Machine Learning II

DS 4420 - Spring 2020

MLE, MAP, & Graphical models

Byron C. Wallace

Probability Spaces

Definition: A probability space (Ω, F, P) consists of

- A sample space Ω (i.e. the set of *outcomes*)
- A set of events F (i.e. the set possible sets)
- A probability measure P (maps events to probabilities)

Axioms of Probability

$$P : \mathcal{F} \rightarrow \mathbb{R} \qquad P(E) \geq 0 \;\; \forall\, E \in \mathcal{F} \qquad P(\Omega) = 1$$

$$P(E_1, E_2) = P(E_1) + P(E_2) \;\; \text{when} \;\; E_1 \cap E_2 = \emptyset$$

Conditional Probabilities

- **Definition:** Joint Probability

Outcomes in both *A* and *B*

$$P(A, B) = P(A \cap B)$$

Events (i.e. sets of outcomes)

- **Definition:** Conditional Probability

$$P(A \mid B) = \frac{P(A, B)}{P(B)}$$

Probability Density Functions

- **Problem:** If X is a *continuous* variable, then $P(X{=}x)$ is **0** for any outcome ***x***

Single Outcome

$$X \sim \mathrm{Normal}(0,1) \qquad \begin{array}{l} P(X=\pi)=0 \\ P(3.1 \leq X \leq 3.2) \neq 0 \end{array}$$

Event

- **Solution:** Define a density function as a derivative

Capital P for probability

$$p_X(x) = \lim_{\delta \to 0} \frac{P(x-\delta < X < x+\delta)}{2\delta}$$

Small p for density

Objectives in Learning

$$\mathbb{E}_{p_\theta(x|y)}[f(x,y)]$$

Model parameters (e.g. neural network weights)

Unknown Variables

Known Variables

Quantity of Interest

Setting	**Self-driving Cars**	**Medical Diagnosis**
$p_\theta(y,x)$	Model for pedestrian behavior	Model for diseases / symptoms
y	Pedestrian motion	Symptoms / Test results
x	Will pedestrian cross road?	Condition of patient
$f(y,x)$	Chance of accident	Treatment outcome

Maximum Likelihood Estimation

MLE Framework

Observe some data $X = x_1, ..., x_n \quad x_i \in R^d$

We assume this is a random draw (sample) from some parameterized distribution P_θ

MLE Framework

Observe some data $X = x_1, ..., x_n \quad x_i \in R^d$

We assume this is a random draw (sample) from some parameterized distribution P_θ

Goal: find θ

MLE Framework

Observe some data $X = x_1, ..., x_n \qquad x_i \in R^d$

We assume this is a random draw (sample) from some parameterized distribution P_θ

Goal: find θ

In MLE we pick

$$\theta_{\mathrm{MLE}} = \mathrm{argmax}_\theta P(X|\theta)$$

MLE Framework

Observe some data $X = x_1, ..., x_n \qquad x_i \in R^d$

We assume this is a random draw (sample) from some parameterized distribution P_θ

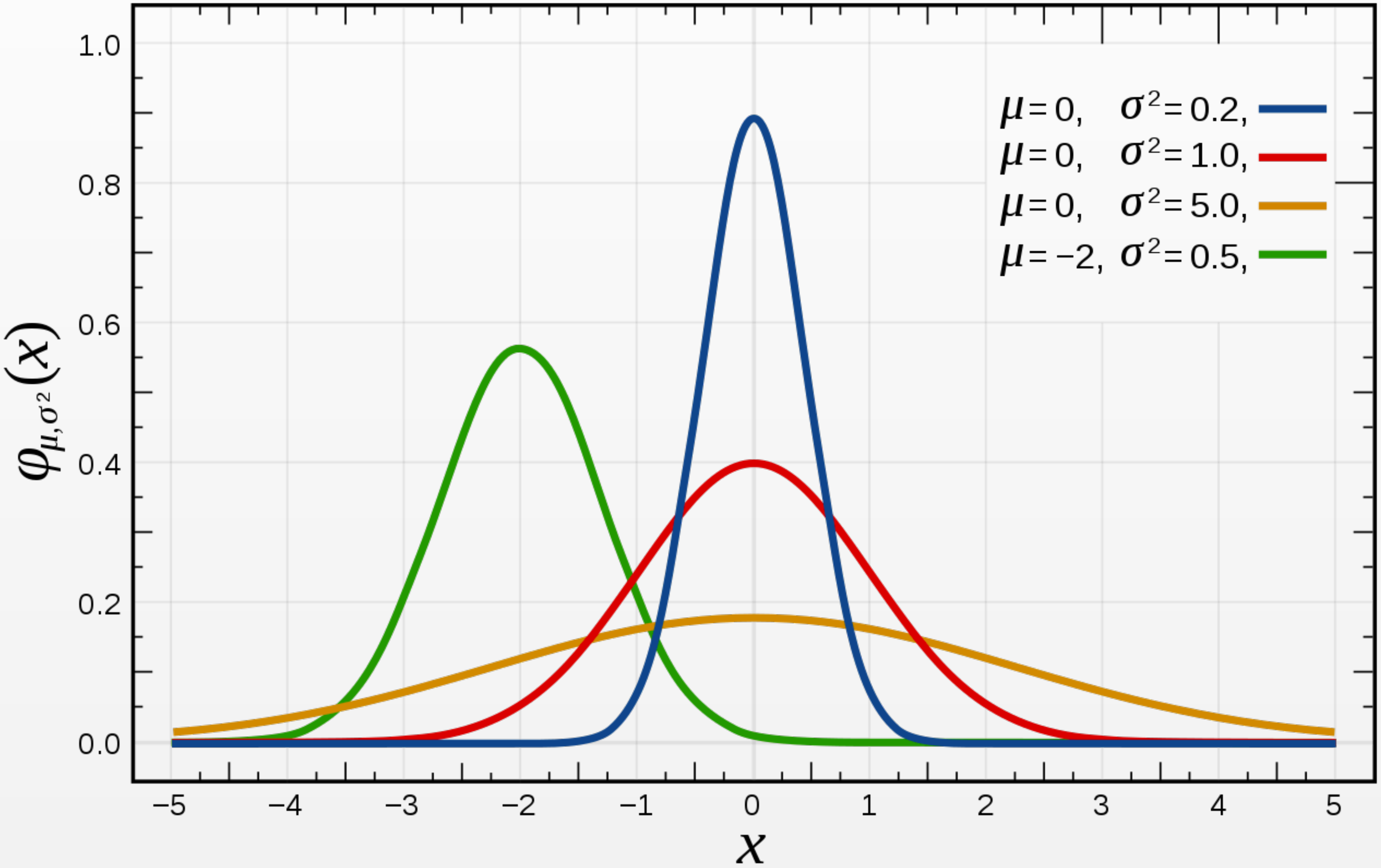
Goal: find θ

In MLE we pick

$$\theta_{\mathrm{MLE}} = \mathrm{argmax}_\theta P(X|\theta)$$

$$P(X|\theta) = \prod_i P(x_i|\theta)$$

Normal



1.0
0.8
0.6
0.4
0.2
0.0
$\varphi_{\mu,\sigma^2}(x)$
$\mu=0, \quad \sigma^2=0.2,$
$\mu=0, \quad \sigma^2=1.0,$
$\mu=0, \quad \sigma^2=5.0,$
$\mu=-2, \quad \sigma^2=0.5,$
−5 −4 −3 −2 −1 0 1 2 3 4 5
x

$$x \sim N(\mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\{-1 \frac{1}{2\sigma^2}(x-\mu)^2\}$$

$$\theta = \{\mu, \sigma^2\}$$

$$x \sim N(\mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\{-1 \frac{1}{2\sigma^2} (x-\mu)^2\}$$

$$\theta = \{\mu, \sigma^2\}$$

$$p(D|\theta) = p(x_1, ..., x_N) = \prod_{i=1}^{N} p(x_i|\theta)$$

$$x \sim N(\mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\{-1 \frac{1}{2\sigma^2}(x-\mu)^2\}$$

