CHAPTER 16:

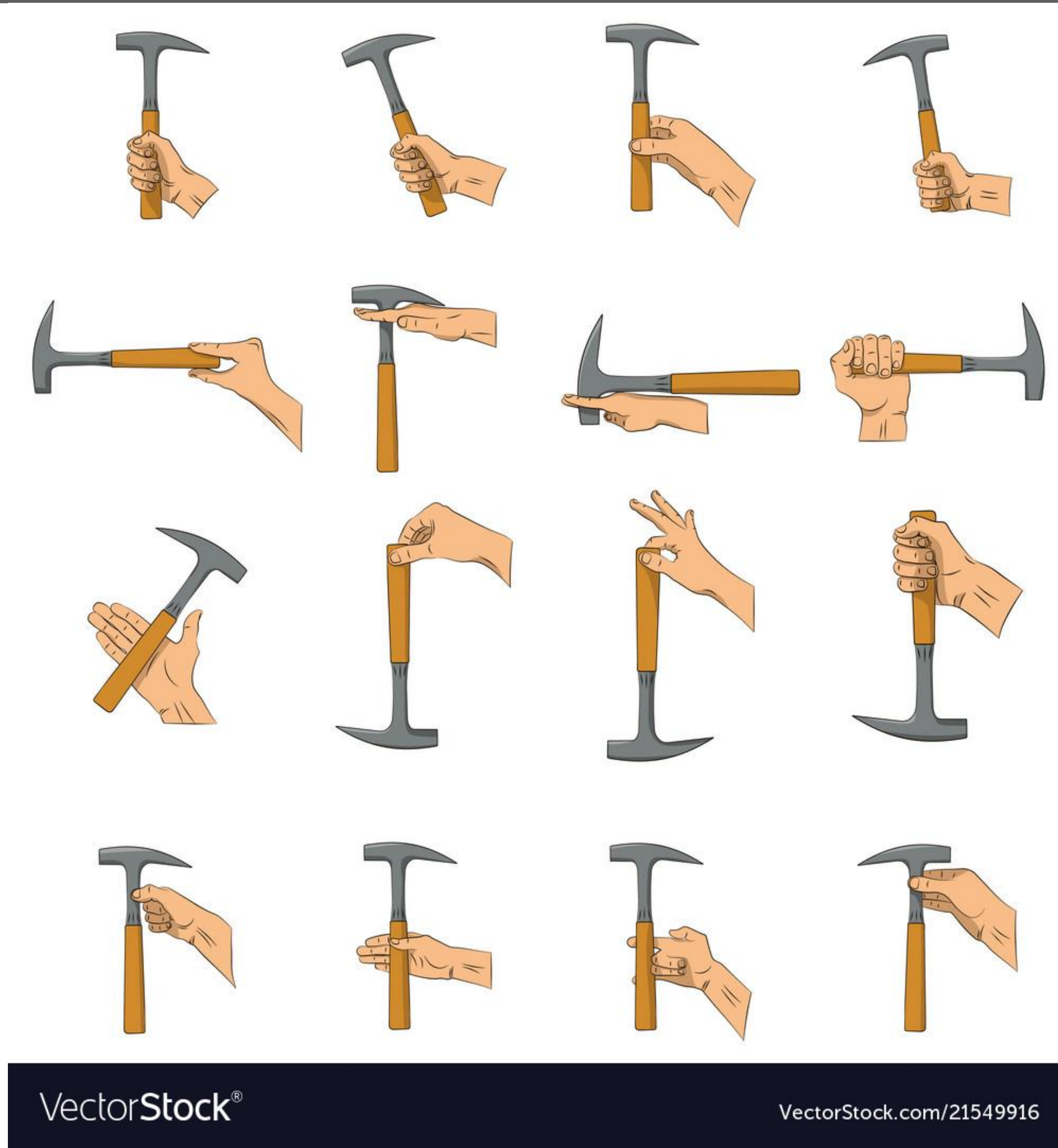
BAYESIAN ESTIMATION

Stella Grasshof

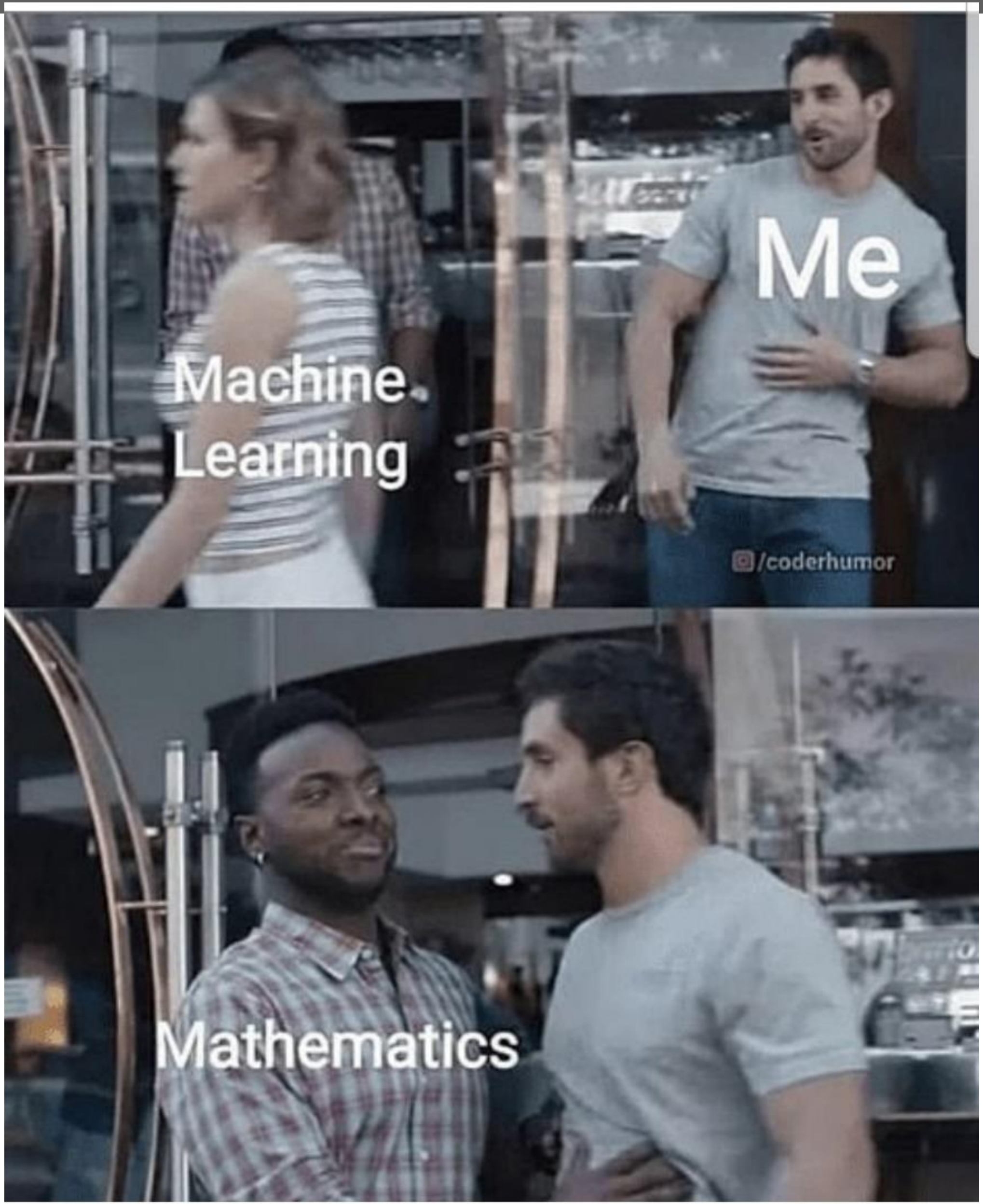
Overview

- Repeat
- Bayesian estimation
- Bayesian estimation of unknown mean
- Bayesian regression
- Howto Prior
- Model Quality

Math is a tool



Math is a tool



Repeat: Distribution

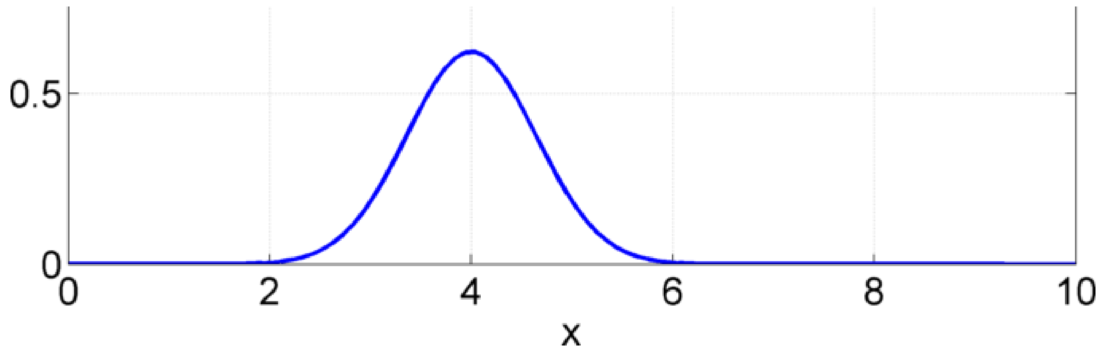
1D Normal

$$X \sim \mathcal{N}(\mu, \sigma^2)$$

$$\mathrm{E}[X] = \mu, \mathrm{Var}[X] = \sigma^2$$

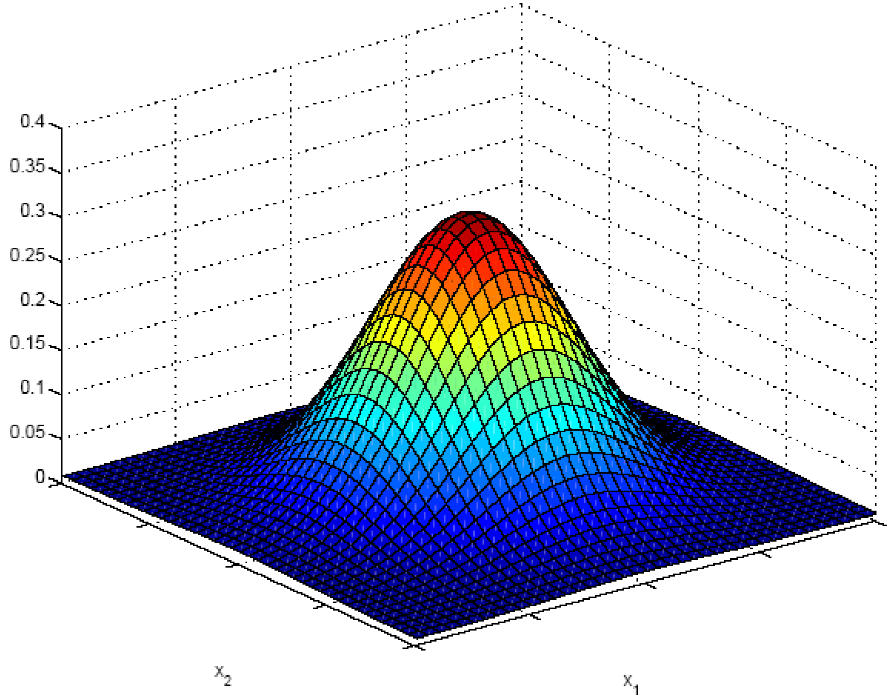
$$\text{p.d.f.} \;\; f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{1}{2} \left(\frac{x-\mu}{\sigma^2} \right)^2 \right]$$

$$\text{c.d.f.} \;\; \Phi(x) = P(X \leq x)$$



Repeat: Distribution

multidimsional Normal



D-dimensional: $\boldsymbol{x} \in \mathbb{R}^D$

$\boldsymbol{x} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$

$$p(\boldsymbol{x}) = \frac{1}{(2\pi)^{D/2}|\boldsymbol{\Sigma}|^{1/2}} \exp\left[-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right]$$

