Applied Bayesian Nonparametrics

Special Topics in Machine Learning
Brown University CSCI 2950-P, Fall 2011

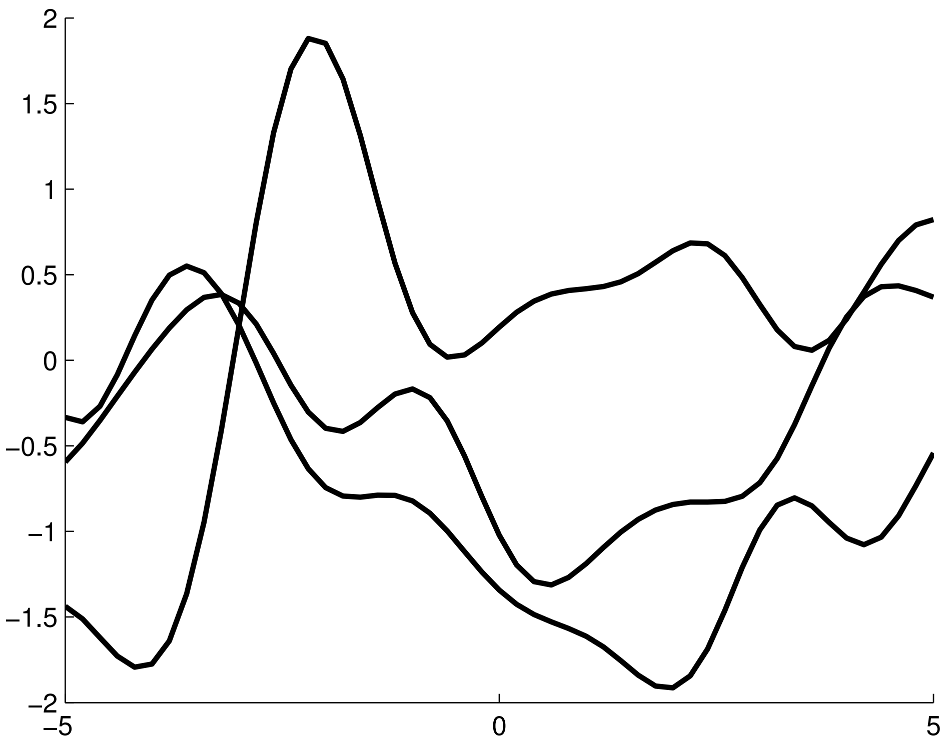
September 20: Gaussian Process Review, Dirichlet Processes and DP Mixture Models

Gaussian Process Kernels & Features

$$p(y) = \mathcal{N}(y \mid 0, \alpha^{-1}\Phi\Phi^T)$$
$$= \mathcal{N}(y \mid 0, K)$$
$$K_{ij} = k(x_i, x_j) = \phi(x_i)^T \phi(x_j)$$

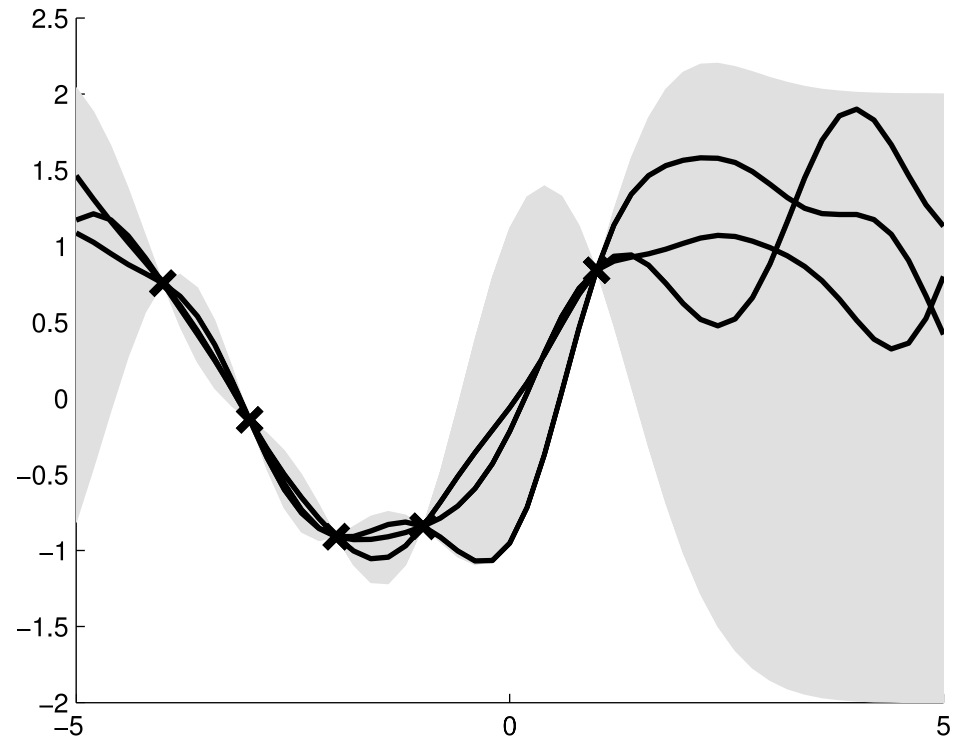
- Features and kernels are dual views of the same models
- Kernel representation useful when the number of features is very large, or even infinite
- Feature representation useful when the amount of data very large, and a moderate number of important features can be identified

1D Gaussian Process Regression



Samples from Prior

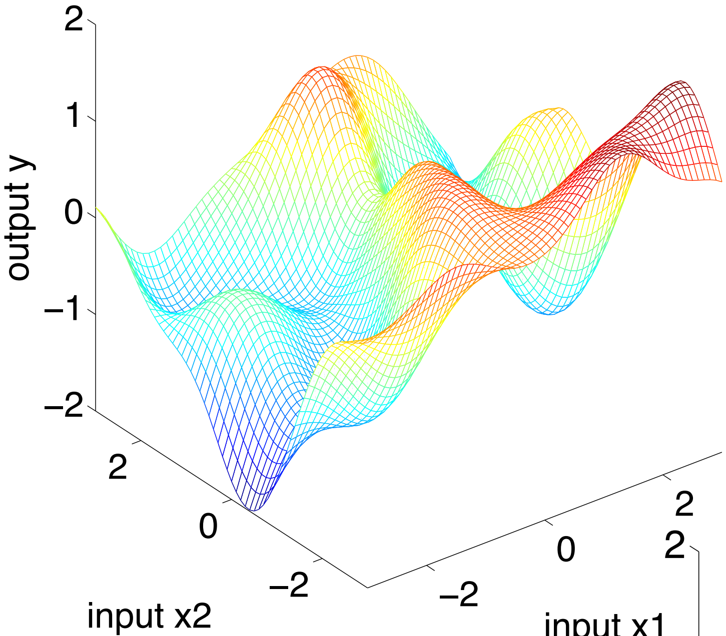
$$\kappa(x, x') = \sigma_f^2 \exp(-\frac{1}{2\ell^2}(x - x')^2)$$

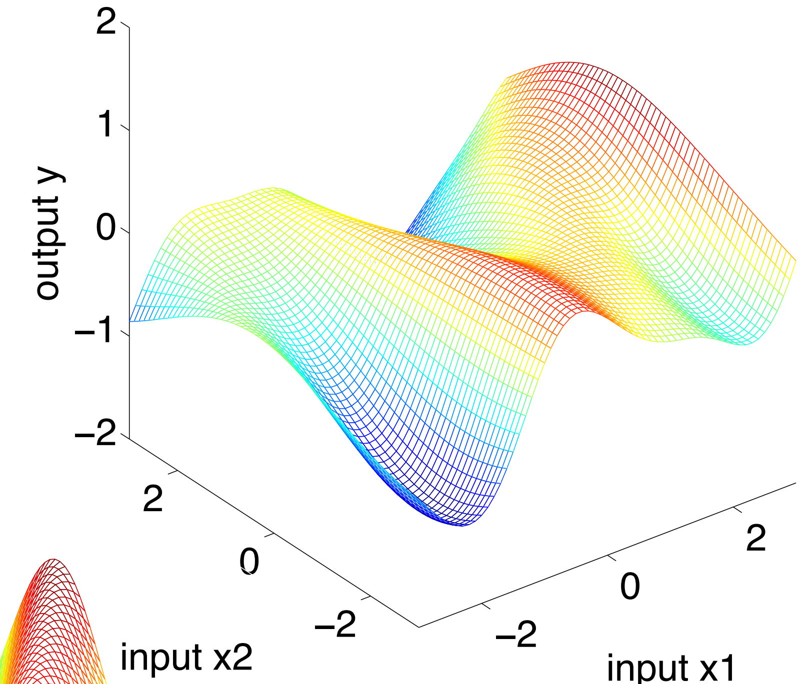


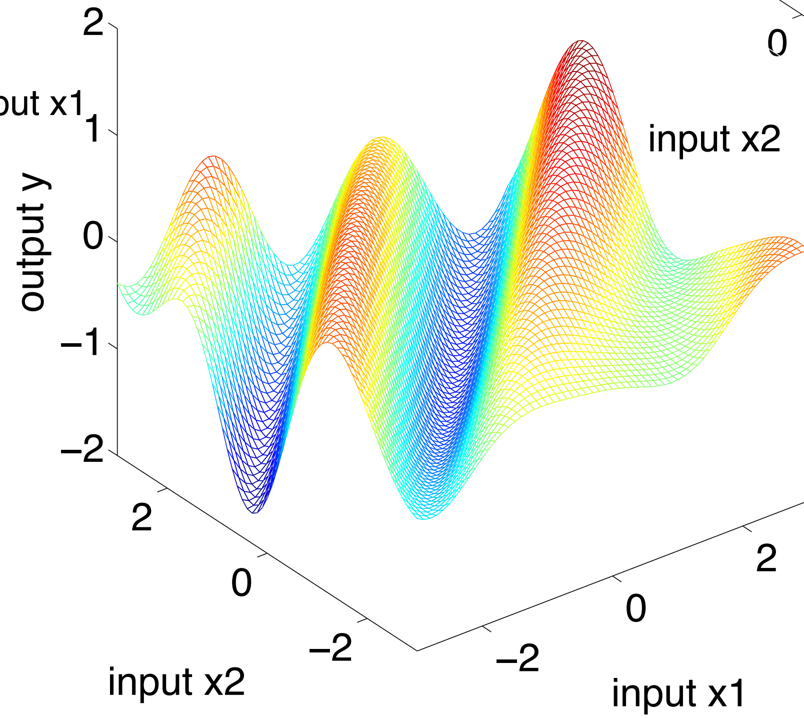
Posterior Given 5 Noise-Free Observations

Squared exponential kernel or radial basis function (RBF) kernel has a countably *infinite* set of underlying feature functions

