Maxent Models and Discriminative Estimation

The maximum entropy model presentation

Maximum Entropy Models

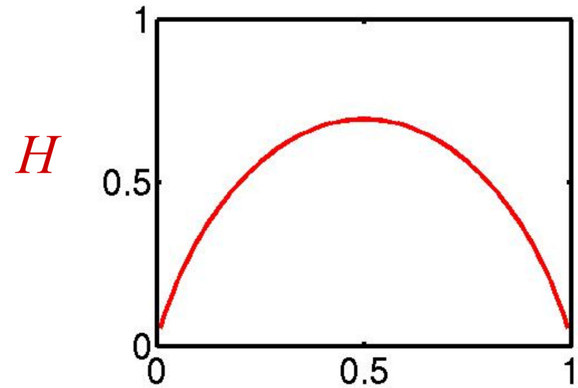
- An equivalent approach:
  - Lots of distributions out there, most of them very spiked, specific, overfit.
  - We want a distribution which is uniform except in specific ways we require.
  - Uniformity means high entropy – we can search for distributions which have properties we desire, but also have high entropy.

Ignorance is preferable to error and he is less remote from the truth who believes nothing than he who believes what is wrong – Thomas Jefferson (1781)

(Maximum) Entropy

- Entropy: the uncertainty of a distribution.
- Quantifying uncertainty ("surprise"):
  - Event x
  - Probability p_x
  - "Surprise" $\log(1/p_x)$
- Entropy: expected surprise (over p):

$$\mathrm{H}(p) = E_p\left[\log_2 \frac{1}{p_x}\right] = -\sum_x p_x \log_2 p_x$$



A coin-flip is most uncertain for a fair coin.

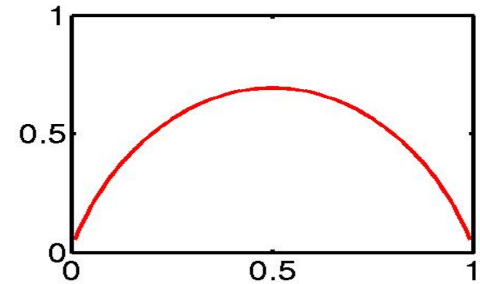
Christopher Manning

Maxent Examples I

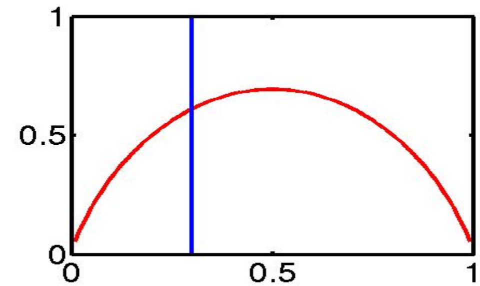
- What do we want from a distribution?
  - Minimize commitment = maximize entropy.
  - Resemble some reference distribution (data).
- Solution: maximize entropy *H*, subject to feature-based constraints:

$$E_p\left[f_i\right] = E_{\hat{p}}\left[f_i\right] \Leftrightarrow \sum_{x \in f_i} p_x = C_i$$

- Adding constraints (features):
  - Lowers maximum entropy
  - Raises maximum likelihood of data
  - Brings the distribution further from uniform
  - Brings the distribution closer to data