Repeat: Distribution

Discrete Binomial

$$X \sim B(n, k)$$

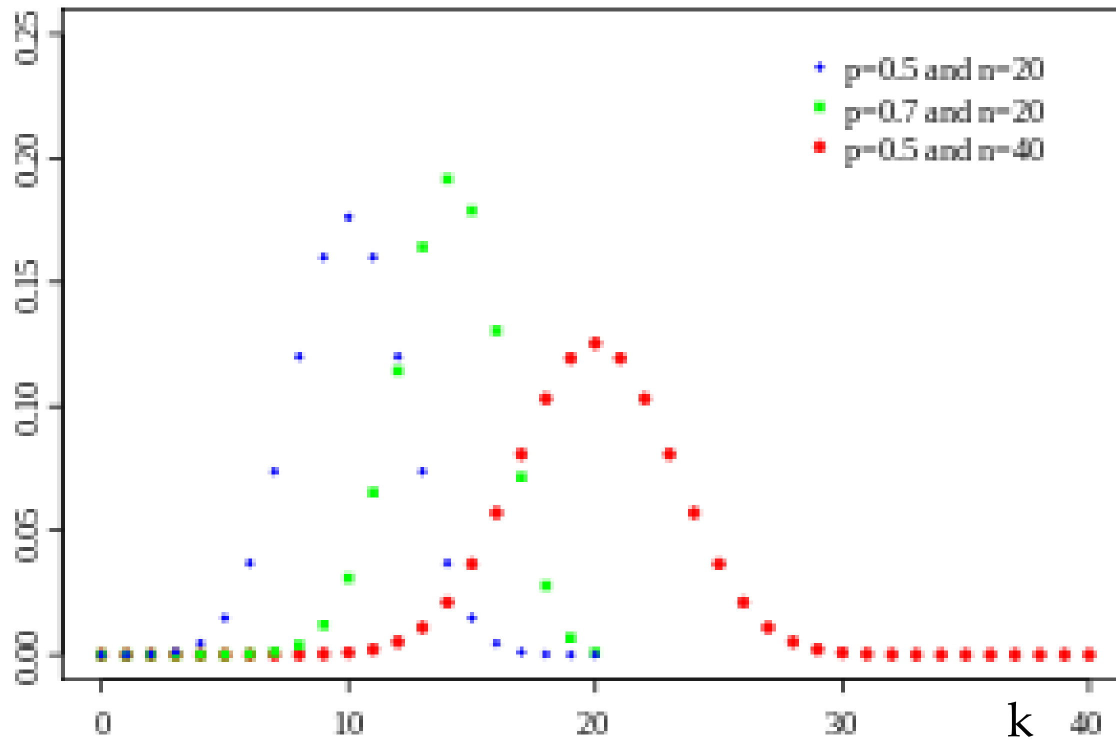
$$\mathrm{E}[X] = np, \mathrm{Var}[X] = np(1-p)$$

p.d.f.

$$f(k) = \binom{n}{k} p^k (1-p)^{n-k}$$

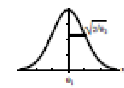
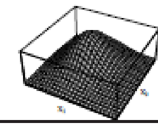
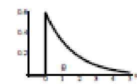
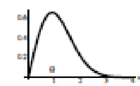
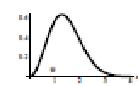
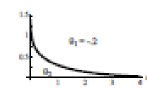
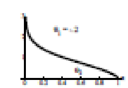
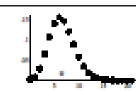
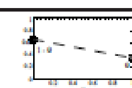
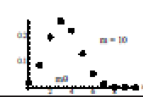
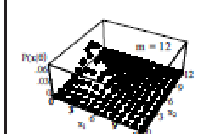
c.d.f.

$$F(x) = P(X \leq x)$$

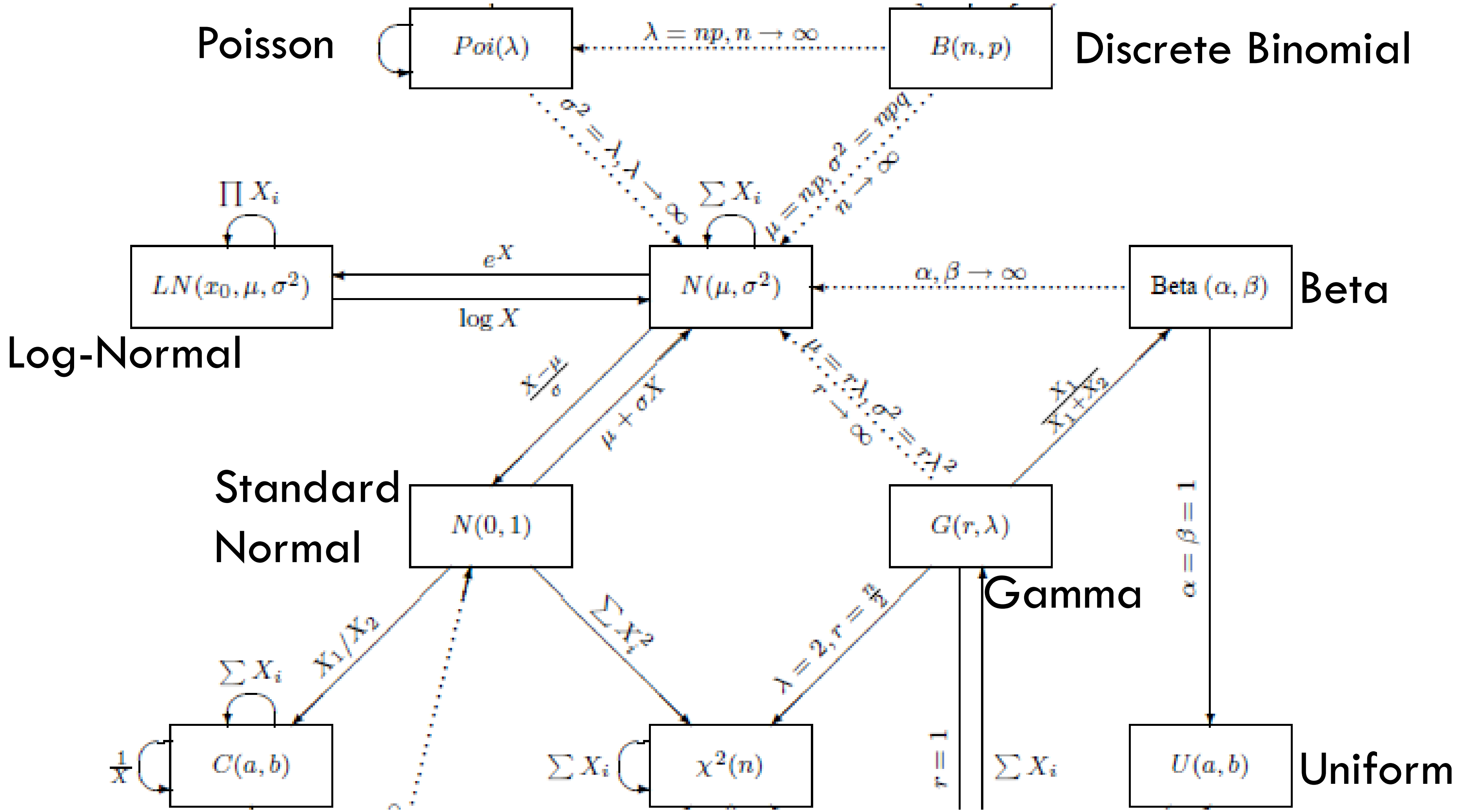


https://en.wikipedia.org/wiki/Binomial_distribution

Table 3.1: Common Exponential Distributions and their Sufficient Statistics.

Name	Distribution	Domain		s
Normal	$p(x\|\boldsymbol{\theta}) = \sqrt{\frac{\theta_2}{2\pi}} e^{-(1/2)\theta_2(x-\theta_1)^2}$	$\theta_2 > 0$		$\begin{bmatrix} \frac{1}{n}\sum_{k=1}^{n} x_k \\ \frac{1}{n}\sum_{k=1}^{n} x_k^2 \end{bmatrix}$
Multi-variate Normal	$p(\mathbf{x}\|\boldsymbol{\theta}) = \frac{\|\boldsymbol{\Theta}_2\|^{1/2}}{(2\pi)^{d/2}} e^{-(1/2)(\mathbf{x}-\boldsymbol{\theta}_1)^t \boldsymbol{\Theta}_2 (\mathbf{x}-\boldsymbol{\theta}_1)}$	$\boldsymbol{\Theta}_2$ positive definite		$\begin{bmatrix} \frac{1}{n}\sum_{k=1}^{n} \mathbf{x}_k \\ \frac{1}{n}\sum_{k=1}^{n} \mathbf{x}_k \mathbf{x}_k^t \end{bmatrix}$
Exponential	$p(x\|\theta) = \begin{cases} \theta e^{-\theta x} & x \geq 0 \\ 0 & \text{otherwise} \end{cases}$	$\theta > 0$		$\frac{1}{n}\sum_{k=1}^{n} x_k$
Rayleigh	$p(x\|\theta) = \begin{cases} 2\theta x e^{-\theta x^2} & x \geq 0 \\ 0 & \text{otherwise} \end{cases}$	$\theta > 0$		$\frac{1}{n}\sum_{k=1}^{n} x_k^2$
Maxwell	$p(x\|\theta) = \begin{cases} \frac{4}{\sqrt{\pi}} \theta^{3/2} x^2 e^{-\theta x^2} & x \geq 0 \\ 0 & \text{otherwise} \end{cases}$	$\theta > 0$		$\frac{1}{n}\sum_{k=1}^{n} x_k^2$
Gamma	$p(x\|\boldsymbol{\theta}) = \begin{cases} \frac{\theta_2^{\theta_1+1}}{\Gamma(\theta_1+1)} x^{\theta_1} e^{-\theta_2 x} & x \geq 0 \\ 0 & \text{otherwise} \end{cases}$	$\theta_1 > -1$ $\theta_2 > 0$		$\begin{bmatrix} \left(\prod_{k=1}^{n} x_k\right)^{1/n} \\ \frac{1}{n}\sum_{k=1}^{n} x_k \end{bmatrix}$
Beta	$p(x\|\boldsymbol{\theta}) = \begin{cases} \frac{\Gamma(\theta_1+\theta_2+2)}{\Gamma(\theta_1+1)\Gamma(\theta_2+1)} x^{\theta_1}(1-x)^{\theta_2} & 0 \leq x \leq 1 \\ 0 & \text{otherwise} \end{cases}$	$\theta_1 > -1$ $\theta_2 > -1$		$\begin{bmatrix} \left(\prod_{k=1}^{n} x_k\right)^{1/n} \\ \left(\prod_{k=1}^{n} (1-x_k)\right)^{1/n} \end{bmatrix}$
Poisson	$P(x\|\theta) = \frac{\theta^x}{x!} e^{-\theta} \quad x = 0, 1, 2, \ldots$	$\theta > 0$		$\frac{1}{n}\sum_{k=1}^{n} x_k$
Bernoulli	$P(x\|\theta) = \theta^x (1-\theta)^{1-x} \quad x = 0, 1$	$0 < \theta < 1$		$\frac{1}{n}\sum_{k=1}^{n} x_k$
Binomial	$P(x\|\theta) = \frac{m!}{x!(m-x)!} \theta^x (1-\theta)^{m-x}$ $x = 0, 1, \ldots, m$	$0 < \theta < 1$		$\frac{1}{n}\sum_{k=1}^{n} x_k$
Multinomial	$P(\mathbf{x}\|\boldsymbol{\theta}) = \frac{m! \prod_{i=1}^{d} \theta_i^{x_i}}{\prod_{i=1}^{d} x_i!} \quad x_i = 0, 1, \ldots, m \quad \sum_{i=1}^{d} x_i = m$	$0 < \theta_i < 1$ $\sum_{i=1}^{d} \theta_i = 1$		$\frac{1}{n}\sum_{k=1}^{n} \mathbf{x}_k$

Repeat: Distribution



Source: script “Biometrie“, 2009, Prof. Ziegler, Lübeck

Repeat: ML vs. MAP

Task: Given data $\mathcal{X}$, estimate parameter θ

- Maximum Likelihood (ML)

$$\theta_{\text{ML}} = \arg\max_{\theta}\, p(\mathcal{X}|\theta)$$

Likelihood

- Maximum A Posteriori (MAP)

$$\begin{aligned}\theta_{\text{MAP}} &= \arg\max_{\theta}\, \boxed{\begin{array}{c} p(\theta|\mathcal{X}) \\ \dfrac{p(\mathcal{X}|\theta)p(\theta)}{p(\mathcal{X})} \end{array}} \\ & \\ &= \arg\max_{\theta}\, p(\mathcal{X}|\theta)p(\theta)\end{aligned}$$

Posterior

Bayes' Rule

give **point estimates: one fixed parameter** θ

Bayesian Approach

1. Prior is pdf $p(\theta)$
   - **high** weight in regions where θ is **likely**
   - **low** weight in regions where θ is **unlikely**
2. Assume parameter θ is not fixed
   generate several estimates θ and average, weighted by probabilities

	1.	2.
$\theta_{\mathrm{ML}} = \arg\max_{\theta} p(\mathcal{X}\|\theta)$	✘	✘
$\theta_{\mathrm{MAP}} = \arg\max_{\theta} p(\mathcal{X}\|\theta)p(\theta)$	✓	✘

Repeat: ML vs. MAP

- Maximum Likelihood (ML)

$$\theta_{\mathrm{ML}} = \arg\max_{\theta} p(\mathcal{X}|\theta)$$

- Maximum A Posteriori (MAP)

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

one fixed parameter

- Bayes

  Parameter θ is random variable with prior $p(\theta)$

$$\theta_{\mathrm{Bayes}} = \mathrm{E}[\theta|\mathcal{X}] = \int \theta p(\theta|\mathcal{X}) d\theta$$

Frequentist vs. Bayesian

Frequentist Approach

- Assumes unknown, fixed parameter θ
- Estimates θ with some confidence
- Prediction by estimated parameter value

Bayesian Approach

- Unknown parameters as **random variables**
- Probability quantifies uncertainty
- Prediction by expectation over unknown parameters

Why is this relevant? Example 1

Consider coin toss example

probability of heads $\widehat{p} = \frac{\#\{\text{Heads}\}}{\#\{\text{Tosses}\}}$

$$\mathcal{X} = \{1,1,1,0,0,1,1,0,1\} \Rightarrow \widehat{p} = \frac{6}{9}$$

$$\mathcal{X} = \{1,1,1\} \Rightarrow \widehat{p} = \frac{3}{3} = 1$$

Why is this relevant? Example 2

Consider a Casino

- Machine A: 3 wins out of 4 plays
- Machine B: 81 wins out of 121 plays

Which Machine would you choose?

