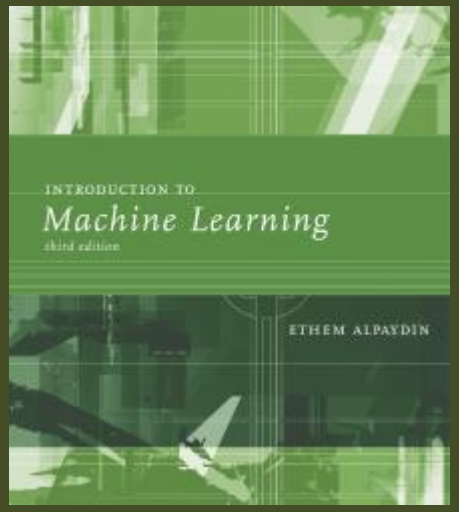


Lecture Slides for

INTRODUCTION TO MACHINE LEARNING

3RD EDITION

ETHEM ALPAYDIN

alpaydin@boun.edu.tr
http://www.cmpe.boun.edu.tr/~ethem/i2ml3e

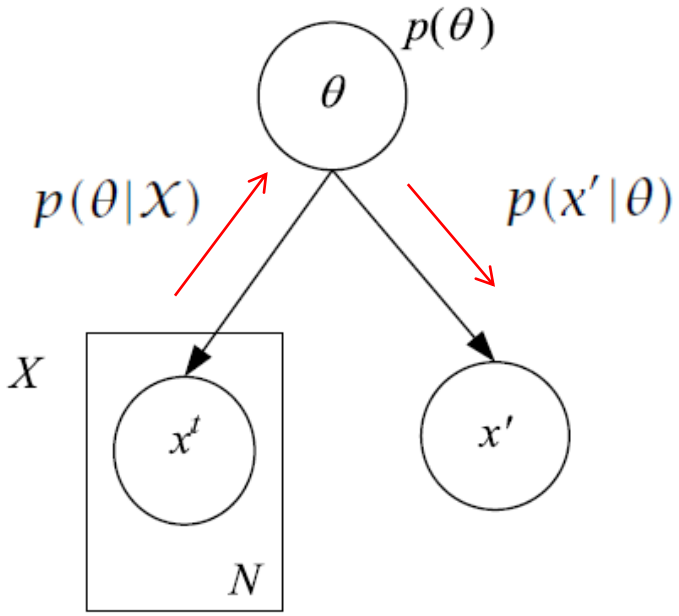
CHAPTER 16:

BAYESIAN ESTIMATION

Rationale

- Parameters θ not constant, but random variables with a prior, $p(\theta)$
- Bayes' Rule: $$p(\theta \mid \mathrm{X}) = \frac{p(\theta)p(\mathrm{X} \mid \theta)}{p(\mathrm{X})}$$

Generative Model



$$p(x', \mathcal{X}, \theta) = p(\theta)p(\mathcal{X}|\theta)p(x'|\theta)$$

$$p(x'|\mathcal{X}) = \frac{p(x', \mathcal{X})}{p(\mathcal{X})} = \frac{\int p(x', \mathcal{X}, \theta)d\theta}{p(\mathcal{X})} = \frac{\int p(\theta)p(\mathcal{X}|\theta)p(x'|\theta)d\theta}{p(\mathcal{X})}$$

$$= \int p(x'|\theta)p(\theta|\mathcal{X})d\theta$$

Bayesian Approach

1. Prior $p(\theta)$ allows us to concentrate on region where θ is likely to lie, ignoring regions where it's unlikely
2. Instead of a single estimate with a single θ, we generate several estimates using several θ and average, weighted by how their probabilities

Even if prior $p(\theta)$ is uninformative, (2) still helps.

MAP estimator does not make use of (2):

$$\theta_{MAP} = \arg\max_{\theta} p(\theta|X)$$

Bayesian Approach

$$p(x'|\mathcal{X}) = \int p(x'|\theta)p(\theta|\mathcal{X})d\theta$$

- In certain cases, it is easy to integrate
- Conjugate prior: Posterior has the same density as prior
- Sampling (Markov Chain Monte Carlo): Sample from the posterior and average
- Approximation: Approximate the posterior with a model easier to integrate
  - Laplace approximation: Use a Gaussian
  - Variational approximation: Split the multivariate density into a set of simpler densities using independencies

Estimating the Parameters of a Distribution: Discrete case

- $x^t_i=1$ if in instance t is in state i, probability of state i is q_i
- Dirichlet prior, α_i are hyperparameters
- Sample likelihood

$$Dirichlet(\mathbf{q}\,|\,\boldsymbol{\alpha}) = \frac{\Gamma(\alpha_0)}{\Gamma(\alpha_1)\cdots\Gamma(\alpha_K)}\prod_{i=1}^{K} q_i^{\alpha_i-1}$$

$$p(X\,|\,\mathbf{q}) = \prod_{t=1}^{N}\prod_{i=1}^{K} q_i^{x_i^t}$$

- Posterior

$$p(\mathbf{q}\,|\,\boldsymbol{\alpha}) = \frac{\Gamma(\alpha_0+N)}{\Gamma(\alpha_1+N_1)\cdots\Gamma(\alpha_K+N_K)}\prod_{i=1}^{K} q_i^{\alpha_i+N_i-1}$$