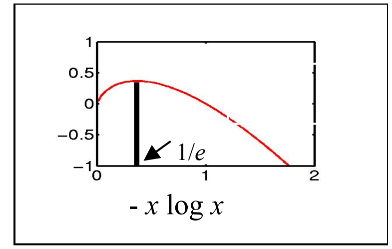
Unconstrained, max at 0.5



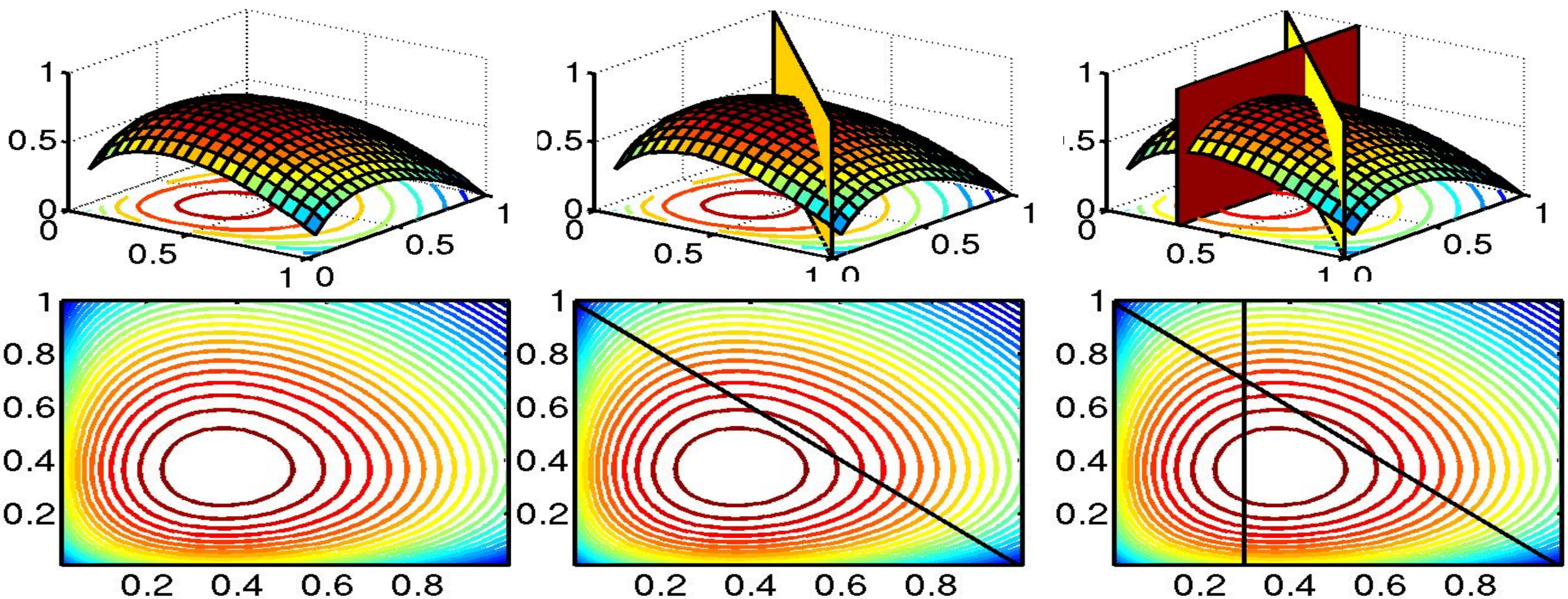
Constraint that $p_{\text{HEADS}} = 0.3$

Christopher Manning

Maxent Examples II



$H(p_H p_T,)$ $\quad$ $p_H + p_T = 1$ $\quad$ $p_H = 0.3$



Christopher Manning

Maxent Examples III

- Let's say we have the following event space:

NN	NNS	NNP	NNPS	VBZ	VBD

- … and the following empirical data:

3	5	11	13	3	1

- Maximize H:

$1/e$	$1/e$	$1/e$	$1/e$	$1/e$	$1/e$

- … want probabilities: E[NN,NNS,NNP,NNPS,VBZ,VBD] = 1

1/6	1/6	1/6	1/6	1/6	1/6

Maxent Examples IV

- Too uniform!
- N* are more common than V*, so we add the feature f_N = {NN, NNS, NNP, NNPS}, with E[f_N] =32/36

