FUNDAMENTALS OF MACHINE LEARNING

PARAMETRIC METHODS

CSCI3320

Prof. John C.S. Lui, CSE Department, CUHK
Introduction to Machine Learning

From Bayesian to Parametric Methods

In Chapter 3, we wrote posterior probability for class C_i

$$P(C_i|x) = \frac{p(x|C_i)P(C_i)}{p(x)} = \frac{p(x|C_i)P(C_i)}{\sum_{k=1}^{K} p(x|C_k)P(C_k)}$$

Using discriminant function

$$g_i(x) = p(x|C_i)P(C_i) \quad \text{or} \quad g_i(x) = \log p(x|C_i) + \log P(C_i)$$

We know how to get the prior $P(C_i)$,
but we have problem in getting $p(x|C_i)$

In parametric methods, we assume $p(x|C_i)$ follows some probability distributions, so we can use the probability density function of x to represent $p(x|C_i)$. The only question is how to "*estimate*" parameters of that distribution.

Outline

- Previously, we studied how to make optimal decision when uncertainty is modeled by probability (Bayes' rule)
- Focus of **SINGLE** dimension (or feature) first
- Now we learn how to *estimate these probabilities* from the training set
- **Parametric approach**
  - Classification
  - Regression
- Introduce bias/variance dilemma
- Introduce model selection methods

Parametric Estimation

- Assume samples are drawn from some distribution (e.g., Gaussian, Bernoulli,…etc)
- Once we know the parameters (e.g., mean, variance), we know the whole distribution
- **IID samples:** $\boldsymbol{X} = \{x^t\}_{t=1}^{N}$ where $x^t \sim p(x|\theta)$
- **Parametric estimation:**

  Assume a form for $p(x|\theta)$ and estimate θ, its sufficient statistics, using X

  e.g., $N(\mu, \sigma^2)$ where $\theta = \{\mu, \sigma^2\}$
- Like to estimate the parameter θ **as likely as possible**

Maximum Likelihood Estimation

- **Likelihood** of θ given the sample X (due to iid)

$$l(\theta|X) \equiv p(X|\theta) = \prod_{t=1}^{N} p(x^t|\theta)$$

find θ which maximizes the likelihood function $l()$

Given example of flipping a coin N times to estimate p

- **Log likelihood** (sometimes to reduce computation)

$$\mathcal{L}(\theta|X) \equiv \log l\,(\theta|X) = \sum_{t=1}^{N} \log p(x^t|\theta)$$

- **Maximum likelihood estimator (MLE)**

$$\theta^* = \text{argmax}_\theta\ \mathcal{L}(\theta|X)$$

Example: Bernoulli

- Bernoulli: Two states, failure/success, N iid x^t

$$P(x) = p^x(1-p)^{1-x} \qquad x \in \{0,1\}$$

$$\mathcal{L}(p|\mathcal{X}) = \log \prod_{t=1}^{N} p^{(x^t)}(1-p)^{(1-x^t)}$$

$$= \sum_t x^t \log p + \left(N - \sum_t x^t\right) \log(1-p)$$

- To maximize $\mathcal{L}(p|\mathcal{X})$, take derivative with p and equate to 0
- MLE:

$$\hat{p} = \frac{\sum_t x^t}{N}$$

How?

Example: Bernoulli (derivation)

$$\mathcal{L}(p|\mathcal{X}) = \log \prod_{t=1}^{N} p^{(x^t)} (1-p)^{(1-x^t)}$$

$$= \sum_t x^t \log p + \left(N - \sum_t x^t \right) \log(1-p)$$

$$\frac{d\mathcal{L}(p|X)}{dp} = \frac{1}{p} \sum_t x^t - \left(N - \sum_t x^t \right) \frac{1}{1-p} = 0$$

$$p \left(N - \sum_t x^t \right) = (1-p) \left(\sum_t x^t \right)$$

$$pN = \sum_t x^t$$

$$p = \frac{1}{N} \sum_t x^t$$

Explain how to get the MLE of Bernoulli parameter with N points

Application

In Chapter 3, we wrote posterior probability for class C_i

$$P(C_i|x) = \frac{p(x|C_i)P(C_i)}{p(x)} = \frac{p(x|C_i)P(C_i)}{\sum_{k=1}^{K} p(x|C_k)P(C_k)}$$