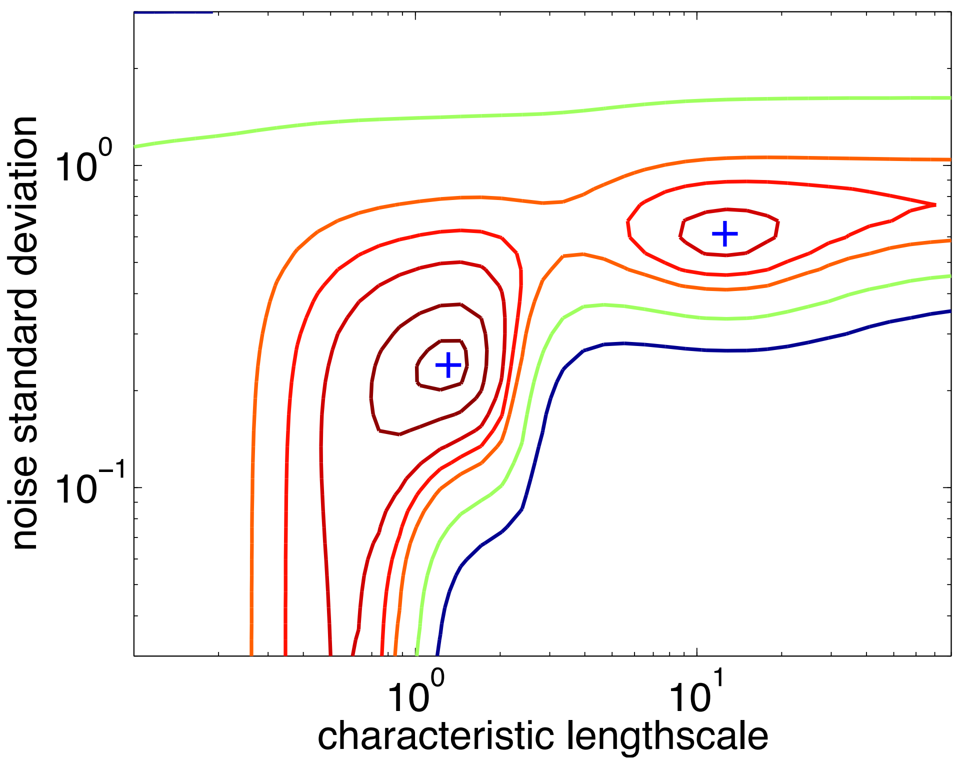
2D Gaussian Processes

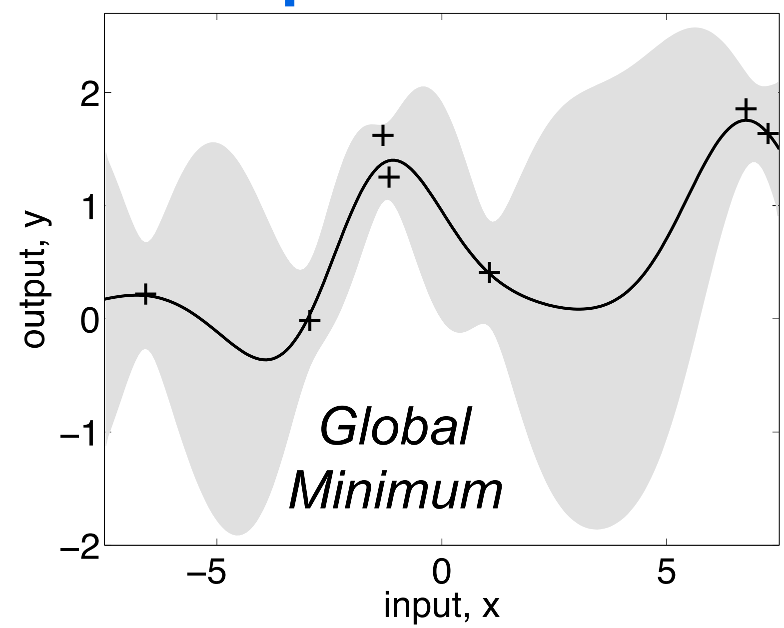


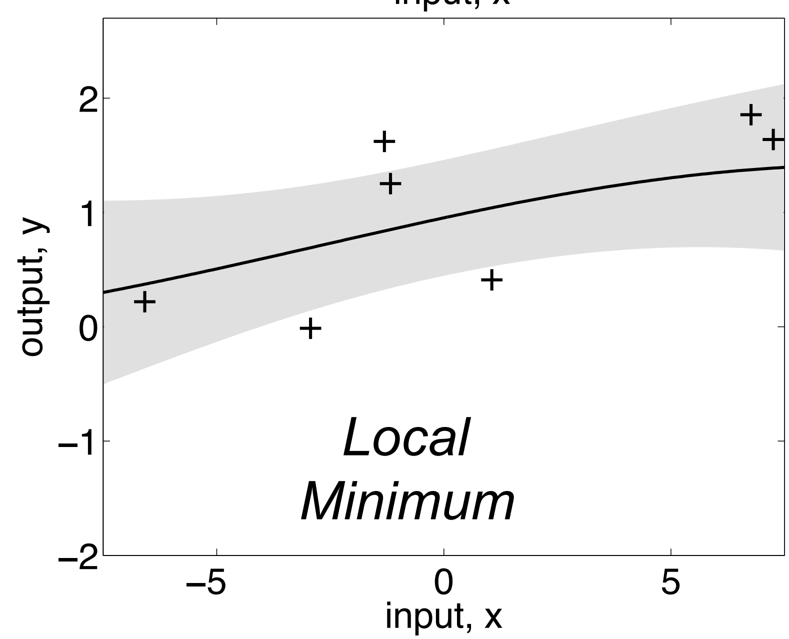




General Issue: Local Optima





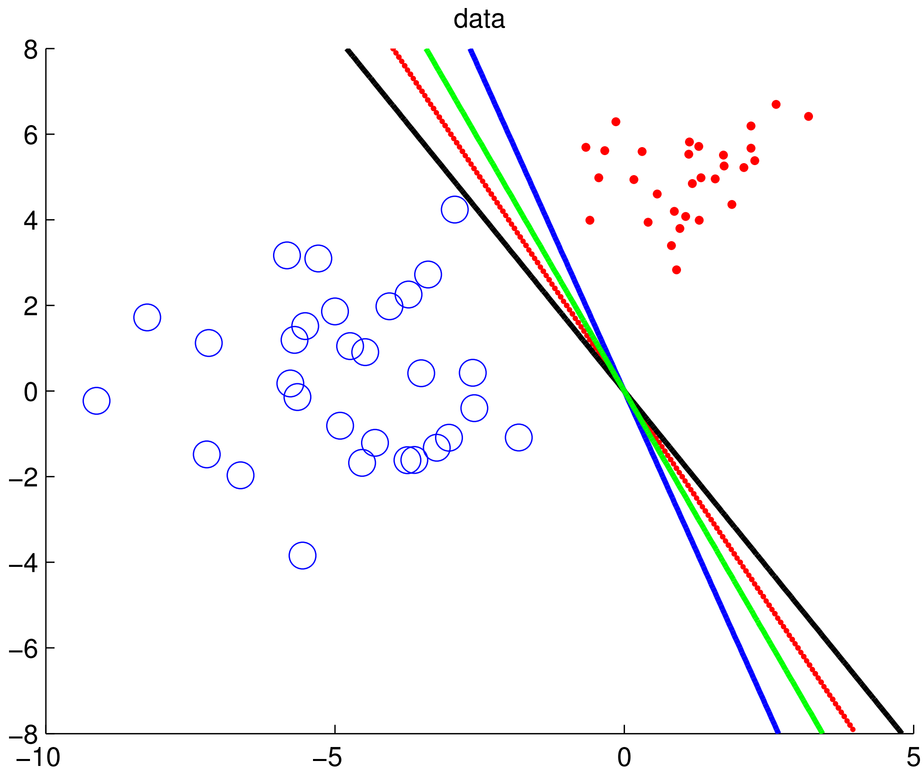


$$p(\mathbf{y}|\mathbf{X}) = \int p(\mathbf{y}|\mathbf{f}, \mathbf{X}) p(\mathbf{f}|\mathbf{X}) d\mathbf{f}$$

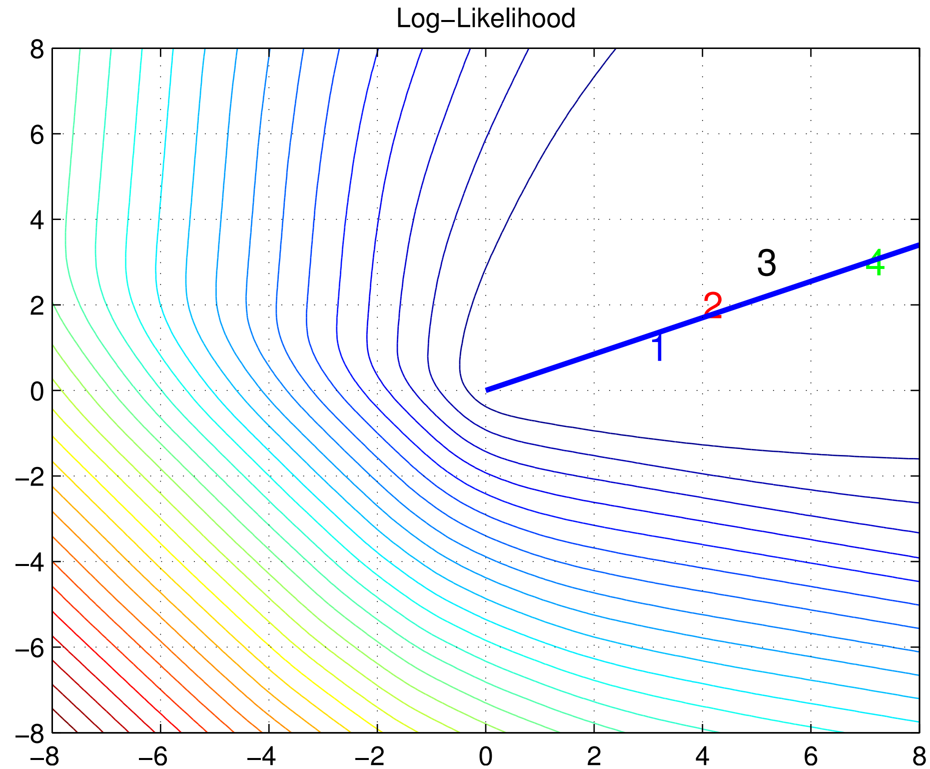
$$\log p(\mathbf{y}|\mathbf{X}) = -\frac{1}{2}\mathbf{y}\mathbf{K}_y^{-1}\mathbf{y} - \frac{1}{2}\log|\mathbf{K}_y| - \frac{N}{2}\ln(2\pi)$$

General Trick: Nonlinear Transforms

Board: Parametric versus nonparametric ***generalized linear models***



Linearly Separable Data

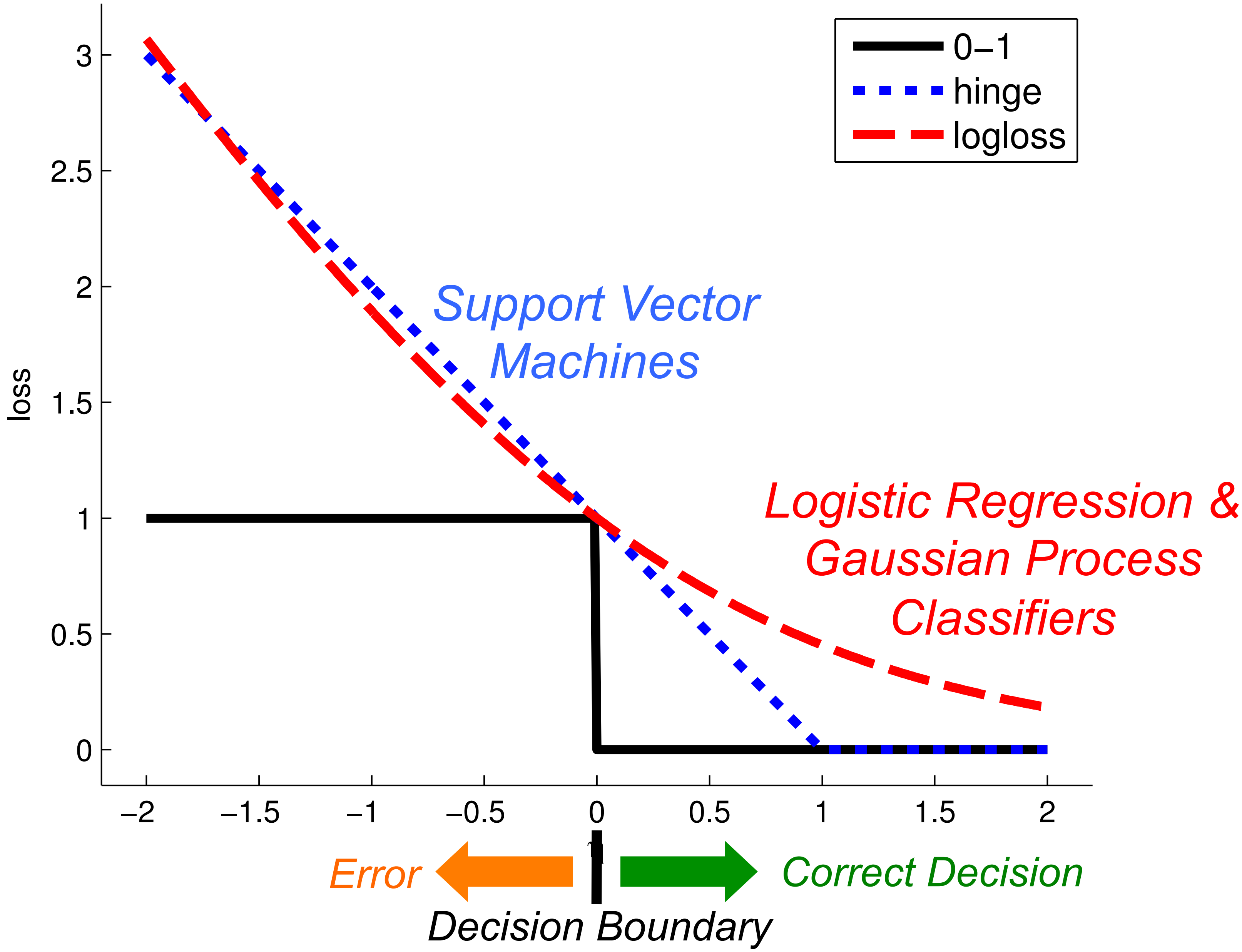


Log-likelihood Function

$$p(\mathbf{y}|\mathbf{X},\mathbf{w}) = \prod_{i:y_i=1} \frac{e^{\mathbf{w}^T\mathbf{x}_i}}{1+e^{\mathbf{w}^T\mathbf{x}_i}} \prod_{i:y_i=0} \frac{1}{1+e^{\mathbf{w}^T\mathbf{x}_i}} = \exp(\mathbf{w}^T \sum_i y_i \mathbf{x}_i) \prod_{i=1}^{N} (1+e^{\mathbf{w}^T\mathbf{x}_i})^{-1}$$

Linear Regression ⟺ **Logistic Regression** ***as*** **GP Regression** ⟺ **GP Classification**

Aside: Loss & Binary Classification



Discrete Distributions

Categorical Distribution:

$$p(x \mid \pi_1, \ldots, \pi_K) = \prod_{k=1}^{K} \pi_k^{\delta(x,k)} \qquad \delta(x,k) \triangleq \begin{cases} 1 & x = k \\ 0 & x \neq k \end{cases}$$

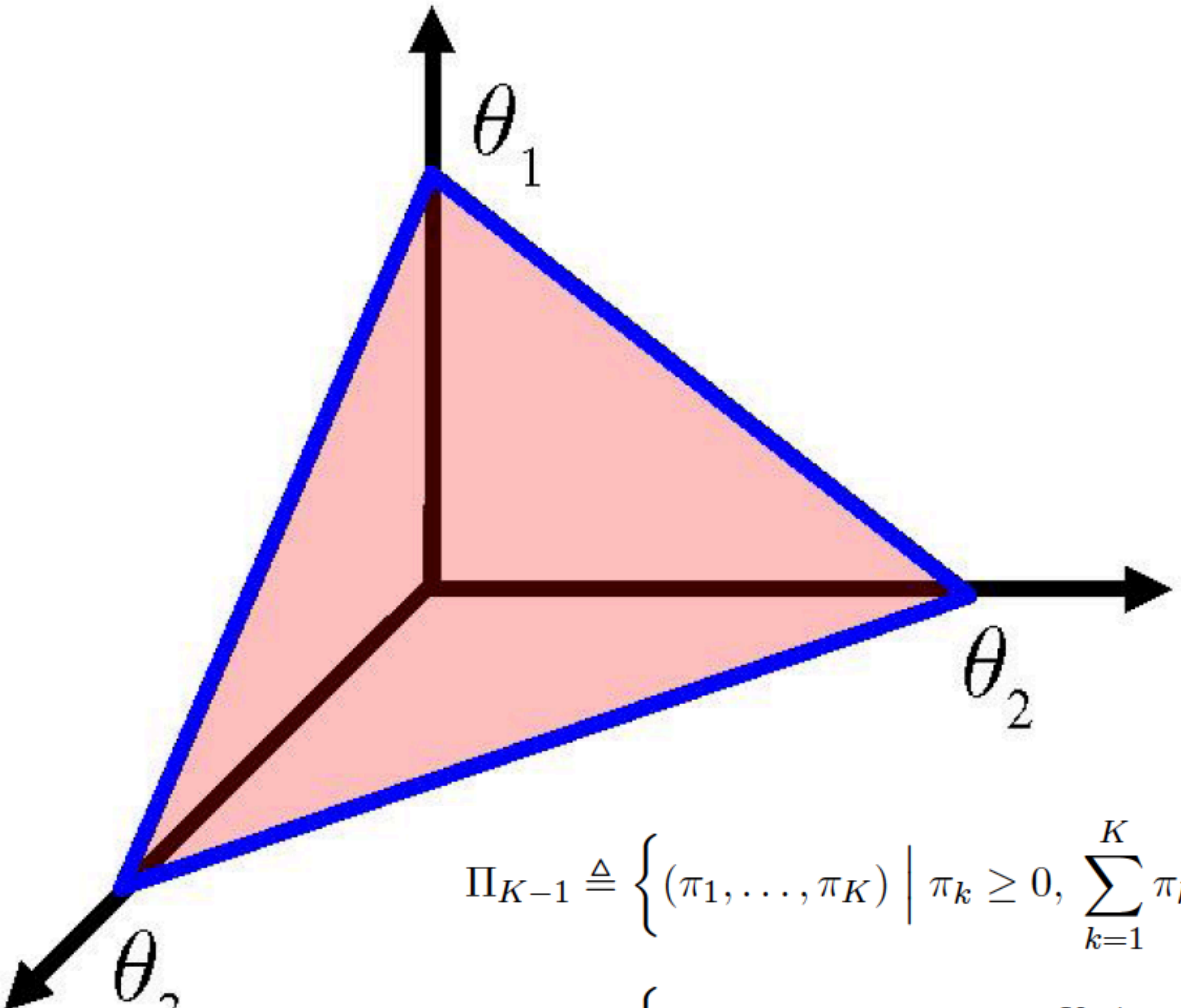
- When K=2, becomes Bernoulli distribution (one parameter)

