

Probability and Statistics

January 17, 2013

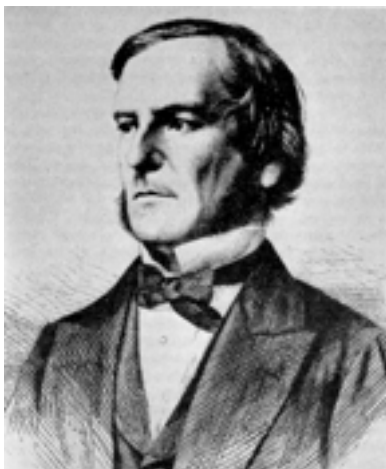


Last time ...

- 1) Formulate a *model* of pairs of sentences.
- 2) *Learn* an instance of the model from *data*.
- 3) Use it to *infer* translations of new inputs.

Why Probability?

- Probability formalizes ...
 - the concept of **models**
 - the concept of **data**
 - the concept of **learning**
 - the concept of **inference** (prediction)



Probability is expectation founded upon partial knowledge.

$p(x \mid \text{partial knowledge})$

“Partial knowledge” is an apt description of what we know about language and translation!

Probability skeptics: please bare with me!

Probability Models

- Key components of a probability model
 - The space of events (Ω or \mathcal{S})
 - The assumptions about conditional independence / dependence among events
 - Functions assigning probability (density) to events
 - *We will assume discrete distributions.*

Events and Random Variables

A **random variable** is a function from a random event from a set of possible outcomes (Ω) and a probability distribution (ρ), a function from outcomes to probabilities.

$$\Omega = \{1, 2, 3, 4, 5, 6\}$$



$$X(\omega) = \omega$$

$$\rho_X(x) = \begin{cases} \frac{1}{6} & \text{if } x = 1, 2, 3, 4, 5, 6 \\ 0 & \text{otherwise} \end{cases}$$

Events and Random Variables

A **random variable** is a function from a random event from a set of possible outcomes (Ω) and a probability distribution (ρ), a function from outcomes to probabilities.

$$\Omega = \{1, 2, 3, 4, 5, 6\}$$

$$Y(\omega) = \begin{cases} 0 & \text{if } \omega \in \{2, 4, 6\} \\ 1 & \text{otherwise} \end{cases}$$

$$\rho_Y(y) = \begin{cases} \frac{1}{2} & \text{if } y = 0, 1 \\ 0 & \text{otherwise} \end{cases}$$



What is our event space?

**What are our random
variables?**

Probability Distributions

A probability distribution (ρ_X) assigns probabilities to the values of a random variable (X).

There are a couple of philosophically different ways to define probabilities, but we will give only the invariants in terms of **random variables**.

$$\sum_{x \in \mathcal{X}} \rho_X(x) = 1$$

$$\rho_X(x) \geq 0 \quad \forall x \in \mathcal{X}$$


Probability distributions of a random variable may be specified in a number of ways.

Specifying Distributions

- Engineering/mathematical convenience
- Important techniques in this course
 - Probability mass functions
 - Tables (“stupid multinomials”)
 - Log-linear (max-ent, random field, multinomial logit) parameterizations
 - **Construct random variables from other r.v.’s with known distributions**
- Important in general
 - Cumulative distribution functions
 - Characteristic functions

Sampling Notation

$$x = 4 \times z + 1.7$$


Variable

Expression

Sampling Notation

$$x = 4 \times z + 1.7$$

$$y \sim \text{Distribution}(\theta)$$

Distribution

Random variable



Parameter



Sampling Notation

$$x = 4 \times z + 1.7$$

$$y \sim \text{Distribution}(\theta)$$

$$y' = y \times x$$

Multivariate r.v.'s

Probability theory is particularly useful because it lets us reason about (cor)related and dependent events.

