Introduction to Machine Learning

Brown University CSCI 1950-F, Spring 2012
Prof. Erik Sudderth

Lecture 4:
Probability: Continuous Random Variables
Naïve Bayes: ML & Bayesian Estimation
Preview of Decision Theory

Many figures courtesy Kevin Murphy's textbook,
Machine Learning: A Probabilistic Perspective

Bayes Rule (Bayes Theorem)

θ ⟶ unknown parameters (many possible models)

$\mathcal{D}$ ⟶ observed data available for learning

$p(\theta)$ ⟶ prior distribution (domain knowledge)

$p(\mathcal{D} \mid \theta)$ ⟶ likelihood function (measurement model)

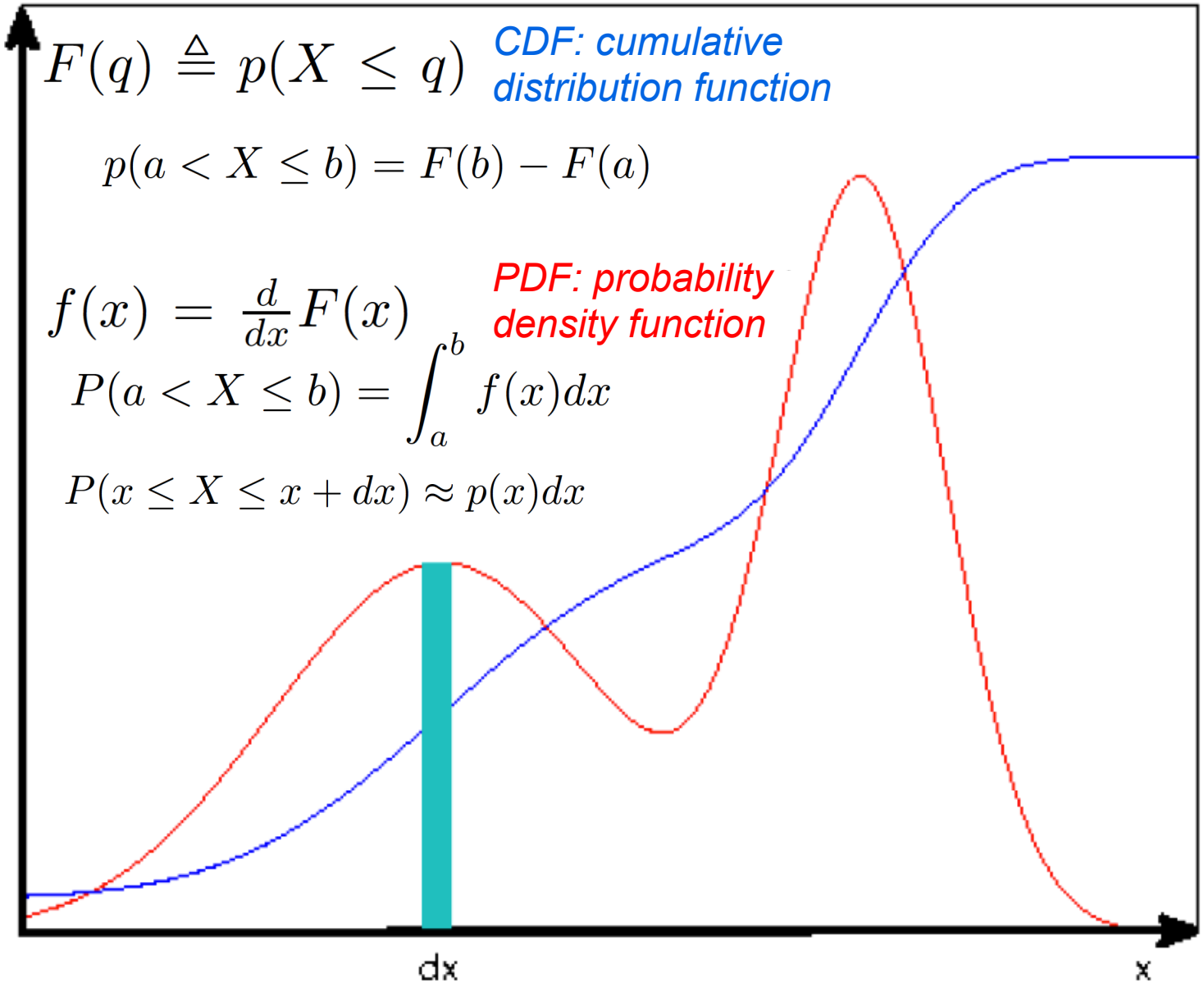
$p(\theta \mid \mathcal{D})$ ⟶ posterior distribution (learned information)

$$p(\theta, \mathcal{D}) = p(\theta)p(\mathcal{D} \mid \theta) = p(\mathcal{D})p(\theta \mid \mathcal{D})$$

$$p(\theta \mid \mathcal{D}) = \frac{p(\theta, \mathcal{D})}{p(\mathcal{D})} = \frac{p(\mathcal{D} \mid \theta)p(\theta)}{\sum_{\theta' \in \Theta} p(\mathcal{D} \mid \theta')p(\theta')}$$

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

Continuous Random Variables



Moments of Random Variables

Mean

$$\mathbb{E}[X] \triangleq \int_{\mathcal{X}} x\ p(x)dx \qquad \mathbb{E}[X] \triangleq \sum_{x\in\mathcal{X}} x\ p(x)$$

Variance

$$\begin{aligned}\mathrm{var}[X] &\triangleq \mathbb{E}\left[(X-\mu)^2\right] = \int (x-\mu)^2 p(x)dx \\ &= \int x^2 p(x)dx + \mu^2 \int p(x)dx - 2\mu \int xp(x)dx = \mathbb{E}\left[X^2\right] - \mu^2\end{aligned}$$

$$\mathbb{E}\left[X^2\right] = \mu^2 + \sigma^2 \qquad \textit{second moment}$$

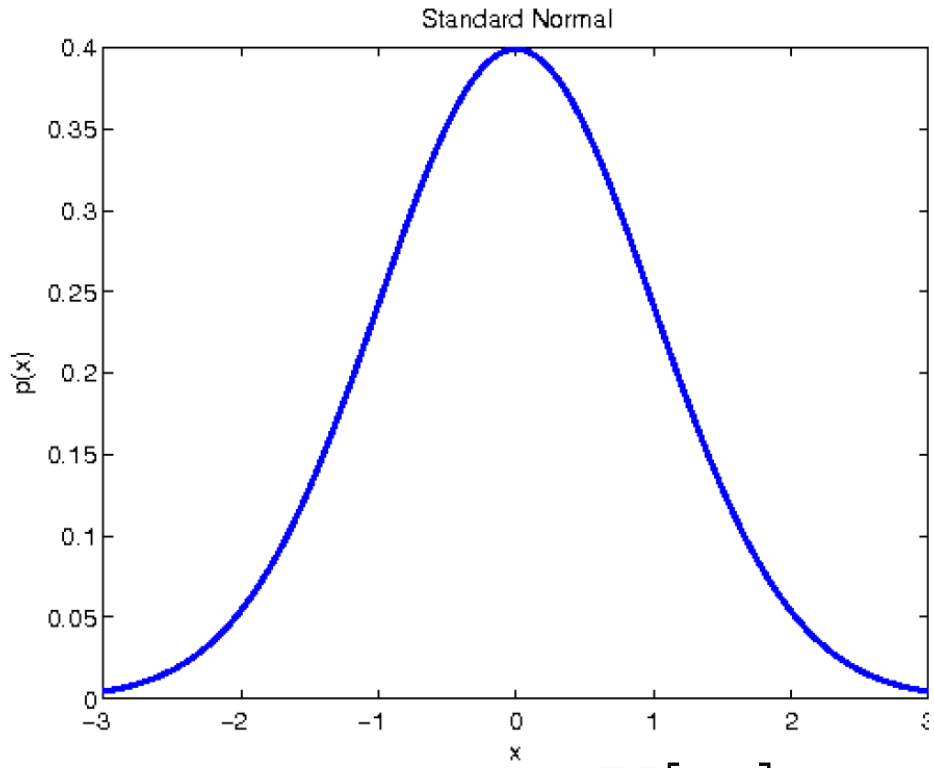
$$\mathrm{std}[X] \triangleq \sqrt{\mathrm{var}[X]} \qquad \textit{standard deviation}$$

