Bayesian Learning

Outline

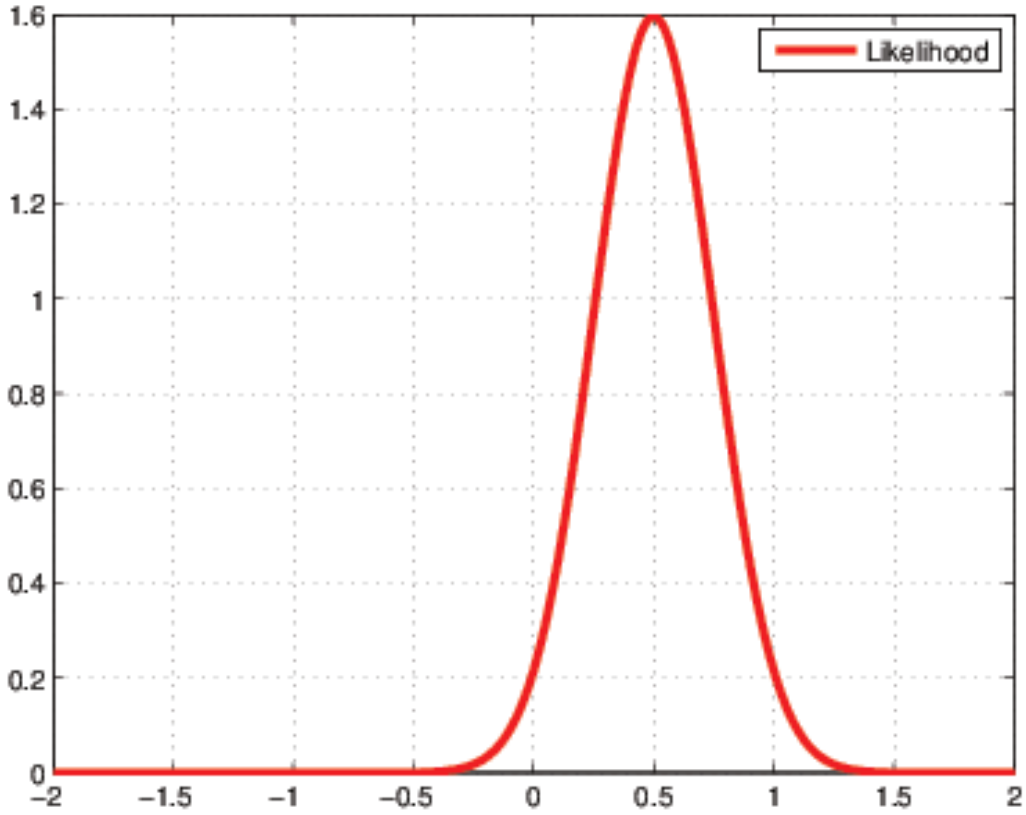
- MLE, MAP vs. Bayesian Learning
- Bayesian Linear Regression
- Bayesian Gaussian Mixture Models
  - Non-parametric Bayes

Take Away ...

1. Maximum Likelihood Estimate (MLE)
   - $\theta^* = \arg\max_\theta p(\mathcal{D}|\theta)$
   - Use θ^* in future to predict y_{n+1} given $\mathbf{x}_{n+1}$
2. Maximum a posteriori estimation (MAP)
   - $\theta^* = \arg\max_\theta p(\theta|\mathcal{D}, \alpha) = \arg\max_\theta p(\mathcal{D}|\theta)p(\theta|\alpha)$
     - α is called Hyperparameter
   - Use θ^* in future to predict y_{n+1} given $\mathbf{x}_{n+1}$
3. Bayesian treatment
   - model $p(\theta|\mathcal{D}, \alpha)$
   - $p(y_{n+1}|\mathbf{x}_{n+1}, \mathcal{D}, \alpha) = \int_\theta p(y_{n+1}|\theta, \mathbf{x_{n+1}})\mathbf{p}(\theta|\mathcal{D}, \alpha)\mathbf{d}\theta$

