Maximum Entropy Models
Improving on N-Grams?

- N-grams don’t combine multiple sources of evidence well

\[ P(\text{construction} \mid \text{After the demolition was completed, the}) \]

- Here:
  - “the” gives syntactic constraint
  - “demolition” gives semantic constraint
  - Unlikely the interaction between these two has been densely observed in this specific n-gram

- We’d like a model that can be more statistically efficient
Some Definitions

**INPUTS**

$x_i$


**CANDIDATE SET**

$\mathcal{Y}(x)$

{door, table, ...}

**CANDIDATES**

$y$

$y^*$

$y_i$

door

**TRUE OUTPUTS**

**FEATURE VECTORS**

$f(x, y)$

$[0 0 1 0 0 0 1 0 0 0 0 0]$
More Features, Less Interaction

\[ x = \text{closing the } \_\_\_\_, \ y = \text{doors} \]

- **N-Grams** \[ x_{-1} = \text{“the” } \land \ y = \text{“doors”} \]
- **Skips** \[ x_{-2} = \text{“closing” } \land \ y = \text{“doors”} \]
- **Lemmas** \[ x_{-2} = \text{“close” } \land \ y = \text{“door”} \]
- **Caching** \[ y \text{ occurs in } x \]
## Data: Feature Impact

<table>
<thead>
<tr>
<th>Features</th>
<th>Train Perplexity</th>
<th>Test Perplexity</th>
</tr>
</thead>
<tbody>
<tr>
<td>3 gram indicators</td>
<td>241</td>
<td>350</td>
</tr>
<tr>
<td>1-3 grams</td>
<td>126</td>
<td>172</td>
</tr>
<tr>
<td>1-3 grams + skips</td>
<td>101</td>
<td>164</td>
</tr>
</tbody>
</table>
Exponential Form

- Weights \( w \)
- Features \( f(x, y) \)

- Linear score \( w^\top f(x, y) \)

- Unnormalized probability

\[
P(y|x, w) \propto \exp(w^\top f(x, y))
\]

- Probability

\[
P(y|x, w) = \frac{\exp(w^\top f(x, y))}{\sum_{y'} \exp(w^\top f(x, y'))}
\]
Likelihood Objective

- **Model form:**

\[ P(y|x, w) = \frac{\exp(w^T f(x, y))}{\sum_{y'} \exp(w^T f(x, y'))} \]

- **Log-likelihood of training data**

\[ L(w) = \log \prod_i P(y_i^* | x_i, w) = \sum_i \log \left( \frac{\exp(w^T f(x_i, y_i^*))}{\sum_{y'} \exp(w^T f(x_i, y'))} \right) \]

\[ = \sum_i \left( w^T f(x_i, y_i^*) - \log \sum_{y'} \exp(w^T f(x_i, y')) \right) \]
Training
History of Training

- **1990’s**: Specialized methods (e.g. iterative scaling)

- **2000’s**: General-purpose methods (e.g. conjugate gradient)

- **2010’s**: Online methods (e.g. stochastic gradient)
What Does LL Look Like?

Example

- Data: xxxy
- Two outcomes, x and y
- One indicator for each
- Likelihood

\[
\log \left( \left( \frac{e^x}{e^x + e^y} \right)^3 \times \frac{e^y}{e^x + e^y} \right)
\]
- The maxent objective is an unconstrained convex problem

\[ L(w) \]

- One optimal value*, gradients point the way
Gradients

\[ L(w) = \sum_i \left( w^\top f(x_i, y_i^*) - \log \sum_y \exp(w^\top f(x_i, y)) \right) \]

\[ \frac{\partial L(w)}{\partial w} = \sum_i \left( f(x_i, y_i^*) - \sum_y P(y|x_i) f(x_i, y) \right) \]

Count of features under target labels

Expected count of features under model predicted label distribution
Gradient Ascent

- The maxent objective is an unconstrained optimization problem

\[ L(w) \]

- Gradient Ascent
  - Basic idea: move uphill from current guess
  - Gradient ascent / descent follows the gradient incrementally
  - At local optimum, derivative vector is zero
  - Will converge if step sizes are small enough, but not efficient
  - All we need is to be able to evaluate the function and its derivative
(Quasi)-Newton Methods

