The Latent Maximum Entropy Principle for Language Modeling

Hsuan-Sheng Chiu

Reference:
Rosenfeld, “A Maximum Entropy Approach to Adaptive Statistical Language Modeling, 1996
Chen, et al. “A Survey of Smoothing Techniques for ME Models”, 2000
Wang, et al. “The Latent Maximum Entropy Principle”, 2003
Wang, et al. “Combining Statistical Language Models via Latent Maximum Entropy Principle, 2005

Outline

- Introduction
- ME
- LME
- Conclusion

History of Maximum Entropy

- Ed. T. Jaynes, “Information Theory and Statistical Mechanics”, 1957
  - *“Information theory provides a constructive criterion for setting up probability distributions on the basis of partial knowledge, and leads to a type of statistical inference which is called the maximum entropy estimate. It is least biased estimate possible on the **given** information; i.e., it is maximally noncommittal with regard to missing information”.*
  - *http://bayes.wustl.edu/etj/etj.html*
- Stephen A. Della Pietra and Vincent J. Della Pietra, “Statistical Modeling by Maximum Entropy”, 1993
  - *Fuzzy ME*
- Adam L. Berger, Stephen A. Della Pietra and Vincent J. Della Pietra, “A Maximum Entropy Approach to Natural Language Processing”,1996
  - *First introduced to NLP (LM)*
- Ronald Rosenfeld, “A Maximum Entropy Approach to Adaptive Statistical Language Modeling”, 1996

History of Maximum Entropy (cont.)

- Shaojun Wang, Ronald Rosenfeld and Yunxin Zhao, “Latent Maximum Entropy Principle For Statistical Language Modeling” ,ASRU 2001
  - LME principle

- Other papers about ME:
  - R. Rosenfeld et al. “Whole-sentence exponential language models: a vehicle for linguistic-statistical integration”, 2001
  - S. Wang et al. “Semantic n-gram Language Modeling with the Latent Maximum Entropy Principle”, ICASSP 2003
  - C-H. Chueh et al. “A maximum Entropy Approach for Integrating Semantic Information in Statistical Language Models”, ISCSLP 2004
  - C-H. Chueh et al. “Discriminative Maximum Entropy Language Model for Speech Recognition”, Interspeech 2005
  - C-H. Chueh et al. “Maximum Entropy Modeling of Acoustic and Linguistic Features”, ICASSP 2006

Introduction

- The quality of a language model M can be judged by its cross entropy with the distribution of some hitherto unseen text T

$$H(P_T; T_M) = -\sum_x P_T(x) \cdot \log P_M(x)$$

- The goal of statistical language modeling is to identify and exploit sources of information in the language stream, so as to bring the cross entropy down, as close as possible to the true entropy

Introduction (cont.)

- Information Sources (in the Document's History)
  - Context-free estimation (unigram)
  - Short-term history (n-gram)
  - Short-term class history (class n-gram)
  - Intermediate distance (skip n-gram)
  - Long distance (trigger)
  - Syntactic constraints
  - (Observed information)

Introduction (cont.)

- Combining information sources
  - Linear interpolation

$$P_{combined}(w \mid h) \overset{def}{=} \sum_{i=1}^{k} \lambda_i P_i(w \mid h), \textbf{ where } 0 \le \lambda_i \le 1 \textbf{ and } \sum_i \lambda_i = 1$$

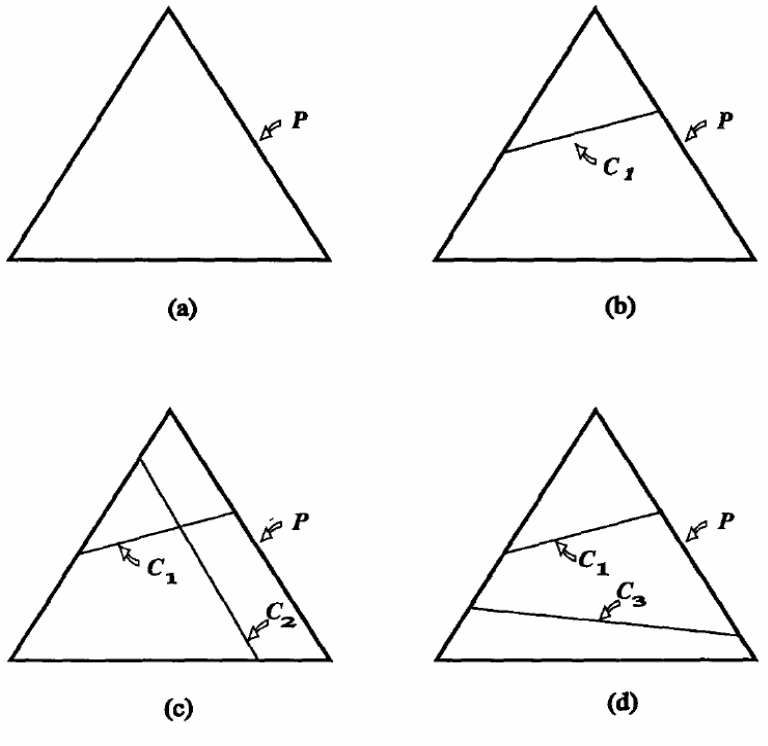
    - Extremely general
    - Easy to implement, experiment with and analyze
    - Can not hurt (no worse than any of its components)
    - Suboptimal ?
    - Inconsistent with components

  - Backoff

The Maximum Entropy Principle

- Under the Maximum Entropy approach, one does not construct separate models. Instead, one builds a single, combined model, which attempts to capture all the information form various knowledge source.

