Maximum Entropy Model (I)

LING 572
Advanced Statistical Methods for NLP
January 28, 2020

MaxEnt in NLP

- The maximum entropy principle has a long history.

- The MaxEnt algorithm was introduced to the NLP field by Berger et. al. (1996).

- Used in many NLP tasks: Tagging, Parsing, PP attachment, …

Readings & Comments

- Several readings:
  - (Berger, 1996), (Ratnaparkhi, 1997)
  - (Klein & Manning, 2003): Tutorial
  - Note: Some of these are very ‘dense’
    - Don’t spend huge amount of time on every detail
    - Take a first pass before class, review after lecture

- Going forward:
  - Techniques more complex
    - Goal: Understand basic model, concepts
    - Training is complex; we’ll discuss, but not implement

Notation

We use this one

	Input	Output	Pair
Berger et al 1996	x	y	(x, y)
Ratnaparkhi 1997	b	a	x
Ratnaparkhi 1996	h	t	(h, t)
Klein and Manning 2003	d	c	(d, c)

Outline

- Overview
- The Maximum Entropy Principle

- Modeling**
- Decoding

- Training**

- Case study: POS tagging

Overview

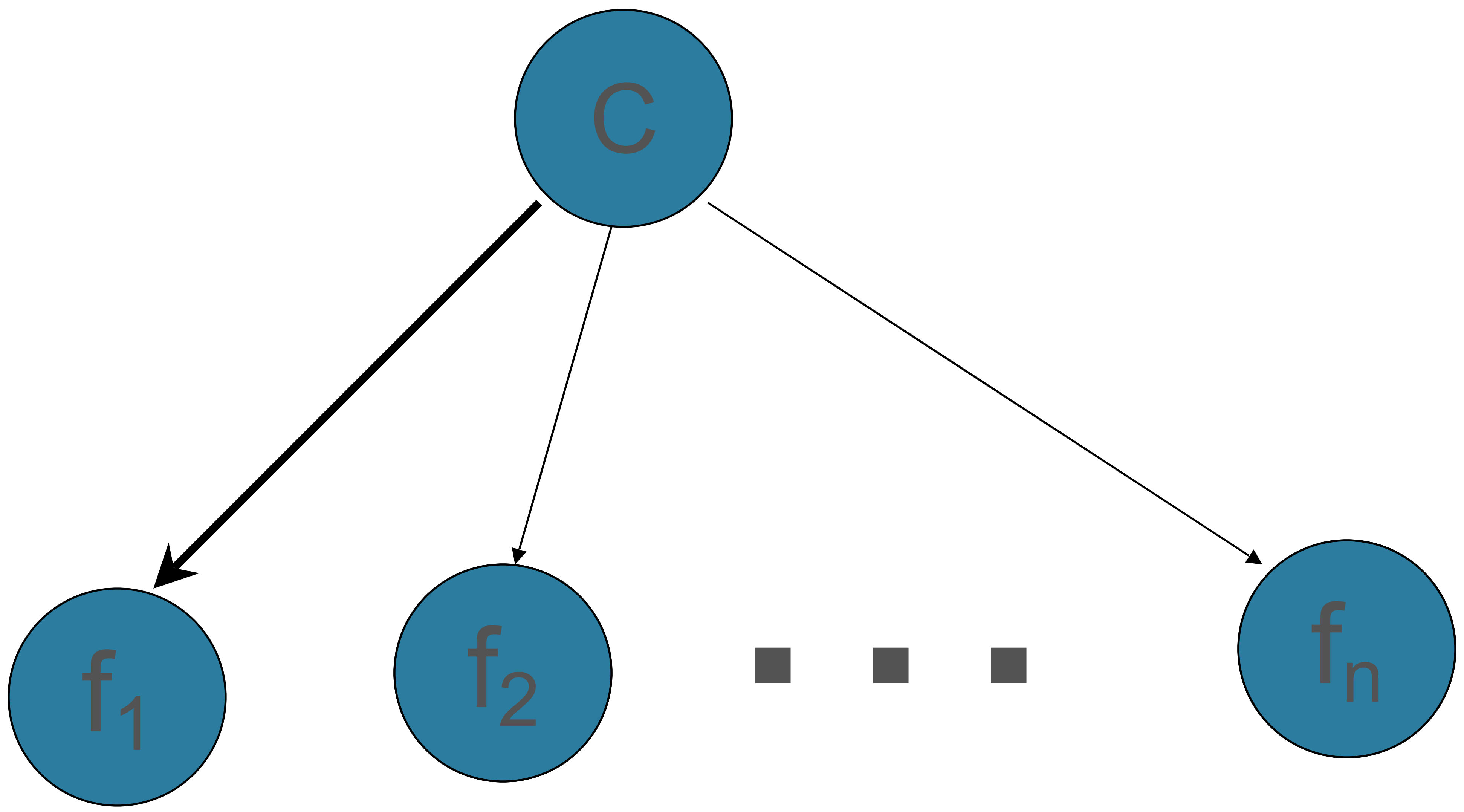
Joint vs. Conditional models

- Given training data {(x,y)}, we want to build a model to predict y for new x's. For each model, we need to estimate the parameters μ.

- Joint (aka generative) models estimate P(x,y) by maximizing the likelihood: $P(X,Y|\mu)$
  - Ex: n-gram models, HMM, Naïve Bayes, PCFG
  - Choosing weights is trivial: just use relative frequencies.

- Conditional (aka discriminative) models estimate P(y | x) by maximizing the conditional likelihood: $P(Y \mid X, \mu)$
  - Ex: MaxEnt, SVM, CRF, etc.
  - Computing weights is more complex.

Naïve Bayes Model



Assumption: each f_i is conditionally independent from f_j given C.

The conditional independence assumption

f_m and f_n are conditionally independent given c:

$$P(f_m \mid c, f_n) = P(f_m \mid c)$$

Counter-examples in the text classification task:

- P("Manchester" | entertainment) !=
  P("Manchester" | entertainment, "Oscar")

Q: How to deal with correlated features?

