03 entropy, KL divergence,

William Trimble

Winter 2023

THE UNIVERSITY OF CHICAGO

Office hours

- Amy Tuesday 12:30-2:00 (maybe also on Wednesday)
- JungHo Tuesday 3:00-4:30pm
- Qiming Wednesday 3:30 – 5:00 pm.
- WT Thursday 2:00-4:00 Crerar 346

The norm definitions

$$L^1 - \text{norm} = |x|_1 = \sum_i |x_i|$$

$$\ell^2 - \text{norm} = |x| = \sqrt{\sum_i |x_i|^2}$$

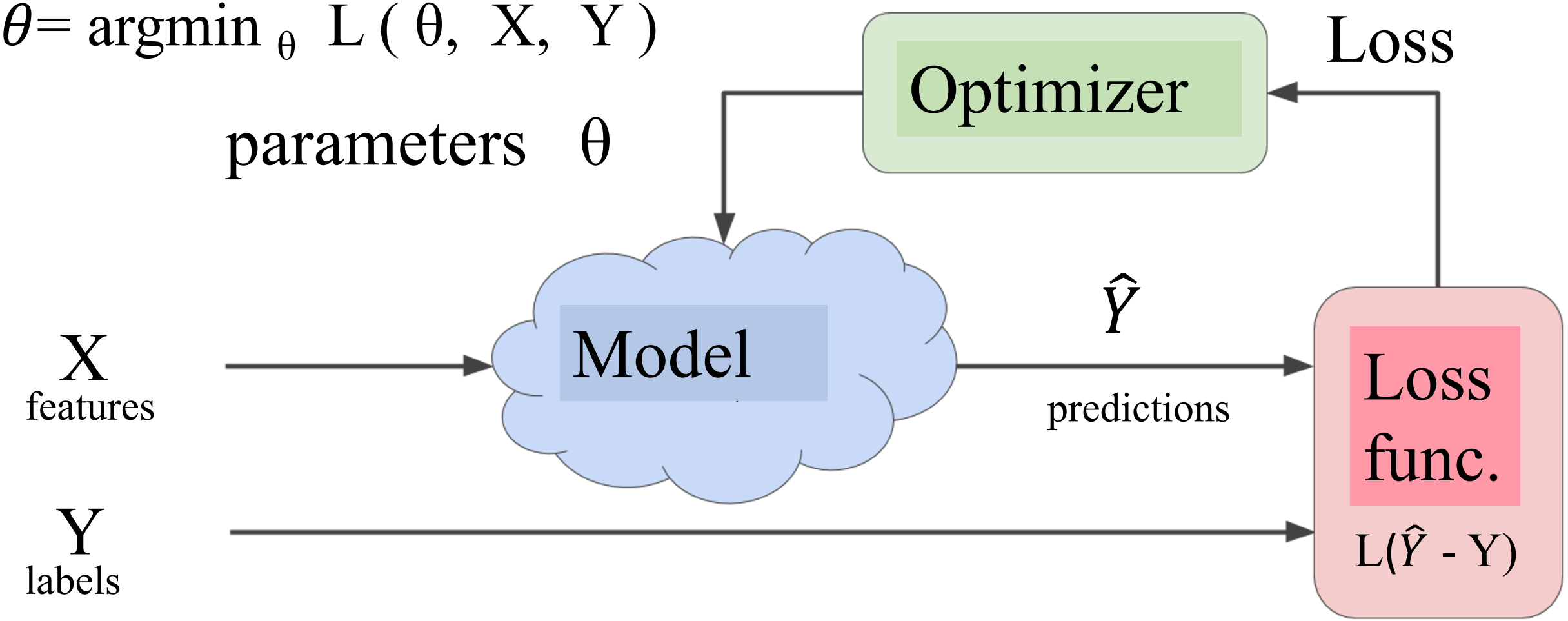
$$\ell^\infty - \text{norm} = |x|_\infty = \max_i |x_i|$$

"norms" always have same units as x!