$$= Dirichlet(\mathbf{q}\,|\,\boldsymbol{\alpha}+\mathbf{n})$$

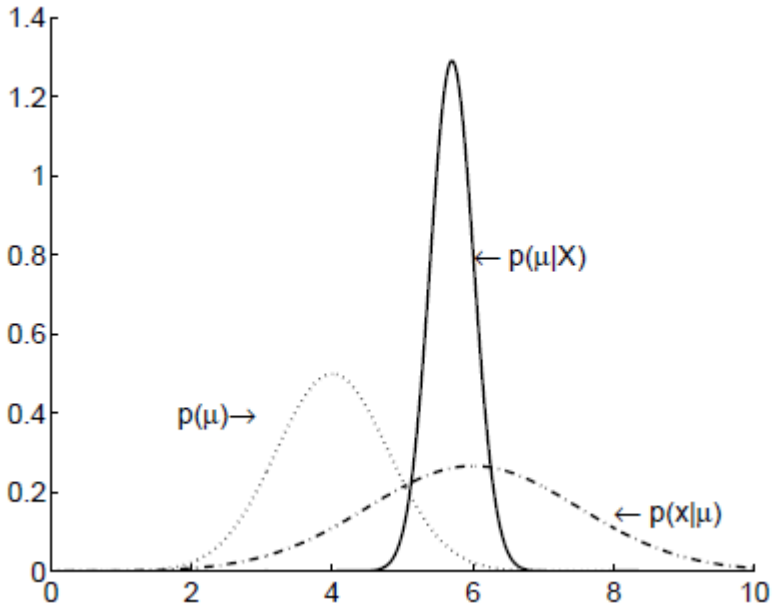
- Dirichlet is a conjugate prior
- With $K=2$, Dirichlet reduced to Beta

Estimating the Parameters of a Distribution: Continuous case

- $p(x^t) \sim N(\mu, \sigma^2)$
- Gaussian prior for μ, $p(\mu) \sim N(\mu_0, \sigma_0^2)$
- Posterior is also Gaussian $p(\mu|X) \sim N(\mu_N, \sigma_N^2)$ where

$$\mu_N = \frac{\sigma^2}{N\sigma_0^2 + \sigma^2}\mu_0 + \frac{N\sigma_0^2}{N\sigma_0^2 + \sigma^2}m$$

$$\frac{1}{\sigma^2} = \frac{1}{\sigma_0^2} + \frac{N}{\sigma^2}$$



Gaussian: Prior on Variance

- Let's define a prior (gamma) on precision $\lambda = 1/\sigma^2$

$$p(\lambda) \sim \text{gamma}(a_0, b_0) = \frac{1}{\Gamma(a_0)} b_0^{a_0} \lambda^{a_0 - 1} \exp(-b_0 \lambda)$$

$$\begin{aligned} p(\mathcal{X}|\lambda) &= \prod_t \frac{\lambda^{1/2}}{\sqrt{2\pi}} \exp\left[-\frac{\lambda}{2}(x^t - \mu)^2\right] \\ &= \lambda^{N/2}(2\pi)^{-N/2} \exp\left[-\frac{\lambda}{2}\sum_t (x^t - \mu)^2\right] \end{aligned}$$

$$\begin{aligned} p(\lambda|\mathcal{X}) &\propto p(\mathcal{X}|\lambda)p(\lambda) \\ &\sim \text{gamma}(a_N, b_N) \end{aligned}$$

$$\begin{aligned} a_N &= a_0 + N/2 = \frac{\nu_0 + N}{2} \\ b_N &= b_0 + \frac{N}{2}s^2 = \frac{\nu_0}{2}s_0^2 + \frac{N}{2}s^2 \end{aligned}$$

Joint Prior and Making a Prediction

$$p(\mu, \lambda) = p(\mu|\lambda)p(\lambda)$$

$$p(\mu, \lambda|\mathcal{X}) \sim \text{normal-gamma}(\mu_N, \kappa_N, a_N, b_N)$$

where

$$\begin{aligned}
\kappa_N &= \kappa_0 + N \\
\mu_N &= \frac{\kappa_0\mu_0 + Nm}{\kappa_N} \\
a_N &= a_0 + N/2 \\
b_N &= b_0 + \frac{N}{2}s^2 + \frac{\kappa_0 N}{2\kappa_N}(m - \mu_0)^2
\end{aligned}$$

$$\begin{aligned}
p(x|\mathcal{X}) &= \int\int p(x|\mu, \lambda)p(\mu, \lambda|\mathcal{X})d\mu d\lambda \\
&\sim t_{2a_N}\left(\mu_N, \frac{b_N(\kappa_N + 1)}{a_N\kappa_N}\right)
\end{aligned}$$

Multivariate Gaussian

$$p(\boldsymbol{x}) \sim \mathcal{N}_d(\boldsymbol{\mu}, \boldsymbol{\Lambda}) \qquad p(\boldsymbol{\mu}|\boldsymbol{\Lambda}) \sim \mathcal{N}_d(\boldsymbol{\mu}_0, (1/\kappa_0)\boldsymbol{\Lambda}) \qquad p(\boldsymbol{\Lambda}) \sim \text{Wishart}(\nu_0, \mathbf{V}_0)$$

$$\begin{aligned} p(\mu, \boldsymbol{\Lambda}) &= p(\boldsymbol{\mu}|\boldsymbol{\Lambda})p(\boldsymbol{\Lambda}) \\ &\sim \text{normal-Wishart}(\boldsymbol{\mu}_0, \kappa_0, \nu_0, \mathbf{V}_0) \end{aligned}$$

$$p(\mu, \boldsymbol{\Lambda}|\mathcal{X}) \sim \text{normal-Wishart}(\boldsymbol{\mu}_N, \kappa_N, \nu_N, \mathbf{V}_N)$$

$$\begin{aligned} \kappa_N &= \kappa_0 + N \\ \boldsymbol{\mu}_N &= \frac{\kappa_0 \boldsymbol{\mu}_0 + N\boldsymbol{m}}{\kappa_N} \\ \nu_N &= \nu_0 + N \\ \mathbf{V}_N &= \left(\mathbf{V}_0^{-1} + \mathbf{C} + \frac{\kappa_0 N}{\kappa_N}(\boldsymbol{m} - \boldsymbol{\mu}_0)(\boldsymbol{m} - \boldsymbol{\mu}_0)^T\right)^{-1} \end{aligned}$$

$$\begin{aligned} p(\boldsymbol{x}|\mathcal{X}) &= \int\int p(\boldsymbol{x}|\boldsymbol{\mu}, \boldsymbol{\Lambda}) p(\boldsymbol{\mu}, \boldsymbol{\Lambda}|\mathcal{X}) d\boldsymbol{\mu} d\boldsymbol{\Lambda} \\ &\sim t_{\nu_N - d + 1}\left(\boldsymbol{\mu}_N, \frac{\kappa_N + 1}{\kappa_N(\nu_N - d + 1)}(\mathbf{V}_N)^{-1}\right) \end{aligned}$$