NN	NNS	NNP	NNPS	VBZ	VBD
8/36	8/36	8/36	8/36	2/36	2/36

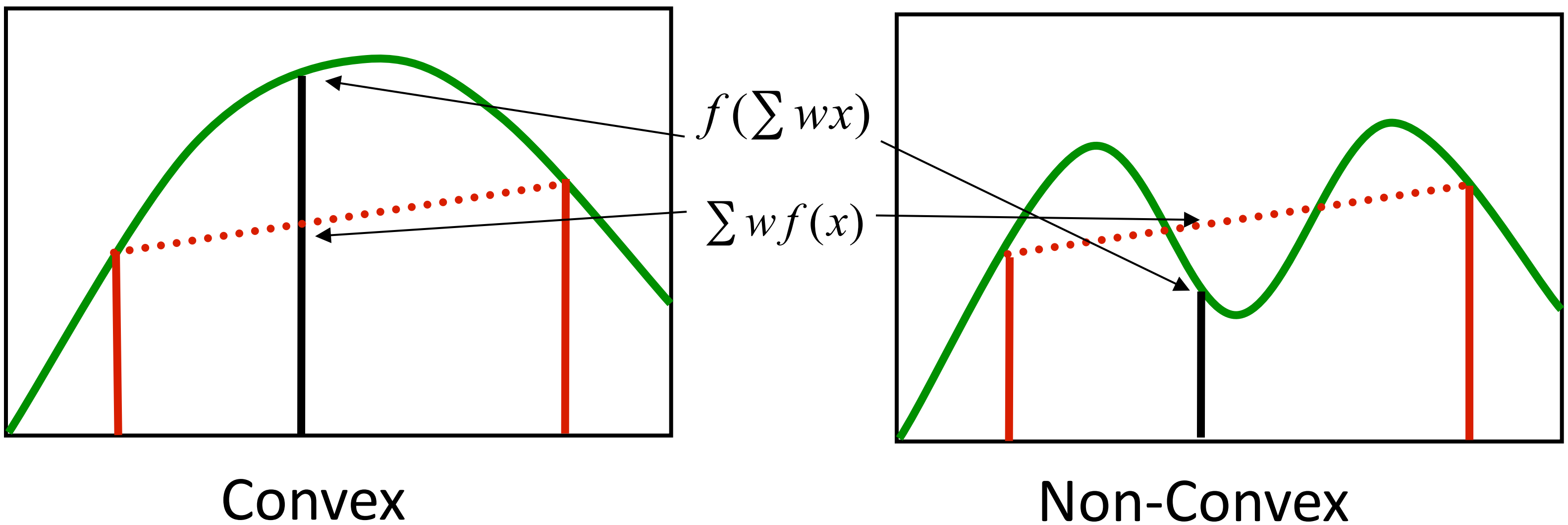
- … and proper nouns are more frequent than common nouns, so we add f_P = {NNP, NNPS}, with E[f_P] =24/36

4/36	4/36	12/36	12/36	2/36	2/36

- … we could keep refining the models, e.g., by adding a feature to distinguish singular vs. plural nouns, or verb types.

Convexity

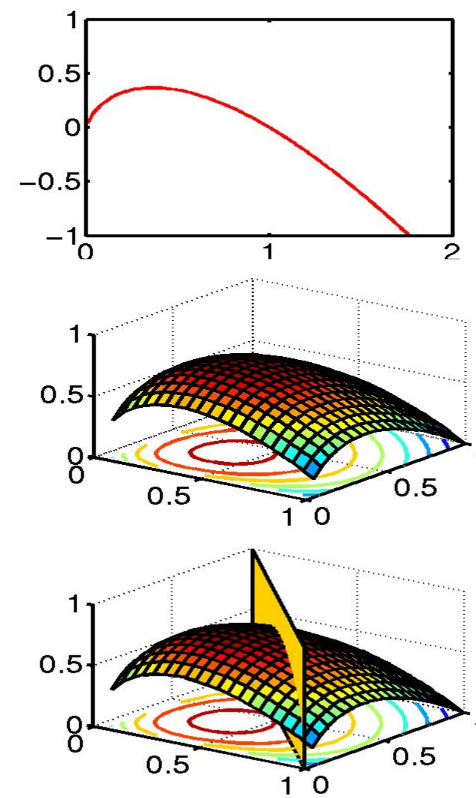
$$f\left(\sum_i w_i x_i\right) \geq \sum_i w_i f(x_i) \qquad \sum_i w_i = 1$$



Convexity guarantees a single, global maximum because any higher points are greedily reachable.

Convexity II

- Constrained $H(p) = -\sum x \log x$ is convex:
  - $-x \log x$ is convex
  - $-\sum x \log x$ is convex (sum of convex functions is convex).
  - The feasible region of constrained H is a linear subspace (which is convex)
  - The constrained entropy surface is therefore convex.
- The maximum likelihood exponential model (dual) formulation is also convex.



Maxent Models and Discriminative Estimation

The maximum entropy model presentation

Feature Overlap/ Feature Interaction

How overlapping features work in maxent models

Feature Overlap

- Maxent models handle overlapping features well.
- Unlike a NB model, there is no double counting!