Multinomial Distribution:

$$p(x^{(1)}, \ldots, x^{(L)} \mid \pi_1, \ldots, \pi_K) = \frac{L!}{\prod_k C_k!} \prod_{k=1}^{K} \pi_k^{C_k} \qquad C_k \triangleq \sum_{\ell=1}^{L} \delta(x^{(\ell)}, k)$$

- Probability of collection of L categorical outcomes, ignoring the order in which those outcomes occurred
- When K=2, becomes binomial distribution

Multinomial Simplex



$$\Pi_{K-1} \triangleq \left\{ (\pi_1, \ldots, \pi_K) \;\middle|\; \pi_k \geq 0,\ \sum_{k=1}^{K} \pi_k = 1 \right\}$$

$$= \left\{ (\pi_1, \ldots, \pi_{K-1}, 1 - \textstyle\sum_{k=1}^{K-1} \pi_k) \;\middle|\; \pi_k \geq 0,\ \displaystyle\sum_{k=1}^{K-1} \pi_k \leq 1 \right\}$$

Exponential Families

- Natural or canonical parameters determine log-linear combination of sufficient statistics:

$$p(x \mid \theta) = \nu(x) \exp\left\{ \sum_{a \in \mathcal{A}} \theta_a \phi_a(x) - \Phi(\theta) \right\}$$

- Log partition function normalizes to produce valid probability distribution:

$$\Phi(\theta) = \log \int_{\mathcal{X}} \nu(x) \exp\left\{ \sum_{a \in \mathcal{A}} \theta_a \phi_a(x) \right\} dx$$

$$\Theta \triangleq \left\{ \theta \in \mathbb{R}^{|\mathcal{A}|} \mid \Phi(\theta) < \infty \right\}$$

Sufficiency

Theorem 2.1.2. *Let $p(x \mid \theta)$ denote an exponential family with canonical parameters θ, and $p(\theta \mid \lambda)$ a corresponding prior density. Given L independent, identically distributed samples $\{x^{(\ell)}\}_{\ell=1}^{L}$, consider the following statistics:*

$$\boldsymbol{\phi}(x^{(1)}, \ldots, x^{(L)}) \triangleq \left\{ \frac{1}{L} \sum_{\ell=1}^{L} \phi_a(x^{(\ell)}) \;\middle|\; a \in \mathcal{A} \right\} \tag{2.24}$$

These empirical moments, along with the sample size L, are then said to be parametric sufficient *for the posterior distribution over canonical parameters, so that*

$$p(\theta \mid x^{(1)}, \ldots, x^{(L)}, \lambda) = p(\theta \mid \boldsymbol{\phi}(x^{(1)}, \ldots, x^{(L)}), L, \lambda) \tag{2.25}$$

Equivalently, they are predictive sufficient *for the likelihood of new data $\bar{x}$:*

$$p(\bar{x} \mid x^{(1)}, \ldots, x^{(L)}, \lambda) = p(\bar{x} \mid \boldsymbol{\phi}(x^{(1)}, \ldots, x^{(L)}), L, \lambda) \tag{2.26}$$

Conjugate Priors

- For any family of distributions with hyperparameters λ:

$$p(\theta \mid x, \lambda) \propto p(x \mid \theta)\, p(\theta \mid \lambda) \propto p\big(\theta \mid \bar{\lambda}\big)$$

- Excluding degeneracies, only possible for exponential families:

$$p(x \mid \theta) = \nu(x) \exp\left\{ \sum_{a \in \mathcal{A}} \theta_a \phi_a(x) - \Phi(\theta) \right\}$$

$$p(\theta \mid \lambda) = \exp\left\{ \sum_{a \in \mathcal{A}} \theta_a \lambda_0 \lambda_a - \lambda_0 \Phi(\theta) - \Omega(\lambda) \right\}$$

$$\Omega(\lambda) = \log \int_{\Theta} \exp\left\{ \sum_{a \in \mathcal{A}} \theta_a \lambda_0 \lambda_a - \lambda_0 \Phi(\theta) \right\} d\theta$$

Conjugate Posteriors

Proposition 2.1.4. *Let $p(x \mid \theta)$ denote an exponential family with canonical parameters θ, and $p(\theta \mid \lambda)$ a family of conjugate priors defined as in eq.* (2.28)*. Given L independent samples $\{x^{(\ell)}\}_{\ell=1}^{L}$, the posterior distribution remains in the same family:*

$$p(\theta \mid x^{(1)}, \ldots, x^{(L)}, \lambda) = p\big(\theta \mid \bar{\lambda}\big) \tag{2.31}$$

$$\bar{\lambda}_0 = \lambda_0 + L \qquad \bar{\lambda}_a = \frac{\lambda_0 \lambda_a + \sum_{\ell=1}^{L} \phi_a(x^{(\ell)})}{\lambda_0 + L} \qquad a \in \mathcal{A} \tag{2.32}$$

Integrating over Θ, the log–likelihood of the observations can then be compactly written using the normalization constant of eq. (2.29)*:*

$$\log p(x^{(1)}, \ldots, x^{(L)} \mid \lambda) = \Omega\big(\bar{\lambda}\big) - \Omega(\lambda) + \sum_{\ell=1}^{L} \log \nu(x^{(\ell)}) \tag{2.33}$$

Finite Dirichlet Distributions

$$p(\pi \mid \alpha) = \frac{\Gamma(\sum_k \alpha_k)}{\prod_k \Gamma(\alpha_k)} \prod_{k=1}^{K} \pi_k^{\alpha_k - 1} \qquad \alpha_k > 0$$

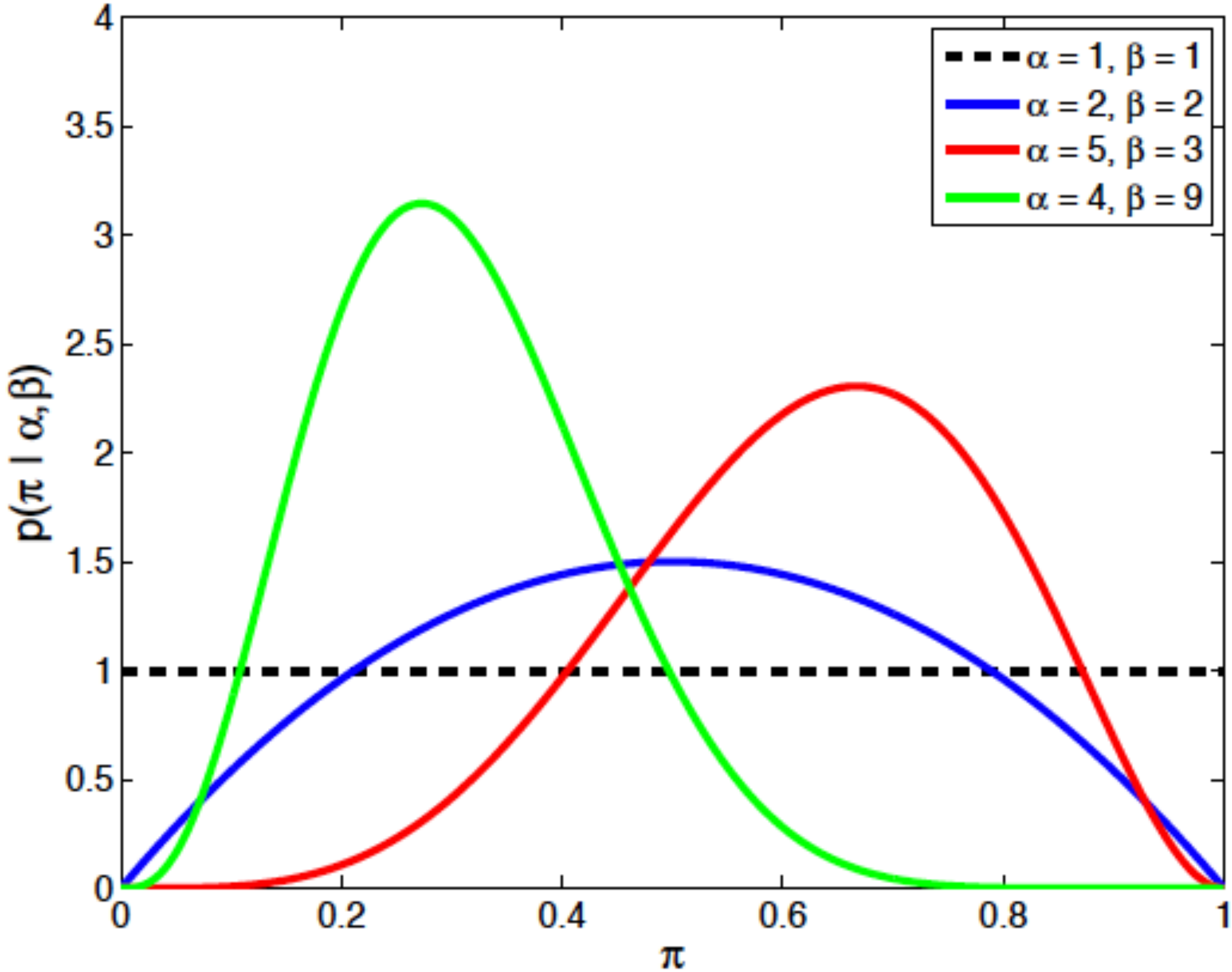
$$\mathbb{E}_\alpha[\pi_k] = \frac{\alpha_k}{\alpha_0} \qquad \alpha_0 \triangleq \sum_{k=1}^{K} \alpha_k$$

$$\mathrm{Var}_\alpha[\pi_k] = \frac{K-1}{K^2(\alpha_0 + 1)} \qquad \alpha_k = \frac{\alpha_0}{K}$$

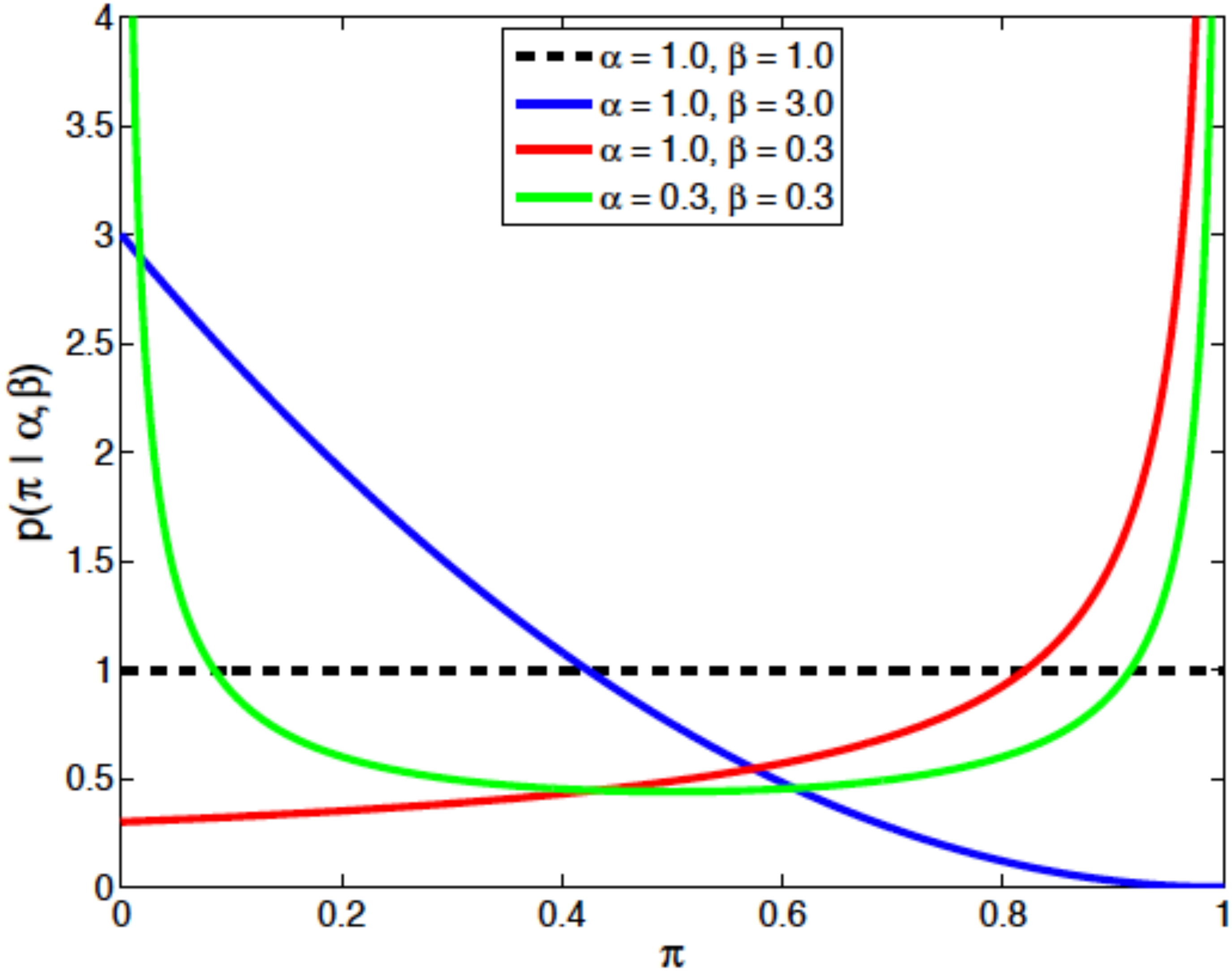
- Beta distribution is special case where K=2:

$$p(\pi \mid \alpha, \beta) = \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)\,\Gamma(\beta)} \pi^{\alpha - 1}(1 - \pi)^{\beta - 1} \qquad \alpha, \beta > 0$$

