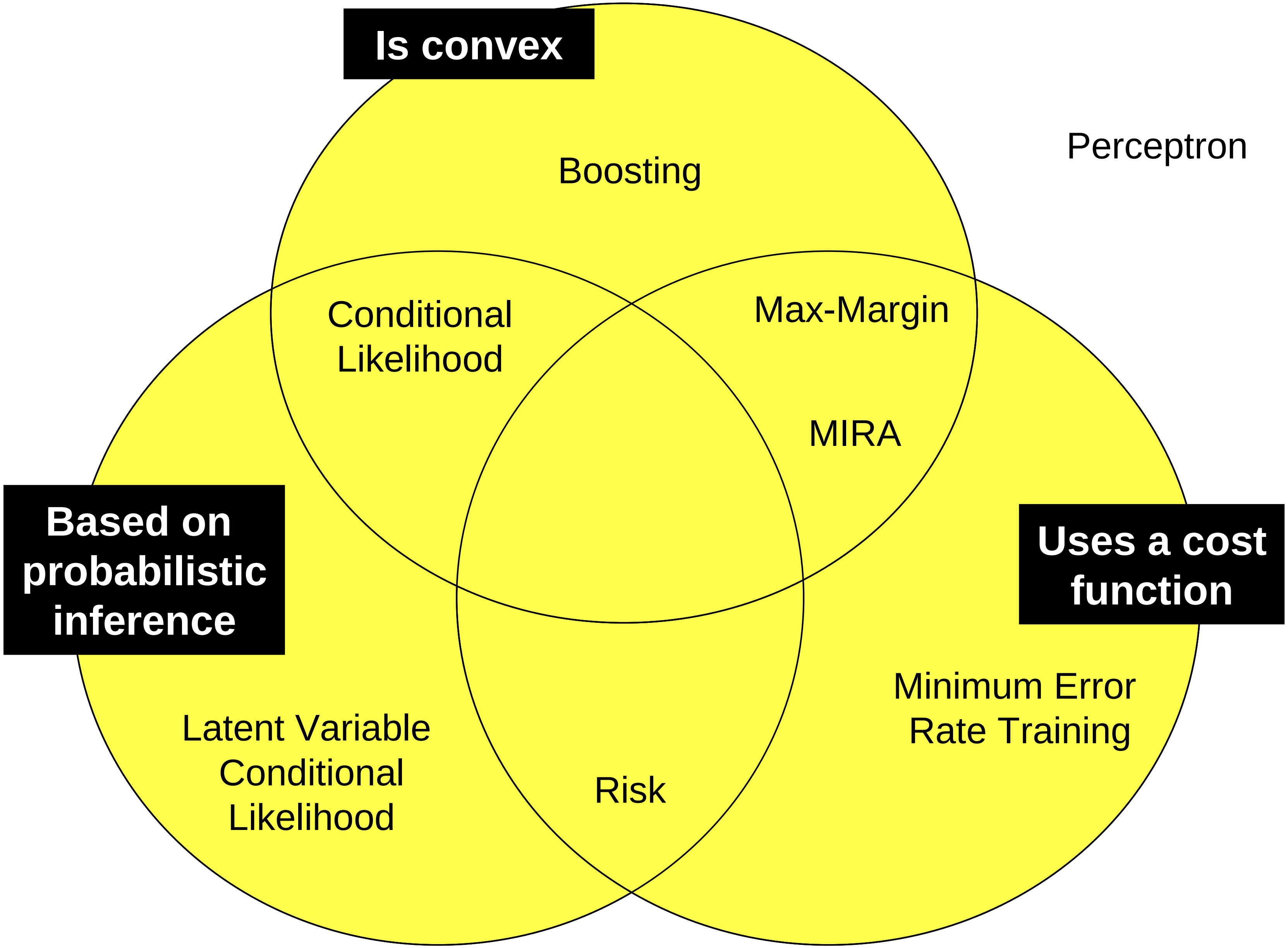
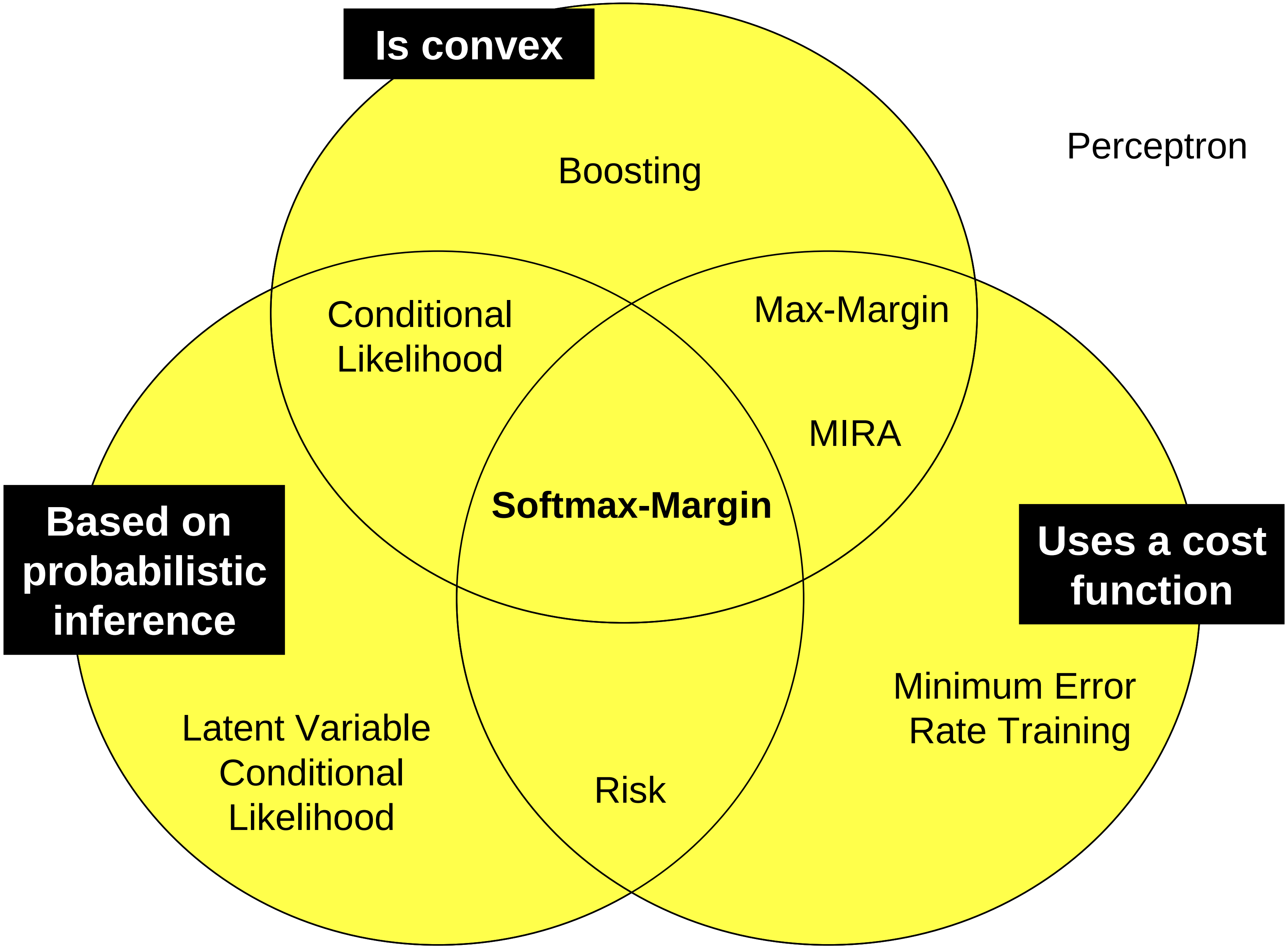
