Unsupervised Model Adaptation using Information-Theoretic Criterion

Ariya Rastrow, Fred Jelinek, Abhinav Sethy and Bhuvana Ramabhadran

Center for Language and Speech Processing (CLSP)
Human Language Technology Center of Excellence (HLTCOE)
Electrical and Computer Engineering Department
Johns Hopkins University (JHU)

Overview

- **Motivation**
- Conditional Entropy based Adaptation
    - Entropy Definition
    - Entropy vs. Classifier Performance
    - Problems
    - Entropy-Stability
    - Proposed Objective Function
- Speech Recognition Task
    - Entropy/Gradient of Entropy for Speech Lattices
    - Language Model Adaptation
    - Experiment / Results / Explanation
- Future Work

Motivation

- The success of all **statistical and machine learning** techniques depends on:

  1. Availability of reasonable amount of **training data**
  2. Similarity between **underlying distribution of training and test** data

- **little amount of (or No) labeled data** for new domains/genres

  - **Frequent scenario** for Automatic Speech Recognition systems
  - The target domain contains **named entity and *N*-gram** sequences unique to the domain

- **Model adaptation** is crucial for these scenarios

Motivation

- The success of all **statistical and machine learning** techniques depends on:

    1. Availability of reasonable amount of **training data**
    2. Similarity between **underlying distribution of training and test** data

- **little amount of (or No) labeled data** for new domains/genres

    - **Frequent scenario** for Automatic Speech Recognition systems
    - The target domain contains **named entity and *N*-gram** sequences unique to the domain

- **Model adaptation** is crucial for these scenarios

Motivation

- The success of all **statistical and machine learning** techniques depends on:

    1. Availability of reasonable amount of **training data**
    2. Similarity between **underlying distribution of training and test** data

- **little amount of (or No) labeled data** for new domains/genres

    - **Frequent scenario** for Automatic Speech Recognition systems
    - The target domain contains **named entity and *N*-gram** sequences unique to the domain

- **Model adaptation** is crucial for these scenarios

Motivation

- The success of all **statistical and machine learning** techniques depends on:
    1. Availability of reasonable amount of **training data**
    2. Similarity between **underlying distribution of training and test** data
- **little amount of (or No) labeled data** for new domains/genres
    - **Frequent scenario** for Automatic Speech Recognition systems
    - The target domain contains **named entity and *N*-gram** sequences unique to the domain
- **Model adaptation** is crucial for these scenarios

Motivation

- The success of all **statistical and machine learning** techniques depends on:
  1. Availability of reasonable amount of **training data**
  2. Similarity between **underlying distribution of training and test** data
- **little amount of (or No) labeled data** for new domains/genres
  - **Frequent scenario** for Automatic Speech Recognition systems
  - The target domain contains **named entity and *N*-gram** sequences unique to the domain
- **Model adaptation** is crucial for these scenarios

Motivation

- The success of all **statistical and machine learning** techniques depends on:

    1. Availability of reasonable amount of **training data**
    2. Similarity between **underlying distribution of training and test** data

- **little amount of (or No) labeled data** for new domains/genres

    - **Frequent scenario** for Automatic Speech Recognition systems
    - The target domain contains **named entity and *N*-gram** sequences unique to the domain

- **Model adaptation** is crucial for these scenarios

Motivation

- The success of all **statistical and machine learning** techniques depends on:

    1. Availability of reasonable amount of **training data**
    2. Similarity between **underlying distribution of training and test** data

- **little amount of (or No) labeled data** for new domains/genres

    - **Frequent scenario** for Automatic Speech Recognition systems
    - The target domain contains **named entity and *N*-gram** sequences unique to the domain

- **Model adaptation** is crucial for these scenarios

Motivation

- The success of all **statistical and machine learning** techniques depends on:

    1. Availability of reasonable amount of **training data**
    2. Similarity between **underlying distribution of training and test** data

- **little amount of (or No) labeled data** for new domains/genres

    - **Frequent scenario** for Automatic Speech Recognition systems
    - The target domain contains **named entity and *N*-gram** sequences unique to the domain

- **Model adaptation** is crucial for these scenarios

J. Jiang, 2008

Motivation

- In this talk, we present a general framework for **unsupervised model adaptation**
    - The proposed method is based on **Conditional Entropy**
    - The idea is to improve the performance of initial model (trained on out-of-domain data) by **adjusting the initial decision boundaries** on in-domain data
- Directions for using the proposed framework as a **Semi-Supervised Learning (SSL)** technique is also presented

Motivation

- In this talk, we present a general framework for **unsupervised model adaptation**
  - The proposed method is based on **Conditional Entropy**
  - The idea is to improve the performance of initial model (trained on out-of-domain data) by **adjusting the initial decision boundaries** on in-domain data
- Directions for using the proposed framework as a **Semi-Supervised Learning (SSL)** technique is also presented

Motivation

- In this talk, we present a general framework for **unsupervised model adaptation**
  - The proposed method is based on **Conditional Entropy**
  - The idea is to improve the performance of initial model (trained on out-of-domain data) by **adjusting the initial decision boundaries** on in-domain data
- Directions for using the proposed framework as a **Semi-Supervised Learning (SSL)** technique is also presented

Motivation

- In this talk, we present a general framework for **unsupervised model adaptation**
  - The proposed method is based on **Conditional Entropy**
  - The idea is to improve the performance of initial model (trained on out-of-domain data) by **adjusting the initial decision boundaries** on in-domain data
- Directions for using the proposed framework as a **Semi-Supervised Learning (SSL)** technique is also presented

Overview

- Motivation
- **Conditional Entropy based Adaptation**
  - **Entropy Definition**
  - **Entropy vs. Classifier Performance**
  - **Problems**
  - **Entropy-Stability**
  - **Proposed Objective Function**
- Speech Recognition Task
  - Entropy/Gradient of Entropy for Speech Lattices
  - Language Model Adaptation
  - Experiment / Results / Explanation
- Future Work

Conditional Entropy

- Entropy: Measure of uncertainty associated with a *random variable*
- definition:

$$H(Y) = -\sum_y p(y) \log p(y)$$

Conditional Entropy

- Entropy: Measure of uncertainty associated with a *random variable*
- definition:

$$H(Y) = -\sum_{y} p(y) \log p(y)$$

- Now, imagine you want to measure uncertainty in Y after observing X

Conditional Entropy

- Entropy: Measure of uncertainty associated with a *random variable*
- definition:

$$H(Y) = -\sum_y p(y) \log p(y)$$

- Now, imagine you want to measure uncertainty in Y after observing X
  - Conditional Entropy:

$$H(Y|X) = E_X[H(Y|X = x)] = -\sum_x p(x) \sum_y p(y|x) \log p(y|x)$$

Classifier Performance

- Fano's Inequality :

$$P_e = P\{\hat{\mathbf{Y}} \neq \mathbf{Y}\} \geq \frac{H(\mathbf{Y}|\mathbf{X}) - 1}{\log|\mathcal{Y}|}$$

- Classification goal:
    - estimate Y from X with a low probability of misclassification

Classifier Performance

- Fano's Inequality :

$$P_e = P\{\hat{\mathbf{Y}} \neq \mathbf{Y}\} \geq \frac{H(\mathbf{Y}|\mathbf{X}) - 1}{\log|\mathcal{Y}|}$$

- Classification goal:
    - estimate Y from X with a low probability of misclassification

- Entropy needs to be low in order to have low probability of misclassification

Classifier Performance

- Fano's Inequality :

$$P_e(\theta) = P\{\hat{\mathbf{Y}} \neq \mathbf{Y} \ |\theta\} \geq \frac{H_\theta(\mathbf{Y}|\mathbf{X}) - 1}{\log|\mathcal{Y}|}$$

- Classification goal:
    - estimate Y from X with a low probability of misclassification

- Entropy needs to be low in order to have low probability of misclassification

Minimum Entropy Criterion

- Entropy Regularization

  Grandvalet and Bengio, NIPS 2004

  - Maximum Likelihood on labeled data + Minimum Conditional Entropy on unlabeled data

- Minimum Entropy Clustering

  Li, Zhang and Jiang, 2004

  - Non-parametric approach which improves over *k*-means clustering.

- Minimum Entropy Solution favors models which have their **decision boundaries** passing through **low-density regions** of the input distribution

Minimum Entropy Criterion

- Entropy Regularization

  Grandvalet and Bengio, NIPS 2004

  - Maximum Likelihood on labeled data + Minimum Conditional Entropy on unlabeled data

- Minimum Entropy Clustering

  Li, Zhang and Jiang, 2004

  - Non-parametric approach which improves over *k*-means clustering.

- Minimum Entropy Solution favors models which have their **decision boundaries** passing through **low-density regions** of the input distribution

Problems with Min. Entropy Solution?

- Trivial solutions:
  - Imagine a model which classifies all the inputs as one class.

$$H_\theta(\mathbf{Y}|\mathbf{X}) = 0$$

- Overlapped Classes and Imbalanced priors:
  - No valid low-density regions for decision boundaries

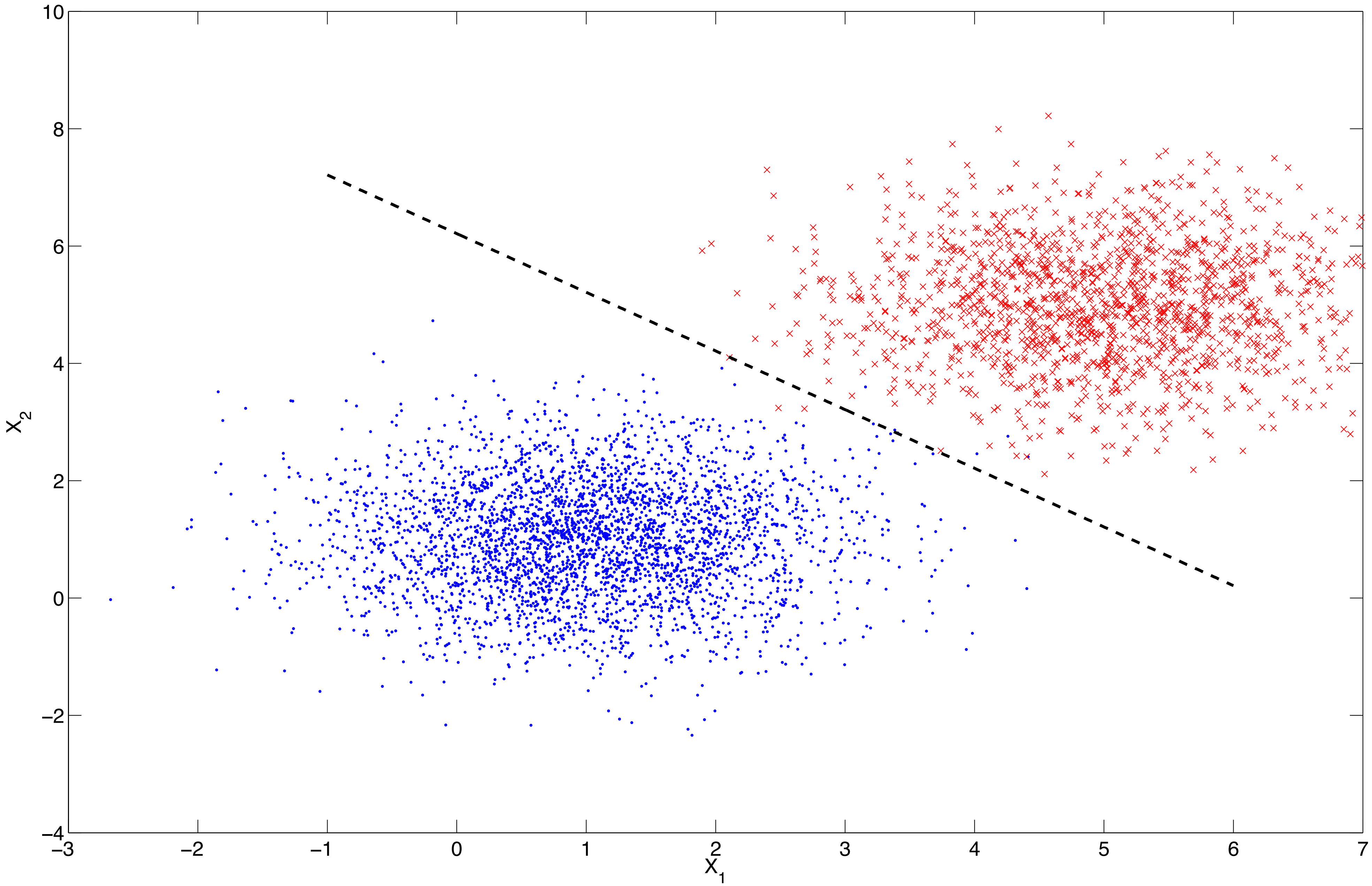
Problems with Min. Entropy Solution?