A: Many models, including MaxEnt, do not assume that features are conditionally independent.

Naïve Bayes highlights

- Choose

  $c^* = \arg\max_c P(c) \prod_k P(f_k \mid c)$

- Two types of model parameters:
  - Class prior: $P(c)$
  - Conditional probability: $P(f_k \mid c)$

- The number of model parameters:

  $|C|+|CV|$

P(f l c) in NB

	f_1	f_2	…	f_j
c_1	$P(f_1 \mid c_1)$	$P(f_2 \mid c_1)$	…	$P(f_j \mid c_1)$
c_2	$P(f_1 \mid c_2)$	…	…	…
…	…			
c_i	$P(f_1 \mid c_i)$	…	…	$P(f_j \mid c_i)$

Each cell is a weight for a particular (class, feat) pair.

Weights in NB and MaxEnt

- In NB
  - P(f | y) are probabilities (i.e., in [0,1])
  - P(f | y) are multiplied at test time

$$P(y|x) = \frac{P(y)\prod_k P(f_k|y)}{Z} = \frac{e^{ln(P(y)\prod_k P(f_k|y))}}{Z}$$

$$= \frac{e^{lnP(y)+ln(\prod_k P(f_k|y))}}{Z} = \frac{e^{lnP(y)+\sum_k lnP(f_k|y)}}{Z}$$

- In MaxEnt
  - the weights are real numbers: they can be negative.
  - the weighted features are added at test time

$$P(y|x) = \frac{e^{\sum_j \lambda_j f_j(x,y)}}{Z}$$

Highlights of MaxEnt

$$P(y|x) = \frac{e^{\sum_j \lambda_j f_j(x,y)}}{Z}$$

$f_j(x,y)$ is a feature function, which normally corresponds to a (feature, class) pair.

Training: to estimate λ_j

Testing: to calculate P(y | x)

Main questions

- What is the maximum entropy principle?

- What is a feature function?

- Modeling: Why does P(y|x) have the form?

$$P(y|x) = \frac{e^{\sum_j \lambda_j f_j(x,y)}}{Z}$$

- Training: How do we estimate λ_j ?

Outline

- Overview
- The Maximum Entropy Principle

- Modeling**
- Decoding

- Training*

- Case study

Maximum Entropy Principle

Maximum Entropy Principle

- Intuitively, model all that is known, and assume as little as possible about what is unknown.

- Related to Occam's razor and other similar justifications for scientific inquiry

- Also: Laplace's *Principle of Insufficient Reason:* when one has no information to distinguish between the probability of two events, the best strategy is to consider them equally likely.