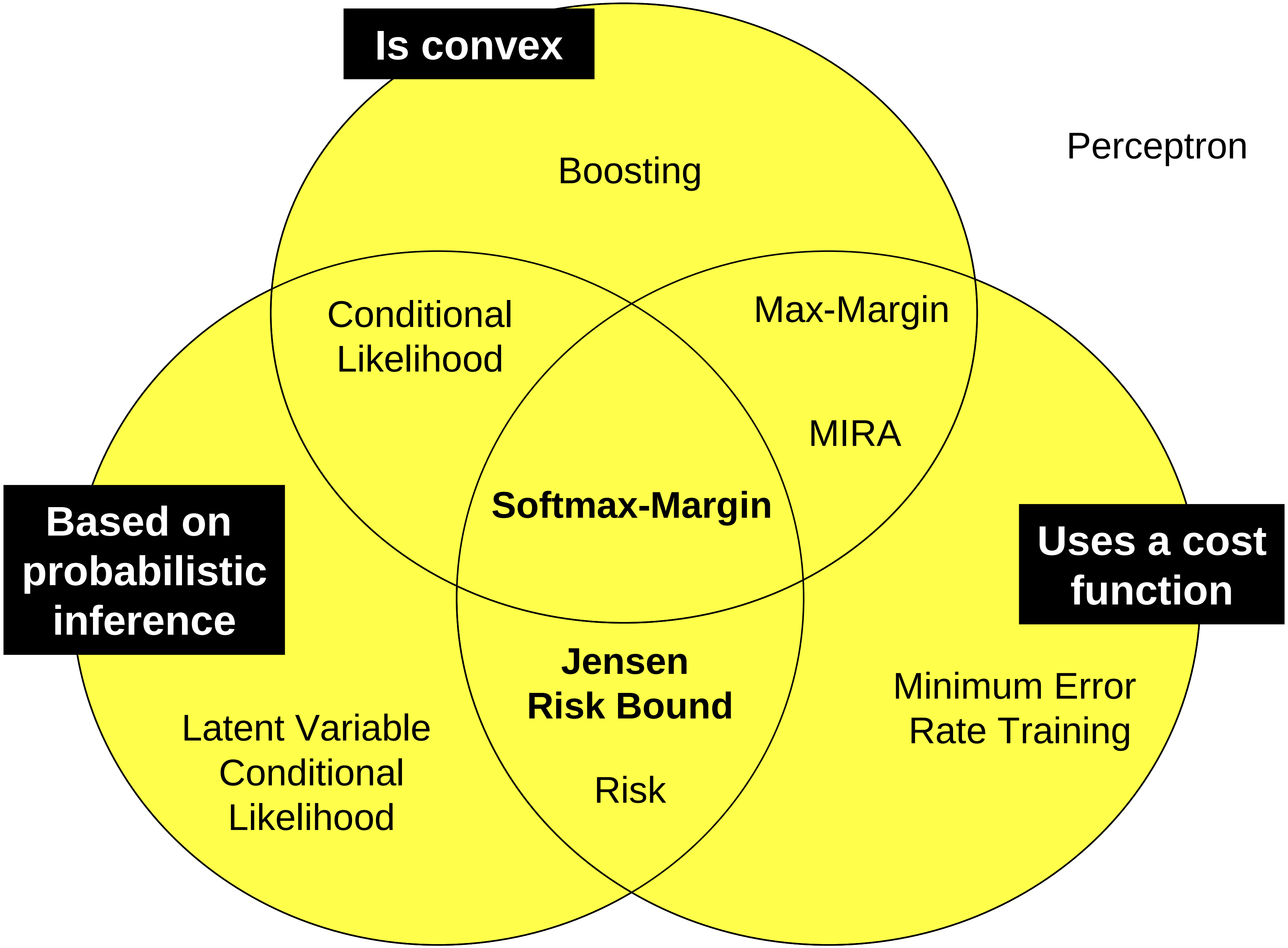
Softmax-Margin CRFs: Training Log-Linear Models with Cost Functions

