Introduction to Bayesian methods
Lecture 16

David Sontag
New York University

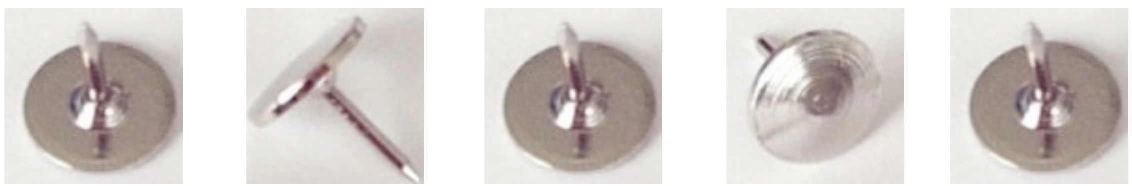
Slides adapted from Luke Zettlemoyer, Carlos Guestrin, Dan Klein, and Vibhav Gogate

Bayesian learning

- Bayesian learning uses **probability** to *model* data and *quantify uncertainty* of predictions
  - Eliminates arbitrary loss functions and regularizers
  - Facilitates incorporation of prior knowledge
  - Gives optimal predictions
    - Allows for decision-theoretic reasoning

Your first consulting job

- A billionaire from the suburbs of Manhattan asks you a question:
  - He says: I have thumbtack, if I flip it, what's the probability it will fall with the nail up?
  - You say: Please flip it a few times:



  - You say: The probability is:
    - P(heads) = 3/5
  - **He says: Why???**
  - You say: Because…

Random Variables

- A random variable is some aspect of the world about which we (may) have uncertainty
  - R = Is it raining?
  - D = How long will it take to drive to work?
  - L = Where am I?

- We denote random variables with capital letters

- Random variables have domains
  - R in {true, false} (sometimes write as {+r, ¬r})
  - D in [0, ∞)
  - L in possible locations, maybe {(0,0), (0,1), …}

Probability Distributions

- Discrete random variables have distributions

$P(T)$

T	P
warm	0.5
cold	0.5

$P(W)$

W	P
sun	0.6
rain	0.1
fog	0.3
meteor	0.0

- A discrete distribution is a TABLE of probabilities of values
- A probability (lower case value) is a single number

$$P(W = rain) = 0.1 \qquad P(rain) = 0.1$$

- Must have:

$$\forall x \, P(x) \geq 0 \qquad \sum_x P(x) = 1$$

Joint Distributions

- A *joint distribution* over a set of random variables: $X_1, X_2, \ldots X_n$ specifies a real number for each assignment:

$$P(X_1 = x_1, X_2 = x_2, \ldots X_n = x_n)$$

$$P(x_1, x_2, \ldots x_n)$$

  – Size of distribution if n variables with domain sizes d?

  – Must obey:

$$P(x_1, x_2, \ldots x_n) \geq 0$$

$$\sum_{(x_1, x_2, \ldots x_n)} P(x_1, x_2, \ldots x_n) = 1$$

$P(T, W)$

T	W	P
hot	sun	0.4
hot	rain	0.1
cold	sun	0.2
cold	rain	0.3

- For all but the smallest distributions, impractical to write out

Marginal Distributions

- Marginal distributions are sub-tables which eliminate variables
- Marginalization (summing out): Combine collapsed rows by adding

$P(T,W)$

T	W	P
hot	sun	0.4
hot	rain	0.1
cold	sun	0.2
cold	rain	0.3

$$P(t) = \sum_w P(t,w)$$

$$P(w) = \sum_t P(t,w)$$

$P(T)$

T	P
hot	0.5
cold	0.5

$P(W)$

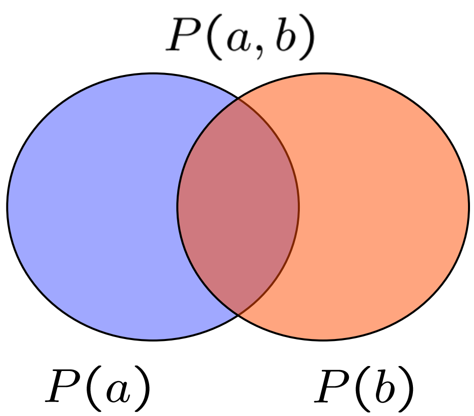
W	P
sun	0.6
rain	0.4

$$P(X_1 = x_1) = \sum_{x_2} P(X_1 = x_1, X_2 = x_2)$$

Conditional Probabilities

- A simple relation between joint and conditional probabilities
  - In fact, this is taken as the *definition* of a conditional probability