Using discriminant function

$$g_i(x) = p(x|C_i)P(C_i) \quad \text{or} \quad g_i(x) = \log p(x|C_i) + \log P(C_i)$$

$$g_i(x) = \log\left(\hat{p}^x(1-\hat{p})^{1-x}\right) + \log P(C_i)$$

Now we remove the limitation of Bayesian method

Application

Values	Class #
1	1
0	1
0	2
1	1
1	2
.....	
1	?

- *N* training data
- Value indicates whether the person is high risk (1) or low risk (0)
- Person can belong to class 1 or class 2

Example: Multinomial

- Outcome of the event is one of *K* mutually exclusive and exhaustive states (or classes) with p_i with $\sum_{i=1}^{K} p_i = 1.$
- Let $x_1, x_2, \ldots, x_K$ are the indicator variables
- x_i is *1* if the outcome is state (class) *i* and *0* otherwise
- We have $P(x_1, x_2, \ldots, x_K) = \prod_{i=1}^{K} p_i^{x_i}$
- *N* independent experiments:

$$\mathcal{X} = \{\boldsymbol{x}^t\}_{t=1}^{N}$$ **Note**: $\boldsymbol{x}^t$ *is a vector !!!*

$$x_i^t = \begin{cases} 1 & \text{if experiment } t \text{ chooses state } i \\ 0 & \text{otherwise} \end{cases} \qquad \text{with } \sum_i x_i^t = 1.$$

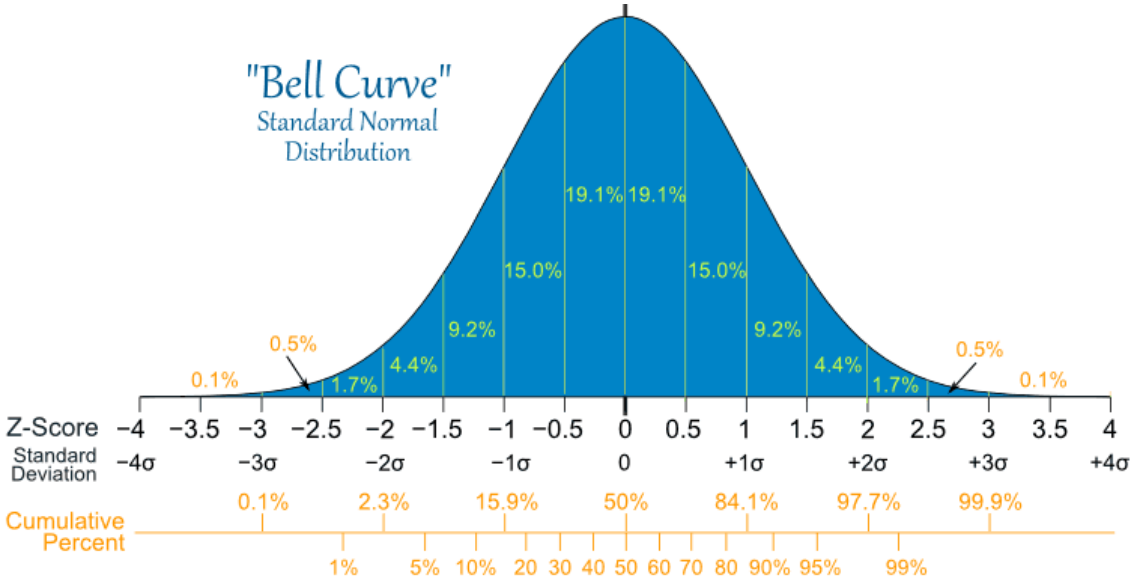
- **MLE:** $\hat{p}_i = \frac{\sum_t x_i^t}{N}$

Application

Values	Class #
1	1
0	1
0	2
2	1
2	2
.....	
2	?

