Bayesian Discriminative Adaptation for Speech Recognition

C. K. Raut, Kai Yu and Mark Gales

2007 April 12

Cambridge University Engineering Department

Overview

- Adaptation and Adaptive Training
  - Speech Recognition in Varying Acoustic Conditions
  - Multistyle and Adaptive Training
- Adaptation from Bayesian Perspective
  - Bayesian Adaptation in ML case
  - Issues and Approximations
- Discriminative Adaptation
  - Estimation of Discriminative Transforms
  - Optimisation of Discriminative Objective Functions
  - Issues with Bayesian Discriminative Adaptation

Speech Recognition in Varying Acoustic Conditions

- Automatic speech recognizers (ASRs) : Operate in different or varying acoustic conditions

  -Model Adaptation: models adapted using linear-transforms

- ASRs training on found data

  -Adaptive Training: speech and non-speech variability modelled seprately, by HMM parameters and transforms respectively

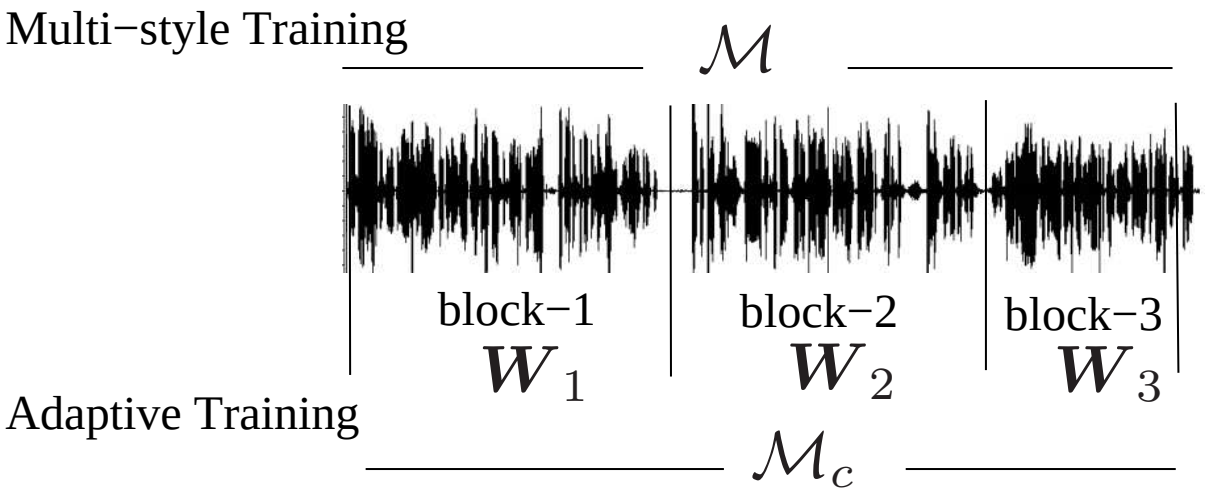
- Maximum likelihood linear transforms widely used

Adaptive Training

- Speech and non-speech variability: Separately modelled
- Speech variability $\Rightarrow$ Canonical model

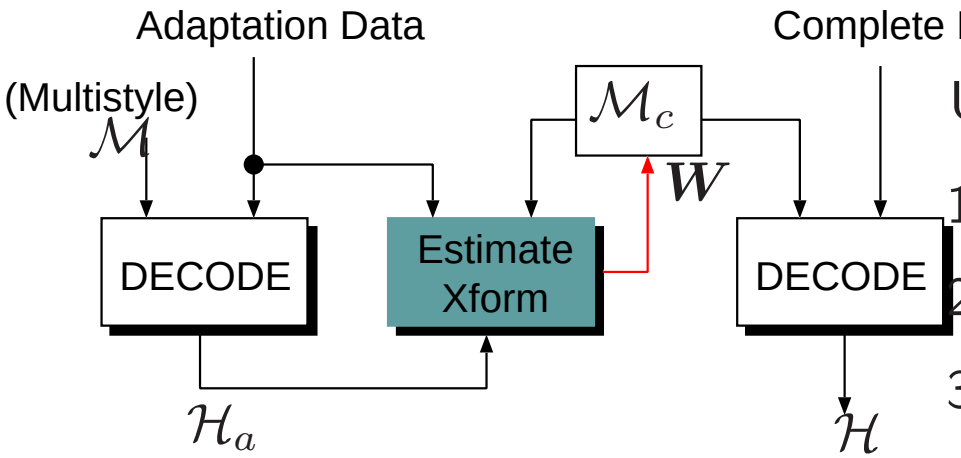
  -global in nature (same for all speech blocks/speakers)
- Non-speech variability $\Rightarrow$ Transforms

  -specific to each homogeneous blocks of speech
- Canonical model and transforms: estimated in interleaved fashion



(Unsupervised) Adaptation

- No transcriptions available for test data in unsupervised mode
- Canonical model not suitable for direct decoding
- Multi-style trained system: can be directly used for decoding