Empirical

	A	a
B	2	1
b	2	1

	A	a
B		
b		

All = 1

	A	a
B	1/4	1/4
b	1/4	1/4

	A	a
B		
b		

	A	a
B		
b		

A = 2/3

	A	a
B	1/3	1/6
b	1/3	1/6

	A	a
B	λ_A	
b	λ_A	

	A	a
B		
b		

A = 2/3

	A	a
B	1/3	1/6
b	1/3	1/6

	A	a
B	$\lambda'_A+\lambda''_A$	
b	$\lambda'_A+\lambda''_A$	

Example: Named Entity Feature Overlap

Grace is correlated with PERSON, but does not add much evidence on top of already knowing prefix features.

Local Context

	Prev	Cur	Next
State	Other	???	???
Word	at	Grace	Road
Tag	IN	NNP	NNP
Sig	x	Xx	Xx

Feature Weights

Feature Type	Feature	PERS	LOC
Previous word	*at*	-0.73	0.94
Current word	*Grace*	0.03	0.00
Beginning bigram	<G	0.45	-0.04
Current POS tag	NNP	0.47	0.45
Prev and cur tags	IN NNP	-0.10	0.14
Previous state	Other	-0.70	-0.92
Current signature	Xx	0.80	0.46
Prev state, cur sig	O-Xx	0.68	0.37
Prev-cur-next sig	x-Xx-Xx	-0.69	0.37
P. state - p-cur sig	O-x-Xx	-0.20	0.82
…			
Total:		**-0.58**	**2.68**

Christopher Manning

Feature Interaction

- Maxent models handle overlapping features well, but do not automatically model feature interactions.

Empirical

	A	a
B	1	1
b	1	0

	A	a
B		
b		

All = 1

	A	a
B	1/4	1/4
b	1/4	1/4

	A	a
B	0	0
b	0	0

	A	a
B		
b		

A = 2/3

	A	a
B	1/3	1/6
b	1/3	1/6

	A	a
B	λ_A	
b	λ_A	

	A	a
B		
b		

B = 2/3

	A	a
B	4/9	2/9
b	2/9	1/9

	A	a
B	$\lambda_A+\lambda_B$	λ_B
b	λ_A	

Feature Interaction

- If you want interaction terms, you have to add them:

Empirical

	A	a
B	1	1
b	1	0

	A	a
B		
b		

A = 2/3

	A	a
B	1/3	1/6
b	1/3	1/6

	A	a
B		
b		

B = 2/3

	A	a
B	4/9	2/9
b	2/9	1/9

	A	a
B		
b		

AB = 1/3

	A	a
B	1/3	1/3
b	1/3	0

- A disjunctive feature would also have done it (alone):

	A	a
B		
b		

	A	a
B	1/3	1/3
b	1/3	0

Christopher Manning

Quiz Question

- Suppose we have a 1 feature maxent model built over observed data as shown.
- What is the constructed model's probability distribution over the four possible outcomes?

Empirical

	A	a
B	2	1
b	2	1

Features

	A	a
B		
b		

Expectations

Probabilities

	A	a
B		
b		

Feature Interaction

- For loglinear/logistic regression models in statistics, it is standard to do a greedy stepwise search over the space of all possible interaction terms.
- This combinatorial space is exponential in size, but that’s okay as most statistics models only have 4–8 features.
- In NLP, our models commonly use hundreds of thousands of features, so that’s not okay.
- Commonly, interaction terms are added by hand based on linguistic intuitions.

Example: NER Interaction

Previous-state and current-signature have interactions, e.g. P=PERS-C=Xx indicates C=PERS much more strongly than C=Xx and P=PERS independently.

This feature type allows the model to capture this interaction.

Local Context

	Prev	Cur	Next
State	Other	???	???
Word	at	Grace	Road
Tag	IN	NNP	NNP
Sig	x	Xx	Xx

Feature Weights

Feature Type	Feature	PERS	LOC
Previous word	*at*	-0.73	0.94
Current word	*Grace*	0.03	0.00
Beginning bigram	*<G*	0.45	-0.04
Current POS tag	NNP	0.47	0.45
Prev and cur tags	IN NNP	-0.10	0.14
Previous state	Other	-0.70	-0.92
Current signature	Xx	0.80	0.46
Prev state, cur sig	O-Xx	0.68	0.37
Prev-cur-next sig	x-Xx-Xx	-0.69	0.37
P. state - p-cur sig	O-x-Xx	-0.20	0.82
…			
Total:		**-0.58**	**2.68**