- N training data
- Value indicates whether the person is high risk (1), low risk (0), or neutral (2)
- Person can belong to class 1 or class 2

Gaussian (Normal) Distribution



- $p(x) : \mathcal{N}(\mu, \sigma^2)$

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

- MLE for μ and σ^2:

Log likelihood function

$$\mathcal{L}(\mu, \sigma | \mathcal{X}) = -\frac{N}{2}\log(2\pi) - N\log\sigma - \frac{\sum_t (x^t - \mu)^2}{2\sigma^2}$$

Show this in class

$$m = \frac{\sum_t x^t}{N}$$

$$s^2 = \frac{\sum_t (x^t - m)^2}{N}$$

Gaussian (Normal) Distribution

Log likelihood function: $\mathcal{L}(\mu, \sigma|\mathcal{X}) = -\frac{N}{2}\log(2\pi) - N\log\sigma - \frac{\sum_t (x^t - \mu)^2}{2\sigma^2}$

$$\frac{d\mathcal{L}(\mu,\sigma|\mathcal{X})}{d\mu} = \frac{1}{2\sigma^2}(2)\sum_t (x^t - \mu) = 0$$

$$N\mu = \sum_t x^t$$

$$\mu = \frac{\sum_t x^t}{N}$$

$$\frac{d\mathcal{L}(\mu,\sigma|\mathcal{X})}{d\sigma} = -\frac{N}{\sigma} + 2\frac{\sum_t (x^t - \mu)^2}{2\sigma^3} = 0$$

$$N\sigma^2 = \sum_t (x^t - \mu)^2$$

$$\sigma^2 = \frac{\sum_t (x^t - \mu)^2}{N}$$

$$m = \frac{\sum_t x^t}{N}$$

$$s^2 = \frac{\sum_t (x^t - m)^2}{N}$$

Parametric Classification

In Chapter 3, we wrote posterior probability for class C_i

$$P(C_i|x) = \frac{p(x|C_i)P(C_i)}{p(x)} = \frac{p(x|C_i)P(C_i)}{\sum_{k=1}^{K} p(x|C_k)P(C_k)}$$

Using discriminant function

$$g_i(x) = p(x|C_i)P(C_i) \quad \text{or} \quad g_i(x) = \log p(x|C_i) + \log P(C_i)$$

If we can assume that $p(x|C_i)$ are Gaussian

$$p(x|C_i) = \frac{1}{\sqrt{2\pi}\sigma_i} \exp\left[-\frac{(x-\mu_i)^2}{2\sigma_i^2}\right] \qquad \sigma_1, \mu_i, P(C_i) = ?$$

$$g_i(x) = -\frac{1}{2}\log 2\pi - \log \sigma_i - \frac{(x-\mu_i)^2}{2\sigma_i^2} + \log P(C_i)$$

Classification of K *cars*

- Given the sample $\mathcal{X} = \{x^t, \boldsymbol{r}^t\}_{t=1}^N$

$$x \in \Re \qquad \boldsymbol{r} \in \{0,1\}^K \text{ such that } \quad r_i^t = \begin{cases} 1 & \text{if } \boldsymbol{x}^t \in C_i \\ 0 & \text{if } \boldsymbol{x}^t \in C_k, k \neq i \end{cases}$$

- Want *posterior probability P(C$_i$|x), use Bayes Rule*
- **ML estimates** are

$$\hat{P}(C_i) = \frac{\sum_t r_i^t}{N} \qquad m_i = \frac{\sum_t x^t r_i^t}{\sum_t r_i^t} \qquad s_i^2 = \frac{\sum_t (x^t - m_i)^2 r_i^t}{\sum_t r_i^t}$$

