Noisy Objective Functions based on the f-Divergence

Markus Nussbaum-Thom[1,2], Ralf Schlüter[2], Vaibhava Goel[1], Hermann Ney[2]

[1]IBM Thomas J. Watson Research Center
Yorktown Heights, NY 10598, USA.
[2]Computer Science Dept. 6, RWTH Aachen University, Aachen, Germany

March 8, 2017

Contents

Goals

- Derive training criteria from bound on the error difference.

Goals

- Derive training criteria from bound on the error difference.
- Family of f-divergence based training criteria:

Goals

- Derive training criteria from bound on the error difference.
- Family of f-divergence based training criteria:
  - Conjugate Power Approximation: $\frac{1}{\alpha(1-\alpha)}(1-q)^{\alpha}$

Goals

- Derive training criteria from bound on the error difference.
- Family of f-divergence based training criteria:
    - Conjugate Power Approximation: $\frac{1}{\alpha(1-\alpha)}(1-q)^{\alpha}$
- How to choose α?

Goals

- Derive training criteria from bound on the error difference.
- Family of f-divergence based training criteria:
  - Conjugate Power Approximation: $\frac{1}{\alpha(1-\alpha)}(1-q)^{\alpha}$
- How to choose α?
  - Iteratively minimize over bounds/criteria.

Goals

- Derive training criteria from bound on the error difference.
- Family of f-divergence based training criteria:
  - Conjugate Power Approximation: $\frac{1}{\alpha(1-\alpha)}\left(1-q\right)^{\alpha}$
- How to choose α?
  - Iteratively minimize over bounds/criteria.
  - Randomly choose bound/criteria.

Statistical Classification Problem

- *Bayes'* decision rule:

$$c_{pr}(x) = \underset{c \in \mathcal{C}}{\operatorname{argmax}} \left\{ \underbrace{pr(c|x)}_{\text{true}} \right\} \tag{1.1}$$

with observations $x \in \mathcal{X}$ and classes $c \in \mathcal{C}$.

Statistical Classification Problem

- *Bayes'* decision rule:

$$c_{pr}(x) = \operatorname*{argmax}_{c \in \mathcal{C}} \Big\{ \underbrace{pr(c|x)}_{\text{true}} \Big\} \tag{1.1}$$

  with observations $x \in \mathcal{X}$ and classes $c \in \mathcal{C}$.

- Model-based decision rule:

$$c_{q}(x) = \operatorname*{argmax}_{c \in \mathcal{C}} \Big\{ \underbrace{q(c|x)}_{\text{model}} \Big\} \tag{1.2}$$

Statistical Classification Problem

- *Bayes'* decision rule:

$$c_{pr}(x) = \operatorname*{argmax}_{c \in \mathcal{C}} \Big\{ \underbrace{pr(c|x)}_{\text{true}} \Big\} \tag{1.1}$$

with observations $x \in \mathcal{X}$ and classes $c \in \mathcal{C}$.

- Model-based decision rule:

$$c_q(x) = \operatorname*{argmax}_{c \in \mathcal{C}} \Big\{ \underbrace{q(c|x)}_{\text{model}} \Big\} \tag{1.2}$$

- Error difference:

$$\Delta(x) = \underbrace{1 - pr(c_{pr}(x)|x)}_{\textit{Bayes}\text{ error}} - \underbrace{(1 - pr(c_q(x)|x))}_{\text{model error}} \qquad \text{"local"}$$

$$\Delta = \int pr(x)\Delta(x)\,\mathrm{d}x \qquad \text{"global"}$$

Relation of the Error and Training Criterion

- What is the relation between the *Bayes* error,

Relation of the Error and Training Criterion

- What is the relation between the *Bayes* error, the model-based error,

Relation of the Error and Training Criterion

- What is the relation between the *Bayes* error, the model-based error, and the training criterion ?

Relation of the Error and Training Criterion

- What is the relation between the *Bayes* error, the model-based error, and the training criterion ?
- *Kullback-Leibler* divergence:

Relation of the Error and Training Criterion

- What is the relation between the *Bayes* error, the model-based error, and the training criterion ?
- *Kullback-Leibler* divergence:

$$\underbrace{\Delta^2}_{\text{error difference}} \leq 2 \int pr(x) \sum_{c \in \mathcal{C}} pr(c|x) \log \left(\frac{pr(c|x)}{q(c|x)} \right) \, \mathrm{d}x \tag{1.3}$$

Relation of the Error and Training Criterion

- What is the relation between the *Bayes* error, the model-based error, and the training criterion ?
- *Kullback-Leibler* divergence:

$$\underbrace{\Delta^2}_{\text{error difference}} \leq 2 \int pr(x) \sum_{c \in \mathcal{C}} pr(c|x) \log \left(\frac{pr(c|x)}{q(c|x)} \right) \, \mathrm{d}x \tag{1.3}$$

- Cross entropy criterion for samples $(x_n, c_n), n = 1, \ldots, N$:

Relation of the Error and Training Criterion

- What is the relation between the *Bayes* error, the model-based error, and the training criterion ?
- *Kullback-Leibler* divergence:

$$\underbrace{\Delta^2}_{\text{error difference}} \leq 2 \int pr(x) \sum_{c \in \mathcal{C}} pr(c|x) \log \left(\frac{pr(c|x)}{q(c|x)} \right) \mathrm{d}x \tag{1.3}$$