$$\theta = \{\mu, \sigma^2\}$$

$$p(D|\theta) = p(x_1, ..., x_N) = \prod_{i=1}^{N} p(x_i|\theta)$$

Let's work this out…

Discrete/Categorical Distribution

Example: Loaded Dice

$$p_\theta(x = k) = \theta_k$$

$$x \in \{1, 2, 3, 4, 5, 6\}$$

$$\theta = \{\theta_1, \ldots, \theta_6\}$$

Discrete/Categorical Distribution

Example: Loaded Dice

$$p_\theta(x = k) = \theta_k$$

$x \in \{1, 2, 3, 4, 5, 6\}$

$\theta = \{\theta_1, \ldots, \theta_6\}$

Equivalent Notation: Indicator variables

$$p_\theta(x) = \prod_{k=1}^{K} \theta_k^{x_k}$$

$x_k := I[x = k]$

$[1, 0, 0, 0, 0, 0] : 1$

$[0, 1, 0, 0, 0, 0] : 2$

$\ldots$

$[0, 0, 0, 0, 0, 1] : 6$

Discrete/Categorical Distribution

Example: Loaded Dice

$$p_\theta(x = k) = \theta_k \qquad x \in \{1, 2, 3, 4, 5, 6\}$$

$$\theta = \{\theta_1, \ldots, \theta_6\}$$

Equivalent Notation: Indicator variables

$$p_\theta(x) = \prod_{k=1}^{K} \theta_k^{x_k} \qquad x_k := I[x = k]$$

$$[1, 0, 0, 0, 0, 0] : 1$$
$$[0, 1, 0, 0, 0, 0] : 2$$
$$\ldots$$
$$[0, 0, 0, 0, 0, 1] : 6$$

Question: If you perform **1000** rolls and get **200** outcomes **x=6**, then how would you estimate θ_6?

Maximum Likelihood Estimation

Likelihood of ***N*** independent events:

$$p_\theta(x_1, \ldots, x_N) = \prod_{n=1}^{N} p_\theta(x_n) \qquad p_\theta(x_n) = \prod_{k=1}^{K} \theta_k^{x_{n,k}}$$

Maximum likelihood estimation

$$\begin{aligned} \theta^* &= \underset{\theta}{\operatorname{argmax}}\ p_\theta(x_1, \ldots, x_N) \\ &= \underset{\theta}{\operatorname{argmax}}\ \log p_\theta(x_1, \ldots, x_N) \end{aligned}$$

Maximum Likelihood Estimation

Likelihood of *N* independent events:

$$p_\theta(x_1, \ldots, x_N) = \prod_{n=1}^{N} p_\theta(x_n) \qquad p_\theta(x_n) = \prod_{k=1}^{K} \theta_k^{x_{n,k}}$$

Maximum likelihood estimation

$$\begin{aligned} \theta^* &= \underset{\theta}{\operatorname{argmax}}\ p_\theta(x_1, \ldots, x_N) \\ &= \underset{\theta}{\operatorname{argmax}}\ \log p_\theta(x_1, \ldots, x_N) \end{aligned}$$

Problem: Express $\boldsymbol{\theta^*}$ in terms of $\{\boldsymbol{x_1}, \ldots, \boldsymbol{x_N}\}$

hint: solve for $\nabla_\theta \log p_\theta(x_1, \ldots, x_N) = 0$

Maximum Likelihood Estimation

Likelihood of ***N*** independent events:

$$p_\theta(x_1, \ldots, x_N) = \prod_{n=1}^{N} p_\theta(x_n) \qquad p_\theta(x_n) = \prod_{k=1}^{K} \theta_k^{x_{n,k}}$$

Maximum likelihood estimation

$$\begin{aligned} \theta^* &= \operatorname*{argmax}_\theta \; p_\theta(x_1, \ldots, x_N) \\ &= \operatorname*{argmax}_\theta \; \log p_\theta(x_1, \ldots, x_N) \\ &= \operatorname*{argmax}_\theta \sum_{k=1}^{K} N_k \log \theta_k \qquad N_k = \sum_{n=1}^{N} x_{n,k} \end{aligned}$$

(known as cross-entropy loss in neural net libraries)

Likelihood: Bernoulli

(a discrete distribution with outcomes x=0 and x=1)

$$
\begin{aligned}
\mathrm{Bern}(x|\mu) &= \mu^x(1-\mu)^{1-x} \\
\mathbb{E}[x] &= \mu \\
\mathrm{var}[x] &= \mu(1-\mu) \\
\mathrm{mode}[x] &= \begin{cases} 1 & \text{if } \mu \geqslant 0.5, \\ 0 & \text{otherwise} \end{cases}
\end{aligned}
$$

$$\theta_1^{x_1}\theta_2^{x_2}$$

$$x_2 = (1 - x_1)$$

$$\theta_2 = (1 - \theta_1)$$

$$x \in \{0, 1\} \qquad \mu \in [0, 1]$$

Likelihood: Bernoulli

(a discrete distribution with outcomes x=0 and x=1)

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

Question: What is the likelihood of N trials?
What is ML estimate for μ?

Likelihood: Bernoulli

(a discrete distribution with outcomes x=0 and x=1)

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

Question: What is the likelihood of N trials?
What is ML estimate for μ?

Let's work this out in the MLE framework

Likelihood: Bernoulli

(a discrete distribution with outcomes x=0 and x=1)

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

Question: What is the likelihood of N trials? What is ML estimate for μ?

$$p(x_1, \ldots, x_N \mid \mu) = \mu^{N_1} (1-\mu)^{N_0}$$

$$\underset{\mu}{\operatorname{argmax}} \; p(x \mid \mu) = \frac{N_1}{N}$$

$$N_1 = \sum_{n=1}^{N} x_n$$

$$N_0 = \sum_{n=1}^{N} (1 - x_n)$$

Problems with MLE?

- Provides a *point estimate*; no notion of uncertainty around parameters

- Does not naturally incorporate prior beliefs (maybe a pro, if you're a frequentist?)

- Other thoughts?

Maximum A Posteriori (MAP)

Problem: Maximum likelihood can overfit when data is limited

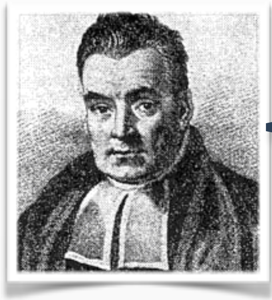
One Solution: Place a prior over parameters

$$
\begin{aligned}
\theta^* &= \operatorname*{argmax}_\theta \ \log p(\theta \mid x) \\
&= \operatorname*{argmax}_\theta \ \log [p(\theta \mid x) p(x)] \\
&= \operatorname*{argmax}_\theta \ \log [p(x \mid \theta) p(\theta)] \\
&= \operatorname*{argmax}_\theta \ \log p(x \mid \theta) + \log p(\theta)
\end{aligned}
$$

ML Objective
(same as minimizing CE loss)

Regularization

Intuition: Bayesian Posterior



$$p(\theta \mid x) = p(x \mid \theta) p(\theta) / p(x)$$

Posterior   Likelihood   Prior

Example: Biased Coin

x Observed data (flip outcomes)

x = 1: Heads

x = 0: Tails

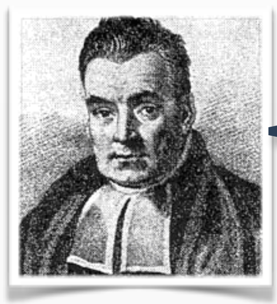
θ Unknown variable (coin bias)

