

Point Estimation

Vibhav Gogate

The University of Texas at Dallas

Some slides courtesy of Carlos Guestrin, Chris Bishop, Dan Weld and Luke Zettlemoyer.

Binary Variables (1)

- Coin flipping: heads=1, tails=0

$$p(x = 1|\mu) = \mu$$

- Bernoulli Distribution

$$\text{Bern}(x|\mu) = \mu^x(1 - \mu)^{1-x}$$

$$\mathbb{E}[x] = \mu$$

$$\text{var}[x] = \mu(1 - \mu)$$

Binary Variables (2)

- N coin flips:

$$p(m \text{ heads} | N, \mu)$$

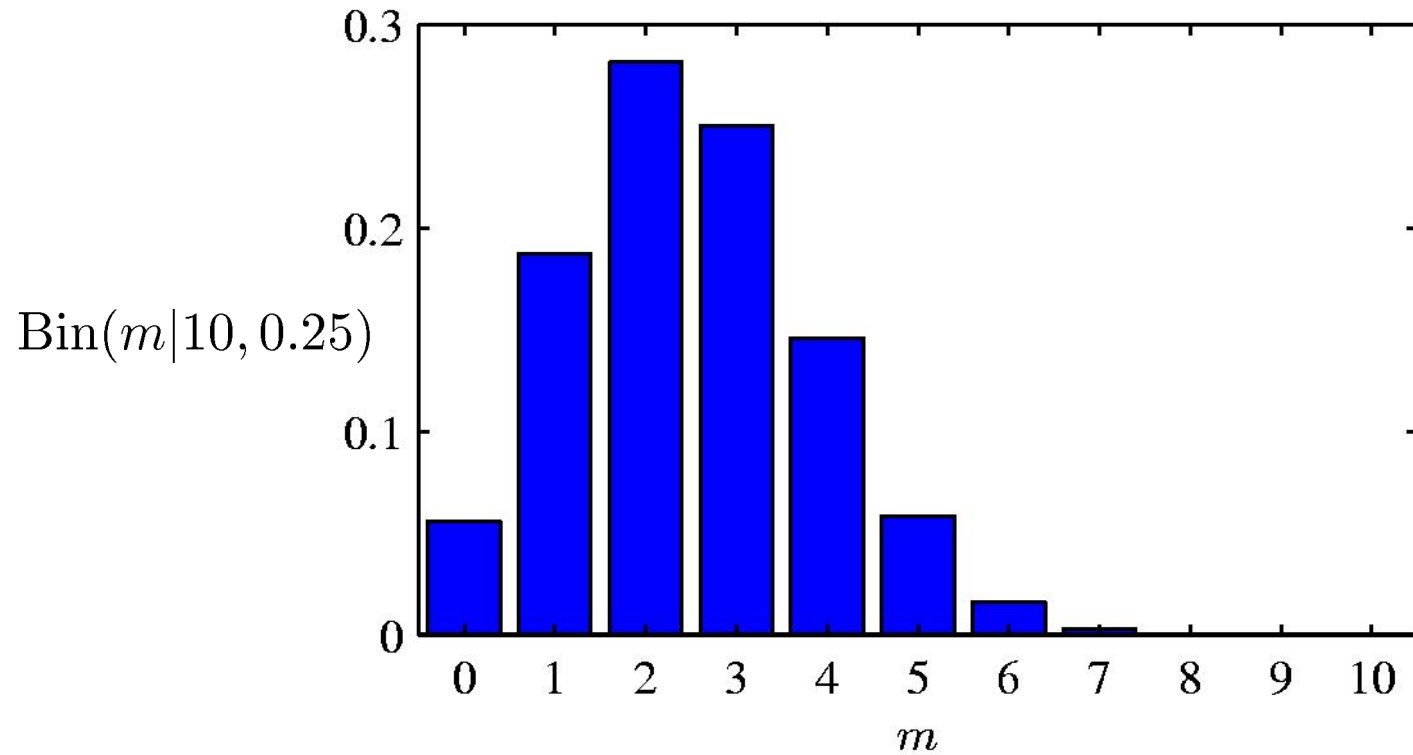
- Binomial Distribution

$$\text{Bin}(m|N, \mu) = \binom{N}{m} \mu^m (1 - \mu)^{N-m}$$

$$\mathbb{E}[m] \equiv \sum_{m=0}^N m \text{Bin}(m|N, \mu) = N\mu$$

$$\text{var}[m] \equiv \sum_{m=0}^N (m - \mathbb{E}[m])^2 \text{Bin}(m|N, \mu) = N\mu(1 - \mu)$$

Binomial Distribution



Your first consulting job

Billionaire in Dallas asks:

- 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:
 - $P(H) = 3/5$
- He says: **Why???**
- You say: Because...