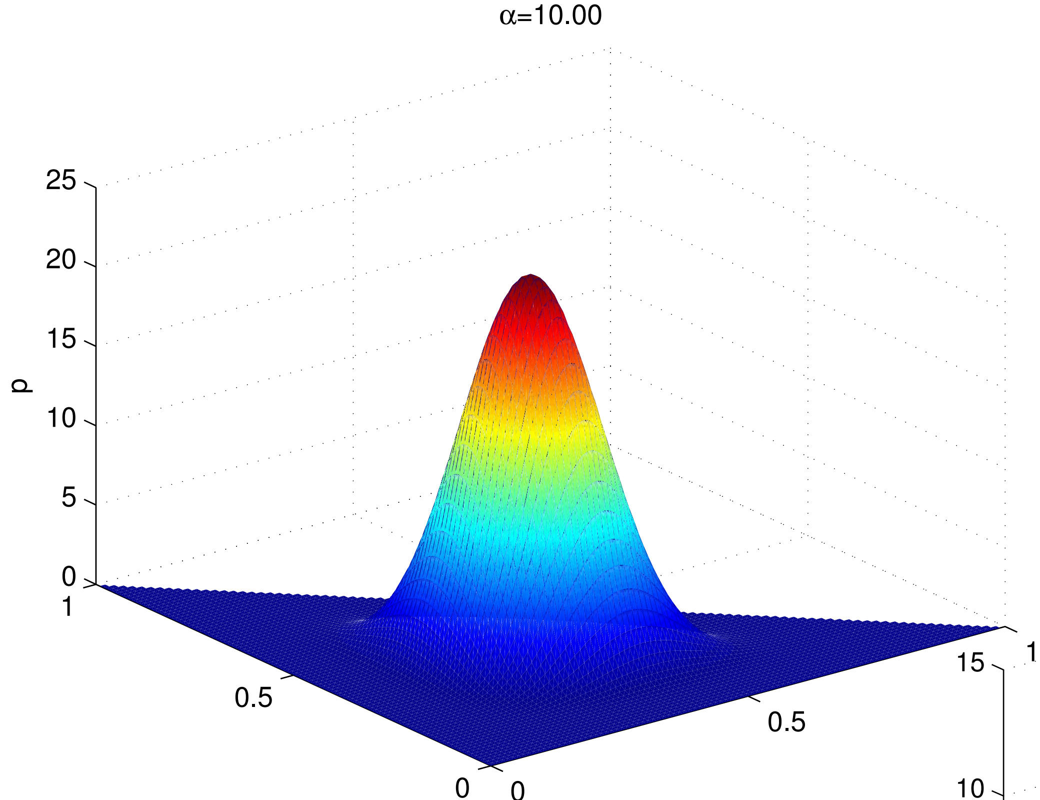
Beta Distributions

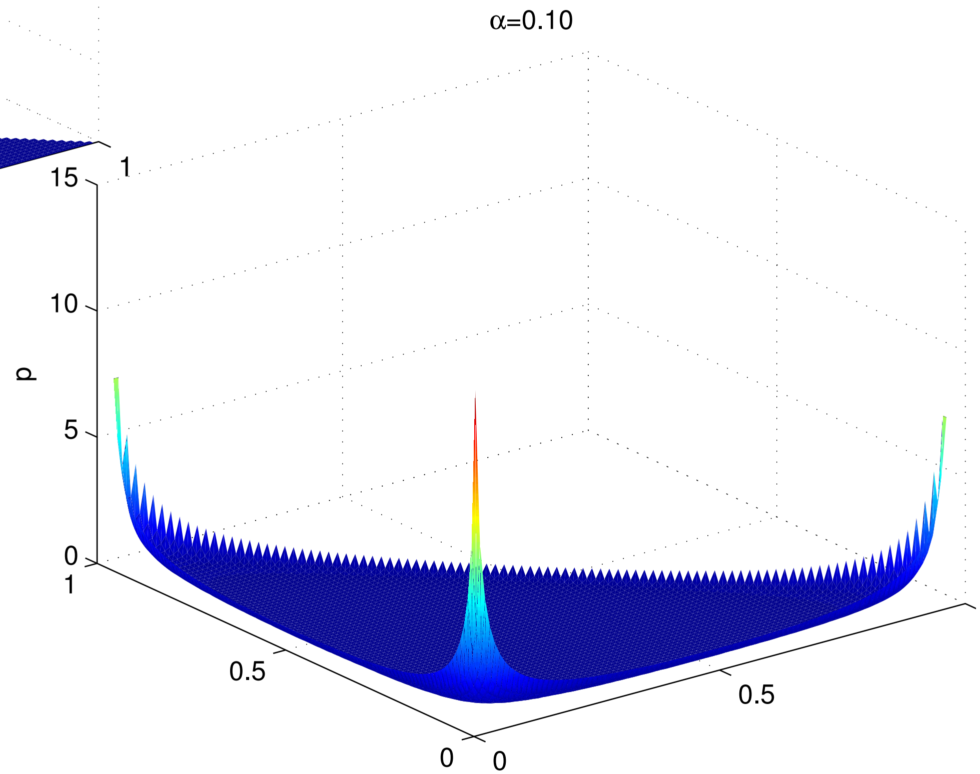


Beta Distributions

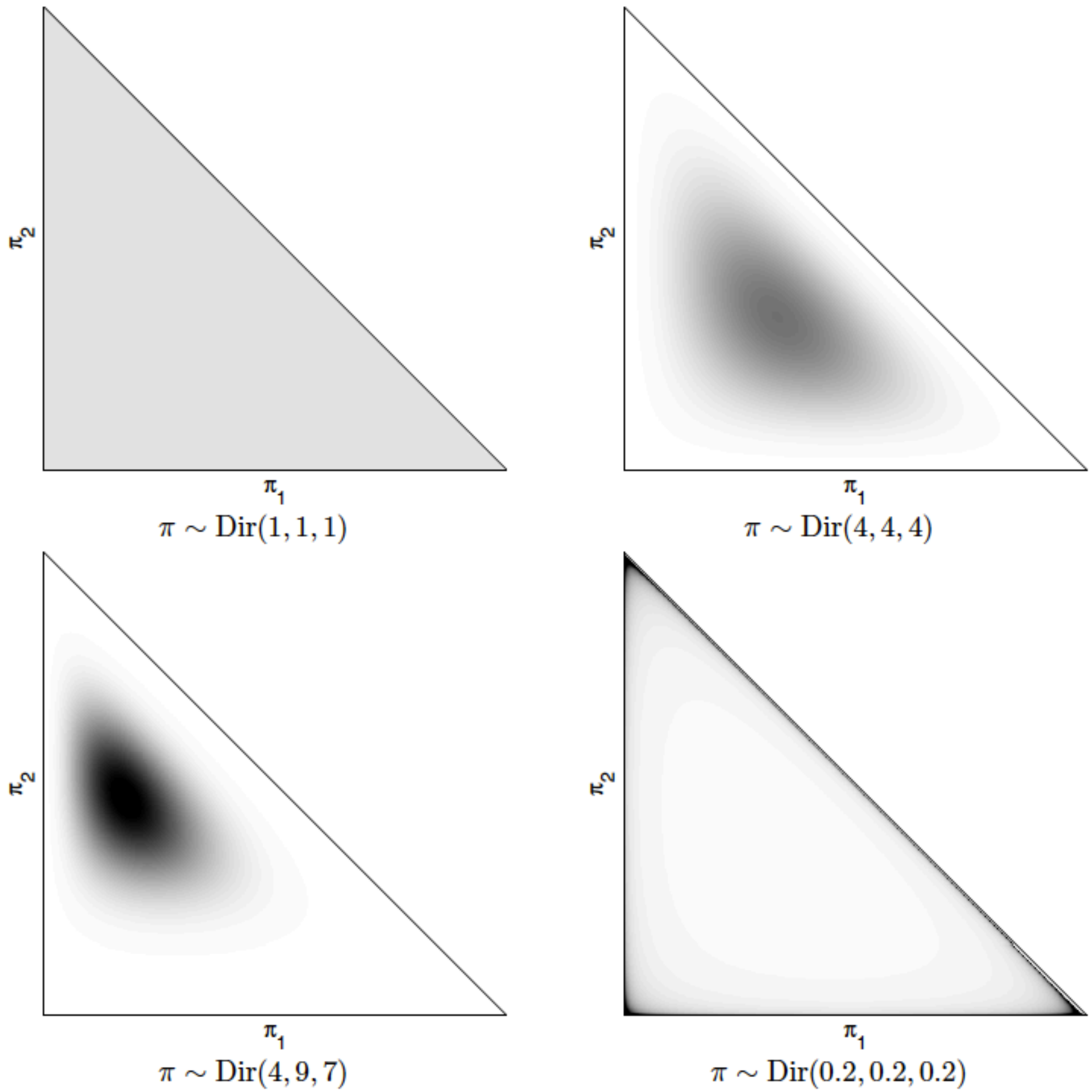


Dirichlet Distributions





Dirichlet Distributions



Posteriors and Marginals

$$p(\pi \mid x^{(1)}, \dots, x^{(L)}, \alpha) \propto p(\pi \mid \alpha)\, p(x^{(1)}, \dots, x^{(L)} \mid \pi)$$

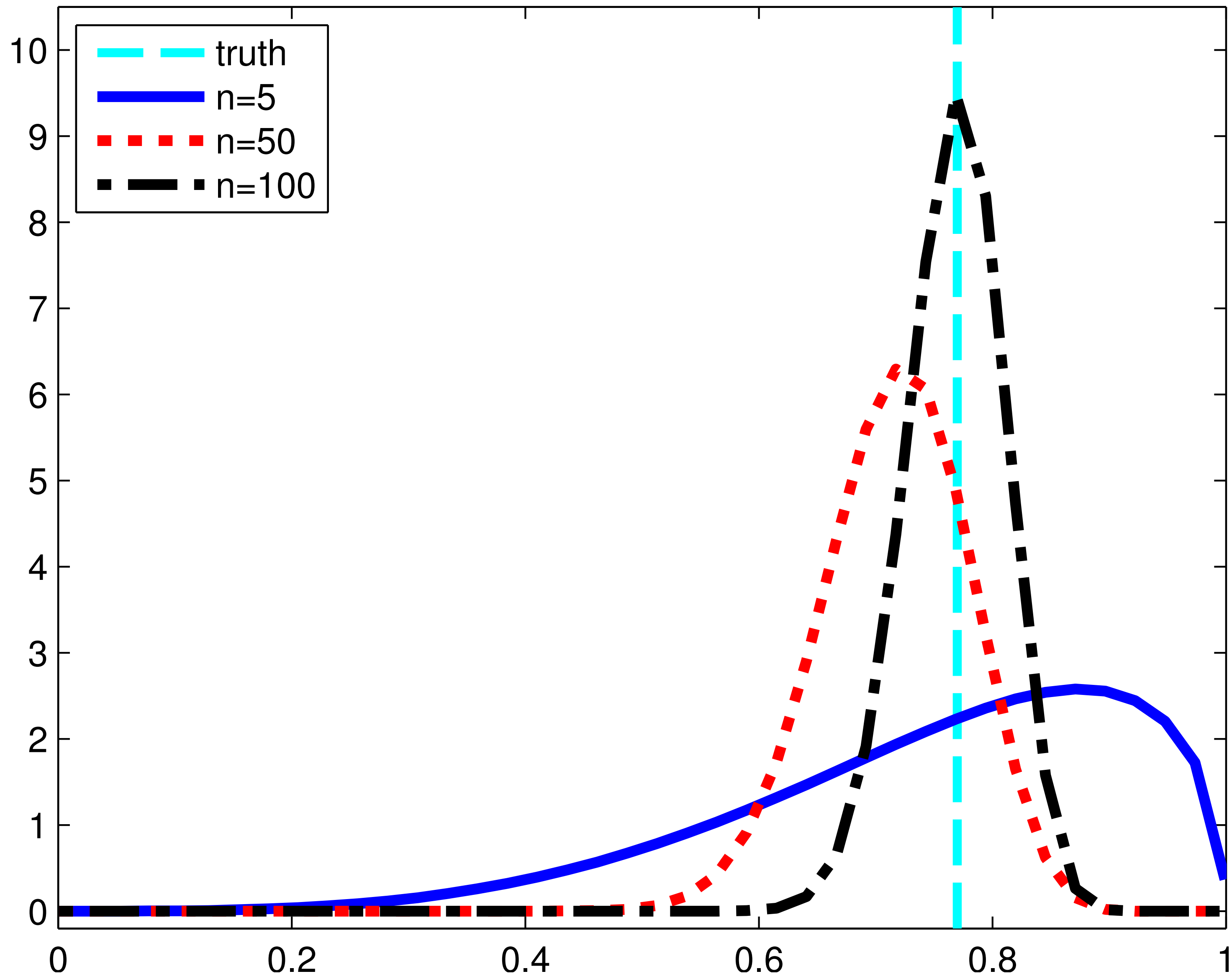
$$\propto \prod_{k=1}^{K} \pi_k^{\alpha_k + C_k - 1} \propto \mathrm{Dir}(\alpha_1 + C_1, \dots, \alpha_K + C_K)$$

$$p(\bar{x} = k \mid x^{(1)}, \dots, x^{(L)}, \alpha) = \frac{C_k + \alpha_k}{L + \alpha_0}$$

$$(\pi_1 + \pi_2, \pi_3, \dots, \pi_K) \sim \mathrm{Dir}(\alpha_1 + \alpha_2, \alpha_3, \dots, \alpha_K)$$

$$\pi_k \sim \mathrm{Beta}(\alpha_k, \alpha_0 - \alpha_k)$$

A Sequence of Beta Posteriors



De Finetti's Theorem

- Finitely exchangeable random variables satisfy:

$$p(x_1, \ldots, x_N) = p\big(x_{\tau(1)}, \ldots, x_{\tau(N)}\big) \qquad \text{for any permutation } \tau(\cdot)$$