θ = 0.5: No bias

θ = 1.0: Always heads

θ = 0.0: Always tails

Intuition: Bayesian Posterior

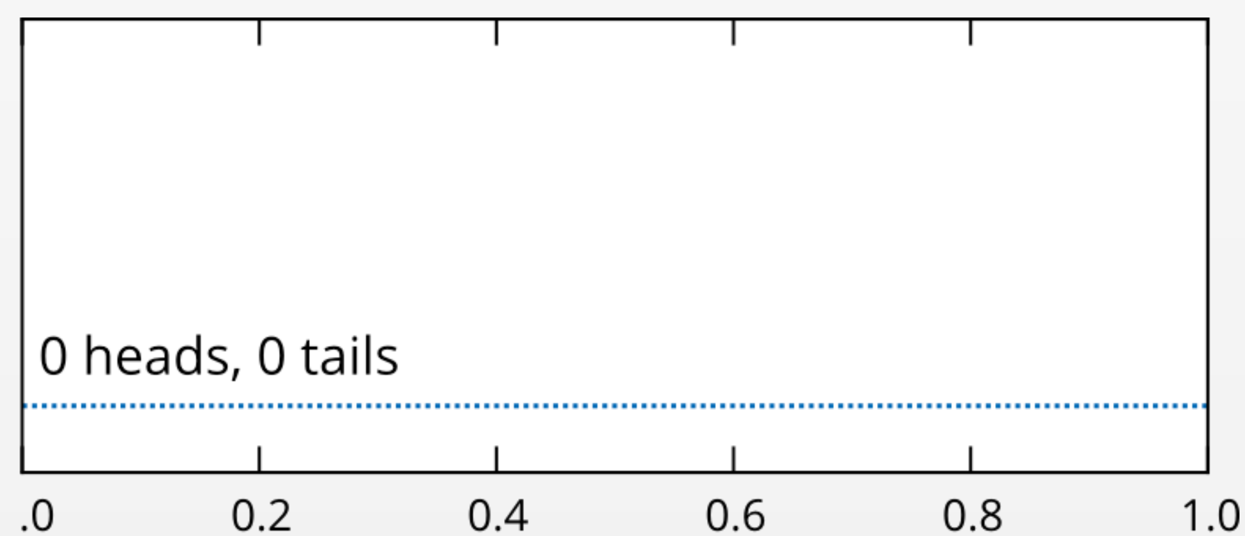


$$p(\theta \mid x) = p(x \mid \theta) p(\theta) / p(x)$$

Posterior Likelihood Prior

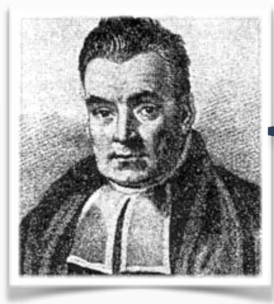
Uninformative Prior

$p(\theta)$



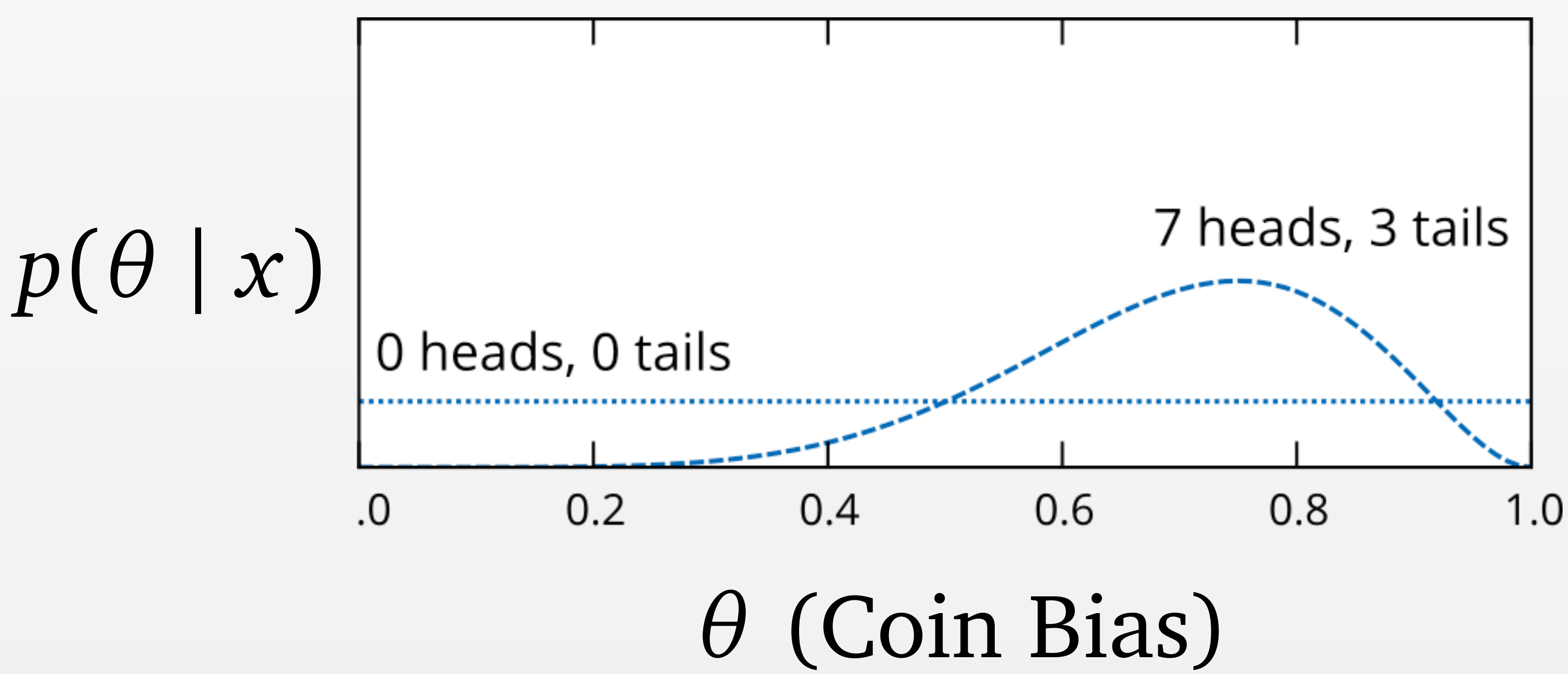
θ (Coin Bias)

Intuition: Bayesian Posterior

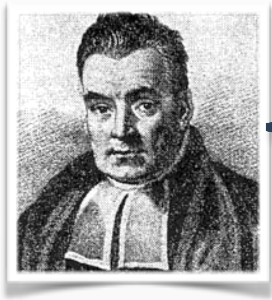


$$\underbrace{p(\theta \mid x)}_{\text{Posterior}} = \underbrace{p(x \mid \theta)}_{\text{Likelihood}}\underbrace{p(\theta)}_{\text{Prior}}/p(x)$$

Posterior after 10 trials



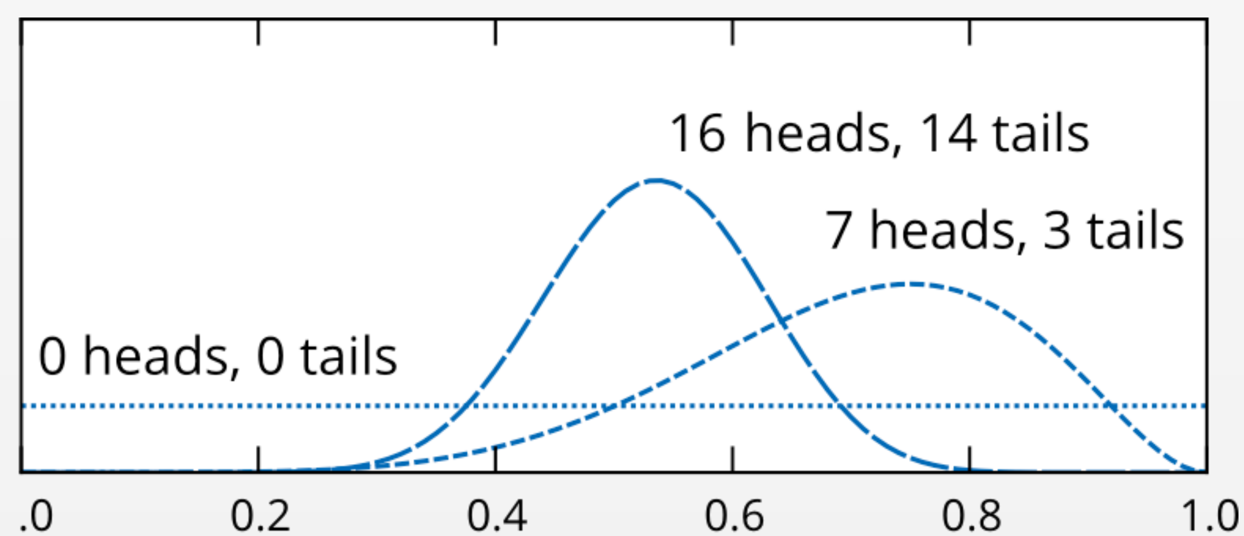
Intuition: Bayesian Posterior



$$\underbrace{p(\theta \mid x)}_{\text{Posterior}} = \underbrace{p(x \mid \theta)}_{\text{Likelihood}}\underbrace{p(\theta)}_{\text{Prior}}/p(x)$$