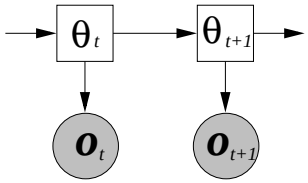
Unsupervised Adaptation:

1. Generate 1-Best supervision
2. Estimate transforms
3. Model adaptation and recognition

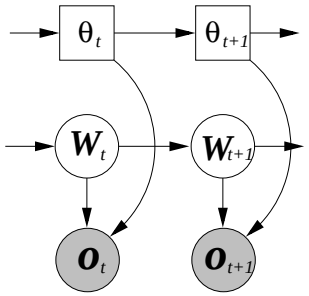
Drawbacks

- Estimates biased towards the hypotheses in unsupervised adaptation
- Assumes sufficient amount of data, whereas no control over adaptation data
  $\Rightarrow$ motivates Bayesian approaches

Adaptive Training From Bayesian Perspective



Standard HMM



Adaptive HMM

- Observation dependent on
  - state/component θ: discreate random variable
  - transform $\boldsymbol{W}$: continuous random variable
- Transform is constant for each testing acoustic condition $\boldsymbol{W}_t = \boldsymbol{W}_{t+1}$
- Output of adaptive training with sufficient data
  - Point estimate of canonical model, $\hat{\mathcal{M}}_c$
  - Prior distribution of transform parameters, $p(\boldsymbol{W})$

Adaptation Using Bayesian Inference in ML case

- Acoustic score - marginal likelihood of the whole sequence

$$p(\mathbf{O}|\mathcal{H}) = \underbrace{\int_{\boldsymbol{W}} p(\mathbf{O}|\mathcal{H},\boldsymbol{W})p(\boldsymbol{W})\,d\boldsymbol{W}}_{\text{intractable}}$$

- Acoustic score calculated for *every* possible hypothesis sequence
  - Observations not conditionally independent due to constant transform
  - Viterbi algorithm is not applicable
  - N-Best rescoring used
- Lower-bound Approximation by Jensen's inequality: Joint variational distribution of state/component sequence $\boldsymbol{\theta}$ and transform $\boldsymbol{W}$ is introduced

$$\begin{aligned}\log p(\mathbf{O}|\mathcal{H}) &= \log \int_{\boldsymbol{W}} p(\mathbf{O}|\mathcal{H},\boldsymbol{W})p(\boldsymbol{W})\,d\boldsymbol{W} \\ &\geq \int_{\boldsymbol{W}} q(\boldsymbol{\theta},\boldsymbol{W}) \log \frac{p(\mathbf{O},\boldsymbol{\theta}|\boldsymbol{W},\mathcal{H})p(\boldsymbol{W})}{q(\boldsymbol{\theta},\boldsymbol{W})}\,d\boldsymbol{W}\end{aligned}$$

Form of Lower Bound Approximations for ML

- **Variational Bayes (VB)** - state/component $\boldsymbol{\theta}$ and transform $\boldsymbol{W}$ conditionally independent

$$q(\boldsymbol{\theta}, \boldsymbol{W}) = P(\boldsymbol{\theta}|\mathbf{O}, \mathcal{H})p(\boldsymbol{W}|\mathbf{O}, \mathcal{H})$$

  - Decoupling of $\boldsymbol{\theta}$ and $\boldsymbol{W}$ $\Rightarrow$ integral tractable

- **Point estimate (MAP)** - Sufficient data assumption: Transform posterior becomes Dirac delta function (prior still a distribution)

$$q(\boldsymbol{\theta}, \boldsymbol{W}) = P(\boldsymbol{\theta}|\mathbf{O}, \mathcal{H}, \hat{\boldsymbol{W}})\delta(\boldsymbol{W} - \hat{\boldsymbol{W}})$$

  - point estimate $\hat{\boldsymbol{W}}$: tractable

- ML case: Tractable LB approximation to Bayesian integral can be found, EM-like algorithm for estimation

Experiments on Conversational Telephone Speech Task

- Switchboard (English): conversational telephone speech task
  - Training dataset: about 290hr, 5446spkr
  - Test dataset: 6hr, 144spkr
  - Front-end: PLP+Energy+1^{st},2^{nd},3^{rd} derivatives
  - HLDA and VTLN used
  - 150-Best list rescoring in inference

- 16 Gaussian components per state systems
  - ML and MPE speaker independent (SI) system - baseline
  - Form of transform

$$\hat{\boldsymbol{\mu}}^{(s)} = \mathbf{A}^{(s)}\boldsymbol{\mu} + \mathbf{b}^{(s)}$$

  - prior distribution - **Single Gaussian distribution**
  - MPE-SAT only discriminatively updated the canonical model given ML estimated transforms

Utterance Level Bayesian Adaptation - ML

Adapt. Approach.	ML Train	
	SI	SAT
—	32.83	—
ML	35.54	35.16
MAP	32.16	31.76
VB	31.77	31.50