- 2nd-Order methods: repeatedly create a quadratic approximation and solve it

\[ L(w) \]

\[ L(w_0) + \nabla L(w)^\top (w - w_0) + (w - w_0)^\top \nabla^2 L(w)(w - w_0) \]

- E.g. LBFGS, which tracks derivative to approximate (inverse) Hessian
Regularization
Regularization Methods

- Early stopping
- L2: $L(w) - |w|^2$
- L1: $L(w) - |w|$
Regularization Effects

- Early stopping: don’t do this

- L2: weights stay small but non-zero

- L1: many weights driven to zero
  - Good for sparsity
  - Usually bad for accuracy for NLP
Scaling
Why is Scaling Hard?

\[ L(w) = \sum_i \left( w^\top f(x_i, y_i^*) - \log \sum_y \exp(w^\top f(x_i, y)) \right) \]

- Big normalization terms
- Lots of data points
Hierarchical Prediction

- Hierarchical prediction / softmax [Mikolov et al 2013]

- Noise-Contrastive Estimation [Mnih, 2013]

- Self-Normalization [Devlin, 2014]
Stochastic Gradient

- View the gradient as an average over data points

\[
\frac{\partial L(w)}{\partial w} = \frac{1}{N} \sum_i \left( f(x_i, y_i^*) - \sum_y P(y|x_i)f(x_i, y) \right)
\]

- Stochastic gradient: take a step each example (or mini-batch)

\[
\frac{\partial L(w)}{\partial w} \approx \frac{1}{1} \left( f(x_i, y_i^*) - \sum_y P(y|x_i)f(x_i, y) \right)
\]

- Substantial improvements exist, e.g. AdaGrad (Duchi, 11)
Log-linear Parameterization

- **Model form:**

  $$P(y|x; w) = \frac{\exp(w^\top f(x, y))}{\sum_{y'} \exp(w^\top f(x, y'))}$$

- **Learn by following gradient of training LL:**

  $$\frac{\partial L(w)}{\partial w} = \sum_i f(x_i, y_i^*) - \sum_i \left( \mathbb{E}_{P(y|x_i; w)} [f(x_i, y)] \right)$$
Mixed Interpolation

- But can’t we just interpolate:
  - $P(w|\text{most recent words})$
  - $P(w|\text{skip contexts})$
  - $P(w|\text{caching})$
  - ...

- Yes, and people do (well, did)
  - But additive combination tends to flatten distributions, not zero out candidates
Neural LMs
Neural LMs

\[ i-th \text{ output } = P(w_t = i \mid \text{context}) \]

Image: (Bengio et al, 03)
Neural vs Maxent

- **Maxent LM**

  \[ P(y|x; w) \propto \exp(w^\top f(x, y)) \]

- **Simple Neural LM**

  \[ P(y|x; A, B, C) \propto \exp \left( A_y^\top \sigma(B^\top C_x) \right) \]

  \( \sigma \) nonlinear, e.g. tanh
Neural LM Example

\[ P(y|x) \propto e^{(A^\top h)} \]

\[ h = \sigma (B^\top C_x) \]

\[ x_{-2} = \text{closing} \]

\[ x_{-1} = \text{the} \]
Neural LMs

\[ i\text{-th output} = P(w_t = i \mid \text{context}) \]

most computation here

\[ \text{softmax} \]

\[ \text{tanh} \]

\[ C(w_{t-n+1}) \]

Table look-up in \( C \)

index for \( w_{t-n+1} \)

\[ C(w_t-2) \]

Matrix \( C \)

shared parameters across words

index for \( w_{t-2} \)

\[ C(w_{t-1}) \]

index for \( w_{t-1} \)

Image: (Bengio et al, 03)
Decision Trees / Forests

- Decision trees?
  - Good for non-linear decision problems
  - Random forests can improve further [Xu and Jelinek, 2004]
  - Paths to leaves basically learn conjunctions
  - General contrast between DTs and linear models
Speech Signals
Speech in a Slide