$$P(a|b) = \frac{P(a,b)}{P(b)}$$



$$P(T,W)$$

T	W	P
hot	sun	0.4
hot	rain	0.1
cold	sun	0.2
cold	rain	0.3

$$P(W=r|T=c) = ???$$

Conditional Distributions

- Conditional distributions are probability distributions over some variables given fixed values of others

Conditional Distributions

$P(W|T)$

$P(W|T = hot)$

W	P
sun	0.8
rain	0.2

$P(W|T = cold)$

W	P
sun	0.4
rain	0.6

Joint Distribution

$P(T, W)$

T	W	P
hot	sun	0.4
hot	rain	0.1
cold	sun	0.2
cold	rain	0.3

The Product Rule

- Sometimes have conditional distributions but want the joint

$$P(x|y) = \frac{P(x,y)}{P(y)} \quad \Longleftrightarrow \quad P(x,y) = P(x|y)P(y)$$

- Example:

$P(W)$

W	P
sun	0.8
rain	0.2

$P(D|W)$

D	W	P
wet	sun	0.1
dry	sun	0.9
wet	rain	0.7
dry	rain	0.3

$\Longleftrightarrow$

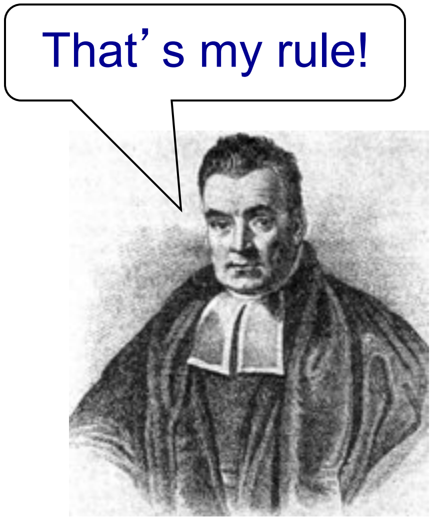
$P(D,W)$

D	W	P
wet	sun	0.08
dry	sun	0.72
wet	rain	0.14
dry	rain	0.06

Bayes' Rule

- Two ways to factor a joint distribution over two variables:

$$P(x,y) = P(x|y)P(y) = P(y|x)P(x)$$



- Dividing, we get:

$$P(x|y) = \frac{P(y|x)}{P(y)}P(x)$$

- Why is this at all helpful?
  - Lets us build one conditional from its reverse
  - Often one conditional is tricky but the other one is simple
  - Foundation of many practical systems (e.g. ASR, MT)

- In the running for most important ML equation!

Returning to thumbtack example…

- P(Heads) = θ, P(Tails) = 1-θ

- Flips are *i.i.d.*: $D=\{x_i \mid i=1\ldots n\},\ P(D \mid \theta) = \Pi_i P(x_i \mid \theta)$
  - Independent events
  - Identically distributed according to Bernoulli distribution
- Sequence *D* of α_H Heads and α_T Tails

$$P(\mathcal{D} \mid \theta) = \theta^{\alpha_H}(1-\theta)^{\alpha_T}$$

Called the "likelihood" of the data under the model

Maximum Likelihood Estimation

- **Data:** Observed set D of α_H Heads and α_T Tails
- **Hypothesis:** Bernoulli distribution
- **Learning:** finding θ is an optimization problem
  - What's the objective function?

$$P(\mathcal{D} \mid \theta) = \theta^{\alpha_H}(1-\theta)^{\alpha_T}$$

- **MLE:** Choose θ to maximize probability of *D*

$$\begin{aligned} \widehat{\theta} &= \arg\max_{\theta} \quad P(\mathcal{D} \mid \theta) \\ &= \arg\max_{\theta} \quad \ln P(\mathcal{D} \mid \theta) \end{aligned}$$

Your first parameter learning algorithm

$$\begin{aligned}\widehat{\theta} &= \arg\max_{\theta} \quad \ln P(\mathcal{D} \mid \theta) \\ &= \arg\max_{\theta} \quad \ln \theta^{\alpha_H}(1-\theta)^{\alpha_T}\end{aligned}$$