Maximum Entropy

- Why maximum entropy?
  - Maximize entropy = Minimize commitment

- Model all that is known and assume nothing about what is unknown.
  - Model all that is known: satisfy a set of constraints that must hold

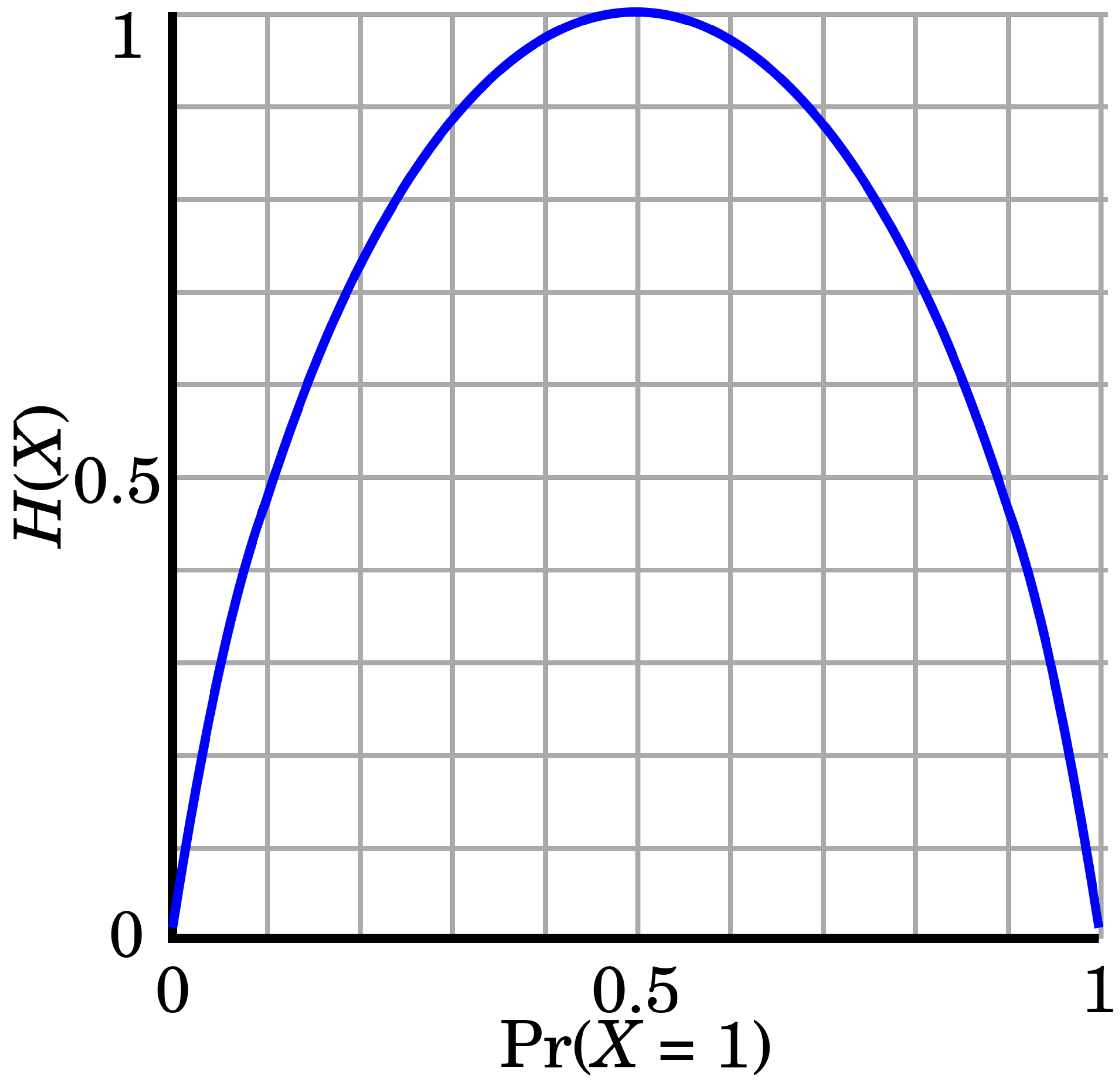
  - Assume nothing about what is unknown:
    choose the most "uniform" distribution

    ➔ choose the one with maximum entropy

Ex1: Coin-flip example (Klein & Manning, 2003)

- Toss a coin: p(H)=p1, p(T)=p2.
- Constraint: p1 + p2 = 1
- Question: what's p(x)? That is, what is the value of p1?
- Answer: choose the p that maximizes

$$H(p) = -\sum_x p(x) \log p(x)$$



Ex2: An MT example (Berger et. al., 1996)