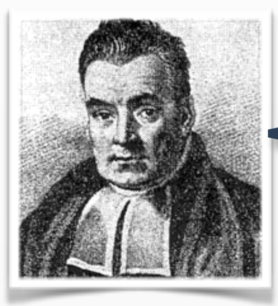
Posterior after 30 trials

$p(\theta \mid x)$



θ (Coin Bias)

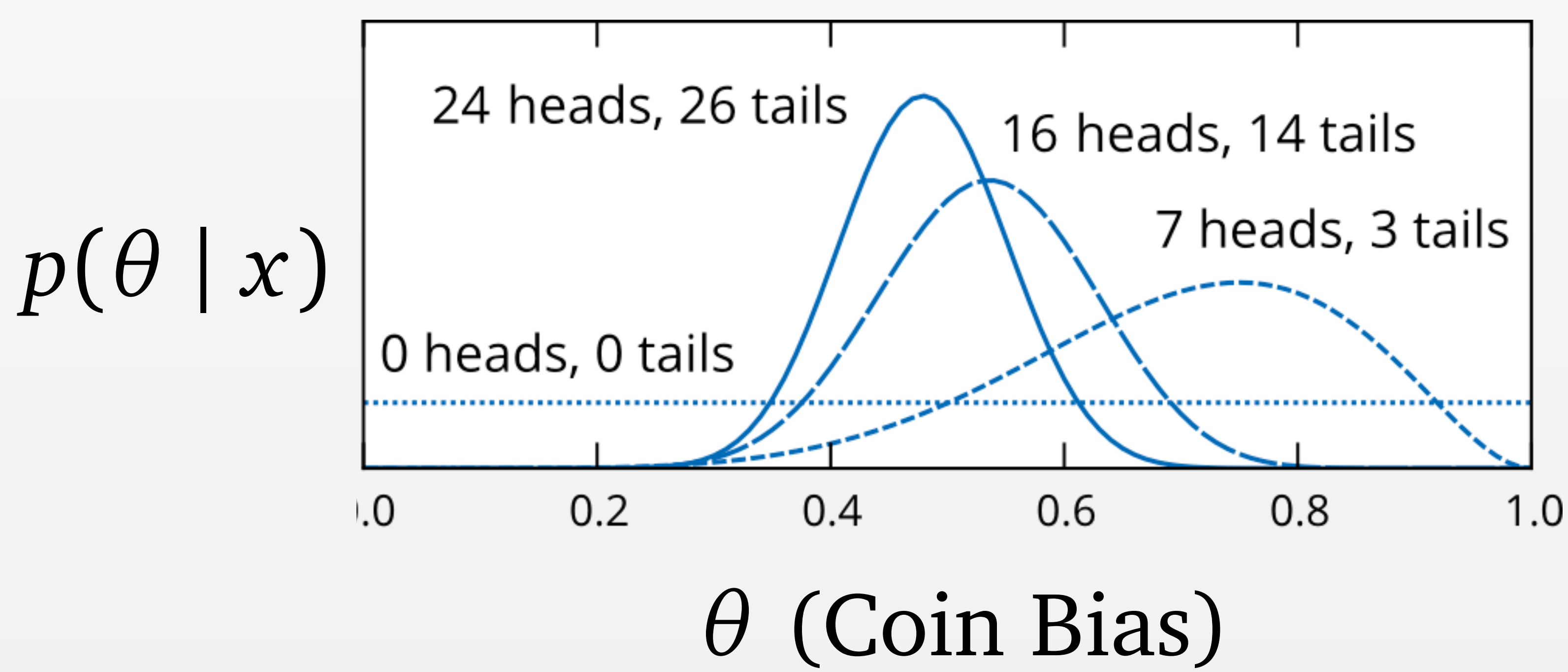
Intuition: Bayesian Posterior



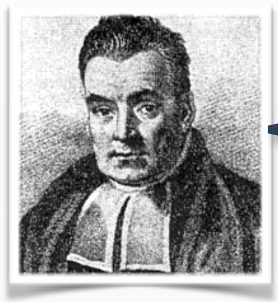
$$p(\theta \mid x) = p(x \mid \theta)p(\theta)/p(x)$$

Posterior Likelihood Prior

Posterior after 50 trials



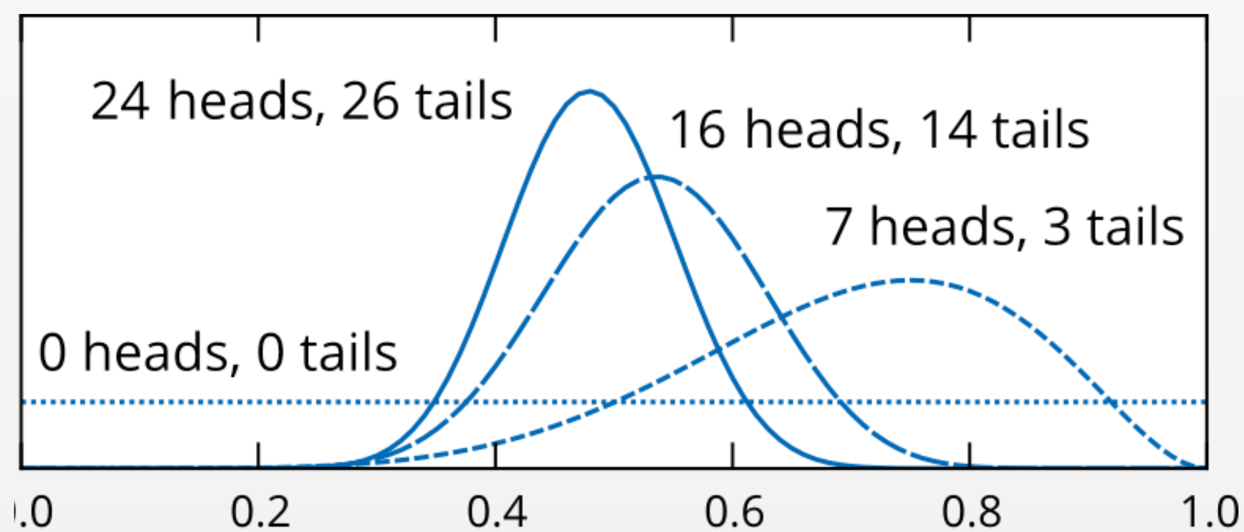
Intuition: Bayesian Posterior



$$\underbrace{p(\theta \mid x)}_{\text{Posterior}} = \underbrace{p(x \mid \theta)}_{\text{Likelihood}}\,\underbrace{p(\theta)}_{\text{Prior}}/p(x)$$

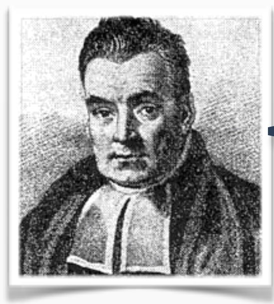
Posterior after 50 trials

$p(\theta \mid x)$



θ (Coin Bias)