- ML adaptation fails - insufficient adaptation data
- MAP improved WER - use prior information
- VB significantly better - non-point distribution, tighter bound
- Bayesian approach for ML case: improves performance. But
  - Maximizing likelihood over adaptation data may not improve WER
    $\Rightarrow$ motivates use of discriminative criteria (related to error rate)
  - State-of-art system uses discriminative criteria for training
    $\Rightarrow$ Consistent discriminative framework both for training and adaptation

Discriminative Adaptive Training

- Half-way discriminative
  - Discriminative models
  - ML transforms
- Full discriminative
  - Discriminative models
  - Discriminative transforms

Half-way Discriminative Adaptive Training

- Half-way Discriminative (Yu et al.): ML transforms, discriminative models

Adapt. Approach	MPE Train	
	SI	SAT
—	29.20	—
ML	32.44	32.27
MAP	29.01	28.80
VB	28.75	28.63

- Not much gain from Bayesian approaches (MAP/VB)
  - ML transforms on discriminative models
  - Prior distribution estimated on ML transforms
  - Prior applied in a non-discriminative way

- Next Step:

1. Estimate and use discriminative transforms
2. Compute prior over discriminative transforms
3. Apply prior in discriminative way (discriminative objective function)

Discriminative Objective Functions

- Discriminative criteria:

$$\text{MMI}: \quad \mathcal{F}(\mathbf{O}, \boldsymbol{H}) = P(\mathcal{H}|\mathbf{O}) \tag{1}$$

$$\text{MPE}: \quad \mathcal{F}(\mathbf{O}, \boldsymbol{H}) = \sum_{\mathcal{H}} P(\mathcal{H}|\mathbf{O}) A(\mathcal{H}, \mathcal{H}_r) \tag{2}$$

- Description in terms of MMI for convenience
- Discriminative objective function for inference/adaptation

$$\text{non} - \text{point}: \quad P(\mathcal{H}|\mathbf{O}) = \int_{\boldsymbol{W}} P(\mathcal{H}|\mathbf{O}, \hat{\mathcal{M}}, \boldsymbol{W}) p(\boldsymbol{W}|\phi)\, d\boldsymbol{W} \tag{3}$$

$$\text{MAP} - \text{estimate}: \quad P(\mathcal{H}|\mathbf{O}) = P(\mathcal{H}|\mathbf{O}, \hat{\mathcal{M}}, \hat{\boldsymbol{W}}) p(\hat{\boldsymbol{W}}|\phi) \tag{4}$$

MAP Estimation of Discriminative Transforms

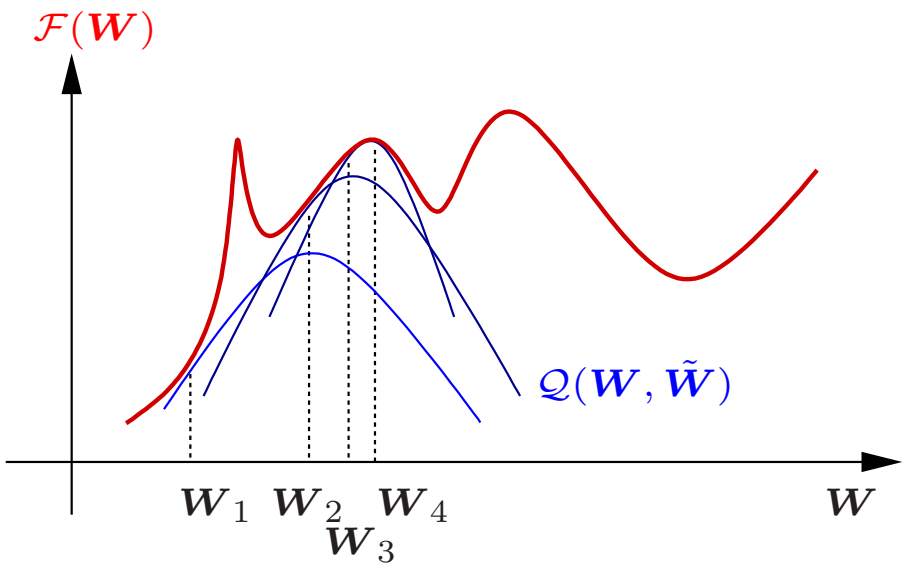
- With MMI-criteria, the MAP-estimation becomes

$$\begin{aligned}\hat{\boldsymbol{W}} &= \underset{\boldsymbol{W}}{\operatorname{argmax}} \left\{ \log P(\mathcal{H}|\mathbf{O}, \hat{\mathcal{M}}, \boldsymbol{W}) + \log p(\boldsymbol{W}|\phi) \right\} \\ &= \underset{\boldsymbol{W}}{\operatorname{argmax}} \left\{ \log \frac{p(\mathbf{O}|\mathcal{H}, \hat{\mathcal{M}}, \boldsymbol{W}) P(\mathcal{H})}{\sum_{\breve{\mathcal{H}}} p(\mathbf{O}|\breve{\mathcal{H}}, \hat{\mathcal{M}}, \boldsymbol{W}) P(\breve{\mathcal{H}})} + \log p(\boldsymbol{W}|\phi) \right\} \end{aligned} \quad (5)$$

- num term minus den term, in objective function

- How to optimise it?