Kevin Gimpel and Noah A. Smith



Carnegie Mellon


Is convex
Boosting
Perceptron
Conditional Likelihood
Max-Margin
MIRA
Softmax-Margin
Based on probabilistic inference
Uses a cost function
Minimum Error Rate Training
Latent Variable Conditional Likelihood
Risk



ARK lti Carnegie Mellon

Linear Models for Structured Prediction

input *output*

$$\hat{y} = \operatorname*{argmax}_{y \in \mathcal{Y}(x)} \boldsymbol{\theta}^\top \boldsymbol{f}(x, y)$$

weights *features*

- For probabilistic interpretation, exponentiate and normalize:

$$p_{\boldsymbol{\theta}}(y|x) = \frac{\exp\{\boldsymbol{\theta}^\top \boldsymbol{f}(x, y)\}}{\sum_{y' \in \mathcal{Y}(x)} \exp\{\boldsymbol{\theta}^\top \boldsymbol{f}(x, y')\}}$$

Training

- Standard approach is to maximize conditional likelihood:

$$\min_{\boldsymbol{\theta}} \sum_{i=1}^{n} \left(-\boldsymbol{\theta}^\top \boldsymbol{f}(x^{(i)}, y^{(i)}) \; + \; \log \sum_{y \in \mathcal{Y}(x^{(i)})} \exp\{\boldsymbol{\theta}^\top \boldsymbol{f}(x^{(i)}, y)\} \right)$$

- Another approach maximizes margin (Taskar et al., 2003):

$$\min_{\boldsymbol{\theta}} \sum_{i=1}^{n} \left(-\boldsymbol{\theta}^\top \boldsymbol{f}(x^{(i)}, y^{(i)}) \; + \max_{y \in \mathcal{Y}(x^{(i)})} \left(\boldsymbol{\theta}^\top \boldsymbol{f}(x^{(i)}, y) + \mathrm{cost}(y^{(i)}, y) \right) \right)$$

task-specific cost function

Training

- Standard approach is to maximize conditional likelihood:

$$\min_{\boldsymbol{\theta}} \sum_{i=1}^{n} \left(-\boldsymbol{\theta}^\top \boldsymbol{f}(x^{(i)}, y^{(i)}) \;+\; \log \sum_{y \in \mathcal{Y}(x^{(i)})} \exp\{\boldsymbol{\theta}^\top \boldsymbol{f}(x^{(i)}, y)\} \right)$$

- Another approach maximizes margin (Taskar et al., 2003):

$$\min_{\boldsymbol{\theta}} \sum_{i=1}^{n} \left(-\boldsymbol{\theta}^\top \boldsymbol{f}(x^{(i)}, y^{(i)}) \;+\; \underbrace{\max_{y \in \mathcal{Y}(x^{(i)})} \left(\boldsymbol{\theta}^\top \boldsymbol{f}(x^{(i)}, y) + \mathrm{cost}(y^{(i)}, y) \right)}_{\textit{cost-augmented decoding}} \right)$$