Feature Overlap/ Feature Interaction

How overlapping features work in maxent models

Conditional Maxent Models for Classification

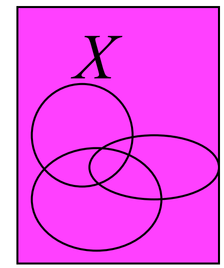
The relationship between conditional and joint maxent/ exponential models

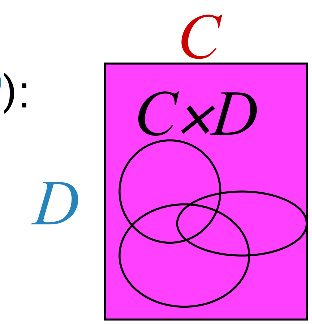
Classification

- What do these joint models of $\mathrm{P}(X)$ have to do with conditional models $\mathrm{P}(C|D)$?
- Think of the space $C \times D$ as a complex X.
  - C is generally small (e.g., 2-100 topic classes)
  - D is generally huge (e.g., space of documents)
- We can, in principle, build models over $\mathrm{P}(C,D)$.
- This will involve calculating expectations of features (over $C \times D$):

$$E(f_i) = \sum_{(c,d) \in (C,D)} P(c,d) f_i(c,d)$$

- Generally impractical: can't enumerate X efficiently.

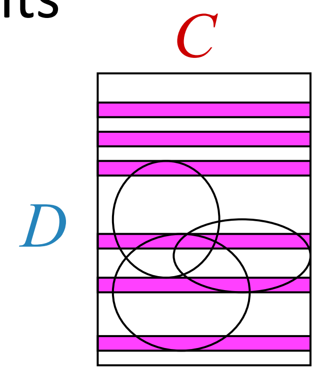




Classification II

- *D* may be huge or infinite, but only a few *d* occur in our data.
- What if we add one feature for each *d* and constrain its expectation to match our empirical data?

$$\forall(d) \in D \quad P(d) = \hat{P}(d)$$



- Now, most entries of $P(c,d)$ will be zero.
- We can therefore use the much easier sum:

$$E(f_i) = \sum_{(c,d)\in(C,D)} P(c,d) f_i(c,d)$$
$$= \sum_{(c,d)\in(C,D)\wedge \hat{P}(d)>0} P(c,d) f_i(c,d)$$

Christopher Manning

Classification III

- But if we've constrained the *D* marginals

$$\forall(d) \in D \quad P(d) = \hat{P}(d)$$

- then the only thing that can vary is the conditional distributions:

$$P(c,d) = P(c \mid d)P(d)$$

$$= P(c \mid d)\hat{P}(d)$$

Classification IV

- This is the connection between joint and conditional maxent / exponential models:
  - Conditional models can be thought of as joint models with marginal constraints.
- Maximizing joint likelihood and conditional likelihood of the data in this model are equivalent!

Conditional Maxent Models for Classification

The relationship between conditional and joint maxent/ exponential models

Smoothing/Priors/ Regularization for Maxent Models

Smoothing: Issues of Scale

- Lots of features:
  - NLP maxent models can have well over a million features.
  - Even storing a single array of parameter values can have a substantial memory cost.

- Lots of sparsity:
  - Overfitting very easy – we need smoothing!
  - Many features seen in training will never occur again at test time.

- Optimization problems:
  - Feature weights can be infinite, and iterative solvers can take a long time to get to those infinities.

Christopher Manning

Smoothing: Issues

- Assume the following empirical distribution:

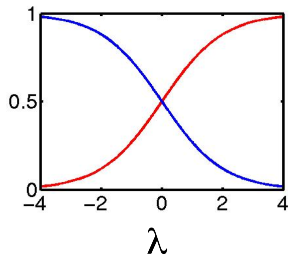
Heads	Tails
h	t

- Features: {Heads}, {Tails}
- We'll have the following model distribution:

$$p_{\mathrm{HEADS}} = \frac{e^{\lambda_{\mathrm{H}}}}{e^{\lambda_{\mathrm{H}}} + e^{\lambda_{\mathrm{T}}}} \qquad p_{\mathrm{TAILS}} = \frac{e^{\lambda_{\mathrm{T}}}}{e^{\lambda_{\mathrm{H}}} + e^{\lambda_{\mathrm{T}}}}$$

