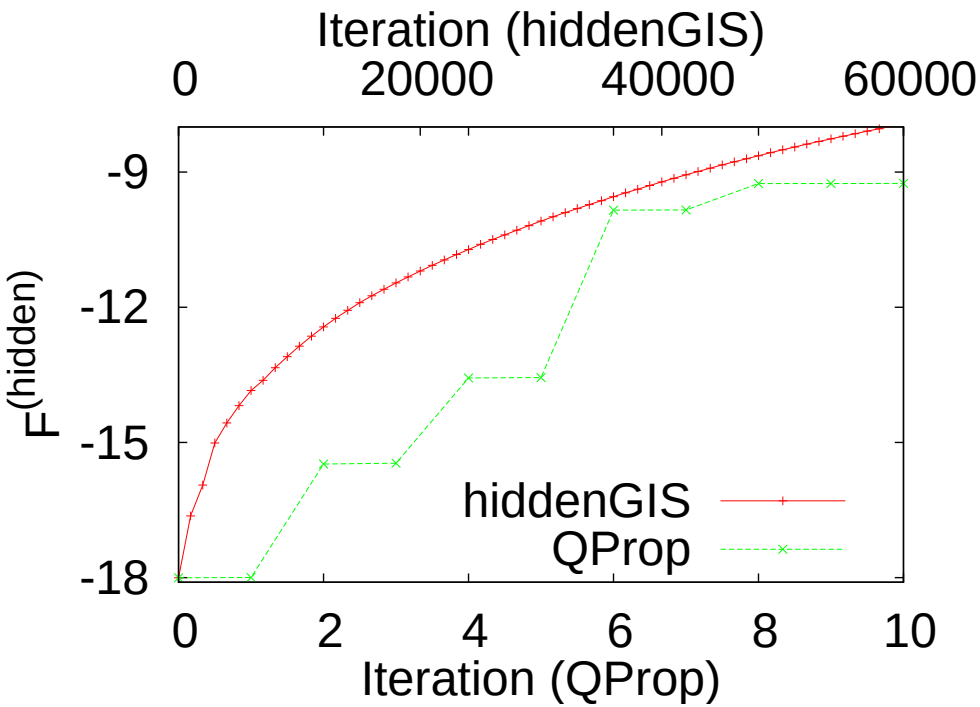


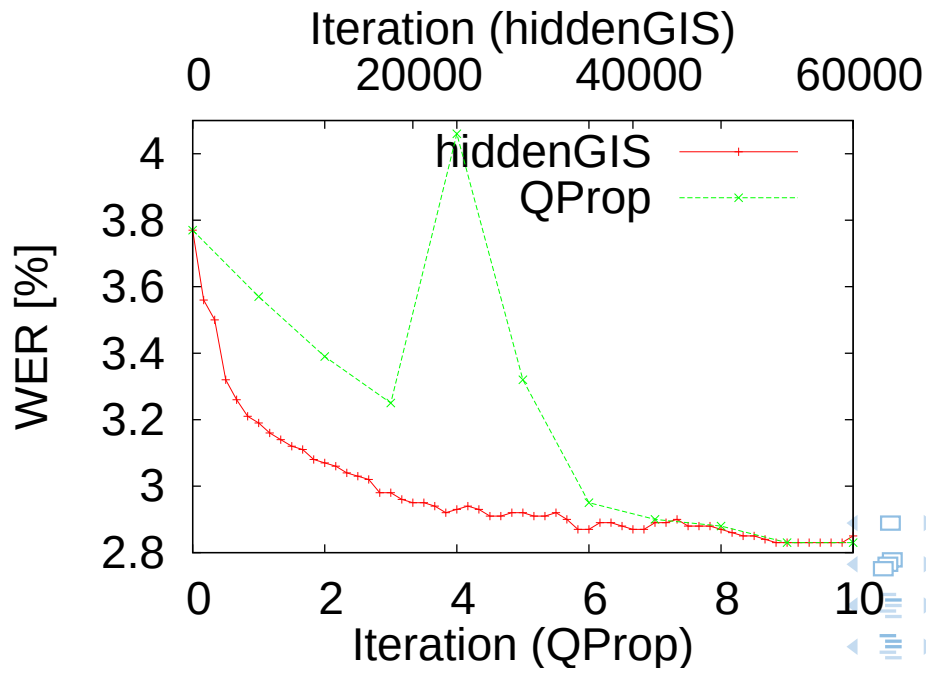


Experiments: ASR/Digit Strings

SieTill: German Digit Recognition

- Whole word Gaussian HMMs.
- Single Gaussians per HMM state with global pooled variances.
- Train and test data: 5.5h/43k runnings words/2M obs. each.
- Maximum Mutual Information (MMI) training criterion.





Conclusions

- Equivalence of GHMMs and GHMM-like log-linear models shown.
- With comparable features, log-linear models cover same posterior space than corresponding generative models.
- Training criteria known from generative models (e.g. error based) can be adopted.

- Small modification to MMI/MPE shows strong relation to SVM objective with hinge loss/margin (phoneme) error.
- Consistent improvements for margin, even for very large scale task.
- From equivalence to HCRF, GHMM can be directly related to SVM.

- Hidden-GIS generalizes GIS to cover HCRF and MPE objective.
- Verification for image and automatic speech recognition
- Smooth progress during iterations.
- Convergence becomes very slow for ASR task, ok for image task.
- Different behaviour: bounded vs. unbounded feature functions?

Conclusions