Thumbtack – Binomial Distribution

- $P(\text{Heads}) = \theta$, $P(\text{Tails}) = 1 - \theta$



- Flips are *i.i.d.*:
 - Independent events
 - Identically distributed according to Binomial distribution
- Sequence \mathcal{D} of α_H Heads and α_T Tails

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

Maximum Likelihood Estimation

- **Data:** Observed set D of α_H Heads and α_T Tails
- **Hypothesis:** Binomial 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

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

Your first parameter learning algorithm

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

- Set derivative to zero, and solve!

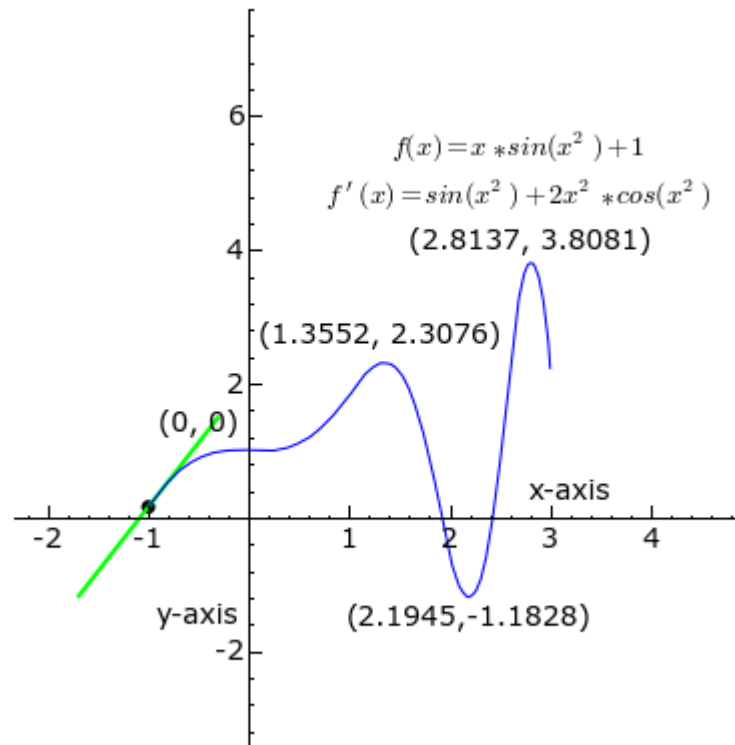
$$\frac{d}{d\theta} \ln P(\mathcal{D} \mid \theta) = \frac{d}{d\theta} [\ln \theta^{\alpha_H} (1 - \theta)^{\alpha_T}]$$