Training

- Standard approach is to maximize conditional likelihood:

$$\min_{\boldsymbol{\theta}} \sum_{i=1}^{n} \left(-\boldsymbol{\theta}^\top \boldsymbol{f}(x^{(i)}, y^{(i)}) \;+\; \log \sum_{y \in \mathcal{Y}(x^{(i)})} \exp\{\boldsymbol{\theta}^\top \boldsymbol{f}(x^{(i)}, y)\} \right)$$

- Another approach maximizes margin (Taskar et al., 2003):

$$\min_{\boldsymbol{\theta}} \sum_{i=1}^{n} \left(-\boldsymbol{\theta}^\top \boldsymbol{f}(x^{(i)}, y^{(i)}) \;+\; \max_{y \in \mathcal{Y}(x^{(i)})} \left(\boldsymbol{\theta}^\top \boldsymbol{f}(x^{(i)}, y) + \mathrm{cost}(y^{(i)}, y) \right) \right)$$

- Softmax-margin: replace “max” with “softmax”

$$\min_{\boldsymbol{\theta}} \sum_{i=1}^{n} \left(-\boldsymbol{\theta}^\top \boldsymbol{f}(x^{(i)}, y^{(i)}) + \underbrace{\log \sum_{y \in \mathcal{Y}(x^{(i)})} \exp\{\boldsymbol{\theta}^\top \boldsymbol{f}(x^{(i)}, y) + \mathrm{cost}(y^{(i)}, y)\}}_{\text{“cost-augmented summing”}} \right)$$

Training

- Standard approach is to maximize conditional likelihood:

$$\min_{\boldsymbol{\theta}} \sum_{i=1}^{n} \left(-\boldsymbol{\theta}^\top \boldsymbol{f}(x^{(i)}, y^{(i)}) \; + \; \log \sum_{y \in \mathcal{Y}(x^{(i)})} \exp\{\boldsymbol{\theta}^\top \boldsymbol{f}(x^{(i)}, y)\} \right)$$

- Another approach maximizes margin (Taskar et al., 2003):

$$\min_{\boldsymbol{\theta}} \sum_{i=1}^{n} \left(-\boldsymbol{\theta}^\top \boldsymbol{f}(x^{(i)}, y^{(i)}) \; + \max_{y \in \mathcal{Y}(x^{(i)})} \left(\boldsymbol{\theta}^\top \boldsymbol{f}(x^{(i)}, y) + \mathrm{cost}(y^{(i)}, y) \right) \right)$$

- Softmax-margin: replace “max” with “softmax”

$$\min_{\boldsymbol{\theta}} \sum_{i=1}^{n} \left(-\boldsymbol{\theta}^\top \boldsymbol{f}(x^{(i)}, y^{(i)}) + \log \sum_{y \in \mathcal{Y}(x^{(i)})} \exp\{\boldsymbol{\theta}^\top \boldsymbol{f}(x^{(i)}, y) + \mathrm{cost}(y^{(i)}, y)\} \right)$$

Sha and Saul (2006), Povey et al. (2008)

Properties of Softmax-Margin

- Has a probabilistic interpretation in the minimum divergence framework (Jelinek, 1997)
  - Details in technical report

- Is a bound on:
  - Max-margin
  - Conditional likelihood
  - Risk

Properties of Softmax-Margin

- Has a probabilistic interpretation in the minimum divergence framework (Jelinek, 1997)
  - Details in technical report
- Is a bound on:
  - Max-margin (because “softmax” bounds “max”) ✓
  - Conditional likelihood
  - Risk

Risk?

- **Risk** is the expected value of the cost function (Smith and Eisner, 2006; Li and Eisner, 2009):

$$\min_{\boldsymbol{\theta}} \sum_{i=1}^{n} \mathbb{E}_{p_{\boldsymbol{\theta}}(\cdot|x^{(i)})}[\mathrm{cost}(y^{(i)}, \cdot)]$$

ARK lti Carnegie Mellon

Bounding Conditional Likelihood and Risk

- Softmax-margin:

$$\sum_{i=1}^{n} \left(-\boldsymbol{\theta}^\top \boldsymbol{f}(x^{(i)}, y^{(i)}) + \log \sum_{y \in \mathcal{Y}(x^{(i)})} \exp\{\boldsymbol{\theta}^\top \boldsymbol{f}(x^{(i)}, y) + \text{cost}(y^{(i)}, y)\} \right)$$

$$= \underbrace{\sum_{i=1}^{n} \left(-\boldsymbol{\theta}^\top \boldsymbol{f}(x^{(i)}, y^{(i)}) + \log Z_i \right)}_{\textbf{Conditional likelihood}} + \underbrace{\sum_{i=1}^{n} \log \mathbb{E}_{p_i}[\exp\{\text{cost}(y^{(i)}, \cdot)\}]}_{\textbf{Bound on risk via Jensen's inequality}}$$