A **joint probability distribution** is a probability distribution over r.v.'s with the following form:

$$Z = \begin{bmatrix} X(\omega) \\ Y(\omega) \end{bmatrix}$$

$$\sum_{x \in \mathcal{X}, y \in \mathcal{Y}} \rho_Z \left(\begin{bmatrix} x \\ y \end{bmatrix} \right) = 1 \quad \rho_Z \left(\begin{bmatrix} x \\ y \end{bmatrix} \right) \geq 0 \quad \forall x \in \mathcal{X}, y \in \mathcal{Y}$$

$$\Omega = \{1, 2, 3, 4, 5, 6\}$$



$$X(\omega) = \omega$$

$$\begin{aligned} \Omega = \{ & (1, 1), (1, 2), (1, 3), (1, 4), (1, 5), (1, 6), \\ & (2, 1), (2, 2), (2, 3), (2, 4), (2, 5), (2, 6), \\ & (3, 1), (3, 2), (3, 3), (3, 4), (3, 5), (3, 6), \\ & (4, 1), (4, 2), (4, 3), (4, 4), (4, 5), (4, 6), \\ & (5, 1), (5, 2), (5, 3), (5, 4), (5, 5), (5, 6), \\ & (6, 1), (6, 2), (6, 3), (6, 4), (6, 5), (6, 6), \} \end{aligned}$$



$$X(\omega) = \omega_1 \quad Y(\omega) = \omega_2$$

$$\rho_{X,Y}(x, y) = \begin{cases} \frac{1}{36} & \text{if } (x, y) \in \Omega \\ 0 & \text{otherwise} \end{cases}$$

$$\Omega = \{1, 2, 3, 4, 5, 6\}$$



$$X(\omega) = \omega$$

$$\begin{aligned} \Omega = \{ & (1, 1), (1, 2), (1, 3), (1, 4), (1, 5), (1, 6), \\ & (2, 1), (2, 2), (2, 3), (2, 4), (2, 5), (2, 6), \\ & (3, 1), (3, 2), (3, 3), (3, 4), (3, 5), (3, 6), \\ & (4, 1), (4, 2), (4, 3), (4, 4), (4, 5), (4, 6), \\ & (5, 1), (5, 2), (5, 3), (5, 4), (5, 5), (5, 6), \\ & (6, 1), (6, 2), (6, 3), (6, 4), (6, 5), (6, 6), \} \end{aligned}$$



$$X(\omega) = \omega_1 \quad Y(\omega) = \omega_2$$

$$\rho_{X,Y}(x, y) = \begin{cases} \frac{x+y}{252} & \text{if } (x, y) \in \Omega \\ 0 & \text{otherwise} \end{cases}$$

Marginal Probability

$$p(X = x, Y = y) = \rho_X(x, y)$$

$$p(X = x) = \sum_{y' \in \mathcal{Y}} p(X = x, Y = y')$$

$$p(Y = y) = \sum_{x' \in \mathcal{X}} p(X = x', Y = y)$$

$\Omega = \{(1, 1), (1, 2), (1, 3), (1, 4), (1, 5), (1, 6),$
 $(2, 1), (2, 2), (2, 3), (2, 4), (2, 5), (2, 6),$
 $(3, 1), (3, 2), (3, 3), (3, 4), (3, 5), (3, 6),$
 $(4, 1), (4, 2), (4, 3), (4, 4), (4, 5), (4, 6),$
 $(5, 1), (5, 2), (5, 3), (5, 4), (5, 5), (5, 6),$
 $(6, 1), (6, 2), (6, 3), (6, 4), (6, 5), (6, 6), \}$

$$p(X = 4) = \sum_{y' \in [1, 6]} p(X = 4, Y = y')$$

$$p(Y = 3) = \sum_{x' \in [1, 6]} p(X = x', Y = 3)$$

$$\rho_{X,Y}(x,y) = \begin{cases} \frac{1}{36} & \text{if } (x,y) \in \Omega \\ 0 & \text{otherwise} \end{cases}$$

$$\Omega = \{(1,1), (1,2), (1,3), (1,4), (1,5), (1,6), \\ (2,1), (2,2), (2,3), (2,4), (2,5), (2,6), \\ (3,1), (3,2), (3,3), (3,4), (3,5), (3,6), \\ (4,1), (4,2), (4,3), (4,4), (4,5), (4,6), \\ (5,1), (5,2), (5,3), (5,4), (5,5), (5,6), \\ (6,1), (6,2), (6,3), (6,4), (6,5), (6,6), \}$$