- Constrained optimization
  - a) all p are allowable
  - b) p lying on the line are allowable
  - c) p at intersection are allowable
  - d) no model can satisfy



The Maximum Entropy Principle (cont.)

- Estimate $P(BANK \mid h)$

	h ends in "THE"	h ends in "OF"	...	...
LOAN $\in h$	. . . . . .	. . . . . .	. . . . . .	. . . . . .
LOAN $\notin h$	. . . . . .	. . . . . .	. . . . . .	. . . . . .

$$P_{BIGRAM}(BANK \mid THE) = K_{\{THE,BANK\}}$$

$$P_{LOAN \to BANK}(BANK \mid LOAN \in h) = K_{\{BANK,LOAN \in h\}}$$

$$K_{\{THE,BANK\}} = \frac{C(THE, BANK)}{C(THE)}$$

  - Mutually inconsistent. How to reconcile?
    - Linear interpolation
    - Backoff
    - ME

The Maximum Entropy Principle (cont.)

- ME does away with the inconsistency by relaxing the conditions imposed by the component sources

$$\underset{\text{h end in "THE"}}{E}\left[P_{COMBINED}(BANK \mid h)\right] = K_{\{THE, BANK\}}$$

$$\underset{\text{"LOAN"} \in \text{h}}{E}\left[P_{COMBINED}(BANK \mid h)\right] = K_{\{BANK, LOAN \in h\}}$$

  - Only requires probability P be equal to K on average in the training data

- There are many different function P that would satisfy it

- Try to find out the model p among all the function in that intersection that has the highest entropy

The Maximum Entropy Principle (cont.)

- Information Sources as Constraint Functions

$$\sum_{(h,w)\in S}\left[P(h,w)\right]=K$$

- Index function (selector function)

$$f_S(h,w)\overset{def}{=}\begin{cases}1 & \textbf{if } (\mathbf{h},\mathbf{w})\in\mathbf{S}\\ 0 & \textbf{otherwise}\end{cases}$$

- So

$$\sum_{(h,w)}\left[P(h,w)f_S(h,w)\right]=K$$

- Any real-valued function can be used.
  - Constraint function

The Maximum Entropy Principle (cont.)

- Constrained optimization (primal)

$$P^* = \arg\max_{P \in C} H(P)$$

$$= \arg\max_{P \in C} \left(-\sum_{h,w} \tilde{P}(h) P(w \mid h) \log P(w \mid h) \right)$$

- Lagrangian

$$\Lambda(P, \Lambda, \gamma) = -\sum_{h,w} \tilde{P}(h) P(w \mid h) \log P(w \mid h)$$

$$+ \sum_i \lambda_i \left(\sum_{h,w} \tilde{P}(h) P(w \mid h) f_i(h,w) - \sum_{h,w} \tilde{P}(h,w) f_i(h,w) \right)$$

$$- \gamma_h \left(\sum_w P(w \mid h) - 1 \right)$$

Derivation of Exponential Form

$$\frac{\partial \Lambda}{\partial P(w \mid h)} = -\tilde{P}(h)\left(1 + \log P(w \mid h)\right) + \sum_i \lambda_i \tilde{P}(h) f_i(h, w) - \gamma = 0$$

$$\Rightarrow \tilde{P}(h)\left(1 + \log P^*(w \mid h)\right) = \sum_i \lambda_i \tilde{P}(h) f_i(h, w) - \gamma$$

$$\Rightarrow \log P^*(w \mid h) = \sum_i \lambda_i f_i(h, w) - \frac{\gamma}{\tilde{P}(h)} - 1$$

$$\Rightarrow P^*(w \mid h) = \exp\left(\sum_i \lambda_i f_i(h, w)\right) \exp\left(-\frac{\gamma}{\tilde{P}(h)} - 1\right)$$

$$\because \forall h, \sum_w P(w \mid h) = 1$$

$$\Rightarrow \sum_w P(w \mid h) = \sum_w \exp\left(\sum_i \lambda_i f_i(h, w)\right) \exp\left(-\frac{\gamma}{\tilde{P}(h)} - 1\right) = 1$$

$$\Rightarrow \exp\left(-\frac{\gamma}{\tilde{P}(h)} - 1\right) \cdot \sum_w \exp\left(\sum_i \lambda_i f_i(h, w)\right) = 1$$

$$\Rightarrow \exp\left(-\frac{\gamma}{\tilde{P}(h)} - 1\right) = \frac{1}{\sum_w \exp\left(\sum_i \lambda_i f_i(h, w)\right)}$$

Exponential Form

- Exponential form

$$P^*(w \mid h) = \frac{1}{\sum_{\hat{w}} \exp\left(\sum_i \lambda_i f_i(h, \hat{w})\right)} \cdot \exp\left(\sum_i \lambda_i f_i(h, w)\right) = \frac{1}{Z(h)} \exp\left(\sum_i \lambda_i f_i(h, w)\right)$$

- Dual function

$$\Psi(\lambda) = -\sum_h p(h) \log Z_\lambda(h) + \sum_i \lambda_i \sum_{h,w} \tilde{P}(h, w) f_i(h, w)$$

- Suppose that λ^* of $\Psi(\lambda)$ is the solution of the dual problem, then p_{λ^*} is the solution of the primal problem
  - The maximum entropy model subject to the constraints hasp parametric form p_{λ^*}, where the parameter values λ^* can be determined by maximizing the dual function $\Psi(\lambda)$