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

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

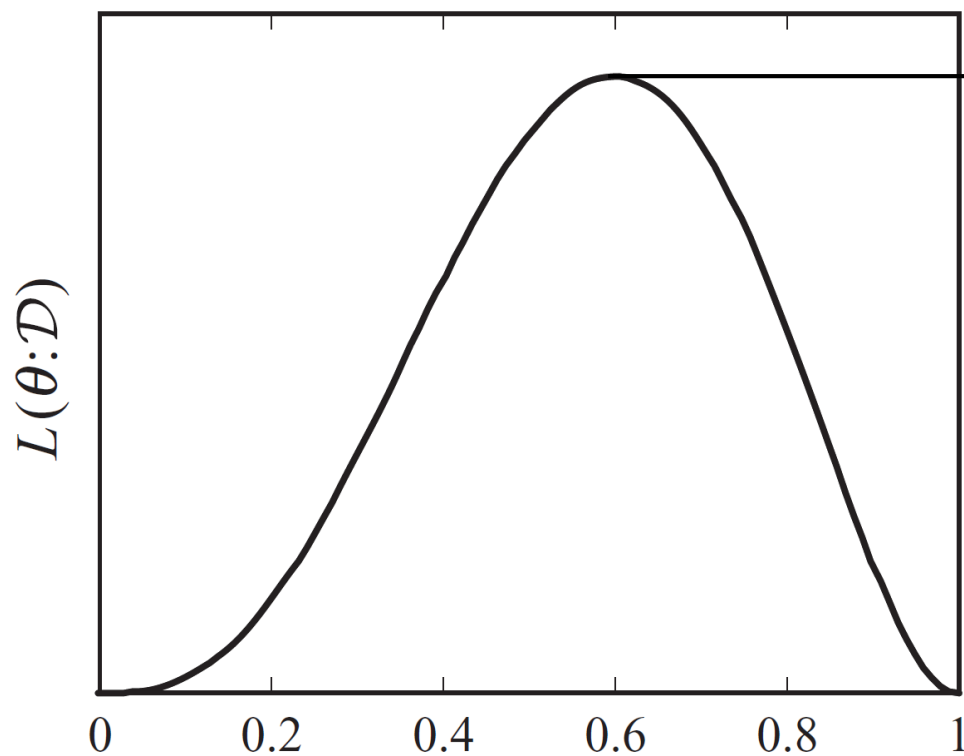
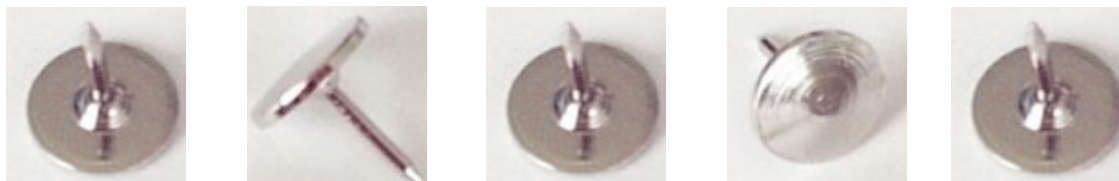
$$= \frac{\alpha_H}{\theta} - \frac{\alpha_T}{1 - \theta} = 0$$

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



At each point, the derivative of is the slope of a line that is tangent to the curve. The line is always tangent to the blue curve; its slope is the derivative. Note derivative is **positive where green**, **negative where red**, and **zero where black**.

Data



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

Parameter Estimation: Summary

- ML for Bernoulli

Given: $\mathcal{D} = \{x_1, \dots, x_N\}$, m heads (1), $N - m$ tails (0)

$$p(\mathcal{D}|\mu) = \prod_{n=1}^N p(x_n|\mu) = \prod_{n=1}^N \mu^{x_n} (1 - \mu)^{1-x_n}$$

$$\ln p(\mathcal{D}|\mu) = \sum_{n=1}^N \ln p(x_n|\mu) = \sum_{n=1}^N \{x_n \ln \mu + (1 - x_n) \ln(1 - \mu)\}$$

$$\mu_{\text{ML}} = \frac{1}{N} \sum_{n=1}^N x_n = \frac{m}{N}$$

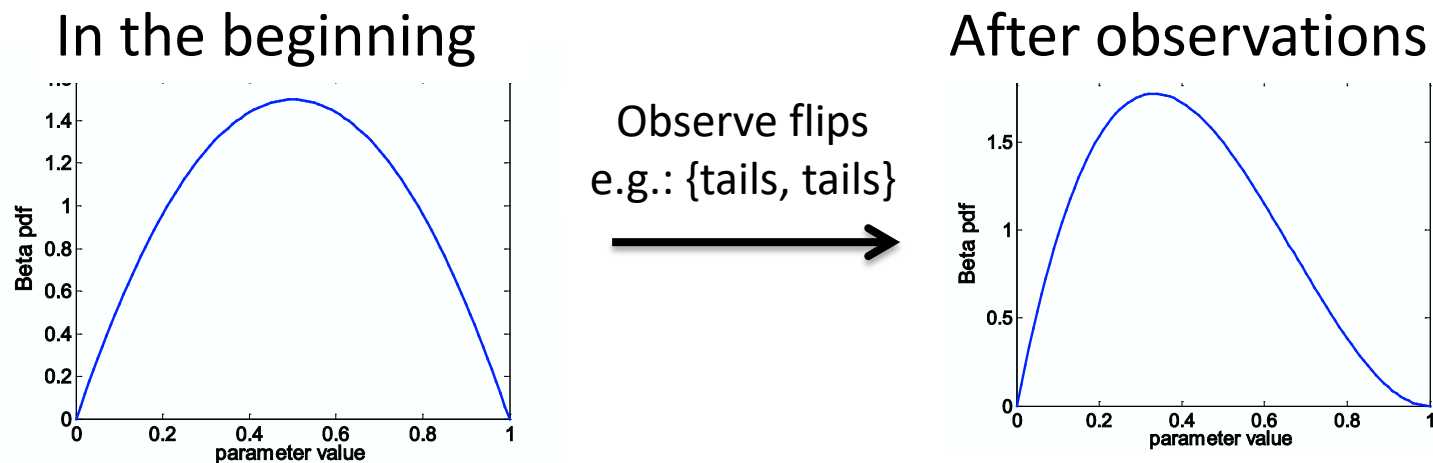
But, how many flips do I need?