$\frac{6}{36} = \frac{1}{6}$

$$\rho_{X,Y}(x,y) = \begin{cases} \frac{x+y}{252} & \text{if } (x,y) \in \Omega \\ 0 & \text{otherwise} \end{cases}$$

$$\Omega = \{(1,1), (1,2), (1,3), (1,4), (1,5), (1,6), \\ (2,1), (2,2), (2,3), (2,4), (2,5), (2,6), \\ (3,1), (3,2), (3,3), (3,4), (3,5), (3,6), \\ (4,1), (4,2), (4,3), (4,4), (4,5), (4,6), \\ (5,1), (5,2), (5,3), (5,4), (5,5), (5,6), \\ (6,1), (6,2), (6,3), (6,4), (6,5), (6,6), \}$$

$\frac{4+1+4+2+4+3+4+4+4+5+4+6}{252} = \frac{45}{252}$

Conditional Probability

The **conditional probability** of one random variable given another is defined* as follows:

$$p(X = x \mid Y = y) = \frac{p(X = x, Y = y)}{p(Y = y)} = \frac{\text{joint probability}}{\text{marginal}}$$

Given that $p(y) \neq 0$

Conditional probability distributions are useful for specifying joint distributions since:

$$p(x \mid y)p(y) = p(x, y) = p(y \mid x)p(x)$$

Why might this be useful?

Conditional Probability Distributions

A **conditional probability distribution** is a probability distribution over r.v.'s X and Y with the form $\rho_{X|Y=y}(x)$.

$$\sum_{x \in \mathcal{X}} \rho_{X|Y=y}(x) \quad \forall y \in \mathcal{Y}$$

Chain rule

The **chain rule** is derived from a repeated application of the definition of conditional probability:

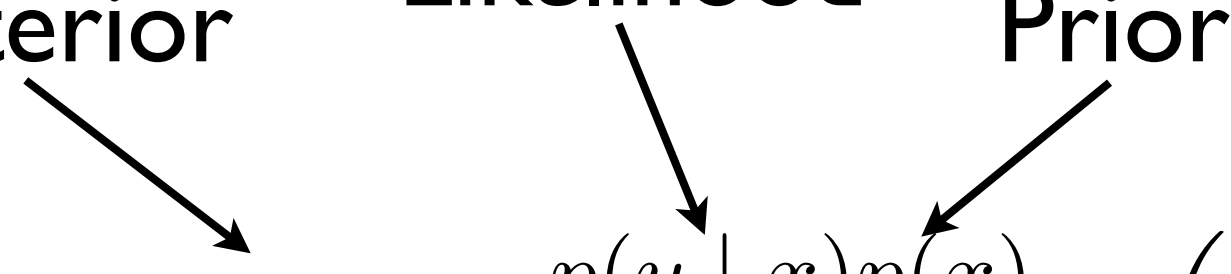
$$p(a, b, c, d)$$




Use as many times as necessary!

Bayes' Rule

Posterior Likelihood Prior


$$p(x | y) = \frac{p(y | x)p(x)}{p(y)} \left(= \frac{p(y | x)p(x)}{\sum_{x'} p(y | x')p(x')} \right)$$