Fairy dust picture of optimization

$$\hat{\theta} = \text{argmin}_{\theta}\ L(\theta, X, Y)$$

parameters θ



Inferring parameters

θ: value of model parameter

Y: training data

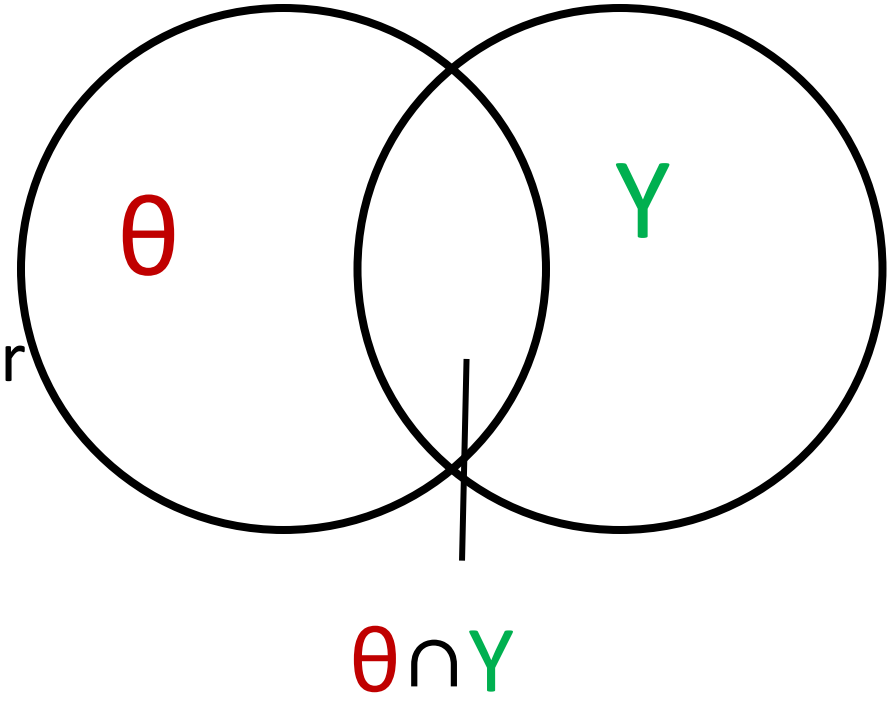
$$P(\theta \mid Y) = P(Y|\theta)\, P(\theta) / P(Y)$$

Distribution to help me find the parameter value

This is the likelihood

Prior distribution for parameter

Doesn't matter for us



posterior *likelihood* *prior* *evidence*

Inferring parameters

θ: value of model parameter

Y: training data

$$\log P(\theta \mid Y) = \log P(Y \mid \theta) + \log P(\theta)$$

log parameter distribution

The log likelihood

Prior distribution on the parameter

log posterior

log likelihood

log prior

Inferring parameters

θ: value of model parameter

Y: training data

$$\log P(\theta \mid Y) = \log P(Y \mid \theta) + \log P(\theta)$$

$$L(\theta \mid Y) = \sum \frac{(\hat{y}_i(\theta) - y_i)^2}{\sigma_i^2} + \sum \frac{(\theta_k)^2}{s_k^2}$$

Additive normal errors on $\hat{y}$

Normal prior on each component of θ

log posterior — *log likelihood* — *log prior*

Dice problem

- Suppose I have two six-sided dice:

A: 1 2 3 4 5 6

B: 1 2 3 4 5 1

P(O|A) = $\frac{1}{6}, \frac{1}{6}, \frac{1}{6}, \frac{1}{6}, \frac{1}{6}, \frac{1}{6}$

P(O|B) = $\frac{1}{3}, \frac{1}{6}, \frac{1}{6}, \frac{1}{6}, \frac{1}{6}, 0$

Dice problem

- Suppose I have two six-sided dice:

A: 1 2 3 4 5 6

B: 1 2 3 4 5 1

$P(O|A) = \frac{1}{6}, \frac{1}{6}, \frac{1}{6}, \frac{1}{6}, \frac{1}{6}, \frac{1}{6}$

$P(O|B) = \frac{1}{3}, \frac{1}{6}, \frac{1}{6}, \frac{1}{6}, \frac{1}{6}, 0$

Imagine the rolls of the dice as symptoms of a denial of service attack, symtoms of cancer, and the identity of the die (A or B) as the underlying state of nature to be estimated.