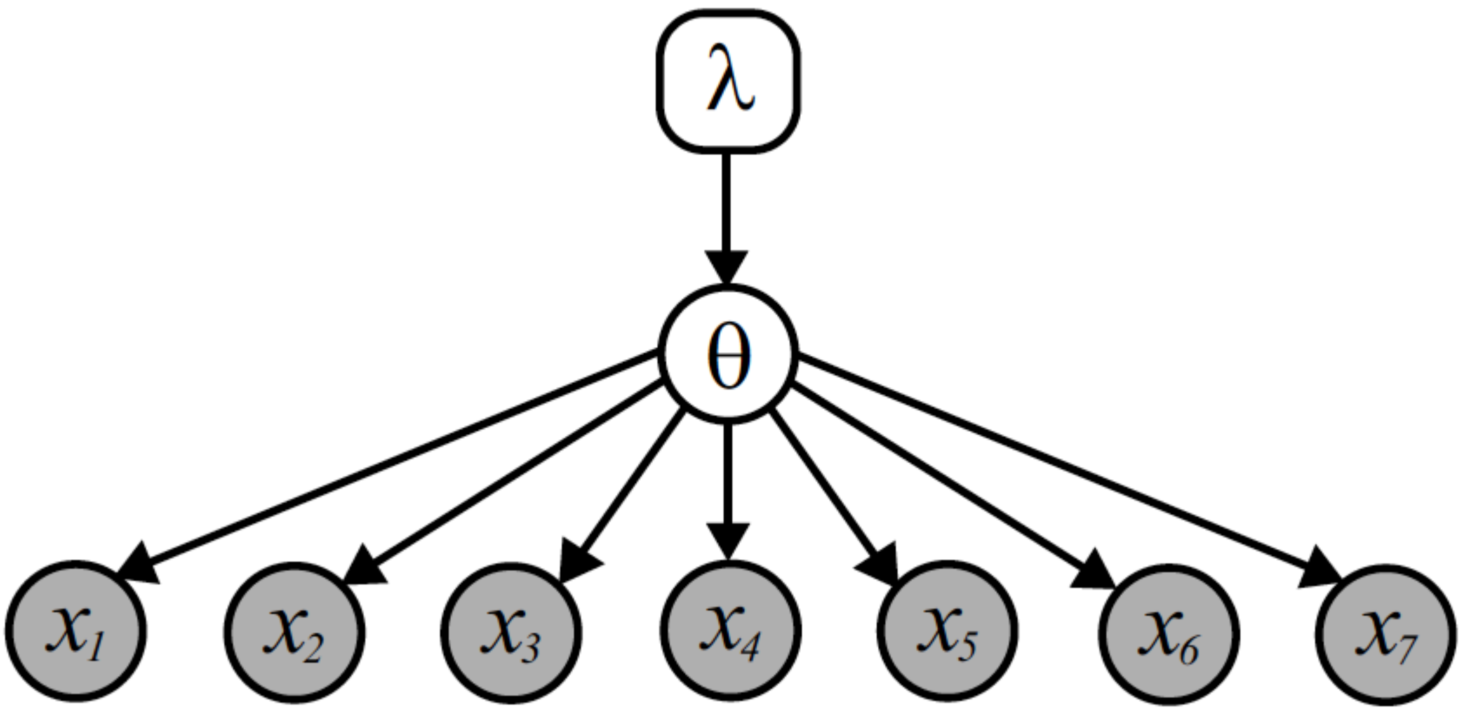
- A sequence is infinitely exchangeable if every finite subsequence is exchangeable
- Exchangeable variables need not be independent, but always have a representation with conditional independencies:

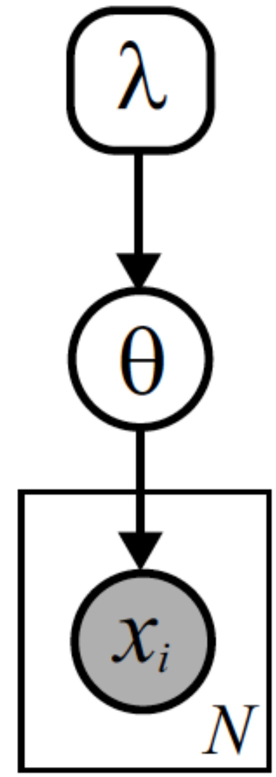
Theorem 2.2.2 (De Finetti). *For any infinitely exchangeable sequence of random variables $\{x_i\}_{i=1}^{\infty}$, $x_i \in \mathcal{X}$, there exists some space Θ, and corresponding density $p(\theta)$, such that the joint probability of any N observations has a mixture representation:*

$$p(x_1, x_2, \ldots, x_N) = \int_\Theta p(\theta) \prod_{i=1}^{N} p(x_i \mid \theta)\; d\theta \tag{2.77}$$

When $\mathcal{X}$ is a K–dimensional discrete space, Θ may be chosen as the $(K-1)$–simplex. For Euclidean $\mathcal{X}$, Θ is an infinite–dimensional space of probability measures.

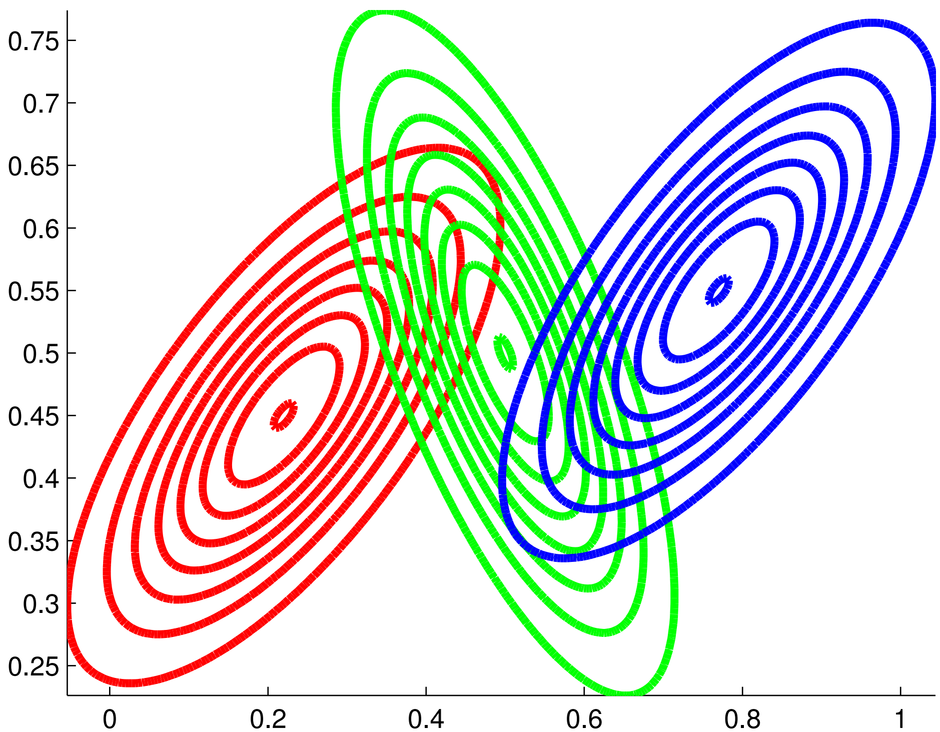
De Finetti's Directed Graph



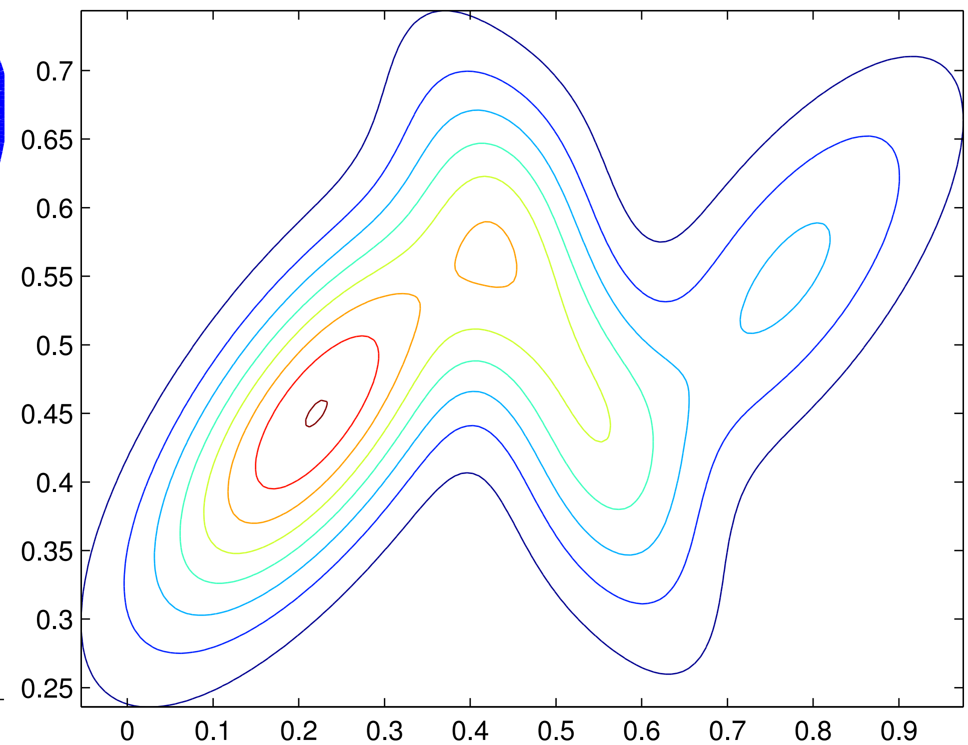


$$p(x_1, \ldots, x_N, \theta \mid \lambda) = p(\theta \mid \lambda) \prod_{i=1}^{N} p(x_i \mid \theta)$$

Gaussian Mixture Models

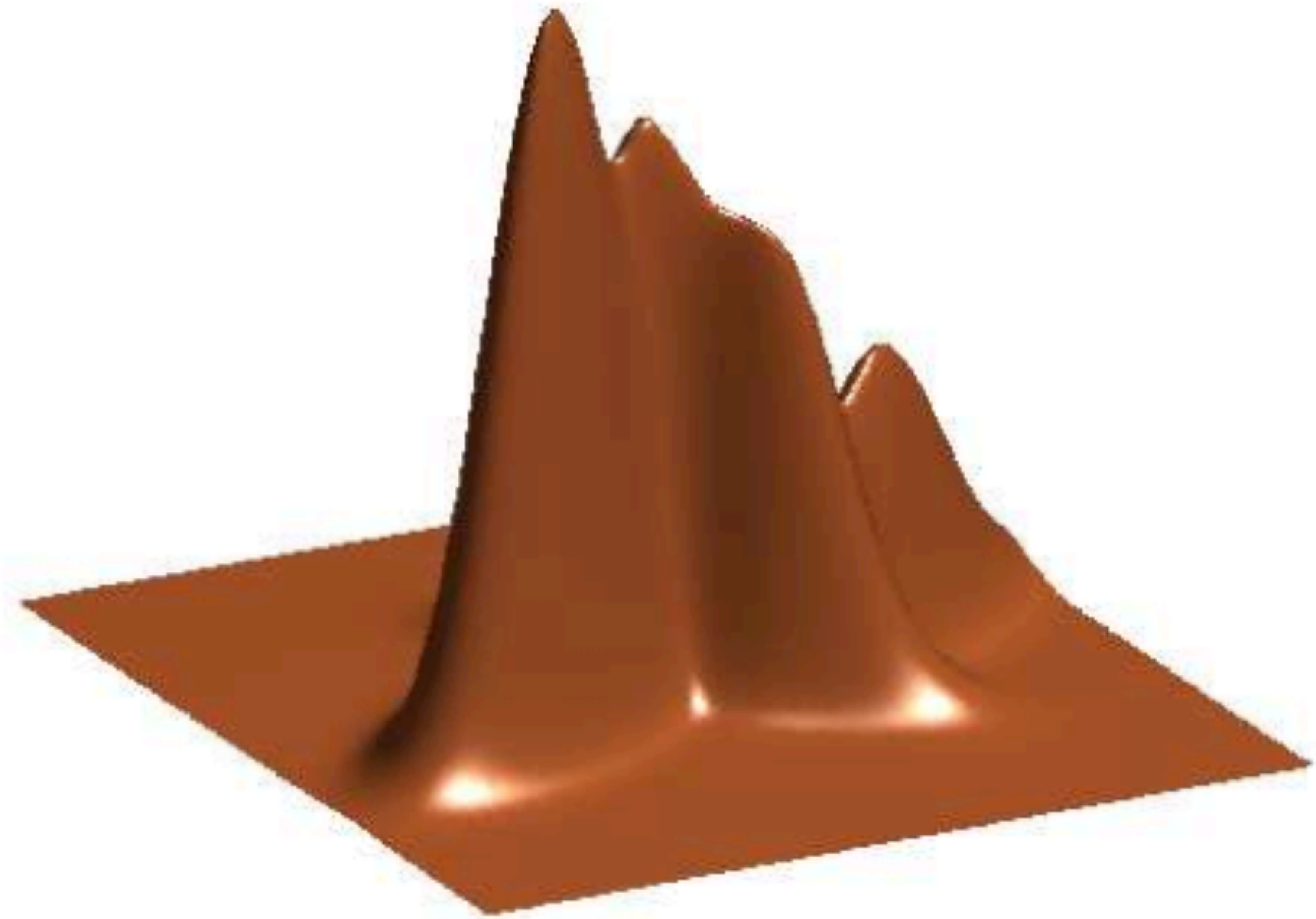


Mixture of 3 Gaussian Distributions in 2D



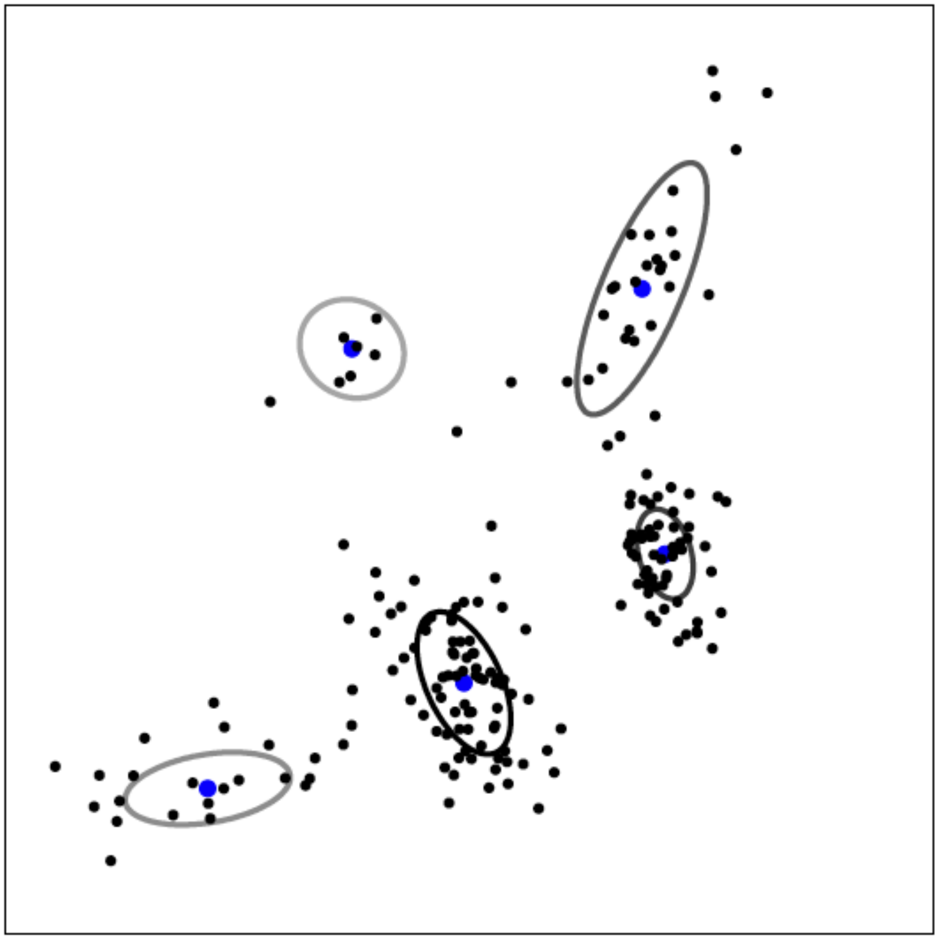
Contour Plot of Joint Density, Marginalizing Cluster Assignments