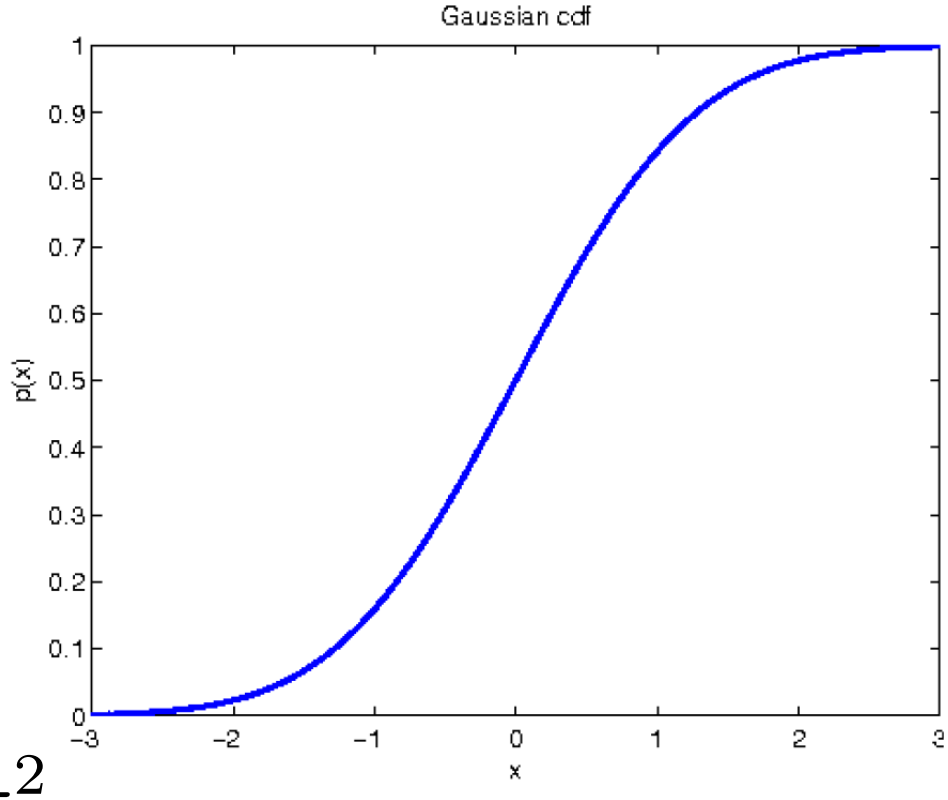
Moments & Conditional Moments

$$\mathbb{E}[g(X)] = \int_{\mathcal{X}} g(x)p(x)\ dx \qquad \mathbb{E}[g(X) \mid Y = y] = \int_{\mathcal{X}} g(x)p(x \mid y)\ dx$$

Gaussian (Normal) Distributions

$$\mathcal{N}(x \mid \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\} \quad \Phi(x \mid \mu, \sigma^2) = \int_{-\infty}^{x} \mathcal{N}(z \mid \mu, \sigma^2)\, dz$$

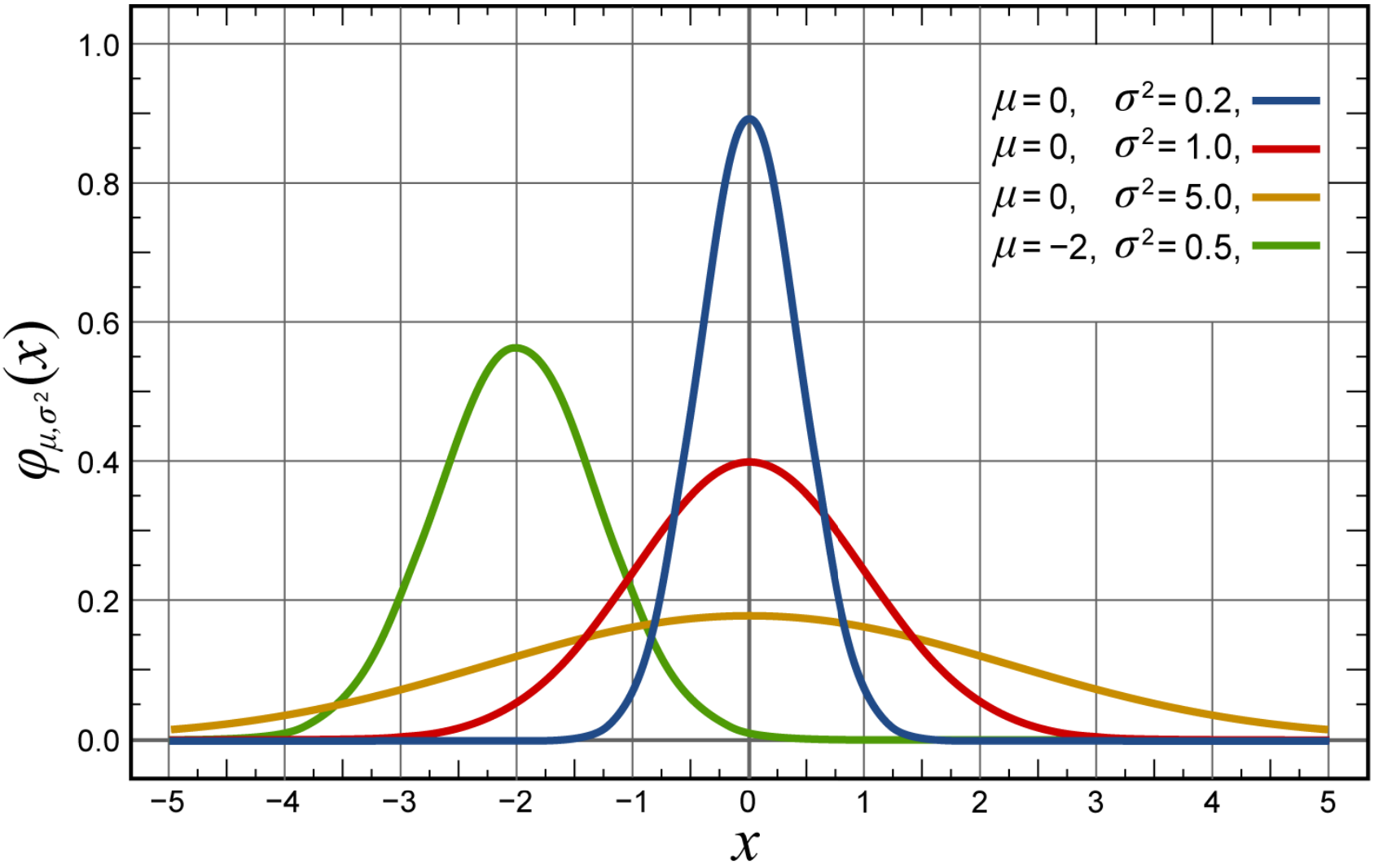




$$\mathbb{E}[X] = \mu \qquad \mathbb{V}[X] = \sigma^2$$

$$\text{Mode}[X] = \arg\max_{x \in \mathbb{R}} \mathcal{N}(x \mid \mu, \sigma^2) = \mu$$