- by intuition: B
  because more samples = reliable

- ML estimate

$$\widehat{\theta}_{ML,A} = \frac{3}{4} \approx 0.75 \qquad \widehat{\theta}_{ML,B} = \frac{81}{121} \approx 0.67$$

Why is this relevant? Example 2

Consider a Casino

- Machine A: 3 wins out of 4 plays
- Machine B: 81 wins out of 121 plays

Binomial distribution pdf:

$$p(k|n,\theta) = P(X=k) = \binom{n}{k}\theta^k(1-\theta)^{n-k}$$

- MAP estimate

$$\widehat{\theta}_{\text{MAP}} = \arg\max_{\theta} p(n,k|\theta)p(\theta)$$

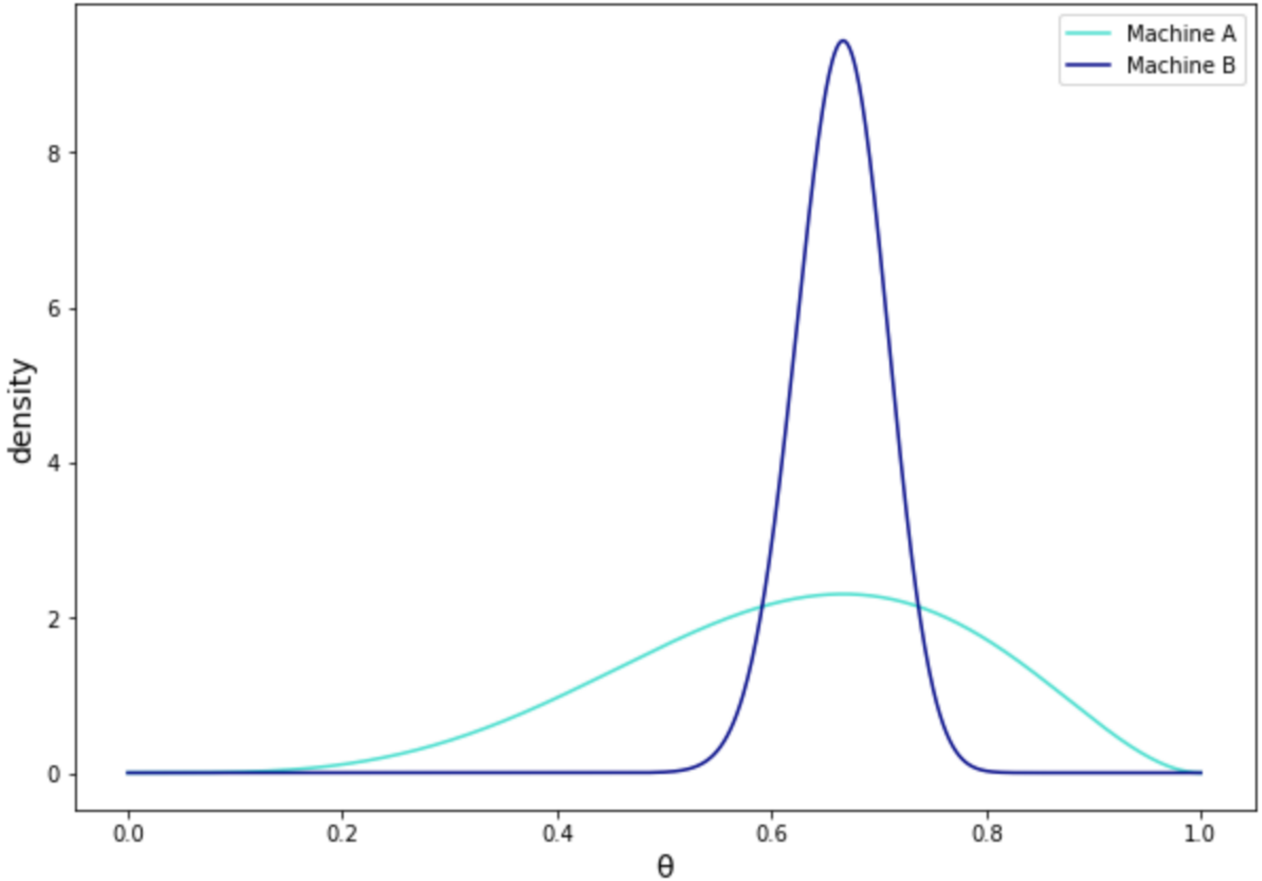
Detailed steps on learnit

$$\widehat{\theta}_{\text{MAP},A} = \widehat{\theta}_{\text{MAP},B} \approx 0.667$$

Why is this relevant? Example 2

Bayes $$p(\theta|\text{Data}) = p(\theta|n,k) = \frac{p(n,k|\theta)p(\theta)}{p(n,k)}$$

$$p_A(\theta|n,k)$$
$$p_B(\theta|n,k)$$



https://towardsdatascience.com/mle-map-and-bayesian-inference-3407b2d6d4d9

Why is this relevant? Example 2

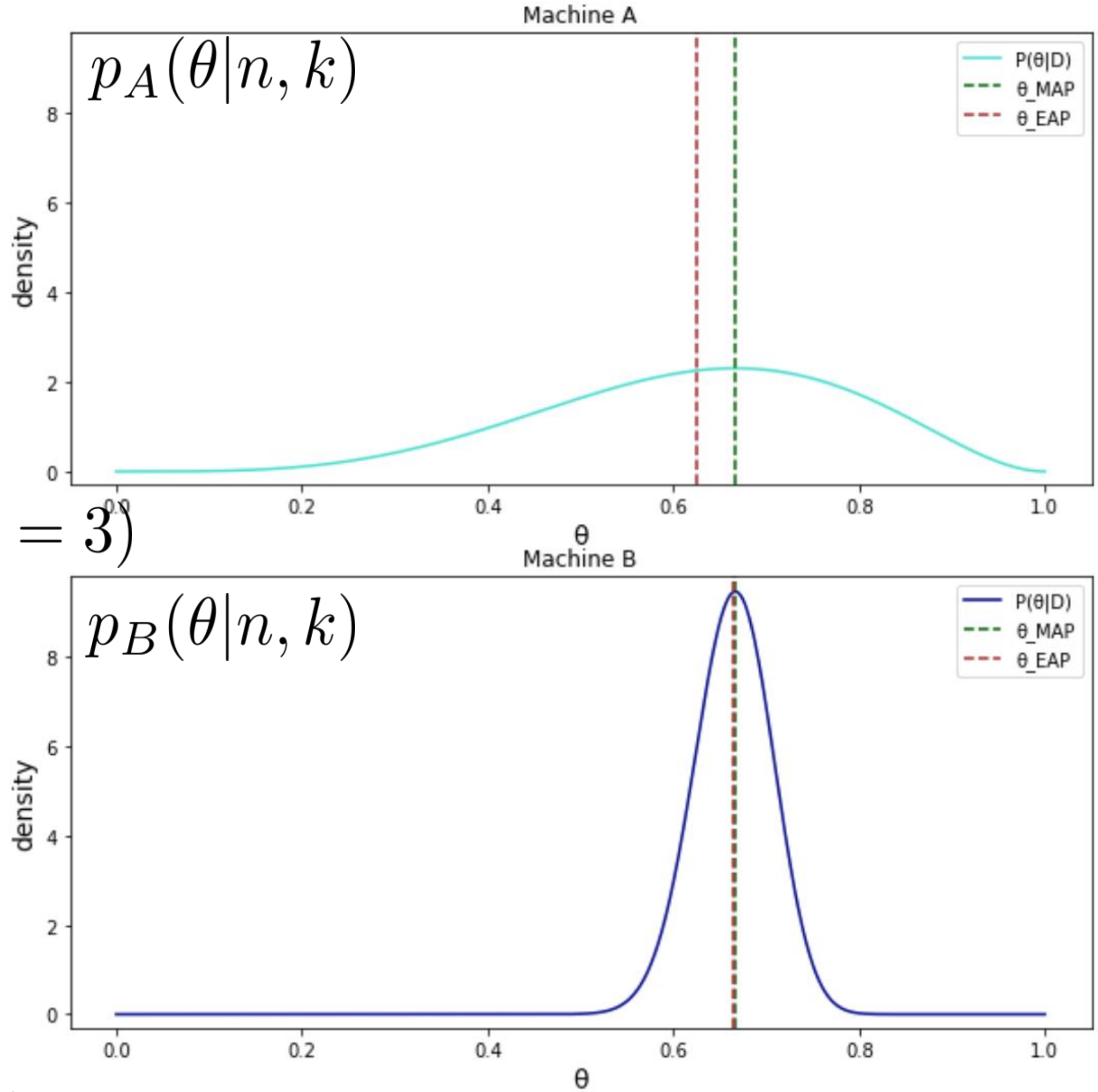
Bayes:

Expected A Posteriori (EAP)

$$\theta_{\mathrm{Bayes,A}} = \mathrm{E}(\theta|\mathcal{X}_A) = \mathrm{E}(\theta|n=4, k=3)$$

$$\theta_{\mathrm{Bayes,A}} \approx 0.625$$

$$\theta_{\mathrm{Bayes,B}} \approx 0.664$$



https://towardsdatascience.com/mle-map-and-bayesian-inference-3407b2d6d4d9

Why is this relevant? Example 2

Consider a Casino

- Machine A: 3 wins out of 4 plays
- Machine B: 81 wins out of 121 plays

Estimated winning probability:

- ML estimate $\widehat{\theta}_{\mathrm{ML},A} = 0.75, \qquad \widehat{\theta}_{\mathrm{ML},B} \approx 0.67$
- MAP estimate $\widehat{\theta}_{\mathrm{MAP},A} \approx 0.667, \qquad \widehat{\theta}_{\mathrm{MAP},B} \approx 0.667$
- Bayes estimate $\theta_{\mathrm{Bayes,A}} \approx 0.625 \qquad \theta_{\mathrm{Bayes,B}} \approx 0.664$

Likelihood or Bayes?

- Bayes uses more information than ML solutions:
  - additional training data changes the estimate
  - Uncertainty of estimate is well reflected
- But Bayes is often complex to compute
  therefore sampling of the posterior

=> Choice of Likelihood or Bayes depends on problem

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
- Approximate the posterior
  with a model easier to integrate

Estimate Parameters of Distribution

- Assume data is Gaussian $f(x) = \mathcal{N}(\mu, \sigma^2)$

$$p(\mathcal{X}|\mu, \sigma^2) = \prod_{n=1}^{N} f(x_n|\mu, \sigma^2)$$

- Gaussian Prior

$$p(\mu) = \mathcal{N}(\mu_0, \sigma_0^2)$$

- Posterior is Gaussian

$$p(\mu|\mathcal{X}) \propto p(\mu)p(\mathcal{X}|\mu) = \mathcal{N}(\mu_N, \sigma_N^2)$$

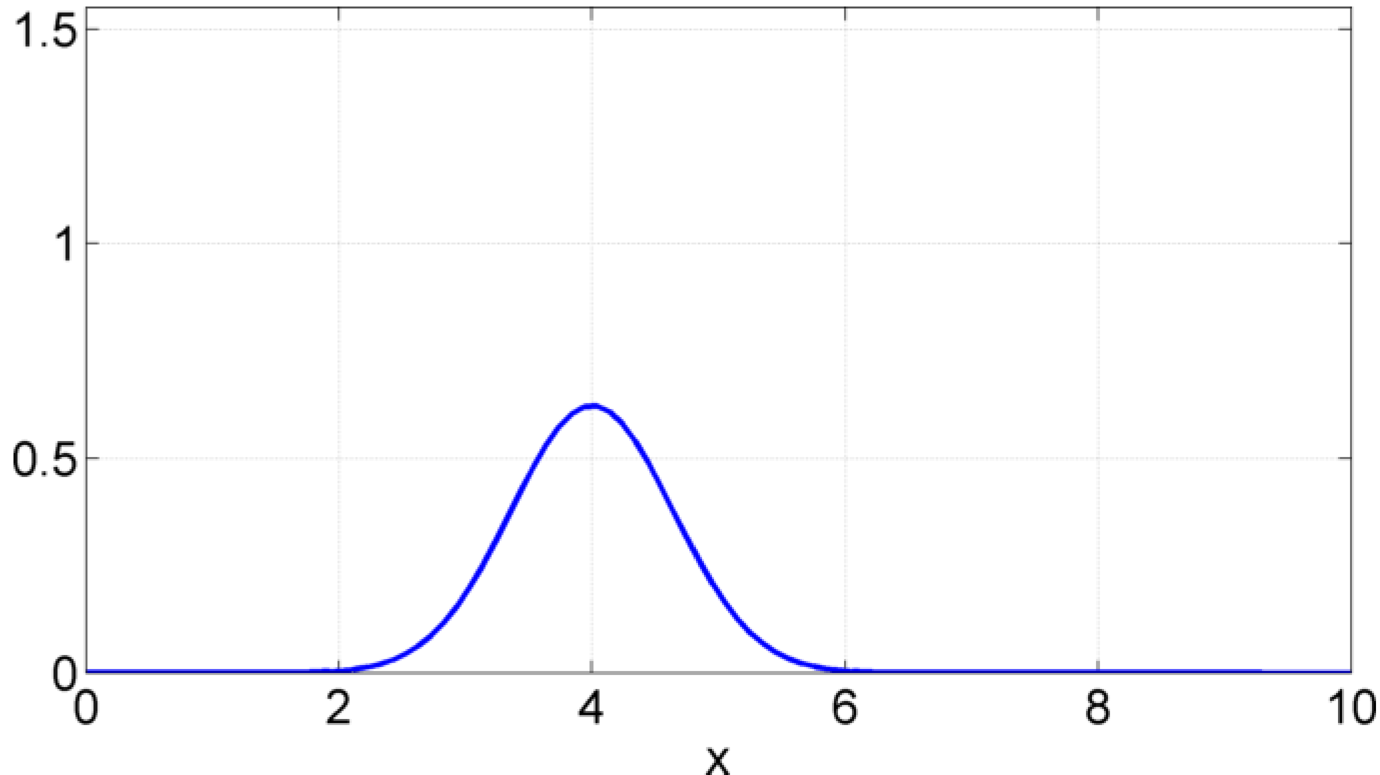
$$\mu_N = \frac{\sigma^2}{N\sigma_0^2 + \sigma^2}\mu_0 + \frac{N\sigma_0^2}{N\sigma_0^2 + \sigma^2}m \qquad m = \frac{1}{N}\sum_{n=1}^{N} x_n$$