- Really, only one degree of freedom ($\lambda = \lambda_H - \lambda_T$)

$$p_{\mathrm{HEADS}} = \frac{e^{\lambda_{\mathrm{H}}} e^{-\lambda_{\mathrm{T}}}}{e^{\lambda_{\mathrm{H}}} e^{-\lambda_{\mathrm{T}}} + e^{\lambda_{\mathrm{T}}} e^{-\lambda_{\mathrm{T}}}} = \frac{e^{\lambda}}{e^{\lambda} + e^{0}} \qquad p_{\mathrm{TAILS}} = \frac{e^{0}}{e^{\lambda} + e^{0}}$$

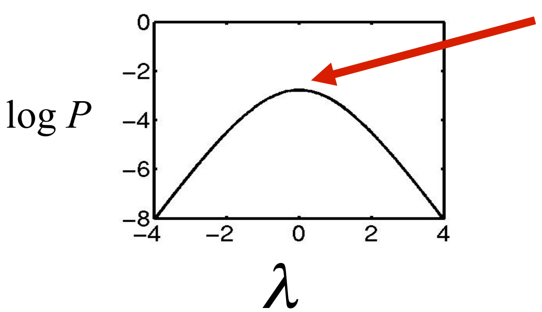


Smoothing: Issues

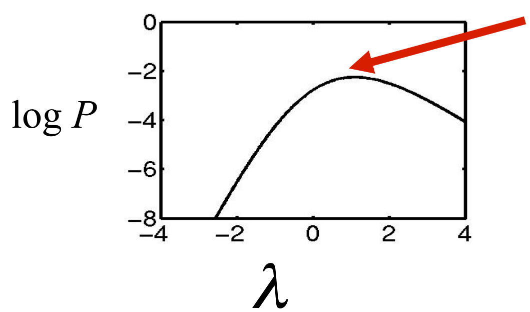
- The data likelihood in this model is:

$$\log P(h,t \mid \lambda) = h \log p_{\text{HEADS}} + t \log p_{\text{TAILS}}$$

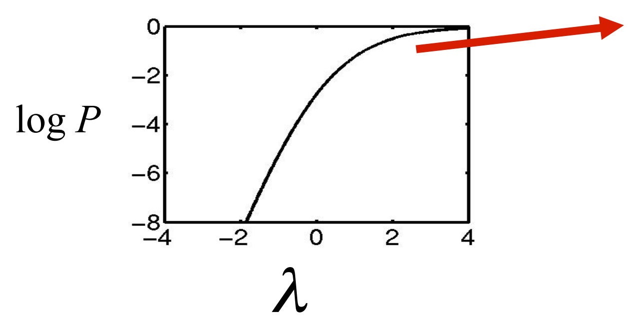
$$\log P(h,t \mid \lambda) = h\lambda - (t+h)\log(1+e^{\lambda})$$



Heads	Tails
2	2



Heads	Tails
3	1

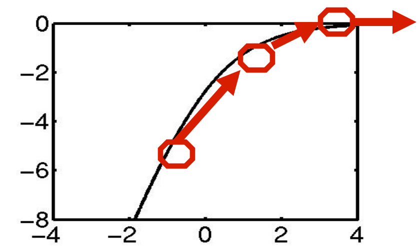


Heads	Tails
4	0

Christopher Manning

Smoothing: Early Stopping

- In the 4/0 case, there were two problems:
  - The optimal value of λ was ∞, which is a long trip for an optimization procedure.
  - The learned distribution is just as spiked as the empirical one – no smoothing.
- One way to solve both issues is to just stop the optimization early, after a few iterations.
  - The value of λ will be finite (but presumably big).
  - The optimization won't take forever (clearly).
  - Commonly used in early maxent work.