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

- Billionaire says: I flipped 3 heads and 2 tails.
- You say: $\theta = 3/5$, 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.

What if I have prior beliefs?

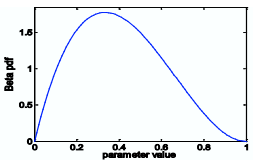
- Billionaire says: Wait, I know that the thumbtack is “close” to 50-50. What can you do for me now?
- You say: I can learn it the Bayesian way...
- Rather than estimating a single θ , we obtain a distribution over possible values of θ



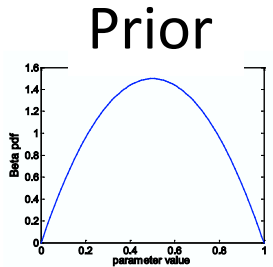
Bayesian Learning

Use Bayes rule:

Posterior


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

Data Likelihood



Prior

Normalization

Or equivalently:

$$P(\theta | \mathcal{D}) \propto P(\mathcal{D} | \theta) P(\theta)$$

Also, for uniform priors:

→ reduces to MLE objective

$$P(\theta) \propto 1 \quad P(\theta | \mathcal{D}) \propto P(\mathcal{D} | \theta)$$

Bayesian Learning for Thumbtacks

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

Likelihood function is Binomial:

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

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

Beta Distribution



- Distribution over $\mu \in [0, 1]$.

$$B(a, b) = \frac{\Gamma(a + b)}{\Gamma(a)\Gamma(b)}$$

$$\text{Beta}(\mu|a, b) = \frac{\Gamma(a + b)}{\Gamma(a)\Gamma(b)} \mu^{a-1} (1 - \mu)^{b-1}$$