Evidence



Independence

Two random variables are independent iff

$$p(X = x, Y = y) = p(X = x)p(Y = y)$$

Equivalently, (use def. of cond. prob to prove)

$$p(X = x \mid Y = y) = p(X = x)$$

Equivalently again:

$$p(Y = y \mid X = x) = p(Y = y)$$

“Knowing about X doesn’t tell me about Y ”

$$\rho_{X,Y}(x,y) = \begin{cases} \frac{1}{36} & \text{if } (x,y) \in \Omega \\ 0 & \text{otherwise} \end{cases}$$

$$\Omega = \{(1,1), (1,2), (1,3), (1,4), (1,5), (1,6), \\ (2,1), (2,2), (2,3), (2,4), (2,5), (2,6), \\ (3,1), (3,2), (3,3), (3,4), (3,5), (3,6), \\ (4,1), (4,2), (4,3), (4,4), (4,5), (4,6), \\ (5,1), (5,2), (5,3), (5,4), (5,5), (5,6), \\ (6,1), (6,2), (6,3), (6,4), (6,5), (6,6), \}$$

$$\rho_{X,Y}(x,y) = \begin{cases} \frac{x+y}{252} & \text{if } (x,y) \in \Omega \\ 0 & \text{otherwise} \end{cases}$$

$$\Omega = \{(1,1), (1,2), (1,3), (1,4), (1,5), (1,6), \\ (2,1), (2,2), (2,3), (2,4), (2,5), (2,6), \\ (3,1), (3,2), (3,3), (3,4), (3,5), (3,6), \\ (4,1), (4,2), (4,3), (4,4), (4,5), (4,6), \\ (5,1), (5,2), (5,3), (5,4), (5,5), (5,6), \\ (6,1), (6,2), (6,3), (6,4), (6,5), (6,6), \}$$

Independence

Independence has **practical benefits**. Think about how many parameters you need for a naive parameterization of $\rho_{X,Y}(x,y)$ vs $\rho_X(x)$ and $\rho_Y(y)$

$$O(xy) \text{ vs } O(x + y)$$

What about the cpd's $\rho_{X|Y=y}(x)$? Could we use $\rho_{X,Y}(x,y) = \rho_{X|Y=y}(x)\rho_Y(y)$ to save parameters?

$$O(xy + y) = O(xy)$$

Conditional Independence

Two equivalent statements of conditional independence:

$$p(a, c \mid b) = p(a \mid b)p(c \mid b)$$

and:

$$p(a \mid b, c) = p(a \mid b)$$

“If I know B, then C doesn’t tell me about A”

Conditional Independence

$$\begin{aligned} p(a, b, c) &= p(a \mid b, c)p(b, c) \\ &= p(a \mid b, c)p(b \mid c)p(c) \end{aligned}$$

“If I know B, then C doesn’t tell me about A”

$$p(a \mid b, c) = p(a \mid b)$$

$$\begin{aligned} p(a, b, c) &= p(a \mid b, c)p(b, c) \\ &= p(a \mid b, \textcolor{red}{c})p(b \mid c)p(c) \\ &= p(a \mid b)p(b \mid c)p(c) \end{aligned}$$

How many parameters do we need now?

Independence

- Some variables are independent In Nature
 - How do we know?
- Some variables we *pretend* are independent for computational convenience
 - Examples?
- Assuming independence is equivalent to letting our model “forget” something that happened in its past
 - What should we forget in language?

A Word About Data

- When we formulate our models there will be two kinds of random variables: observed and latent
- Observed: words, sentences(?), parallel corpora, web pages, formatting...
- Latent: parameters, syntax, “meaning”, word alignments, translation dictionaries...

Interlingua

“meaning”

```
report_event[
  factivity=true
  explode(e, bomb, car)
  loc(e, downtown)
]
```

explodieren
:a 0 Bombe
:a 1 Innenstadt
:loc Innenstadt
:tempus imperi

Hidden

detonate
:arg0 bomb
:a 1 car
:loc downtown
:time past

In der Innenstadt explodierte eine Autobombe

A car bomb exploded downtown

\$

\$

In der Innenstadt explodierte eine Autobombe

A car bomb exploded downtown

Learning

- Let's say we have formulated a model of a phenomenon
 - Made independence assumptions
 - Figured out what kinds of parameters we want
- Let's say we have collected data we assume to be generated by this model
 - E.g. some parallel data

What do we do now?