Dice problem

- Suppose I have two six-sided dice:

A: 1 2 3 4 5 6

B: 1 2 3 4 5 1

P(O|A) = $\frac{1}{6}, \frac{1}{6}, \frac{1}{6}, \frac{1}{6}, \frac{1}{6}, \frac{1}{6}$

P(O|B) = $\frac{1}{3}, \frac{1}{6}, \frac{1}{6}, \frac{1}{6}, \frac{1}{6}$, 0

The first time I see a six, the game is over.

Dice problem

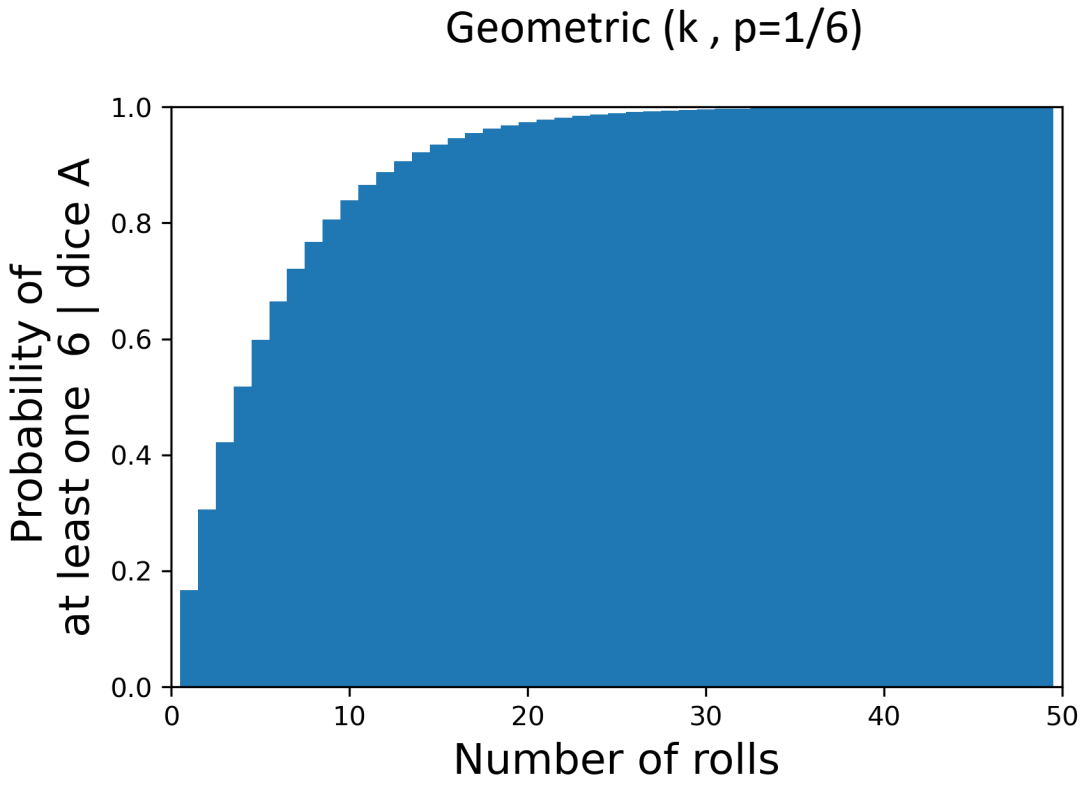
- Suppose I have two six-sided dice:

A: 1 2 3 4 5 6

B: 1 2 3 4 5 1

$P(O|A) = \frac{1}{6}, \frac{1}{6}, \frac{1}{6}, \frac{1}{6}, \frac{1}{6}, \frac{1}{6}$

$P(O|B) = \frac{1}{3}, \frac{1}{6}, \frac{1}{6}, \frac{1}{6}, \frac{1}{6}, 0$



Dice problem

- Suppose I have two six-sided dice:

A: 1 2 3 4 5 6

B: 1 2 3 4 5 1

$P(O|A) = \frac{1}{6}, \frac{1}{6}, \frac{1}{6}, \frac{1}{6}, \frac{1}{6}, \frac{1}{6}$

$P(O|B) = \frac{1}{3}, \frac{1}{6}, \frac{1}{6}, \frac{1}{6}, \frac{1}{6}, 0$

$P(A \mid roll) = P(roll|A)\, P(A) / P(roll)$

$P(B \mid roll) = P(roll|B)\, P(B) / P(roll)$

Suggestion: Why don't I keep track of the ratio of the probability of A to the probability of B?

Dice problem

- Suppose I have two six-sided dice:

A: 1 2 3 4 5 6

B: 1 2 3 4 5 1

$P(O|A) = \frac{1}{6}, \frac{1}{6}, \frac{1}{6}, \frac{1}{6}, \frac{1}{6}, \frac{1}{6}$

$P(O|B) = \frac{1}{3}, \frac{1}{6}, \frac{1}{6}, \frac{1}{6}, \frac{1}{6}, 0$

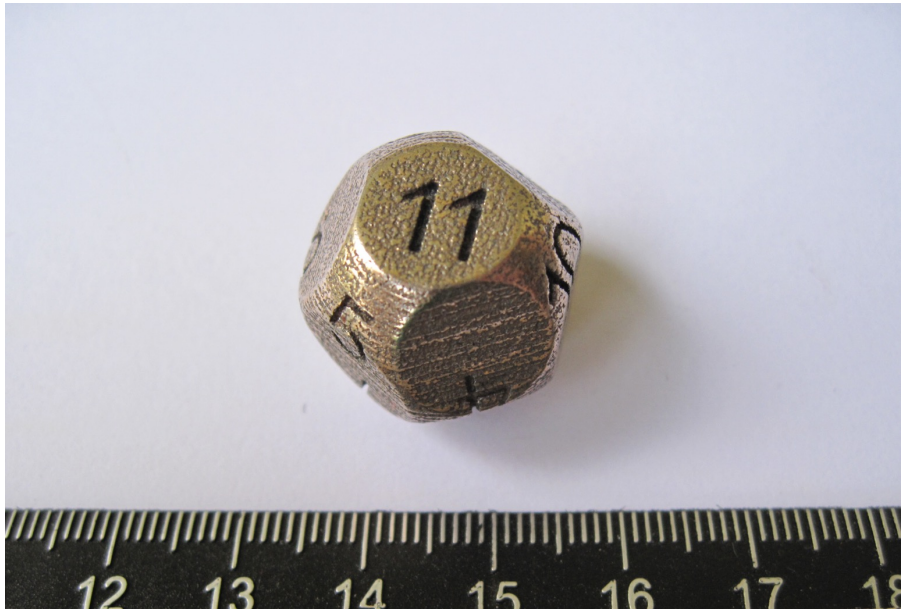
$P(A \mid roll) = P(roll|A)\, P(A) / P(roll)$

$P(B \mid roll) = P(roll|B)\, P(B) / P(roll)$

Suggestion: Why don't I keep track of the ratio of the probability of A to the probability of B?

Consequence: I'm going to need a lookup table for P(A|R) / P(B|R) for all the relevant values of R.

Dice problem



- Suppose I have two 11-sided dice, one of which is unfair, but they have the same faces:

A: 2 3 4 5 6 7 8 9 10 11 12 (fair 11-sided die)

C: 2 3 4 5 6 7 8 9 10 11 12 (sum of 2 six-sided dice)

$$P(O|A) = \frac{1}{11}, \frac{1}{11}, \frac{1}{11}, \frac{1}{11}, \frac{1}{11}, \frac{1}{11}, \frac{1}{11}, \frac{1}{11}, \frac{1}{11}, \frac{1}{11}, \frac{1}{11}$$

$$P(O|C) = \frac{1}{36}, \frac{2}{36}, \frac{3}{36}, \frac{4}{36}, \frac{5}{36}, \frac{6}{36}, \frac{5}{36}, \frac{4}{36}, \frac{3}{36}, \frac{2}{36}, \frac{1}{36}$$