- Trivial solutions:
  - Imagine a model which classifies all the inputs as one class.

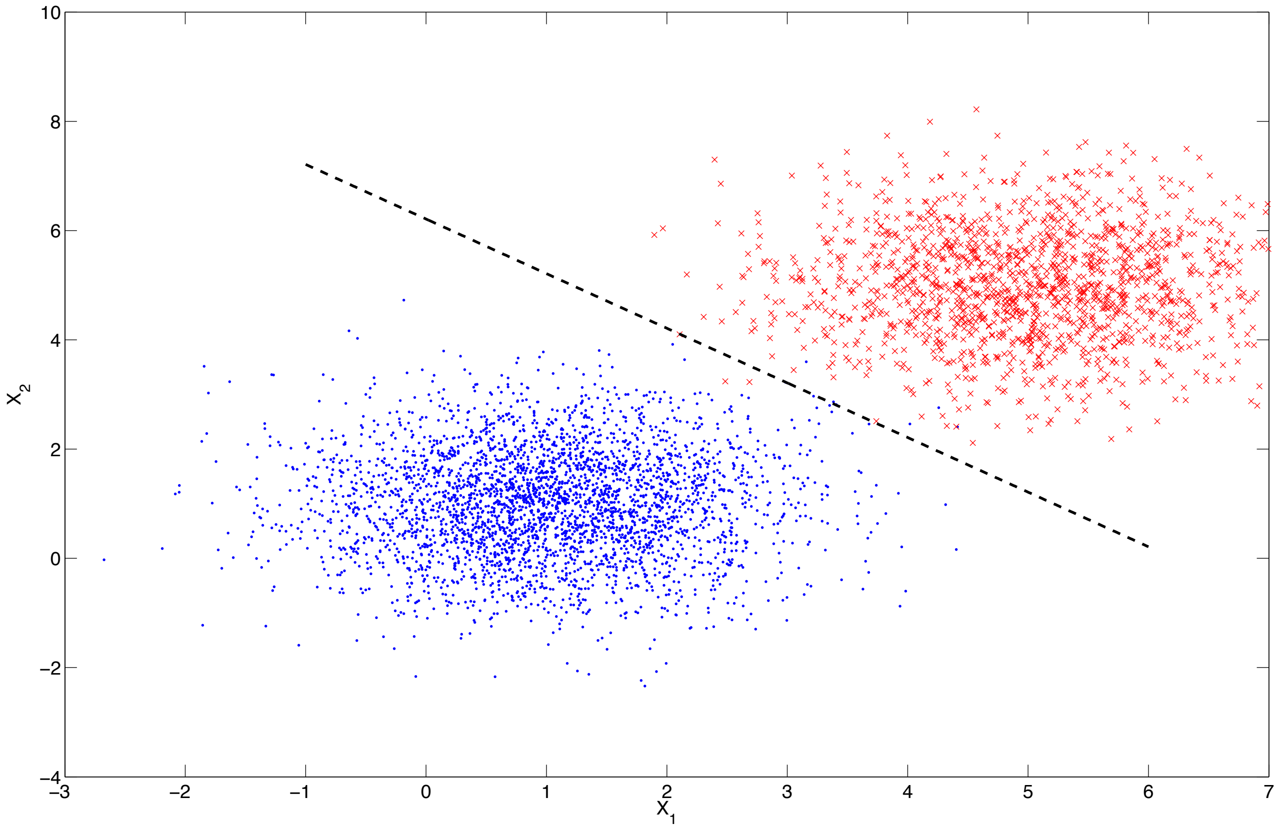
$$H_\theta(\mathbf{Y}|\mathbf{X}) = 0$$

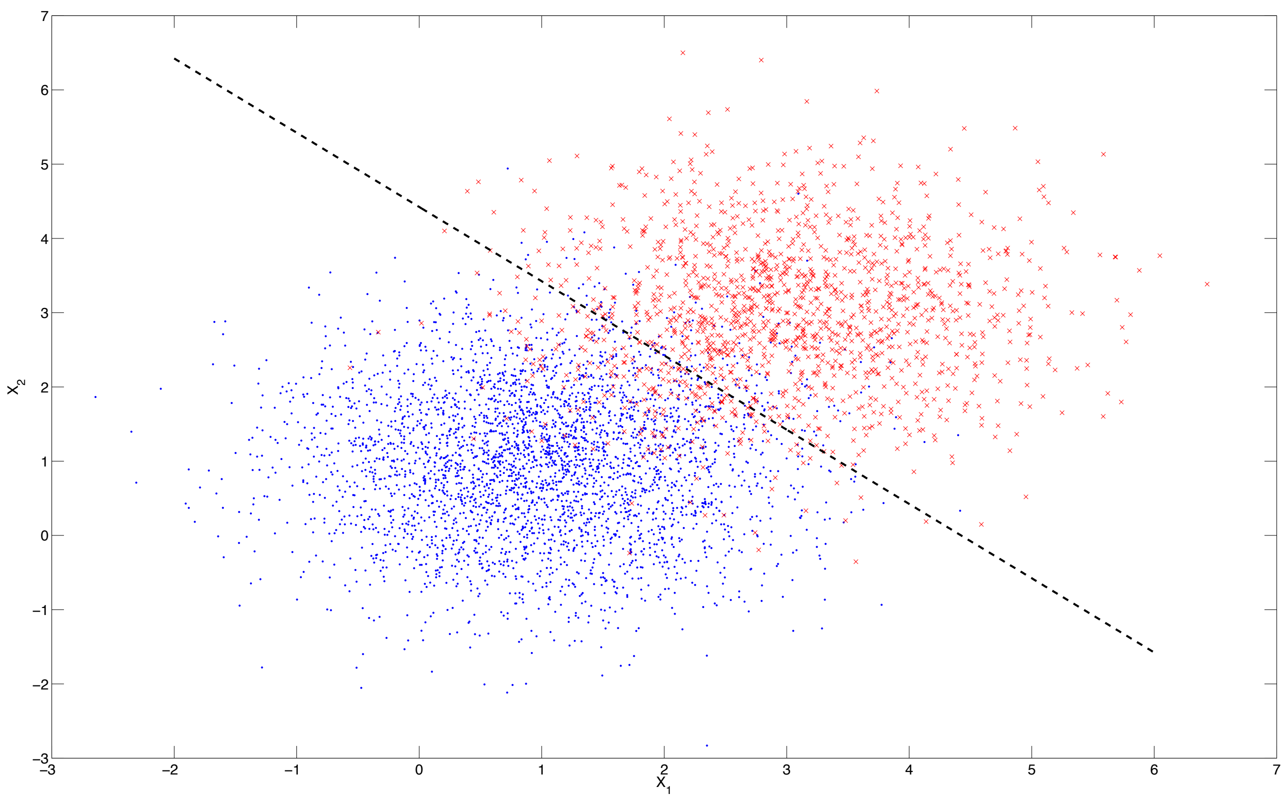
- Overlapped Classes and Imbalanced priors:
  - No valid low-density regions for the decision boundaries

Problems with Min. Entropy Solution?



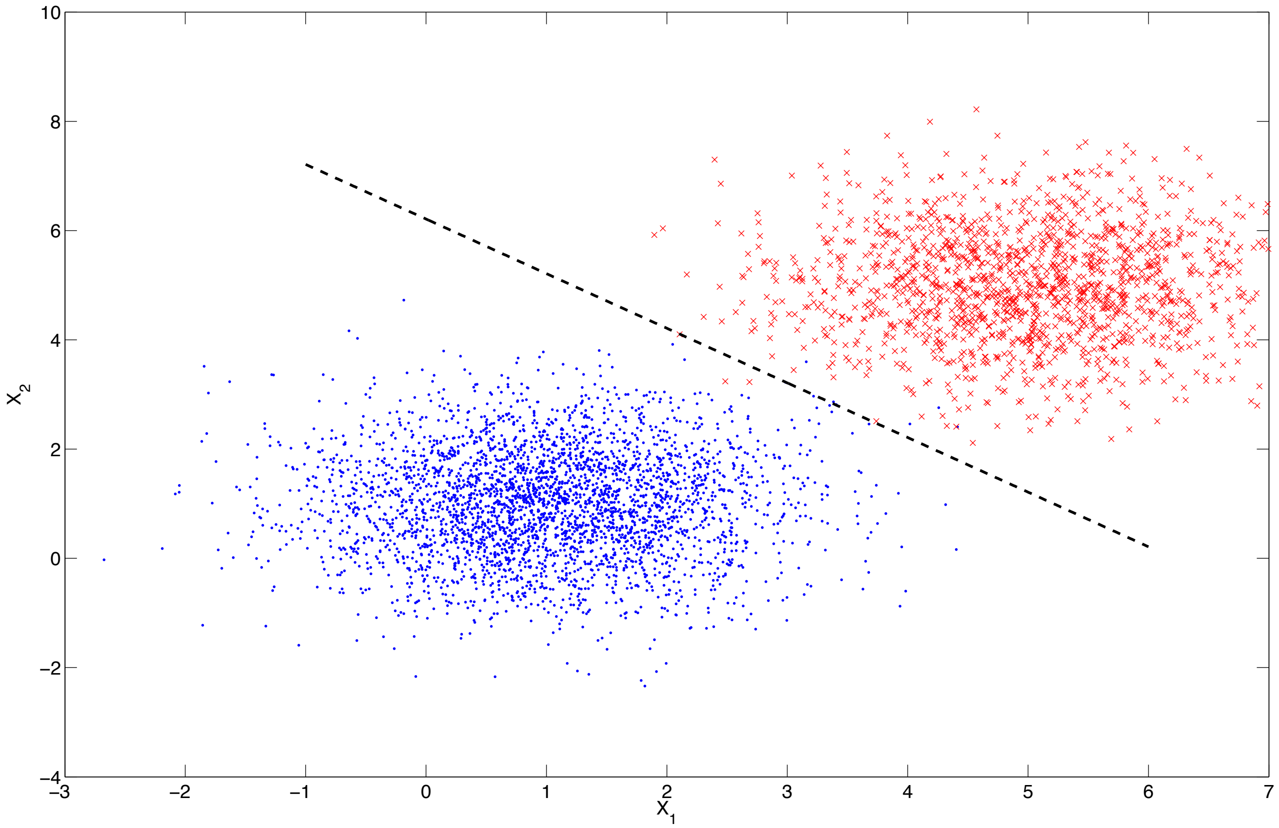
Problems with Min. Entropy Solution?