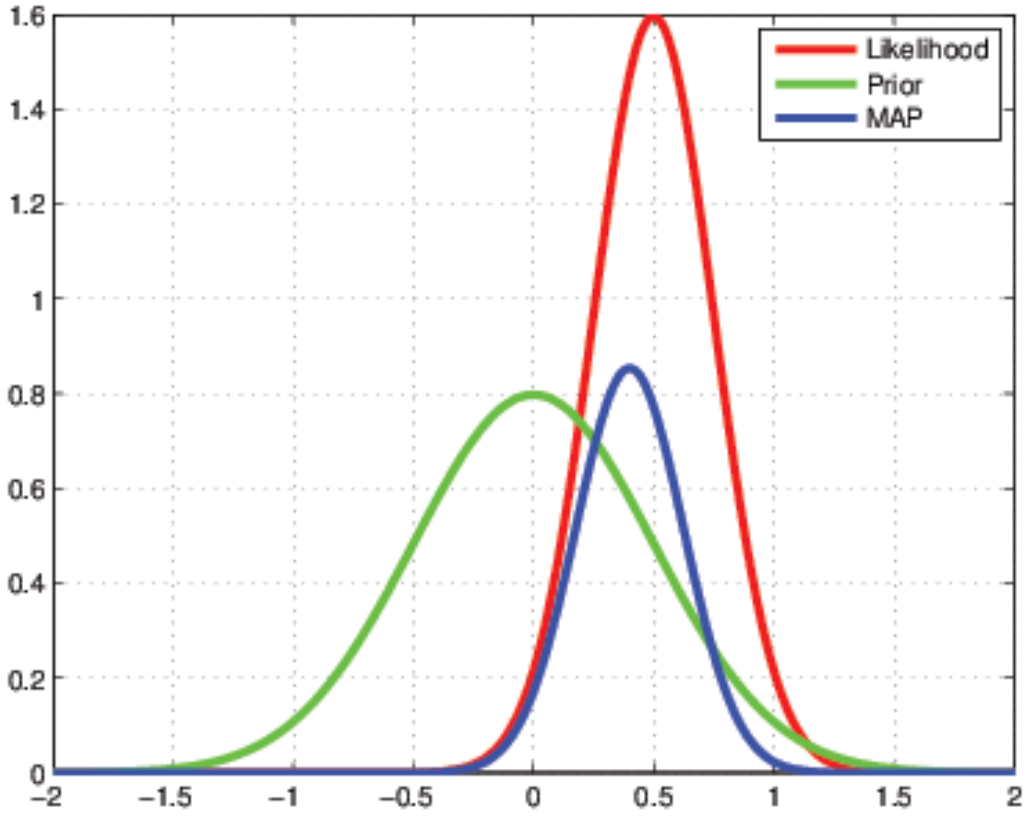
MLE Estimate

$$\theta^* = \arg\max_\theta p(\mathcal{D}|\theta)$$



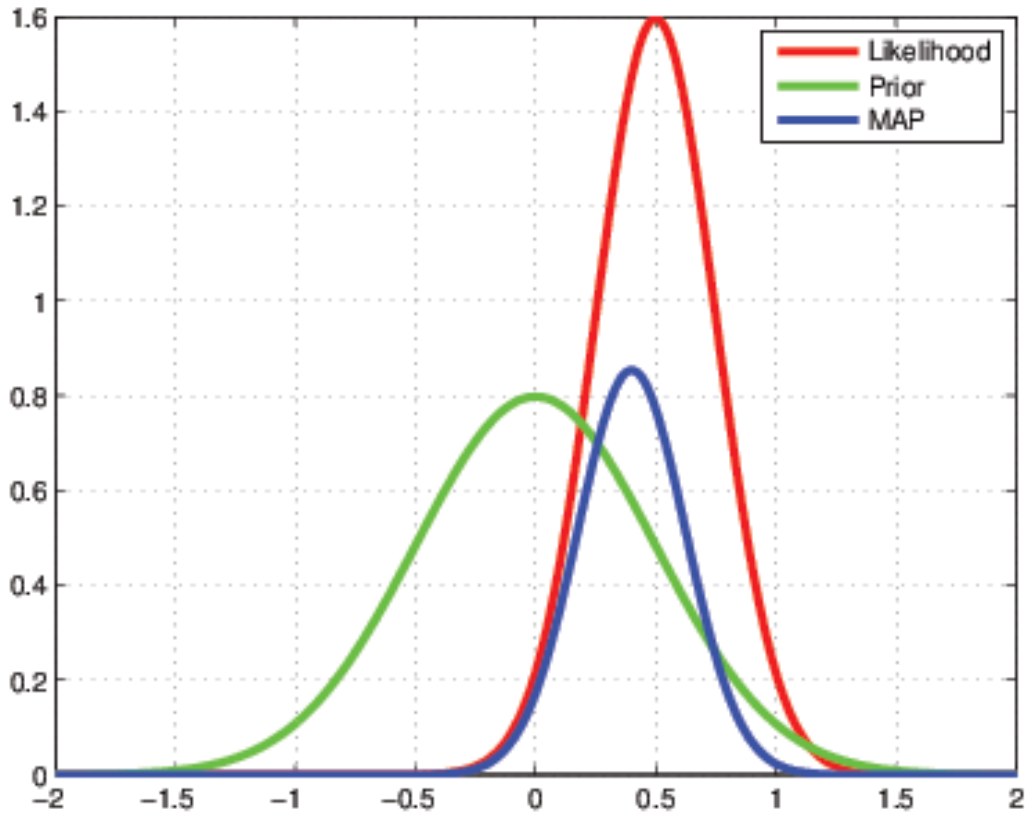
MAP Estimate

$$\theta^* = \arg\max_\theta p(\mathcal{D}|\theta)p(\theta)$$



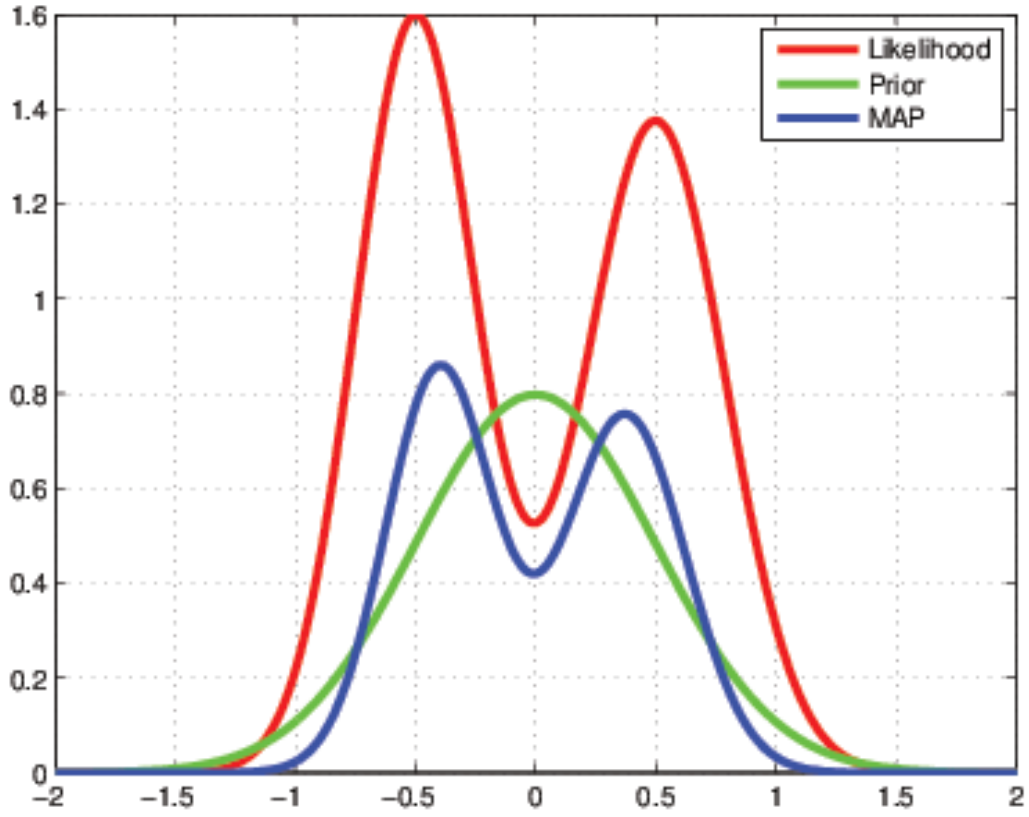
Bayesian Learning

$$p(y_{n+1}|\mathbf{x}_{n+1}, \mathcal{D}) = \int_\theta p(y_{n+1}|\theta, \mathbf{x_{n+1}})\mathbf{p}(\theta|\mathcal{D})\mathbf{d}\theta$$



Bayesian Learning

$$p(y_{n+1}|\mathbf{x}_{n+1}, \mathcal{D}) = \int_\theta p(y_{n+1}|\theta, \mathbf{x_{n+1}})\mathbf{p}(\theta|\mathcal{D})\mathbf{d}\theta$$



Linear Regression

- $\mathcal{D} = \{(\mathbf{x}_i, y_i)\}\ i = 1 \cdots N$
- Assume that $y = f(\mathbf{x}, \mathbf{w}) + \epsilon$