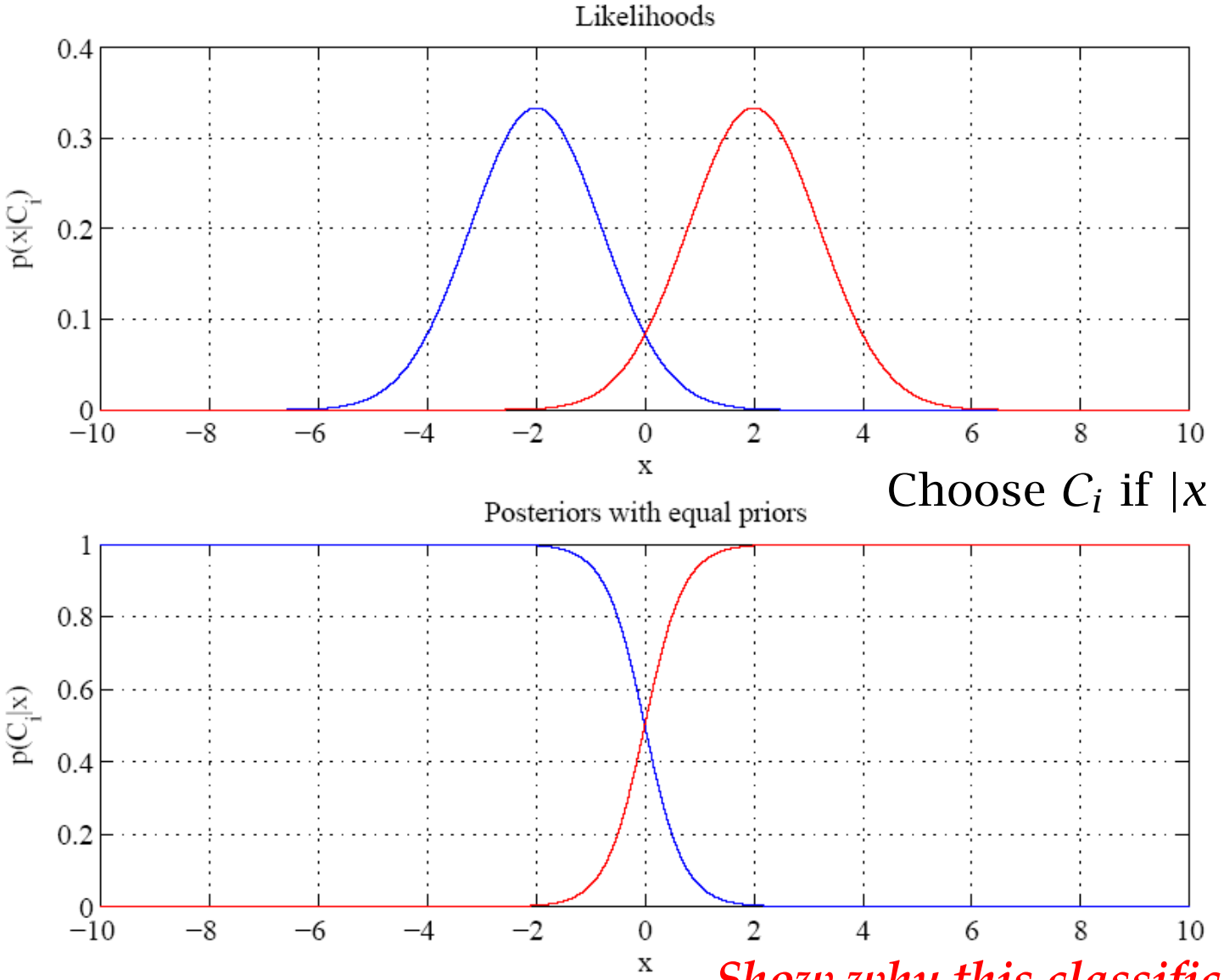
- Discriminant

$$g_i(x) = -\frac{1}{2}\log 2\pi - \log s_i - \frac{(x - m_i)^2}{2s_i^2} + \log \hat{P}(C_i)$$

- If prior probability and variances for **ALL CLASSES ARE THE SAME:** $g_i(x) = -(x - m_i)^2$

What is the physical meaning?