Parameter Estimation

- Inputs
 - Given a model with unspecified parameters
 - Given some data
- Goal: learn model parameters
- How?
 - Find parameters that make the model make predictions that look like the data do
 - What do we mean “look like the data?”
 - Probability (other options: accuracy, moment matching)

Strategies

- **Maximum likelihood estimation**
 - What is the *probability* of generating the data?
- **Accuracy**
 - Using an auxiliary similarity function, find parameters that maximize the (expected?) accuracy of data
- **Bayesian techniques**
 - Construct a probability distribution over the model parameters (how?)
 - Use Bayes rule to find the posterior distribution over the parameters, given the data
 - Pick a set of parameters using the posterior distribution (how? just one?)



$p(\text{heads})$

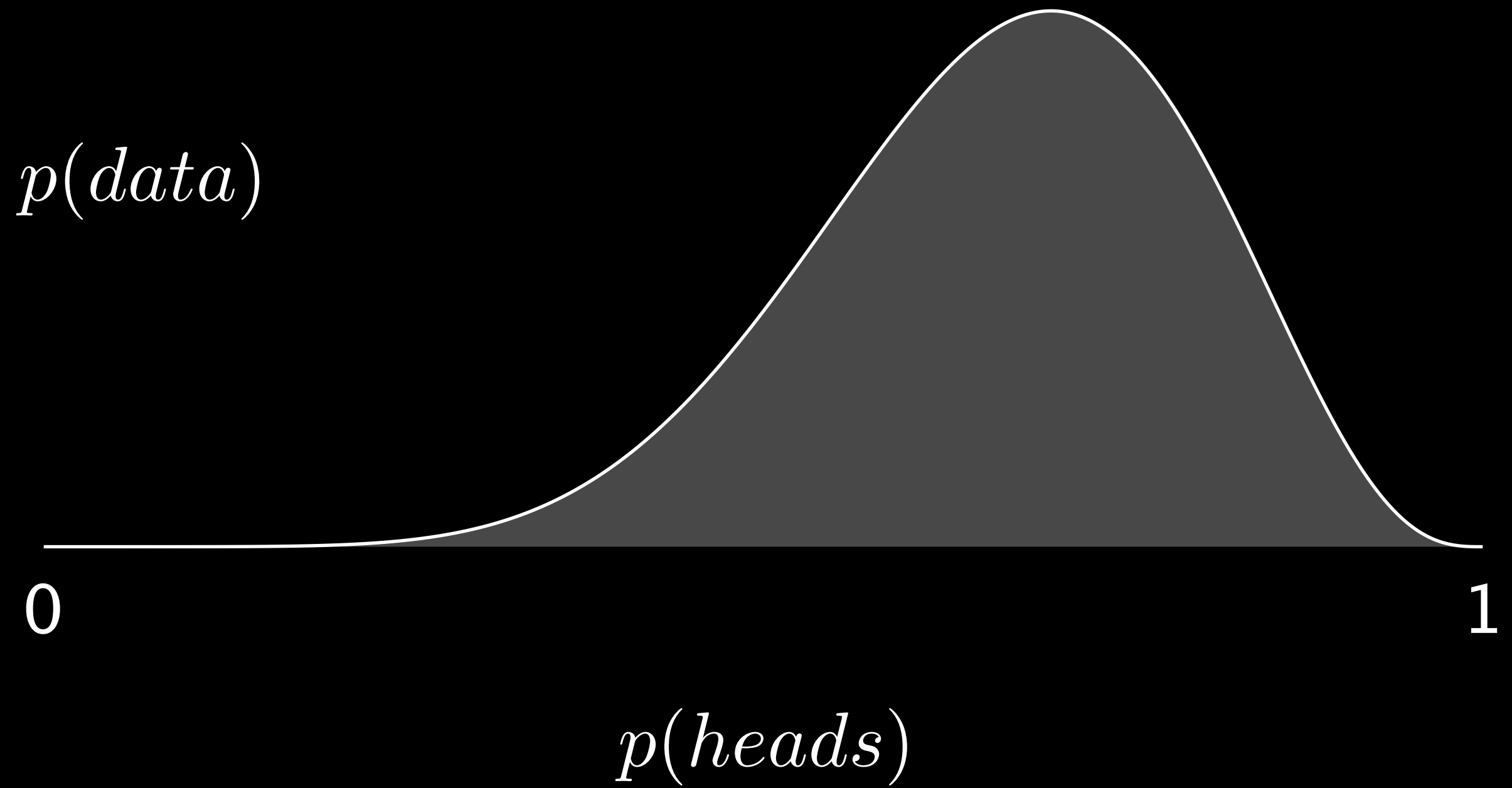


$1 - p(\text{heads})$

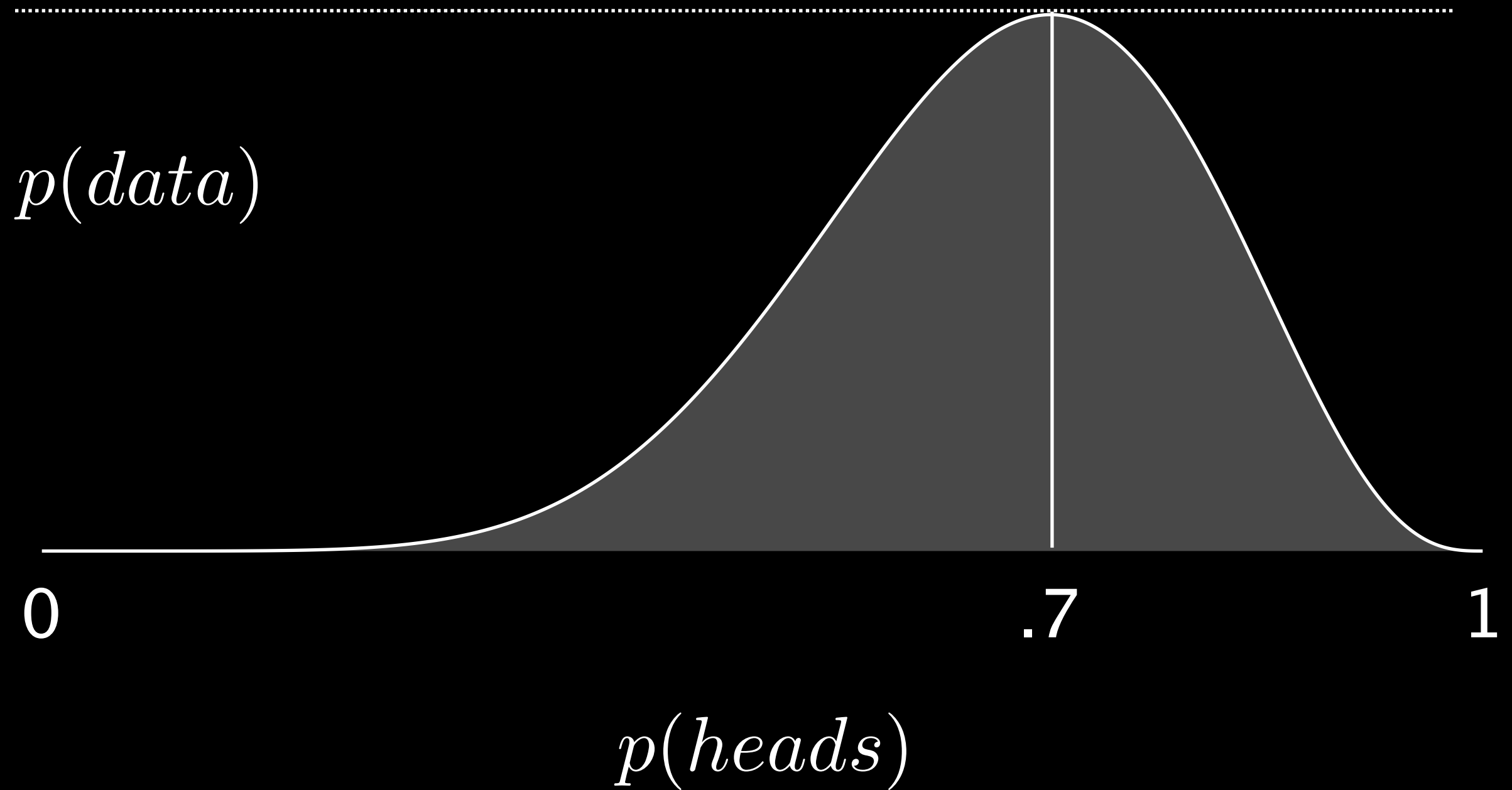
$p(\text{heads})$?