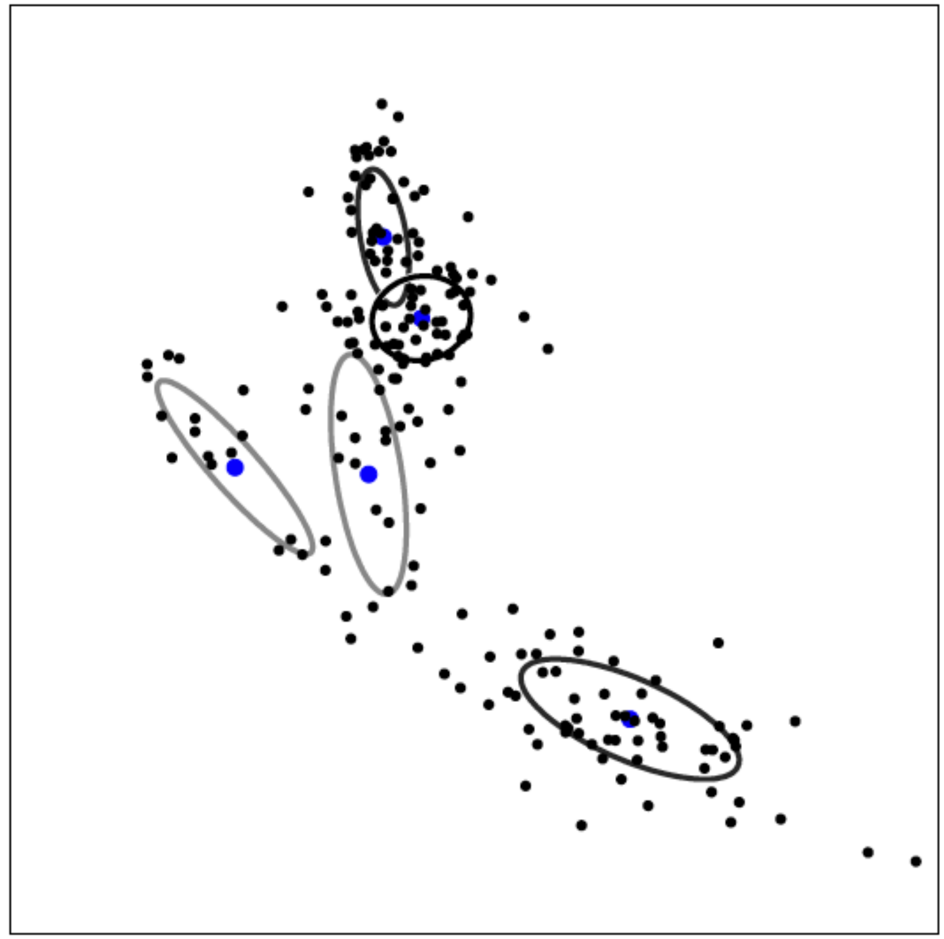
Gaussian Mixture Models



Surface Plot of Joint Density,
Marginalizing Cluster Assignments

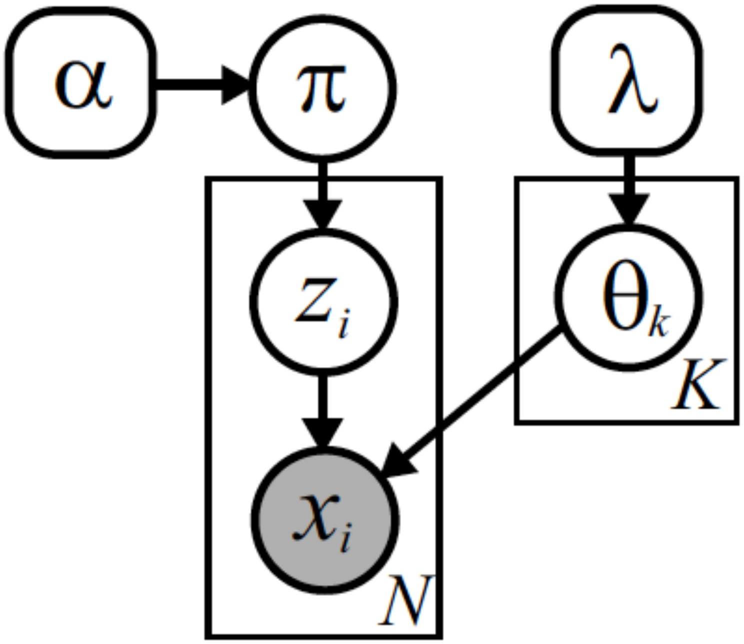
Generative Gaussian Mixture Samples

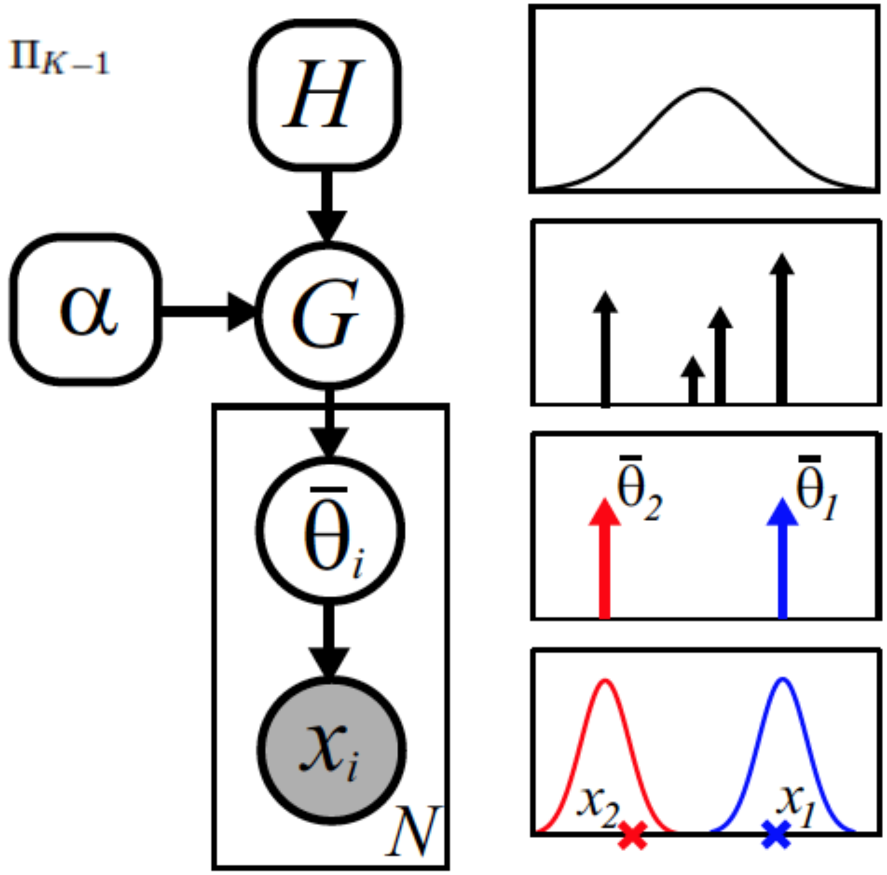




Finite Bayesian Mixture Models

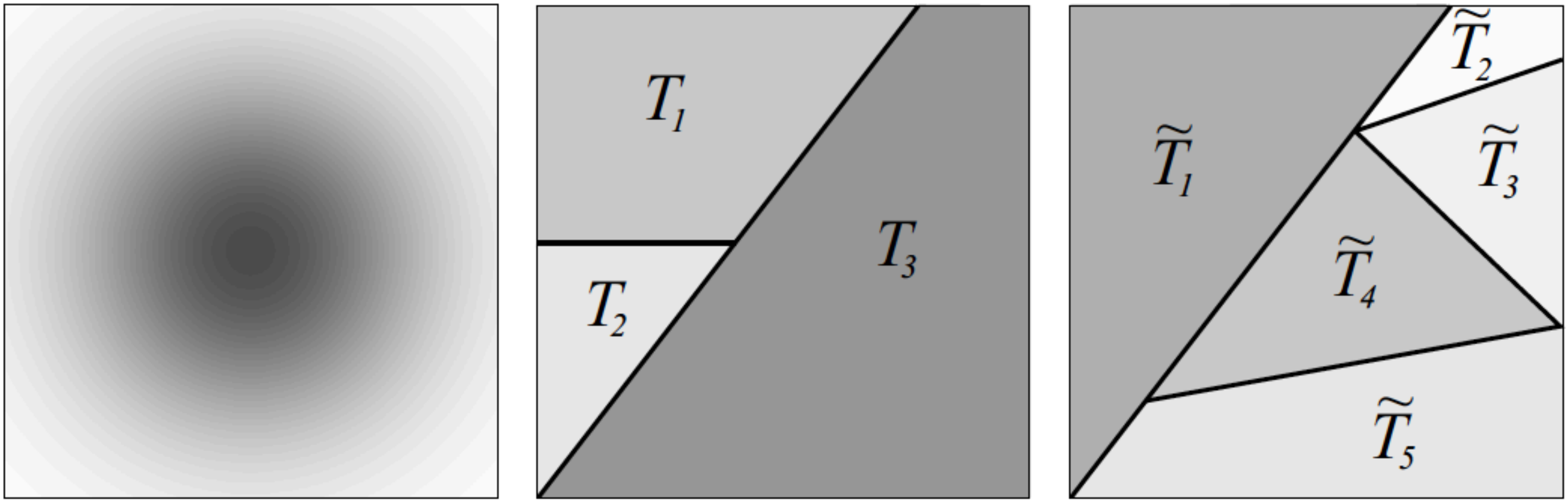
$$p(x \mid \pi, \theta_1, \ldots, \theta_K) = \sum_{k=1}^{K} \pi_k f(x \mid \theta_k) \qquad \pi \in \Pi_{K-1}$$





- Board: Assignment variable and distribution representations

Dirichlet Processes



$$\mathbb{E}[G(T)] = H(T) \qquad\qquad G \sim \mathrm{DP}(\alpha, H)$$

Dirichlet Processes

Theorem 2.5.1. *Let H be a probability distribution on a measurable space Θ, and α a positive scalar. Consider a finite partition $(T_1, \ldots, T_K)$ of Θ:*

$$\bigcup_{k=1}^{K} T_k = \Theta \qquad T_k \cap T_\ell = \emptyset \qquad k \neq \ell \tag{2.165}$$

A random probability distribution G on Θ is drawn from a Dirichlet process if its measure on every finite partition follows a Dirichlet distribution:

$$(G(T_1), \ldots, G(T_K)) \sim \mathrm{Dir}(\alpha H(T_1), \ldots, \alpha H(T_K)) \tag{2.166}$$

For any base measure H and concentration parameter α, there exists a unique stochastic process satisfying these conditions, which we denote by $\mathrm{DP}(\alpha, H)$.

Proof Hint: Kolmogorov's Theorem requires consistency of the specified finite-dimensional marginal distributions

$$(\pi_1 + \pi_2, \pi_3, \ldots, \pi_K) \sim \mathrm{Dir}(\alpha_1 + \alpha_2, \alpha_3, \ldots, \alpha_K)$$

$$\pi_k \sim \mathrm{Beta}(\alpha_k, \alpha_0 - \alpha_k)$$

DP Posteriors and Conjugacy

Proposition 2.5.1. *Let $G \sim \mathrm{DP}(\alpha, H)$ be a random measure distributed according to a Dirichlet process. Given N independent observations $\bar{\theta}_i \sim G$, the posterior measure also follows a Dirichlet process:*

$$p(G \mid \bar{\theta}_1, \dots, \bar{\theta}_N, \alpha, H) = \mathrm{DP}\left(\alpha + N, \frac{1}{\alpha + N}\left(\alpha H + \sum_{i=1}^{N} \delta_{\bar{\theta}_i}\right)\right) \tag{2.169}$$

Proof Hint: For any finite partition, we have

$$p\big((G(T_1), \dots, G(T_K)) \mid \bar{\theta} \in T_k\big) = \mathrm{Dir}(\alpha H(T_1), \dots, \alpha H(T_k) + 1, \dots, \alpha H(T_K))$$

DPs are Neutral: "Almost" independent

The distribution of a random probability measure G is neutral with respect to a finite partition $(T_1, \ldots, T_K)$ *iff*

$$G(T_k) \quad \text{is independent of} \quad \left\{ \frac{G(T_\ell)}{1 - G(T_k)} \;\middle|\; \ell \neq k \right\}$$

$$\text{given that } G(T_k) < 1.$$

Theorem 2.5.2. *Consider a distribution $\mathcal{P}$ on probability measures G for some space Θ. Assume that $\mathcal{P}$ assigns positive probability to more than one measure G, and that with probability one samples $G \sim \mathcal{P}$ assign positive measure to at least three distinct points $\theta \in \Theta$. The following conditions are then equivalent:*

(i) $\mathcal{P} = \mathrm{DP}(\alpha, H)$ *is a Dirichlet process for some base measure H on Θ.*

(ii) $\mathcal{P}$ *is neutral with respect to every finite, measurable partition of Θ.*

(iii) For every measurable $T \subset \Theta$, and any N observations $\bar{\theta}_i \sim G$, the posterior distribution $p\big(G(T) \mid \bar{\theta}_1, \ldots, \bar{\theta}_N\big)$ depends only on the number of observations that fall within T (and not their particular locations).