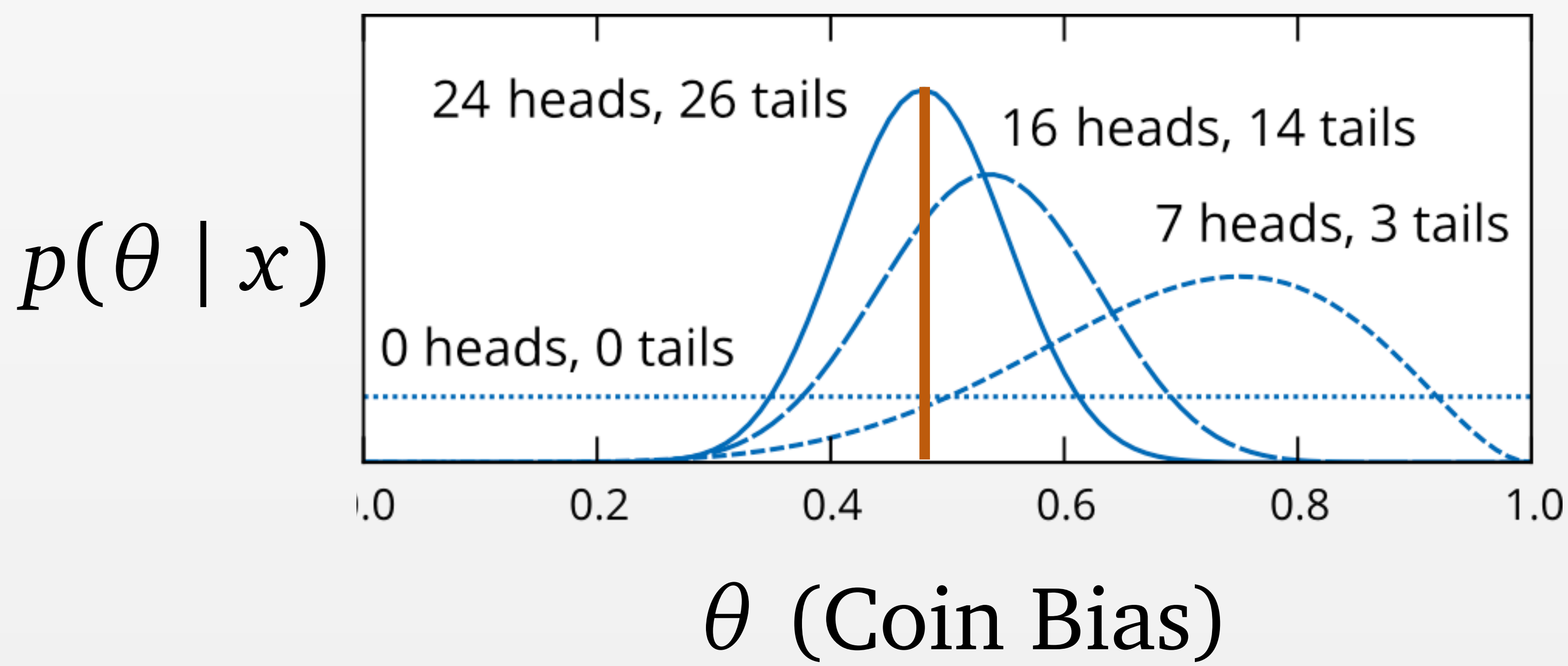
Intuition: Bayesian Posterior



$$p(\theta \mid x) = p(x \mid \theta)p(\theta)/p(x)$$

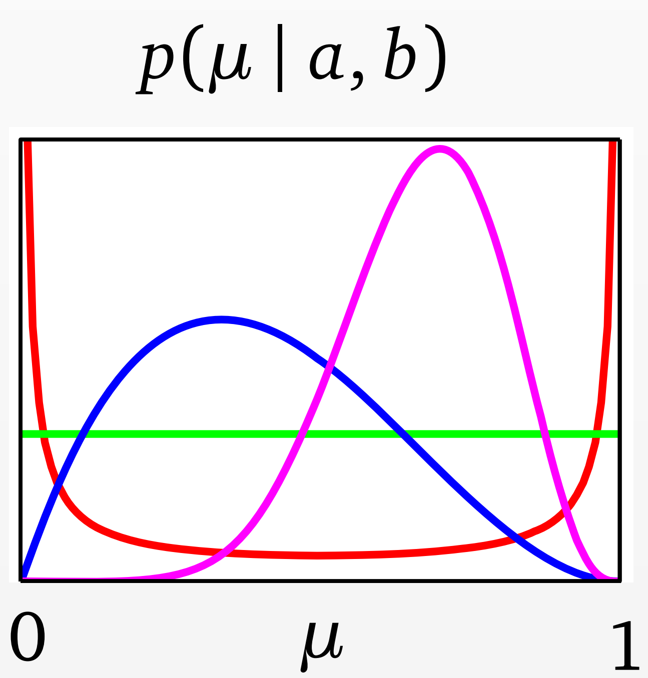
Posterior Likelihood Prior

MAP estimate after 50 trials



Prior: Beta Distribution

(a distribution on values in the range 0 to 1)



$$
\begin{aligned}
\mathrm{Beta}(\mu|a,b) &= \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\mu^{a-1}(1-\mu)^{b-1} \\
\mathbb{E}[\mu] &= \frac{a}{a+b} \\
\mathrm{var}[\mu] &= \frac{ab}{(a+b)^2(a+b+1)} \\
\mathrm{mode}[\mu] &= \frac{a-1}{a+b-2}.
\end{aligned}
$$

ML and MAP estimation

Maximum Likelihood Estimation

$$\theta^* = \operatorname*{argmax}_{\theta} \; p_\theta(x_1, \ldots, x_N)$$
$$= \operatorname*{argmax}_{\theta} \; \log p_\theta(x_1, \ldots, x_N)$$

Maximum A Posterior Estimation

$$\theta^* = \operatorname*{argmax}_{\theta} \; \log p(\theta \mid x)$$
$$= \operatorname*{argmax}_{\theta} \; \log p(x \mid \theta) + \log p(\theta)$$

ML Objective
(same as minimizing a loss function)

Regularization

The Marginal Likelihood

Marginal likelihood
(a.k.a. the Evidence)

Likelihood

Prior

$$p(x \mid a, b) = \int p(x \mid \mu)\, p(\mu \mid a, b)\, d\mu$$

$$= \mathbb{E}_{p(\mu \mid a, b)}[p(x \mid \mu)]$$

"Average" likelihood of data (under prior)

$$p(x \mid a, b) = \frac{B(N_1 + a, N_0 + b)}{B(a, b)}$$

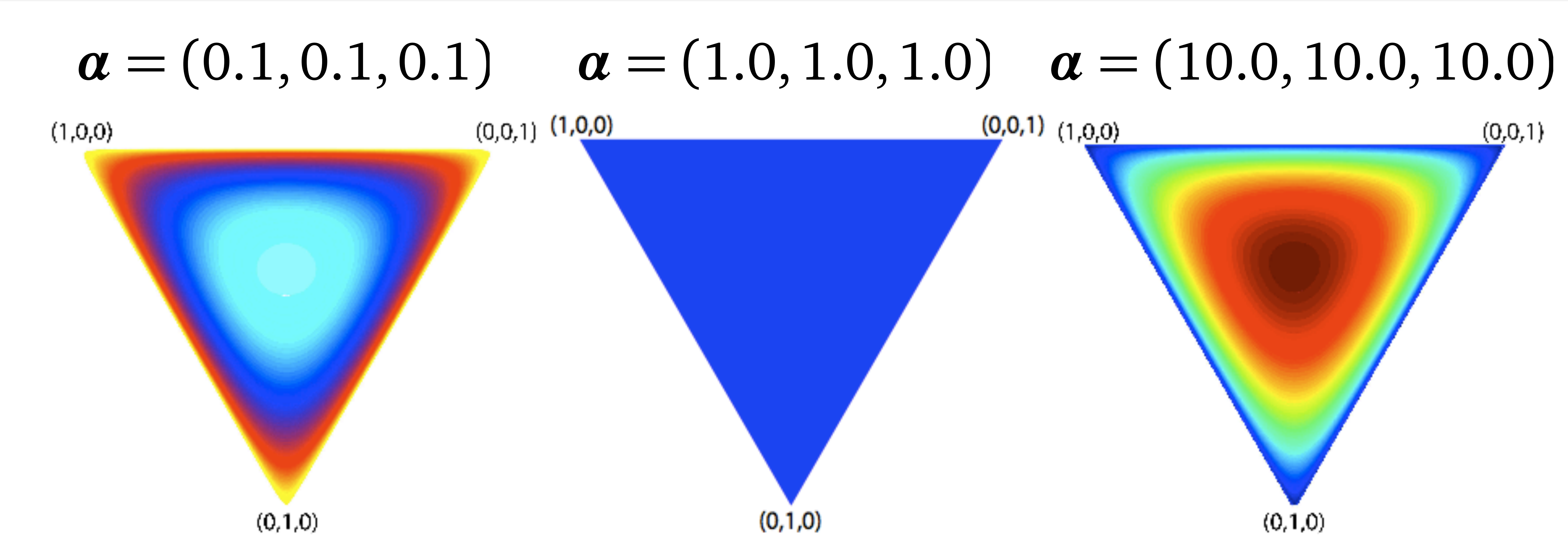
Can calculate in closed form for conjugate priors