  - $\epsilon \sim \mathcal{N}(0, \beta^{-1})$
- Linear models assume that
  - $f(\mathbf{x}, \mathbf{w}) = \langle \mathbf{w}, \mathbf{x} \rangle = \mathbf{w}^T \mathbf{x}$
- The aim is to find the appropriate weight vector $\mathbf{w}$

Maximum Likelihood Estimate (MLE)

1. Write the Likelihood
   - $p(y|\mathbf{x}, \mathbf{w}, \beta) = \mathcal{N}(y|f(\mathbf{x}, \mathbf{w}), \beta^{-1}) = \mathcal{N}(y|\mathbf{w}^T\mathbf{x}, \beta^{-1})$

$$
\begin{aligned}
\mathcal{L}(\mathbf{w}) &= p(y_1..y_N|\mathbf{x}_1..\mathbf{x}_N, \mathbf{w}, \beta) \\
&= \prod_{i=1}^{N} \mathcal{N}(y_i|\mathbf{w}^T\mathbf{x}_i, \beta^{-1}) \\
&= \prod_{i} \frac{\sqrt{\beta}}{2\pi} \exp\Big(-\frac{\beta}{2}(y_i - \mathbf{w}^T\mathbf{x})^2\Big) \\
\mathbf{w}^* &= \arg\min_{\mathbf{w}} \sum_{i} \Big(y_i - \mathbf{w}^T\mathbf{x}\Big)^2 \qquad (1)
\end{aligned}
$$

2. Solve for $\mathbf{w}^*$ and use it for future predictions.

MAP Estimate

1. Introduce Priors on the parameters
   - What are the parameters in this model ?
   - Conjugate Priors
     - Prior and Posterior have same form.
     - Beta is conjugate to Bernoulli dist.
     - Normal with known variance is conjugate to Normal dist.
   - Hyperparameter
     - The parameters of the prior distribution
2. Model the posterior distribution – $p(\theta|\mathcal{D}, \alpha)$

$$\theta^* = \arg\max_{\theta} p(\theta|\mathcal{D}, \alpha) = \arg\max_{\theta} p(\mathcal{D}|\theta)p(\theta|\alpha)$$