So let's calculate the odds ratio A:B

It is convenient to look at the ratio of probabilities (odds) favoring A over B:

$$\frac{P(A\,|O)}{P(B|\,O)} = \frac{P(O\,|A)}{P(O\,|B)} \quad \frac{P(A)}{P(B)}$$

and its logarithm, log-odds A over B:

$$\log\left(\frac{P(A\,|O)}{P(B|\,O)}\right) = \log(P(O\,|A) - \log(P(O\,|B)) + (\log P(A) - \log P(B))$$

Why? Because products become sums, and I can just add a term to my log odds each time the die is thrown.

How long will it take?

$$P(O|A) = \frac{1}{11}, \frac{1}{11}, \frac{1}{11}, \frac{1}{11}, \frac{1}{11}, \frac{1}{11}, \frac{1}{11}, \frac{1}{11}, \frac{1}{11}, \frac{1}{11}, \frac{1}{11}$$

$$P(O|C) = \frac{1}{36}, \frac{2}{36}, \frac{3}{36}, \frac{4}{36}, \frac{5}{36}, \frac{6}{36}, \frac{5}{36}, \frac{4}{36}, \frac{3}{36}, \frac{2}{36}, \frac{1}{36}$$

$$\log_2\left(\frac{P(A|O)}{P(C|O)}\right) = 1.71, 0.71, 0.13, -0.29, -0.61, -0.97, -0.61, -0.29, 0.13, 0.71, 1.77$$

Mean log-odds | A = 0.22 bits / roll

Mean log-odds | B = -0.19 bits / roll

$\log_2 99 = 6.63$ bits (threshold 99:1)

How long will it take?

$$P(O|A) = \frac{1}{11}, \frac{1}{11}, \frac{1}{11}, \frac{1}{11}, \frac{1}{11}, \frac{1}{11}, \frac{1}{11}, \frac{1}{11}, \frac{1}{11}, \frac{1}{11}, \frac{1}{11}$$

$$P(O|C) = \frac{1}{36}, \frac{2}{36}, \frac{3}{36}, \frac{4}{36}, \frac{5}{36}, \frac{6}{36}, \frac{5}{36}, \frac{4}{36}, \frac{3}{36}, \frac{2}{36}, \frac{1}{36}$$

$$\log_2\left(\frac{P(A|O)}{P(C|O)}\right) = 1.71, 0.71, 0.13, -0.29, -0.61, -0.97, -0.61, -0.29, 0.13, 0.71, 1.77$$

Mean log-odds | A = 0.22 bits / roll

Mean log-odds | B = -0.19 bits / roll

$\log_2 99 = 6.63$ bits (threshold 99:1)

6.63 bits / 0.22 bits / roll = 30.2 rolls |A

6.63 bits / 0.19 bits / roll = 35.8 rolls |B

Wait, what assumptions did I make here?

- 0.4047 bits per roll is the difference between the expected value of $\log_2\left(\frac{P(A\,|O)}{P(C|\,O)}\right)$ in the universe where A is true its value when C is true.

Why can I trust this ? Central limit theorem.

Sums of 30 independent likelihood terms? Well-behaved mean.

Why can't I trust this? Central limit theorem.

Wait, what assumptions did I make here?

- 0.4047 bits per roll is the difference between the expected value of $\log_2\left(\frac{P(A|O)}{P(C|O)}\right)$ in the universe where A is true its value when C is true.

Why can I trust this ? Central limit theorem.

Sums of 30 independent likelihood terms? Well-behaved mean.

Why can't I trust this? Central limit theorem.

- CLT requires finite variance; I can't use this trick if any part of the distribution confers certainty (p=0, 1/p = inf).

Wait, what assumptions did I make here?

- 0.4047 bits per roll is the difference between the expected value of $\log_2\left(\frac{P(A\,|O)}{P(C|\,O)}\right)$ in the universe where A is true its value when C is true.

Why can I trust this ? Central limit theorem.

Sums of 30 independent likelihood terms? Well-behaved mean.