Gaussian (Normal) Distributions



Summaries: Mean, median, mode, variance, standard deviation

Learning Binary Probabilities

Bernoulli Distribution: Single toss of a (possibly biased) coin

$$\mathrm{Ber}(x \mid \theta) = \theta^{\mathbb{I}(x=1)}(1-\theta)^{\mathbb{I}(x=0)} \qquad 0 \leq \theta \leq 1$$

- Suppose we observe N samples from a Bernoulli distribution with unknown mean:

$$X_i \sim \mathrm{Ber}(\theta), i = 1, \ldots, N$$

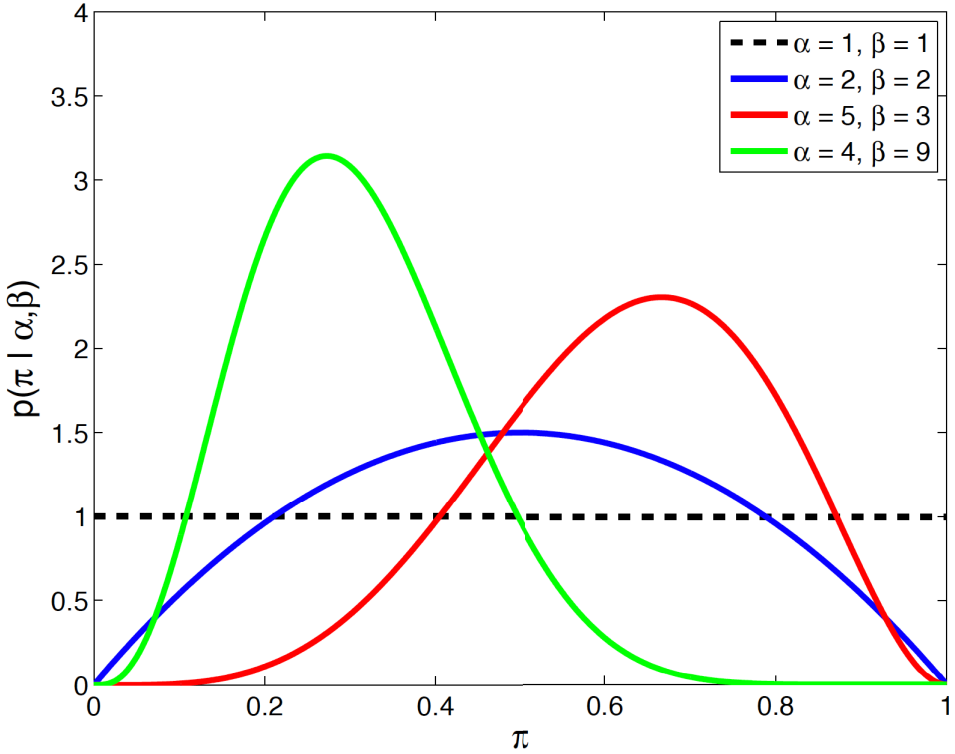
$$p(x_1, \ldots, x_N \mid \theta) = \theta^{N_1}(1-\theta)^{N_0}$$

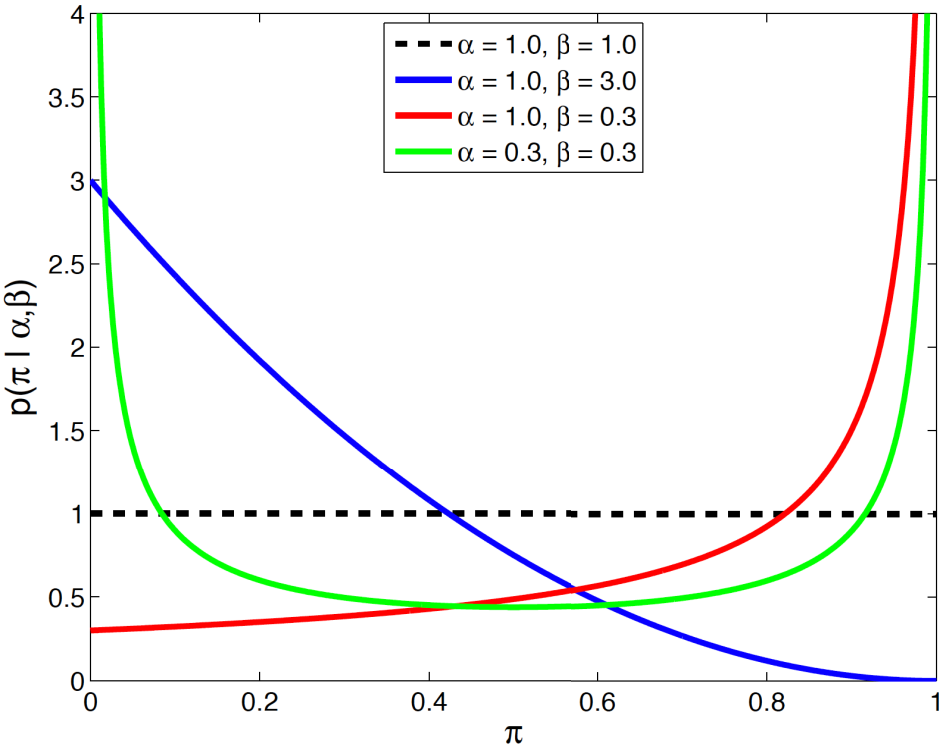
$$N_1 = \sum_{i=1}^{N} \mathbb{I}(x_i = 1) \qquad N_0 = \sum_{i=1}^{N} \mathbb{I}(x_i = 0)$$

- What is the *maximum likelihood* parameter estimate?