Duality

- Robert M. Freund, "Applied Lagrange Duality for Constrained Optimization", 2004

$$S := \left\{ (r, z) \in \Re^{m+1} \mid (r, z) = (g(x), f(x)) \text{ for some } x \in P \right\} .$$

$$H = H_{u,\alpha} = \left\{ (r, z) \in \Re^{m+1} \mid z + u^T r = \alpha \right\} .$$

$$\begin{pmatrix} r \\ z \end{pmatrix} = \begin{pmatrix} r_1 \\ r_2 \\ \vdots \\ r_m \\ z \end{pmatrix} = \begin{pmatrix} g_1(x) \\ g_2(x) \\ \vdots \\ g_m(x) \\ f(x) \end{pmatrix} = \begin{pmatrix} g(x) \\ f(x) \end{pmatrix} .$$

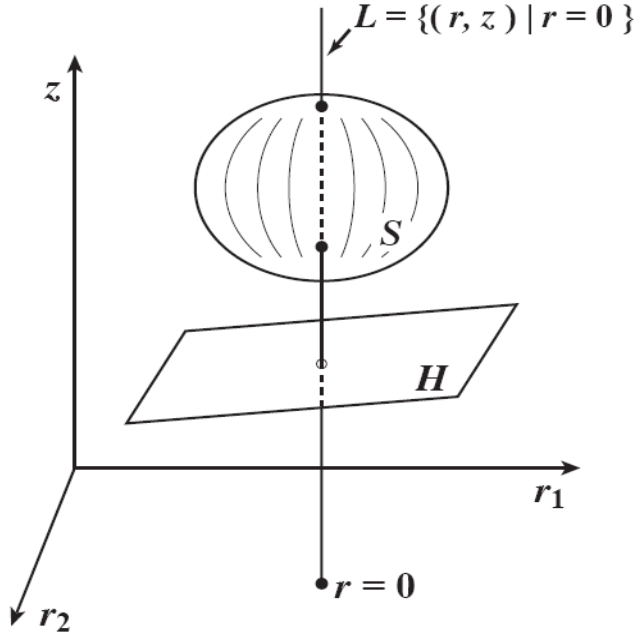


Figure 10: The column geometry of the primal and dual problem.

Relation to Maximum Likelihood

- Definition of log-likelihood

$$L_{\tilde{P}}(P) \equiv \log \prod_{h,w} P(w \mid h)^{\tilde{P}(h,w)} = \sum_{h,w} \tilde{P}(h,w) \log P(w \mid h)$$

- Replace p with exponential form

$$L_{\tilde{P}}(P) = \sum_{h,w} \tilde{P}(h,w) \left(\sum_i \lambda_i f_i(h,w) \right) - \sum_{h,w} \tilde{P}(h,w) \log \sum_{\hat{w}} \exp\left(\sum_i \lambda_i f_i(h,\hat{w}) \right)$$

$$= \sum_{h,w} \tilde{P}(h,w) \sum_i \lambda_i f_i(h,w) - \sum_h \tilde{P}(h) \log \sum_w \exp\left(\sum_i \lambda_i f_i(h,w) \right)$$

  - Dual function equals to log-likelihood of the training data

- The model with maximum entropy is the model in the parametric family that maximizes the likelihood of the training data

Improved Iterative Scaling (IIS)

- IIS performs hill-climbing in the log-likelihood space with enforcement of two relax lower bounds
  - Adam Berger, “The Improved Iterative Scaling Algorithm: A Gentle Introduction”, 1997
  - Rong Yan, “A variant of IIS algorithm”

- Definition of difference of likelihood function

$$L_{\tilde{P}}(\Lambda+\Delta)-L_{\tilde{P}}(\Lambda)=\sum_{h,w}\tilde{P}(h,w)\log P_{\Lambda+\Delta}(w\mid h)-\sum_{h,w}\tilde{P}(h,w)\log P_{\Lambda}(w\mid h)$$

Improved Iterative Scaling (IIS) (cont.)

- derivation

$$
\begin{aligned}
& L_{\tilde{P}}(\Lambda+\Delta)-L_{\tilde{P}}(\Lambda) \\
&= \sum_{h,w} \tilde{P}(h,w)\log P_{\Lambda+\Delta}(w\mid h)-\sum_{h,w} \tilde{P}(h,w)\log P_{\Lambda}(w\mid h) \\
&= \sum_{h,w} \tilde{P}(h,w)\log\left[\frac{1}{Z_{\Lambda+\Delta}(h)}\exp\left(\sum_i(\lambda_i+\delta_i)f_i(h,w)\right)\right]-\sum_{h,w} \tilde{P}(h,w)\log\left[\frac{1}{Z_{\Lambda}(h)}\exp\left(\sum_i\lambda_i f_i(h,w)\right)\right] \\
&= \sum_{h,w} \tilde{P}(h,w)\left[\sum_i(\lambda_i+\delta_i)f_i(h,w)+\log\left(\frac{1}{Z_{\Lambda+\Delta}(h)}\right)\right]-\sum_{h,w} \tilde{P}(h,w)\left[\sum_i\lambda_i f_i(h,w)+\log\left(\frac{1}{Z_{\Lambda}(h)}\right)\right] \\
&= \sum_{h,w} \tilde{P}(h,w)\left[\sum_i\delta_i f_i(h,w)+\log\left(\frac{1}{Z_{\Lambda+\Delta}(h)}\right)-\log\left(\frac{1}{Z_{\Lambda}(h)}\right)\right] \\
&= \sum_{h,w} \tilde{P}(h,w)\sum_i\delta_i f_i(h,w)-\sum_{h,w} \tilde{P}(h,w)\log\left(\frac{Z_{\Lambda+\Delta}(h)}{Z_{\Lambda}(h)}\right) \\
&= \sum_{h,w} \tilde{P}(h,w)\sum_i\delta_i f_i(h,w)-\sum_{h} \tilde{P}(h)\log\left(\frac{Z_{\Lambda+\Delta}(h)}{Z_{\Lambda}(h)}\right)
\end{aligned}
$$