- Cross entropy criterion for samples $(x_n, c_n), n = 1, \ldots, N$:

$$\rightsquigarrow F_{CE}(q) = -\frac{1}{N} \sum_{n=1}^{N} \log q(c_n|x_n) \tag{1.4}$$

Relation of the Error and Training Criterion

- What is the relation between the *Bayes* error, the model-based error, and the training criterion ?
- *Kullback-Leibler* divergence:

$$\underbrace{\Delta^2}_{\text{error difference}} \leq 2 \int pr(x) \sum_{c \in \mathcal{C}} pr(c|x) \log \left(\frac{pr(c|x)}{q(c|x)} \right) \, \mathrm{d}x \tag{1.3}$$

- Cross entropy criterion for samples $(x_n, c_n), n = 1, \ldots, N$:

$$\rightsquigarrow F_{CE}(q) = -\frac{1}{N} \sum_{n=1}^{N} \log q(c_n|x_n) \tag{1.4}$$

- Non-parametric solution:

$$q(c|x) \rightsquigarrow pr(c|x) \tag{1.5}$$

Error Bounds based on the f-Divergence

- If $f : \mathbb{R}^+ \rightarrow \mathbb{R}$ is a convex function and $f(1) = 0$ then the *f-Divergence* is defined by:

$$D_f^x(pr||q) := \sum_{c \in \mathcal{C}} q(c|x) f\left(\frac{pr(c|x)}{q(c|x)}\right).$$

- Implicit error bounds based on the *f-Divergence* [2013]:

$$2D_f^x(pr|q) \geq f(1 + \Delta(x)) + f(1 - \Delta(x)).$$

Global Classification Error Bound

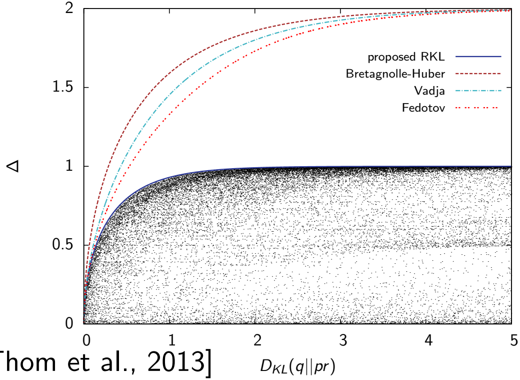
- Global error bound:

$$\Delta \leq \sqrt{1 - \exp(-2D_{\mathrm{KL}}(q||pr))}.$$

"Reversed KL $f(u) = -\log(u)$"

$$\Delta \leq 2\sqrt{1 - \exp(-D_{\mathrm{KL}}(pr||p))}.$$

"Bretagnolle-Huber"



[Nußbaum-Thom et al., 2013]

Explicit Error Bound

- Assume $f(1) = 0$, $f'''(u), u \in [1, 2]$ exists monotonically increasing:

$$\Delta^2(x) \le \frac{1}{f''(1)} D_f^x(pr||q) \tag{2.6}$$

(2.7)

(2.8)

Explicit Error Bound

- Assume $f(1) = 0$, $f'''(u), u \in [1, 2]$ exists monotonically increasing:

$$\Delta^2(x) \leq \frac{1}{f''(1)} D_f^x(pr||q) \tag{2.6}$$

- Conjugate: $f(u) = ug(1/u), u = pr(c|x)/q(c|x)$, g monotonically increasing.

(2.7)

(2.8)

Explicit Error Bound

- Assume $f(1) = 0$, $f'''(u), u \in [1, 2]$ exists monotonically increasing:

$$\Delta^2(x) \leq \frac{1}{f''(1)} D_f^x(pr||q) \tag{2.6}$$

- Conjugate: $f(u) = ug(1/u), u = pr(c|x)/q(c|x)$, g monotonically increasing.
- Explicit error bound:

(2.7)

(2.8)

Explicit Error Bound

- Assume $f(1) = 0$, $f'''(u), u \in [1, 2]$ exists monotonically increasing:

$$\Delta^2(x) \leq \frac{1}{f''(1)} D_f^x(pr||q) \tag{2.6}$$

- Conjugate: $f(u) = ug(1/u), u = pr(c|x)/q(c|x)$, g monotonically increasing.
- Explicit error bound:

$$\Delta^2 \leq \frac{1}{f''(1)} \int pr(x) D_f^x(pr||q)\, \mathrm{d}x \tag{2.7}$$

$$\tag{2.8}$$

Explicit Error Bound

- Assume $f(1) = 0$, $f'''(u), u \in [1, 2]$ exists monotonically increasing:

$$\Delta^2(x) \leq \frac{1}{f''(1)} D_f^x(pr||q) \tag{2.6}$$

- Conjugate: $f(u) = ug(1/u), u = pr(c|x)/q(c|x)$, g monotonically increasing.
- Explicit error bound:

$$\Delta^2 \leq \frac{1}{f''(1)} \int pr(x) D_f^x(pr||q) \, \mathrm{d}x \tag{2.7}$$

$$\leq \frac{1}{f''(1)} \int \sum_{c \in \mathcal{C}} pr(x, c) g(q(c|x)) \, \mathrm{d}x \tag{2.8}$$

From Classification Error Bounds to Training Criteria

- Empirical/True distribution for samples $(x_n, c_n), n = 1, \ldots, N$:

$$pr(x, c) = \frac{1}{N} \sum_{n=1}^{N} \underbrace{\delta(x - x_n)}_{\text{Dirac}} \underbrace{\delta(c, c_n)}_{\text{Kronecker}}$$