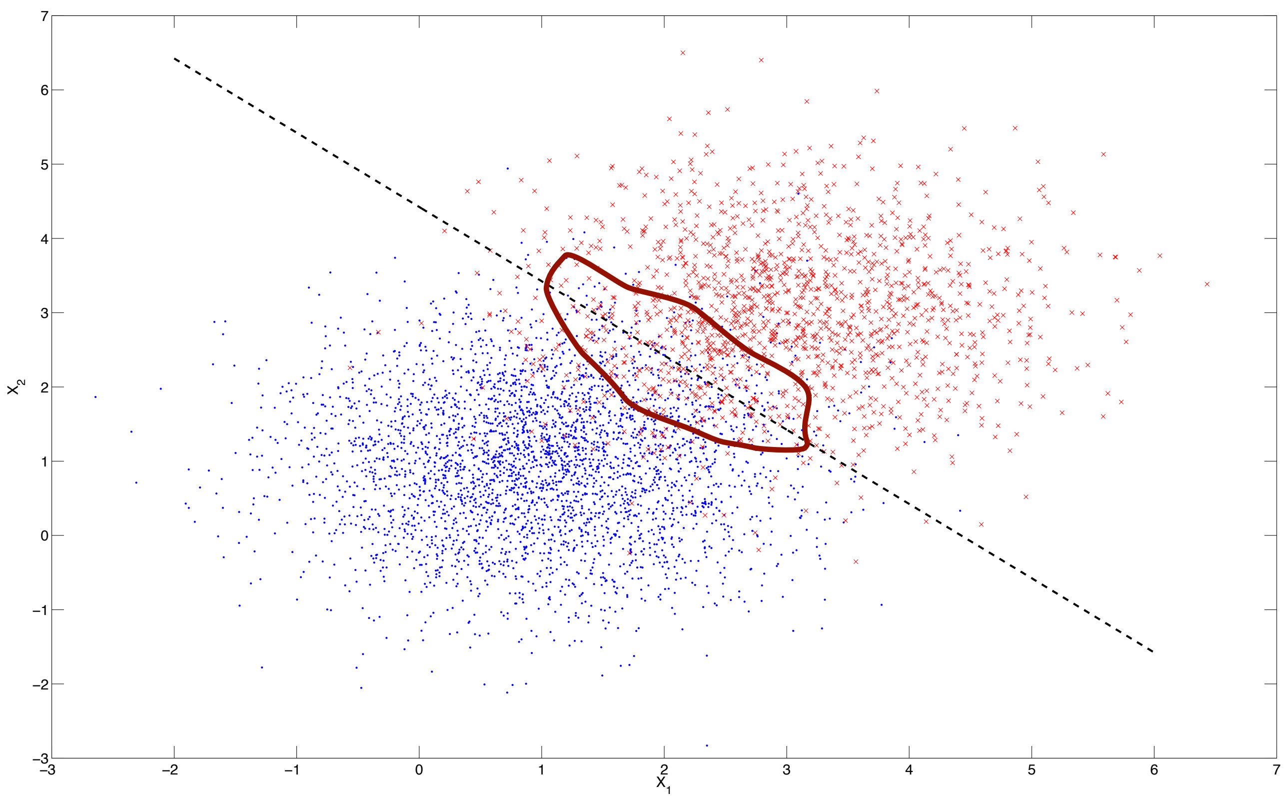




- Minimum Entropy Solution is favoring models which have their **decision boundaries** passing through **low-density regions** of the input distribution

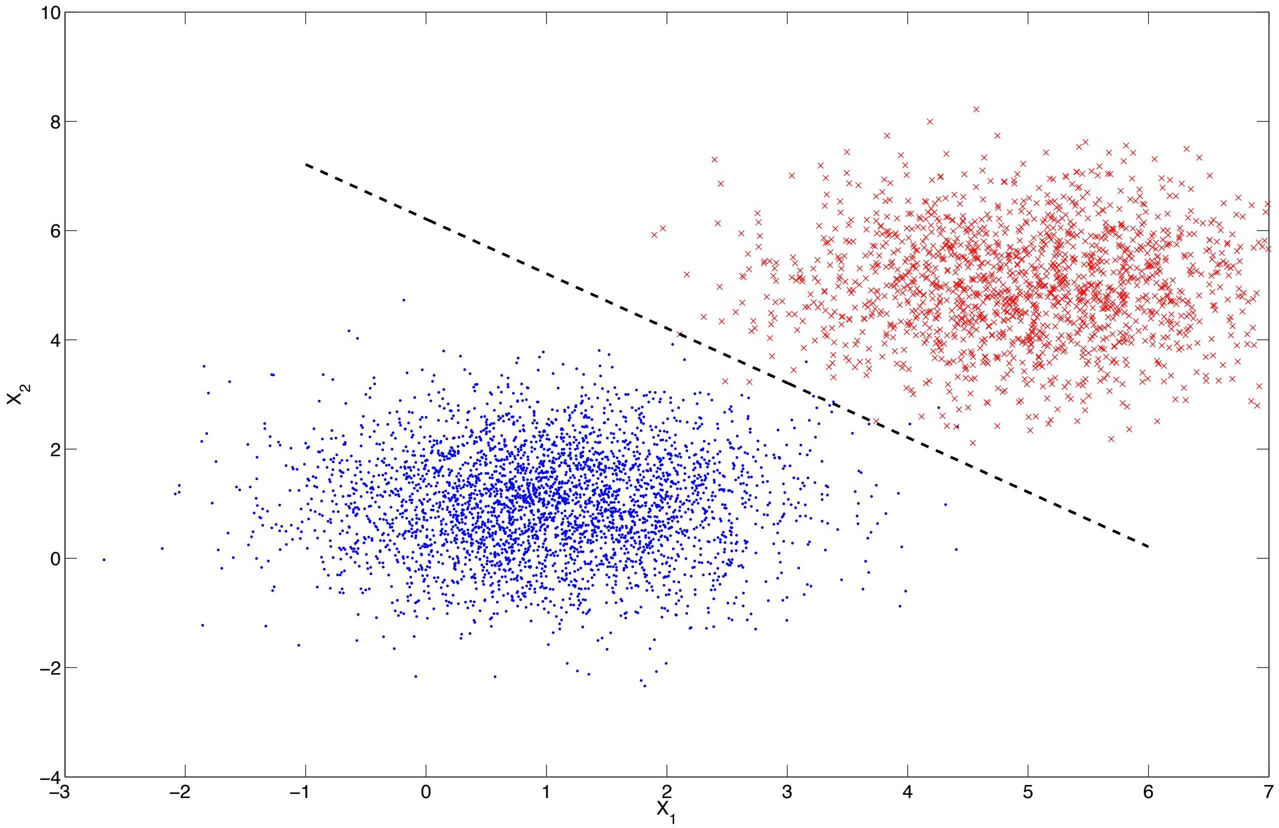
Problems with Min. Entropy Solution?

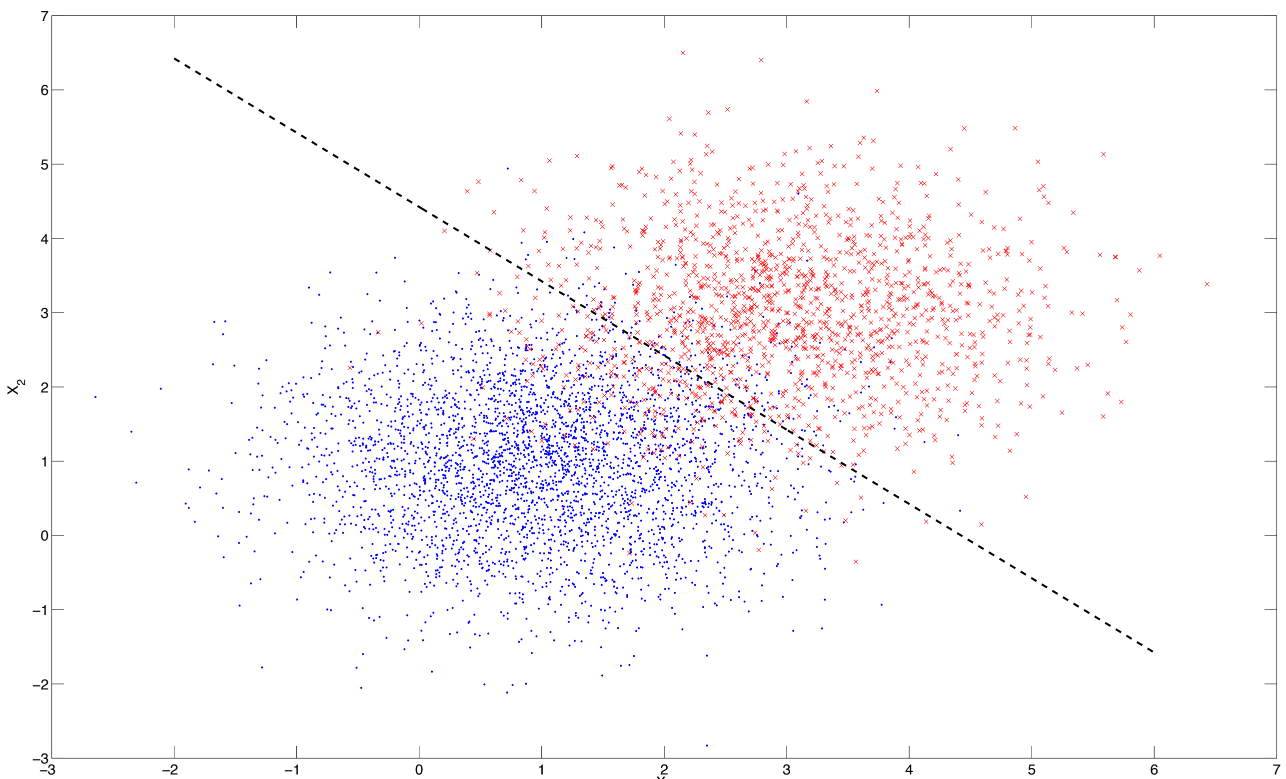


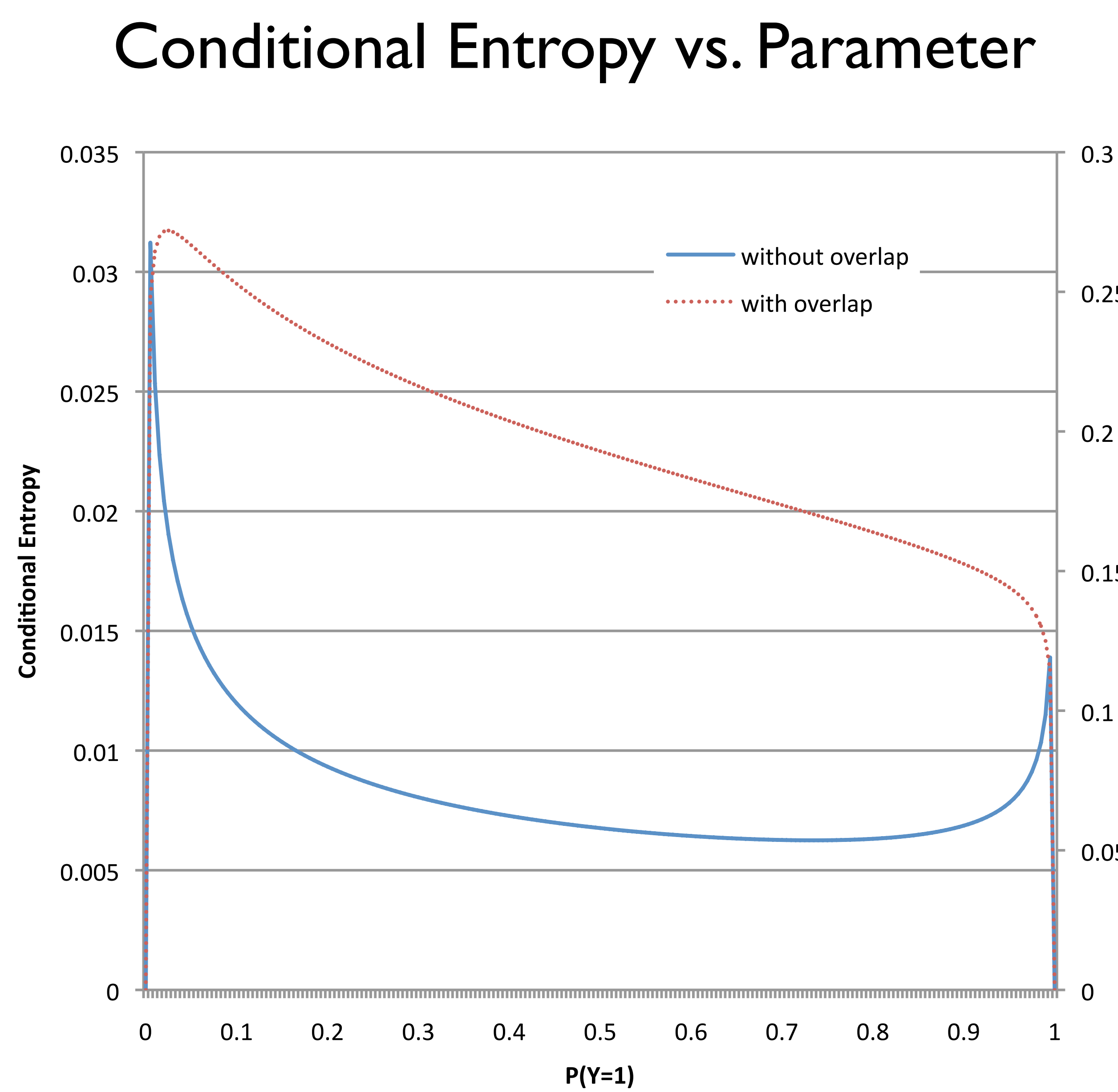


- For the overlapped classes, there is **no low-density region** at the boundary of the classes

Conditional Entropy. Problems?