When prior ($P(C_i)$)and variance (s_i^2) are equal



$$g_i(x) = -(x - m_i)^2$$

Choose C_i if $|x - m_i| = \min_k |x - m_k|$

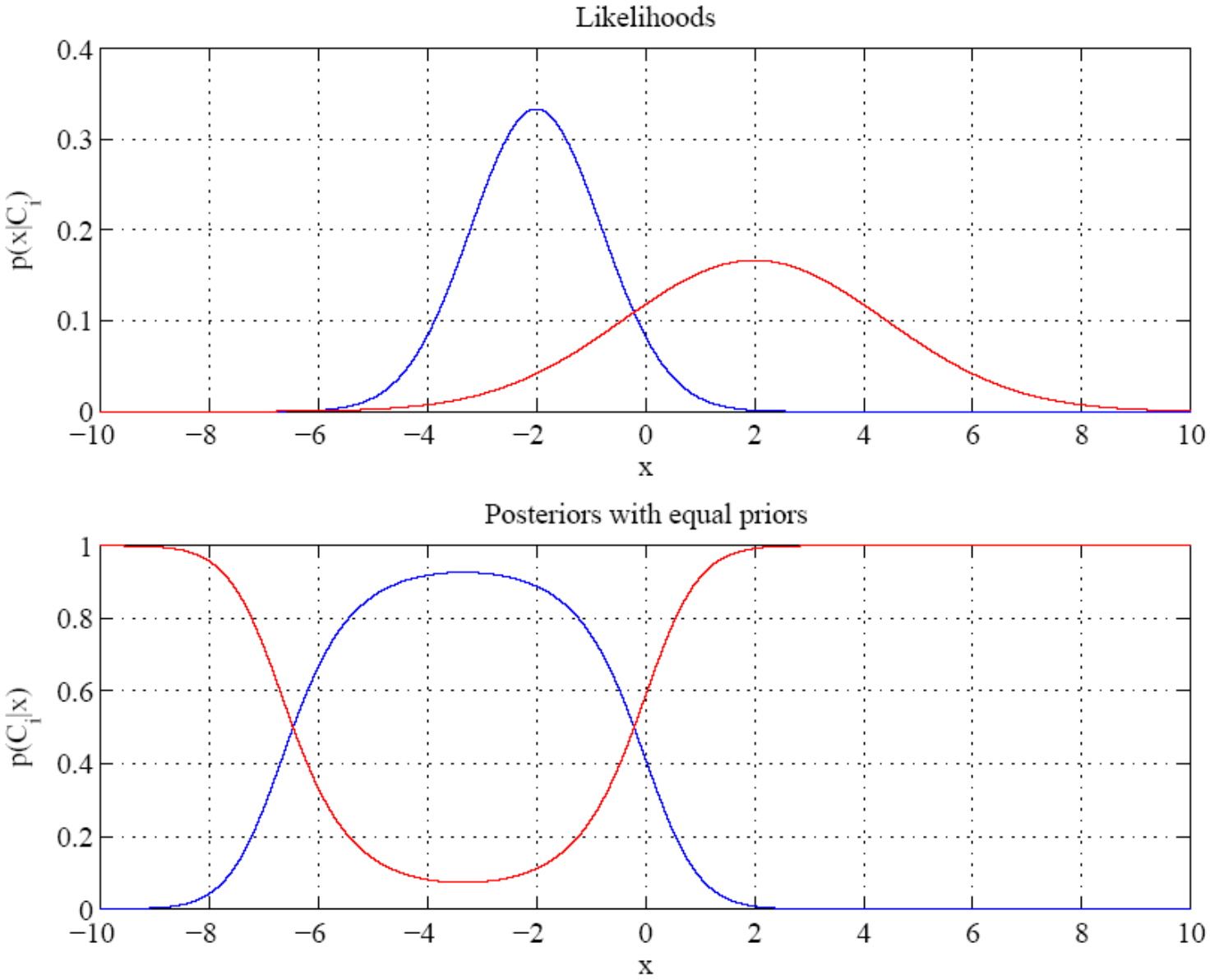
Single boundary at halfway between means, or when $g_1(x)=g_2(x)$