Estimating the Parameters of a Function: Regression

- $r=\mathbf{w}^T\mathbf{x}+\varepsilon$, $p(\varepsilon)\sim N(0,1/\beta)$, and $p(r^t \mid x^t,\mathbf{w},\beta)\sim N(\mathbf{w}^T\mathbf{x}^t, 1/\beta)$
- Log likelihood

$$L(\mathbf{r}\mid\mathbf{X},\mathbf{w},\beta)=\log\prod_t p(r^t\mid\mathbf{x}^t,\mathbf{w},\beta)$$

$$=-N\log\left(\sqrt{2\pi}\right)+N\log\beta-\frac{\beta}{2}\sum_t\left(r^t-\mathbf{w}^T\mathbf{x}^t\right)$$

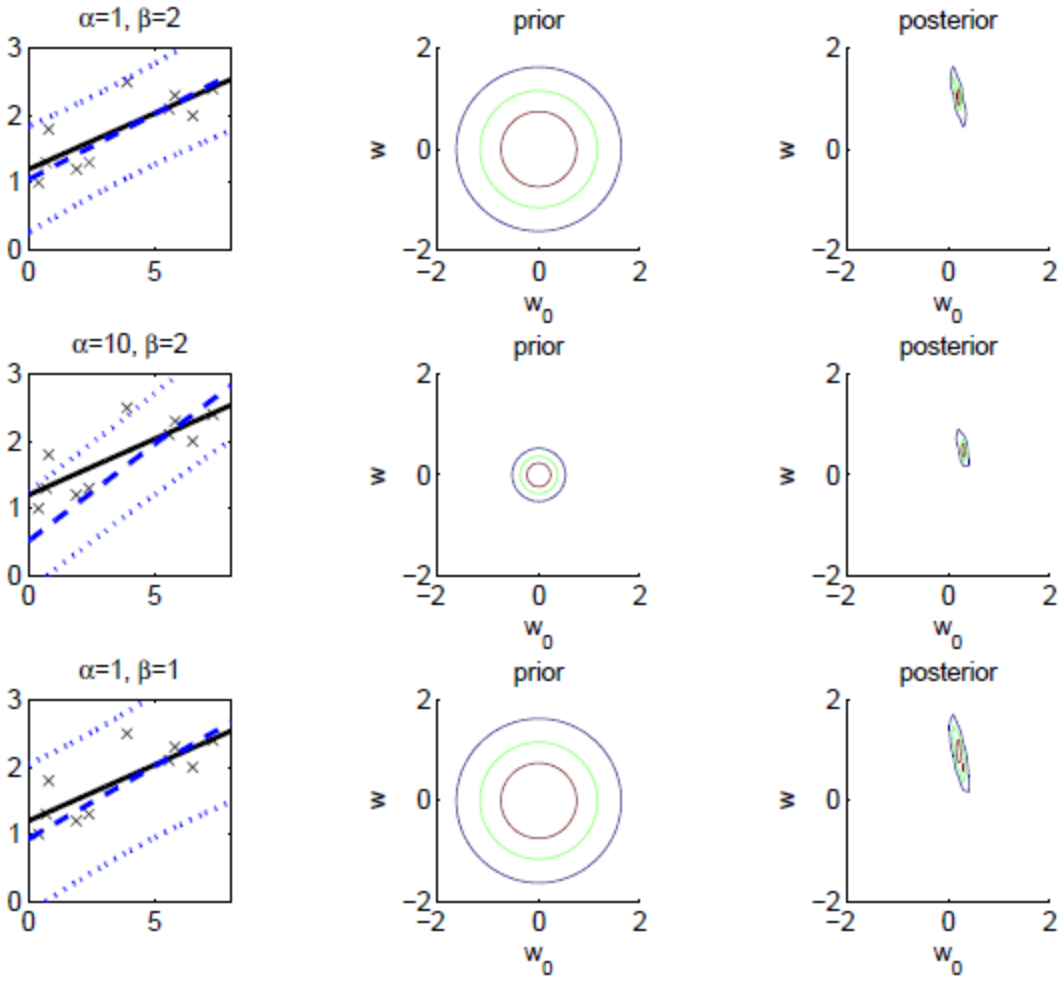
ML solution $\mathbf{w}_{ML}=(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{r}$

- Gaussian conjugate prior: $p(\mathbf{w})\sim N(0,1/\alpha)$
- Posterior: $p(\mathbf{w}\mid\mathbf{X})\sim N(\boldsymbol{\mu}_N,\Sigma_N)$ where