MAP Estimation of Discriminative Transforms



- If lower-bound can be found: EM(-like) algorithm for maximisation
- increasing auxiliary function guarantees increase in objective function
- $\Delta\mathcal{F} \geq \Delta\mathcal{Q}$: always true

$$\mathcal{F}(\boldsymbol{W}) - \mathcal{F}(\tilde{\boldsymbol{W}}) \geq \mathcal{Q}(\boldsymbol{W}, \tilde{\boldsymbol{W}}) - \mathcal{Q}(\tilde{\boldsymbol{W}}, \tilde{\boldsymbol{W}}) \tag{6}$$

- How discriminative criteria alone (without prior) optimised?
- Does it have a lower-bound?

Function Optimisation by Weak-sense Aux. Function

- "Weak-sense" Auxiliary function widely used in training of discriminative HMMs

- Weak-sense auxiliary function is defined by the property:

  – same gradient at current estimate

$$\left.\frac{\partial \mathcal{Q}(\boldsymbol{W}, \tilde{\boldsymbol{W}})}{\partial \boldsymbol{W}}\right|_{\boldsymbol{W}=\tilde{\boldsymbol{W}}} = \left.\frac{\partial \mathcal{F}(\boldsymbol{W})}{\partial \boldsymbol{W}}\right|_{\boldsymbol{W}=\tilde{\boldsymbol{W}}} \tag{7}$$

  – same value at current estimate of parameters

$$\mathcal{Q}(\boldsymbol{W}, \tilde{\boldsymbol{W}}) = \mathcal{F}(\tilde{\boldsymbol{W}})$$

Function Optimisation by Weak-sense Aux. Function

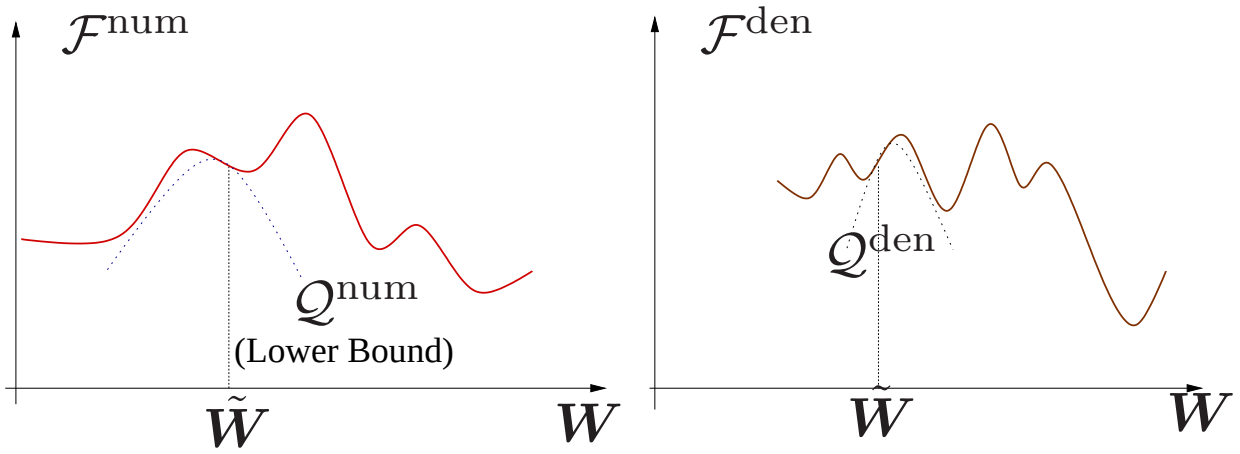
- Objective function:

$$\begin{aligned}
\mathcal{F}(\boldsymbol{W}) &= \sum_t \log \sum_m P(m) p(\boldsymbol{o}_t|m, \mathcal{M}^{\text{num}}) - \sum_t \log \sum_m P(m) p(\boldsymbol{o}_t|m, \mathcal{M}^{\text{den}}) \\
&\not\geqslant K^{\text{num}} + \sum_{t,m} \gamma_m^{\text{num}}(t) \log p(\boldsymbol{o}_t|m, \mathcal{M}^{\text{num}}) \quad \text{LB of num} \\
&\quad - \left(K^{\text{den}} + \sum_{t,m} \gamma_{jm}^{\text{den}}(t) \log p(\boldsymbol{o}_t|m, \mathcal{M}^{\text{den}}) \right) \quad \text{LB of den} \\
&\quad \underbrace{+K^{\text{sm}} + \sum_m D_m \left(-\frac{1}{2} (\boldsymbol{W} - \tilde{\boldsymbol{W}}) \boldsymbol{\zeta}_m^{\text{T}} \boldsymbol{\Sigma}_m^{-1} \boldsymbol{\zeta}_m (\boldsymbol{W} - \tilde{\boldsymbol{W}}) \right)}_{\mathcal{Q}^{\text{sm}}\text{:smoothing term}} \qquad (8)
\end{aligned}$$