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

Estimate Parameters of Distribution

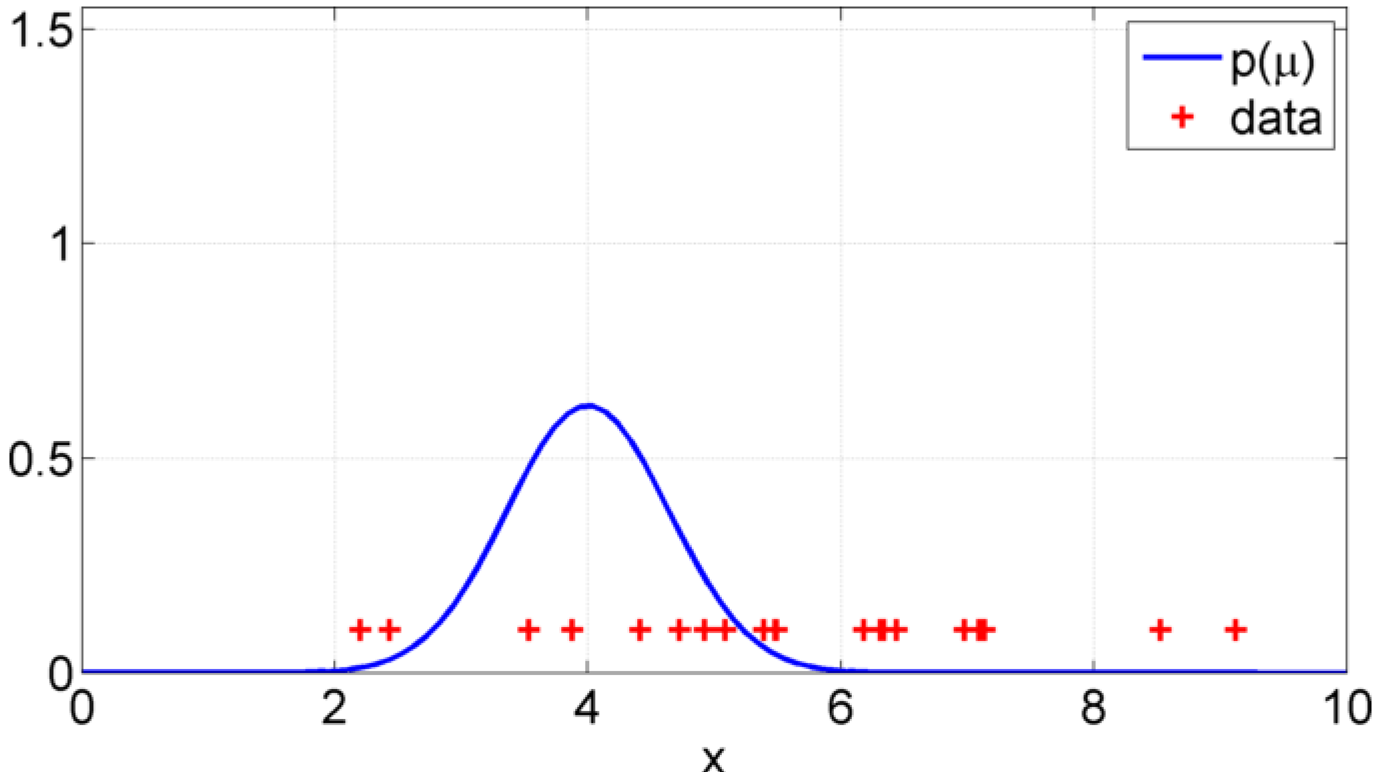
- Gaussian Prior $p(\mu) = \mathcal{N}(\mu_0 = 4, \sigma_0^2 = 0.8^2)$

  assumption of prior knowledge!



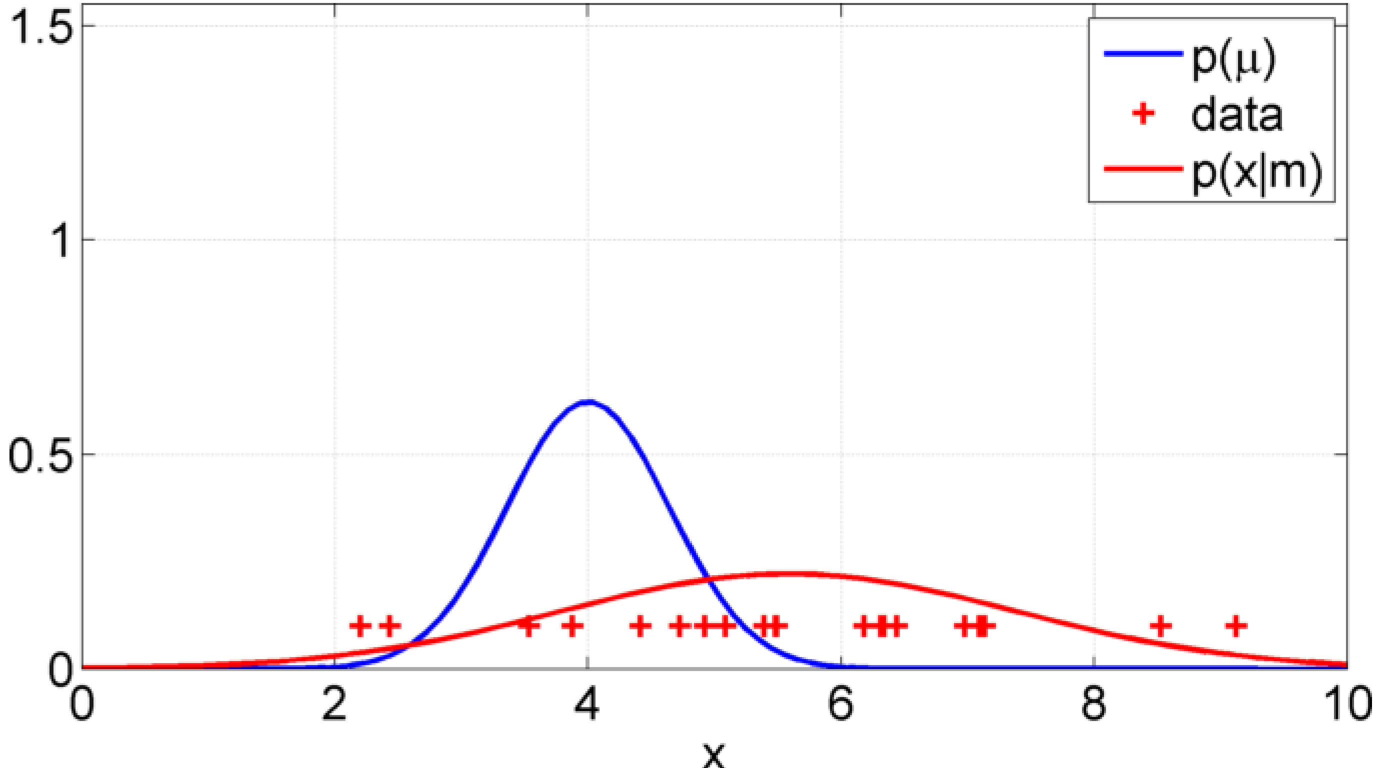
Estimate Parameters of Distribution

- Gaussian Prior $p(\mu) = \mathcal{N}(\mu_0 = 4, \sigma_0^2 = 0.8^2)$
- Data $p(\mathcal{X}|\mu) = \mathcal{N}(m = 6, \sigma^2 = 1.5^2)$



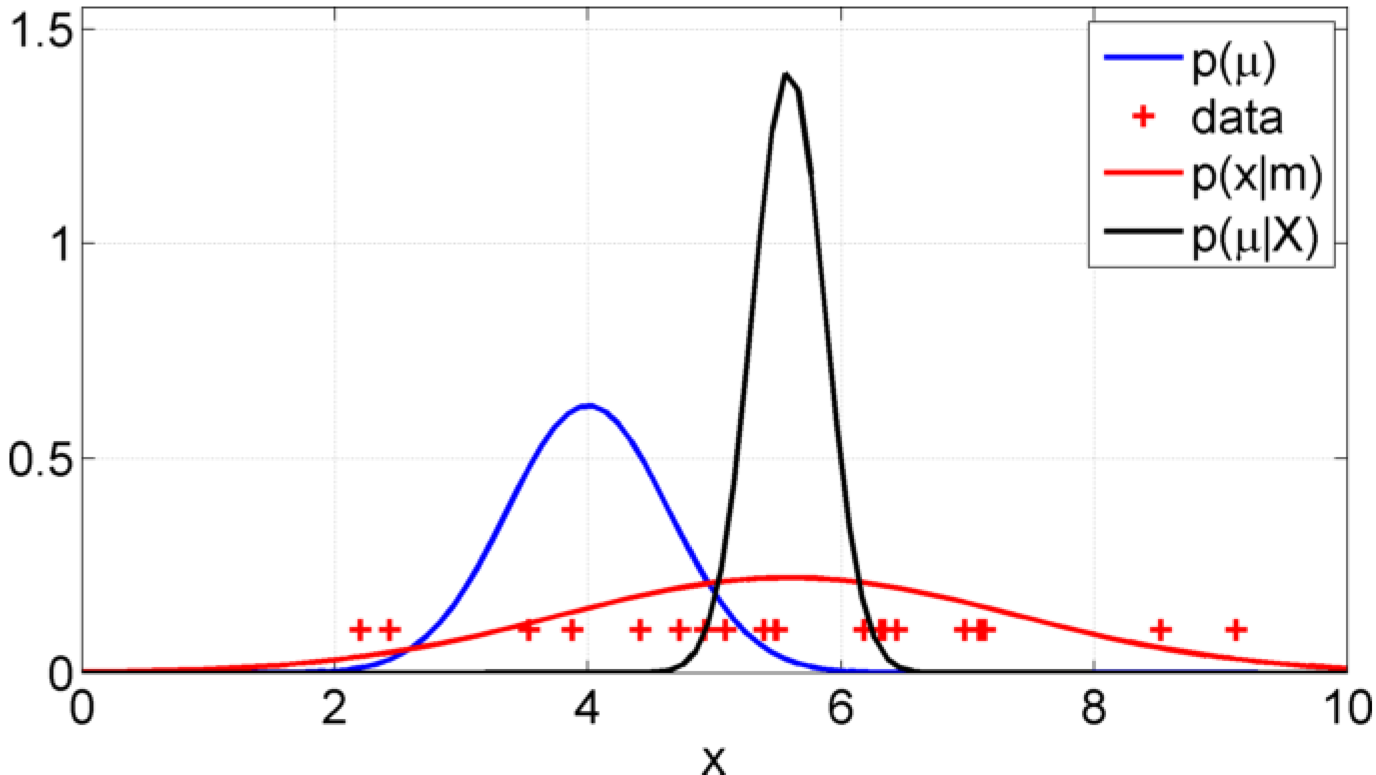
Estimate Parameters of Distribution

- Gaussian Prior $p(\mu) = \mathcal{N}(\mu_0 = 4, \sigma_0^2 = 0.8^2)$
- Data $p(\mathcal{X}|\mu) = \mathcal{N}(m = 6, \sigma^2 = 1.5^2)$