MAP Estimate

For Linear Regression, $p(y|\mathbf{x}, \mathbf{w}, \beta) = \mathcal{N}(y|\mathbf{w}^T\mathbf{x}, \beta^{-1})$

1. Introduce Prior distribution
   - Identify the Parameters
   - We put a Gaussian prior on $\mathbf{w}$

$$p(\mathbf{w}|\alpha) = \mathcal{N}(\mathbf{w}|\mathbf{0}, \alpha^{-1}\mathbf{I})$$

2. Model Posterior distribution
   - $p(\mathbf{w}|\mathbf{y}, X, \alpha) \propto p(\mathbf{y}|\mathbf{w}, X)\ p(\mathbf{w}|\alpha)$
     - Likelihood $\mathcal{L}(\mathbf{w}) = p(\mathbf{y}|\mathbf{w}, X)$ is :

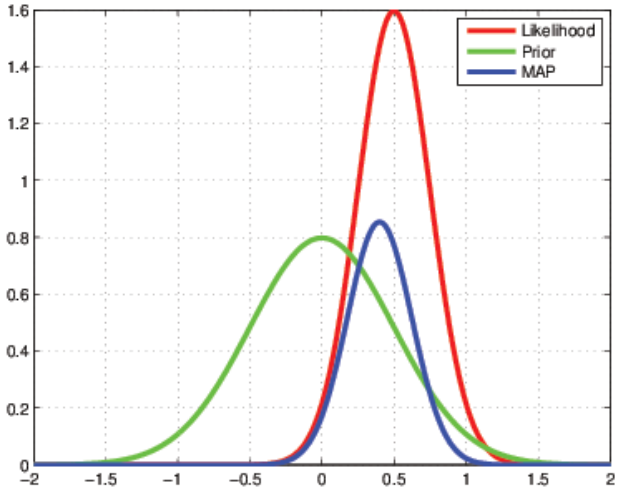
$$\prod_i p(y_i|\mathbf{x}_i, \mathbf{w}, \beta) = \prod_i \mathcal{N}(y_i|\mathbf{w}^T\mathbf{x}_i, \beta^{-1})$$

MAP Estimate

With the above choice of prior,

$$p(\mathbf{w}|\mathbf{y}, X, \alpha, \beta) = \mathcal{N}(\mathbf{w}|\mu_N, \Sigma_N)$$

- $\Sigma_N = \alpha\mathbf{I} + \beta X^T X$
- $\mu_N = \beta\Sigma_N^{-1} X^T \mathbf{y}$



Since this is Gaussian, mode is same as the mean.

$$\mathbf{w}^*_{MAP} = \mu_N = \beta\Sigma_N^{-1} X^T \mathbf{y}$$

Bayesian Treatment

1. Introduce prior on the parameters
   - For Linear Regression, $p(\mathbf{w}|\alpha) = \mathcal{N}(\mathbf{w}, \mathbf{0}, \alpha^{-1}\mathbf{I})$
2. Model the posterior distribution of parameters
   - $p(\mathbf{w}|\mathbf{y}, X, \alpha) \propto p(\mathbf{y}|\mathbf{w}, X)\ p(\mathbf{w}|\alpha)$
   - For Linear Regression, $p(\mathbf{w}|\mathbf{y}, X, \alpha, \beta) = \mathcal{N}(\mathbf{w}|\mu_N, \Sigma_N)$
3. Predictive Distribution
   - $p(y_{n+1}|\mathbf{x}_{n+1}, \mathbf{y}, X, \alpha, \beta)$

The first two steps are common to the MAP estimate process.