$$\hat{\theta} = \arg\max_{\theta} \log p(x \mid \theta) = \frac{N_1}{N}$$

Beta Distributions

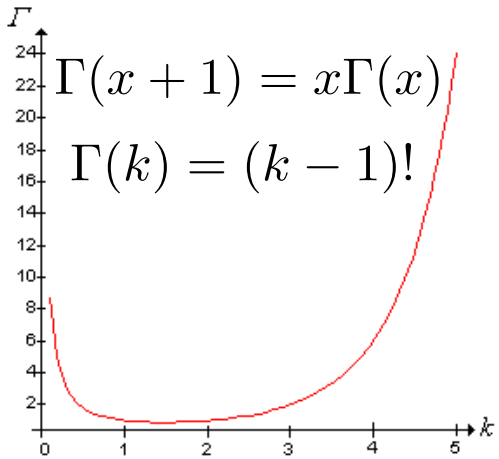




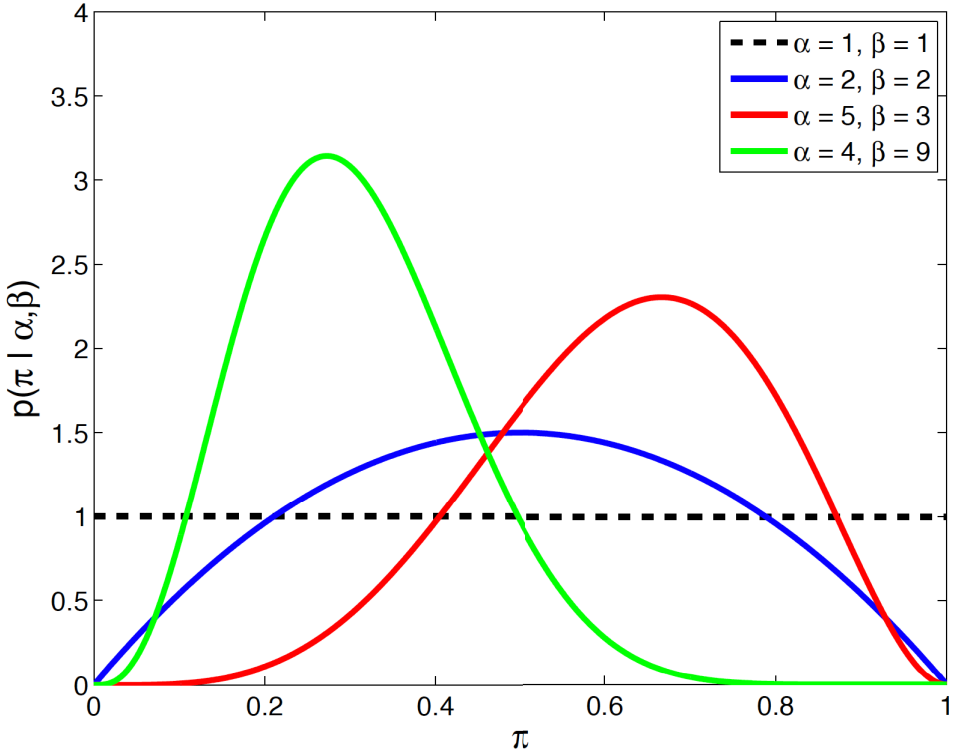
Probability density function: $x \in [0, 1]$

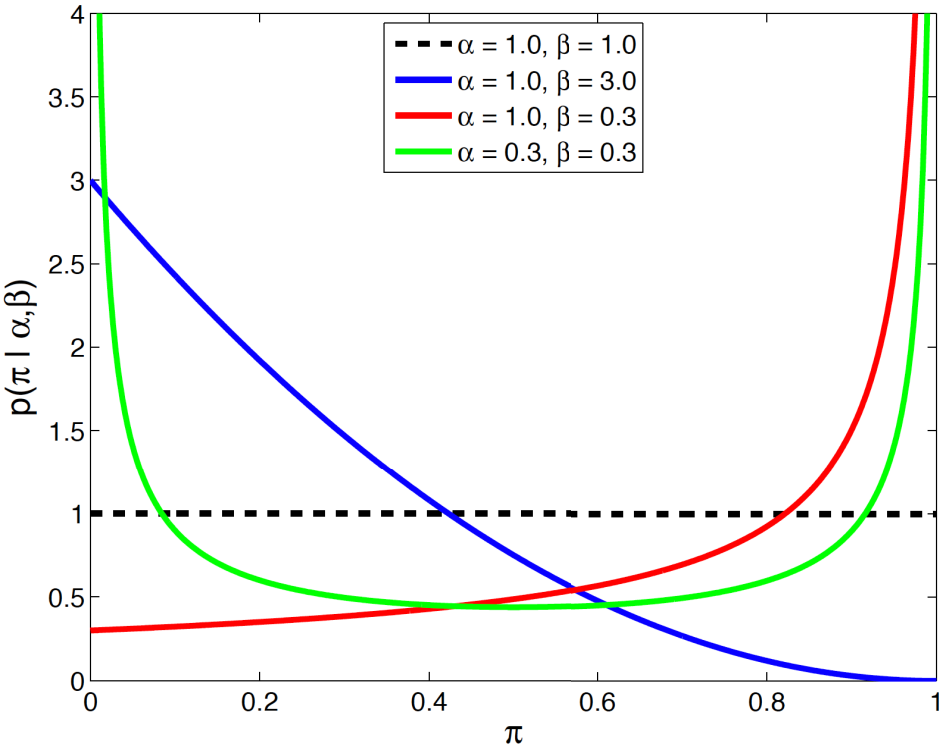
$$\text{Beta}(x|a, b) = \frac{1}{B(a, b)} x^{a-1}(1-x)^{b-1}$$

$$B(a, b) \triangleq \frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)} \qquad a, b > 0$$



Beta Distributions





$$\mathbb{E}[x] = \frac{a}{a+b} \qquad \mathbb{V}[x] = \frac{ab}{(a+b)^2(a+b+1)}$$

$$\text{Mode}[x] = \arg\max_{x \in [0,1]} \text{Beta}(x \mid a, b) = \frac{a-1}{(a-1)+(b-1)}$$

Bayesian Learning of Probabilities

Bernoulli Likelihood: Single toss of a (possibly biased) coin

$$\text{Ber}(x \mid \theta) = \theta^{\mathbb{I}(x=1)}(1-\theta)^{\mathbb{I}(x=0)} \quad 0 \leq \theta \leq 1$$

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

Beta Prior Distribution:

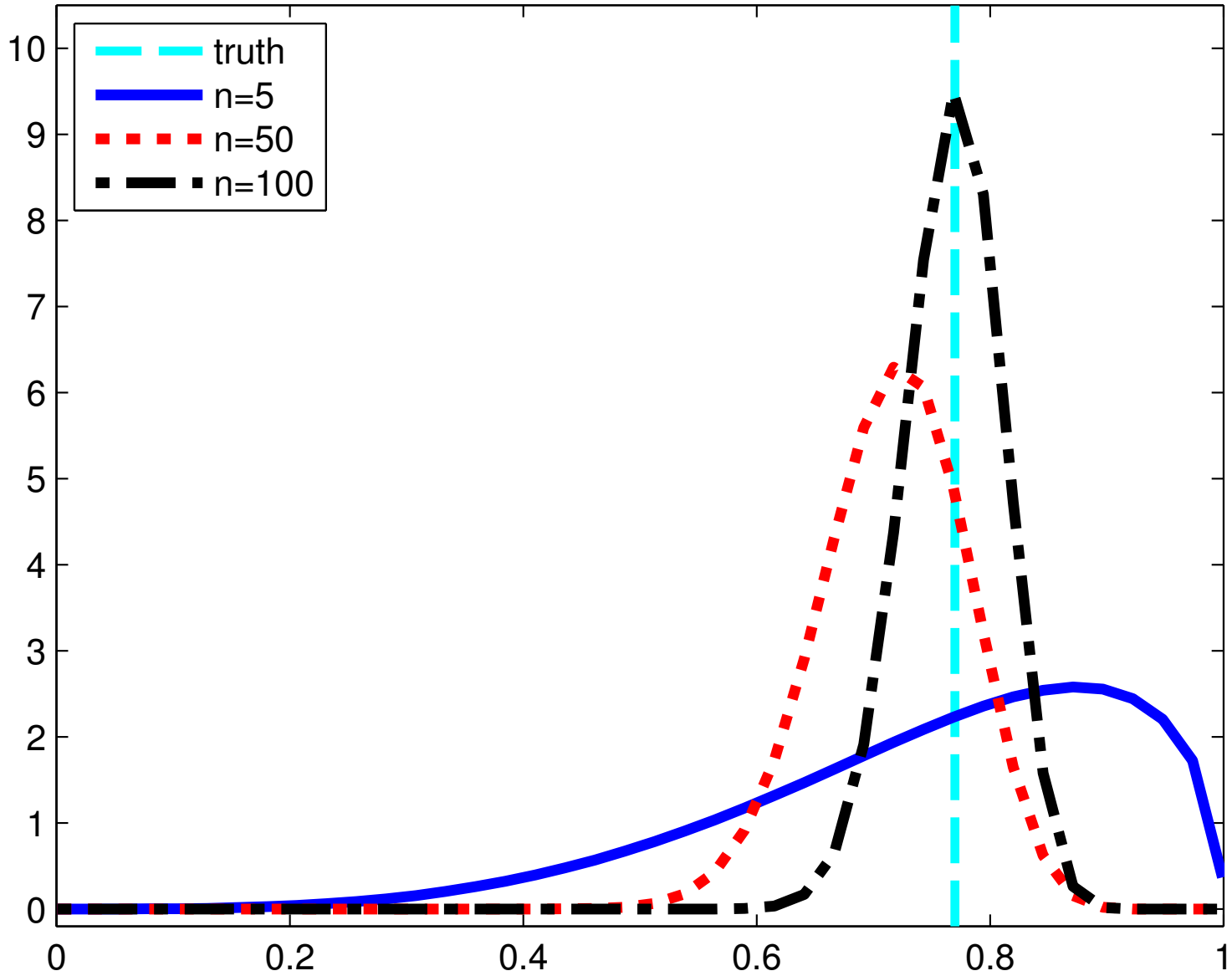
$$p(\theta) = \text{Beta}(\theta \mid a, b) \propto \theta^{a-1}(1-\theta)^{b-1}$$

Posterior Distribution:

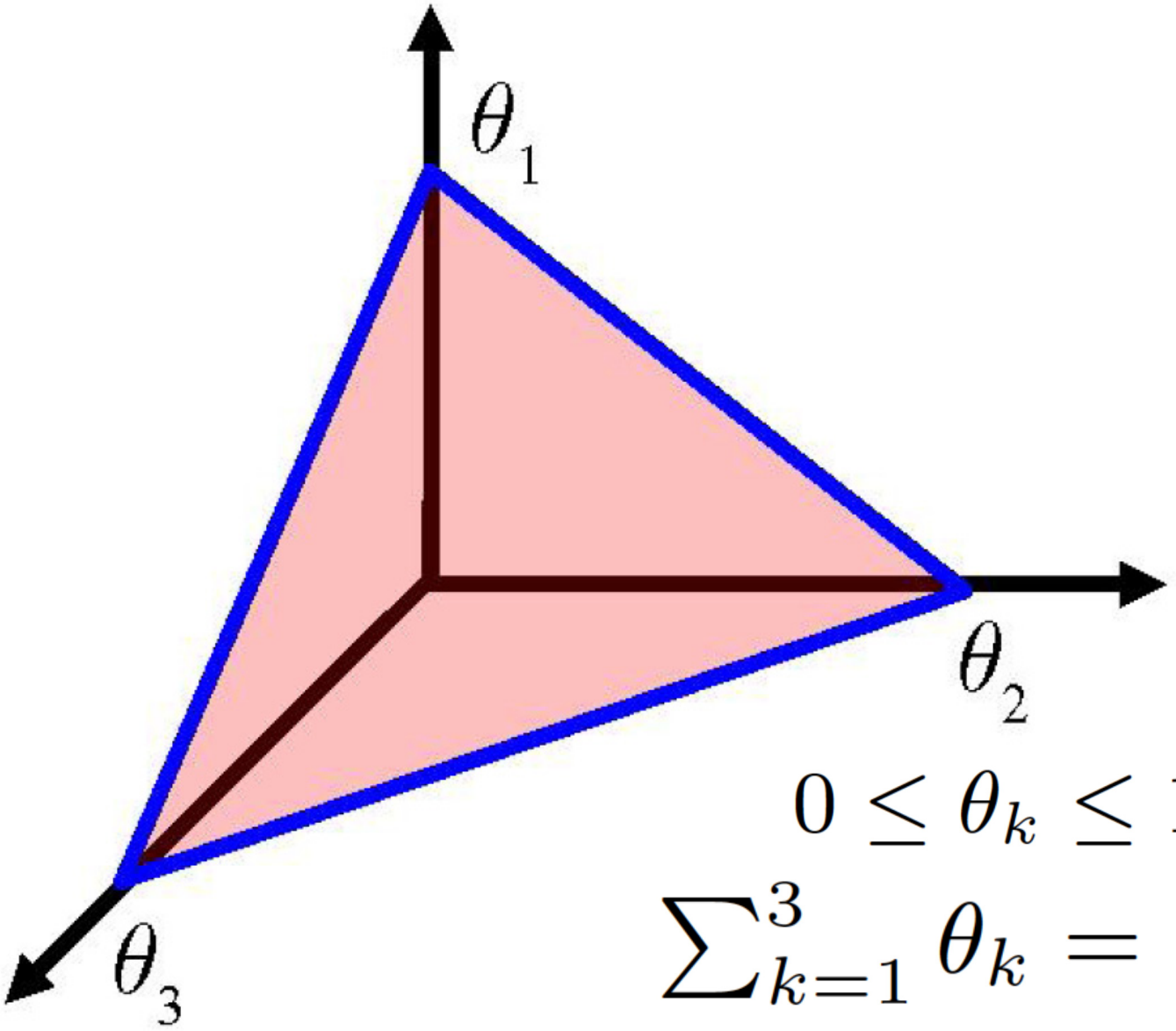
$$p(\theta \mid x) \propto \theta^{N_1+a-1}(1-\theta)^{N_0+b-1} \propto \text{Beta}(\theta \mid N_1 + a, N_0 + b)$$

- *This is a conjugate prior, because posterior is in same family*
- *Estimate by posterior mode (MAP) or mean (preferred)*
- *Here, posterior predictive equivalent to mean estimate*