DPs and Stick Breaking

Theorem 2.5.3. *Let $\pi = \{\pi_k\}_{k=1}^{\infty}$ be an infinite sequence of mixture weights derived from the following stick–breaking process, with parameter $\alpha > 0$:*

$$\beta_k \sim \mathrm{Beta}(1, \alpha) \qquad\qquad k = 1, 2, \ldots \tag{2.174}$$

$$\pi_k = \beta_k \prod_{\ell=1}^{k-1} (1 - \beta_\ell) = \beta_k \left(1 - \sum_{\ell=1}^{k-1} \pi_\ell \right) \tag{2.175}$$

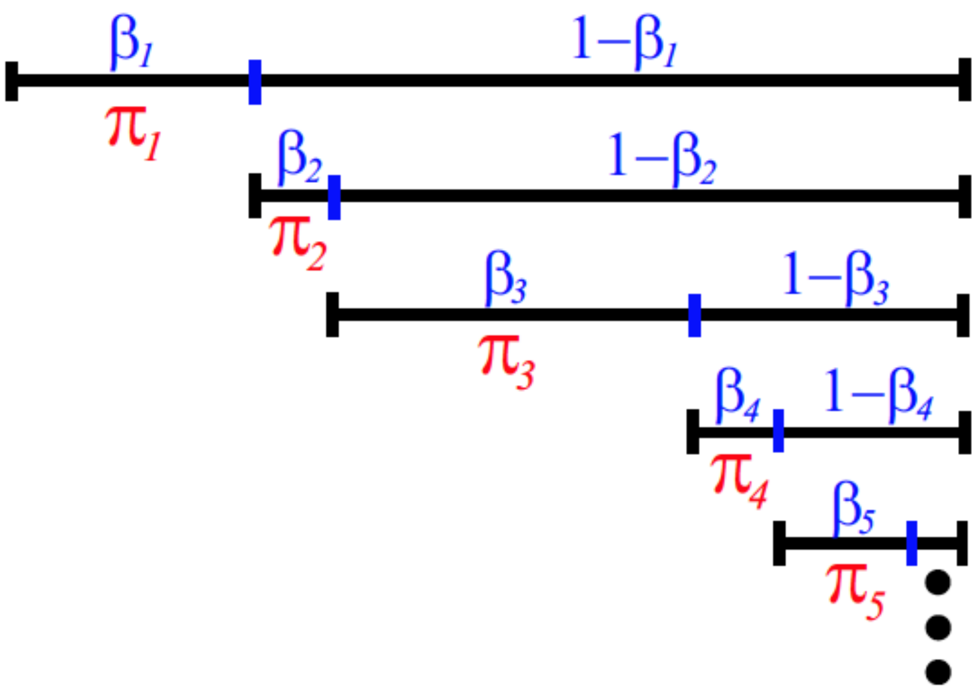
Given a base measure H on Θ, consider the following discrete random measure:

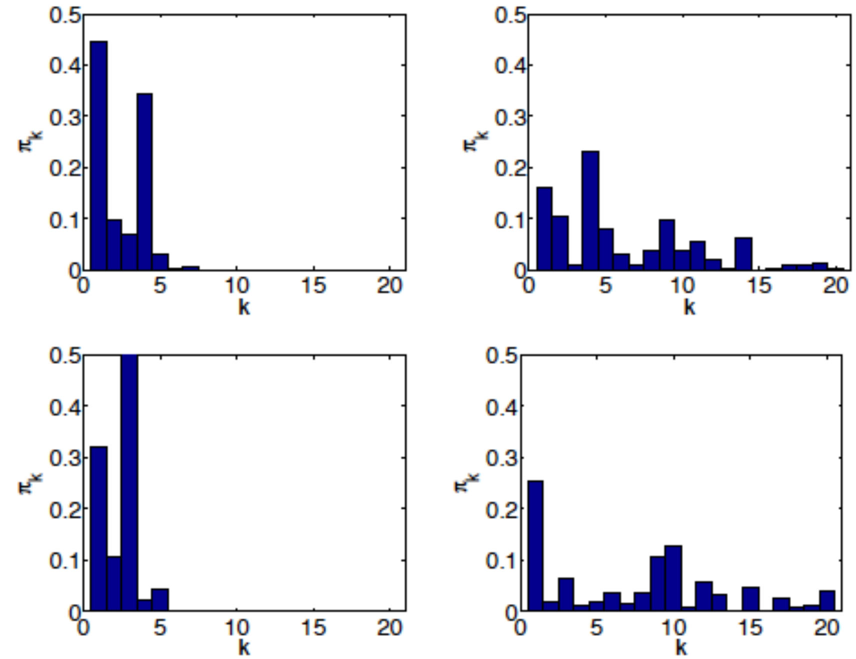
$$G(\theta) = \sum_{k=1}^{\infty} \pi_k \delta(\theta, \theta_k) \qquad\qquad \theta_k \sim H \tag{2.176}$$

This construction guarantees that $G \sim \mathrm{DP}(\alpha, H)$. Conversely, samples from a Dirichlet process are discrete with probability one, and have a representation as in eq. (2.176).

- Board: Intuition for why DP samples must be discrete

DPs and Stick Breaking





$\alpha = 1$ $\alpha = 5$

$$\pi_k = \beta_k \prod_{\ell=1}^{k-1}(1-\beta_\ell) = \beta_k \left(1 - \sum_{\ell=1}^{k-1} \pi_\ell\right)$$

$$1 - \sum_{k=1}^{K} \pi_k = \prod_{k=1}^{K}(1-\beta_k) \longrightarrow 0$$

$$\beta_k \sim \mathrm{Beta}(1, \alpha)$$

$$\mathbb{E}[\beta_k] = \frac{1}{1+\alpha}$$

DPs and Polya Urns

Theorem 2.5.4. *Let $G \sim \mathrm{DP}(\alpha, H)$ be distributed according to a Dirichlet process, where the base measure H has corresponding density $h(\theta)$. Consider a set of N observations $\bar{\theta}_i \sim G$ taking K distinct values $\{\theta_k\}_{k=1}^K$. The predictive distribution of the next observation then equals*

$$p(\bar{\theta}_{N+1} = \theta \mid \bar{\theta}_1, \ldots, \bar{\theta}_N, \alpha, H) = \frac{1}{\alpha + N}\left(\alpha h(\theta) + \sum_{k=1}^{K} N_k \delta(\theta, \theta_k)\right) \qquad (2.180)$$

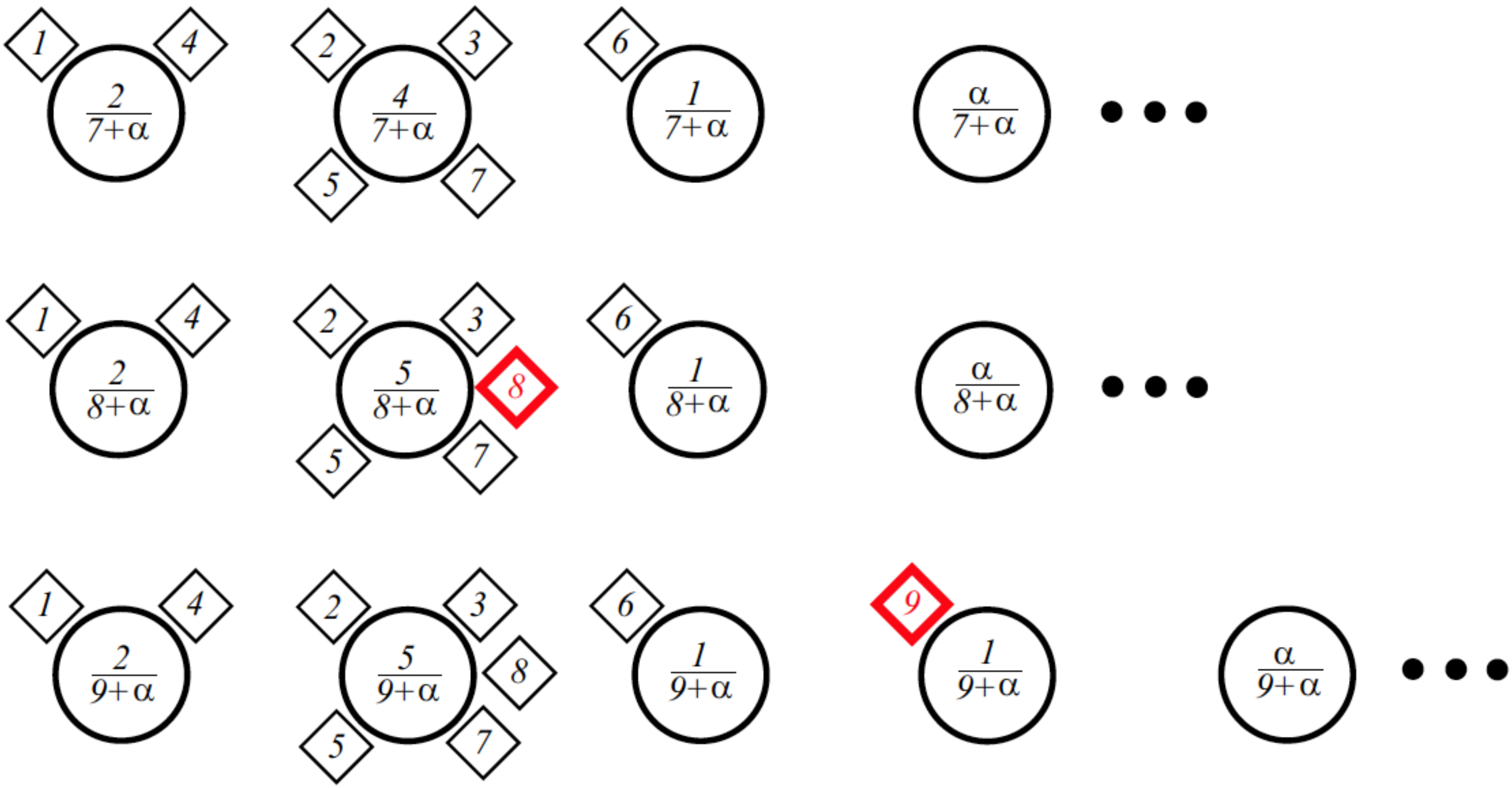
where N_k is the number of previous observations of θ_k, as in eq. (2.179).

Proof Hint: *Posterior mean after N observations equals*

$$\mathbb{E}\left[G(T) \mid \bar{\theta}_1, \ldots, \bar{\theta}_N, \alpha, H\right] = \frac{1}{\alpha + N}\left(\alpha H(T) + \sum_{k=1}^{K} N_k \delta_{\theta_k}(T)\right)$$

$$N_k \triangleq \sum_{i=1}^{N} \delta(\bar{\theta}_i, \theta_k) \qquad k = 1, \ldots, K$$

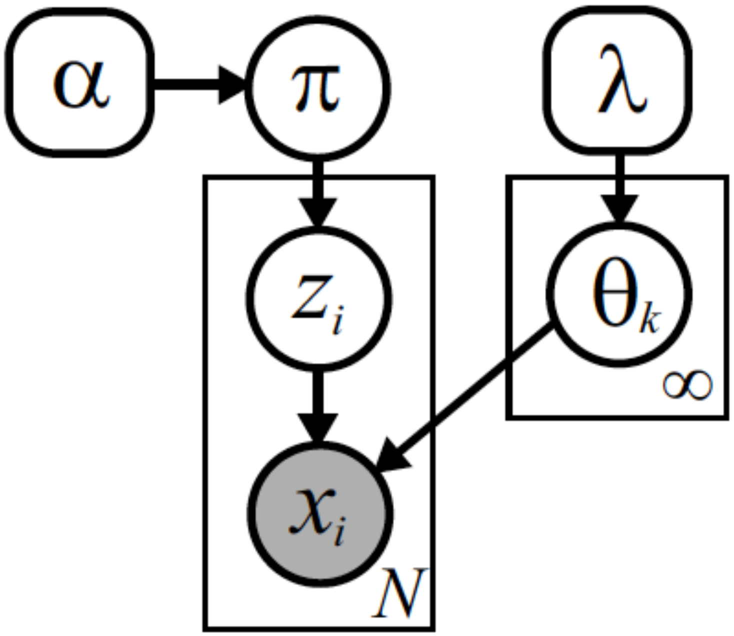
Chinese Restaurant Process

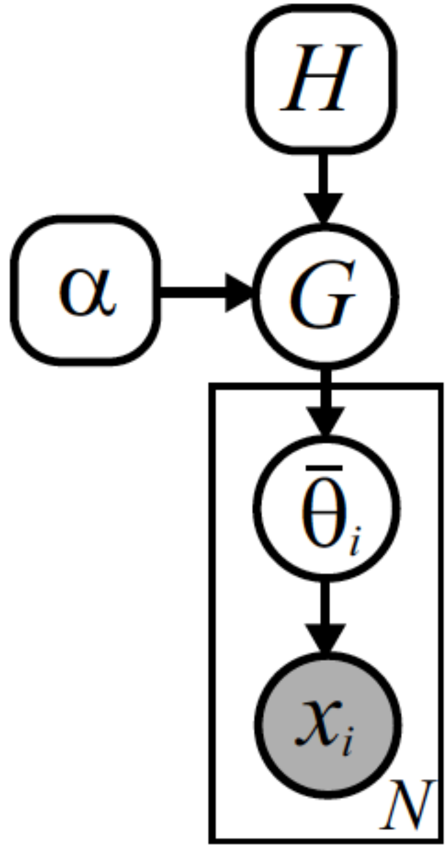


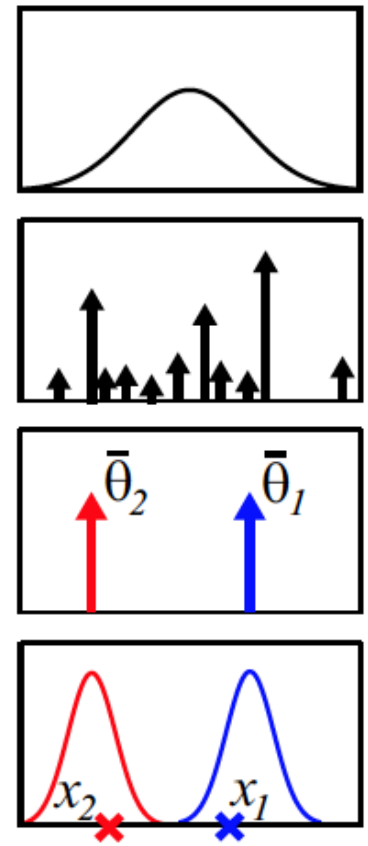
$$p(z_{N+1} = z \mid z_1, \ldots, z_N, \alpha) = \frac{1}{\alpha + N}\left(\sum_{k=1}^{K} N_k \delta(z, k) + \alpha\delta(z, \bar{k})\right)$$