Improved Iterative Scaling (IIS) (cont.)

$$L_{\tilde{P}}(\Lambda+\Delta)-L_{\tilde{P}}(\Lambda)=\sum_{h,w}\tilde{P}(h,w)\sum_i \delta_i f_i(h,w)-\sum_h \tilde{P}(h)\log\frac{Z_{\Lambda+\Delta}(h)}{Z_\Lambda(h)}$$

$$\geq \sum_{h,w}\tilde{P}(h,w)\sum_i \delta_i f_i(h,w)+\sum_h \tilde{P}(h)\left(1-\frac{Z_{\Lambda+\Delta}(h)}{Z_\Lambda(h)}\right) \qquad -\log a \geq 1-a, a>0$$

$$=\sum_{h,w}\tilde{P}(h,w)\sum_i \delta_i f_i(h,w)+1-\sum_h \tilde{P}(h)\left(\frac{Z_{\Lambda+\Delta}(h)}{Z_\Lambda(h)}\right)$$

$$=\sum_{h,w}\tilde{P}(h,w)\sum_i \delta_i f_i(h,w)+1-\sum_h \tilde{P}(h)\frac{\sum_w \exp\left(\sum_i (\lambda_i+\delta_i) f_i(h,w)\right)}{\sum_{\hat{w}} \exp\left(\sum_i \lambda_i f_i(h,\hat{w})\right)}$$

$$=\sum_{h,w}\tilde{P}(h,w)\sum_i \delta_i f_i(h,w)+1-\sum_h \tilde{P}(h)\frac{\sum_w \exp\left(\sum_i \lambda_i f_i(h,w)\right)\cdot\exp\left(\sum_i \delta_i f_i(h,w)\right)}{\sum_{\hat{w}} \exp\left(\sum_i \lambda_i f_i(h,\hat{w})\right)}$$

$$=\sum_{h,w}\tilde{P}(h,w)\sum_i \delta_i f_i(h,w)+1-\sum_h \tilde{P}(h)\sum_w \frac{\exp\left(\sum_i \lambda_i f_i(h,w)\right)}{\sum_{\hat{w}} \exp\left(\sum_i \lambda_i f_i(h,\hat{w})\right)}\cdot\exp\left(\sum_i \delta_i f_i(h,w)\right)$$

$$=\sum_{h,w}\tilde{P}(h,w)\sum_i \delta_i f_i(h,w)+1-\sum_h \tilde{P}(h)\sum_w p_\Lambda(w\mid h)\exp\left(\sum_i \delta_i f_i(h,w)\right)$$

Improved Iterative Scaling (IIS) (cont.)

$$A(\Delta \mid \Lambda) =$$

$$\sum_{h,w} \tilde{P}(h,w) \sum_i \delta_i f_i(h,w) + 1 - \sum_h \tilde{P}(h) \sum_w p_\Lambda(w \mid h) \exp\left(f^{\#}(h,w) \sum_i \delta_i \frac{f_i(h,w)}{f^{\#}(h,w)} \right)$$

$$* \; f^{\#}(h,w) = \sum_i f_i(h,w)$$

Jensen Inequality： $\exp\left(\sum_x P(x) Q(x) \right) \le \sum_x P(x) \exp(Q(x))$

$$A(\Delta \mid \Lambda)$$

$$= \sum_{h,w} \tilde{P}(h,w) \sum_i \delta_i f_i(h,w) + 1 - \sum_h \tilde{P}(h) \sum_w P_\Lambda(w \mid h) \exp\left(f^{\#}(h,w) \sum_i \delta_i \frac{f_i(h,w)}{f^{\#}(h,w)} \right)$$

$$= \sum_{h,w} \tilde{P}(h,w) \sum_i \delta_i f_i(h,w) + 1 - \sum_h \tilde{P}(h) \sum_w P_\Lambda(w \mid h) \exp\left(\sum_i \frac{f_i(h,w)}{f^{\#}(h,w)} \left(\delta_i f^{\#}(h,w) \right) \right)$$

$$\ge \sum_{h,w} \tilde{P}(h,w) \sum_i \delta_i f_i(h,w) + 1 - \sum_h \tilde{P}(h) \sum_w P_\Lambda(w \mid h) \left(\sum_i \frac{f_i(h,w)}{f^{\#}(h,w)} \exp\left(\delta_i f^{\#}(h,w) \right) \right)$$

Improved Iterative Scaling (IIS) (cont.)

$$B(\Delta \mid \Lambda)$$

$$= \sum_{h,w} \tilde{P}(h,w) \sum_i \delta_i f_i(h,w) + 1 - \sum_h \tilde{P}(h) \sum_w P_\Lambda(w \mid h) \left(\sum_i \frac{f_i(h,w)}{f^\#(h,w)} \exp\left(\delta_i f^\#(h,w)\right) \right)$$

$$\frac{\partial B(\Delta \mid \Lambda)}{\partial \delta_i} = \sum_{h,w} \tilde{P}(h,w) f_i(h,w) - \sum_h \tilde{P}(h) \sum_w P_\Lambda(w \mid h) f_i(h,w) \exp\left(\delta_i f^\#(h,w)\right)$$

- It is straightforward to solve for each of the n free parameters individually by differentiating with respect to δ in turn
- In case $f^\#(h,w)$ is constant for each (h,w) pair, IIS can be degraded to the GIS algorithm and simply solved in close-form
- Otherwise, this can solve with numeric root-finding procedure

IIS (cont.)