Sequence of Beta Posteriors



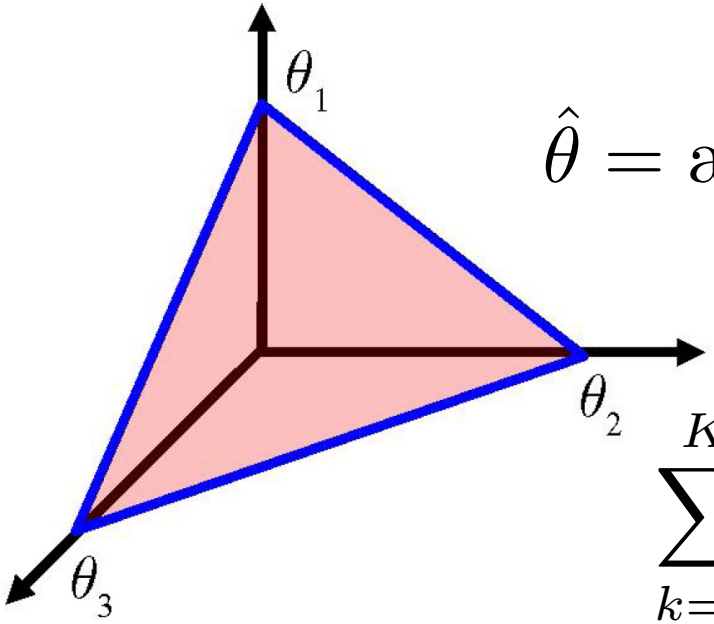
Multinomial Simplex



$$0 \leq \theta_k \leq 1$$

$$\sum_{k=1}^{3} \theta_k = 1$$

Constrained Optimization



$$\hat{\theta} = \arg\max_{\theta} \sum_{k=1}^{K} a_k \log \theta_k \qquad a_k \geq 0$$

subject to

$$\sum_{k=1}^{K} \theta_k = 1 \qquad \theta_k \geq 0$$

- Solution: $\hat{\theta}_k = \dfrac{a_k}{a_0} \qquad a_0 = \sum_{k=1}^{K} a_k$
- Proof for K=2: Change of variables to unconstrained problem
- Proof for general K: Lagrange multipliers (see textbook)

Learning Categorical Probabilities

Multinoulli Distribution: Single roll of a (possibly biased) die

$$\mathrm{Cat}(x \mid \theta) = \prod_{k=1}^{K} \theta_k^{x_k} \qquad \mathcal{X} = \{0,1\}^K, \sum_{k=1}^{K} x_k = 1$$

- If we have N_k observations of outcome k in N trials:

$$p(x_1, \ldots, x_N \mid \theta) = \prod_{k=1}^{K} \theta_k^{N_k}$$

- The *maximum likelihood* parameter estimates are then:

$$\hat{\theta} = \arg\max_{\theta} \log p(x \mid \theta) \qquad \hat{\theta}_k = \frac{N_k}{N}$$

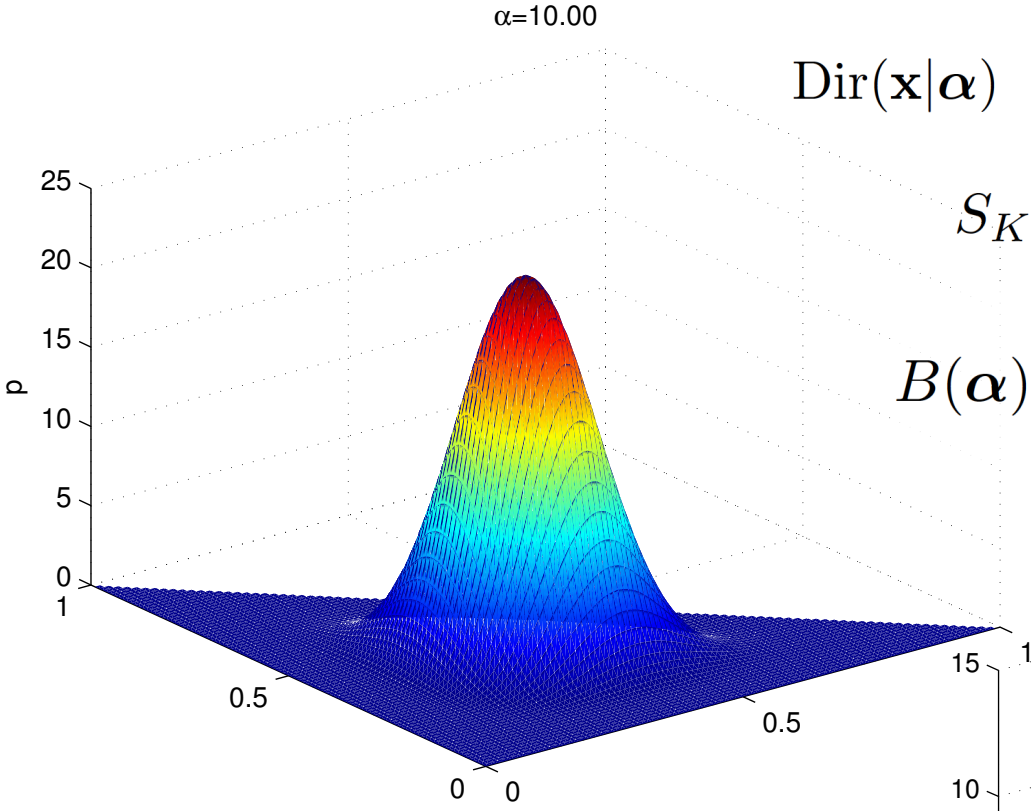
- Will this produce sensible predictions when K is large?

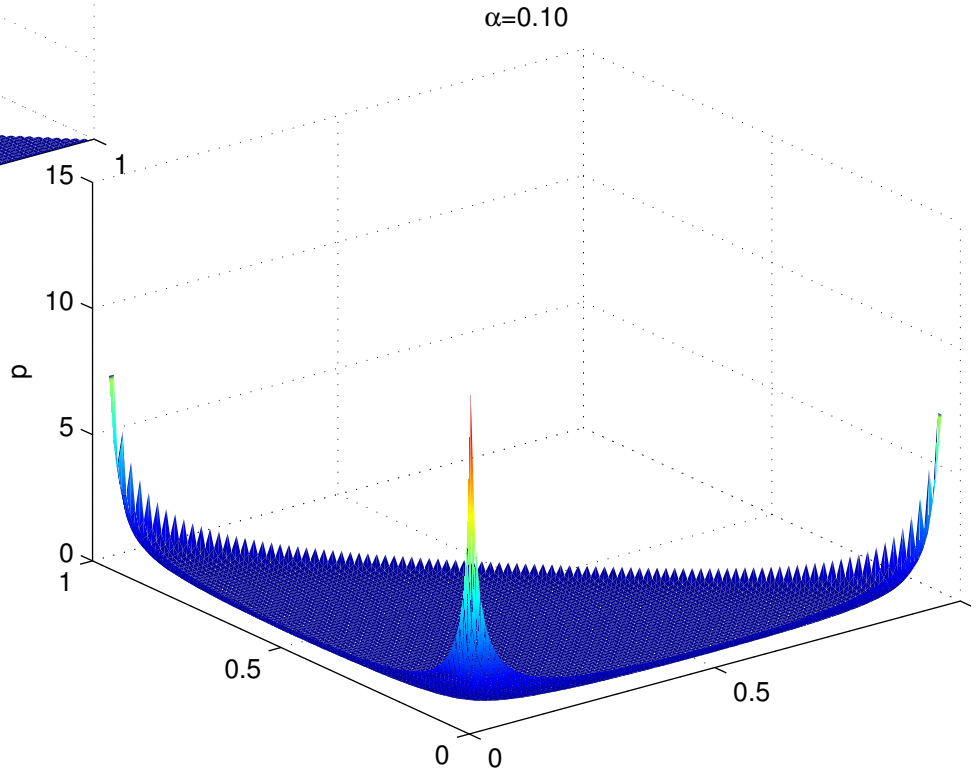
Dirichlet Probability Densities

$$\mathrm{Dir}(\mathbf{x}|\boldsymbol{\alpha}) \triangleq \frac{1}{B(\boldsymbol{\alpha})} \prod_{k=1}^{K} x_k^{\alpha_k - 1} \mathbb{I}(\mathbf{x} \in S_K)$$

$$S_K = \{\mathbf{x} : 0 \leq x_k \leq 1, \sum_{k=1} x_k = 1\}$$

$$B(\boldsymbol{\alpha}) \triangleq \frac{\prod_{k=1}^{K} \Gamma(\alpha_k)}{\Gamma(\alpha_0)} \qquad \alpha_0 \triangleq \sum_{k=1}^{K} \alpha_k$$