- Frequency gives pitch; amplitude gives volume

- Frequencies at each time slice processed into observation vectors
Articulation
Articulatory System

Nasal cavity
Oral cavity
Pharynx
Vocal folds (in the larynx)
Trachea
Lungs

Sagittal section of the vocal tract (Techmer 1880)
Text from Ohala, Sept 2001, from Sharon Rose slide
Space of Phonemes

<table>
<thead>
<tr>
<th>Nasal</th>
<th>m</th>
<th>n</th>
<th>η</th>
<th>η</th>
<th>η</th>
<th>η</th>
<th>n</th>
</tr>
</thead>
<tbody>
<tr>
<td>Plosive</td>
<td>p</td>
<td>b</td>
<td>t</td>
<td>d</td>
<td>t</td>
<td>d</td>
<td>k</td>
</tr>
<tr>
<td>Fricative</td>
<td>f</td>
<td>v</td>
<td>θ</td>
<td>δ</td>
<td>s</td>
<td>z</td>
<td>s</td>
</tr>
<tr>
<td>Approximant</td>
<td>v</td>
<td>j</td>
<td>ʃ</td>
<td>ʃ</td>
<td>ʃ</td>
<td>ʃ</td>
<td>ʃ</td>
</tr>
<tr>
<td>Trill</td>
<td>B</td>
<td>r</td>
<td>ɾ</td>
<td>ɾ</td>
<td>ɾ</td>
<td>ɾ</td>
<td>ɾ</td>
</tr>
<tr>
<td>Tap, Flap</td>
<td>v</td>
<td>r</td>
<td>ɾ</td>
<td>ɾ</td>
<td>ɾ</td>
<td>ɾ</td>
<td>ɾ</td>
</tr>
<tr>
<td>Lateral fricative</td>
<td>ɻ</td>
<td>ɻ</td>
<td>ɻ</td>
<td>ɻ</td>
<td>ɻ</td>
<td>ɻ</td>
<td>ɻ</td>
</tr>
<tr>
<td>Lateral approximant</td>
<td>l</td>
<td>l</td>
<td>ɿ</td>
<td>ɿ</td>
<td>ɿ</td>
<td>ɿ</td>
<td>ɿ</td>
</tr>
<tr>
<td>Lateral flap</td>
<td>l</td>
<td>l</td>
<td>l</td>
<td>l</td>
<td>l</td>
<td>l</td>
<td>l</td>
</tr>
</tbody>
</table>

- Standard international phonetic alphabet (IPA) chart of consonants
Place
Places of Articulation

- labial
- dental
- alveolar
- post-alveolar/palatal
- velar
- uvular
- pharyngeal
- laryngeal/glottal