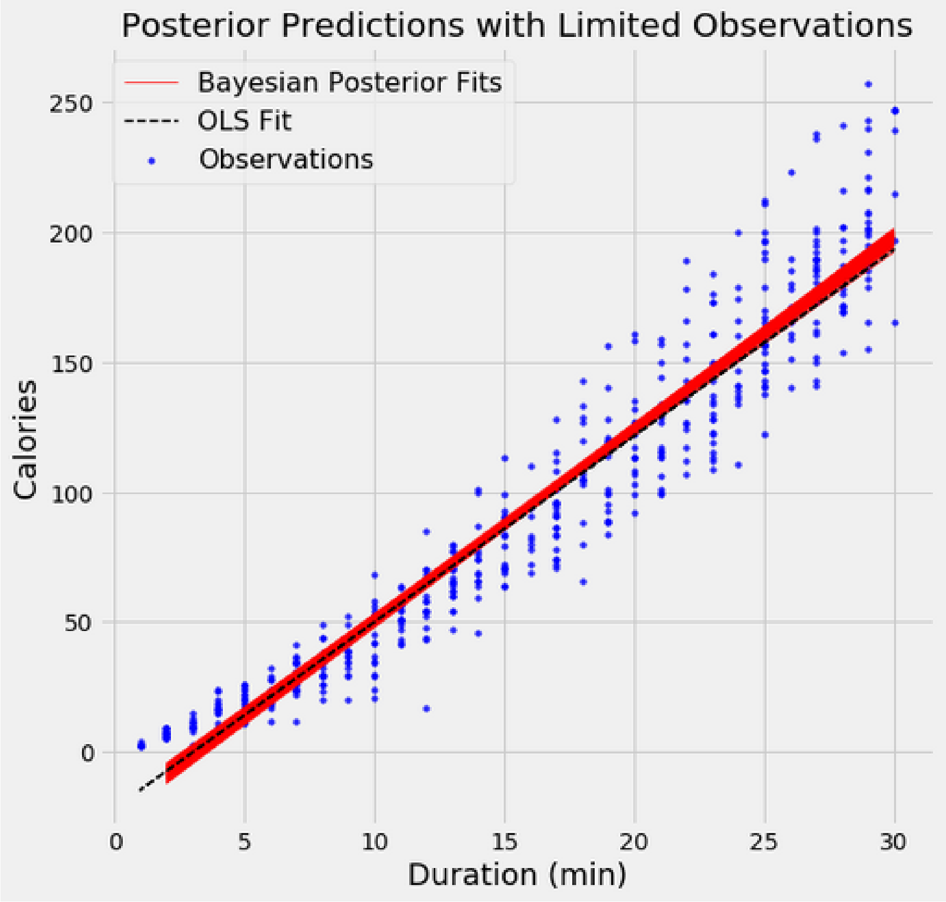


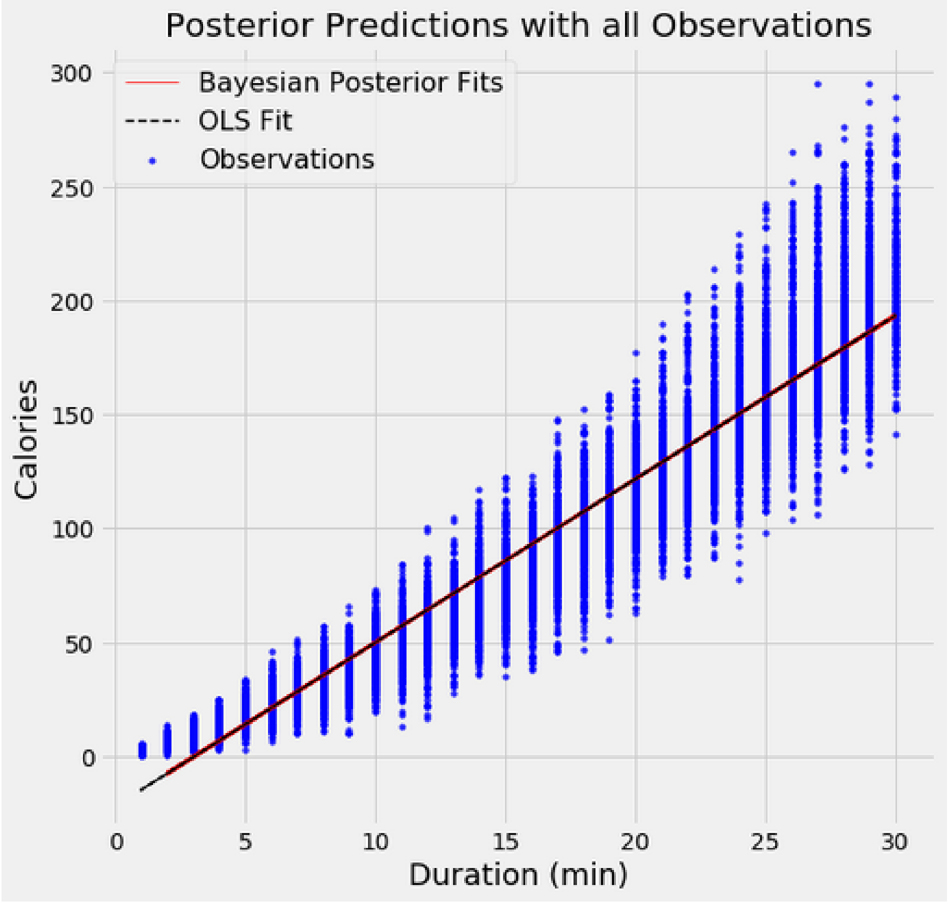
Estimate Parameters of Distribution

- Gaussian Prior $p(\mu) = \mathcal{N}(\mu_0 = 4, \sigma_0^2 = 0.8^2)$
- Data $p(\mathcal{X}|\mu) = \mathcal{N}(m = 6, \sigma^2 = 1.5^2)$
- Posterior $p(\mu|\mathcal{X}) = \mathcal{N}(\mu_N = 5.7, \sigma_N^2 = 0.3^2)$



Bayesian Regression





https://towardsdatascience.com/introduction-to-bayesian-linear-regression-e66e60791ea7

Bayesian Regression

- Line with Gaussian error

$$r_n = f(\boldsymbol{x}_n) = w_0 + w_1\boldsymbol{x}_n + \epsilon, \quad \epsilon \sim \mathcal{N}(0, 1/\beta)$$
$$r_n = f(\boldsymbol{x}_n) = \boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}_n + \epsilon, \quad \epsilon \sim \mathcal{N}(0, 1/\beta),\ \boldsymbol{w} = (w_0, w_1)^{\mathrm{T}}$$

- Assume samples $\boldsymbol{x}_n, r_n,\ n = 1, \ldots, N$

$$p(r_n|\boldsymbol{x}_n, \boldsymbol{w}, \beta) \sim \mathcal{N}(\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}_n, 1/\beta)$$

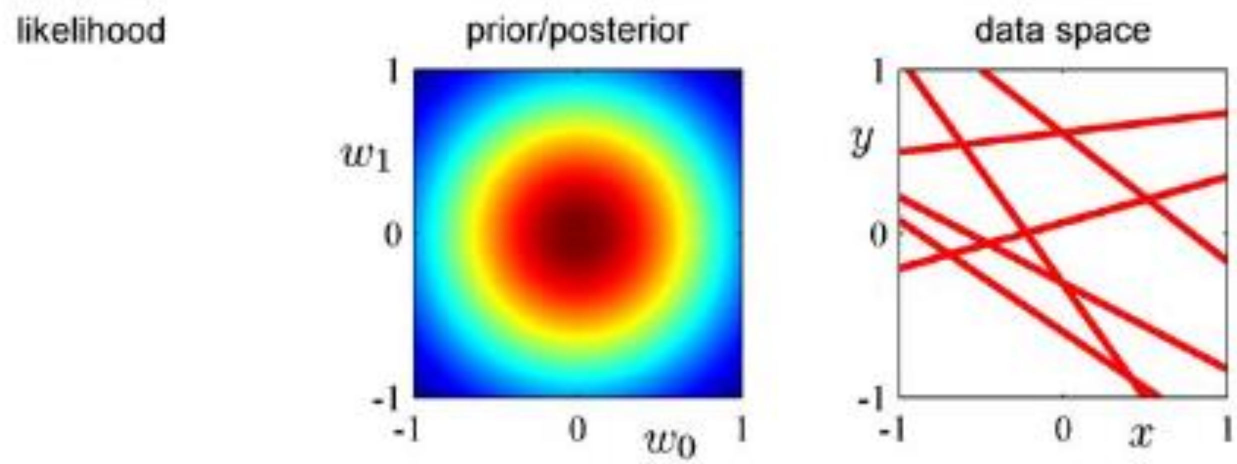
- ML estimate

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

- Assume prior $p(\boldsymbol{w}) \sim \mathcal{N}(\mathbf{0}, 1/\alpha\boldsymbol{I})$

Bayesian Regression

No data
N=0



$$p(\boldsymbol{w}) \sim \mathcal{N}(\mathbf{0}, 1/\alpha \boldsymbol{I})$$

Duda, “Pattern Recognition“

Bayesian Regression

- Line with Gaussian error

$$r_n = f(\boldsymbol{x}_n) = \boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}_n + \epsilon, \quad \epsilon \sim \mathcal{N}(0, 1/\beta), \;\; \boldsymbol{w} = (w_0, w_1)^{\mathrm{T}}$$

- Assume samples $\boldsymbol{x}_n, r_n, \;\; n = 1, \ldots, N$

$$p(r_n|\boldsymbol{x}_n, \boldsymbol{w}, \beta) \sim \mathcal{N}(\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}_n, 1/\beta)$$

- ML and MAP estimate

$$\boldsymbol{w}_{\mathrm{ML}} = (\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X})^{-1}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{r} \quad \boldsymbol{w}_{\mathrm{MAP}} = \beta((\alpha\boldsymbol{I} + \beta\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X})^{-1})\boldsymbol{X}^{\mathrm{T}}\boldsymbol{r}$$

- Assume prior $p(\boldsymbol{w}) \sim \mathcal{N}(\boldsymbol{0}, 1/\alpha\boldsymbol{I})$
- posterior $p(\boldsymbol{w}|\boldsymbol{X}, \boldsymbol{r}) \propto \underbrace{p(\boldsymbol{X}, \boldsymbol{r}|\boldsymbol{w})}_{\text{likelihood}}\underbrace{p(\boldsymbol{w})}_{\text{prior}}$

prior

likelihood

Bayesian Regression

- Line with Gaussian error

$$r_n = f(\boldsymbol{x}_n) = \boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}_n + \epsilon, \quad \epsilon \sim \mathcal{N}(0, 1/\beta), \ \boldsymbol{w} = (w_0, w_1)^{\mathrm{T}}$$

- Assume samples $\boldsymbol{x}_n, r_n, \ n = 1, \ldots, N$

$$p(r_n|\boldsymbol{x}_n, \boldsymbol{w}, \beta) \sim \mathcal{N}(\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}_n, 1/\beta)$$

- ML and MAP estimate

$$\boldsymbol{w}_{\mathrm{ML}} = (\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X})^{-1}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{r} \quad \boldsymbol{w}_{\mathrm{MAP}} = \beta((\alpha\boldsymbol{I} + \beta\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X})^{-1})\boldsymbol{X}^{\mathrm{T}}\boldsymbol{r}$$

- Assume prior $p(\boldsymbol{w}) \sim \mathcal{N}(\mathbf{0}, 1/\alpha\boldsymbol{I})$
- posterior $p(\boldsymbol{w}|\boldsymbol{X}, \boldsymbol{r}) \sim \mathcal{N}(\boldsymbol{\mu}_N, \boldsymbol{\Sigma}_N)$

$$\boldsymbol{\mu}_N = \beta\boldsymbol{\Sigma}_{\boldsymbol{N}}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{r}$$