Show derivation

$$x = \frac{(m_1+m_2)}{2}$$

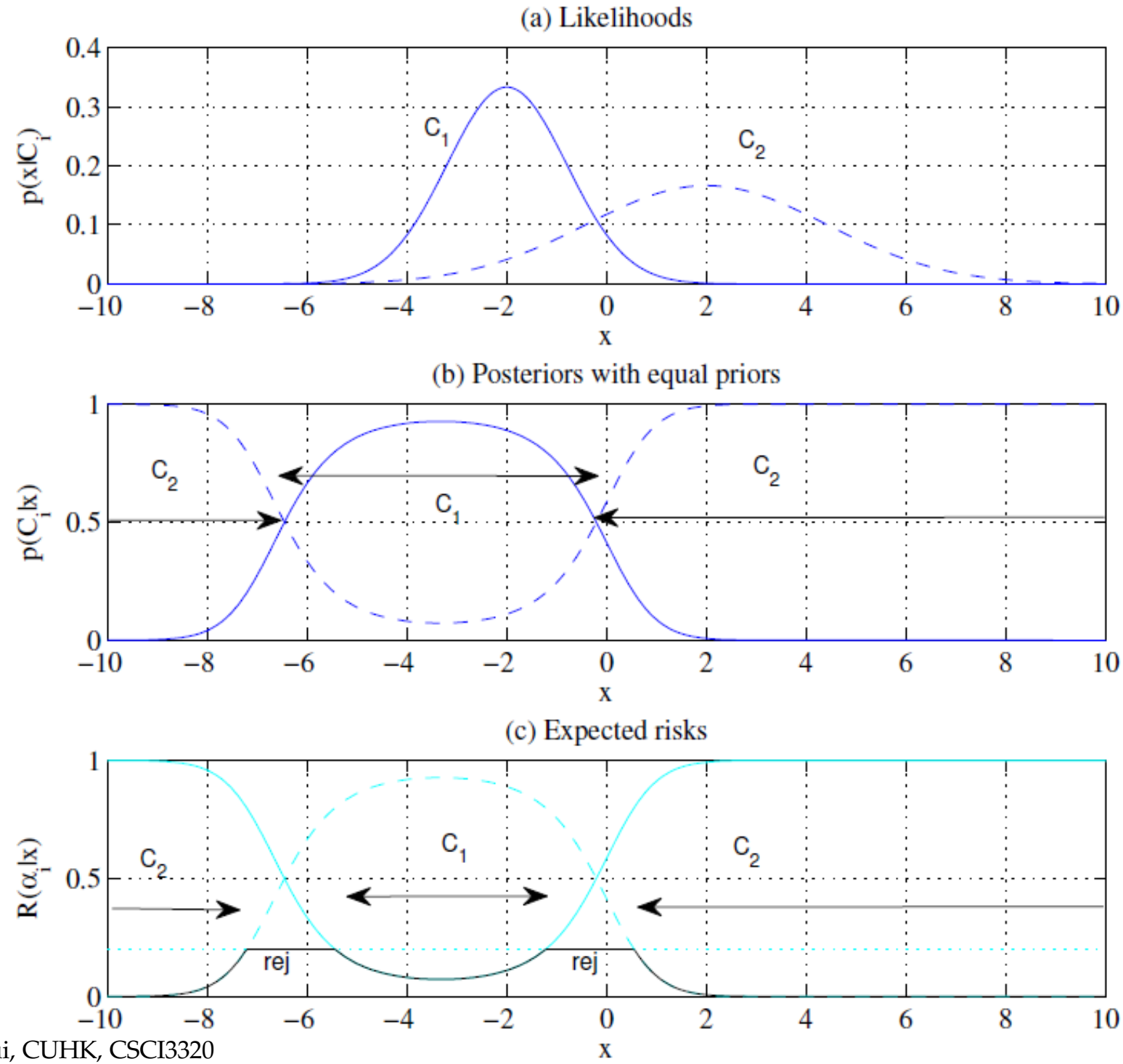
Show why this classification is intuitive

When variance (s_i^2) are different



Two boundaries.

Homework !!!