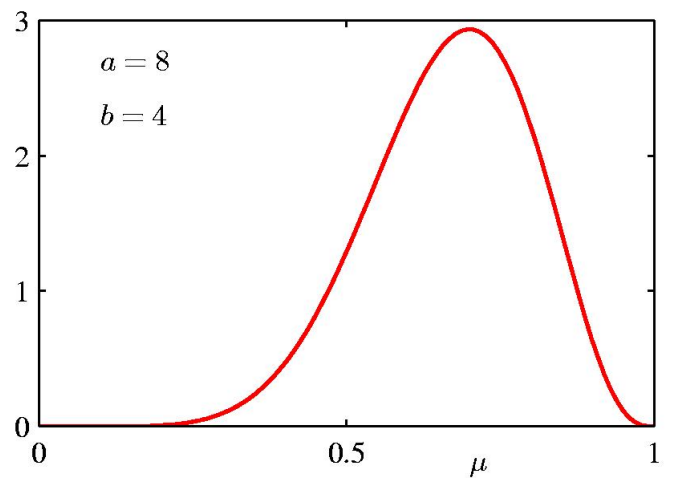
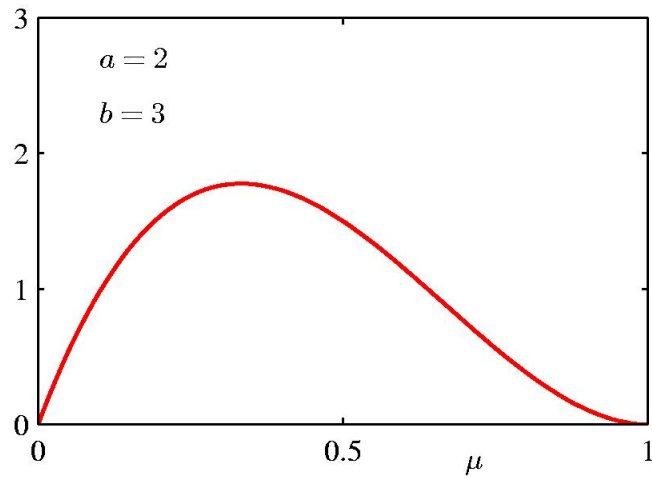
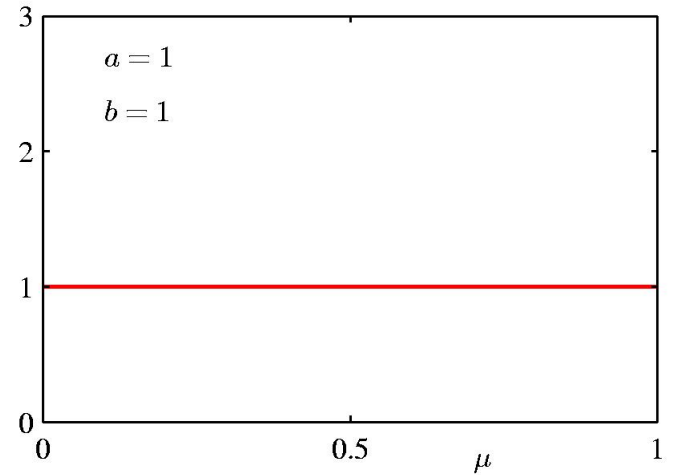
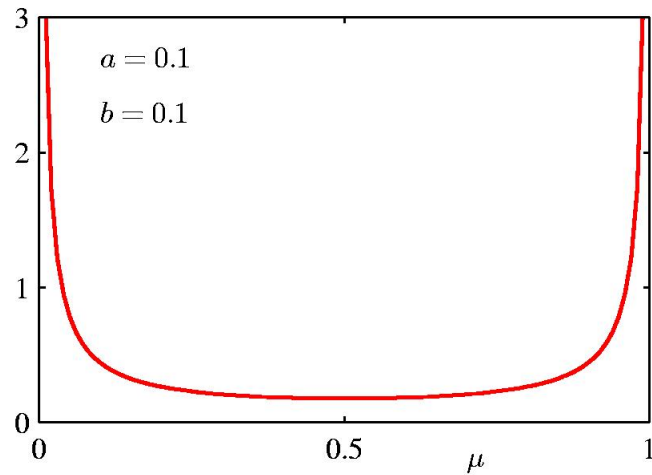
$$\mathbb{E}[\mu] = \frac{a}{a + b}$$

$$\text{var}[\mu] = \frac{ab}{(a + b)^2(a + b + 1)}$$

$$B(a, b) = \int_0^1 u^{a-1} (1 - u)^{b-1} du, \quad a > 0, b > 0$$

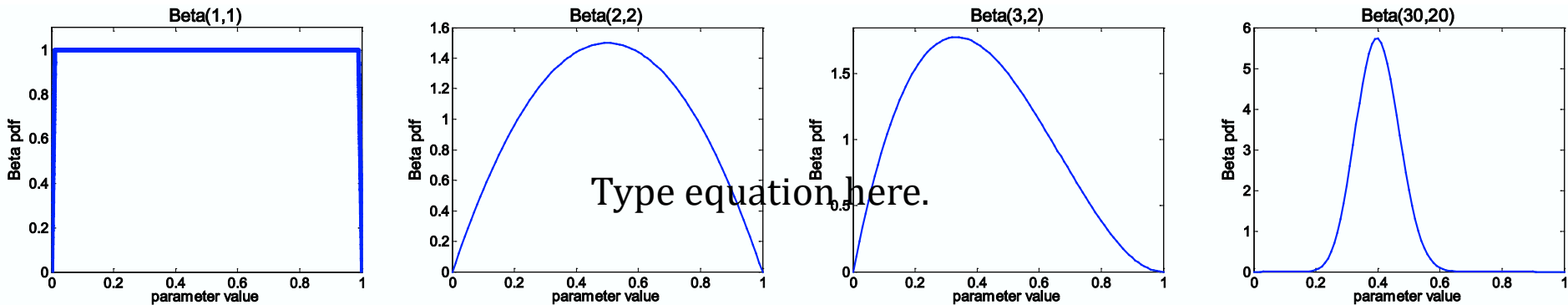
$$\Gamma(a) = \int_0^\infty u^{a-1} e^{-u} du$$