Entropy Stability

- Entropy Stability:
  - reciprocal of:

$$\left\|\frac{\partial H_\theta(\mathbf{Y}|\mathbf{X})}{\partial \theta}\right\|_p$$

  - Measures how **stable posterior probabilities are w.r.t the model parameter** through the following equation:

$$\left\|\int p(x)\left(\sum_y \frac{\partial p_\theta(y|x)}{\partial \theta} \log p_\theta(y|x)\right) dx\right\|_p$$

  - A high value indicates regions where posterior probabilities are **sensitive to parameters**

Entropy Stability

- Entropy Stability:
  - reciprocal of:

$$\left\| \frac{\partial H_\theta(\mathbf{Y}|\mathbf{X})}{\partial \theta} \right\|_p$$

  - Measures how **stable posterior probabilities are w.r.t the model parameter** through the following equation:

$$\left\| \int p(x) \left(\sum_y \frac{\partial p_\theta(y|x)}{\partial \theta} \log p_\theta(y|x) \right) dx \right\|_p$$

  - A high value indicates regions where posterior probabilities are **sensitive to parameters**

Entropy Stability

- Entropy Stability:
  - reciprocal of:

$$\left\| \frac{\partial H_\theta(\mathbf{Y}|\mathbf{X})}{\partial \theta} \right\|_p$$

  - Measures how **stable posterior probabilities are w.r.t the model parameter** through the following equation:

$$\left\| \int p(x) \left(\sum_y \frac{\partial p_\theta(y|x)}{\partial \theta} \log p_\theta(y|x) \right) dx \right\|_p$$

  - A high value indicates regions where posterior probabilities are **sensitive to parameters**

Objective Function

- Unsupervised Model Adaptation

$$\theta_{\mathbf{new}} = \underset{\theta}{\operatorname{argmin}} \left(H_\theta(\mathbf{Y}|\mathbf{X}) + \gamma \left|\left| \frac{\partial H_\theta(\mathbf{Y}|\mathbf{X})}{\partial \theta} \right|\right|_{p'} + \lambda \left|\left| \theta - \theta_{\mathbf{init}} \right|\right|_p \right)$$

  - Using Entropy Stability only regions close to **the overlapped parts** of the input distribution are accepted
  - Then using **minimum entropy criterion**, we find the optimum solutions for the model parameters
  - The L_p regularizer prevents the model parameters to get **too deviated from initial model** (supervision)

Objective Function

- Unsupervised Model Adaptation

$$\theta_{\mathbf{new}} = \underset{\theta}{\operatorname{argmin}} \left(H_\theta(\mathbf{Y}|\mathbf{X}) + \gamma \boxed{\left|\left| \frac{\partial H_\theta(\mathbf{Y}|\mathbf{X})}{\partial \theta} \right|\right|_{p'}} + \lambda \left|\left|\theta - \theta_{\mathbf{init}}\right|\right|_p \right)$$

  - Using Entropy Stability only regions close to **the overlapped parts** of the input distribution are accepted
  - Then using **minimum entropy criterion**, we find the optimum solutions for the model parameters
  - The L_p regularizer prevents the model parameters to get **too deviated from initial model** (supervision)

Objective Function

- Unsupervised Model Adaptation

$$\theta_{\mathbf{new}} = \underset{\theta}{\operatorname{argmin}} \left(H_\theta(\mathbf{Y}|\mathbf{X}) + \gamma \left|\left| \frac{\partial H_\theta(\mathbf{Y}|\mathbf{X})}{\partial \theta} \right|\right|_{p'} + \lambda \left|\left| \theta - \theta_{\mathbf{init}} \right|\right|_p \right)$$

  - Using Entropy Stability only regions close to **the overlapped parts** of the input distribution are accepted
  - Then using **minimum entropy criterion**, we find the optimum solutions for the model parameters
  - The L_p regularizer prevents the model parameters to get **too deviated from initial model** (supervision)

Objective Function

- Unsupervised Model Adaptation

$$\theta_{\mathbf{new}} = \underset{\theta}{\operatorname{argmin}} \left(H_\theta(\mathbf{Y}|\mathbf{X}) + \gamma \left|\left| \frac{\partial H_\theta(\mathbf{Y}|\mathbf{X})}{\partial \theta} \right|\right|_{p'} + \lambda \boxed{||\theta - \theta_{\mathbf{init}}||_p} \right)$$

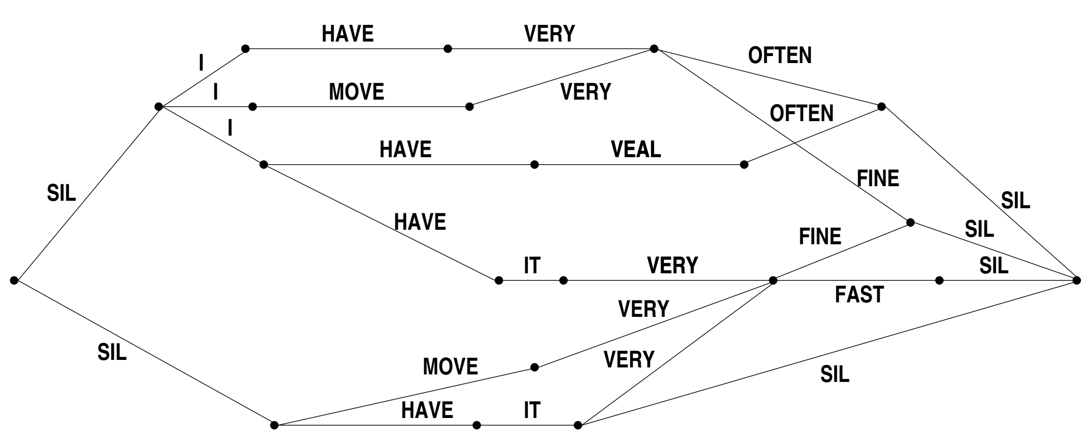
  - Using Entropy Stability only regions close to **the overlapped parts** of the input distribution are accepted
  - Then using **minimum entropy criterion**, we find the optimum solutions for the model parameters
  - The L_p regularizer prevents the model parameters to get **too deviated from initial model** (supervision)

Overview

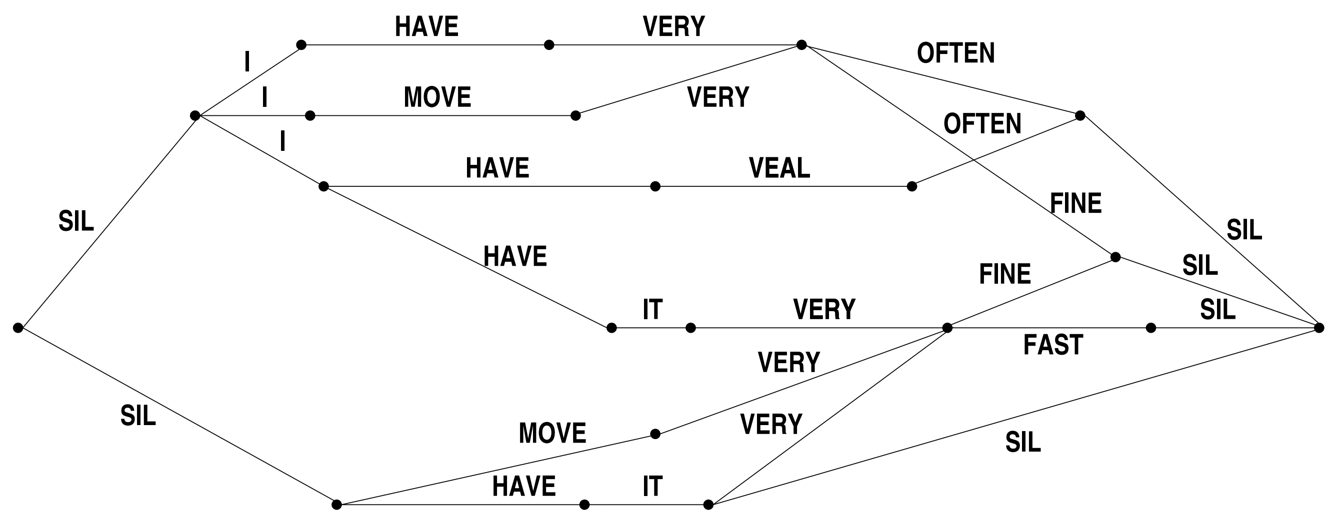
- Motivation
- Conditional Entropy based Adaptation
  - Entropy Definition
  - Entropy vs. Classifier Performance
  - Problems
  - Entropy-Stability
  - Proposed Objective Function
- **Speech Recognition Task**
  - **Entropy/Gradient of Entropy for Speech Lattices**
  - **Language Model Adaptation**
  - **Experiment / Results / Explanation**
- Future Work

Speech Recognition

- Moving to speech recognition task:
  - Y is now sequence of words ($\mathbf{W}$)
  - For a given chunk of speech data, almost every $\mathbf{W}$ is possible (with different likelihoods)
  - Need for compact representation of space
- Lattice is acyclic directed graph which represents the most likely paths (sequence of words)