- Training criterion:

$$F_f(q)$$

From Classification Error Bounds to Training Criteria

- Empirical/True distribution for samples $(x_n, c_n), n = 1, \ldots, N$:

$$pr(x, c) = \frac{1}{N} \sum_{n=1}^{N} \underbrace{\delta(x - x_n)}_{\text{Dirac}} \underbrace{\delta(c, c_n)}_{\text{Kronecker}}$$

- Training criterion:

$$F_f(q) = \frac{1}{f''(1)} \int \sum_{c \in \mathcal{C}} pr(x, c) g(q(c|x)) \, \mathrm{d}x \quad \text{"emp. } pr(x, c)\text{"}$$

From Classification Error Bounds to Training Criteria

- Empirical/True distribution for samples $(x_n, c_n), n = 1, \ldots, N$:

$$pr(x, c) = \frac{1}{N} \sum_{n=1}^{N} \underbrace{\delta(x - x_n)}_{\text{Dirac}} \underbrace{\delta(c, c_n)}_{\text{Kronecker}}$$

- Training criterion:

$$F_f(q) = \frac{1}{f''(1)} \int \sum_{c \in \mathcal{C}} pr(x, c) g(q(c|x)) \, \mathrm{d}x \quad \text{"emp. } pr(x, c)\text{"}$$

$$= \frac{1}{f''(1)} \int \sum_{c \in \mathcal{C}} \frac{1}{N} \sum_{n=1}^{N} \delta(x - x_n) \delta(c, c_n) g(q(c|x)) \, \mathrm{d}x$$

From Classification Error Bounds to Training Criteria

- Empirical/True distribution for samples $(x_n, c_n), n = 1, \ldots, N$:

$$pr(x, c) = \frac{1}{N} \sum_{n=1}^{N} \underbrace{\delta(x - x_n)}_{\text{Dirac}} \underbrace{\delta(c, c_n)}_{\text{Kronecker}}$$

- Training criterion:

$$F_f(q) = \frac{1}{f''(1)} \int \sum_{c \in \mathcal{C}} pr(x, c) g(q(c|x)) \, \mathrm{d}x \quad \text{"emp. } pr(x, c)\text{"}$$

$$= \frac{1}{f''(1)} \int \sum_{c \in \mathcal{C}} \frac{1}{N} \sum_{n=1}^{N} \delta(x - x_n) \delta(c, c_n) g(q(c|x)) \, \mathrm{d}x$$

$$= \frac{1}{f''(1)} \frac{1}{N} \sum_{n=1}^{N} g(q(c_n|x_n))$$

$$\text{"} \int h(x) \delta(x - x_n) \, \mathrm{d}x = h(x_n) \text{"}$$

Conjugate Power Approximation Criterion

- Logarithm:

$$-\log(u) = \lim_{\alpha \to 0} \frac{1}{\alpha} (1-u)^{\alpha} \,. \tag{2.9}$$

Conjugate Power Approximation Criterion

- Logarithm:

$$-\log(u) = \lim_{\alpha \to 0} \frac{1}{\alpha} (1 - u)^{\alpha} \,. \tag{2.9}$$

- The power-approximation *f-Divergence* is defined by:

$$f(u) = u \underbrace{\frac{\left(1 - \frac{1}{u^{\alpha}}\right)}{\alpha}}_{g\left(\frac{1}{u}\right)} . \tag{2.10}$$

Conjugate Power Approximation Criterion

- Logarithm:

$$-\log(u) = \lim_{\alpha \to 0} \frac{1}{\alpha} (1-u)^{\alpha}. \qquad (2.9)$$

- The power-approximation *f-Divergence* is defined by:

$$f(u) = u \underbrace{\frac{\left(1 - \frac{1}{u^{\alpha}}\right)}{\alpha}}_{g\left(\frac{1}{u}\right)}. \qquad (2.10)$$

- Conjugate Power approximation criterion:

$$F_{\alpha}(q) = \frac{1}{\alpha(1-\alpha)} \frac{1}{N} \sum_{n=1}^{N} (1 - q^{\alpha}(c_n|x_n)) \overset{\alpha \to 0}{\rightsquigarrow} F_{CE}(q) \qquad (2.11)$$

Conjugate Power Approximation Criterion

- Logarithm:

$$-\log(u) = \lim_{\alpha \to 0} \frac{1}{\alpha} (1-u)^{\alpha} \, . \tag{2.9}$$

- The power-approximation *f-Divergence* is defined by:

$$f(u) = u \underbrace{\frac{\left(1 - \frac{1}{u^{\alpha}}\right)}{\alpha}}_{g\left(\frac{1}{u}\right)} . \tag{2.10}$$

- Conjugate Power approximation criterion:

$$F_{\alpha}(q) = \frac{1}{\alpha(1-\alpha)} \frac{1}{N} \sum_{n=1}^{N} (1 - q^{\alpha}(c_n|x_n)) \overset{\alpha \to 0}{\rightsquigarrow} F_{CE}(q) \tag{2.11}$$

- Non-parametric solution: $q(k|y) \rightsquigarrow \frac{\sqrt[1-\alpha]{pr(k|y)}}{\sum_{c \in \mathcal{C}} \sqrt[1-\alpha]{pr(c|y)}}$

Experimental Setup

Table: Corpus statistics (RW : running words).