Beta Distribution



Beta prior distribution – $P(\theta)$

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



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

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

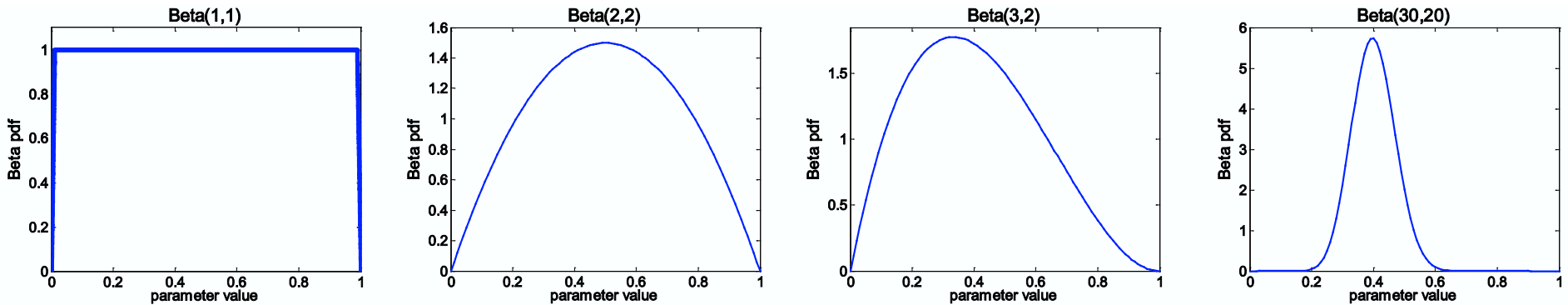
$$= \theta^{\alpha_H+\beta_H-1}(1-\theta)^{\alpha_T+\beta_T-1}$$

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

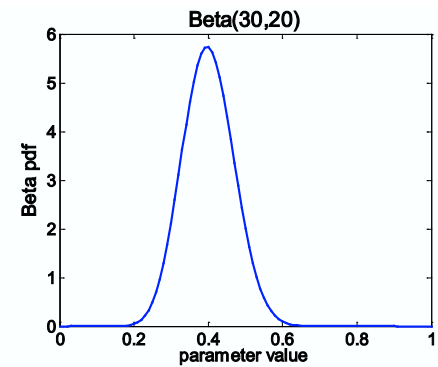
Posterior Distribution

- 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:

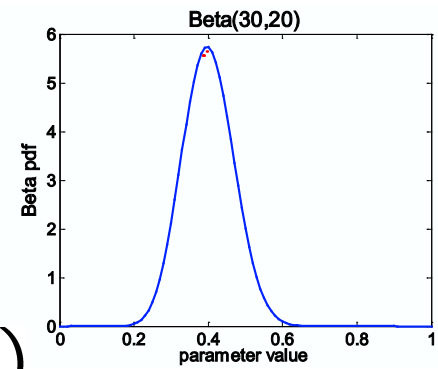
$$P(\theta \mid \mathcal{D}) \sim \text{Beta}(\beta_H + \alpha_H, \beta_T + \alpha_T)$$

- 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



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

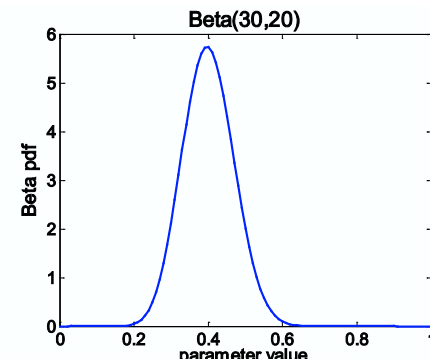
$$E[f(\theta)] = \int_0^1 f(\theta) P(\theta \mid \mathcal{D}) d\theta$$