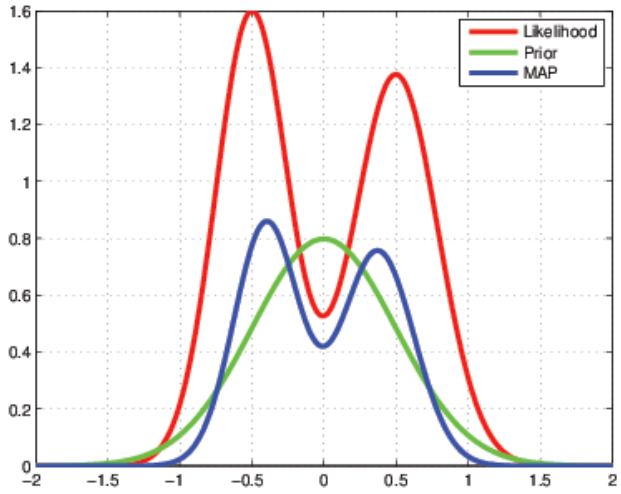
Predictive Distribution

Model the posterior distribution

$$p(\mathbf{w}|\mathbf{y}, X, \alpha, \beta) = \mathcal{N}(\mathbf{w}|\mu_N, \Sigma_N)$$

unlike MAP estimate, *we sum over all possible parameter values*

$$p(y_{n+1}|\mathbf{x}_{n+1}, \mathbf{y}, X, \alpha, \beta) = \int_{\mathbf{w}} p(y_{n+1}|\mathbf{w}, \mathbf{x}_{n+1}, \beta) p(\mathbf{w}|\mathbf{y}, X, \alpha, \beta)$$



Predictive Distribution

$$
\begin{aligned}
p(y_{n+1}|\mathbf{x}_{n+1}, \mathbf{y}, X, \alpha, \beta) &= \int_{\mathbf{w}} p(y_{n+1}|\mathbf{w}, \mathbf{x}_{n+1}, \beta) p(\mathbf{w}|\mathbf{y}, X, \alpha, \beta) \\
&= \mathcal{N}\Big(y \mid \mu_N^T \mathbf{x}, \sigma_N^2(\mathbf{x_{n+1}})\Big)
\end{aligned}
$$

- The variance decreases with the N
- In the limit, $y_{n+1} = \mu_N^T \mathbf{x}_{n+1} = \mathbf{w}_{MAP}^T \mathbf{x}_{n+1}$
- Hyperparameter estimation
  - Put prior on the hyperparameters ?
  - Empirical Bayes or EM

Hyperparameter Estimation – Empirical Bayes

$$p(y|\mathbf{y}, X) = \int_{\mathbf{w}} \int_{\alpha} \int_{\beta} p(y|\mathbf{w}, X, \beta) p(\mathbf{w}|\mathbf{y}, X, \alpha, \beta) p(\alpha, \beta|\mathbf{y}) d\alpha \; d\beta \; d\mathbf{w}$$

- Relatively less sensitive to the hyperparameters
- If posterior $p(\alpha, \beta|\mathbf{y}, X)$ is sharply peaked, then

  $$p(y|\mathbf{y}, X) \approx p(y|\mathbf{y}, X, \alpha^*, \beta^*) = \int_{\mathbf{w}} p(y|\mathbf{w}, X, \beta^*) p(\mathbf{w}|\mathbf{y}, X, \alpha^*, \beta^*)$$

- If the prior is relatively flat, then
  - α^* and β^* are obtained by maximizing the likelihood.

Bayesian Treatment

1. Introduce prior distribution
   - Conjugacy
2. Model the posterior distribution
   - Hyperparameters can be estimated using Empirical Bayes
     - Avoids the Cross-validation step
     - Hence, we can use all the training data
3. Predictive Distribution
   - Integrate over the parameters
   - Draw few samples from posterior and sum over them.

Outline

- MLE, MAP vs. Bayesian Learning
- Bayesian Linear Regression
- Bayesian Gaussian Mixture Models
  - Non-parametric Bayes

Mixture Models (Recap)

- Finite Gaussian Mixture Model
  - $z = 1 \cdots K$ mixture components
  - parameters for each component (μ_k, β).

$$\begin{aligned} p(x, z) &= p(z)p(x|z) \\ p(x) &= \sum_{z=1\ldots K} p(z = k)p(x|\mu_k, \beta) \\ &= \sum_k \phi_k p(x|\mu_k, \beta) \end{aligned}$$

- What are the parameters in Gaussian Mixture Model ?