- Set derivative to zero, and solve!

$$\frac{d}{d\theta} \ln P(\mathcal{D} \mid \theta) = \frac{d}{d\theta} [\ln \theta^{\alpha_H}(1-\theta)^{\alpha_T}]$$

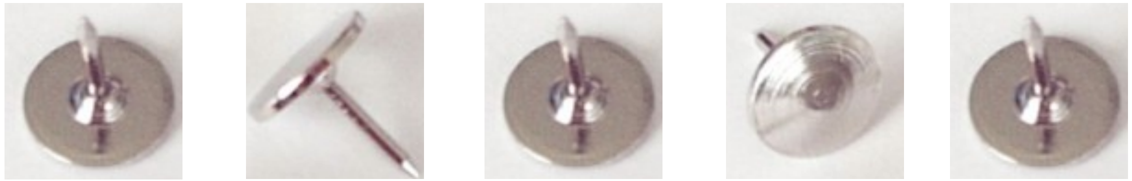
$$= \frac{d}{d\theta} [\alpha_H \ln \theta + \alpha_T \ln(1-\theta)]$$

$$= \alpha_H \frac{d}{d\theta} \ln \theta + \alpha_T \frac{d}{d\theta} \ln(1-\theta)$$

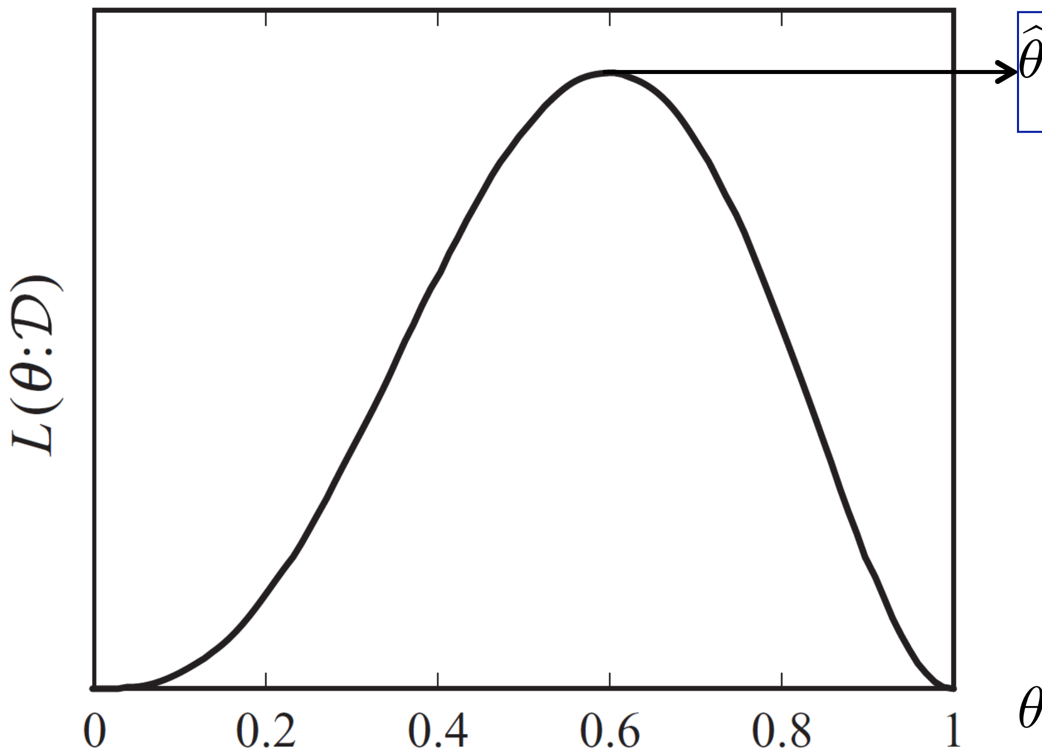
$$= \frac{\alpha_H}{\theta} - \frac{\alpha_T}{1-\theta} \quad = 0$$

$$\boxed{\widehat{\theta}_{MLE} \quad = \quad \frac{\alpha_H}{\alpha_H + \alpha_T}}$$