Corpus	Train/ Dev/ Eval		
	Data[h]	#Segments	#Words
WSJ0	15.28/ 0.76/ 0.66	7k/ 410/ 330	130k/ 6k/ 5k

Models:

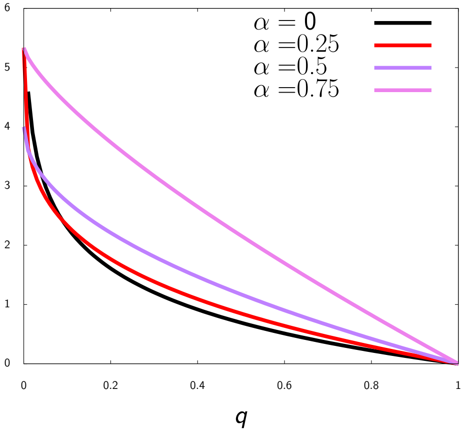
- Bidirectional Gated Recurrent Units (BGRUs).
- BGRUs with dropout.

Overlapping Conjugate Power-Approximation Criteria

$$F_\alpha(q) = \frac{1}{\alpha(1-\alpha)} \frac{1}{N} \sum_{n=1}^{N} (1 - q^\alpha(c_n|x_n))$$

How to choose α ?

$\frac{1}{\alpha(1-\alpha)}(1 - q^\alpha)$

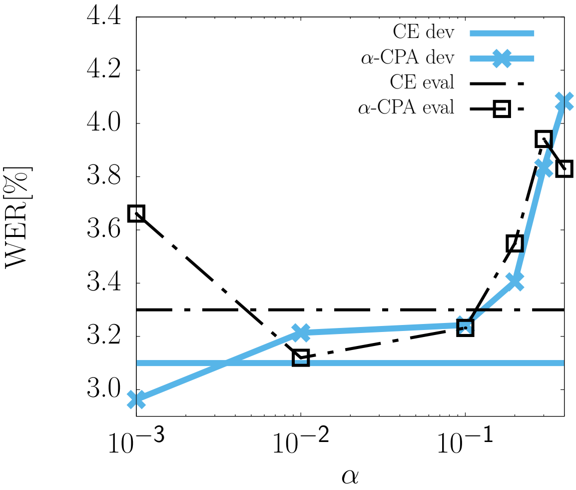


How to Choose α ?

- Grid search: Evaluate α-CPA and (CE+α-CPA)/2

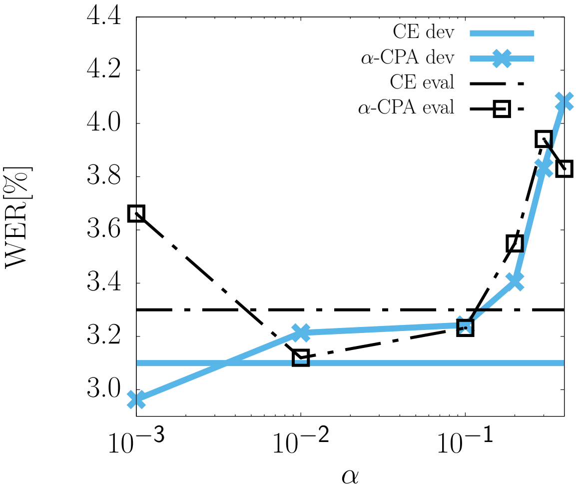
How to Choose α ?

- Grid search: Evaluate α-CPA and (CE+α-CPA)/2



How to Choose α ?

- Grid search: Evaluate α-CPA and (CE+α-CPA)/2
  - Does only result in no or a very small improvement.



Other Strategies to choose α

- Minimize over criteria:

$$\mathcal{G}(q) = \min_{\alpha \in [0,1]} \{\mathcal{F}_\alpha(q)\}$$

Other Strategies to choose α

- Minimize over criteria:

$$\mathcal{G}(q) = \min_{\alpha \in [0,1]} \{\mathcal{F}_\alpha(q)\}$$

- Randomly choose criteria:

$$\mathcal{G}(q) = \min_{\alpha \sim \mathcal{N}(\mu,\sigma^2)} \{\mathcal{F}_\alpha(q)\}$$

Other Strategies to choose α

- Minimize over criteria:

$$\mathcal{G}(q) = \min_{\alpha \in [0,1]} \{\mathcal{F}_\alpha(q)\}$$

- Randomly choose criteria:

$$\mathcal{G}(q) = \min_{\alpha \sim \mathcal{N}(\mu, \sigma^2)} \{\mathcal{F}_\alpha(q)\}$$

- Different α per sample / mini-batch.

Other Strategies to choose α

- Minimize over criteria:

$$\mathcal{G}(q) = \min_{\alpha \in [0,1]} \{\mathcal{F}_\alpha(q)\}$$

- Randomly choose criteria:

$$\mathcal{G}(q) = \min_{\alpha \sim \mathcal{N}(\mu,\sigma^2)} \{\mathcal{F}_\alpha(q)\}$$

- Different α per sample / mini-batch.
- Choose a cutoff $\alpha \in [0, \beta]$.