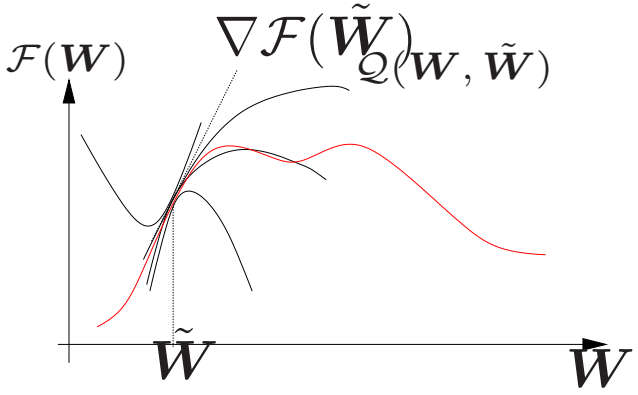
$\mathcal{Q}^{\text{sm}}$ max at current estimate:

$$\left. \frac{\partial \mathcal{Q}^{\text{sm}}(\boldsymbol{W}, \tilde{\boldsymbol{W}})}{\partial \boldsymbol{W}} \right|_{\boldsymbol{W}=\tilde{\boldsymbol{W}}} = 0 \qquad (9)$$

Characteristics of Weak-sense Aux. Function



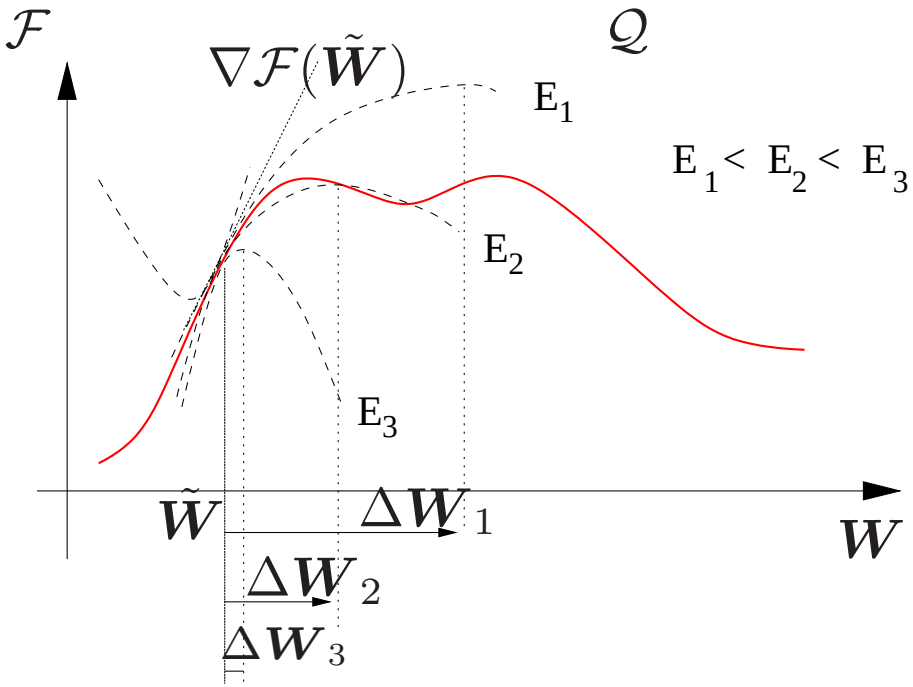
- Not a lower-bound (as LB taken in both num and den terms)



It can be anywhere, except

- gradient is same at the current estimate
- passes via current estimate of parameter