Bounding Conditional Likelihood and Risk

- Softmax-margin:

$$\sum_{i=1}^{n}\left(-\boldsymbol{\theta}^{\top}\boldsymbol{f}(x^{(i)},y^{(i)})+\log\sum_{y\in\mathcal{Y}(x^{(i)})}\exp\{\boldsymbol{\theta}^{\top}\boldsymbol{f}(x^{(i)},y)+\text{cost}(y^{(i)},y)\}\right)$$

$$=\underbrace{\sum_{i=1}^{n}\left(-\boldsymbol{\theta}^{\top}\boldsymbol{f}(x^{(i)},y^{(i)})+\log Z_i\right)}_{\textbf{Conditional likelihood}}+\underbrace{\sum_{i=1}^{n}\log\mathbb{E}_{p_i}[\exp\{\text{cost}(y^{(i)},\cdot)\}]}_{\textbf{Bound on risk via Jensen's inequality}}$$

Softmax-margin is a convex bound on max-margin, conditional likelihood, and risk

Bounding Conditional Likelihood and Risk

- Softmax-margin:

$$\sum_{i=1}^{n} \left(-\boldsymbol{\theta}^\top \boldsymbol{f}(x^{(i)}, y^{(i)}) + \log \sum_{y \in \mathcal{Y}(x^{(i)})} \exp\{\boldsymbol{\theta}^\top \boldsymbol{f}(x^{(i)}, y) + \mathrm{cost}(y^{(i)}, y)\} \right)$$

$$= \underbrace{\sum_{i=1}^{n} \left(-\boldsymbol{\theta}^\top \boldsymbol{f}(x^{(i)}, y^{(i)}) + \log Z_i \right)}_{\text{Conditional likelihood}} + \underbrace{\sum_{i=1}^{n} \log \mathbb{E}_{p_i}[\exp\{\mathrm{cost}(y^{(i)}, \cdot)\}]}_{\text{Jensen Risk Bound}}$$

Jensen Risk Bound

Easier to optimize than risk

(cf. Li and Eisner, 2009)

Implementation

- Conditional likelihood → Softmax-margin
  - If cost function factors the same way as the features, it's easy:
    - Add additional features for the cost function
    - Keep their weights fixed
  - If not, use a simpler cost function or use approximate inference

Experiments

- English named-entity recognition (CoNLL 2003)

- Compared softmax-margin and Jensen risk bound with five baselines:
  - Perceptron (Collins, 2002)
  - 1-best MIRA with cost-augmented decoding (Crammer et al., 2006)
  - Max-margin via subgradient descent (Ratliff et al., 2006)
  - Conditional likelihood (Lafferty et al., 2001)
  - Risk (Xiong et al., 2009)

- For risk and Jensen risk bound, initialized using output of conditional likelihood training

- Used Hamming cost for cost function

Results

Method	Test F_1
Perceptron	83.98*
MIRA	85.72
Max-Margin	85.28*
Conditional Likelihood	85.46*
Risk	85.59
Jensen Risk Bound	85.65
Softmax-Margin	85.84

*** Indicates significance (compared with softmax-margin)**

Results

Significant improvement with equal training time and implementation difficulty

Method	Test F_1
Perceptron	83.98*
MIRA	85.72
Max-Margin	85.28*
Conditional Likelihood	85.46*
Risk	85.59
Jensen Risk Bound	85.65
Softmax-Margin	85.84