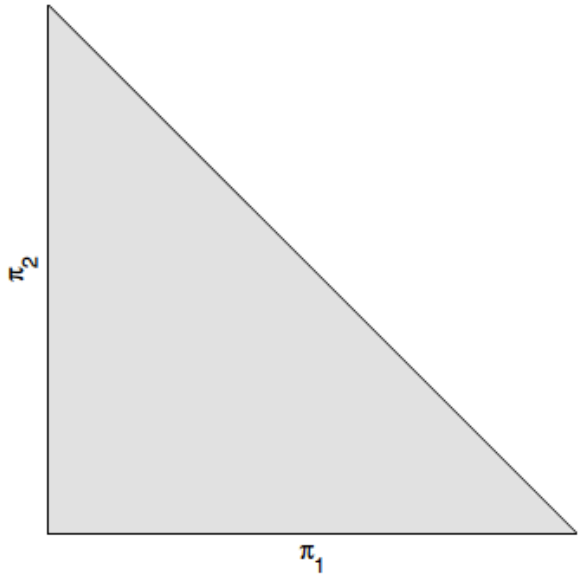




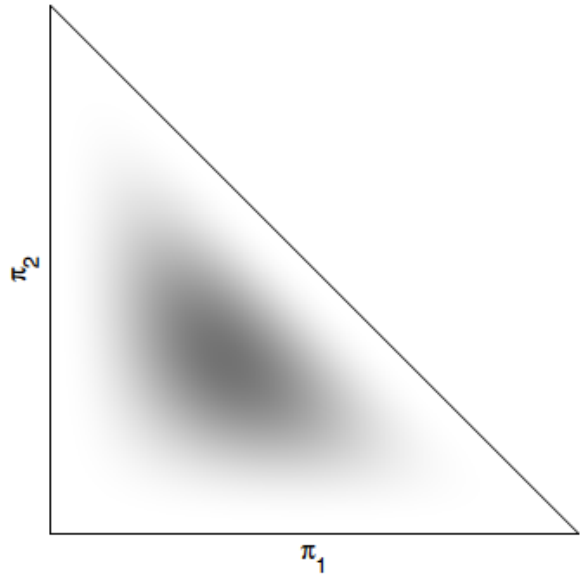
Mean: $\mathbb{E}[x_k] = \frac{\alpha_k}{\alpha_0}$