$$\boldsymbol{\mu}_N=\beta\boldsymbol{\Sigma}_N\mathbf{X}^T\mathbf{r}$$

$$\boldsymbol{\Sigma}_N=(\alpha\mathbf{I}+\beta\mathbf{X}^T\mathbf{X})^{-1}$$

Aka ridge regression/parameter shrinkage/ L2 regularization/weight decay


α=1, β=2
prior
posterior
α=10, β=2
prior
posterior
α=1, β=1
prior
posterior
w
w0

Prior on Noise Variance

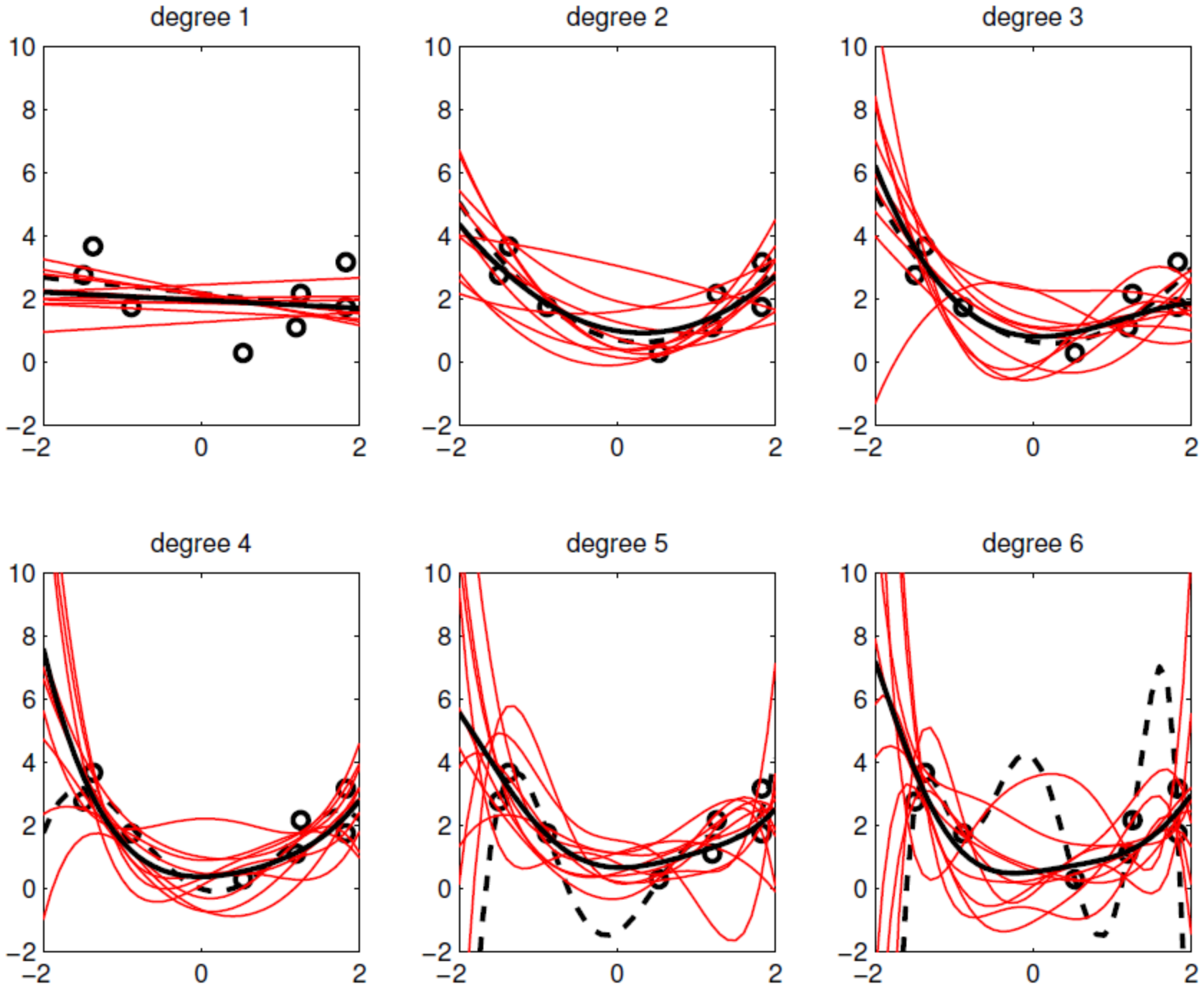
$$p(\beta) \sim \text{gamma}(a_0, b_0) \qquad p(\boldsymbol{w}|\beta) \sim \mathcal{N}(\boldsymbol{\mu}_0, \beta\Sigma_0)$$

$$p(\boldsymbol{w}, \beta) = p(\beta)p(\boldsymbol{w}|\beta) \sim \text{normal-gamma}(\boldsymbol{\mu}_0, \Sigma_0, a_0, b_0)$$

$$p(\boldsymbol{w}, \beta|\mathbf{X}, \boldsymbol{r}) \sim \text{normal-gamma}(\boldsymbol{\mu}_N, \Sigma_N, a_N, b_N)$$

$$\begin{aligned}
\Sigma_N &= (\mathbf{X}^T\mathbf{X} + \Sigma_0)^{-1} \\
\boldsymbol{\mu}_N &= \Sigma_N(\mathbf{X}^T\boldsymbol{r} + \Sigma_0\boldsymbol{\mu}_0) \\
a_N &= a_0 + N/2 \\
b_N &= b_0 + \frac{1}{2}(\boldsymbol{r}^T\boldsymbol{r} + \boldsymbol{\mu}_0^T\Sigma_0\boldsymbol{\mu}_0 - \boldsymbol{\mu}_N^T\Sigma_N\boldsymbol{\mu}_N)
\end{aligned}$$