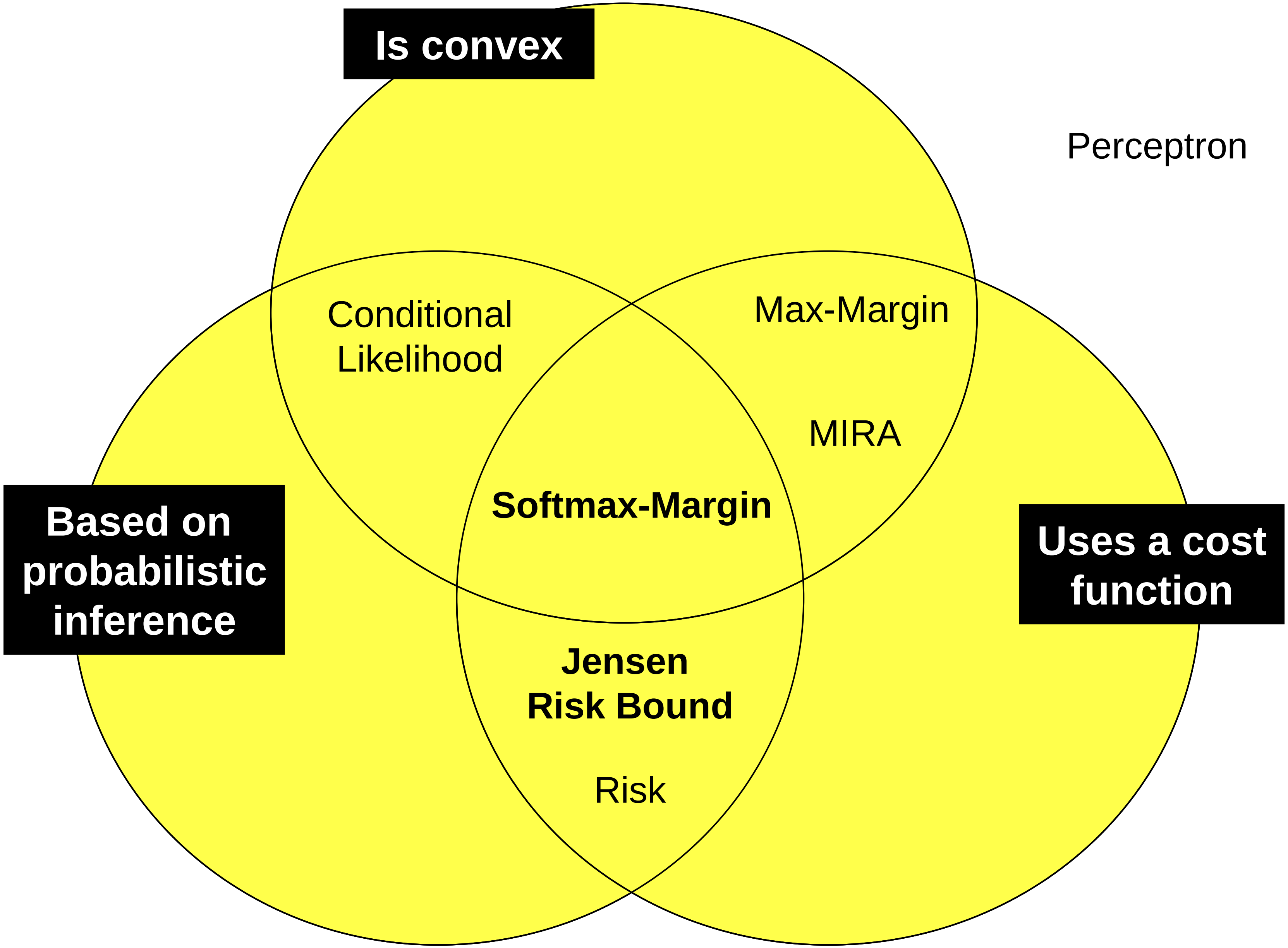
*** Indicates significance (compared with softmax-margin)**