Mode: $\hat{x}_k = \frac{\alpha_k - 1}{\alpha_0 - K}$

Dirichlet Probability Densities

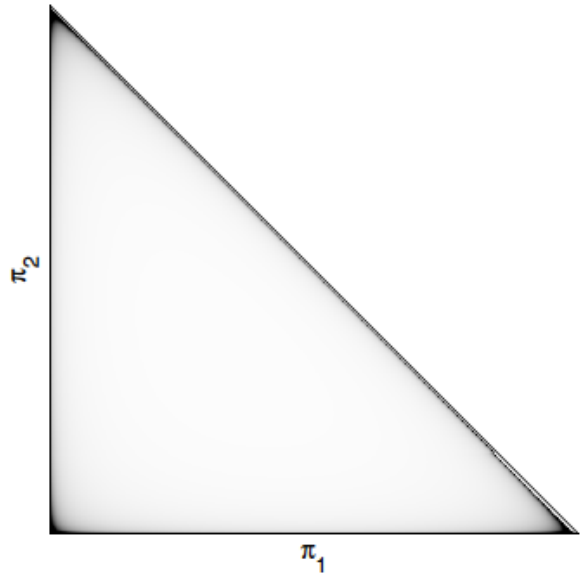


$\pi \sim \mathrm{Dir}(1, 1, 1)$



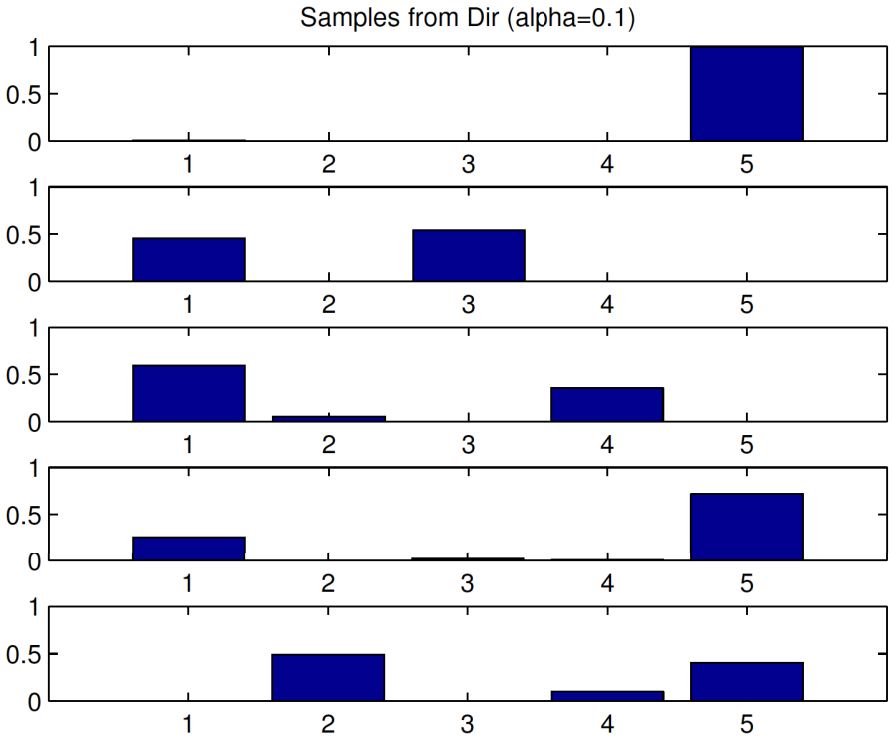
$\pi \sim \mathrm{Dir}(4, 4, 4)$

$\pi \sim \mathrm{Dir}(4, 9, 7)$

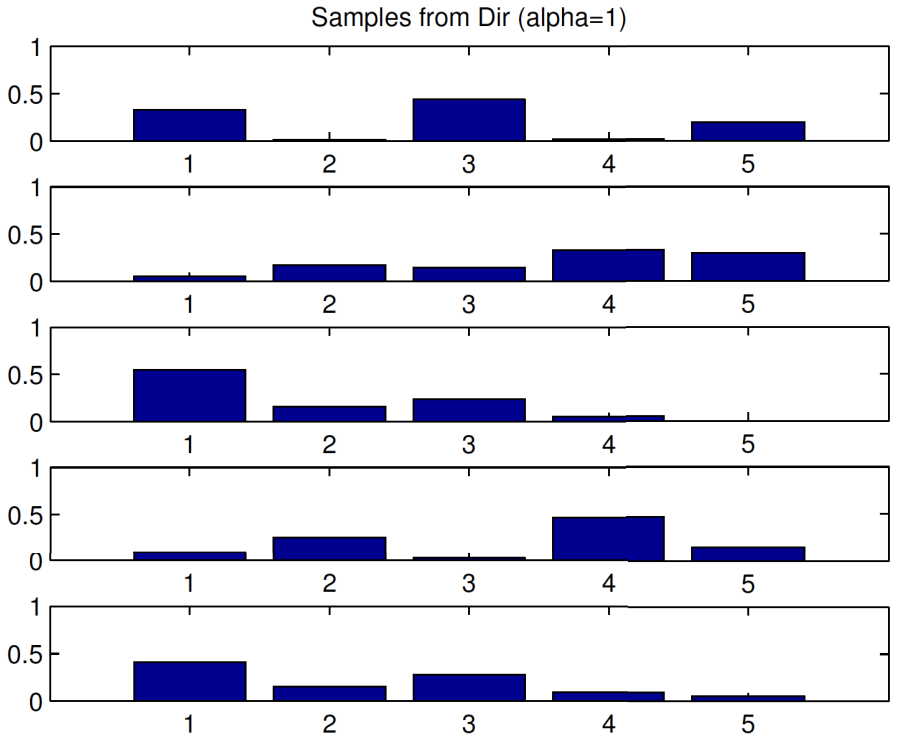


$\pi \sim \mathrm{Dir}(0.2, 0.2, 0.2)$

Dirichlet Samples



$$\mathrm{Dir}(\theta \mid 0.1, 0.1, 0.1, 0.1, 0.1)$$



$$\mathrm{Dir}(\theta \mid 1.0, 1.0, 1.0, 1.0, 1.0)$$

Bayesian Learning of Probabilities

Multinoulli Distribution: Single roll of a (possibly biased) die

$$\mathrm{Cat}(x \mid \theta) = \prod_{k=1}^{K} \theta_k^{x_k} \qquad \mathcal{X} = \{0,1\}^K, \sum_{k=1}^{K} x_k = 1$$

$$p(x_1, \ldots, x_N \mid \theta) = \prod_{k=1}^{K} \theta_k^{N_k}$$

Dirichlet Prior Distribution:

$$p(\theta) = \mathrm{Dir}(\theta \mid \alpha) \propto \prod_{k=1}^{K} \theta_k^{\alpha_k - 1}$$

Posterior Distribution:

$$p(\theta \mid x) \propto \prod_{k=1}^{K} \theta_k^{N_k + \alpha_k - 1} \propto \mathrm{Dir}(\theta \mid N_1 + \alpha_1, \ldots, N_K + \alpha_K)$$

- *This is a conjugate prior, because posterior is in same family*

Learning a Naïve Bayes Model

$$p(\theta \mid y, x) \propto p(\theta, y, x) = p(\theta)p(y \mid \theta)p(x \mid y, \theta)$$

model *class* *features*

- Assume we have N training examples independently sampled from an unknown naïve Bayes model:

$y_i \longrightarrow$ observed class label for training example *i*

$x_{ij} \longrightarrow$ value of feature *j* for training example *i*

$$p(\theta \mid y, x) \propto p(\theta) \prod_{i=1}^{N} p(y_i \mid \theta)p(x_i \mid y_i, \theta)$$

$$\propto p(\theta) \prod_{i=1}^{N} p(y_i \mid \theta) \prod_{j=1}^{D} p(x_{ij} \mid y_i, \theta)$$

- Learning: ML estimate, MAP estimate, or posterior prediction

Naïve Bayes: ML & Bayes

$$p(\mathbf{x}_i, y_i|\boldsymbol{\theta}) = p(y_i|\boldsymbol{\pi}) \prod_j p(x_{ij}|\boldsymbol{\theta}_j) = \prod_c \pi_c^{\mathbb{I}(y_i=c)} \prod_j \prod_c p(x_{ij}|\boldsymbol{\theta}_{jc})^{\mathbb{I}(y_i=c)}$$

$$\log p(\mathcal{D}|\boldsymbol{\theta}) = \sum_{c=1}^{C} N_c \log \pi_c + \sum_{j=1}^{D} \sum_{c=1}^{C} \sum_{i:y_i=c} \log p(x_{ij}|\boldsymbol{\theta}_{jc})$$

N_c ⟶ number of examples of training class *c*

- Maximizing the sum of functions of independent parameters can be done by maximizing them independently:

Maximum Likelihood

$$\hat{\pi}_c = \frac{N_c}{N} \qquad \hat{\theta}_{jc} = \frac{N_{jc}}{N_c} \quad \textit{if} \quad x_j|y = c \sim \mathrm{Ber}(\theta_{jc})$$

- Similarly, if the parameters for different features are independent under the prior, they remain independent under the posterior, and Bayesian analysis decomposes

Generative Classifiers

$y \longrightarrow$ class label in *{1,…,C}*, observed in training

$x \in \mathcal{X} \longrightarrow$ observed features to be used for classification

$\theta \longrightarrow$ parameters indexing family of models

$$p(y, x \mid \theta) = p(y \mid \theta) p(x \mid y, \theta)$$

prior distribution *likelihood function*

- Compute class *posterior distribution* via Bayes rule:

$$p(y = c \mid x, \theta) = \frac{p(y = c \mid \theta) p(x \mid y = c, \theta)}{\sum_{c'=1}^{C} p(y = c' \mid \theta) p(x \mid y = c', \theta)}$$

- *Inference:* Find label distribution for some input example
- *Classification:* Make decision based on inferred distribution
- *Learning:* Estimate parameters θ from labeled training data

Decision Theory

$y \in \mathcal{Y}$ ⟶ unknown hidden state of "nature"

$x \in \mathcal{X}$ ⟶ observed data

$a \in \mathcal{A}$ ⟶ set of possible actions we can take

$L(y, a)$ ⟶ real-valued loss function: the price we pay if we choose action *a*, and *y* is the true hidden state

- Goal: Choose the action which minimizes the expected loss

$$\delta(\mathbf{x}) = \underset{a \in \mathcal{A}}{\operatorname{argmin}} \, \mathbb{E}\left[L(y, a)\right] \qquad \delta : \mathcal{X} \to \mathcal{A}$$

  - Some averaging is necessary because we don't know *y*
  - Two notions of expectation: Bayesian versus frequentist
- Some communities speak of maximizing expected utility, which is equivalent if utility equals negative loss

Losses for Classification

$y \in \mathcal{Y}$ ⟶ unknown class or category, finite set of options

$x \in \mathcal{X}$ ⟶ observed data, can take values in any space

$\mathcal{A} = \mathcal{Y}$ ⟶ action is to choose one of the categories

$L(y, a)$ ⟶ table giving loss for all possible mistakes

- Most common default choice is the 0-1 loss:

$$L(y, a) = \mathbb{I}(y \neq a) = \begin{cases} 0 & \text{if } a = y \\ 1 & \text{if } a \neq y \end{cases}$$

- For the special case of binary classification:

predicted label $\hat{y}$	true label y	
	0	1
0	0	λ_{01}
1	λ_{10}	0