- Input: feature functions $f_1, f_2, ..., f_n$; empircal distribution $\tilde{p}(x, y)$
- Output: Optimal parameter values λ_i^* ; optimal model p_λ^*

  - 1. start with $\lambda_i = 0$ for all $i \in \{1,2,...,n\}$
  - 2. do for each $i \in \{1,2,...,n\}$:
    a. Let δ_i be the solution to

$$\sum_{h,w} \tilde{p}(h) p(w \mid h) f_i(h,w) \exp\left(\delta_i f^{\#}(h,w)\right) = \tilde{p}(f_i)$$

    where $f^{\#}(h,w) = \sum_i f_i(h,w)$

    b. Update the value of λ_i according to: $\lambda_i \leftarrow \lambda_i + \delta_i$
  - 3. Go to step 2 if not all the λ_i have converged

IIS (cont.)

- Assume $\exp(\delta_i) = x$, considered trigram model only
  - For each pair (h,w) , $f^{\#}(h, w)$ may be 3,2,or 1

- We can arrange the coefficient of x by the order

$$(a_1 + a_2 + \ldots)x^3 + (b_1 + b_2 + \ldots)x^2 + (c_1 + c_2 + \ldots)x = d$$

a_i, b_i, c_i are model probabilities of one pair (h,w)

d is empirical distribution

- Now it becomes a root-solving problem

- If we get x, we can obtain $\delta_i = \log(x)$

IIS vs. GIS

$$\sum_{h,w} \tilde{P}(h) P_{\Lambda}(w|h) f_i(h,w) \exp\left(\delta_i f^{\#}(h,w)\right) = \sum_{h,w} \tilde{P}(h,w) f_i(h,w)$$

assume $f^{\#}(h,w) = M$ **is a constant for all** $(\mathbf{h}, \mathbf{w})$

$a_1 e^{\delta M} + a_2 e^{\delta M} + a_3 e^{\delta M} + \ldots + a_n e^{\delta M}$ **a_i is the value** $\tilde{P}(h) P_{\Lambda}(w|h) f_i(h,w)$ **for each pair(h, w)**

$= (a_1 + a_2 + a_3 + \ldots + a_n) e^{\delta M}$

$a_1 + a_2 + a_3 + \ldots + a_n = P(f_i)$

$P(f_i) \exp(\delta_i M) = \tilde{P}(f_i)$ **homogeneous function**

$$\exp(\delta_i M) = \frac{\tilde{P}(f_i)}{P(f_i)}, \delta_i M = \log\left(\frac{\tilde{P}(f_i)}{P(f_i)}\right), \delta_i = \frac{\log\left(\frac{\tilde{P}(f_i)}{P(f_i)}\right)}{M}$$

assume $M = 1, \delta_i = \log\left(\frac{\tilde{P}(f_i)}{P(f_i)}\right), \lambda_{i+1} = \lambda_i + \delta_i = \lambda_i + \log\left(\frac{\tilde{P}(f_i)}{P(f_i)}\right)$

take exponential fuction, $\lambda_{i+1} = \lambda_i \left(\frac{\tilde{P}(f_i)}{P(f_i)}\right)$ **= GIS update criterion**

Smoothing Maximum Entropy Models

- A. Constraint Exclusion
  - Exclude constraints for ME n-gram models that occur fewer than a certain number of times in the training data
  - Do not rely on the true frequency of training data
  - Analogous to using count cutoffs for n-gram models
  - Higher entropy than original
  - Feature induction or feature selection criteria

Smoothing Maximum Entropy Models (cont.)

- B. Good-Turing Discounting
  - Use discounted distribution to instead of empirical distribution for exact feature expectation

$$\sum_{h,w} \tilde{p}(h)p(w \mid h)f_i(h,w) = \tilde{p}_{GT}(h,w)$$

  - Katz variation of GT
  - These constraints may no longer be consistent
    - trigram $\tilde{p}(A,B)p(C \mid A,B) = \tilde{p}_{GT}(A,B,C)$
    - bigram $\sum_{w_2} \tilde{p}(A,B)p(C \mid w_2,B) = \tilde{p}_{GT}(B,C)$
    - If B only follows A, but $\tilde{p}_{GT}(A,B,C) \neq \tilde{p}_{GT}(B,C)$
  - Knesey-Ney smoothing (?)
  - Inconsistency is symptomatic of constraints that will lead to poor parameter estimates

Smoothing Maximum Entropy Models (cont.)

- C. Fuzzy ME
  - Instead of requiring that constraints are satisfied exactly, a penalty is associated with inexact constraint satisfaction
  - Find model satisfying constraints that minimizes the KL distance from the uniform model

$$D(q \,\|\, p_{unif}) + U(q)$$

  - Penalty function:

$$U(q) = \sum_{i=1}^{F} \frac{1}{2\rho_i^2} \left[\sum_x q(x) f_i(x) - \sum_x \tilde{p}(x) f_i(x) \right]^2$$

  - View fuzzy ME as imposing Gaussian prior centered around 0 on the parameters: **MAP** estimation
    - The prior nudges the parameters toward zero, thereby making the model more uniform

Smoothing Maximum Entropy Models (cont.)