Effect of Smoothing Factor



- Smoothing factor

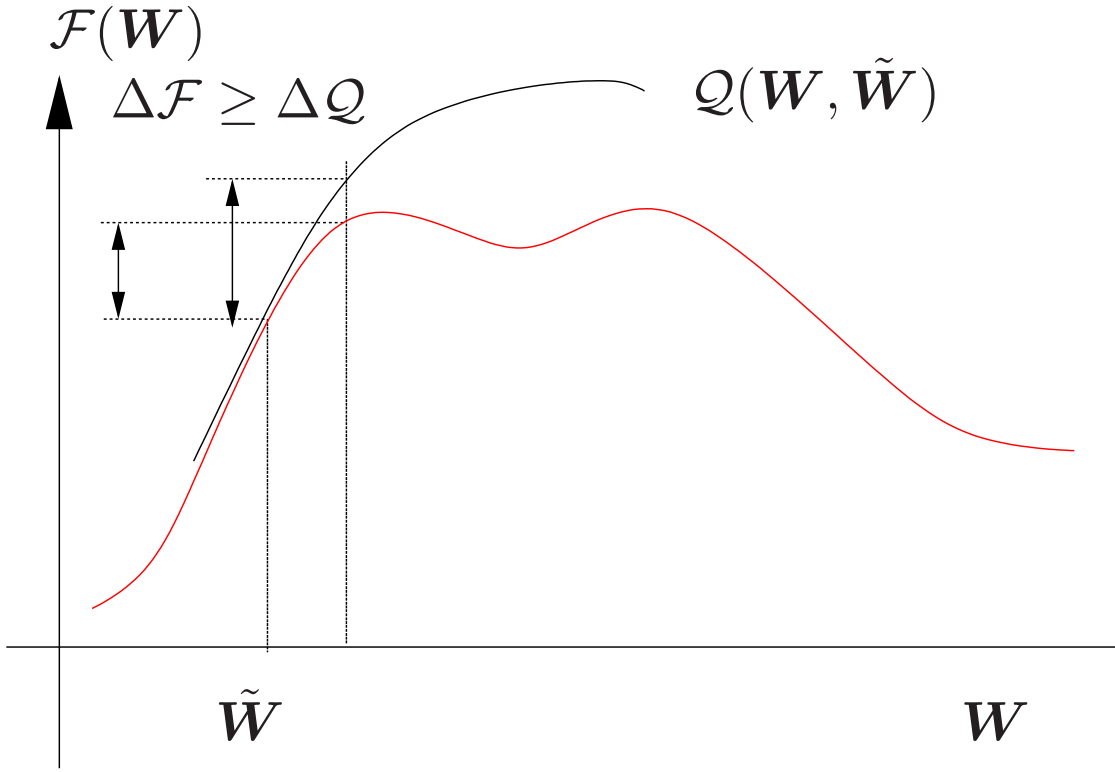
  $D_m = E \sum_t \gamma_m(t)$.

- Value of E: Controls step of update
- As E increases, it holds at current estimate
- Empirically $E = 1$ to 2: used discriminative training (weak-sense auxiliary function)

Problem with MAP Objective Function Optimisation

- Weak-sense auxiliary function, though increases MMI/MPE objective function, is not guaranteed to be a lower-bound

- $\Delta\mathcal{F} \ngeq \Delta\mathcal{Q}$: possible



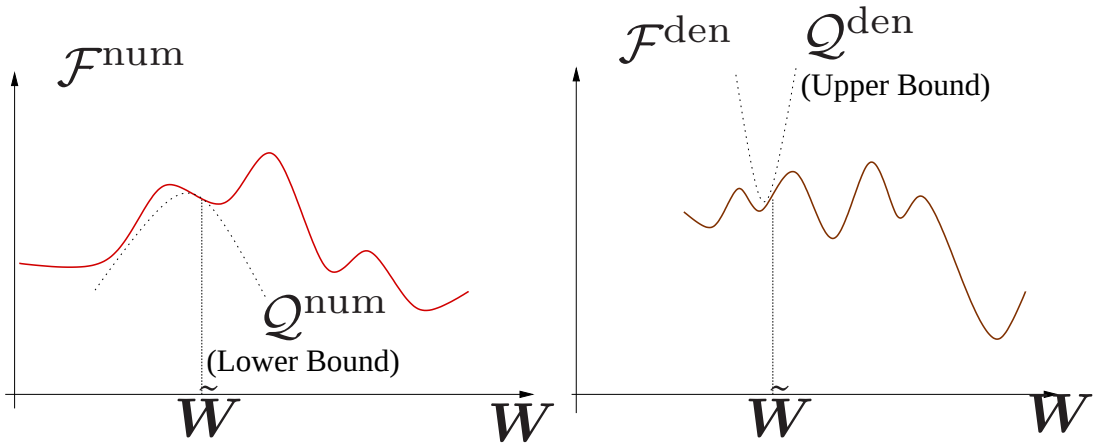
$$\mathcal{F}(\hat{\boldsymbol{W}}) - \mathcal{F}(\boldsymbol{W}) \ngeq \mathcal{Q}(\boldsymbol{W}, \hat{\boldsymbol{W}}) - \mathcal{Q}(\boldsymbol{W}, \boldsymbol{W}) \tag{10}$$

- This implies adding prior term to such an auxiliary function, as it is, to optimise overall criteria will not necessarily increase the overall objective function.