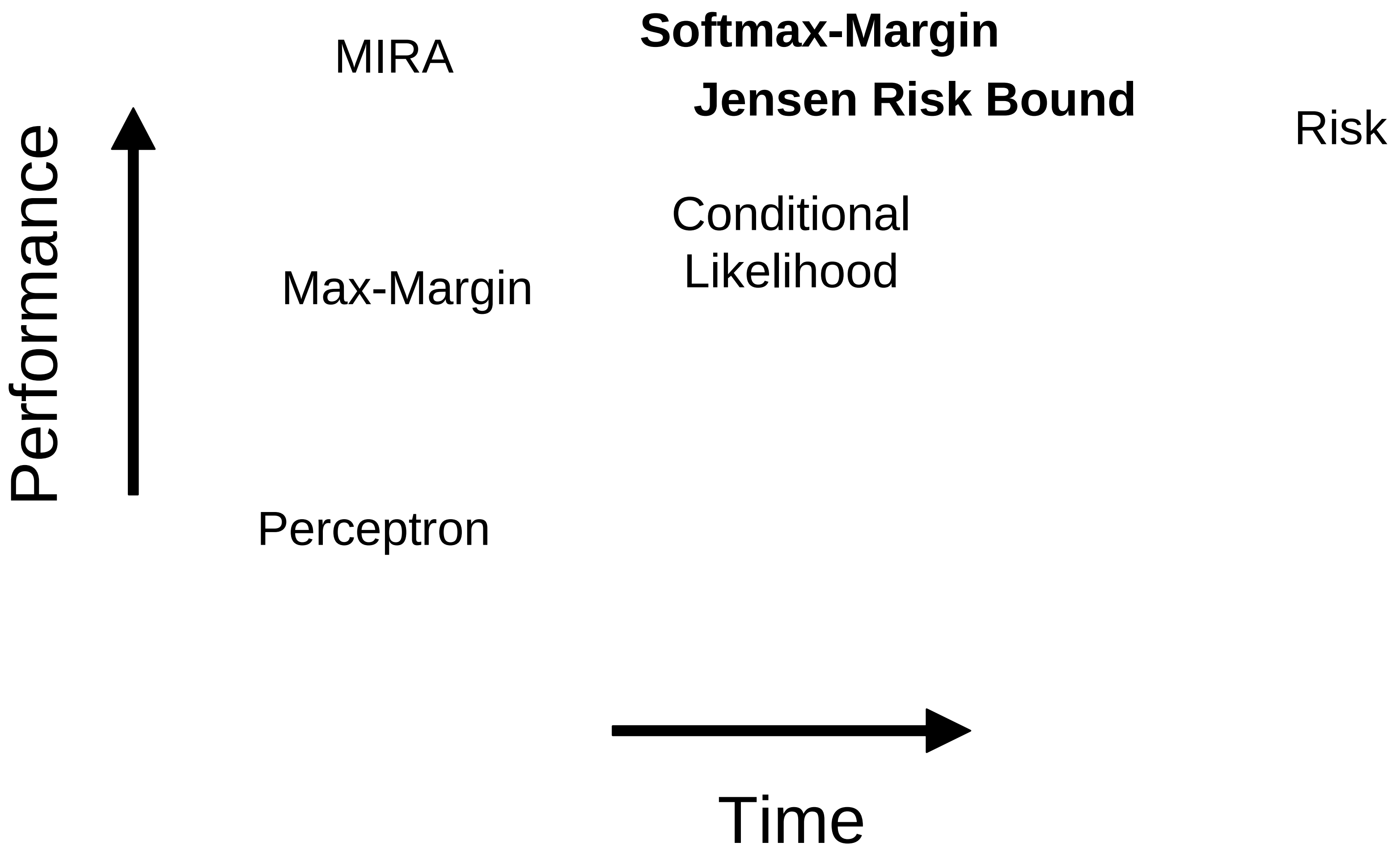
Results

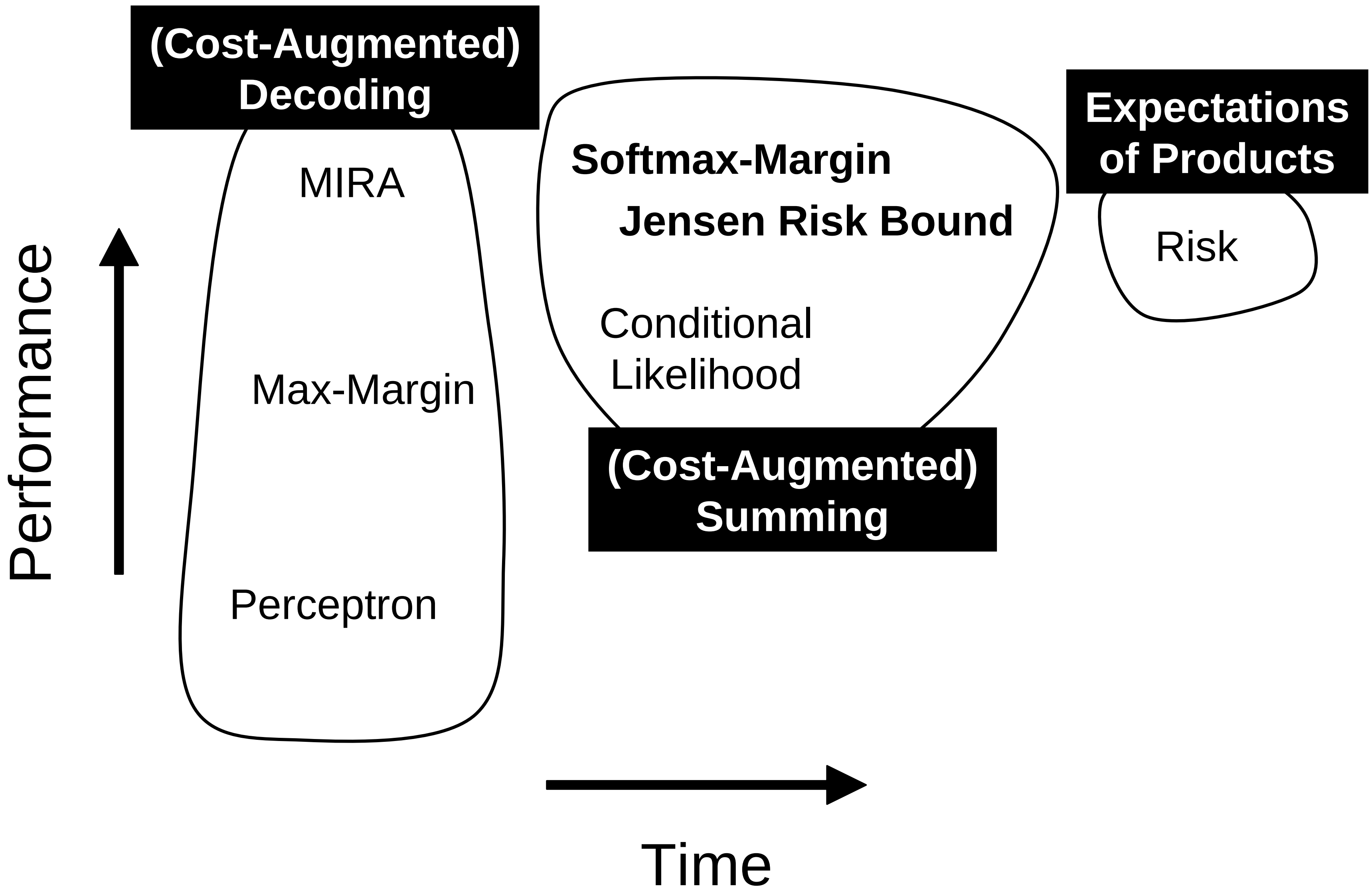
Method	Test F_1
Perceptron	83.98*
MIRA	85.72
Max-Margin	85.28*
Conditional Likelihood	85.46*
Risk	85.59
Jensen Risk Bound	85.65
Softmax-Margin	85.84