- Regular ME model with finding parameters that maximize the log-likelihood of the training data

$$L_X(\Lambda) = \sum_x \tilde{p}(x) \log q_\Lambda(x)$$

- With Gaussian Prior:

$$L_x'(\Lambda) = L_X(\Lambda) + \sum_{i=1}^{F} \log\left(\frac{1}{\sqrt{2\pi\sigma_i^2}} \exp\left(-\frac{\lambda_i^2}{2\sigma_i^2} \right) \right) = L_X(\Lambda) - \sum_{i=1}^{F} \log\left(\frac{\lambda_i^2}{2\sigma_i^2} \right) + const(\Lambda)$$

- Modification of IIS:
  - Update $\lambda_i^{(t+1)} = \lambda_i^{(t)} + \delta_i^{(t)}$

    where $\delta_i^{(t)}$ satisfies

$$\sum_x \tilde{p}(x) f_i(x) = \sum_x q_\Lambda^{(t)}(x) f_i(x) \exp\left(\delta_i^{(t)} f^{\#}(x) \right) + \frac{\lambda_i^{(t)} + \delta_i^{(t)}}{\sigma_i^2}$$

Smoothing Maximum Entropy Models (cont.)

- Constraint of Fuzzy ME

$$\sum_x \tilde{p}(x) f_i(x) - \frac{\lambda_i}{\sigma^2} = \sum_x q_\Lambda(x) f_i(x)$$

  - Can be derived from penalty function (see reference 4 appendix)
  - Fuzzy ME smoothing is like as logarithmic discounting (ideal average discounting method)

- D. Fat Constraints
  - I.e. $\sum_{i=1}^{F} W_i \left[\sum_x q(x) f_i(x) - \sum_x \tilde{p}(x) f_i(x) \right]^2 \le \sigma^2$
  - or $\alpha_i \le \sum_x q(x) f_i(x) \le \beta_i, \quad i = 1, \ldots, F$

Derivation of fuzzy ME constraints

from: Chen, et al. “A Survey of Smoothing Techniques for ME Models”

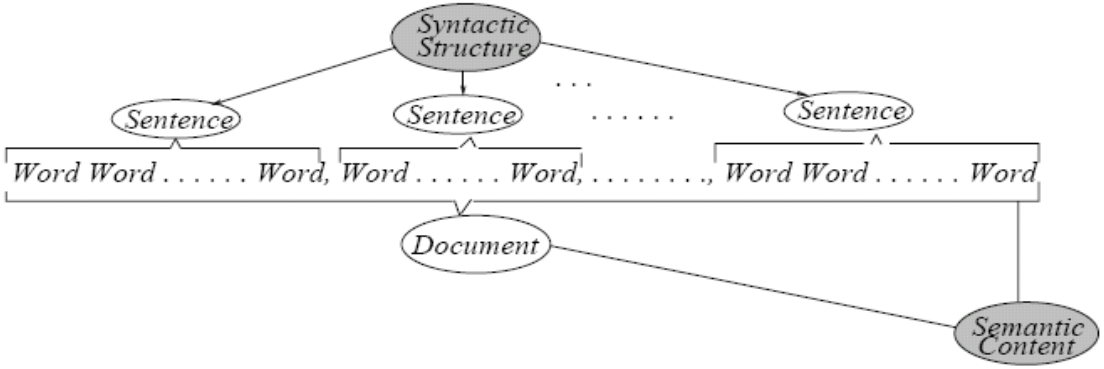
$$
\begin{aligned}
L'_X(\Lambda) &= \sum_{x,\,y} \tilde{p}(x,\,y)\,\log\, q_\Lambda(y|x) - \sum_{i=1}^{F} \frac{\lambda_i^2}{2\sigma_i^2} + \text{const}(\Lambda) \\
&= \sum_{x,\,y} \tilde{p}(x,\,y) \sum_i \lambda_i f_i(x,\,y) \\
&\quad - \sum_{x,\,y} \tilde{p}(x,\,y)\,\log \sum_{y'} \exp\left(\sum_i \lambda_i f_i(x,\,y')\right) \\
&\quad - \sum_{i=1}^{F} \frac{\lambda_i^2}{2\sigma_i^2} + \text{const}(\Lambda)
\end{aligned}
$$

Derivation of fuzzy ME constraints (cont.)

$$
\begin{aligned}
\frac{\partial L'_X(\Lambda)}{\partial \lambda_i} &= \sum_{x,\,y} \tilde{p}(x,\,y) f_i(x,\,y) - \sum_{x,\,y} \tilde{p}(x,\,y) \sum_{y'} \\
&\quad \cdot \frac{\exp\left(\sum_i \lambda_i f_i(x,\,y')\right)}{Z_\Lambda(x)} f_i(x,\,y') - \frac{\lambda_i}{\sigma_i^2} \\
&= \sum_{x,\,y} \tilde{p}(x,\,y) f_i(x,\,y) - \sum_{x,\,y} \tilde{p}(x,\,y) \\
&\quad \cdot \sum_{y'} q_\Lambda(y'|x) f_i(x,\,y') - \frac{\lambda_i}{\sigma_i^2} \\
&= \sum_{x,\,y} \tilde{p}(x,\,y) f_i(x,\,y) - \sum_{x} \tilde{p}(x) \\
&\quad \cdot \sum_{y'} q_\Lambda(y'|x) f_i(x,\,y') \sum_{y} \tilde{p}(y|x) - \frac{\lambda_i}{\sigma_i^2} \\
&= \sum_{x,\,y} \tilde{p}(x,\,y) f_i(x,\,y) \\
&\quad - \sum_{x,\,y} \tilde{p}(x) q_\Lambda(y|x) f_i(x,\,y) - \frac{\lambda_i}{\sigma_i^2}.
\end{aligned}
$$