Markov Chain Monte Carlo (MCMC) sampling


degree 1
degree 2
degree 3
degree 4
degree 5
degree 6

Basis/Kernel Functions

- For new $\mathbf{x}'$, the estimate r' is calculated as

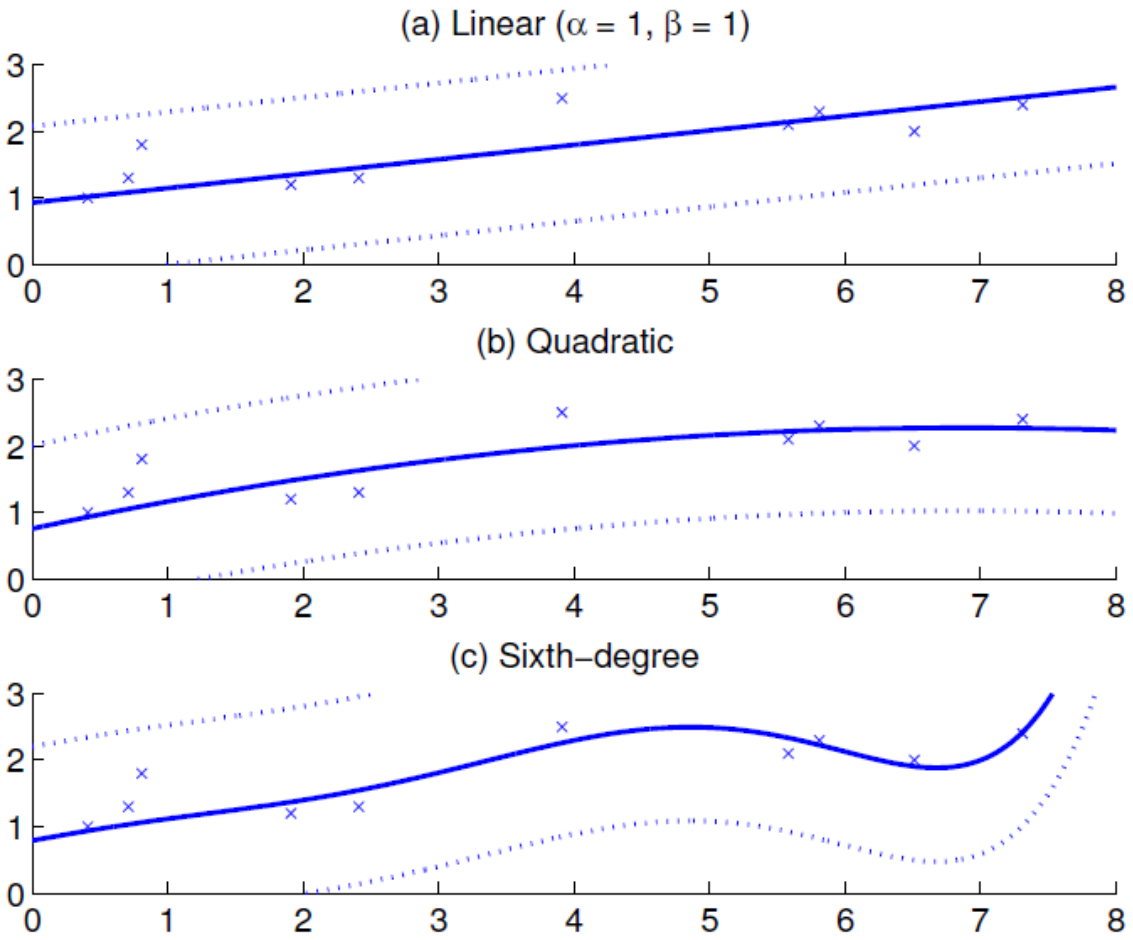
$$r' = (\mathbf{x}')^T \beta \boldsymbol{\Sigma}_N \mathbf{X}^T \mathbf{r} = \sum_t \beta (\mathbf{x}')^T \boldsymbol{\Sigma}_N \mathbf{x}^t r^t$$

Dual representation

- Linear kernel
- For any other $\phi(\mathbf{x})$, we can write $K(\mathbf{x}',\mathbf{x}) = \phi(\mathbf{x}')^T \phi(\mathbf{x})$

$$r' = \sum_t \beta (\mathbf{x}')^T \boldsymbol{\Sigma}_N \mathbf{x}^t r^t \sum_t \beta K(\mathbf{x}', \mathbf{x}^t) r^t$$

Kernel Functions



What's in a Prior?

- Defining a prior is subjective
- Uninformative prior if no prior preference
- How high to go?

$$\text{Level I: } p(x|\mathcal{X}) = \int p(x|\theta)p(\theta|\mathcal{X})d\theta$$

$$\text{Level II: } p(x|\mathcal{X}) = \int p(x|\theta)p(\theta|\mathcal{X},\alpha)p(\alpha)d\theta d\alpha$$

- Empirical Bayes: Use one good α^*

$$\text{Level II ML: } p(x|\mathcal{X}) = \int p(x|\theta)p(\theta|\mathcal{X},\alpha^*)d\theta$$

Bayesian Model Comparison

- Marginal likelihood of a model:

$$p(\mathcal{X}|\mathcal{M}) = \int p(\mathcal{X}|\theta, \mathcal{M}) p(\theta|\mathcal{M}) d\theta$$

- Posterior probability of model given data:

$$p(\mathcal{M}|\mathcal{X}) = \frac{p(\mathcal{X}|\mathcal{M}) p(\mathcal{M})}{p(\mathcal{X})}$$

- Bayes' factor:

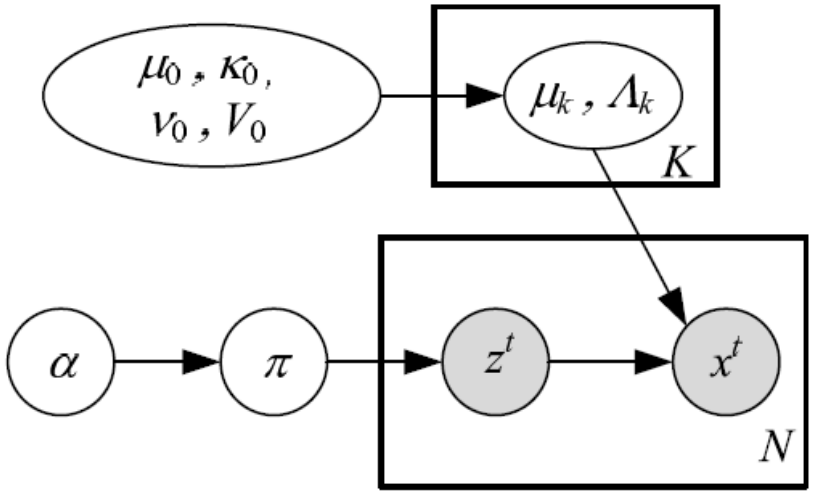
$$\frac{P(\mathcal{M}_1|\mathcal{X})}{P(\mathcal{M}_0|\mathcal{X})} = \frac{P(\mathcal{X}|\mathcal{M}_1)}{P(\mathcal{X}|\mathcal{M}_0)} \frac{P(\mathcal{M}_1)}{P(\mathcal{M}_0)}$$