Data



$$L(\theta;\mathcal{D}) = \ln P(\mathcal{D}|\theta)$$



$$\boxed{\widehat{\theta}_{MLE} = \frac{\alpha_H}{\alpha_H + \alpha_T}}$$

But, how many flips do I need?

$$\widehat{\theta}_{MLE} = \frac{\alpha_H}{\alpha_H + \alpha_T}$$

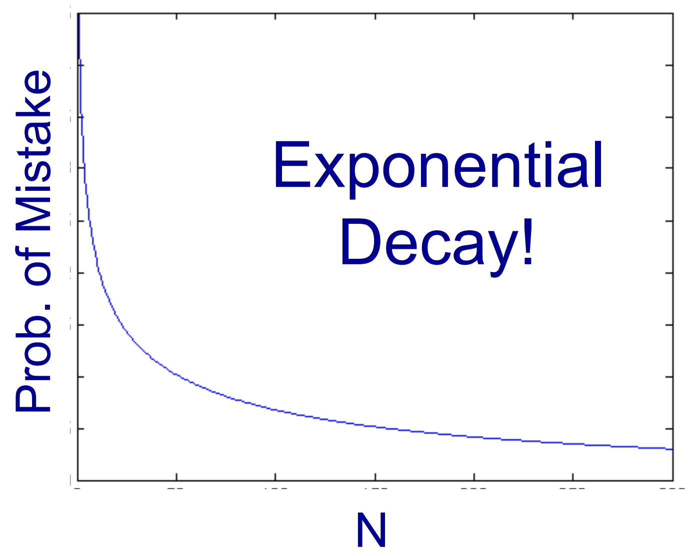
- Billionaire says: I flipped 3 heads and 2 tails.
- You say: θ = 3/5, I can prove it!
- He says: What if I flipped 30 heads and 20 tails?
- You say: Same answer, I can prove it!
- **He says: What's better?**
- You say: Umm… The more the merrier???
- He says: Is this why I am paying you the big bucks???
- You say: I will give you a theoretical bound.

A bound (from Hoeffding's inequality)

- Let $N = \alpha_H + \alpha_T$, and $\widehat{\theta}_{MLE} = \dfrac{\alpha_H}{\alpha_H + \alpha_T}$

- Let θ^* be the true parameter. For any $\varepsilon > 0$,

$$P(|\,\widehat{\theta} - \theta^*\,| \geq \epsilon) \leq 2e^{-2N\epsilon^2}$$



PAC Learning

- PAC: Probably Approximate Correct
- Billionaire says: I want to know the thumbtack θ, within ε = 0.1, with probability at least 1-δ = 0.95.
- How many flips? Or, how big do I set N?

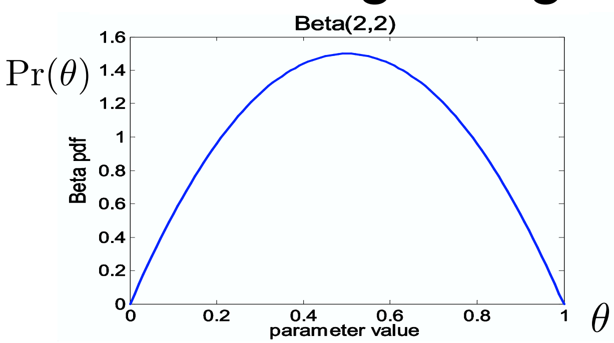
$$P(|\ \widehat{\theta} - \theta^* \ | \geq \underbrace{\epsilon}_{.1}) \ \leq \ \underbrace{2e^{-2N\epsilon^2}}_{= .05}$$

$$N \geq \frac{\ln(2/0.05)}{2 \times 0.1^2} \approx \frac{3.8}{0.02} = 190$$

What if I have prior beliefs?