λ

Heads	Tails
4	0

Input

Heads	Tails
1	0

Output

Smoothing: Priors (MAP)

- What if we had a prior expectation that parameter values wouldn't be very large?
- We could then balance evidence suggesting large parameters (or infinite) against our prior.
- The evidence would never totally defeat the prior, and parameters would be smoothed (and kept finite!).
- We can do this explicitly by changing the optimization objective to maximum posterior likelihood:

$$\log P(C,\lambda \mid D) = \log P(\lambda) + \log P(C \mid D,\lambda)$$

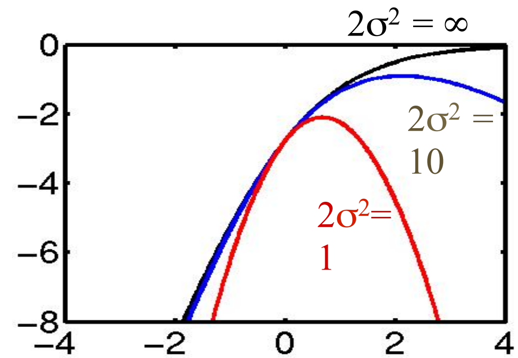
Posterior Prior Evidence

Smoothing: Priors

- Gaussian, or quadratic, or L_2 priors:
  - Intuition: parameters shouldn't be large.
  - Formalization: prior expectation that each parameter will be distributed according to a gaussian with mean μ and variance σ^2.

$$P(\lambda_i) = \frac{1}{\sigma_i \sqrt{2\pi}} \exp\left(-\frac{(\lambda_i - \mu_i)^2}{2\sigma_i^2} \right)$$

  - Penalizes parameters for drifting to far from their mean prior value (usually μ=0).
  - $2\sigma^2$=1 works surprisingly well.



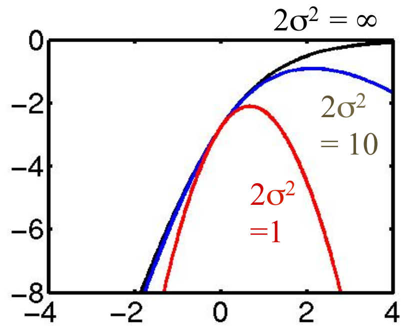
They don't even capitalize my name anymore!

Smoothing: Priors

- If we use gaussian priors:
  - Trade off some expectation-matching for smaller parameters.
  - When multiple features can be recruited to explain a data point, the more common ones generally receive more weight.
  - Accuracy generally goes up!
- Change the objective:

$$\log P(C,\lambda \mid D) = \log P(C \mid D,\lambda) - \log P(\lambda)$$

$$\log P(C,\lambda \mid D) = \sum_{(c,d)\in(C,D)} P(c \mid d,\lambda) - \sum_i \frac{(\lambda_i - \mu_i)^2}{2\sigma_i^2} + k$$



- Change the derivative:

$$\partial \log P(C,\lambda \mid D) / \partial\lambda_i = \text{actual}(f_i, C) - \text{predicted}(f_i,\lambda) - (\lambda_i - \mu_i)/\sigma^2$$

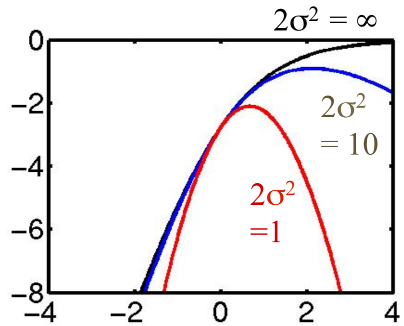
Christopher Manning

Smoothing: Priors

- If we use gaussian priors:
  - Trade off some expectation-matching for smaller parameters.
  - When multiple features can be recruited to explain a data point, the more common ones generally receive more weight.
  - Accuracy generally goes up!
- Change the objective:

$$\log P(C,\lambda \mid D) = \log P(C \mid D,\lambda) - \log P(\lambda)$$

$$\log P(C,\lambda \mid D) = \sum_{(c,d)\in(C,D)} P(c \mid d,\lambda) - \sum_i \frac{\lambda_i^2}{2\sigma_i^2} + k$$



- Change the derivative:

$$\partial \log P(C,\lambda \mid D) / \partial\lambda_i = \text{actual}(f_i, C) - \text{predicted}(f_i,\lambda) - \lambda_i / \sigma^2$$

Taking prior mean as 0

Example: NER Smoothing

Because of smoothing, the more common prefix and single-tag features have larger weights even though entire-word and tag-pair features are more specific.