Let’s see this in action…

Example: Dirichlet Distribution

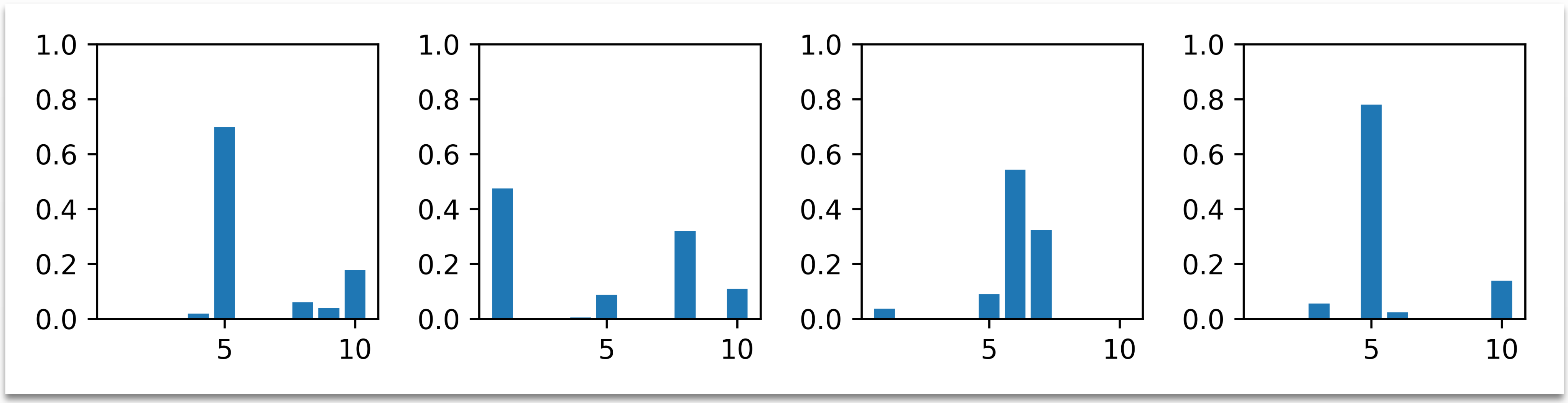
(Conjugate to the Discrete Distribution)

$$p(\boldsymbol{\theta}) = \frac{1}{B(\boldsymbol{\alpha})} \prod_{k=1}^{K} \theta_k^{\alpha_k - 1} \qquad B(\boldsymbol{\alpha}) := \frac{\prod_{k=1}^{K} \Gamma(\alpha_k)}{\Gamma\left(\sum_{k=1}^{K} \alpha_k\right)}$$