- Approximations:

BIC: $\log p(\mathcal{X}|\mathcal{M}) \approx \text{BIC} \equiv \log p(\mathcal{X}|\theta_{ML}, \mathcal{M}) - \frac{|\mathcal{M}|}{2} \log N$

AIC: $\text{AIC} \equiv \log p(\mathcal{X}|\theta_{ML}, \mathcal{M}) - |\mathcal{M}|$

Mixture Model



$$p(\boldsymbol{x}) = \sum_{i=1}^{k} P(\mathcal{G}_i) p(\boldsymbol{x}|\mathcal{G}_i)$$

$$p(\Phi) = p(\boldsymbol{\pi}) \prod_i p(\boldsymbol{\mu}_i, \boldsymbol{\Lambda}_i)$$

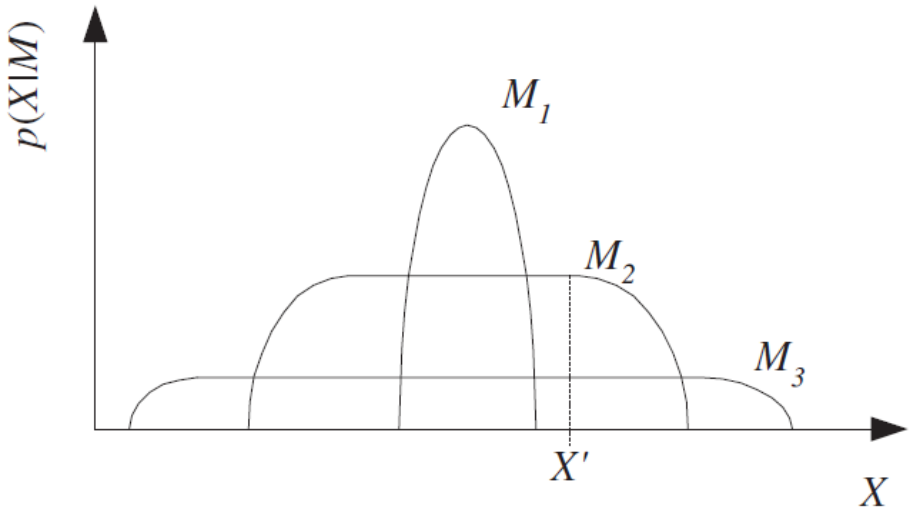
$$= \text{Dirichlet}(\boldsymbol{\pi}|\boldsymbol{\alpha}) \prod_i \text{normal-Wishart}(\boldsymbol{\mu}_0, \kappa_0, \nu_0, \mathbf{V}_0)$$

$$\mathcal{Q}(\Phi|\Phi^l) = \sum_t \sum_i h_i^t \log \pi_i + \sum_t \sum_i h_i^t \log p_i(\boldsymbol{x}^t|\Phi^l) + \log p(\boldsymbol{\pi}) + \sum_i \log p(\boldsymbol{\mu}_i, \boldsymbol{\Lambda}_i)$$

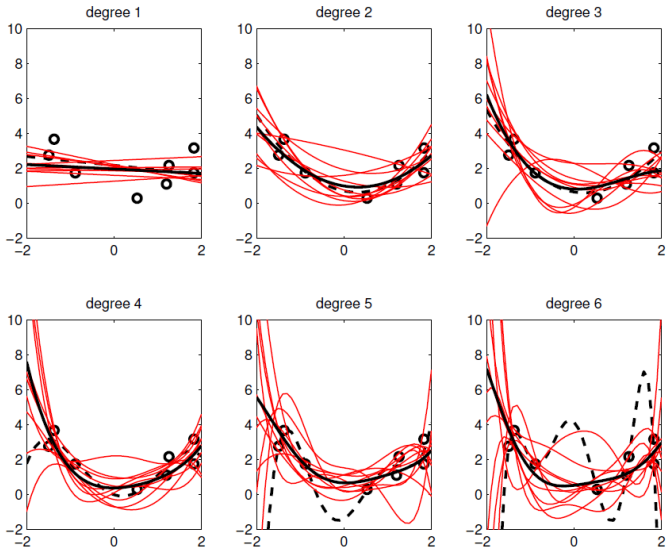
$$\pi_i^{l+1} = \frac{\alpha_i + N_i - 1}{\sum_i \alpha_i + N - k}$$

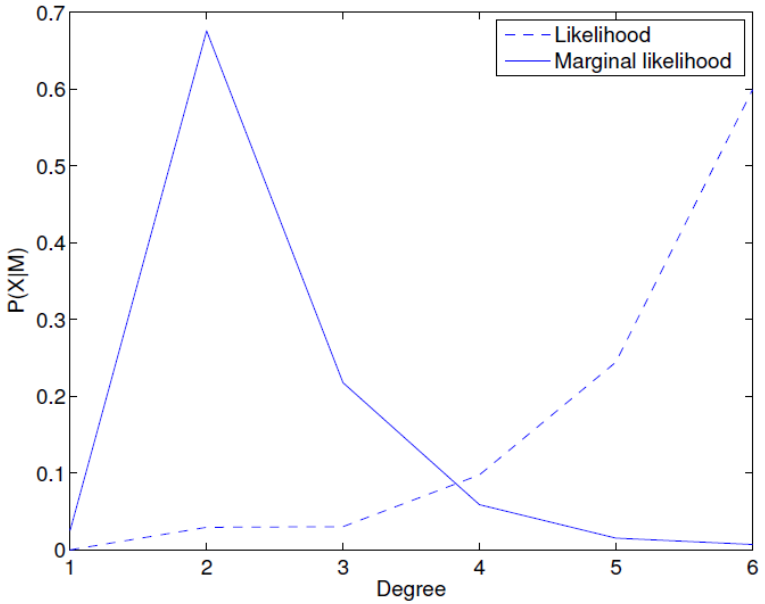
$$\boldsymbol{\mu}_i^{l+1} = \frac{\kappa_0 \boldsymbol{\mu}_0 + N_i \boldsymbol{m}_i}{\kappa_0 + N_i}$$

$$\boldsymbol{\Lambda}_i^{l+1} = \left(\frac{\mathbf{V}_0^{-1} + \mathbf{C}_i + \mathbf{S}_i}{\nu_0 + N_i + d + 2} \right)^{-1}$$