(a) Likelihoods
p(x|C_i)
C_1
C_2
x
(b) Posteriors with equal priors
p(C_i|x)
C_2
C_1
C_2
x
(c) Expected risks
R(α_i|x)
C_2
C_1
C_2
rej
rej
x

More comment:

- If we have prior information about the means, we can use a Bayesian estimate of $p(x|C_i)$ with prior on μ_i. We will cover this in later chapter
- When x is continuous, don't assume it is Gaussian
- Use the data and test whether it is Gaussian
  - Via plotting to visually check (**remember QQ plot?**)
  - Use ***normality test*** in statistics
- Summary of current "***likelihood approach***" in classification
  - **Use data to estimate densities separately**
  - **Calculate posterior densities with Bayes' rules**
  - **Derive discriminant**

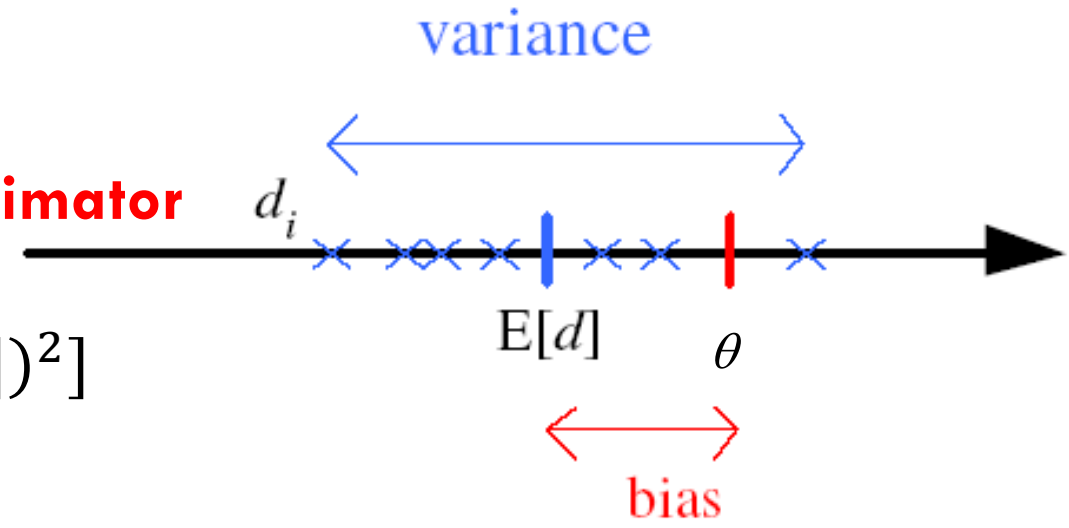
Evaluating an Estimator: Bias and Variance

Unknown parameter θ

Estimator $d_i = h(X_i)$ on sample X_i

Bias: $b_{\theta(d)} = E[d] - \theta$

when $b_{\theta(d)} = 0$, **unbiased estimator**



Variance of RV d: $E[(d - E[d])^2]$

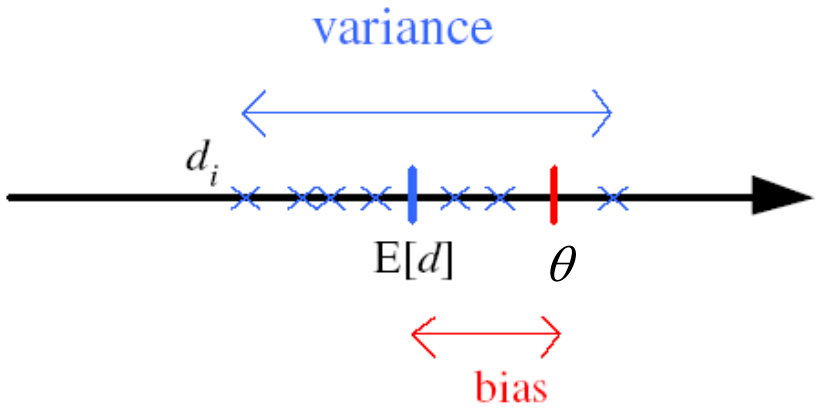
Mean square error:

$$r(d,\theta) = E[(d-\theta)^2]$$
$$= (E[d]-\theta)^2 + E[(d-E[d])^2]$$
$$= \text{Bias}^2 + \text{Variance}$$