Minimum Conjugate Power Approximation

- Minimize per sample over criterion over $\alpha \in [0, \beta]$:

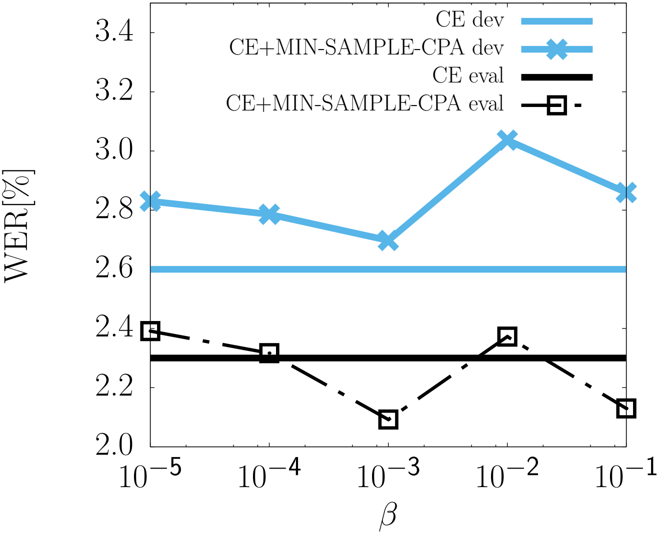
$$\frac{1}{N}\sum_{n=1}^{N} \min_{\alpha\in[0,\beta]} \left\{ \frac{(1 - q^{\alpha}(c_n|x_n))}{\alpha(1-\alpha)} \right\} \qquad \text{(MIN-SAMP-CPA)}$$

- Minimize per batch over criterion over $\alpha \in [0, \beta]$:

$$\min_{\alpha\in[0,\beta]} \left\{ \frac{1}{N}\sum_{n=1}^{N} \frac{(1 - q^{\alpha}(c_n|x_n))}{\alpha(1-\alpha)} \right\} \qquad \text{(MIN-BATCH-CPA)}$$

Choose $\beta \in \{10^{-i} | i \in \{1, 2, 3, 4, 5\}\}$.

Experimental Results Minimization



Noisy Conjugate Power Approximation

- Randomly choose criterion with $\alpha \sim \mathcal{N}(\mu, \sigma^2)$ per sample:

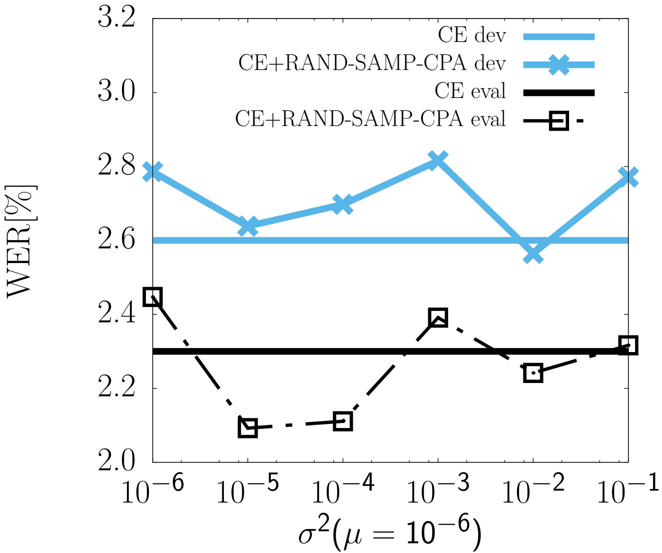
$$\frac{1}{N} \sum_{n=1}^{N} \underset{\substack{\alpha \sim \mathcal{N}(\mu,\sigma^2) \\ \alpha \in [0,1]}}{} \left\{ \frac{(1 - q^{\alpha}(c_n|x_n))}{\alpha(1-\alpha)} \right\} \quad \text{(RAND-SAMP-CPA)}$$

- Randomly choose criterion $\alpha \sim \mathcal{N}(\mu, \sigma^2)$ per batch:

$$\underset{\substack{\alpha \sim \mathcal{N}(\mu,\sigma^2) \\ \alpha \in [0,1]}}{} \left\{ \frac{1}{N} \sum_{n=1}^{N} \frac{(1 - q^{\alpha}(c_n|x_n))}{\alpha(1-\alpha)} \right\} \quad \text{(RAND-BATCH-CPA)}$$

- Choose $\mu \in \{0.1, 0.01, \ldots, 0.000001\}$
- and $\sigma^2 \in \{0.1, 0.01, \ldots, 0.000001\}$.

Experimental Results Randomization



Experimental Results for BGRUs

- Among the single criteria the minimization per samples performs best.

MODEL	CRITERION	WER[%]	
		DEV	EVAL
BGRU	CE	2.6	2.4
	MIN-SAMP-CPA	2.7	2.0
	MIN-BATCH-CPA	2.7	2.2
	RAND-SAMP-CPA	2.6	2.2
	RAND-BATCH-CPA	2.5	2.2

Experimental Results for BGRUs

- In combination with cross-entropy the randomization per sample performs best.

MODEL	CRITERION	WER[%]	
		DEV	EVAL
BGRU	CE	2.6	2.4
	CE+MIN-SAMP-CPA	2.7	2.2
	CE+MIN-BATCH-CPA	2.7	2.1
	CE+RAND-SAMP-CPA	2.6	2.0
	CE+RAND-BATCH-CPA	2.5	2.2

Experimental Results for BGRUs and Dropout

- With dropout the error rate increases.

MODEL	CRITERION	WER[%]	
		DEV	EVAL
BGRU	CE	2.6	2.4
+Dropout(0.1)	CE	2.6	2.3
	CE+MIN-SAMP-CPA	2.7	2.1
	CE+RAND-SAMP-CPA	2.6	2.1

Conclusion

- Scheme to derive training criteria from error bounds.
- Novel regularization schemes to avoid local minima.
- Application to sequence training in automatic speech recognition ?