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

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

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

MAP for Beta distribution



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

MAP: use most likely parameter:

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

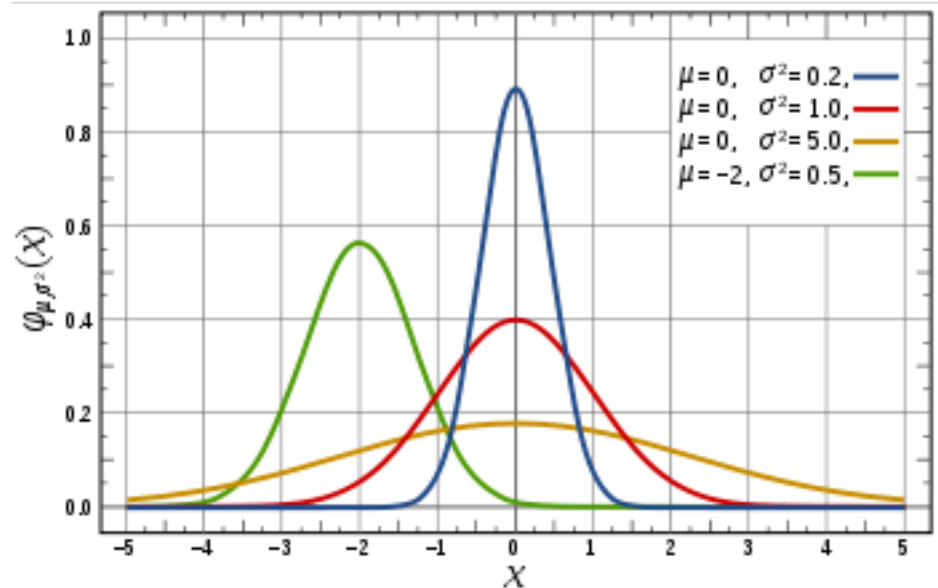
Beta prior equivalent to extra thumbtack flips


As $N \rightarrow \infty$, prior is “forgotten”

But, for small sample size, prior is important!

What about continuous variables?

- Billionaire says: If I am measuring a continuous variable, what can you do for me?
- You say: Let me tell you about Gaussians...





$$P(x \mid \mu, \sigma) = \frac{1}{\sigma \sqrt{2\pi}} e^{\frac{-(x-\mu)^2}{2\sigma^2}}$$

Some properties of Gaussians

- Affine transformation (multiplying by scalar and adding a constant) are Gaussian

- $X \sim N(\mu, \sigma^2)$

- $Y = aX + b \rightarrow Y \sim N(a\mu + b, a^2\sigma^2)$

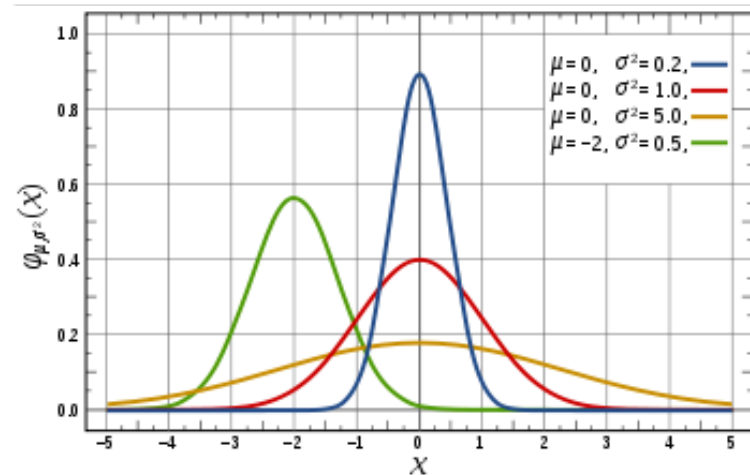
- Sum of Gaussians is Gaussian

- $X \sim N(\mu_X, \sigma_X^2)$

- $Y \sim N(\mu_Y, \sigma_Y^2)$

- $Z = X + Y \rightarrow Z \sim N(\mu_X + \mu_Y, \sigma_X^2 + \sigma_Y^2)$

- Easy to differentiate, as we will see soon!