Why can't I trust this? Central limit theorem.

- CLT requires finite variance; I can't use this trick if any part of the distribution confers certainty (p=0, 1/p = inf).
- CLT requires independence of random variables.

Dice problem

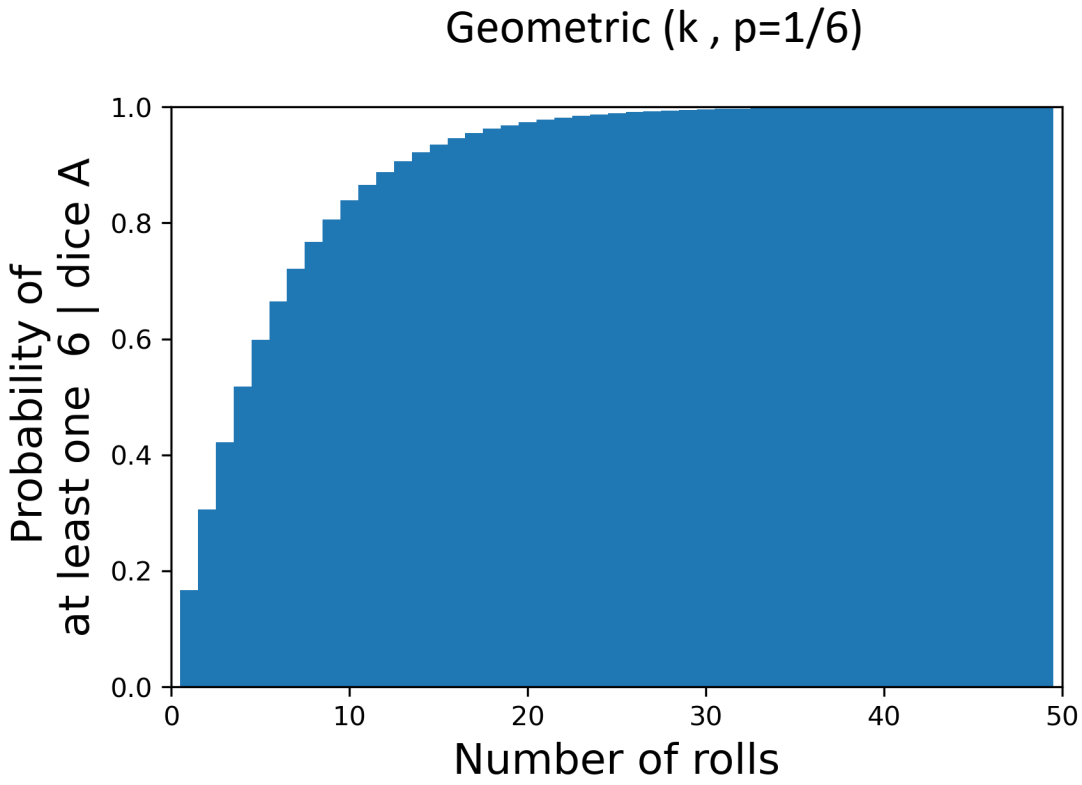
- Suppose I have two six-sided dice:

A: 1 2 3 4 5 6

B: 1 2 3 4 5 1

$P(O|A) = \frac{1}{6}, \frac{1}{6}, \frac{1}{6}, \frac{1}{6}, \frac{1}{6}, \frac{1}{6}$

$P(O|B) = \frac{1}{3}, \frac{1}{6}, \frac{1}{6}, \frac{1}{6}, \frac{1}{6}, 0$



Entropy

$$H = -\sum_i p_i \ log(p_i)$$

Entropy is a property of a probability distribution (or of an element within a probability distribution, like an outcome) that describes how concentrated the distribution is.