Models in increasing complexity. A complex model can fit more datasets but is spread thin, a simple model can fit few datasets but has higher marginal likelihood where it does (MacKay 2003)





Nonparametric Bayes

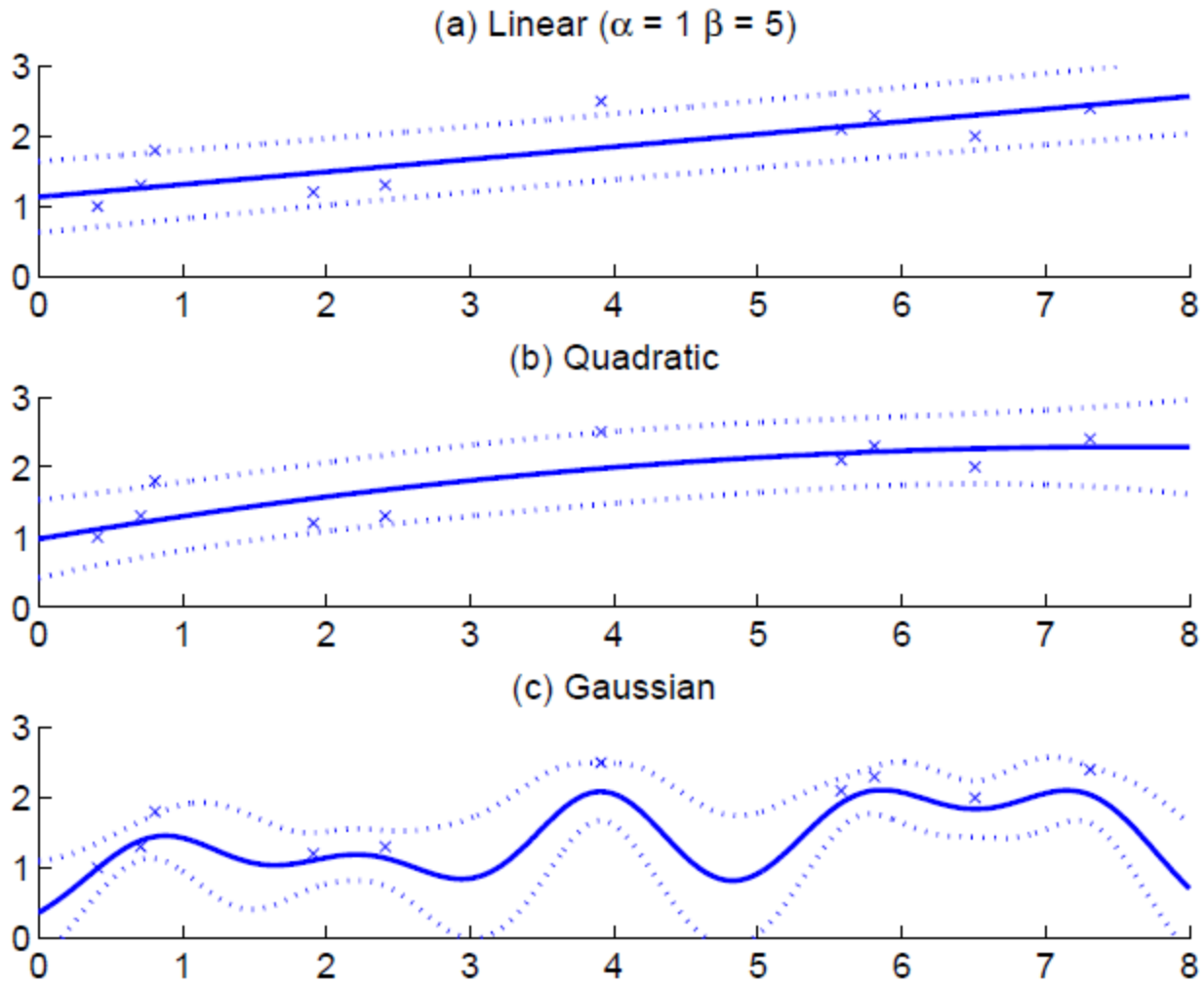
- Model complexity can increase with more data (in practice up to N, potentially to infinity)
- Similar to k-NN and Parzen windows we saw before where training set is the parameters

Gaussian Processes

- Nonparametric model for supervised learning
- Assume Gaussian prior $p(\mathbf{w})\sim N(0,1/\alpha)$
  $\mathbf{y}=\mathbf{Xw}$, where $E[\mathbf{y}]=0$ and $Cov(\mathbf{y})=\mathbf{K}$ with $\mathbf{K}_{ij}=(\mathbf{x}^i)^T\mathbf{x}^i$
  $\mathbf{K}$ is the covariance function, here linear
- With basis function $\phi(\mathbf{x})$, $\mathbf{K}_{ij}=(\phi(\mathbf{x}^i))^T\phi(\mathbf{x}^i)$
  $r\sim N_N(\mathbf{0},C_N)$ where $C_N=(1/\beta)\mathbf{I}+\mathbf{K}$
- With new $\mathbf{x}'$ added as $\mathbf{x}_{N+1}$, $r_{N+1}\sim N_{N+1}(0,C_{N+1})$

$$\mathbf{C}_{N+1}=\begin{bmatrix}\mathbf{C}_N & \mathbf{k}\\ \mathbf{k} & c\end{bmatrix}$$

where $\mathbf{k}=[K(\mathbf{x}',\mathbf{x}^t)_t]^T$ and $c=K(\mathbf{x}',\mathbf{x}')+1/\beta$.
$p(r'|\mathbf{x}',\mathbf{X},\mathbf{r})\sim N(\mathbf{k}^T\mathbf{C}_{N-1}\mathbf{r},c-\mathbf{k}^T\mathbf{C}_{N-1}\mathbf{k})$