Learning a Gaussian

- Collect a bunch of data
 - Hopefully, i.i.d. samples
 - e.g., exam scores
- Learn parameters
 - Mean: μ
 - Variance: σ

X_i $i =$	Exam Score
0	85
1	95
2	100
3	12
...	...
99	89

$$P(x \mid \mu, \sigma) = \frac{1}{\sigma \sqrt{2\pi}} e^{\frac{-(x-\mu)^2}{2\sigma^2}}$$

MLE for Gaussian: $P(x \mid \mu, \sigma) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$

- Prob. of i.i.d. samples $D=\{x_1, \dots, x_N\}$:

$$P(\mathcal{D} \mid \mu, \sigma) = \left(\frac{1}{\sigma\sqrt{2\pi}} \right)^N \prod_{i=1}^N e^{-\frac{(x_i-\mu)^2}{2\sigma^2}}$$

$$\mu_{MLE}, \sigma_{MLE} = \arg \max_{\mu, \sigma} P(\mathcal{D} \mid \mu, \sigma)$$

- Log-likelihood of data:

$$\begin{aligned} \ln P(\mathcal{D} \mid \mu, \sigma) &= \ln \left[\left(\frac{1}{\sigma\sqrt{2\pi}} \right)^N \prod_{i=1}^N e^{-\frac{(x_i-\mu)^2}{2\sigma^2}} \right] \\ &= -N \ln \sigma\sqrt{2\pi} - \sum_{i=1}^N \frac{(x_i - \mu)^2}{2\sigma^2} \end{aligned}$$

Your second learning algorithm: MLE for mean of a Gaussian

- What's MLE for mean?

$$\begin{aligned}\frac{d}{d\mu} \ln P(\mathcal{D} \mid \mu, \sigma) &= \frac{d}{d\mu} \left[-N \ln \sigma \sqrt{2\pi} - \sum_{i=1}^N \frac{(x_i - \mu)^2}{2\sigma^2} \right] \\ &= \frac{d}{d\mu} \left[-N \ln \sigma \sqrt{2\pi} \right] - \sum_{i=1}^N \frac{d}{d\mu} \left[\frac{(x_i - \mu)^2}{2\sigma^2} \right] \\ &= - \sum_{i=1}^N \frac{(x_i - \mu)}{\sigma^2} = 0 \\ &= - \sum_{i=1}^N x_i + N\mu = 0\end{aligned}$$



$$\hat{\mu}_{MLE} = \frac{1}{N} \sum_{i=1}^N x_i$$

MLE for variance

- Again, set derivative to zero:

$$\begin{aligned}\frac{d}{d\sigma} \ln P(\mathcal{D} \mid \mu, \sigma) &= \frac{d}{d\sigma} \left[-N \ln \sigma \sqrt{2\pi} - \sum_{i=1}^N \frac{(x_i - \mu)^2}{2\sigma^2} \right] \\ &= \frac{d}{d\sigma} \left[-N \ln \sigma \sqrt{2\pi} \right] - \sum_{i=1}^N \frac{d}{d\sigma} \left[\frac{(x_i - \mu)^2}{2\sigma^2} \right] \\ &= -\frac{N}{\sigma} + \sum_{i=1}^N \frac{(x_i - \mu)^2}{\sigma^3} = 0\end{aligned}$$



$$\hat{\sigma}_{MLE}^2 = \frac{1}{N} \sum_{i=1}^N (x_i - \hat{\mu})^2$$

Learning Gaussian parameters

- MLE:

$$\hat{\mu}_{MLE} = \frac{1}{N} \sum_{i=1}^N x_i$$

$$\hat{\sigma}_{MLE}^2 = \frac{1}{N} \sum_{i=1}^N (x_i - \hat{\mu})^2$$

- BTW. MLE for the variance of a Gaussian is **biased**
 - Expected result of estimation is **not** true parameter!
 - Unbiased variance estimator:

$$\hat{\sigma}_{unbiased}^2 = \frac{1}{N-1} \sum_{i=1}^N (x_i - \hat{\mu})^2$$

Bayesian learning of Gaussian parameters

- Conjugate priors
 - Mean: Gaussian prior
 - Variance: Wishart Distribution

- Prior for mean:

$$P(\mu \mid \eta, \lambda) = \frac{1}{\lambda \sqrt{2\pi}} e^{\frac{-(\mu - \eta)^2}{2\lambda^2}}$$