Thanks

Thanks for your attention.

Experimental Setup

Table: Corpus statistics (RW : running words).

Corpus	Train/ Dev/ Eval		
	Data[h]	#Segments	#Words
WSJ0	15.28/ 0.76/ 0:66	7k/ 410/ 330	130k/ 6k/ 5k

- 5 k recognition lexicon,
- 3-gram recognition language model.
- 1500 context-dependent states,
- alignment for DNN training derived from a speaker independent gaussian mixture model,
- 200k GMM densities,
- GMM features: 40 dimensional LDA-PLP features,
- DNNs: 40 dimensional VTLN-warped Log-Mel features augmented with delta + double-delta.

Recurrent Neural Network Training Recipe

- Truncated back-propagation through time:
  - Utterances sorted by length.
  - Split into subsequences of 21 frames.
  - Uniform sampling with overlap of 10 frames.
  - Starting point shifted by offset of 0 to 9.
  - Minibatch composed of subsequences from same time period.

Recurrent Neural Network Training Recipe

- Back-propagation through time truncated to 21 frames.
- Minibatch composed of subsequences from same time period of utterances with similar length.
- Training:
  - Bidirectional RNNs are unrolled on sequence of 21 frames.
  - Alignment target presented to each frame.
- Testing:
  - Unrolled on spectral window.
  - Center is returned from the last BRNN layer. [Mohamed et al., 2015, Deep Bi-directional Recurrent Networks over Spectral Windows]

Recurrent Neural Network Training Recipe

- Stochastic Gradient Descent Optimizer:
  - Adam.
  - Initial learning rate 0.001.
  - Clock reset after decay on heldout set.
- Learning rate schedule:
  - 50 epochs.
  - Newbob.
  - Decay 0.85
- Early stopping.

Mohamed, A., Seide, F., Yu, D., Droppo, J., Stolcke, A., Zweig, G., and Penn, G. (2015).
Deep bi-directional recurrent networks over spectral windows.
In *2015 IEEE Workshop on Automatic Speech Recognition and Understanding, ASRU 2015, Scottsdale, AZ, USA, December 13-17, 2015*, pages 78–83.

Nußbaum-Thom, M., Beck, E., Alkhouli, T., Schlüter, R., and Ney, H. (2013).
Relative Error Bounds for Statistical Classifiers Based on the f-Divergence.
In *Interspeech*, Lyon, France.

Two Class f-Divergence Aggregation

- Proof Assume $a_1, \ldots, a_I, b_1, \ldots b_I \geq 0, a = \sum_{i=1}^{I} a_i, b = \sum_{i=1}^{I} b_i$:

$$\sum_{i=1}^{I} b_i f\left(\frac{a_i}{b_i}\right) \geq \sum_{i=1}^{I} b\frac{b_i}{b} f\left(\frac{a}{b}\frac{\frac{a_i}{a}}{\frac{b_i}{b}}\right) \tag{6.12}$$

$$\boxed{\textit{Jensens} \text{ inequality: } E(f(X)) \geq f(E(X))} \tag{6.13}$$

$$\geq bf\left(\frac{a}{b}\sum_{i=1}^{I}\frac{b_i}{b}\frac{\frac{a_i}{a}}{\frac{b_i}{b}}\right) \tag{6.14}$$

$$= bf\left(\frac{a}{b}\sum_{i=1}^{I}\frac{a_i}{a}\right) \tag{6.15}$$

$$= bf\left(\frac{a}{b}\right) \tag{6.16}$$

$$\tag{6.17}$$

Two Class f-Divergence Aggregation (I)

- *Jensens* inequality:

$$f\left(\int x p(x)\,\mathrm{d}x\right) \leq \int p(x)\, f(x)\,\mathrm{d}x$$

- Define for $\pi \in \{pr, q\}$ and $r \in R = \{c_{pr}, c_q\}$:

$$\overline{\pi}(r) = \int \pi(x, r(x))\,\mathrm{d}x$$

- Then:

$$D_f(pr||q) = \int \sum_{c \in \mathcal{C}} q(x,c) f\left(\frac{pr(x,c)}{q(x,c)}\right)\,\mathrm{d}x \tag{6.18}$$

Two Class f-Divergence Aggregation (II)

$$= \int \sum_{r \in R} q(x, r(x)) f\left(\frac{pr(x, r(x))}{q(x, r(x))}\right) \mathrm{d}x$$

$$+ \int \sum_{r \in R} \sum_{c \neq r(x)} q(x, c) f\left(\frac{pr(x, c)}{q(x, c)}\right) \mathrm{d}x$$

Two Class f-Divergence Aggregation (III)

$$\geq \sum_{r \in R} \overline{q}(r) f\left(\frac{\overline{pr}(r)}{\overline{q}(r)}\right) + \left(1 - \sum_{r \in R} \overline{q}(r)\right) f\left(\frac{1 - \sum_{r \in R} \overline{pr}(r)}{1 - \sum_{r \in R} \overline{q}(r)}\right)$$

"aggregation"

$$\geq \sum_{r \in R} \overline{q}(r) f\left(\frac{\frac{\overline{pr}(r)}{2}}{\frac{\overline{q}(r)}{2}}\right) + 2 \frac{\left(1 - \sum_{r \in R} \overline{q}(r)\right)}{2} f\left(\frac{\frac{1 - \sum_{r \in R} \overline{pr}(r)}{2}}{\frac{1 - \sum_{r \in R} \overline{q}(r)}{2}}\right)$$