Entropy

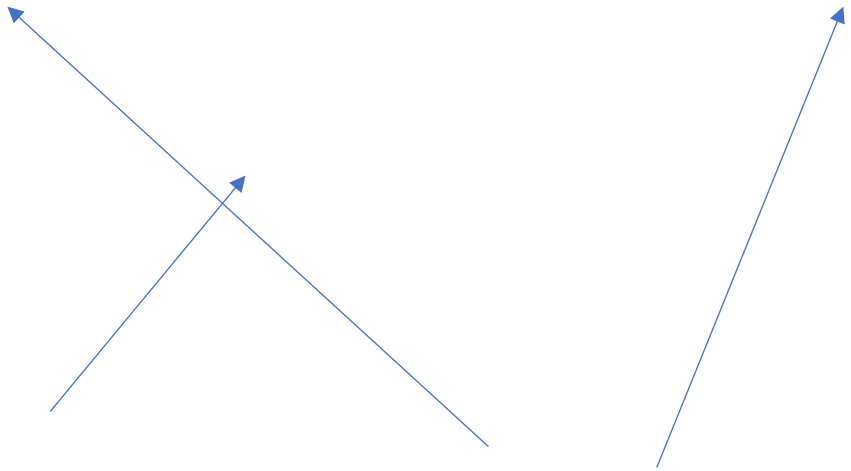
$$H = -\sum_i p_i \, log(p_i)$$

It has been constructed to have some nice properties (entropy of independent events is additive) and is, up to a multiplicative constant (which is the same as choosing the base of logarithm) the unique measure for which products are sums.

Entropy – discrete formulation

$$H = \Sigma_i \; p_i \; \log \left(\frac{1}{p_i}\right)$$

$$\mathrm{H} = -\int p(x) \log(p(x)) dx$$



Expected value of the –log probability

Entropy – continuous formulation

$$H = -\int p(\boldsymbol{x}) log\ p(\boldsymbol{x}) d\boldsymbol{x}$$

Expected value of the –log probability

Entropy – parts of the definition

- Expected value
- log – entropies of independent events additive
- negative – probabilities are all < 1; logs of all these probabilities are < 0; convention makes low-probabilities high entropy and high probabilities low-entropy.

letter	freq p	-log(p)
1 a	0.0575	4.13
2 b	0.0128	6.28
3 c	0.0263	5.25
4 j	0.0006	10.70
5 d	0.0285	5.13

Properties of entropy

- If all of the probability is in one point, H = 0
- Bernoulli distribution: entropy is maximum at p=0.5
- Uniform categorical distribution {1/n, 1/n … 1/n} has entropy log n

- Watch out for reporting entropy; logarithms have “units” that communicate the base of the log. log10 “ban” log 2 “bits” natural log “nats” Antilog entropy (sometimes called “perplexity”) doesn’t have that problem, since it’s a unitless number.

Tool to characterize and compare distributions; tool for use in loss function.

Kullback-Leibler divergence

also called relative entropy

"Expectation of the logarithmic difference between the two probability distributions:"

$$\mathrm{D_{KL}}(\mathrm{P} \parallel \mathrm{Q}) = \sum P(x) \ \log\left(\frac{P(x)}{Q(x)}\right)$$

But we just saw this in the nonstandard dice example:

This is the expected odds favoring the P distribution over the Q distribution assuming the P distribution is correct.

Properties of the Kullback-Liebler divergence

- Why is it called a divergence? Because it's not symmetrical, and the word "distance" is reserved for metrics that are.
- $D_{KL}(P \parallel Q) = \sum P(x) \log(\frac{P(x)}{Q(x)})$
- $D_{KL}(Q \parallel P) = \sum Q(x) \log(\frac{Q(x)}{P(x)})$
- That logarithm term is the same, but the point of measuring the divergence between two distributions is that P and Q are different.
- If P and Q are the same, divergence is zero.
- Otherwise, D_{KL} is always positive

So about that K-L divergence asymmetry…

- But we saw an asymmetry with the dice:
  - There was different qualitative behavior between the two true states of the dice:
  - The die with only 5 different sides causes odds to accrue slowly
  - The six-sided die will stumble into infinite odds in favor after an average of 6 rolls.
- Doesn't handle 0 probability elegantly; the infinite negative logarithm makes the expected value infinite, which is inconvenient.
- There is a symmetrized version that satisfies the triangle inequality, goes by the name Jensen-Shannon divergence if you need it. It imagines an equal mixture model of P and Q.