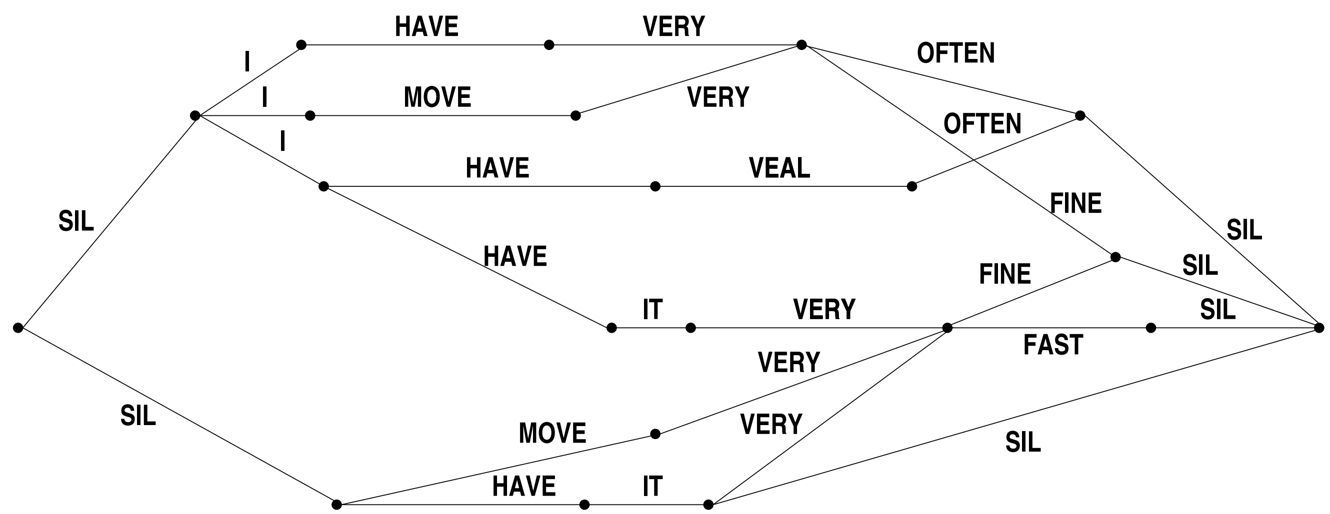


Entropy/Gradient of Entropy for Speech Lattices



$$H_\theta(\mathbf{W}|\mathbf{X}=x) \approx H_\theta(\mathbf{W}|\mathcal{L})$$

Entropy/Gradient of Entropy for Speech Lattices

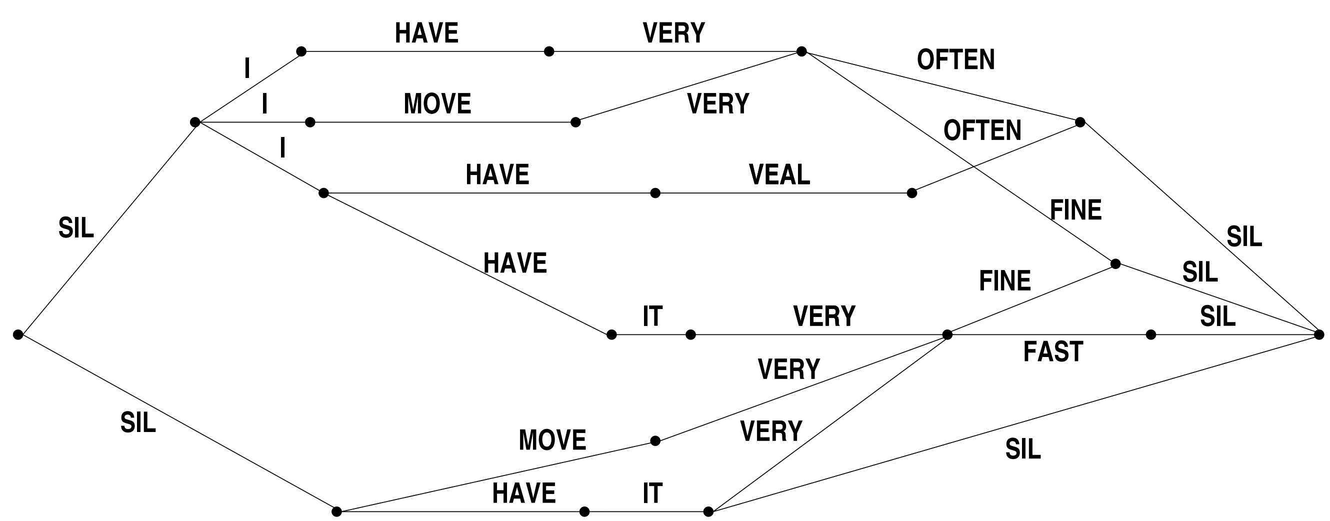


$$H_\theta(\mathbf{W}|\mathbf{X}=x) \approx \boxed{H_\theta(\mathbf{W}|\mathcal{L})}$$

Enumerating over all the paths is intractable!

$$-\sum_{d\in\mathcal{L}} \frac{p(d)}{Z} \log \frac{p(d)}{Z}$$

Entropy/Gradient of Entropy for Speech Lattices



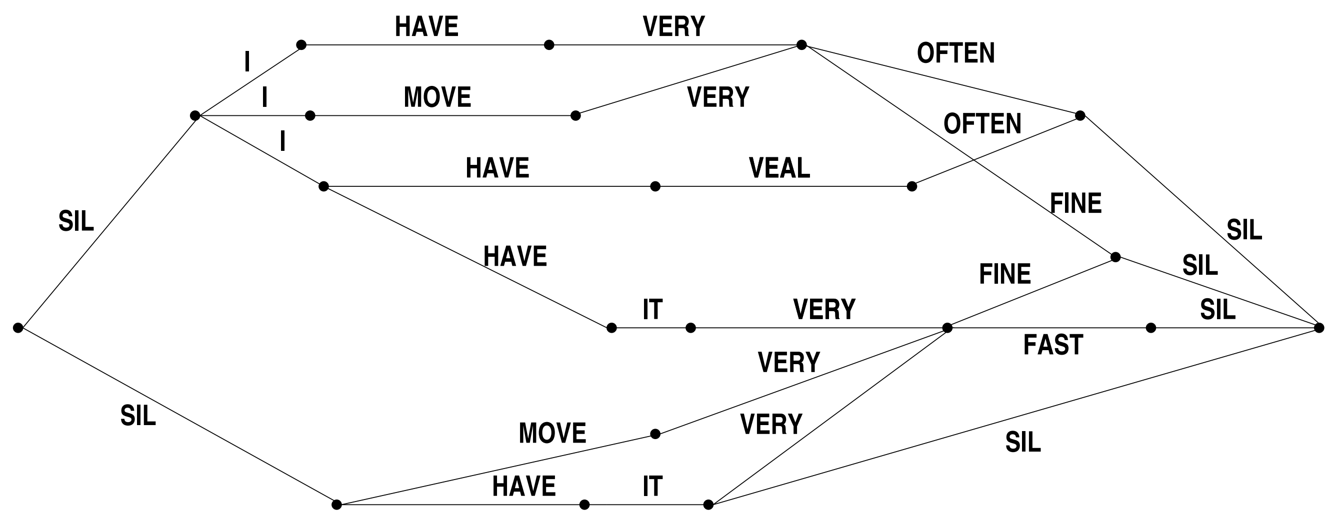
$$H_\theta(\mathbf{W}|\mathbf{X}=x) \approx H_\theta(\mathbf{W}|\mathcal{L})$$

- Entropy (the gradient of entropy) can be computed efficiently on the lattices using Finite-State Machines and **First- and Second-order Expectation Semirings**

Li and Eisner, EMNLP 2009

  - The implementation based on OpenFST^{TM} will be released

Entropy/Gradient of Entropy for Speech Lattices



$$H_\theta(\mathbf{W}|\mathbf{X}=x) \approx H_\theta(\mathbf{W}|\mathcal{L})$$

- Entropy (the gradient of entropy) can be computed efficiently on the lattices using Finite-State Machines and **First- and Second-order Expectation Semirings**

Li and Eisner, EMNLP 2009

$$H_\theta(\mathbf{W}|\mathbf{X}) \approx \frac{1}{N}\sum_{i=1}^{N} H_\theta(\mathbf{W}|\mathcal{L}_i)$$

First/Second Order Expectation Semirings

$$H(p) = -\sum_{d \in \mathcal{L}} \frac{p(d)}{Z} \log(\frac{p(d)}{Z})$$

$$= \log Z - \frac{1}{Z} \sum_{d \in \mathcal{L}_i} p(d) \log p(d)$$

$$= \log Z - \frac{\bar{r}}{Z}$$

First/Second Order Expectation Semirings

$$H(p) = -\sum_{d \in \mathcal{L}} \frac{p(d)}{Z} \log(\frac{p(d)}{Z})$$

$$= \log Z - \frac{1}{Z} \sum_{d \in \mathcal{L}_i} p(d) \log p(d)$$

$$= \log Z - \frac{\bar{r}}{Z}$$

We need to calculate $< Z, \bar{r} >$

First/Second Order Expectation Semirings

$$H(p) = -\sum_{d \in \mathcal{L}} \frac{p(d)}{Z} \log(\frac{p(d)}{Z})$$

$$= \log Z - \frac{1}{Z} \sum_{d \in \mathcal{L}_i} p(d) \log p(d)$$

$$= \log Z - \frac{\bar{r}}{Z}$$

We need to calculate $< Z, \bar{r} >$

First-order Expectation Semiring

First/Second Order Expectation Semirings

$$H(p) = -\sum_{d \in \mathcal{L}} \frac{p(d)}{Z} \log(\frac{p(d)}{Z})$$

$$= \log Z - \frac{1}{Z} \sum_{d \in \mathcal{L}_i} p(d) \log p(d)$$

$$= \log Z - \frac{\bar{r}}{Z}$$

We need to calculate $< Z, \bar{r} >$

First-order Expectation Semiring

$$< p, r > = < p_e, p_e \log p_e >$$

Element	$\langle p, r \rangle$
$\langle p_1, r_1 \rangle \otimes \langle p_2, r_2 \rangle$	$\langle p_1 p_2,\ p_1 r_2 + p_2 r_1 \rangle$
$\langle p_1, r_1 \rangle \oplus \langle p_2, r_2 \rangle$	$\langle p_1 + p_2,\ r_1 + r_2 \rangle$
0	$\langle 0, 0 \rangle$
1	$\langle 1, 0 \rangle$

First/Second Order Expectation Semirings

$$H(p) = -\sum_{d \in \mathcal{L}} \frac{p(d)}{Z} \log(\frac{p(d)}{Z})$$

$$= \log Z - \frac{1}{Z} \sum_{d \in \mathcal{L}_i} p(d) \log p(d)$$

$$= \log Z - \frac{\bar{r}}{Z}$$

We need to calculate $< Z, \bar{r} >$

First-order Expectation Semiring

$$< p, r >=< p_e, p_e \log p_e >$$

Element	$\langle p, r \rangle$
$\langle p_1, r_1 \rangle \otimes \langle p_2, r_2 \rangle$	$\langle p_1 p_2,\ p_1 r_2 + p_2 r_1 \rangle$
$\langle p_1, r_1 \rangle \oplus \langle p_2, r_2 \rangle$	$\langle p_1 + p_2,\ r_1 + r_2 \rangle$
$\mathbf{0}$	$\langle 0, 0 \rangle$
$\mathbf{1}$	$\langle 1, 0 \rangle$

Forward algorithm will return $< Z, \bar{r} >$ as the weight of the final node.

Language Model Adaptation

- **little amount of labeled data** for new domain/ genre

Language Model Adaptation

- **little amount of labeled data** for new domain/genre
    - **LM interpolation** is most commonly used for adaptation:

$$P(w_i|h) = \lambda P_B(w_i|h) + (1-\lambda) P_A(w_i|h)$$

Language Model Adaptation

- **little amount of labeled data** for new domain/genre
    - **LM interpolation** is most commonly used for adaptation:

$$P(w_i|h) = \lambda \underbrace{\boxed{P_B(w_i|h)}}_{\text{out-of-domain } N\text{-grams}} + (1-\lambda)P_A(w_i|h)$$

Language Model Adaptation

- **little amount of labeled data** for new domain/genre
    - **LM interpolation** is most commonly used for adaptation:

$$P(w_i|h) = \lambda P_B(w_i|h) + (1-\lambda)\underbrace{\boxed{P_A(w_i|h)}}_{\text{in-domain } N\text{-grams}}$$

Language Model Adaptation

- **little amount of labeled data** for new domain/ genre
  - **LM interpolation** is most commonly used for adaptation:

$$P(w_i|h) = \lambda P_B(w_i|h) + (1-\lambda)P_A(w_i|h)$$

Interpolation weight

Language Model Adaptation

- **little amount of labeled data** for new domain/genre
  - **LM interpolation** is most commonly used for adaptation:

$$P(w_i|h) = \lambda P_B(w_i|h) + (1-\lambda) P_A(w_i|h)$$

- λ is optimized using the following criterion:

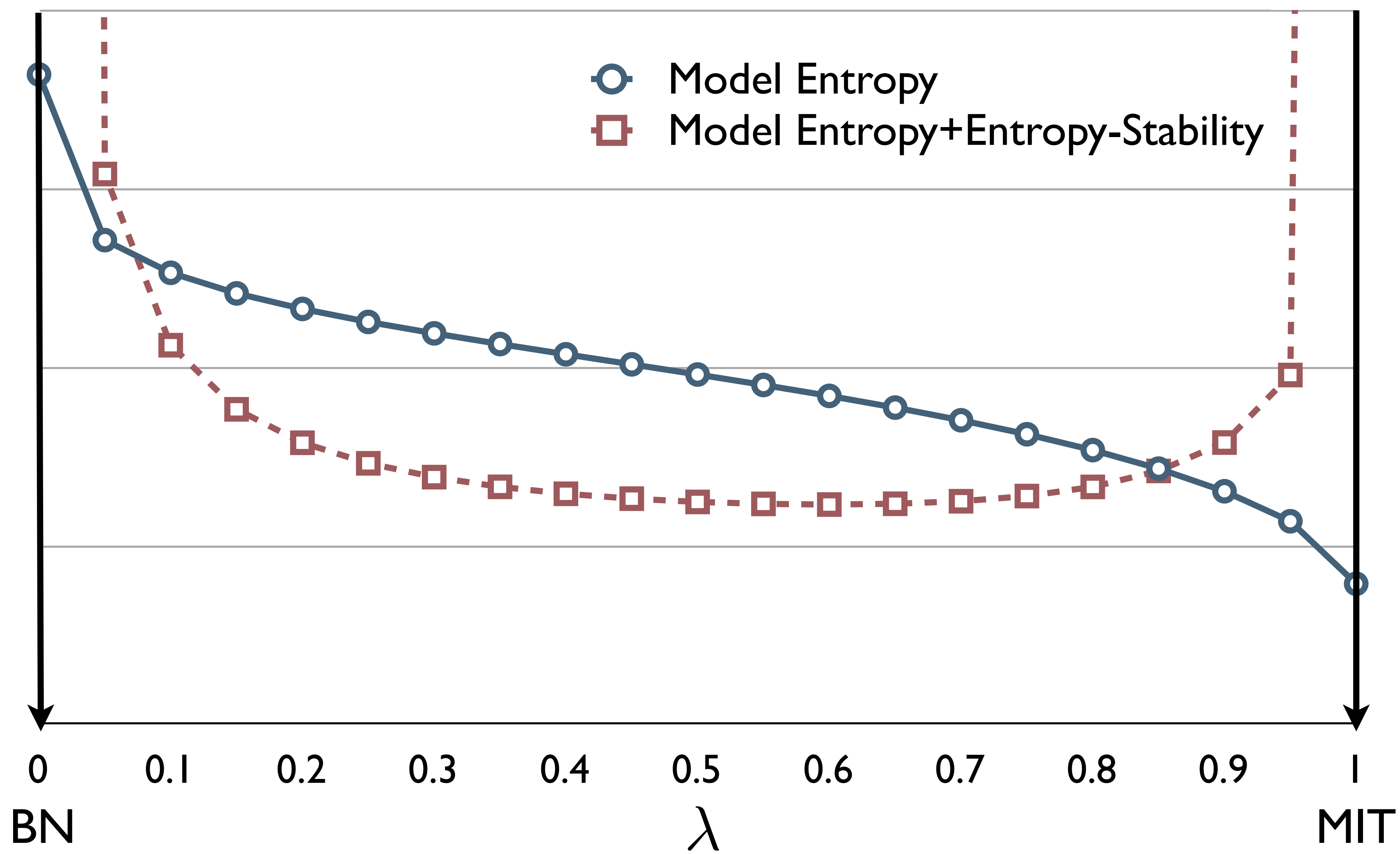
$$\hat{\lambda} = \underset{0 \leq \lambda \leq 1}{\operatorname{argmin}} H_\lambda(\mathbf{Y}|\mathbf{X}) + \left|\frac{\partial H_\lambda(\mathbf{Y}|\mathbf{X})}{\partial \lambda}\right|$$

Experiments

- The LVCSR system is based on the *2008 IBM Speech recognition system*.
    - The acoustic models are state-of-the-art discriminatively trained
- The out-of-domain LM (P_B) is built on 340M words (8 BN corpora)
    - 8 hours for building target specific LM (P_A)
    - 8 hours for evaluation and calculation of our objective function
    - 2.5 hours as development set (for supervised tuning of weight)

Experiments

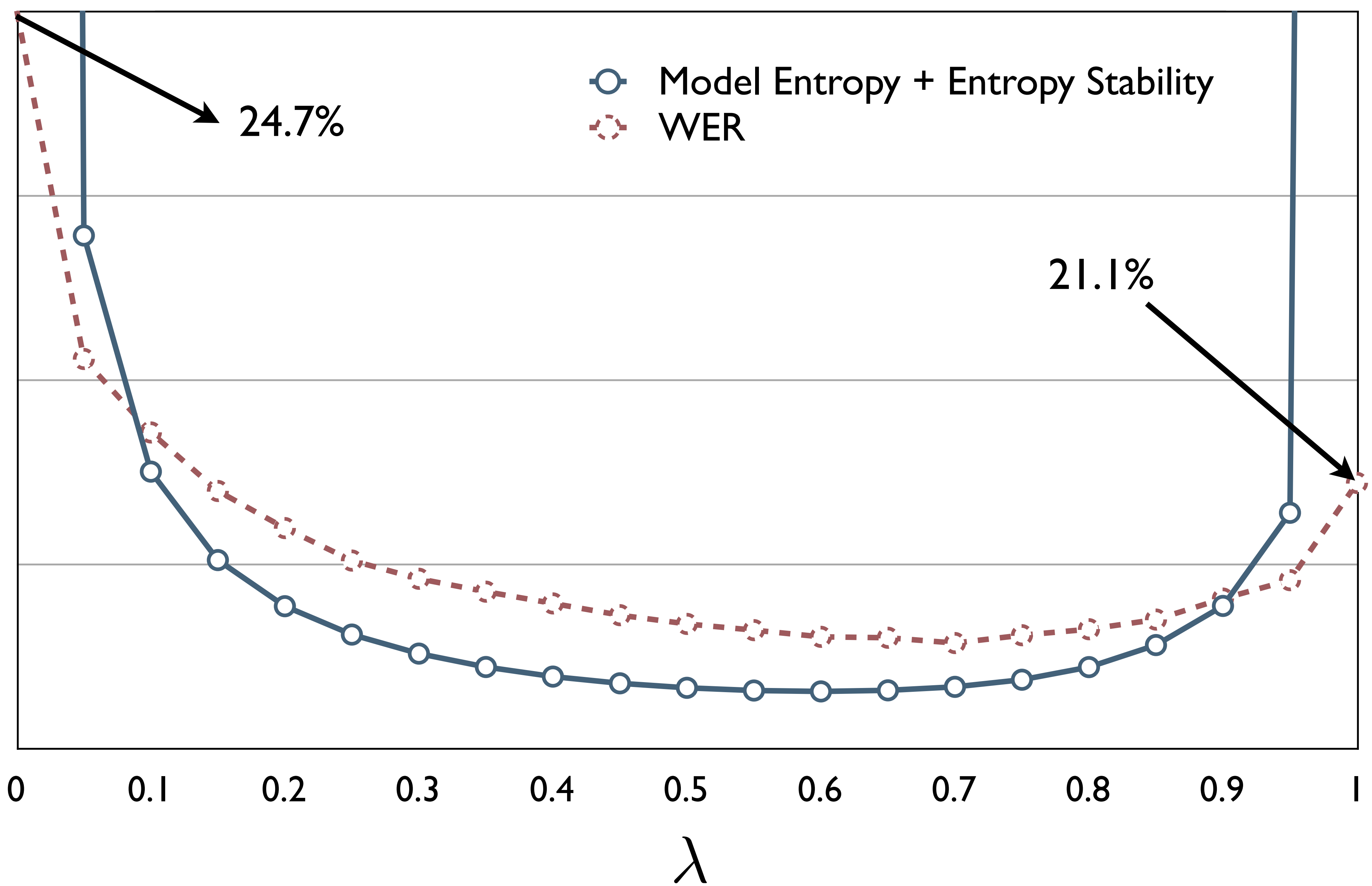
$$P(w_i|h) = \lambda P_B(w_i|h) + (1-\lambda)P_A(w_i|h)$$