DP Mixture Models

$$p(x \mid \pi, \theta_1, \theta_2, \ldots) = \sum_{k=1}^{\infty} \pi_k f(x \mid \theta_k)$$







$$G(\theta) = \sum_{k=1}^{\infty} \pi_k \delta(\theta, \theta_k)$$

$$\pi \sim \mathrm{GEM}(\alpha)$$

$$\theta_k \sim H(\lambda) \qquad k = 1, 2, \ldots$$

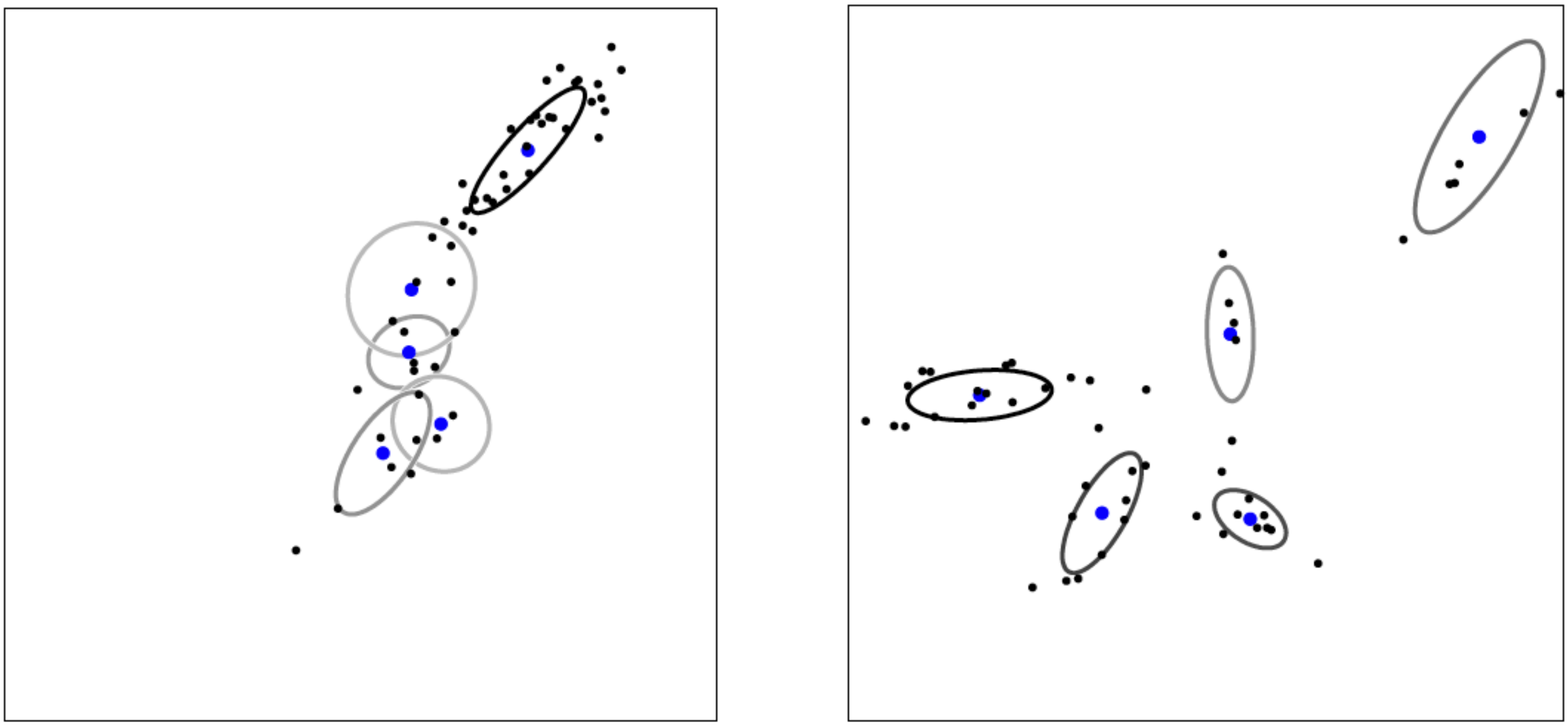
$$\bar{\theta}_i \sim G$$

$$x_i \sim F(\bar{\theta}_i)$$

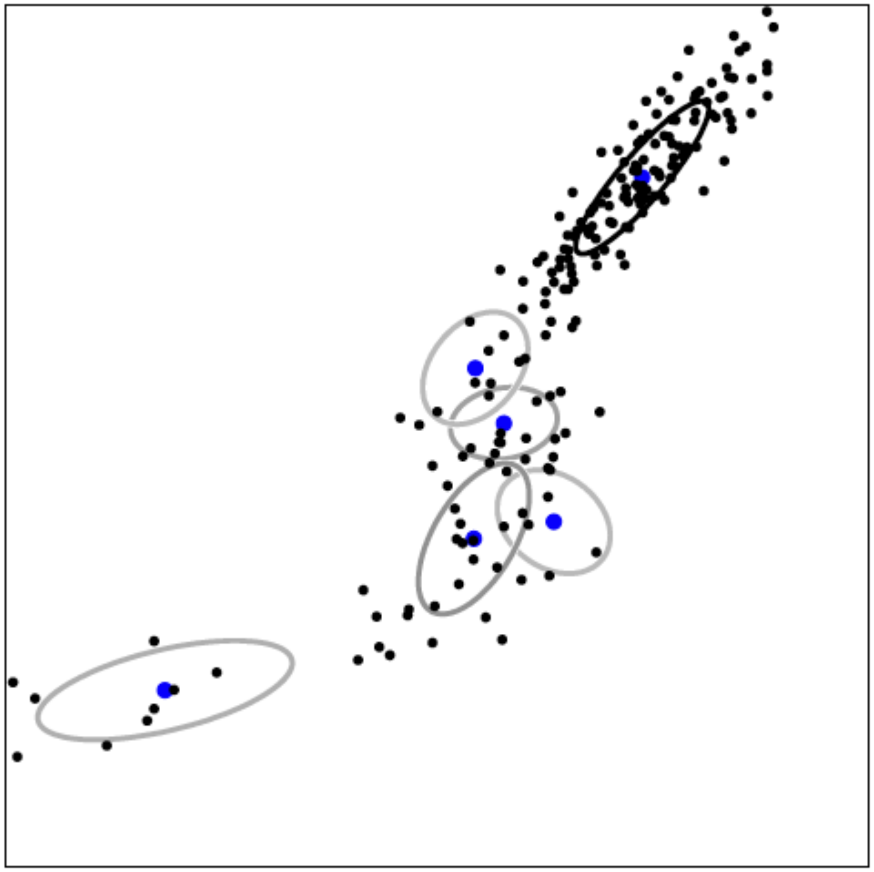
$$z_i \sim \pi$$

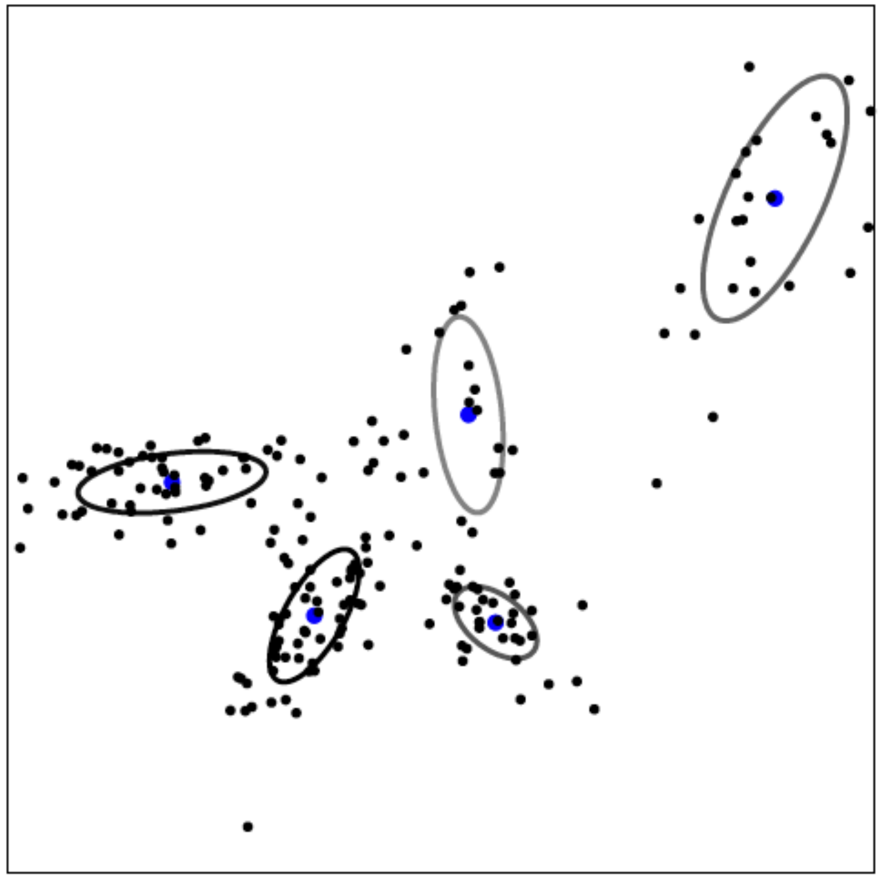
$$x_i \sim F(\theta_{z_i})$$

Samples from DP Mixture Priors

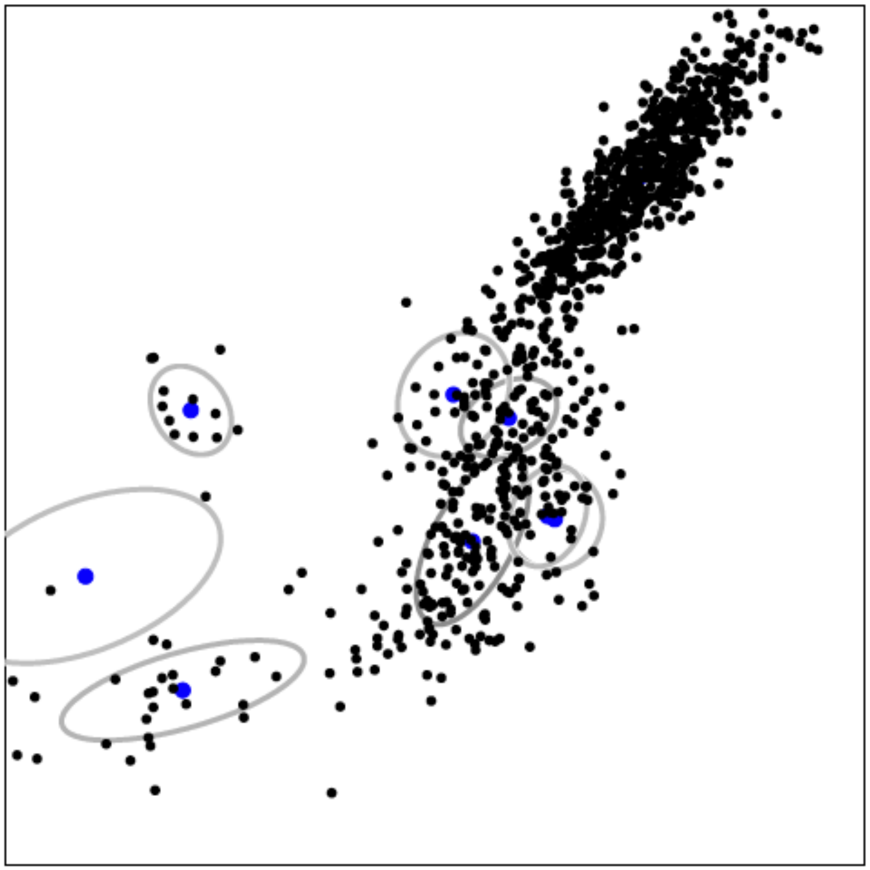


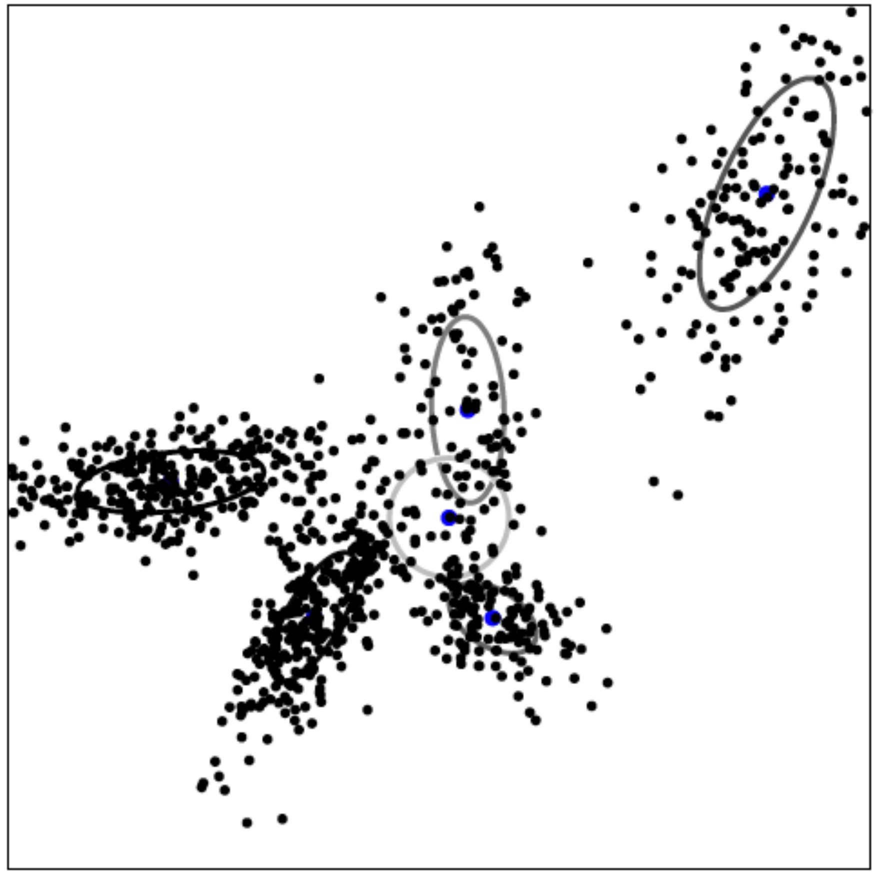
Samples from DP Mixture Priors





Samples from DP Mixture Priors





Views of the Dirichlet Process

- Implicit stochastic process: Finite Dirichlet marginals
- *Explicit stochastic process: Normalized gamma process*
- Implicit stochastic process: Neutrality
- Stick-breaking construction
- Marginalized predictions: Polya urn, or (almost) equivalently the Chinese restaurant process

Later in this course:

- Modeling: Generalize one of these representations, to get a fancier (but usually less tractable) process
- Inference: Deal with infinite-dimensional processes by analytic integration, or finite truncation (static or dynamic)