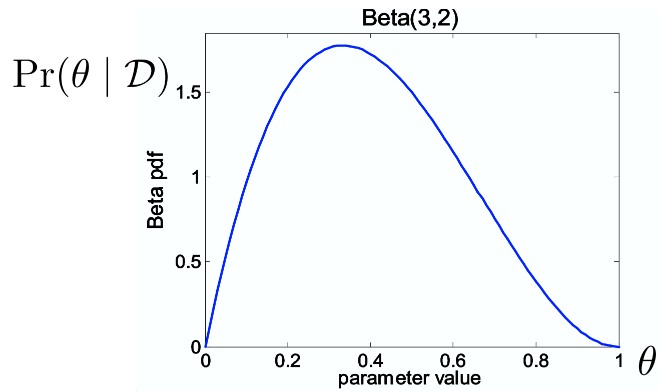
- Billionaire says: Wait, I know that the thumbtack is "close" to 50-50. What can you do for me now?
- **You say: I can learn it the Bayesian way…**
- Rather than estimating a single θ, we obtain a distribution over possible values of θ

In the beginning



Observe flips
e.g.: {tails, tails}

After observations

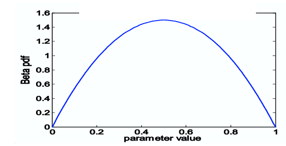


Bayesian Learning

- Use Bayes' rule!

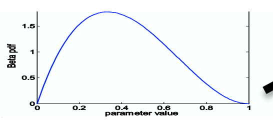
Data Likelihood

Prior



$$P(\theta \mid \mathcal{D}) = \frac{P(\mathcal{D} \mid \theta) P(\theta)}{P(\mathcal{D})}$$

Posterior



Normalization

- Or equivalently: $P(\theta \mid \mathcal{D}) \propto P(\mathcal{D} \mid \theta) P(\theta)$
- For *uniform* priors, this reduces to maximum likelihood estimation!

$$P(\theta) \propto 1 \qquad P(\theta \mid \mathcal{D}) \propto P(\mathcal{D} \mid \theta)$$

Bayesian Learning for Thumbtacks

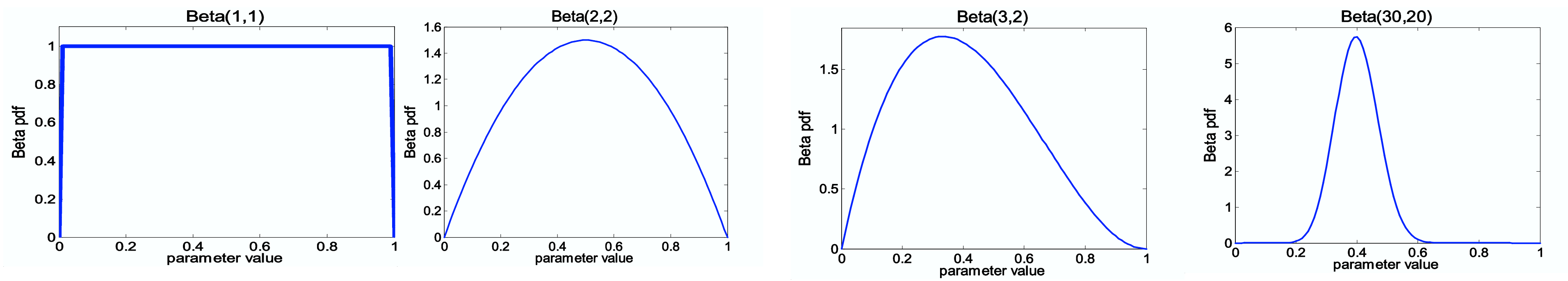
$$P(\theta \mid \mathcal{D}) \;\propto\; P(\mathcal{D} \mid \theta) P(\theta)$$

Likelihood: $P(\mathcal{D} \mid \theta) = \theta^{\alpha_H}(1-\theta)^{\alpha_T}$

- What should the prior be?
  - Represent expert knowledge
  - Simple posterior form
- Conjugate priors:
  - Closed-form representation of posterior
  - **For Binomial, conjugate prior is Beta distribution**

Beta prior distribution – P(θ)

$$P(\theta) = \frac{\theta^{\beta_H - 1}(1-\theta)^{\beta_T - 1}}{B(\beta_H, \beta_T)} \sim Beta(\beta_H, \beta_T)$$