"aggregation"

Two Class f-Divergence Aggregation (IV)

$$\geq \overline{q}(c_{pr}) f\left(\frac{\overline{pr}(c_{pr})}{\overline{q}(c_{pr})}\right) + \frac{\left(1 - \sum_{r \in R} \overline{q}(r)\right)}{2} f\left(\frac{\frac{1 - \sum_{r \in R} \overline{pr}(r)}{2}}{\frac{1 - \sum_{r \in R} \overline{q}(r)}{2}}\right)$$

$$+ \overline{q}(c_p) f\left(\frac{\overline{q}(c_q)}{\overline{q}(c_q)}\right) + \frac{\left(1 - \sum_{r \in R} \overline{q}(r)\right)}{2} f\left(\frac{\frac{1 - \sum_{r \in R} \overline{pr}(r)}{2}}{\frac{1 - \sum_{r \in R} \overline{q}(r)}{2}}\right)$$

"aggregation"

Two Class f-Divergence Aggregation (V)

$$\geq \left(\overline{q}(c_{pr}) + \frac{1 - \sum\limits_{r \in R} \overline{q}(r)}{2} \right) f \left(\frac{\overline{pr}(c_{pr}) + \frac{1 - \sum\limits_{r \in R} \overline{pr}(r)}{2}}{\overline{q}(c_{pr}) + \frac{1 - \sum\limits_{r \in R} \overline{q}(r)}{2}} \right)$$

$$+ \left(\overline{q}(c_q) + \frac{1 - \sum\limits_{r \in R} \overline{q}(r)}{2} \right) f \left(\frac{\overline{pr}(c_q) + \frac{1 - \sum\limits_{r \in R} \overline{pr}(r)}{2}}{\overline{q}(c_q) + \frac{1 - \sum\limits_{r \in R} \overline{q}(r)}{2}} \right)$$

Two Class f-Divergence Aggregation (VI)

$$= \left(\frac{1}{2} + \frac{1}{2}\overline{q}(c_{pr}) - \frac{1}{2}\overline{q}(c_q)\right) f\left(\frac{\frac{1}{2} + \frac{1}{2}\overline{pr}(c_{pr}) - \frac{1}{2}\overline{pr}(c_q)}{\frac{1}{2} + \frac{1}{2}\overline{q}(c_{pr}) - \frac{1}{2}\overline{q}(c_q)}\right)$$

$$+ \left(\frac{1}{2} + \frac{1}{2}\overline{q}(c_q) - \frac{1}{2}\overline{q}(c_{pr})\right) f\left(\frac{\frac{1}{2} + \frac{1}{2}\overline{pr}(c_q) - \frac{1}{2}\overline{pr}(c_{pr})}{\frac{1}{2} + \frac{1}{2}\overline{q}(c_q) - \frac{1}{2}\overline{q}(c_{pr})}\right)$$

[with $\Delta^q = \overline{q}(c_q) - \overline{q}(c_{pr})$]

$$= \left(\frac{1}{2} - \frac{1}{2}\Delta^q)\right) f\left(\frac{\frac{1}{2} + \frac{1}{2}\Delta}{\frac{1}{2} - \frac{1}{2}\Delta^q}\right) + \left(\frac{1}{2} + \frac{1}{2}\Delta^q\right) f\left(\frac{\frac{1}{2} - \frac{1}{2}\Delta}{\frac{1}{2} + \frac{1}{2}\Delta^q}\right)$$

back

Taylor's theorem

- Let $k \in \mathbb{N}$ and $f : \mathbb{R} \to \mathbb{R}$ be k times differentiable in $y_0 \in \mathbb{R}$.
- Then exists a $\mu_y \in [y_0, y]$ with

$$R_k(y) = \frac{f^{(k+1)}(\mu_y)}{k!}(y - y_0)^{k+1}$$

such that:

$$f(y) = \sum_{n=0}^{k-1} \frac{f^{(n)}(y)(y - y_0)^n}{n!} + R_k(y)$$

Explicit Classification Error Bound (I)

- Assume $f(1) = 0$, $f'''(u), u \in [1, 2]$ monotonically increasing.
- Then the following explicit bound can be formulated:

$$f''(1)\Delta^2(x) \le D_f^x(pr||q)$$

- Proof by *Taylor* expansion in in $y_0 = 1$:

$$\begin{aligned}
&2D_f^x(pr||q) \\
\ge& f\left(1+\Delta(x)\right) + f\left(1-\Delta(x)\right) \\
=& f(1) + f'(1)\Delta(x) + \frac{f''(1)\Delta^2(x)}{2!} + \frac{f'''(\mu_{1+\Delta(x)})}{3!}\Delta^3(x) \\
&+ f(1) - f'(1)\Delta(x) + \frac{f''(1)\Delta^2(x)}{2!} - \frac{f'''(\mu_{1-\Delta(x)})}{3!}\Delta^3(x) \\
=& 2\underbrace{f(1)}_{=0} + 2f''(1)\Delta^2(x) + \frac{\Delta^3(x)}{3!}\left(f'''(\mu_{1+\Delta(x)}) - f'''(\mu_{1-\Delta(x)})\right)
\end{aligned}$$