$$p(\text{data}) = p(\text{heads})^7 \times [1 - p(\text{heads})]^3$$



can be derived analytically using Lagrange multipliers



Optimization

- For the first month, we will be working with maximum likelihood estimation
- The general recipe is:
 - Come up with an expression of the likelihood of your probability model, as a function of **data** and the model **parameters**
 - Set the parameters to maximize the likelihood (using the **log likelihood** is usually easier)
 - This optimization is generally difficult
 - You must respect any constraints on the parameters (>0 , sum to 1, etc)
 - There may not be analytical solutions (log-linear models)

Multinomials

- The good news is, you can go a very long way in Machine Translation with “stupid multinomials”
- Why is this good news if they’re stupid?

$$\mathcal{L}(\boldsymbol{\theta}, \mathbf{x}) = \sum_{i=1}^{|\mathbf{x}|} \log \theta_{x_i}$$

$$\boldsymbol{\theta}^{\text{MLE}} = \arg \max_{\boldsymbol{\theta}} \mathcal{L}(\boldsymbol{\theta}, \mathbf{x})$$

$$\theta_w^{\text{MLE}} = \frac{1}{|\mathbf{x}|} \sum_{i=1}^{|\mathbf{x}|} \delta_w(x_i) \quad \forall w \in \mathbf{x}$$

Less Stupid Multinomials

$$\theta_w = \frac{\exp \lambda^\top \mathbf{f}(w)}{\sum_{w'} \exp \lambda^\top \mathbf{f}(w')}$$

Parameters

Features of w

Ends in *-ing*?

Contains a digit?

Found in Gigaword?

Contains a capital letter?

Less Stupid Multinomials

$$\theta_w = \frac{\exp \boldsymbol{\lambda}^\top \mathbf{f}(w)}{\sum_{w'} \exp \boldsymbol{\lambda}^\top \mathbf{f}(w')}$$

$$\mathcal{L}(\boldsymbol{\lambda}, \mathbf{x}) = \sum_{i=1}^{|\mathbf{x}|} \boldsymbol{\lambda}^\top \mathbf{f}(w_i) - \log \sum_{w'} \exp \boldsymbol{\lambda}^\top \mathbf{f}(w')$$

Less Stupid Multinomials

$$\theta_w = \frac{\exp \boldsymbol{\lambda}^\top \mathbf{f}(w)}{\sum_{w'} \exp \boldsymbol{\lambda}^\top \mathbf{f}(w')}$$

$$\mathcal{L}(\boldsymbol{\lambda}, \mathbf{x}) = \sum_{i=1}^{|\mathbf{x}|} \boldsymbol{\lambda}^\top \mathbf{f}(w) - \log \sum_{w'} \exp \boldsymbol{\lambda}^\top \mathbf{f}(w')$$

$$\frac{\partial \mathcal{L}}{\partial \lambda_k}(\boldsymbol{\lambda}, \mathbf{x}) = \sum_{i=1}^{|\mathbf{x}|} f_k(w) - \frac{\sum_{w'} f_k(w') \exp \boldsymbol{\lambda}^\top \mathbf{f}(w')}{\sum_{w''} \exp \boldsymbol{\lambda}^\top \mathbf{f}(w')}$$