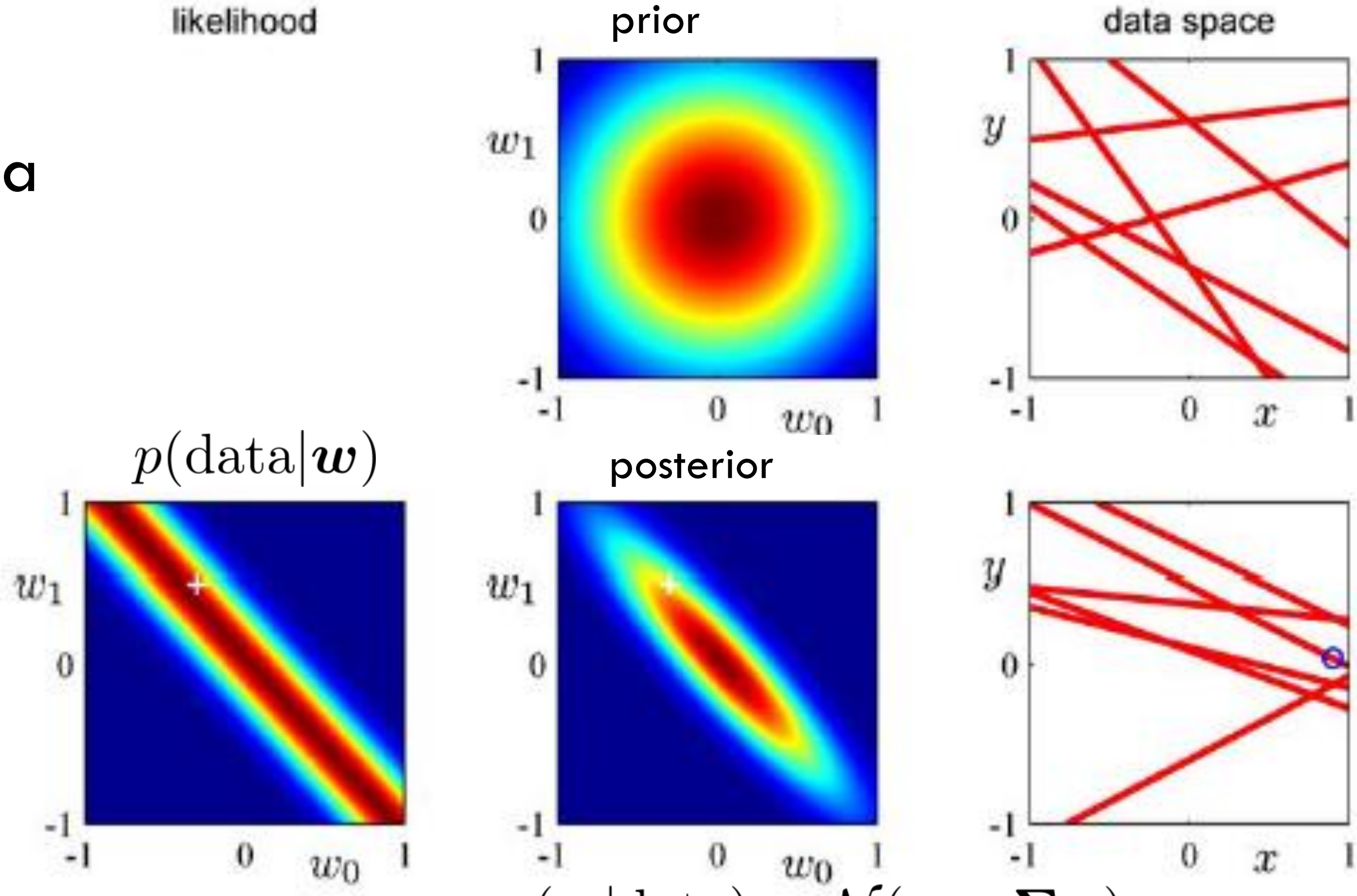
$$\boldsymbol{\Sigma}_N = (\alpha\boldsymbol{I} + \beta\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X})^{-1}$$

Bayesian Regression

$p(\boldsymbol{w}) \sim \mathcal{N}(\mathbf{0}, 1/\alpha \boldsymbol{I})$

No data
N=0

N=1



$p(\boldsymbol{w}|\text{data}) \sim \mathcal{N}(\boldsymbol{\mu}_N, \boldsymbol{\Sigma}_N)$

$p(\boldsymbol{w}|\boldsymbol{X}, \boldsymbol{r}) \propto p(\boldsymbol{X}, \boldsymbol{r}|\boldsymbol{w})p(\boldsymbol{w})$

Posterior =likelihood x prior

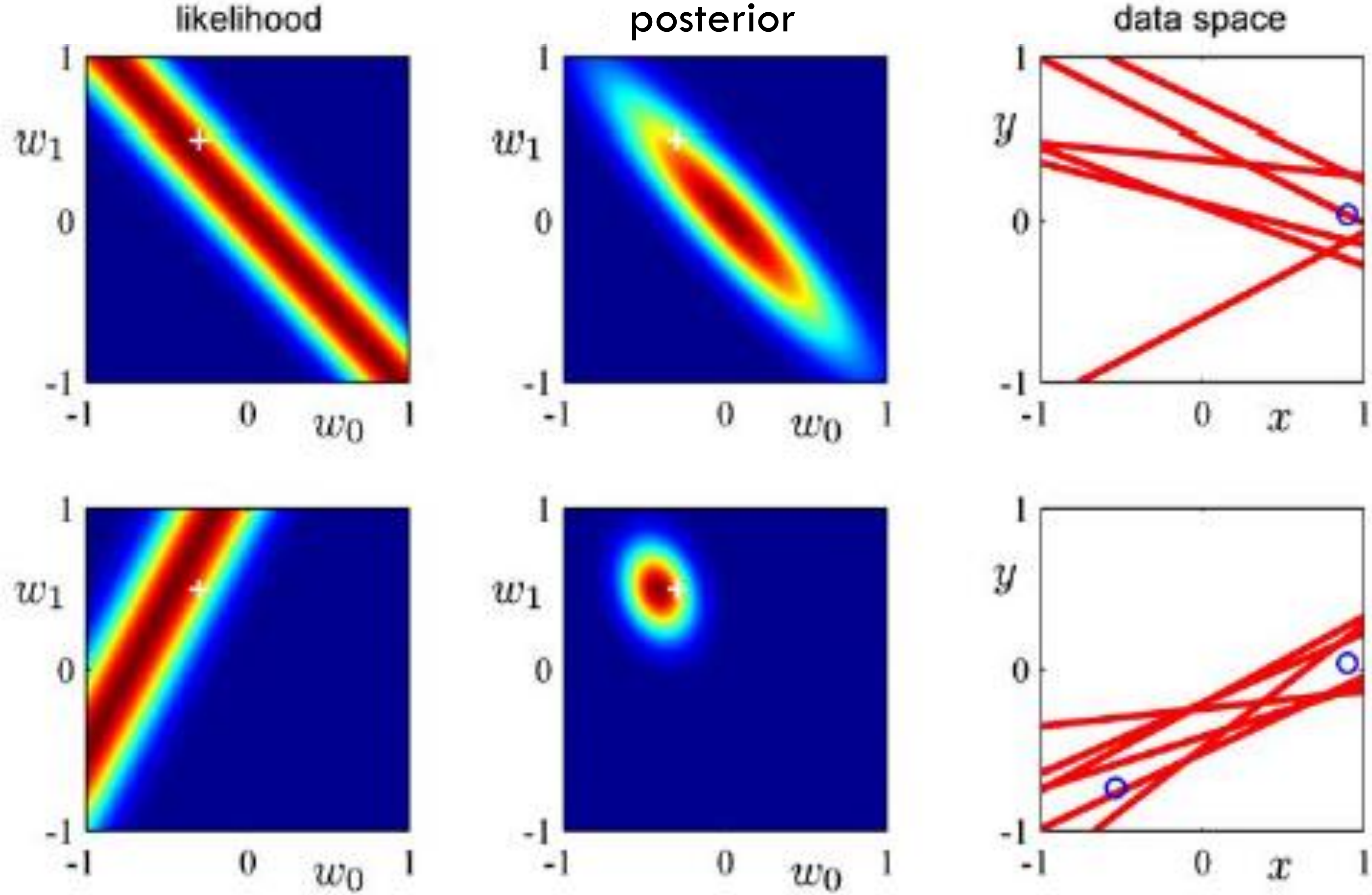
Duda, “Pattern Recognition“

Bayesian Regression

$p(\boldsymbol{X}, \boldsymbol{r} | \boldsymbol{w})$ $\quad$ $p(\boldsymbol{w} | \boldsymbol{X}, \boldsymbol{r}) \propto p(\boldsymbol{X}, \boldsymbol{r} | \boldsymbol{w}) p(\boldsymbol{w})$

N=1

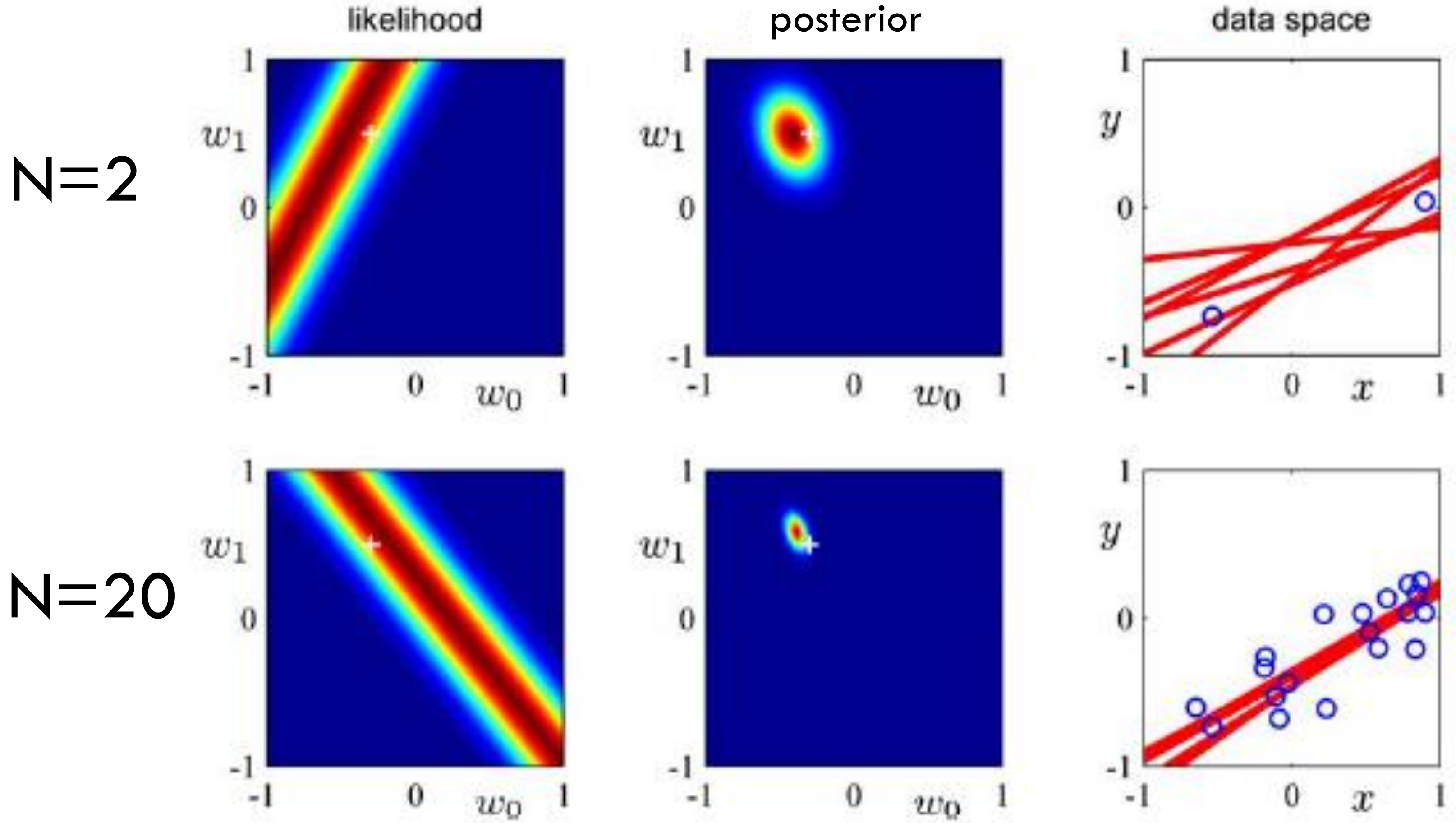
N=2



Duda, “Pattern Recognition“

Bayesian Regression

$p(\boldsymbol{X}, \boldsymbol{r}|\boldsymbol{w}) \quad p(\boldsymbol{w}|\boldsymbol{X}, \boldsymbol{r}) \propto p(\boldsymbol{X}, \boldsymbol{r}|\boldsymbol{w})p(\boldsymbol{w})$



Duda, “Pattern Recognition“

Bayesian Regression

- Line with Gaussian error

$$r_n = f(\boldsymbol{x}_n) = \boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}_n + \epsilon, \quad \epsilon \sim \mathcal{N}(0, 1/\beta), \; \boldsymbol{w} = (w_0, w_1)^{\mathrm{T}}$$

- Assume samples $\boldsymbol{x}_n, r_n, \; n = 1, \ldots, N$

$$p(r_n | \boldsymbol{x}_n, \boldsymbol{w}, \beta) \sim \mathcal{N}(\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}_n, 1/\beta)$$