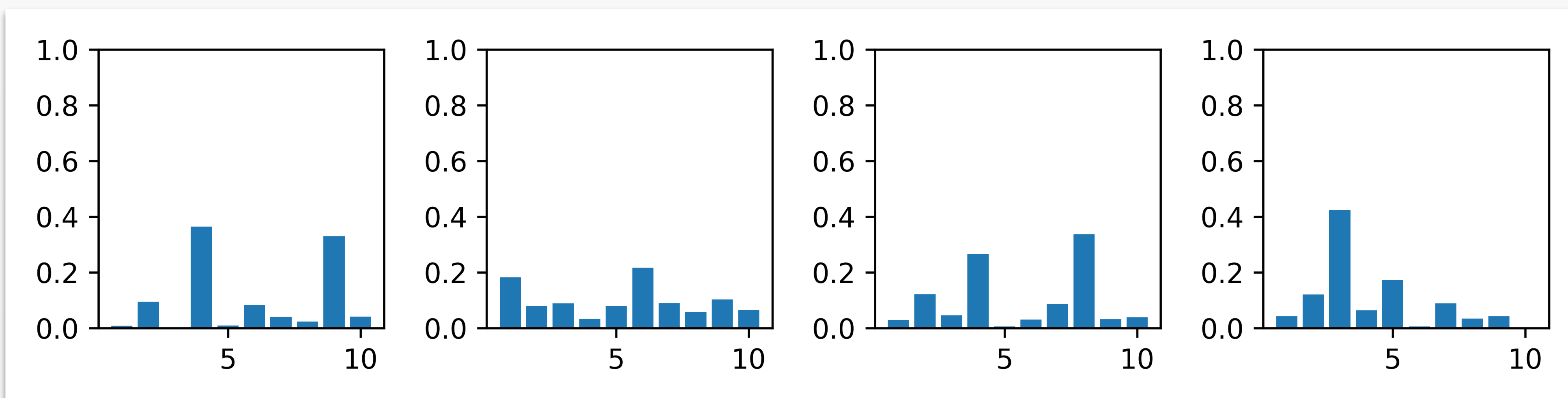


Example: Dirichlet Distribution

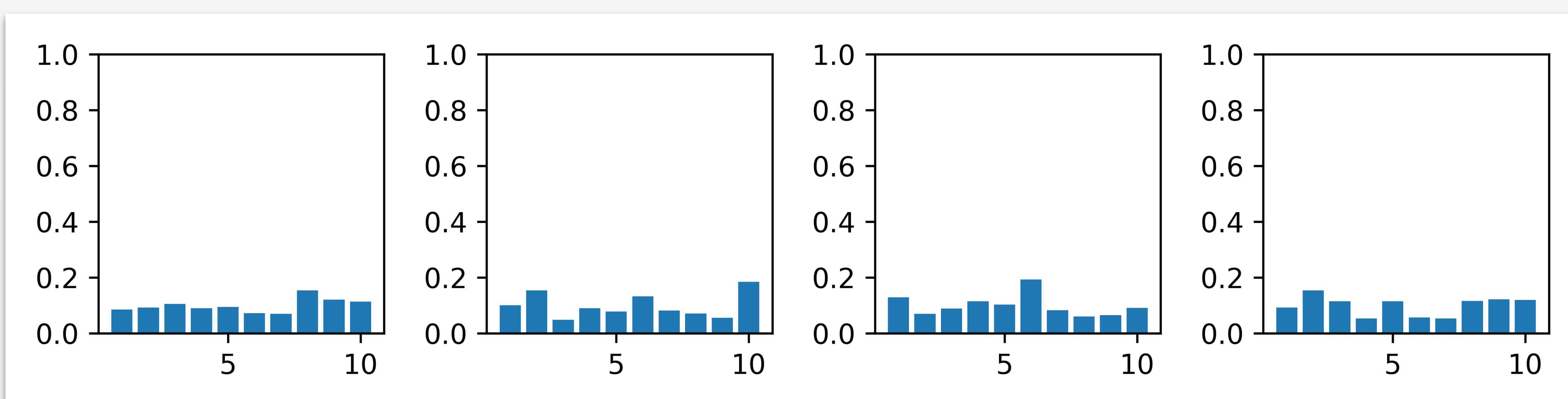
$\alpha_k = 0.1$



$\alpha_k = 1.0$



$\alpha_k = 10.0$



Conjugate Priors

https://en.wikipedia.org/wiki/Conjugate_prior

Likelihood	Model parameters	Conjugate prior distribution	Prior hyperparameters	Posterior hyperparameters	Interpretation of hyperparameters	Posterior predictive[note 4]
Normal with known variance σ^2	μ (mean)	Normal	μ_0, σ_0^2	$\frac{1}{\frac{1}{\sigma_0^2}+\frac{n}{\sigma^2}}\left(\frac{\mu_0}{\sigma_0^2}+\frac{\sum_{i=1}^n x_i}{\sigma^2}\right), \left(\frac{1}{\sigma_0^2}+\frac{n}{\sigma^2}\right)^{-1}$	mean was estimated from observations with total precision (sum of all individual precisions) $1/\sigma_0^2$ and with sample mean μ_0	$\mathcal{N}(\tilde{x}\|\mu_0', {\sigma_0^2}' + \sigma^2)$[5]
Normal with known precision τ	μ (mean)	Normal	μ_0, τ_0	$\frac{\tau_0\mu_0+\tau\sum_{i=1}^n x_i}{\tau_0+n\tau}, \tau_0+n\tau$	mean was estimated from observations with total precision (sum of all individual precisions) τ_0 and with sample mean μ_0	$\mathcal{N}\left(\tilde{x} \mid \mu_0', \frac{1}{\tau_0'}+\frac{1}{\tau}\right)$[5]
Normal with known mean μ	σ^2 (variance)	Inverse gamma	α, β [note 5]	$\alpha+\frac{n}{2}, \beta+\frac{\sum_{i=1}^n (x_i-\mu)^2}{2}$	variance was estimated from 2α observations with sample variance β/α (i.e. with sum of squared deviations 2β, where deviations are from known mean μ)	$t_{2\alpha'}(\tilde{x}\|\mu, \sigma^2 = \beta'/\alpha')$ [5]
Normal with known mean μ	σ^2 (variance)	Scaled inverse chi-squared	ν, σ_0^2	$\nu+n, \frac{\nu\sigma_0^2+\sum_{i=1}^n (x_i-\mu)^2}{\nu+n}$	variance was estimated from ν observations with sample variance σ_0^2	$t_{\nu'}(\tilde{x}\|\mu, {\sigma_0^2}')$[5]
Normal with known mean μ	τ (precision)	Gamma	α, β[note 3]	$\alpha+\frac{n}{2}, \beta+\frac{\sum_{i=1}^n (x_i-\mu)^2}{2}$	precision was estimated from 2α observations with sample variance β/α (i.e. with sum of squared deviations 2β, where deviations are from known mean μ)	$t_{2\alpha'}(\tilde{x} \mid \mu, \sigma^2 = \beta'/\alpha')$ [5]

Graphical Models

Motivation: Spam Filtering

Features: Words in E-mail

$$x = \begin{bmatrix} 1 \\ 0 \\ 0 \\ \vdots \\ 1 \\ \vdots \\ 0 \end{bmatrix} \quad \begin{array}{l} \text{a} \\ \text{aardvark} \\ \text{aardwolf} \\ \vdots \\ \text{buy} \\ \vdots \\ \text{zygmurgy} \end{array}$$

Labels: Spam or not Spam

$$y_n \in \{0, 1\}$$

Input: Labeled Data

$$\{(x_1, y_1), \ldots, (x_N, y_N)\}$$

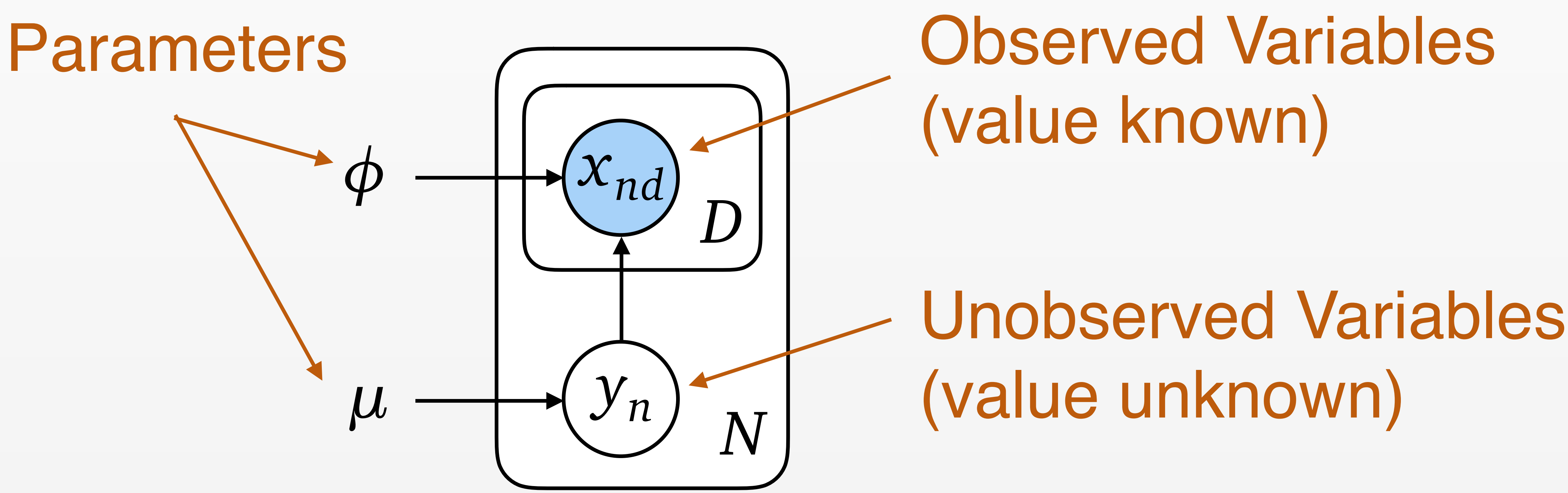
Goal: Predict Unlabeled Data

Naive Bayes
(on board)

Graphical Model: Naive Bayes

$$y_n \sim \text{Bernoulli}(\mu) \qquad n = 1, \ldots, N$$

$$\boldsymbol{x}_{nd} \mid y_n = k \sim \text{Bernoulli}(\phi_{kd}) \qquad k = 0, 1 \quad d = 1, \ldots, D$$



$$p(x, y \mid \mu, \phi) = \prod_{n=1}^{N} p(y_n \mid \mu) \prod_{d=1}^{D} p(x_{nd} \mid y_n, \phi)$$

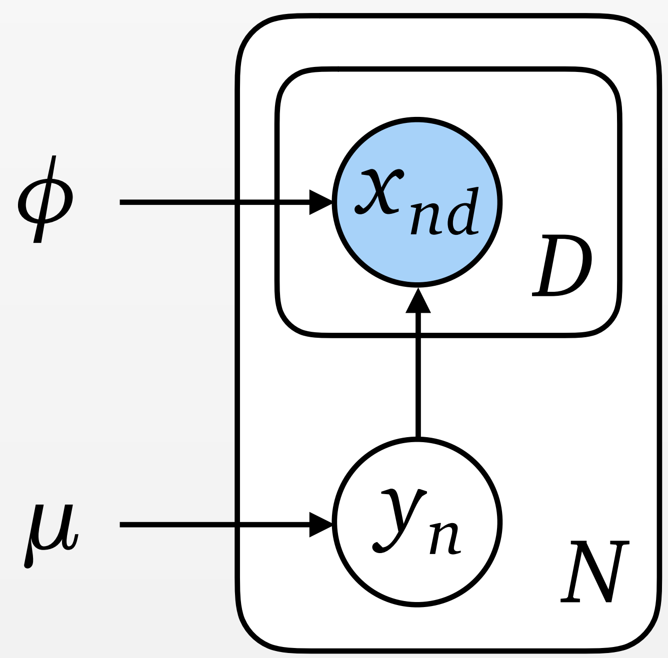
Graphical Model: Naive Bayes

Generative Model

$$y_n \sim \text{Bernoulli}(\mu) \qquad n = 1, \ldots, N$$

$$\boldsymbol{x}_{nd} \mid y_n = k \sim \text{Bernoulli}(\phi_{kd}) \qquad k = 0, 1 \quad d = 1, \ldots, D$$

Graphical Model



Joint Distribution

$$p(x, y \mid \mu, \phi)$$

Goal: Predict Labels

$$p(y \mid x, \mu, \phi) = \frac{p(x \mid y, \phi) p(y \mid \mu)}{p(x \mid \mu, \phi)}$$

Prediction with Bayesian Posterior

Posterior on Label (Spam=1, Not Spam=0)

$$p(y' \mid x', \mu, \phi) = \frac{p(x', y' \mid \mu, \phi)}{p(x' \mid \mu, \phi)}$$

Marginal Likelihood

$$p(x' \mid \mu, \phi) = \sum_{k=\{0,1\}} p(x', y' = k \mid \mu, \phi)$$

Joint

$$p(x', y' \mid \mu, \phi) = p(y' \mid \mu) \prod_{d=1} p(x'_d \mid y', \phi)$$

$$p(y' \mid \mu) = \mu^{y'} (1 - \mu)^{(1-y')}$$

$$p(x'_d \mid y' = k, \phi) = \phi_{kd}^{x'_d} (1 - \phi_{kd})^{1-x'_d}$$

Parameter Estimation

Maximum Likelihood

$$\underset{\mu,\phi}{\operatorname{argmax}} \sum_{n=1}^{N} \log p(y_n, x_n \mid \mu, \phi)$$