Less Stupid Multinomials

$$\theta_w = \frac{\exp \boldsymbol{\lambda}^\top \mathbf{f}(w)}{\sum_{w'} \exp \boldsymbol{\lambda}^\top \mathbf{f}(w')}$$

$$\mathcal{L}(\boldsymbol{\lambda}, \mathbf{x}) = \sum_{i=1}^{|\mathbf{x}|} \boldsymbol{\lambda}^\top \mathbf{f}(w) - \log \sum_{w'} \exp \boldsymbol{\lambda}^\top \mathbf{f}(w')$$

$$\frac{\partial \mathcal{L}}{\partial \lambda_k}(\boldsymbol{\lambda}, \mathbf{x}) = \sum_{i=1}^{|\mathbf{x}|} f_k(w) - \frac{\sum_{w'} f_k(w') \exp \boldsymbol{\lambda}^\top \mathbf{f}(w')}{\sum_{w''} \exp \boldsymbol{\lambda}^\top \mathbf{f}(w')}$$

$$= \sum_{i=1}^{|\mathbf{x}|} f_k(w) - \mathbb{E}_{p(w'; \boldsymbol{\lambda})} f_k(w')$$

Less Stupid Multinomials

$$\theta_w = \frac{\exp \boldsymbol{\lambda}^\top \mathbf{f}(w)}{\sum_{w'} \exp \boldsymbol{\lambda}^\top \mathbf{f}(w')}$$

$$\mathcal{L}(\boldsymbol{\lambda}, \mathbf{x}) = \sum_{i=1}^{|\mathbf{x}|} \boldsymbol{\lambda}^\top \mathbf{f}(w) - \log \sum_{w'} \exp \boldsymbol{\lambda}^\top \mathbf{f}(w')$$

$$\frac{\partial \mathcal{L}}{\partial \lambda_k}(\boldsymbol{\lambda}, \mathbf{x}) = \sum_{i=1}^{|\mathbf{x}|} f_k(w) - \frac{\sum_{w'} f_k(w') \exp \boldsymbol{\lambda}^\top \mathbf{f}(w')}{\sum_{w''} \exp \boldsymbol{\lambda}^\top \mathbf{f}(w')}$$

$$= \sum_{i=1}^{|\mathbf{x}|} f_k(w) - \mathbb{E}_{p(w'; \boldsymbol{\lambda})} f_k(w')$$

$$\nabla \mathcal{L}(\boldsymbol{\lambda}, \mathbf{x}) = 0$$

Less Stupid Multinomials

$$\theta_w = \frac{\exp \boldsymbol{\lambda}^\top \mathbf{f}(w)}{\sum_{w'} \exp \boldsymbol{\lambda}^\top \mathbf{f}(w')}$$

$$\mathcal{L}(\boldsymbol{\lambda}, \mathbf{x}) = \sum_{i=1}^{|\mathbf{x}|} \boldsymbol{\lambda}^\top \mathbf{f}(w) - \log \sum_{w'} \exp \boldsymbol{\lambda}^\top \mathbf{f}(w')$$

$$\frac{\partial \mathcal{L}}{\partial \lambda_k}(\boldsymbol{\lambda}, \mathbf{x}) = \sum_{i=1}^{|\mathbf{x}|} f_k(w) - \frac{\sum_{w'} f_k(w') \exp \boldsymbol{\lambda}^\top \mathbf{f}(w')}{\sum_{w''} \exp \boldsymbol{\lambda}^\top \mathbf{f}(w')}$$

$$= \sum_{i=1}^{|\mathbf{x}|} f_k(w) - \mathbb{E}_{p(w'; \boldsymbol{\lambda})} f_k(w')$$

$$\nabla \mathcal{L}(\boldsymbol{\lambda}, \mathbf{x}) = 0$$

No analytic solution! :(

Announcements

- Complete HW 0 by Friday 11:59pm
- You may remain anonymous on GitHub
- Sign up for language-in-10 minutes (due **Tuesday, Jan 22, at noon**)
- Have a nice weekend!