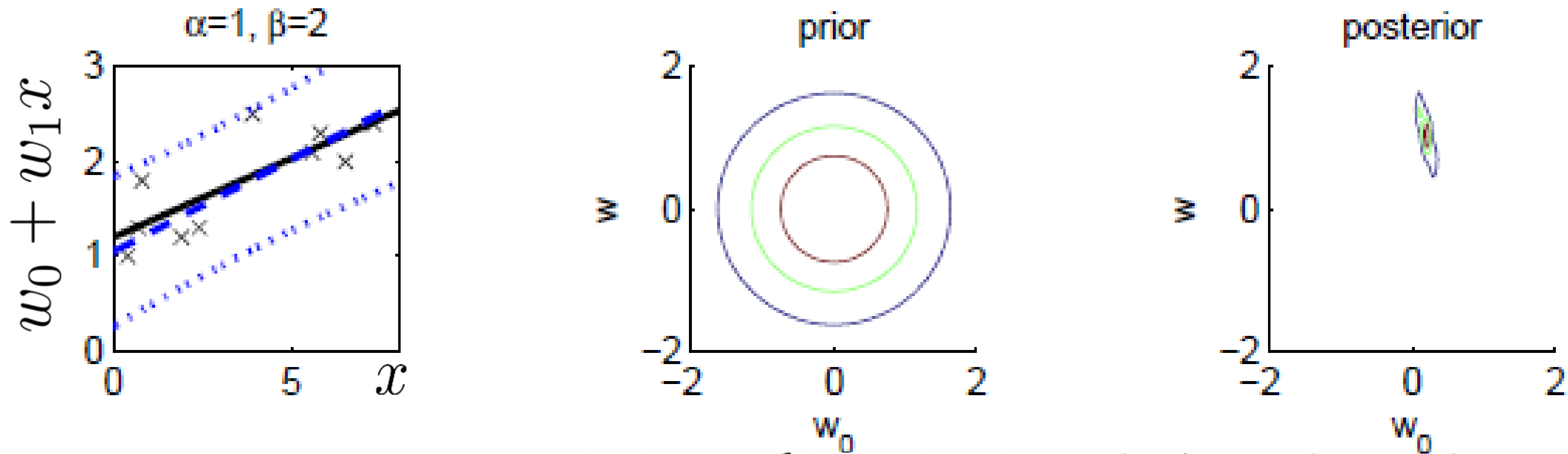
- ML and MAP estimate

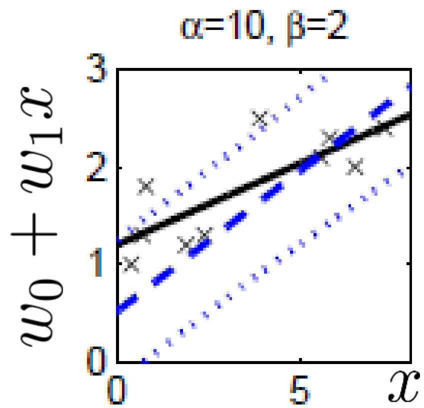
$$\boldsymbol{w}_{\mathrm{ML}} = (\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X})^{-1}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{r} \quad \boldsymbol{w}_{\mathrm{MAP}} = \beta((\alpha\boldsymbol{I} + \beta\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X})^{-1})\boldsymbol{X}^{\mathrm{T}}\boldsymbol{r}$$

- Assume prior $p(\boldsymbol{w}) \sim \mathcal{N}(\boldsymbol{0}, 1/\alpha\boldsymbol{I})$
- posterior $p(\boldsymbol{w}|\boldsymbol{X}, \boldsymbol{r}) \sim \mathcal{N}(\boldsymbol{\mu}_N, \boldsymbol{\Sigma}_N)$

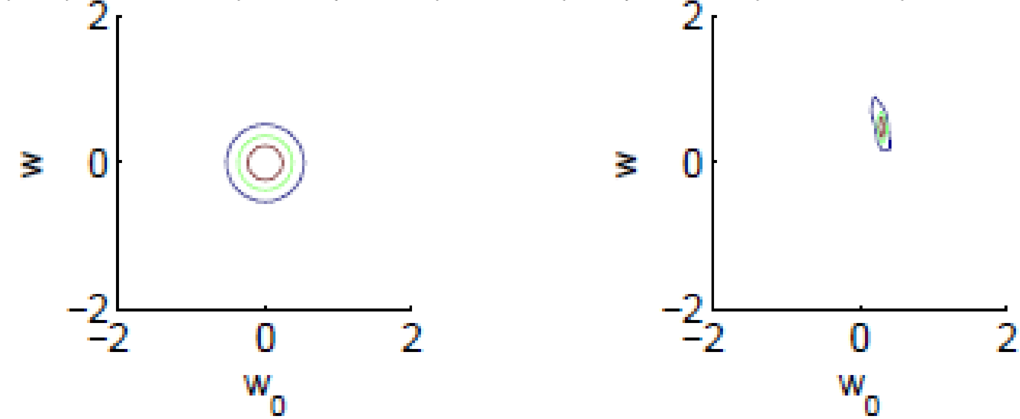
$$\boldsymbol{\mu}_N = \beta\boldsymbol{\Sigma}_N\boldsymbol{X}^{\mathrm{T}}\boldsymbol{r}$$

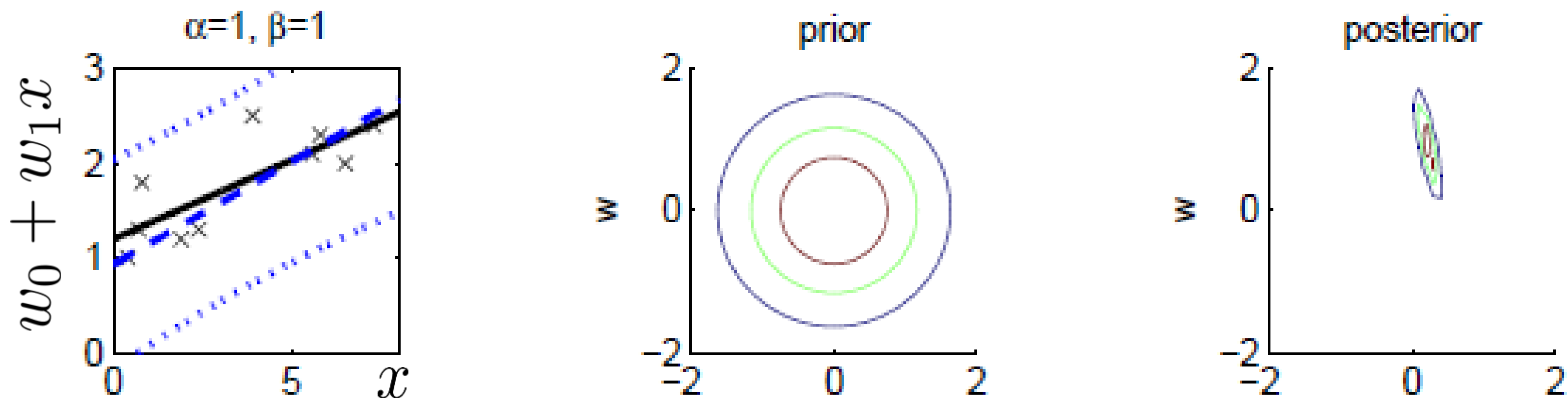
$$\boldsymbol{\Sigma}_N = (\alpha\boldsymbol{I} + \beta\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X})^{-1}$$



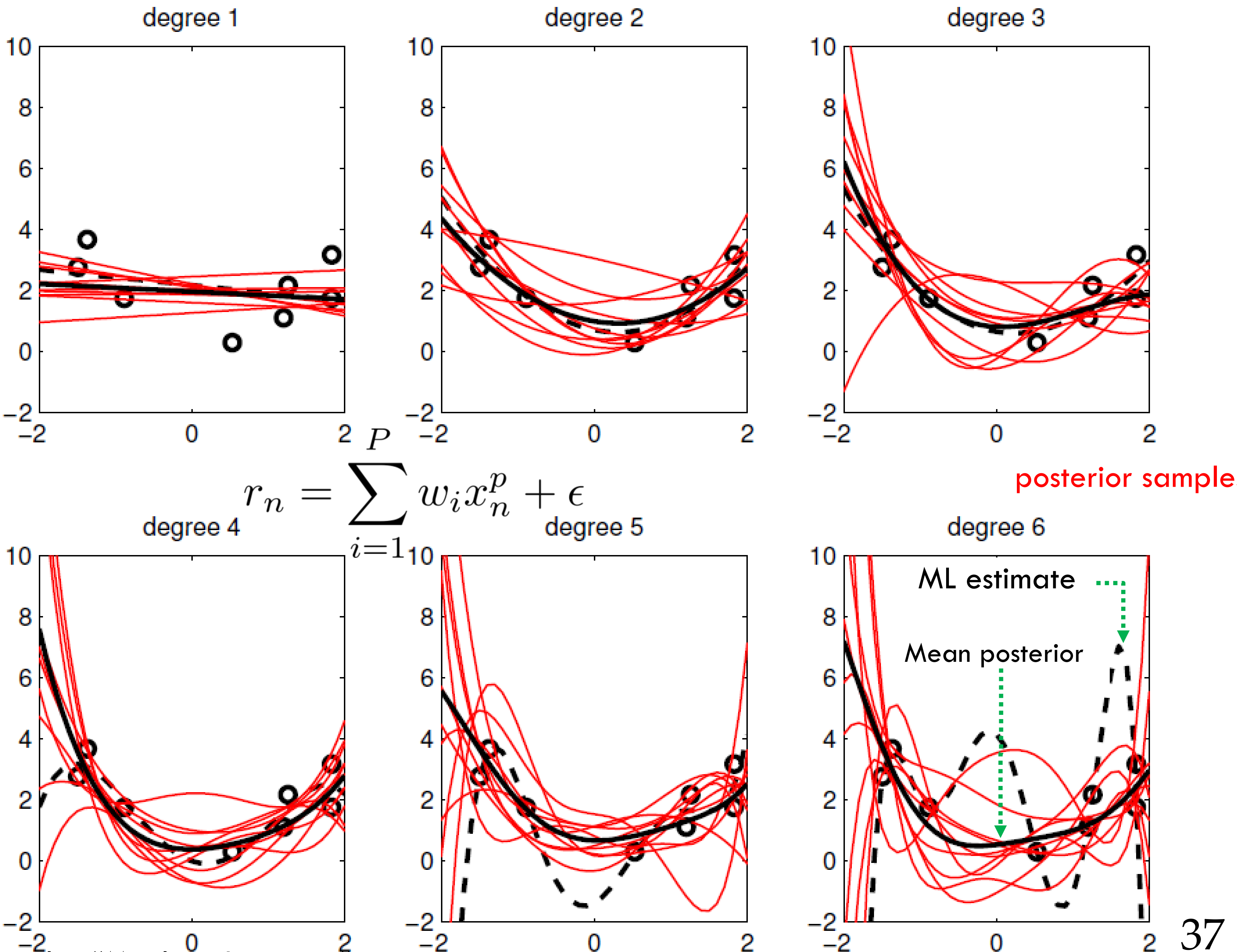


$p(\boldsymbol{w}) \sim \mathcal{N}(\mathbf{0}, 1/\alpha \boldsymbol{I})$ $p(\boldsymbol{w}|\boldsymbol{X}, \boldsymbol{r}) \propto p(\boldsymbol{X}, \boldsymbol{r}|\boldsymbol{w})p(\boldsymbol{w})$





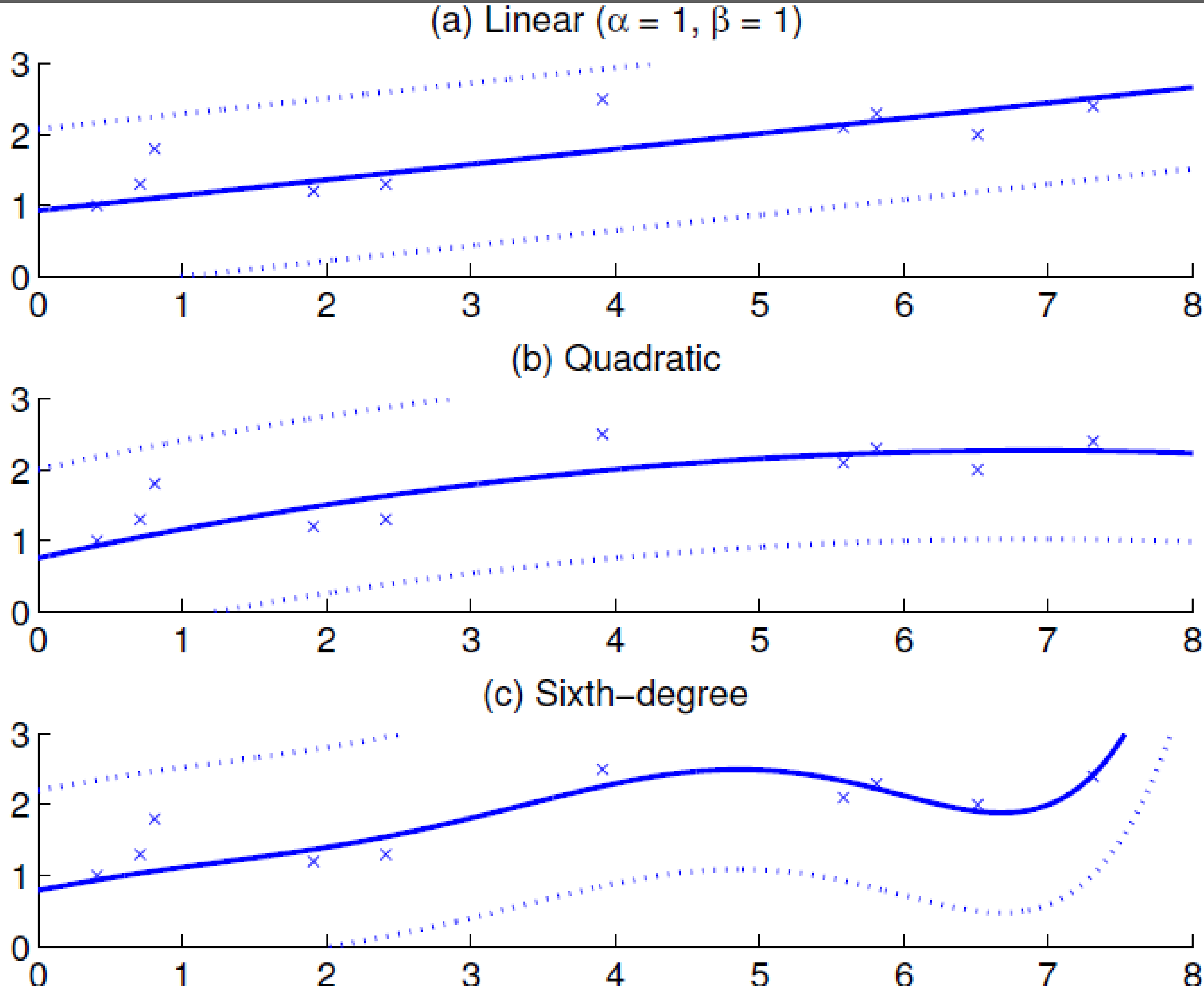
Alpaydin, "Machine Learning"



$$r_n = \sum_{i=1}^{P} w_i x_n^p + \epsilon$$

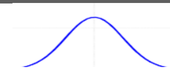
Alpaydin, “Machine Learning”