LME principle

- Illustration



- The problem of ME principle with latent variables is to select a model from a set of allowed probability distributions to maximize the entropy

$$\max_{p} H(p) = -\sum_{x} p(x)\log(x)$$

subject to

$$\sum_{x} p(x) f_i(x) = \sum_{y \in Y} \tilde{p}(y) \sum_{z \in Z} f_i(x) p(z \mid y)$$

  - More general than ME

RLME principle

- Regularized LME principle

$$\max_{p,a} H(p) - U(a) = -\sum_{x} p(x)\log(x) - U(a)$$

subject to

$$\sum_{x} p(x) f_i(x) = \sum_{y \in Y} \tilde{p}(y) \sum_{z \in Z} f_i(x) p(z \mid y) + a_i$$

  - One way to gain robustness to these errors from empirical data is to relax the constraints but add a penalty to the entropy of the joint model

Training algorithm for RLME

- Constraints are nonlinear and the feasible is no longer convex and minima and saddle points may exist
- Restrict p to be an exponential model
  - Approximation
- The key observation to finding feasible solutions is to note that they are intimately related to finding locally MAP solutions
  - Gaussian prior , Laplacian prior

Training algorithm for RLME (cont.)

- R-EM-IS
  - Penalized log-likelihood function

$$R(\lambda)=\sum_y \tilde{p}(y)\log p_\lambda(y)-U^*(\lambda)$$

$$=\sum_y \tilde{p}(y)\sum_z p_{\lambda'}(z\mid y)\log p_\lambda(x)-U^*(\lambda)-\sum_y \tilde{p}(y)\sum_z p_{\lambda'}(z\mid y)\log p_\lambda(z\mid y)$$

$$=-\log(\Phi_\lambda)+\sum_{i=1}^{N}\lambda_i\left(\sum_y \tilde{p}(y)\sum_z p_{\lambda'}(z\mid y)f_i(x)\right)-U^*(\lambda)-\sum_y \tilde{p}(y)\sum_z p_{\lambda'}(z\mid y)\log p_\lambda(z\mid y)$$

$$=Q(\lambda,\lambda')+H(\lambda,\lambda')$$

  - E-step: Compute $\sum_y \tilde{p}(y)\sum_z p_{\lambda'}(z\mid y)f_i(x)$ for i=1,…,N
  - M-step: perform K parallel updates of the parameter values λ_i for i=1,…,N by iterative scaling (GIS or IIS) as follows

$$\lambda_i^{(j+s/K)}=\lambda_i^{(j+(s-1)/K)}+\gamma_i^{(j+s/K)},\ s=1,\ldots,K$$

where $\gamma_i^{(j+s/K)}$ satisfies

$$\sum_y \tilde{p}(y)\sum_z p_{\lambda^{(j)}}(z\mid y)f_i(x)=\sum_x p_{\lambda^{(j+(s-1)/K)}}(x)f_i(x)\exp\left(\gamma_i^{(j+s/K)}f^{\#}(x)\right)+\frac{\lambda_i^{(j+(s-1)/K)}+\gamma_i^{(j+s/K)}}{\sigma_i^2}$$

Combining N-gram and PLSA models

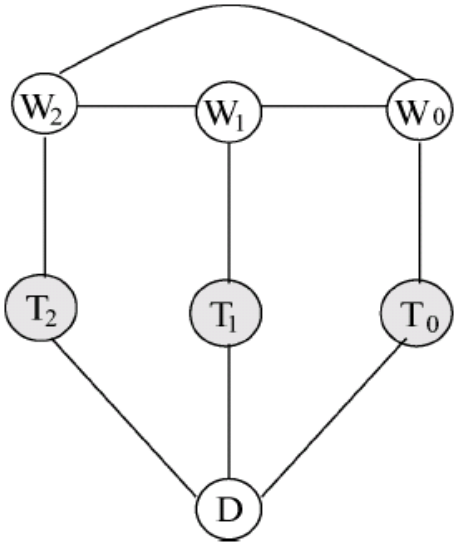
- Complete data $x = (w_2, w_1, w_0, D, T_2, T_1, T_0)$

- For the trigram portion of the model

$$\sum_x p(x)\delta(W_2 = w_i, W_1 = w_j, W_0 = w_k) = \sum_d \tilde{p}(d)p(w_i w_j w_k \mid d)$$

$$\sum_x p(x)\sum_{l=0}^{1}\delta(W_{l+1} = w_i, W_l = w_j) = \sum_d \tilde{p}(d)\sum_{l=0}^{1} p(W_{l+1=} w_i, W_l = w_j \mid d)$$

$$\sum_x p(x)\sum_{l=0}^{2}\delta(W_l = w_i) = \sum_d \tilde{p}(d)\sum_{l=0}^{2} p(W_l = w_i \mid d)$$



- For PLSA portion of the model

$$\sum_x p(x)\sum_{l=0}^{2}\delta(T_l = t, D = d) = \tilde{p}(d)\sum_{l=0}^{2}\tilde{p}(W_l \mid d)p(T_l = t \mid W_l = w_i, D = d)$$

$$\sum_x p(x)\sum_{l=0}^{2}\delta(T_l = t, W_l = w_i) = \sum_d \tilde{p}(d)\sum_{l=0}^{2}\tilde{p}(W_l = w_i \mid d)p(T_l = t \mid W_l = w_i, D = d)$$

Combining N-gram and PLSA models (cont.)