Figure thanks to Jennifer Venditti
Labial place

Bilabial: p, b, m
Labiodental: f, v

Figure thanks to Jennifer Venditti
Coronal place

Dental:
- th/dh

Alveolar:
- t/d/s/z/l/n

Post:
- sh/zh/y

Figure thanks to Jennifer Venditti
Dorsal Place

Velar:
k/g/ng

Figure thanks to Jennifer Venditti
### Space of Phonemes

<table>
<thead>
<tr>
<th>LABIAL</th>
<th>CORONAL</th>
<th>DORSAL</th>
<th>RADICAL</th>
<th>LARYNGEAL</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bilabial</td>
<td>Labio- dental</td>
<td>Dental</td>
<td>Palato-alveolar</td>
<td>Retroflex</td>
</tr>
<tr>
<td>Nasal</td>
<td>m n</td>
<td>n n</td>
<td>n n</td>
<td>n n</td>
</tr>
<tr>
<td>Plosive</td>
<td>p b</td>
<td>t d</td>
<td>t d</td>
<td>c j</td>
</tr>
<tr>
<td>Fricative</td>
<td>f v</td>
<td>θ s z</td>
<td>θ s z</td>
<td>c j</td>
</tr>
<tr>
<td>Approximant</td>
<td>u j</td>
<td>j j</td>
<td>j j</td>
<td>j j</td>
</tr>
<tr>
<td>Trill</td>
<td>B r</td>
<td>r r</td>
<td>r r</td>
<td>r r</td>
</tr>
<tr>
<td>Tap, Flap</td>
<td>v r</td>
<td>r r</td>
<td>r r</td>
<td>r r</td>
</tr>
<tr>
<td>Lateral fricative</td>
<td>ɹ ɹ</td>
<td>ɹ ɹ</td>
<td>ɹ ɹ</td>
<td>ɹ ɹ</td>
</tr>
<tr>
<td>Lateral approximant</td>
<td>ɹ ɹ</td>
<td>ɹ ɹ</td>
<td>ɹ ɹ</td>
<td>ɹ ɹ</td>
</tr>
<tr>
<td>Lateral flap</td>
<td>ɹ ɹ</td>
<td>ɹ ɹ</td>
<td>ɹ ɹ</td>
<td>ɹ ɹ</td>
</tr>
</tbody>
</table>

- **Standard international phonetic alphabet (IPA) chart of consonants**
Manner
Manner of Articulation

- In addition to varying by place, sounds vary by manner

- Stop: complete closure of articulators, no air escapes via mouth
  - Oral stop: palate is raised (p, t, k, b, d, g)
  - Nasal stop: oral closure, but palate is lowered (m, n, ng)

- Fricatives: substantial closure, turbulent: (f, v, s, z)

- Approximants: slight closure, sonorant: (l, r, w)

- Vowels: no closure, sonorant: (i, e, a)
### Standard international phonetic alphabet (IPA) chart of consonants

<table>
<thead>
<tr>
<th>LABIAL</th>
<th>CORONAL</th>
<th>DORSAL</th>
<th>RADICAL</th>
<th>LARYNGEAL</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bilabial</td>
<td>Labiodental</td>
<td>Dental</td>
<td>Alveolar</td>
<td>Palato-alveolar</td>
</tr>
<tr>
<td>Nasal</td>
<td>m</td>
<td>n</td>
<td>η</td>
<td>ƞ</td>
</tr>
<tr>
<td>Plosive</td>
<td>p</td>
<td>b</td>
<td>t</td>
<td>d</td>
</tr>
<tr>
<td>Fricative</td>
<td>f β</td>
<td>f ν</td>
<td>θ δ</td>
<td>s z</td>
</tr>
<tr>
<td>Approximant</td>
<td>v</td>
<td>j</td>
<td>j</td>
<td>w</td>
</tr>
<tr>
<td>Trill</td>
<td>B</td>
<td>r</td>
<td></td>
<td>R</td>
</tr>
<tr>
<td>Tap, Flap</td>
<td>v</td>
<td>t</td>
<td></td>
<td>R</td>
</tr>
<tr>
<td>Lateral fricative</td>
<td>l l</td>
<td>t t</td>
<td>θ θ</td>
<td>θ θ</td>
</tr>
<tr>
<td>Lateral approximant</td>
<td>l l</td>
<td>l l</td>
<td>l l</td>
<td>L L</td>
</tr>
<tr>
<td>Lateral flap</td>
<td>l l</td>
<td>l l</td>
<td>l l</td>
<td>L L</td>
</tr>
</tbody>
</table>
Vowels
Vowel Space

Vowels at right & left of bullets are rounded & unrounded.
Acoustics
“She just had a baby”

- What can we learn from a wavefile?
  - No gaps between words (!)
  - Vowels are voiced, long, loud
  - Length in time = length in space in waveform picture
  - Voicing: regular peaks in amplitude
  - When stops closed: no peaks, silence
  - Peaks = voicing: .46 to .58 (vowel [iy], from second .65 to .74 (vowel [ax]) and so on
  - Silence of stop closure (1.06 to 1.08 for first [b], or 1.26 to 1.28 for second [b])
  - Fricatives like [sh]: intense irregular pattern; see .33 to .46
Time-Domain Information