Kernel Functions

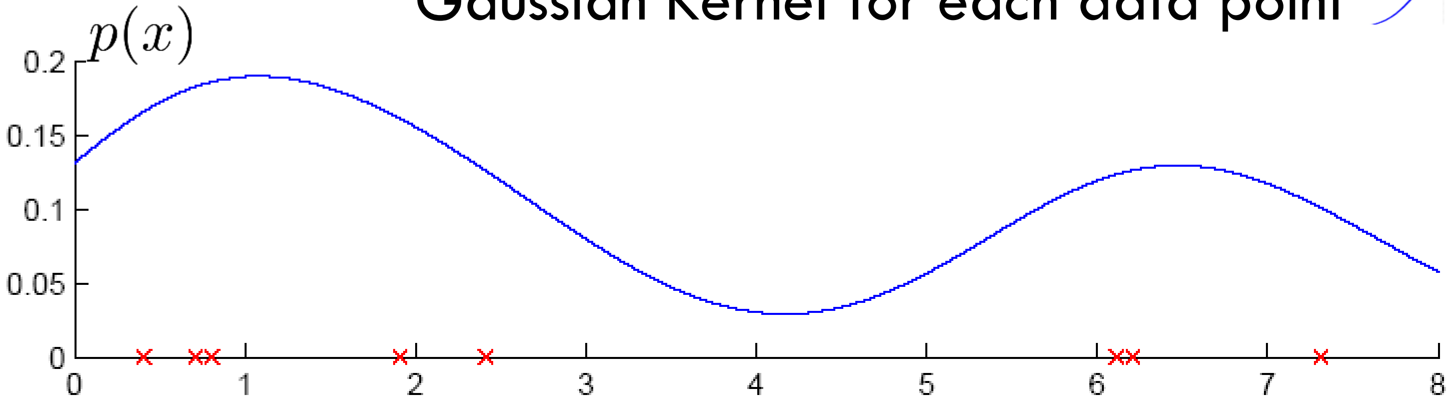


Alpaydin, “Machine Learning“

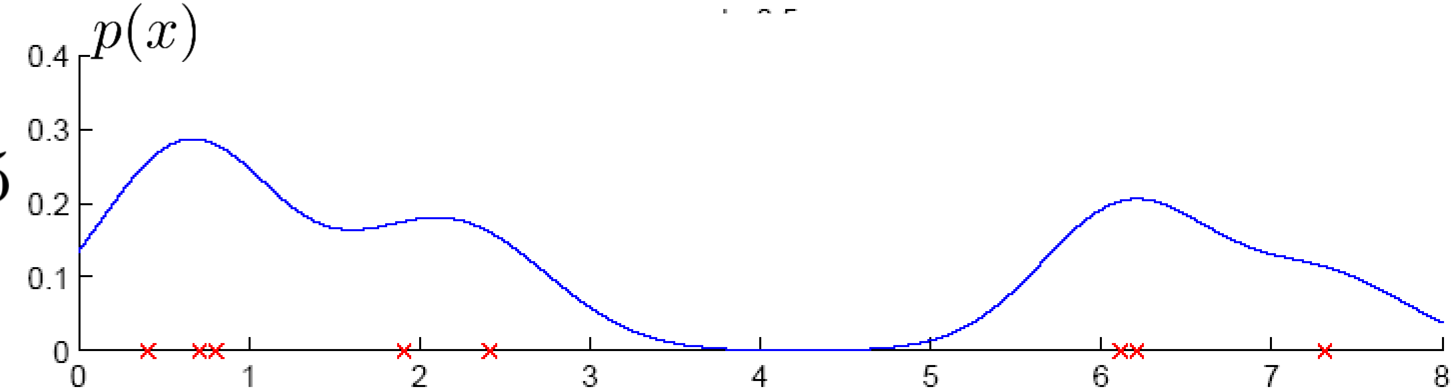
Reminder: Nonparametric Kernel Estimator

Gaussian Kernel for each data point 

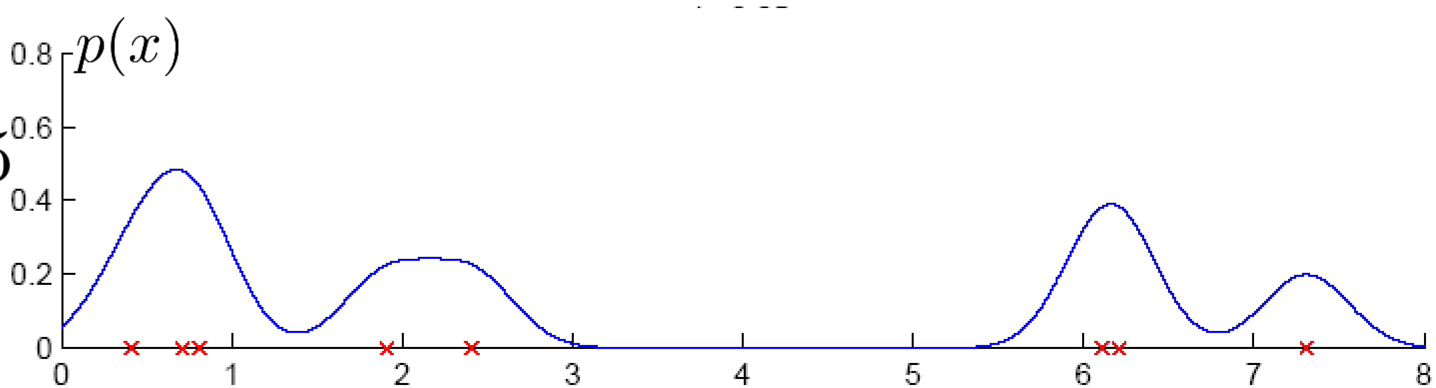
$h = 1$



$h = 0.5$



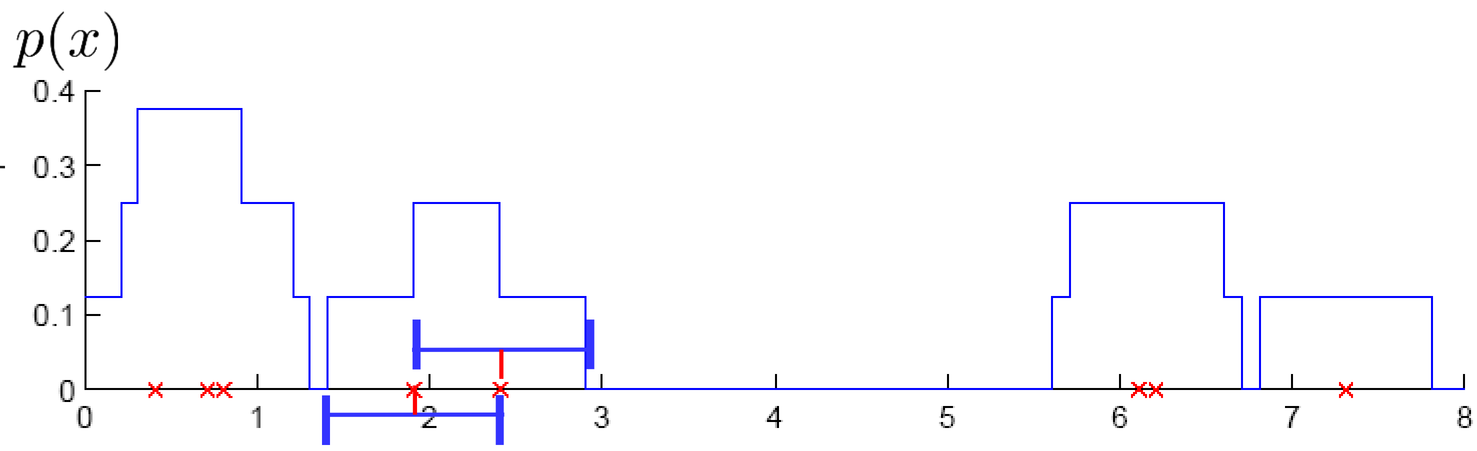
$h = 0.25$



Nonparametric Bayes

- Similar to *k*-NN and Parzen windows: **training set = parameters**
- Model complexity can increases more data (in practice up to *N*, potentially to infinity)

$h = 1$

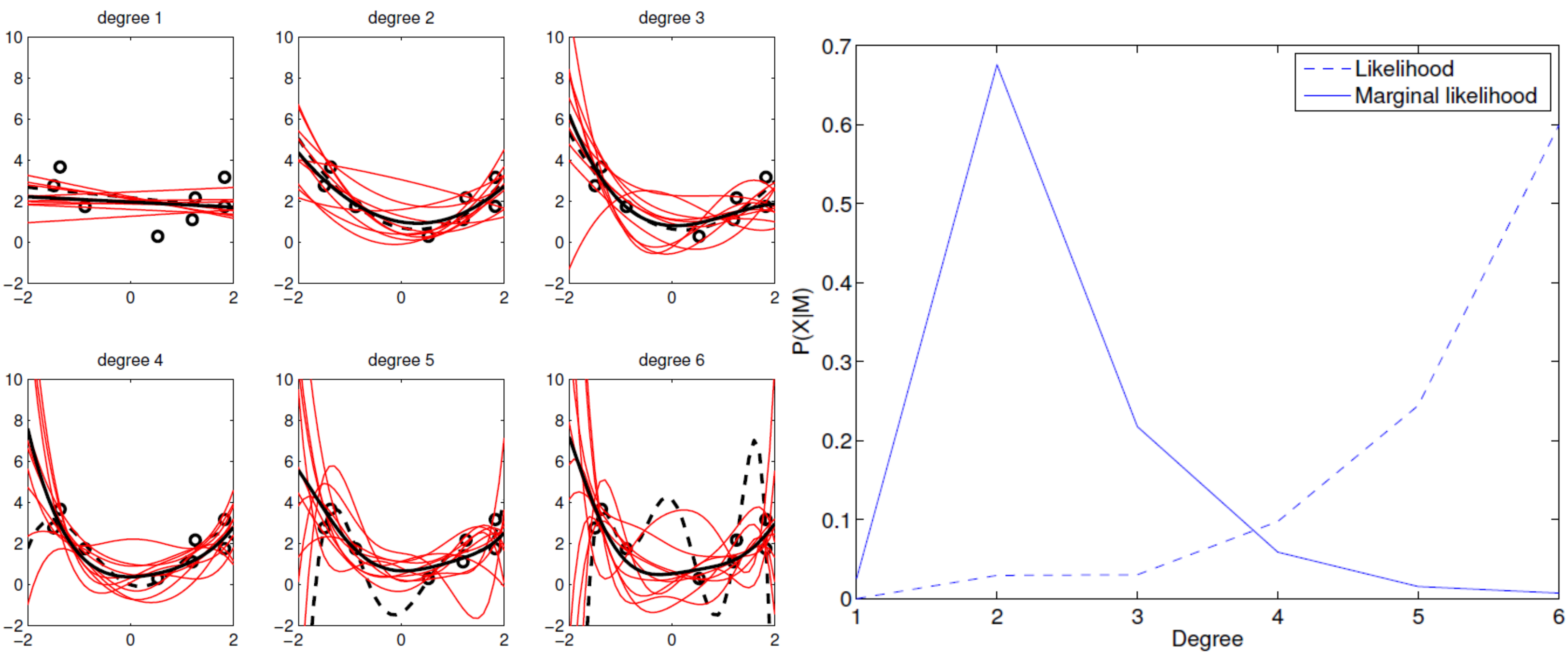


Howto Prior

- Defining a prior is subjective
- Uninformative prior if no prior preference
- Consider prediction
  - Level I $$p(x|\mathcal{X}) = \int p(x|\theta)\overbrace{p(\theta|\mathcal{X})}^{p(\theta|\mathcal{X}) \propto p(\mathcal{X}|\theta)p(\theta)}\, d\theta$$
  - Level II $$p(x|\mathcal{X}) = \int p(x|\theta)p(\theta|\mathcal{X},\alpha)p(\alpha)\, d\alpha$$
  - Level II – ML/Empirical Bayes: Use one good α^*

$$p(x|\mathcal{X}) = \int p(x|\theta)p(\theta|\mathcal{X},\alpha^*)\, d\theta$$

Bayesian Model Comparison



Bayesian Model Comparison

- Marginal likelihood of a model $\mathcal{M}$

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

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

Summary

- ML and MAP give fixed parameter estimates
- Bayes
  - assumes parameter is random variable
  - gives more information
  - is more complex
  - we can sample more data