Bayesian treatment of Mixture Models

Non-parametric Bayes

- What should we do ?

Bayesian treatment of Mixture Models

1. Introduce prior distribution
2. Model the posterior distribution
3. Predictive Distribution

- $p(x) = \sum_k \phi_k p(x|\mu_k, \beta)$
- For GMM, we keep the variance fixed.
  - $p(x|\mu_k, \beta) = \mathcal{N}(\mu_k, \beta^{-1})$
- Put prior on the mixing weights (ϕ_k) and the mean parameters (μ_k).

Dirichlet Process

$$G \sim DP(\alpha_0, G_0)$$

Treat this as a collection of samples

$$\{\theta_1, \theta_2, \cdots\} \text{ with weights } \{\phi_1, \phi_2, \cdots\}$$

- $\theta_i \sim G_0$ can be scalar or vector depending on G_0
  - Countably infinite collection of i.i.d samples
- $\sum_k \phi_k = 1$
  - Stick-breaking construction gives these weights.
  - ϕ_k values depend on α_0
- $\theta \sim G \Rightarrow$ choose a θ_i with weight ϕ_i

Dirichlet Process for GMM

1. Prior on the parameters
   - The base distribution G_0 be $\mathcal{N}(\psi, \gamma\mathbf{I})$
   - $\mu_i \sim G_0 \Rightarrow \mu_i \sim \mathcal{N}(\psi, \gamma\mathbf{I})$
   - Stick-breaking process is used as prior for ϕ_i
   - Allows arbitrary number of mixing components.

$$\begin{aligned} G &\sim DP(\alpha, \mathcal{N}(\psi, \gamma\mathbf{I})) \\ \mu_i|G &\sim G \\ x_i|\mu_i, \beta &\sim \mathcal{N}(\mu_i, \beta^{-1}) \end{aligned}$$

   - Chinese Restaurant Process

Dirichlet Process for GMM

1. Modeling the posterior
   - c_i denote the cluster indicator of i^{th} example
   - $p(\mathbf{c}, \mu|X) \propto p(\mathbf{c}|\alpha)p(\mu|\mathbf{c}, X)$
   - Run Gibbs sampler.
   - Estimate the hyperparameters (α and γ)
2. Predictive distribution
   - Draw samples from the posterior.
   - Sum over those samples.

- Doesn't need to specify the number of components.

Non-parametric Bayes

- Stick-breaking construction gives prior on mixing components.
- Learns the number of components from the data.
- Hyperparameters are estimated using Empirical Bayes
- Hierarchical Dirichlet Process (HDP)
  - Possible to design hierarchical models

Take Away ...

1. Maximum Likelihood Estimate (MLE)
   - $\theta^* = \arg\max_\theta p(\mathcal{D}|\theta)$
   - Use θ^* in future to predict y_{n+1} given $\mathbf{x}_{n+1}$
2. Maximum a posteriori estimation (MAP)
   - $\theta^* = \arg\max_\theta p(\theta|\mathcal{D}, \alpha) = \arg\max_\theta p(\mathcal{D}|\theta)p(\theta|\alpha)$
     - α is called Hyperparameter
   - Use θ^* in future to predict y_{n+1} given $\mathbf{x}_{n+1}$
3. Bayesian treatment
   - model $p(\theta|\mathcal{D}, \alpha)$
   - $p(y_{n+1}|\mathbf{x}_{n+1}, \mathcal{D}, \alpha) = \int_\theta p(y_{n+1}|\theta, \mathbf{x_{n+1}})\mathbf{p}(\theta|\mathcal{D}, \alpha)\mathbf{d}\theta$

Questions ?