Example from Ladefoged
Simple Periodic Waves of Sound

- Y axis: Amplitude = amount of air pressure at that point in time
  - Zero is normal air pressure, negative is rarefaction
- X axis: Time.
- Frequency = number of cycles per second.
- 20 cycles in .02 seconds = 1000 cycles/second = 1000 Hz
Complex Waves: 100Hz + 1000Hz
Spectrum

Frequency components (100 and 1000 Hz) on x-axis
- Note complex wave repeating nine times in figure
- Plus smaller waves which repeats 4 times for every large pattern
- Large wave has frequency of 250 Hz (9 times in 0.036 seconds)
- Small wave roughly 4 times this, or roughly 1000 Hz
- Two little tiny waves on top of peak of 1000 Hz waves
Spectrum of an Actual Soundwave
Spectrum represents these frequency components.

Composed by Fourier transform, an algorithm which separates out each frequency component of a wave.

- x-axis shows frequency, y-axis shows magnitude (in decibels, a log measure of amplitude)
- Peaks at 930 Hz, 1860 Hz, and 3020 Hz.
Source / Channel
Why these Peaks?

- **Articulation process:**
  - The vocal cord vibrations create harmonics
  - The mouth is an amplifier
  - Depending on shape of mouth, some harmonics are amplified more than others
Vowel [i] at increasing pitches

Figures from Ratree Wayland
Resonances of the Vocal Tract

- The human vocal tract as an open tube:

  ![Diagram of the vocal tract](image)

  - Closed end
  - Open end

  Length 17.5 cm.

- Air in a tube of a given length will tend to vibrate at resonance frequency of tube.

- Constraint: Pressure differential should be maximal at (closed) glottal end and minimal at (open) lip end.

Figure from W. Barry
Computing the 3 Formants of Schwa

Let the length of the tube be L

- \( F_1 = c/\lambda_1 = c/(4L) = \frac{35,000}{4 \times 17.5} = 500\text{Hz} \)
- \( F_2 = c/\lambda_2 = c/(4/3L) = \frac{3c}{4L} = \frac{3 \times 35,000}{4 \times 17.5} = 1500\text{Hz} \)
- \( F_3 = c/\lambda_3 = c/(4/5L) = \frac{5c}{4L} = \frac{5 \times 35,000}{4 \times 17.5} = 2500\text{Hz} \)

So we expect a neutral vowel to have 3 resonances at 500, 1500, and 2500 Hz

These vowel resonances are called formants
Seeing Formants: the Spectrogram

[Diagram showing spectrograms for different vowels: i, I, ɛ, æ, a, ɔ, ʊ, u]
Vowel Space

Front  Near front  Central  Near back  Back

Close  i  y  i  u  u

Near close

Close mid  e  ø  e  ø  ø  ø  ø

Mid

Open mid  æ  ø  æ  æ  æ  æ

Near open

Open

Vowels at right & left of bullets are rounded & unrounded.
Spectrograms
How to Read Spectrograms

- **[bab]**: closure of lips lowers all formants: so rapid increase in all formants at beginning of "bab"
- **[dad]**: first formant increases, but F2 and F3 slight fall
- **[gag]**: F2 and F3 come together: this is a characteristic of velars. Formant transitions take longer in velars than in alveolars or labials

From Ladefoged “A Course in Phonetics”
“She came back and started again”

1. lots of high-freq energy  
3.  
4.  
5.  
6.  
7.  
8.  
9. From Ladefoged “A Course in Phonetics”
Deriving Schwa

- Reminder of basic facts about sound waves
  - \( f = \frac{c}{\lambda} \)
  - \( c \) = speed of sound (approx 35,000 cm/sec)
  - A sound with \( \lambda = 10 \) meters: \( f = 35 \) Hz \((35,000/1000)\)
  - A sound with \( \lambda = 2 \) centimeters: \( f = 17,500 \) Hz \((35,000/2)\)
American English Vowel Space

Figures from Jennifer Venditti, H. T. Bunnell
Dialect Issues

- Speech varies from dialect to dialect (examples are American vs. British English)
  - Syntactic ("I could" vs. "I could do")
  - Lexical ("elevator" vs. "lift")
  - Phonological
  - Phonetic

- Mismatch between training and testing dialects can cause a large increase in error rate