- Efficient feature expectation and inference
  - Normalization factor

$$\Phi = \sum_{w_2,w_{,1},w_0,t_2,t_1,t_0,d} \left(e^{\lambda_{w_2}} e^{\lambda_{w_1}} e^{\lambda_{w_0}} e^{\lambda_{w_2 w_1}} e^{\lambda_{w_1 w_0}} e^{\lambda_{w_2 w_1 w_0}} e^{\lambda_{w_2 t_2}} e^{\lambda_{w_1 t_1}} e^{\lambda_{w_0 t_0}} e^{\lambda_{t_2 d}} e^{\lambda_{t_1 d}} e^{\lambda_{t_0 d}} \right)$$

$$= \sum_{w_0} e^{\lambda_{w_0}} \sum_{t_0} e^{\lambda_{w_0 t_0}} \sum_{w_1} e^{\lambda_{w_1}} e^{\lambda_{w_1 w_0}} \sum_{t_1} e^{\lambda_{w_1 t_1}} \sum_{w_2} e^{\lambda_{w_2}} e^{\lambda_{w_2 w_1}} e^{\lambda_{w_2 w_1 w_0}} \sum_{t_2} e^{\lambda_{w_2 t_2}} \sum_{d} e^{\lambda_{t_2 d}} e^{\lambda_{t_1 d}} e^{\lambda_{t_0 d}}$$

  - Trigram feature expectation

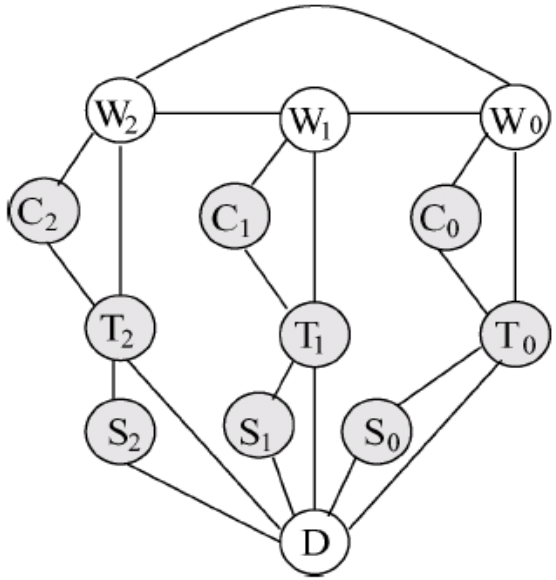
$$\sum_x p(x) \delta(W_2 = w_i, W_1 = w_j, W_0 = w_k) =$$

$$\Phi^{-1} e^{\lambda_{w_0}} e^{\lambda_{w_1}} e^{\lambda_{w_1 w_0}} e^{\lambda_{w_2}} e^{\lambda_{w_2 w_1}} e^{\lambda_{w_2 w_1 w_0}} \sum_{t_0} e^{\lambda_{w_0 t_0}} \sum_{t_1} e^{\lambda_{w_1 t_1}} \sum_{t_2} e^{\lambda_{w_2 t_2}} \sum_{d} e^{\lambda_{t_2 d}} e^{\lambda_{t_1 d}} e^{\lambda_{t_0 d}}$$

Combining N-gram and PLSA models (cont.)

- Semantic smoothing
  - Make use of semantic similarity and subtle variation between words and words and subtopic variation within documents

    add C and S nodes



$$\sum_x p(x)\sum_{l=0}^{2}\delta(T_l=t,C_l=c,D=d)=\tilde{p}(d)\sum_{l=0}^{2}\tilde{p}(W_l\mid d)p(T_l=t,C_l=c\mid W_l=w_i,D=d)$$

$$\sum_x p(x)\sum_{l=0}^{2}\delta(T_l=t,S_l=s,W_l=w_i)=\sum_d\tilde{p}(d)\sum_{l=0}^{2}\tilde{p}(W_l=w_i\mid d)p(T_l=t,S_l=s\mid W_l=w_i,D=d)$$

Computation in testing

$$
\begin{aligned}
p(w_1 \ldots w_{|d|}, d) &= \prod_{\ell=1}^{|d|} p_{D_\ell}(w_\ell, d_\ell \mid w_1 \ldots w_{\ell-1}) \\
&= \prod_{\ell=1}^{|d|} \sum_{T_2,T_1,T_0} p_{D_\ell}(w_\ell, d_\ell, T_2, T_1, T_0 \mid w_1 \ldots w_{\ell-1}) \\
&= \prod_{\ell=1}^{|d|} \sum_{T_2,T_1,T_0} p_{D_\ell}(w_\ell, d_\ell, T_2, T_1, T_0 \mid w_{\ell-2}, w_{\ell-1})
\end{aligned}
$$

Conclusion

- Propose RLME principle for estimating sophisticated mixed chain-table graphical models of natural language

- Computation issue