- Likelihood function: $P(\mathcal{D} \mid \theta) = \theta^{\alpha_H}(1-\theta)^{\alpha_T}$
- Posterior: $P(\theta \mid \mathcal{D}) \;\propto\; P(\mathcal{D} \mid \theta) P(\theta)$

$$P(\theta \mid \mathcal{D}) \propto \theta^{\alpha_H}(1-\theta)^{\alpha_T} \; \theta^{\beta_H - 1}(1-\theta)^{\beta_T - 1}$$

$$= \theta^{\alpha_H + \beta_H - 1}(1-\theta)^{\alpha_T + \beta_t + 1}$$

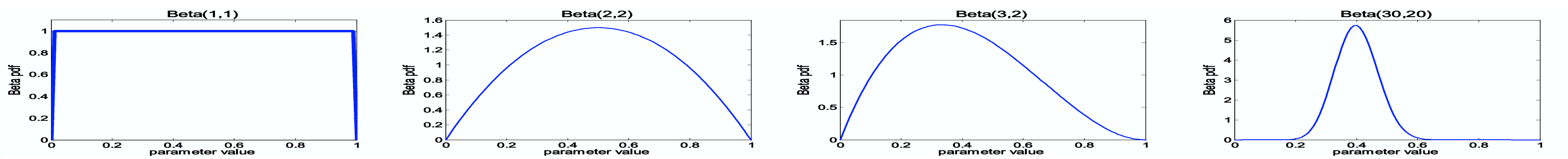
$$= Beta(\alpha_H + \beta_H, \alpha_T + \beta_T)$$

Posterior distribution

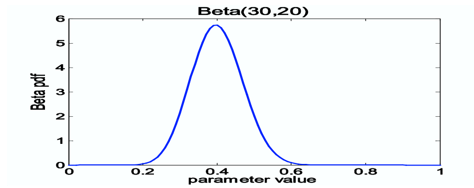
- Prior: $Beta(\beta_H, \beta_T)$
- Data: α_H heads and α_T tails

- Posterior distribution:

$$P(\theta \mid \mathcal{D}) \sim Beta(\beta_H + \alpha_H, \beta_T + \alpha_T)$$



Bayesian Posterior Inference



- Posterior distribution:

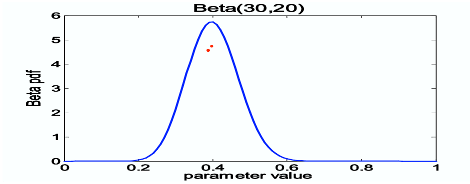
$$P(\theta \mid \mathcal{D}) \sim Beta(\beta_H + \alpha_H, \beta_T + \alpha_T)$$

- Bayesian inference:
  - No longer single parameter
  - For any specific f, the function of interest
  - Compute the expected value of f

$$E[f(\theta)] = \int_0^1 f(\theta) P(\theta \mid \mathcal{D}) d\theta$$

  - Integral is often hard to compute

MAP: Maximum a posteriori approximation



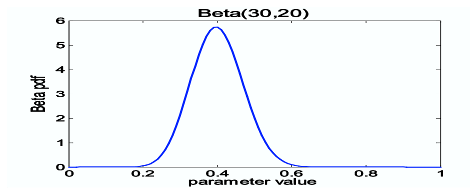
$$P(\theta \mid \mathcal{D}) \sim Beta(\beta_H + \alpha_H, \beta_T + \alpha_T)$$

$$E[f(\theta)] = \int_0^1 f(\theta) P(\theta \mid \mathcal{D}) d\theta$$

- As more data is observed, Beta is more certain
- MAP: use most likely parameter to approximate the expectation

$$\widehat{\theta} = \arg\max_{\theta} P(\theta \mid \mathcal{D})$$

$$E[f(\theta)] \approx f(\widehat{\theta})$$



MAP for Beta distribution

$$P(\theta \mid \mathcal{D}) = \frac{\theta^{\beta_H+\alpha_H-1}(1-\theta)^{\beta_T+\alpha_T-1}}{B(\beta_H+\alpha_H,\beta_T+\alpha_T)} \sim Beta(\beta_H+\alpha_H,\beta_T+\alpha_T)$$

- MAP: use most likely parameter:

$$\widehat{\theta} = \arg\max_{\theta} P(\theta \mid \mathcal{D}) = \frac{\alpha_H+\beta_H-1}{\alpha_H+\beta_H+\alpha_T+\beta_T-2}$$

- Beta prior equivalent to extra thumbtack flips
- As $N \rightarrow \infty$, prior is "forgotten"
- **But, for small sample size, prior is important!**