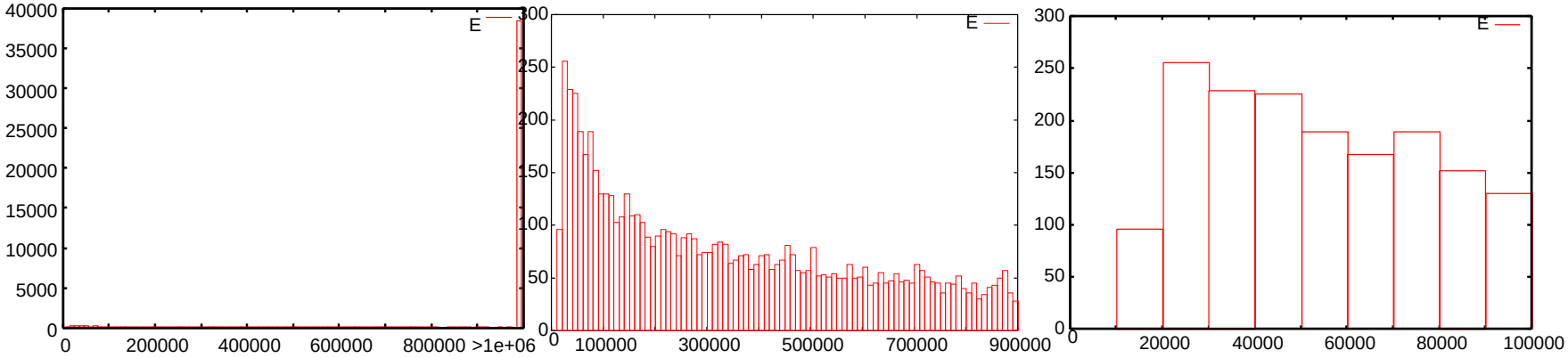
Strict Lower Bound by Reverse-Jensen Inequality

$$\begin{aligned}
\mathcal{F}(\boldsymbol{W}) &= \sum_t \log \sum_m P(m) p(\boldsymbol{o}_t|m, \mathcal{M}^{\text{num}}) - \sum_t \log \sum_m P(m) p(\boldsymbol{o}_t|m, \mathcal{M}^{\text{den}}) \quad (11) \\
&\geq K^{\text{num}} + \sum_{t,m} \gamma_{m,t}^{\text{num}} \log p(\boldsymbol{o}_t|m, \mathcal{M}^{\text{num}}) \quad \text{LB of num} \\
&\quad - \left(K^{\text{den}} + \sum_{t,m} (-w_{m,t}) \log p(\boldsymbol{z}_{m,t}|m, \mathcal{M}^{\text{den}}) \right) \quad \text{UB of den} \quad (12)
\end{aligned}$$

- reverse-Jensen applied to den term to form Upper Bound
- $\boldsymbol{z}_{m,t}$: modified observation, $w_{m,t}$: positive weights
- gives overall lower-bound



Smoothing Factor with Strict Lower-bound



- For sake of comparison: $E_m = \frac{D_m}{\sum_t \gamma_m^{\text{den}}(t)}$
- 38508 (84%) out of 45699 values of E were $\geq 10^6$!!
- cf. E chosen between 1-2 in weak-sense auxiliary function
- Such extremely large values of smoothing constant: holds tightly at current estimate, leads nowhere

Hessian-constrained Auxiliary Function

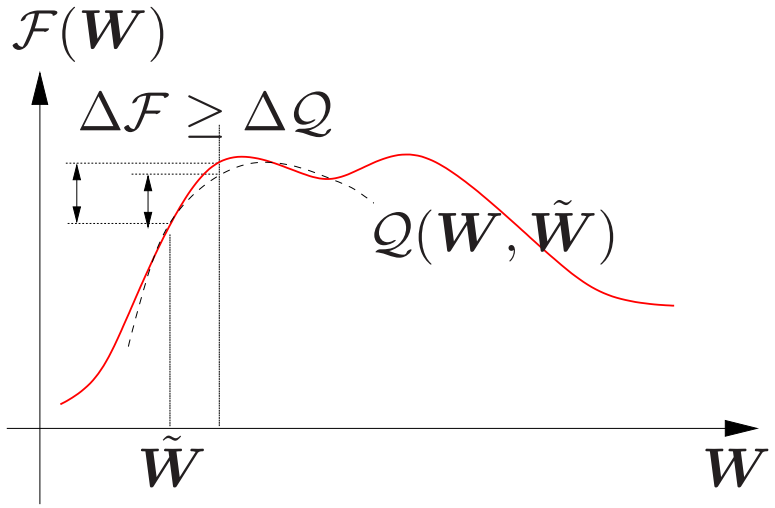
- Weak-sense auxiliary function:
  - sufficient update with low values of E, but not a strict lower-bound.

- Strong-sense auxiliary function:
  - strict-lower bound, but very high value of E. Leads no where.