- Considering only conditional entropy is not useful

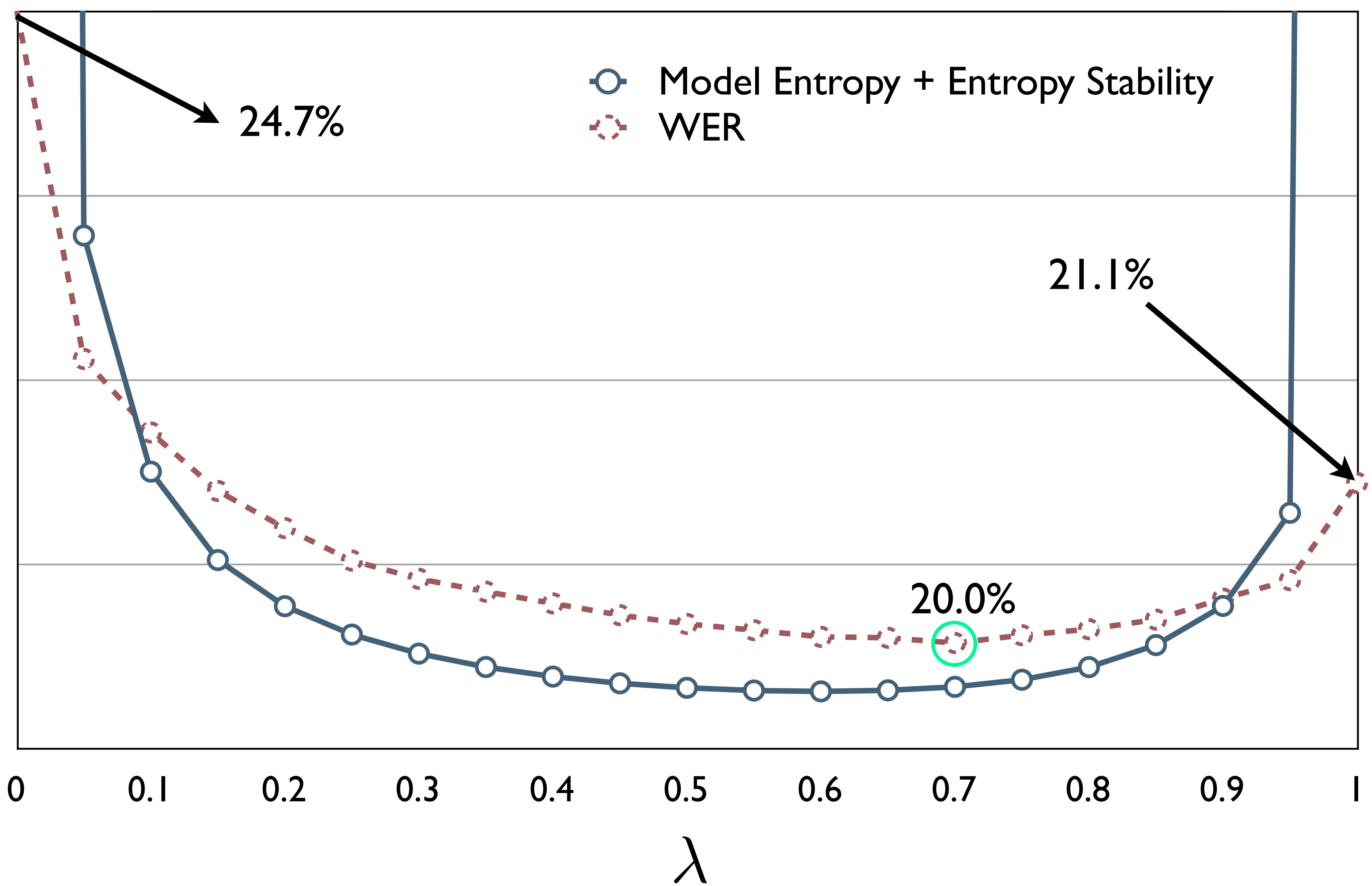
Experiments

$$P(w_i|h) = \lambda P_B(w_i|h) + (1-\lambda)P_A(w_i|h)$$



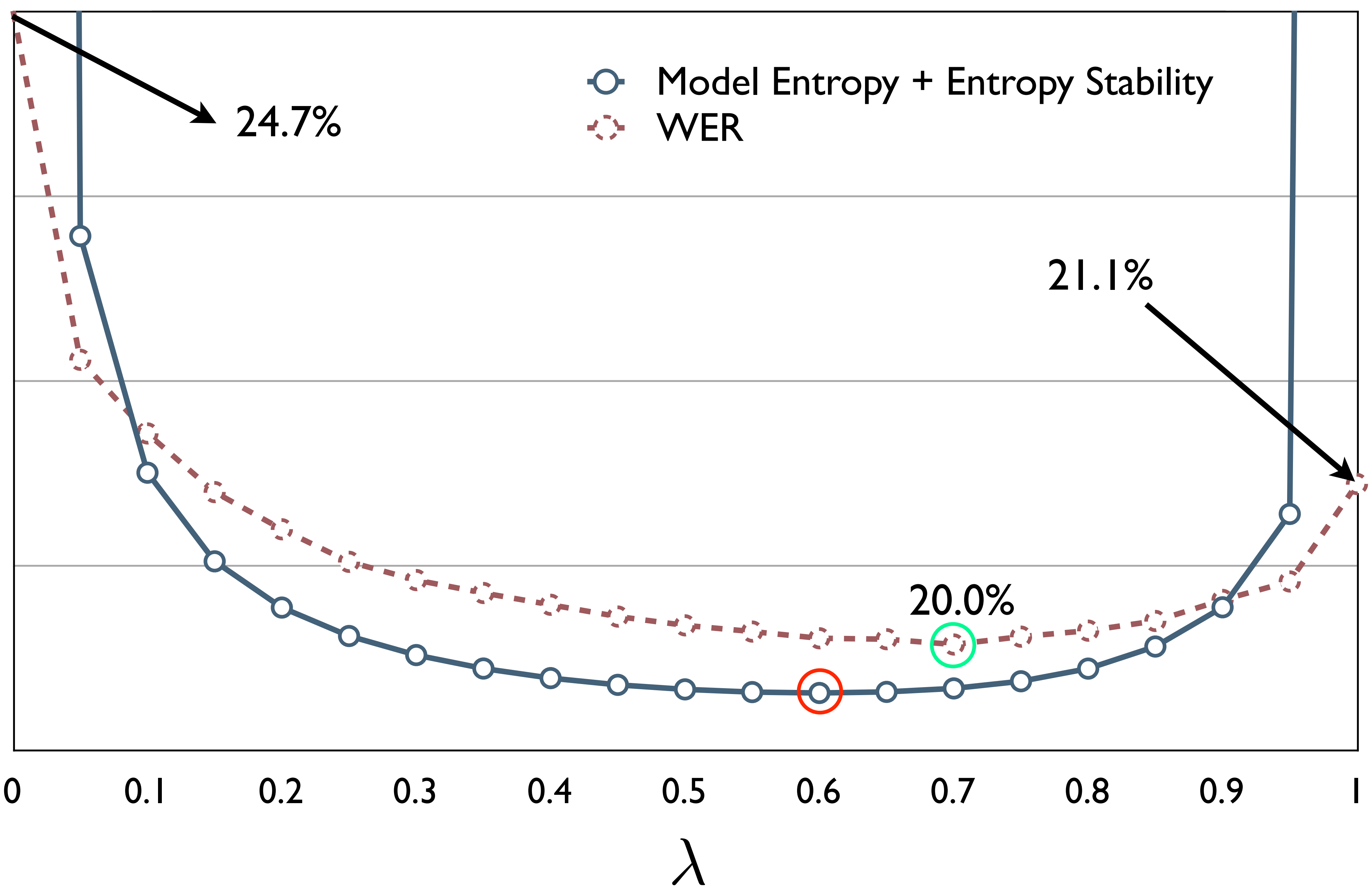
Experiments

$$P(w_i|h) = \lambda P_B(w_i|h) + (1-\lambda)P_A(w_i|h)$$



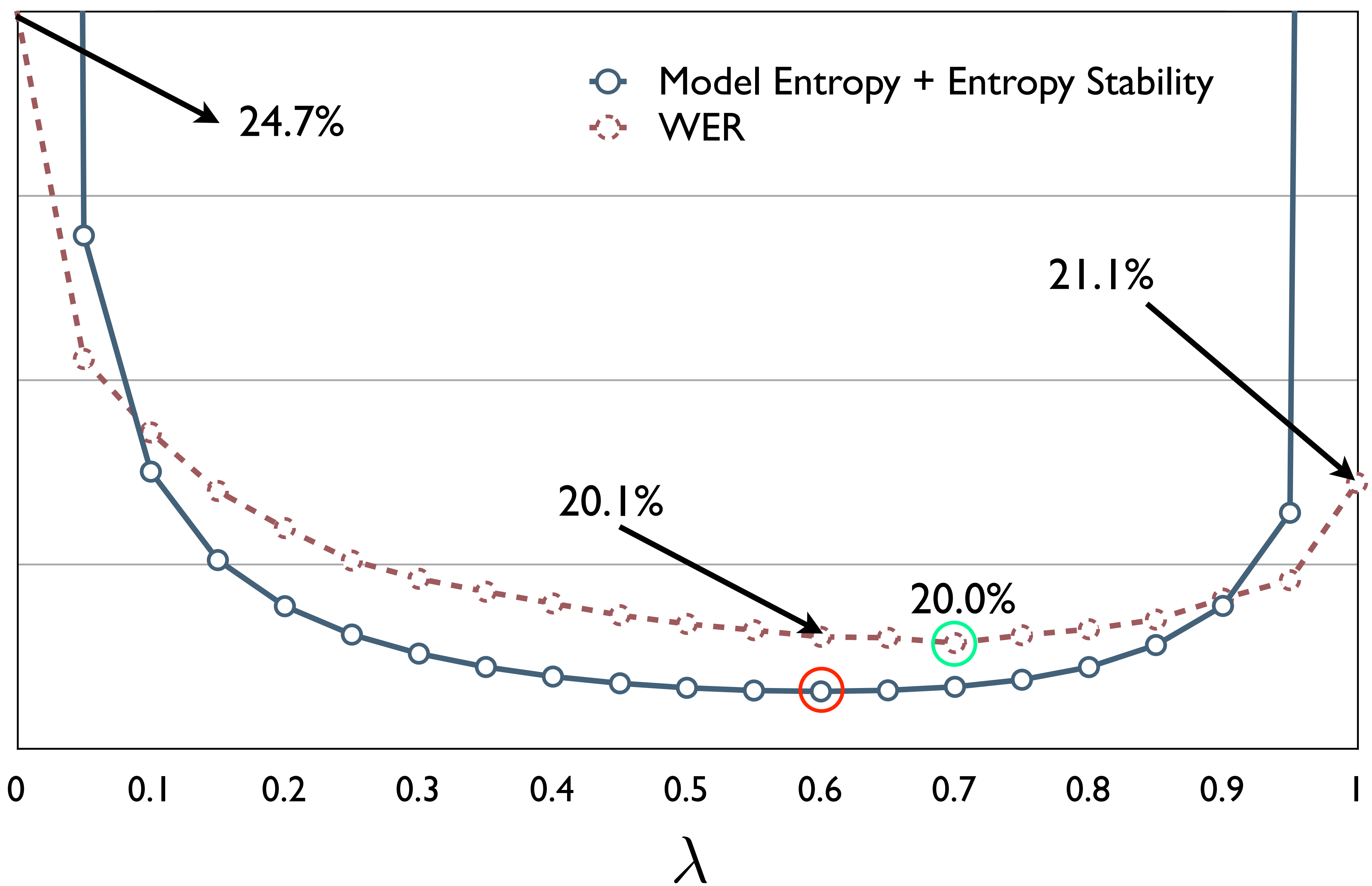
Experiments

$$P(w_i|h) = \lambda P_B(w_i|h) + (1-\lambda)P_A(w_i|h)$$



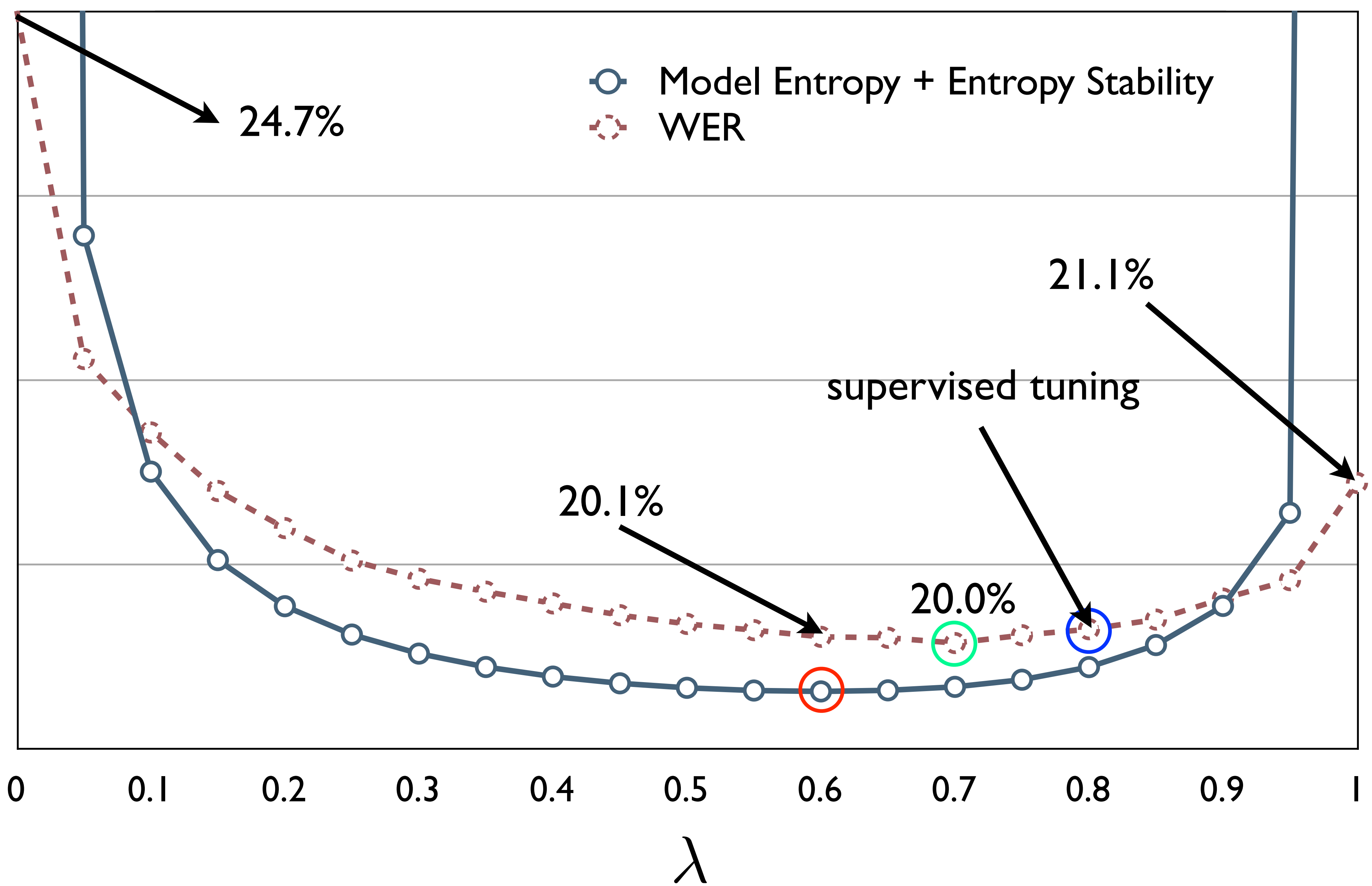
Experiments

$$P(w_i|h) = \lambda P_B(w_i|h) + (1-\lambda)P_A(w_i|h)$$

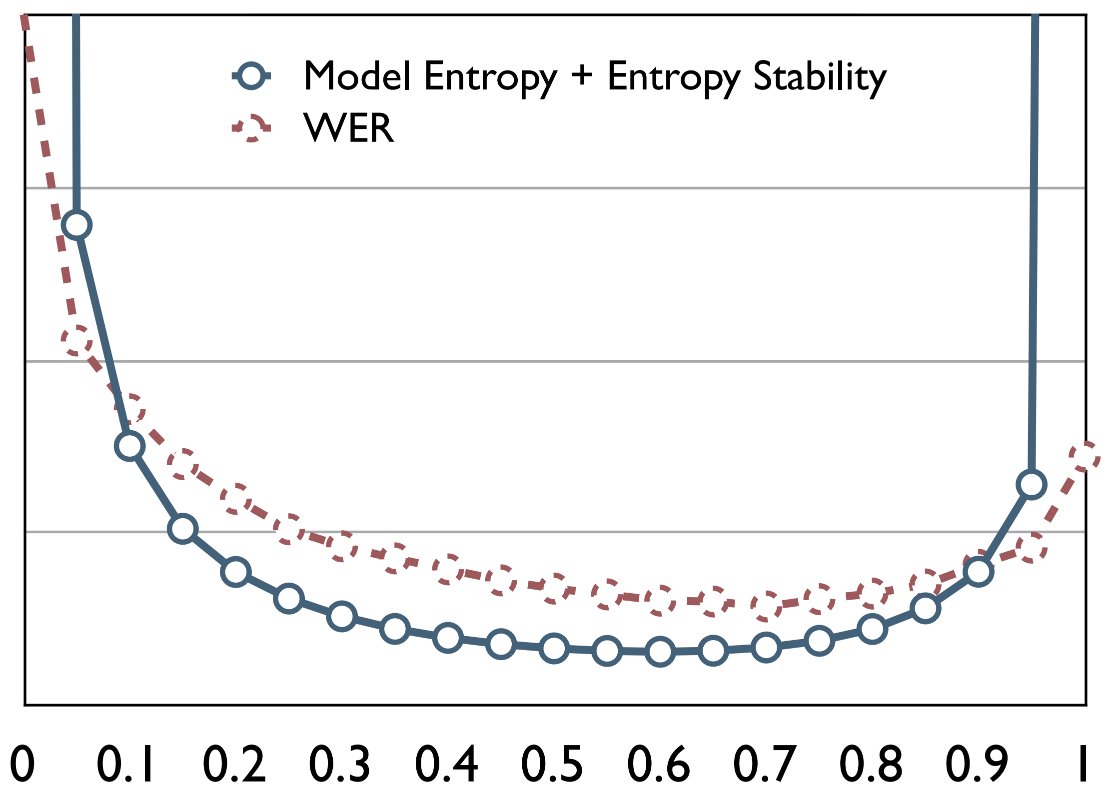


Experiments

$$P(w_i|h) = \lambda P_B(w_i|h) + (1 - \lambda) P_A(w_i|h)$$

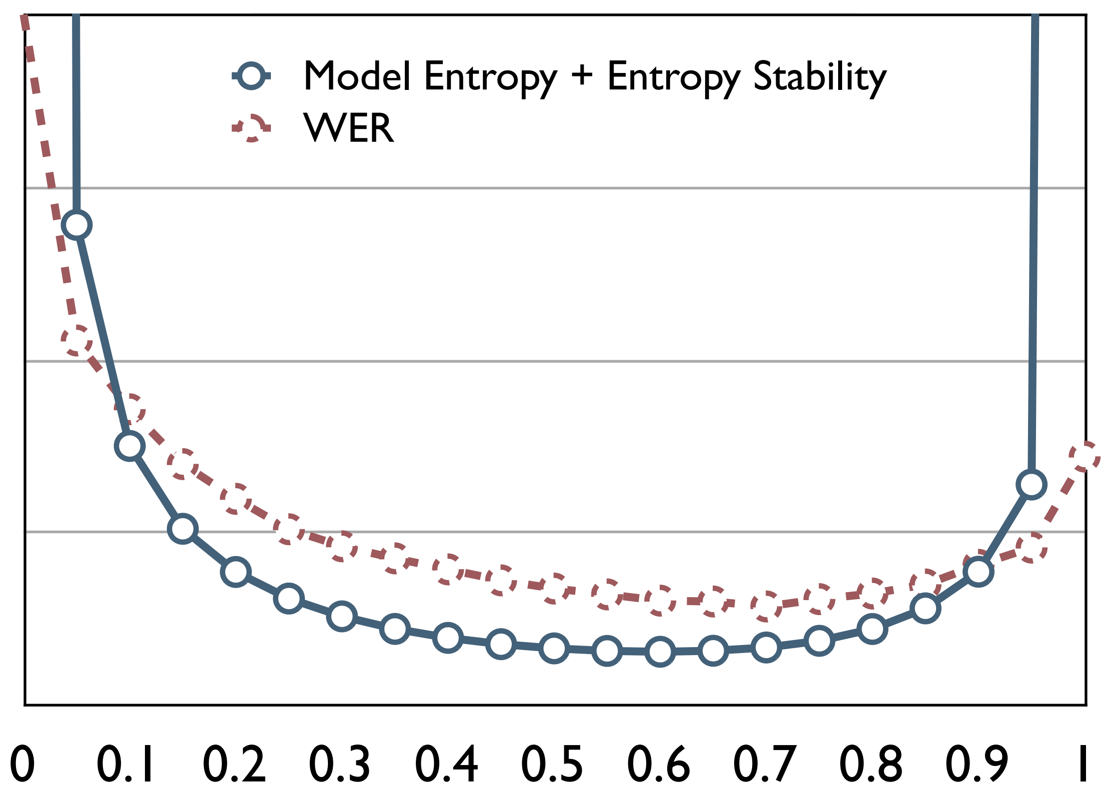


Experiments



- The proposed unsupervised framework results in the **same performance as supervised adaptation**

Experiments



- The proposed unsupervised framework results in the same performance as supervised adaptation
- The **WER trend** is almost perfectly **predicted** by the proposed objective function

Overview

- Motivation
- Conditional Entropy based Adaptation
  - Entropy Definition
  - Entropy vs. Classifier Performance
  - Problems
  - Entropy-Stability
  - Proposed Objective Function
- Speech Recognition Task
  - Entropy/Gradient of Entropy for Speech Lattices
  - Language Model Adaptation
  - Experiment / Results / Explanation
- **Future Work**

Future: Context-Dependent Weights

- The interpolation model is too **simple** and has only a **global weight**

Future: Context-Dependent Weights

- The interpolation model is too **simple** and has only a **global weight**
  - Factors such as *N*-gram modeling resolution, generalization, **topics** and **styles** can affect the contribution of sources on a **local**, **context-dependent** basis

Liu, Gales and Woodland, Interspeech 2008

Future: Context-Dependent Weights

- The interpolation model is too **simple** and has only a **global weight**
  - Factors such as *N*-gram modeling resolution, generalization, **topics** and **styles** can affect the contribution of sources on a **local**, **context-dependent** basis

Liu, Gales and Woodland, Interspeech 2008

  - As the history length grow, the **number of context-dependent weights** to be estimated **increases exponentially**

Future: Context-Dependent Weights

- The interpolation model is too **simple** and has only a **global weight**
  - Factors such as *N*-gram modeling resolution, generalization, **topics** and **styles** can affect the contribution of sources on a **local**, **context-dependent** basis

Liu, Gales and Woodland, Interspeech 2008

  - As the history length grow, the **number of context-dependent weights** to be estimated **increases exponentially**

$$p(w_i|h) = \lambda(\phi(h))p_B(w_i|h) + (1 - \lambda(\phi(h))p_A(w_i|h)$$
$$\phi : h \rightarrow \text{clusters}$$

Future: Context-Dependent Weights

- The interpolation model is too **simple** and has only a **global weight**
  - Factors such as *N*-gram modeling resolution, generalization, **topics** and **styles** can affect the contribution of sources on a **local**, **context-dependent** basis