Explicit Classification Error Bound (II)

$$\geq 2f''(1)\Delta^2(x) + \frac{\Delta^3(x)}{3!}\left(\underbrace{\min_{a\in[1,1+\Delta(x)]} f'''(a)}_{\geq f'''(1)} + \underbrace{\max_{b\in[1-\Delta(x),1]} -f'''(b)}_{\geq -f'''(1)}\right)$$

$$\geq 2f''(1)\Delta^2(x) + \frac{\Delta^3(x)}{3!}(f'''(1) - f'''(1))$$

$$= 2\underbrace{f''(1)}_{\geq 0}\Delta^2(x)$$

back

Explicit Classification Error Bound (III)

$$\frac{1}{f''(1)} \int pr(x) D_f^x(pr||q)\,\mathrm{d}x \qquad \text{"} f(u) = ug\left(\tfrac{1}{u}\right)\text{"}$$

$$= \frac{1}{f''(1)} \int pr(x) \sum_{c \in \mathcal{C}} q(c|x) \frac{pr(c|x)}{q(c|x)} g\left(\frac{q(c|x)}{pr(c|x)}\right) \mathrm{d}x$$

$$= \frac{1}{f''(1)} \int \sum_{c \in \mathcal{C}} pr(x,c) g\left(\frac{q(c|x)}{pr(c|x)}\right) \mathrm{d}x$$

$$\leq \frac{1}{f''(1)} \int \sum_{c \in \mathcal{C}} pr(x,c) g\left(\frac{q(c|x)}{1}\right) \mathrm{d}x$$

"g monotonically decreasing"

$$= \frac{1}{f''(1)} \int \sum_{c \in \mathcal{C}} pr(x,c) g(q(c|x))\,\mathrm{d}x$$

back

Optimal Non-parametric Solution (I)

- $pr(x,c) = \lim_{N\to\infty} \frac{1}{N} \sum_{n=1}^{N} \delta(c, c_n)\delta(x - x_n)$
- Constrained training criterion $\mathcal{F}_f(q)$ in $y \in \mathcal{X}$.

$$\begin{aligned}\overline{\mathcal{F}}_f(q) =& \mathcal{F}_f(q) - \mu \left(\sum_{c\in\mathcal{C}} q(c|y) - 1 \right) \\ =& \frac{1}{f''(1)} \frac{1}{N} \sum_{n=1}^{N} g(q(c_n|x_n)) - \mu \left(\sum_{c\in\mathcal{C}} q(c|y) - 1 \right)\end{aligned}$$

$$\text{"}h(x_n) = \int \delta(x - x_n) h(x)\, \mathrm{d}x\text{"}$$

$$\begin{aligned}=& \frac{1}{f''(1)} \frac{1}{N} \sum_{n=1}^{N} \sum_{c\in\mathcal{C}} \int g(q(c|x))\delta(c, c_n)\delta(x - x_n)\, \mathrm{d}x \\ & - \mu \left(\sum_{c\in\mathcal{C}} q(c|y) - 1 \right)\end{aligned}$$

Optimal Non-parametric Solution (II)

$$= \frac{1}{f''(1)} \sum_{c \in \mathcal{C}} \int g(q(c|x)) \frac{1}{N} \sum_{n=1}^{N} \delta(c, c_n) \delta(x - x_n) \, \mathrm{d}x$$

$$- \mu \left(\sum_{c \in \mathcal{C}} q(c|y) - 1 \right)$$

$$= \frac{1}{f''(1)} \sum_{c \in \mathcal{C}} \int g(q(c|x)) \frac{1}{N} \sum_{n=1}^{N} \delta(c, c_n) \delta(x - x_n) \, \mathrm{d}x$$

$$- \mu \left(\sum_{c \in \mathcal{C}} q(c|y) - 1 \right)$$

Optimal Non-parametric Solution (III)

- Derivative w.r.t $q(k|y)$ and μ and consider $N \to \infty$:

$$\nabla_{q(k|y)} \overline{\mathcal{F}}_f(q) = \frac{1}{f''(1)} q^{\alpha-1}(k|y) \frac{1}{N} \sum_{n=1}^{N} \delta(k, c_n)\delta(y - x_n) - \mu$$

$$= \frac{1}{f''(1)} q^{\alpha-1}(k|y) pr(y, k) - \mu \stackrel{!}{=} 0$$

$$\nabla_{\mu} \overline{\mathcal{F}}_f(q) = \left(\sum_{c \in \mathcal{C}} q(c|y) - 1 \right) \stackrel{!}{=} 0$$

Optimal Non-parametric Solution (IV)

- Recombine

$$q(k|y) = \sqrt[1-\alpha]{\frac{\frac{1}{f''(1)} pr(y,k)}{\mu}} \tag{6.19}$$

$$1 = \sum_{c \in \mathcal{C}} \sqrt[1-\alpha]{\frac{\frac{1}{f''(1)} pr(y,c)}{\mu}} \tag{6.20}$$

$$\Rightarrow \sqrt[1-\alpha]{\mu} = \sum_{c \in \mathcal{C}} \sqrt[1-\alpha]{\frac{1}{f''(1)} pr(y,c)} \tag{6.21}$$

Optimal Non-parametric Solution (V)

- Optimal non-parametric solution:

$$q(k|y) = \frac{\sqrt[1-\alpha]{\frac{1}{f''(1)} pr(y,k)}}{\sum_{c \in \mathcal{C}} \sqrt[1-\alpha]{\frac{1}{f''(1)} pr(y,c)}}$$

$$= \frac{\sqrt[1-\alpha]{pr(k|y)}}{\sum_{c \in \mathcal{C}} \sqrt[1-\alpha]{pr(c|y)}}$$

back