Comparable performance with half the training time

*** Indicates significance (compared with softmax-margin)**


Is convex
Perceptron
Conditional Likelihood
Max-Margin
MIRA
Softmax-Margin
Based on probabilistic inference
Uses a cost function
Jensen Risk Bound
Risk


Softmax-Margin
MIRA
Jensen Risk Bound
Risk
Performance
Conditional Likelihood
Max-Margin
Perceptron
Time

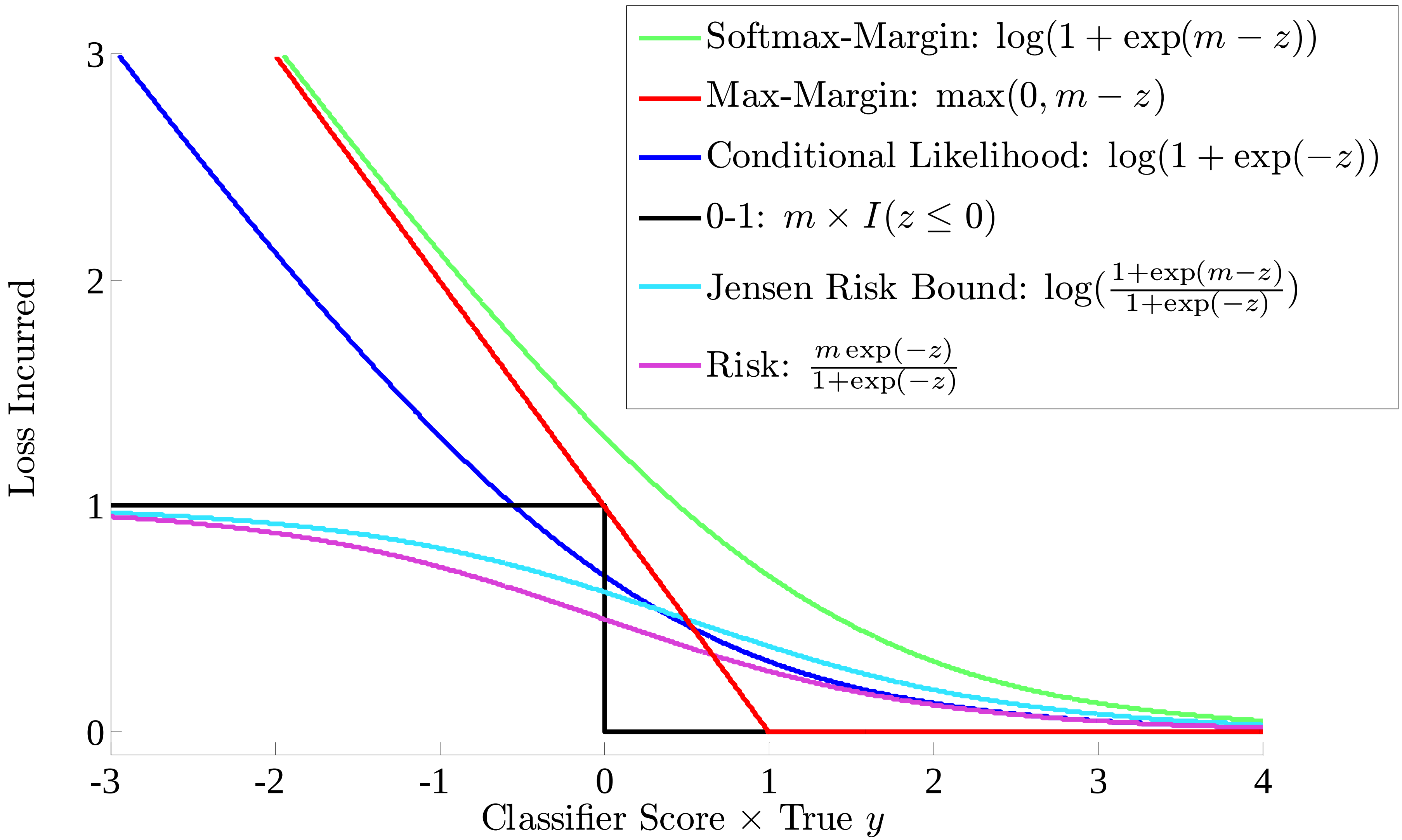


Carnegie Mellon

Thank you!

- See extended technical report for:
  - Probabilistic interpretation for softmax-margin in minimum divergence framework (Jelinek, 1997)
  - Softmax-margin training with hidden variables
  - Additional experiments

Loss Functions for Binary Classification



Training Method	Requirements	Cost Function	Convex	Prob. Interp.
Perceptron	decoding		N/A	
MIRA	cost-augmented decoding	✓	✓	
Max-Margin	cost-augmented decoding	✓	✓	
Conditional Likelihood	summing		✓	✓
Risk	expectations of products	✓		✓
Jensen Risk Bound	cost-augmented summing	✓		✓
Softmax-Margin	cost-augmented summing	✓	✓	✓