Liu, Gales and Woodland, Interspeech 2008

  - As the history length grow, the **number of context-dependent weights** to be estimated **increases exponentially**

$$p(w_i|h) = \boxed{\lambda(\phi(h))} p_B(w_i|h) + (1 - \lambda(\phi(h)) p_A(w_i|h)$$
$$\phi : h \rightarrow \text{clusters}$$

$$\lambda_1, \lambda_2, \cdots, \lambda_{|C|}$$

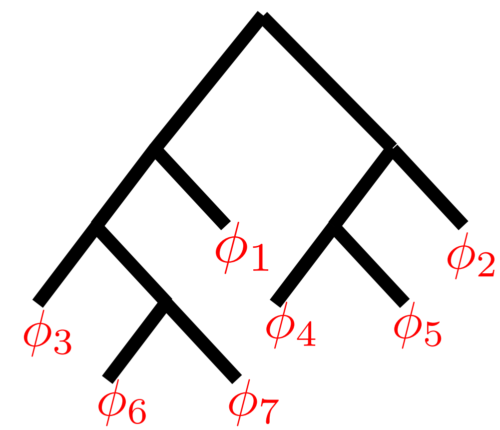
Future: Context-Dependent Weights

Proposed Framework

1. Clustering histories using Decision Tree algorithm
2. Training context-dependent weight using our unsupervised objective function

$$\min_{\bar{\lambda}} \left(H_{\bar{\lambda}}(\mathbf{W}|\mathbf{X}) + \gamma \left|\left| \frac{\partial H_{\bar{\lambda}}(\mathbf{W}|\mathbf{X})}{\partial \bar{\lambda}} \right|\right|_p \right)$$

- The procedure is **unsupervised**
- It will also take into account, **acoustic confusion**



Future: Context-Dependent Weights

Simple Clustering:

$$\phi(h) = \phi(w_{i-1}, w_{i-2}, \cdots, w_{i-N}) = \begin{cases} \phi_1 & \text{if } C(w_{i-1}) > 0 \\ \phi_2 & \text{if } C(w_{i-1}) = 0 \end{cases}$$

Future: Context-Dependent Weights

Simple Clustering:

$$\phi(h) = \phi(w_{i-1}, w_{i-2}, \cdots, w_{i-N}) = \begin{cases} \phi_1 & \text{if } C(w_{i-1}) > 0 \\ \phi_2 & \text{if } C(w_{i-1}) = 0 \end{cases}$$

Optimizing using L-BFGS method:

# Clusters	WER%
1 (Global Weight)	20.1
2 ($\lambda_{1,2}$)	**19.9**

Weight	Value
λ_1	0.63
λ_2	0.32

Future: Context-Dependent Weights

Simple Clustering:

$$\phi(h) = \phi(w_{i-1}, w_{i-2}, \cdots, w_{i-N}) = \begin{cases} \phi_1 & \text{if } C(w_{i-1}) > 50 \\ \phi_2 & \text{if } C(w_{i-1}) > 0 \\ \phi_3 & \text{if } C(w_{i-1}) = 0 \end{cases}$$

Optimizing using L-BFGS method:

# Clusters	WER%
1 (Global Weight)	20.1
2 ($\lambda_{1,2}$)	19.9
3 ($\lambda_{1,2,3}$)	**19.8**

Weight	Value
λ_1	0.68
λ_2	0.57
λ_3	0.31

Future: Semi-Supervised Learning (SSL)

- Using the proposed objective function as a **regularizer** for SSL:

$$\sum_{i=1}^{N} \log p_\theta(y_i|x_i) + \gamma \sum_{i=N+1}^{M} H_\theta(\mathbf{Y}|x_i)$$

- As an application, we are currently working on Semi-supervised CRF-based Named Entity Recognition

Future: Semi-Supervised Learning (SSL)

- Using the proposed objective function as a **regularizer** for SSL:

$$\sum_{i=1}^{N} \log p_\theta(y_i|x_i) + \gamma \sum_{i=N+1}^{M} H_\theta(\mathbf{Y}|x_i) + \lambda \left\| \frac{\partial \left(\sum_{i=N+1}^{M} H_\theta(\mathbf{Y}|x_i) \right)}{\partial \theta} \right\|_p$$

- As an application, we are currently working on Semi-supervised CRF-based Named Entity Recognition

Future: Semi-Supervised Learning (SSL)

- Using the proposed objective function as a **regularizer** for SSL:

$$\sum_{i=1}^{N} \log p_\theta(y_i|x_i) + \gamma \sum_{i=N+1}^{M} H_\theta(\mathbf{Y}|x_i) + \lambda \left\| \frac{\partial \left(\sum_{i=N+1}^{M} H_\theta(\mathbf{Y}|x_i) \right)}{\partial \theta} \right\|_p$$

- As an application, we are currently working on Semi-supervised CRF-based Named Entity Recognition

Thank You!