(a) Linear (α = 1 β = 5)
(b) Quadratic
(c) Gaussian
0
1
2
3
4
5
6
7
8

Dirichlet Processes

- Nonparametric Bayesian approach for clustering
- Chinese restaurant process
- Customers arrive and either join one of the existing tables or start a new one, based on the table occupancies:

$$\text{Join existing table } i \text{ with } P(z_i = 1) = \frac{n_i}{\alpha + n - 1}, i = 1, \ldots, k$$

$$\text{Start new table with } P(z_{k+1} = 1) = \frac{\alpha}{\alpha + n - 1}$$

Nonparametric Gaussian Mixture

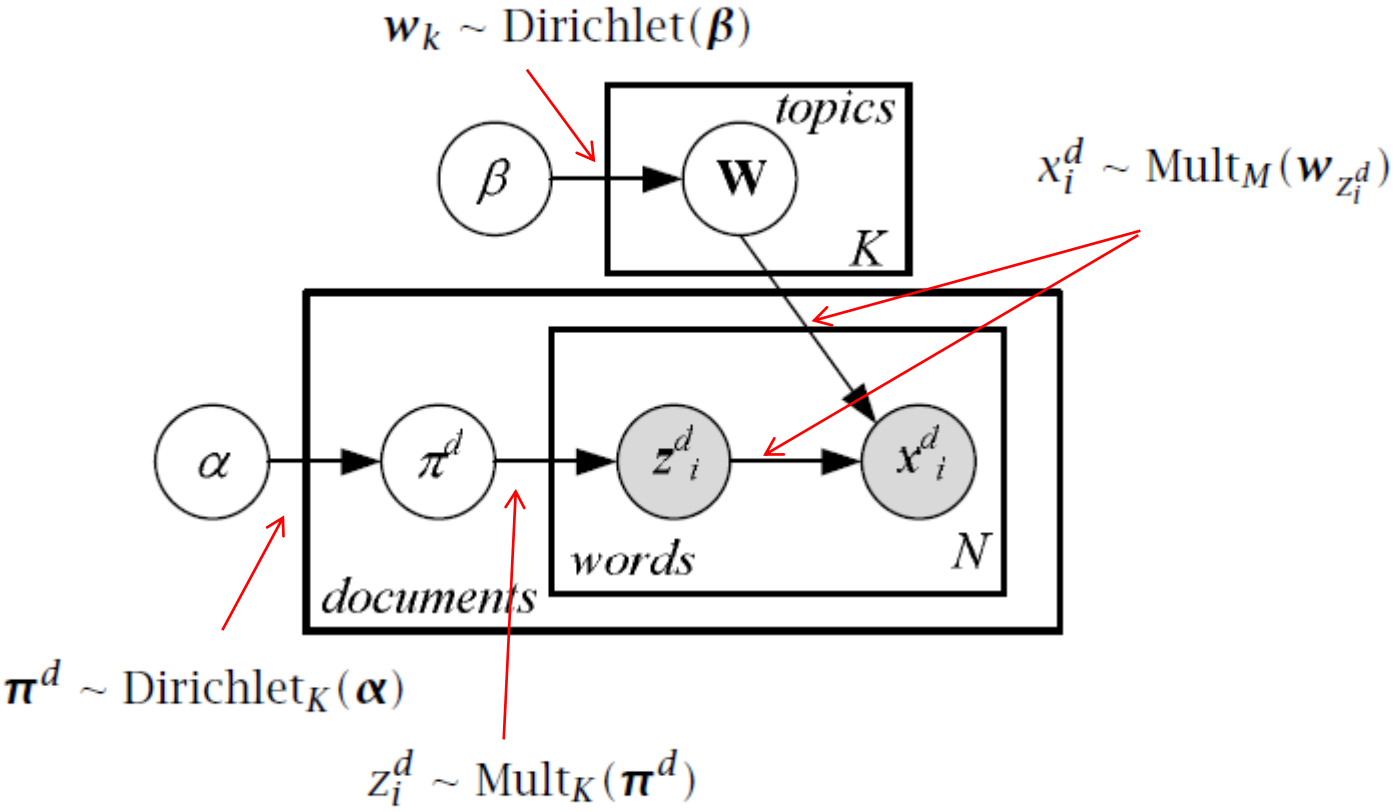
- Tables are Gaussian components and decisions based both on prior and also on input x:

$$\text{Join component } i \text{ with } P(z_i^t = 1) \quad \propto \quad \frac{n_i}{\alpha + n - 1} p(\mathbf{x}^t | \mathcal{X}_i), i = 1, \ldots, k$$

$$\text{Start new component with } P(z_{k+1}^t) \quad \propto \quad \frac{\alpha}{\alpha + n - 1} p(\mathbf{x}^t)$$

Latent Dirichlet Allocation

- Bayesian feature extraction



Beta Processes

- Nonparametric Bayesian approach for feature extraction
- Matrix factorization:

$$\mathbf{X} = \mathbf{ZA} \qquad z_j^t = \begin{cases} 1 & \text{with probability } \mu_j \\ 0 & \text{with probability } 1 - \mu_j \end{cases}$$

$$\mu_j \sim \text{beta}(\alpha, 1)$$

- Nonparametric version: Allow *j* to increase with more data probabilistically
- Indian buffet process: Customer can take one of the existing dishes with prob μ_j or add a new dish to the buffet