Labeled Data

$$\{(x_1, y_1), \ldots, (x_N, y_N)\}$$

Maximum A Posteriori

$$\underset{\mu,\phi}{\operatorname{argmax}} \; \log p(\mu, \phi) + \sum_{n=1}^{N} \log p(y_n, x_n \mid \mu, \phi)$$

Test-time Prediction

$$p(y' \mid x', \mu^*, \phi^*)$$

Maximum Likelihood Estimation

Objective

$$\underset{\mu,\phi}{\operatorname{argmax}} \sum_{n=1}^{N} \log p(y_n, x_n \mid \mu, \phi)$$

Training Data (Labeled)

$$\{(x_1, y_1), \ldots, (x_N, y_N)\}$$

Optimum Parameters

$$\mu = \frac{1}{N} \sum_{n=1}^{N} I[y_n = 1]$$

Interpretation: Fraction of Spam in training set

$$\phi_{kd} = \frac{1}{N_k} \sum_{n: y_n = k} I[x_{nd} = 1]$$

Interpretation: Fraction of *non-spam* **(k=0)** and *spam* **(k=1)** messages that contain term.

Maximum Likelihood Estimation

Objective

$$\operatorname*{argmax}_{\mu,\phi} \sum_{n=1}^{N} \log p(y_n, x_n \mid \mu, \phi)$$

Training Data (Labeled)

$$\{(x_1, y_1), \ldots, (x_N, y_N)\}$$

Optimum Parameters

$$\mu = \frac{N_1^y}{N}$$

$$\phi_{kd} = \frac{N_{kd}^x}{N_k^y}$$

Problem:

What do you do for words not found in training set?

$$\sum_n I[x_{nd} = 1] = 0$$

$$\phi_{0d} = \phi_{1d} = 0$$

Maximum Likelihood Estimation

Objective

$$\underset{\mu,\phi}{\operatorname{argmax}} \sum_{n=1}^{N} \log p(y_n, x_n \mid \mu, \phi)$$

Training Data (Labeled)

$$\{(x_1, y_1), \ldots, (x_N, y_N)\}$$

Optimum Parameters

(with Laplace smoothing)

$$\mu = \frac{N_1^y + 1}{N + 2}$$

$$\phi_{kd} = \frac{N_{kd}^x + 1}{N_k^y + D}$$

Problem:

What do you do for words not found in training set?

$$\sum_n I[x_{nd} = 1] = 0$$

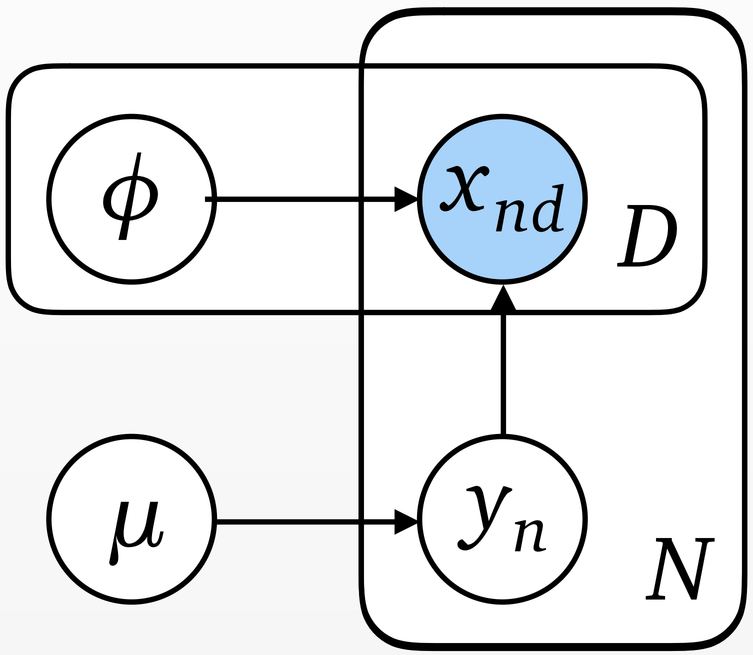
$$\phi_{0d} = \phi_{1d} = 0$$

Naive Bayes with Priors

Generative Model

$$\mu \sim \mathrm{Beta}(1,1)$$
$$\phi_{kd} \sim \mathrm{Beta}(1,1)$$
$$y_n \sim \mathrm{Bernoulli}(\mu)$$
$$x_{nd} \mid y_n = k \sim \mathrm{Bernoulli}(\phi_{kd})$$

Graphical Model



Maximum A Posteriori

$$y' = \underset{y'}{\mathrm{argmax}}\; p(y' \mid x', \mu^*, \phi^*)$$

$$\mu^*, \phi^* = \underset{\mu,\phi}{\mathrm{argmax}}\; \log p(\mu, \phi) + \sum_{n=1}^{N} \log p(x_n, y_n \mid \mu, \phi)$$

Exponential Families

Exponential Family Distributions

Base Measure

Log Normalizer

$$p(x \mid \eta) = h(x) \exp\left[\sum_i \eta_i t_i(x) - a(\eta)\right]$$

Natural Parameters

Sufficient Statistics

Product of terms:

x-dependent, xη-dependent, η-dependent

Example: Discrete Distribution

$$p(x \mid \eta) = h(x) \exp\left[\sum_i \eta_i t_i(x) - a(\eta)\right]$$

$$p(x \mid \theta) = \prod_{k=1}^{K} \theta_k^{x_k} = \exp\left[\sum_{k=1}^{K} \log(\theta_k) x_k\right]$$

$$h(x) = 1$$

$$\eta_k = \log \theta_k$$

$$t_k(x) = x_k$$

$$a(\eta) = 0$$

Example: Gaussian Distribution

$$p(x \mid \eta) = h(x) \exp\left[\sum_i \eta_i t_i(x) - a(\eta)\right]$$

$$h(x) = \frac{1}{\sqrt{2\pi}}$$

$$\eta = \left(\frac{\mu}{\sigma^2}, -\frac{1}{\sigma^2}\right)$$

$$p(x \mid \mu, \sigma) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{1}{2}\frac{(x-\mu)^2}{\sigma^2}\right]$$

$$t(x) = (x, x^2)$$

$$a(\eta) = \frac{\mu^2}{\sigma^2} + \log\sigma$$

Exponential Families

https://en.wikipedia.org/wiki/Exponential_family

Distribution	Parameter(s) $\boldsymbol\theta$	Natural parameter(s) $\boldsymbol\eta$	Inverse parameter mapping	Base measure $h(x)$	Sufficient statistic $T(x)$	Log-partition $A(\boldsymbol\eta)$	Log-partition $A(\boldsymbol\theta)$
Bernoulli distribution	p	$\ln \frac{p}{1-p}$ • This is the logit function.	$\frac{1}{1+e^{-\eta}} = \frac{e^\eta}{1+e^\eta}$ • This is the logistic function.	1	x	$\ln(1+e^\eta)$	$-\ln(1-p)$
binomial distribution with known number of trials *n*	p	$\ln \frac{p}{1-p}$	$\frac{1}{1+e^{-\eta}} = \frac{e^\eta}{1+e^\eta}$	$\binom{n}{x}$	x	$n\ln(1+e^\eta)$	$-n\ln(1-p)$
Poisson distribution	λ	$\ln\lambda$	e^η	$\frac{1}{x!}$	x	e^η	λ
negative binomial distribution with known number of failures *r*	p	$\ln p$	e^η	$\binom{x+r-1}{x}$	x	$-r\ln(1-e^\eta)$	$-r\ln(1-p)$
exponential distribution	λ	$-\lambda$	$-\eta$	1	x	$-\ln(-\eta)$	$-\ln\lambda$
Pareto distribution with known minimum value x_m	α	$-\alpha-1$	$-1-\eta$	1	$\ln x$	$-\ln(-1-\eta)+(1+\eta)\ln x_{\mathrm{m}}$	$-\ln\alpha-\alpha\ln x_{\mathrm{m}}$

Most of the “normally” used distributions:
Normal, Gamma, Dirichlet, Discrete, Poisson, Cauchy, etc.