Local Context

	Prev	Cur	Next
State	Other	???	???
Word	at	Grace	Road
Tag	IN	NNP	NNP
Sig	x	Xx	Xx

Feature Weights

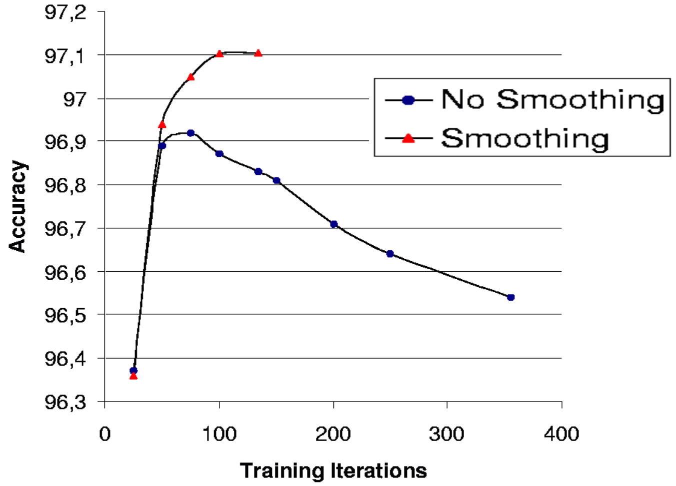
Feature Type	Feature	PERS	LOC
Previous word	*at*	-0.73	0.94
Current word	*Grace*	0.03	0.00
Beginning bigram	*<G*	0.45	-0.04
Current POS tag	NNP	0.47	0.45
Prev and cur tags	IN NNP	-0.10	0.14
Previous state	Other	-0.70	-0.92
Current signature	Xx	0.80	0.46
Prev state, cur sig	O-Xx	0.68	0.37
Prev-cur-next sig	x-Xx-Xx	-0.69	0.37
P. state - p-cur sig	O-x-Xx	-0.20	0.82
…			
Total:		**-0.58**	**2.68**

Christopher Manning

Example: POS Tagging

- From (Toutanova et al., 2003):

	Overall Accuracy	Unknown Word Acc
Without Smoothing	96.54	85.20
With Smoothing	97.10	88.20



- Smoothing helps:
  - Softens distributions.
  - Pushes weight onto more explanatory features.
  - Allows many features to be dumped safely into the mix.
  - Speeds up convergence (if both are allowed to converge)!

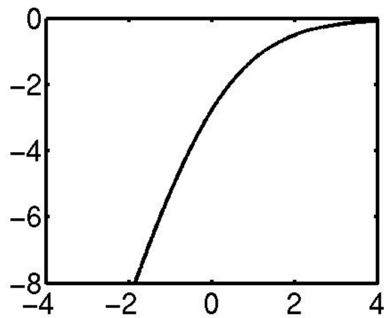
Smoothing: Regularization

- Talking of “priors” and “MAP estimation” is Bayesian language
- In frequentist statistics, people will instead talk about using “regularization”, and in particular, a gaussian prior is “L_2 regularization”

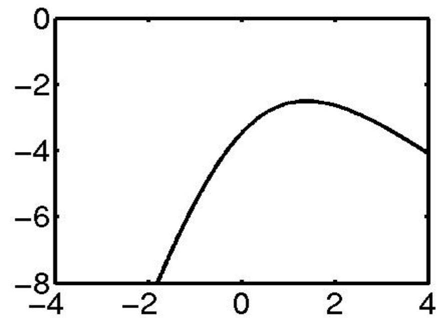
- The choice of names makes no difference to the math

Smoothing: Virtual Data

- Another option: smooth the data, not the parameters.
- Example:



Heads	Tails
4	0



Heads	Tails
5	1

  - Equivalent to adding two extra data points.
  - Similar to add-one smoothing for generative models.
- Hard to know what artificial data to create!

Smoothing: Count Cutoffs

- In NLP, features with low empirical counts are often dropped.
  - Very weak and indirect smoothing method.
  - Equivalent to locking their weight to be zero.
  - Equivalent to assigning them gaussian priors with mean zero and variance zero.
  - Dropping low counts does remove the features which were most in need of smoothing…
  - … and speeds up the estimation by reducing model size …
  - … but count cutoffs generally hurt accuracy in the presence of proper smoothing.
- We recommend: don't use count cutoffs unless absolutely necessary for memory usage reasons.

Smoothing/Priors/ Regularization for Maxent Models