Possible translation for the word "in" is: {dans, en, à, au cours de, pendant}

Constraint: $p(dans) + p(en) + p(à) + p(au\ cours\ de) + p(pendant) = 1$

Intuitive answer:

$$
\begin{aligned}
p(dans) &= 1/5 \\
p(en) &= 1/5 \\
p(à) &= 1/5 \\
p(au\ cours\ de) &= 1/5 \\
p(pendant) &= 1/5
\end{aligned}
$$

An MT example (cont)

Constraints:

$$p(dans) + p(en) = 3/10$$

$$p(dans) + p(en) + p(à) + p(au\ cours\ de) + p(pendant) = 1$$

Intuitive answer:

$$p(dans) = 3/20$$

$$p(en) = 3/20$$

$$p(à) = 7/30$$

$$p(au\ cours\ de) = 7/30$$

$$p(pendant) = 7/30$$

An MT example (cont)

Constraints:

$$p(dans) + p(en) = 3/10$$

$$p(dans) + p(en) + p(\grave{a}) + p(au\ cours\ de) + p(pendant) = 1$$

$$p(dans) + p(\grave{a}) = 1/2$$

Intuitive answer: ??

Ex3: POS tagging (Klein and Manning, 2003)

- Lets 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

Ex3 (cont)

- 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

NN	NNS	NNP	NNPS	VBZ	VBD
4/36	4/36	12/36	12/36	2/36	2/36

Ex4: Overlapping features (Klein and Manning, 2003)

Empirical

	A	a
B	1	1
b	1	0

	A	a
B		
b		

All = 1

	A	a
B	p1	p2
b	p3	p4

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

Ex4 (cont)

Empirical

	A	a
B	1	1
b	1	0

	A	a
B	p1	p2
b	$\frac{2}{3} - p_1$	$\frac{1}{3} - p_2$

A = 2/3

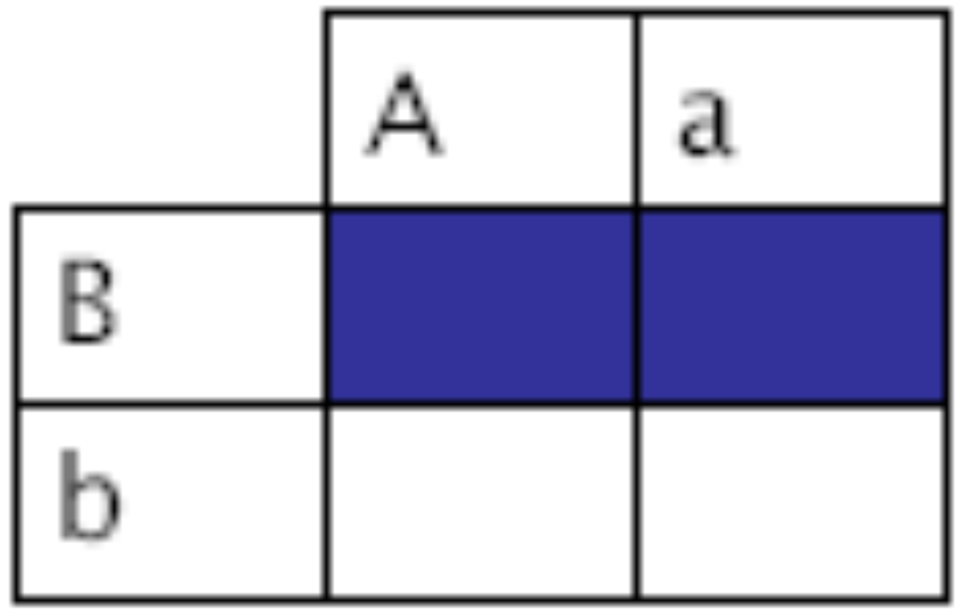
	A	a
B	1/3	1/6
b	1/3	1/6

Ex4 (cont)

Empirical

	A	a
B	1	1
b	1	0

A = 2/3



B = 2/3

	A	a
B	p1	$\frac{2}{3} - p_1$
b	$\frac{2}{3} - p_1$	$p_1 - \frac{1}{3}$

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

The MaxEnt Principle summary

- Goal: Among all the distributions that satisfy the constraints, choose the one, p*, that maximizes H(p).

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

- Q1: How to represent constraints?

- Q2: How to find such distributions?