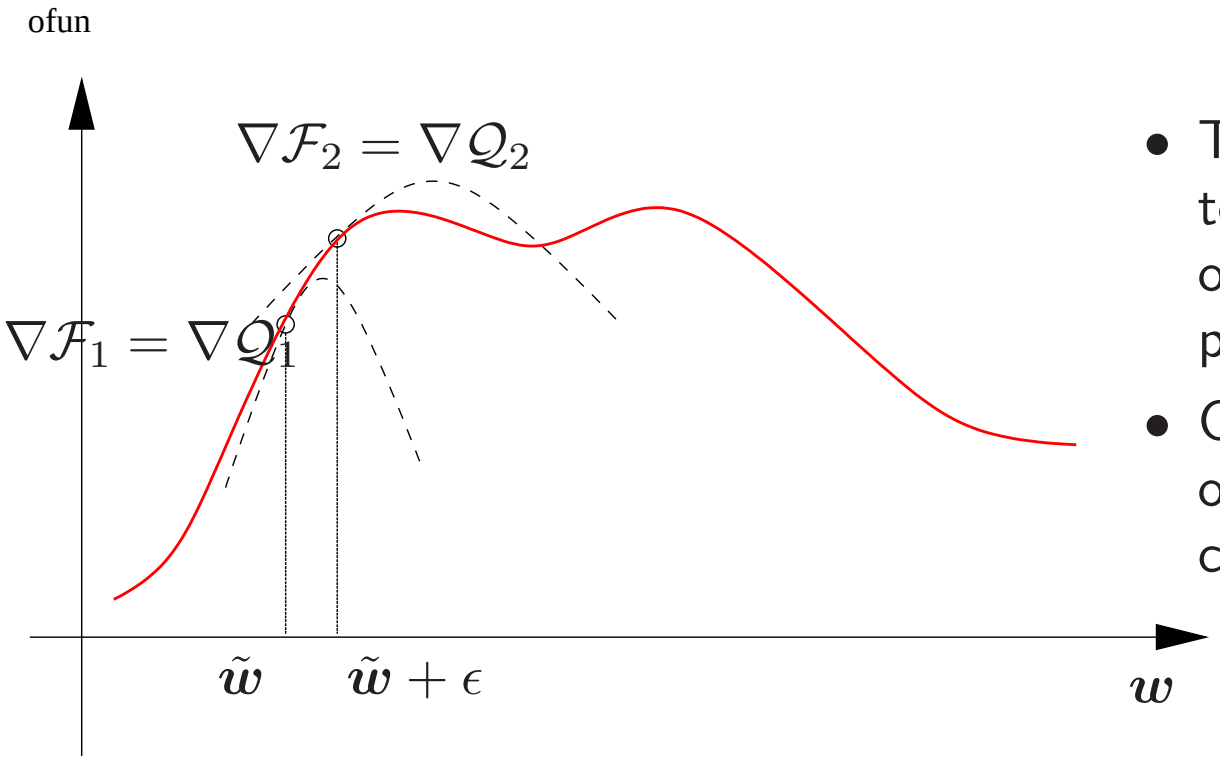
- Compromise:
  - local lower-bound by checking 2nd-order statistics

- Value of E chosen such that

$$\left.\frac{\partial^2 Q(\boldsymbol{w}_i, \tilde{\boldsymbol{w}}_i)}{\partial^2 \boldsymbol{w}_i}\right|_{\boldsymbol{w}_i=\tilde{\boldsymbol{w}}_i} \leq \left.\frac{\partial^2 \mathcal{F}(\boldsymbol{w}_i)}{\partial^2 \boldsymbol{w}_i}\right|_{\boldsymbol{w}_i=\tilde{\boldsymbol{w}}_i} \quad (13)$$



Hessian-constrained Auxiliary Function



- The auxiliary function forced to be more concave than the objective function at current parameter estimate
- Computatoin of curvature of objective function computationally costly

Finite-difference approximation through auxiliary function

$$\left.\frac{\partial^2\mathcal{F}(\boldsymbol{w}_i)}{\partial^2\boldsymbol{w}_i}\right|_{\boldsymbol{w}_i=\tilde{\boldsymbol{w}}_i} \approx \frac{\nabla Q_{\tilde{\boldsymbol{w}}_i+\boldsymbol{\epsilon}}(\boldsymbol{w}_i) - \nabla Q_{\tilde{\boldsymbol{w}}_i}(\boldsymbol{w}_i)}{\boldsymbol{\epsilon}} \tag{14}$$

Experiments with Discriminative Transforms

System	Adaptation		dev01sub		eval03	
	Training	Testing	ML	MPE	ML	MPE
SI	-	MLLR	31.1	28.5	30.2	27.0
	-	DLT	31.0	28.3	30.0	26.9
	-	DLT-MAP	-	-	-	-
SAT	MLLR	MLLR	30.4	27.4	29.3	26.4
	MLLR	DLT	30.3	27.3	29.4	26.3
	MLLR	DLT-MAP	-	-	-	-

- Discriminative transforms improve WER
- MAP estimate by strict lower bound: no significant update (result same as MLLR)
- MAP estimate by Hessian-constrained objective function: under experimentation

Conclusion

- Adaptation and Adaptive Training for varying acoustical environment
- Bayesian approaches effective for insufficient amount of adaptation data and unsupervised adaptation
- State-of-art systems use discriminative models
- Discriminative transforms found to improve WER
- Formulation of discriminative adaptive framework
  - Estimation of discriminative transform
  - Form of prior and its estimation
  - Integrating prior information discriminatively
  - Approximations for discriminative Bayesian inference