Evaluating an Estimator: Bias and Variance

Mean square error:

$$r(d,\theta) = E[(d-\theta)^2]$$
$$= (E[d]-\theta)^2 + E[(d-E[d])^2]$$
$$= \text{Bias}^2 + \text{Variance}$$



$$\begin{aligned}
r(d,\theta) &= E\left[(d-\theta)^2\right] \\
&= E\left[(d-E[d]+E[d]-\theta)^2\right] \\
&= E\left[(d-E[d])^2 + (E[d]-\theta)^2 + 2(E[d]-\theta)(d-E[d])\right] \\
&= E\left[(d-E[d])^2\right] + E\left[(E[d]-\theta)^2\right] + 2E\left[(E[d]-\theta)(d-E[d])\right] \\
&= E\left[(d-E[d])^2\right] + (E[d]-\theta)^2 + 2(E[d]-\theta)E[d-E[d]] \\
&= \underbrace{E\left[(d-E[d])^2\right]}_{variance} + \underbrace{(E[d]-\theta)^2}_{bias^2}
\end{aligned}$$

$E[d-E[d]] = 0$

Explain physical meaning

Evaluating an Estimator: Unbiased Estimator

x^t drawn from some density with mean μ

sample average, m, is an unbiased estimator

$$E[m] = E\left[\frac{\sum_t x^t}{N}\right] = \frac{1}{N}\sum_t E[x^t] = \frac{N\mu}{N} = \mu$$ **meaning**

m is also a *consistent* estimator, that is, $\text{Var}(m) \to 0$ as $N \to \infty$.

$$\text{Var}(m) = \text{Var}\left(\frac{\sum_t x^t}{N}\right) = \frac{1}{N^2}\sum_t \text{Var}(x^t) = \frac{N\sigma^2}{N^2} = \frac{\sigma^2}{N}$$ **meaning**

Evaluating an Estimator: Unbiased Estimator

s^2, the MLE of σ^2:

$$s^2 = \frac{\sum_t (x^t - m)^2}{N} = \frac{\sum_t (x^t)^2 - Nm^2}{N}$$

show in class

$$E[s^2] = \frac{\sum_t E[(x^t)^2] - N \cdot E[m^2]}{N}$$

Given that $\mathrm{Var}(X) = E[X^2] - E[X]^2$, we get $E[X^2] = \mathrm{Var}(X) + E[X]^2$,

show in class

$$E[(x^t)^2] = \sigma^2 + \mu^2 \text{ and } E[m^2] = \sigma^2/N + \mu^2$$

plugging these in, we get

$$E[s^2] = \frac{N(\sigma^2 + \mu^2) - N(\sigma^2/N + \mu^2)}{N} = \left(\frac{N-1}{N}\right)\sigma^2 \neq \sigma^2$$

s^2 is a biased estimator of σ^2

But an asymptotically unbiased estimator

Unbiased estimator: $(N/(N-1))s^2$

Bayes' Estimator (my notes)

- Sometimes, we know the possible value range of θ
- Treat θ as a random variable with prior $p(\theta)$
- Bayes' rule: $p(\theta|\mathcal{X}) = \dfrac{p(\mathcal{X}|\theta)p(\theta)}{p(\mathcal{X})}$
- To estimate density at x, we have

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

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

Bayes' Estimator (my note)

- **Maximum a Posteriori (MAP):** $\theta_{MAP} = \arg\max_{\theta} p(\theta|\mathcal{X})$

Thus, we have $p(x|\mathcal{X}) \quad = \quad p(x|\theta_{MAP})$

- If no prior values of θ, prior density $p(\theta)$ is flat, MAP estimate is equivalent to maximum likelihood estimate
- **Maximum Likelihood (ML):** $\theta_{ML} = \arg\max_{\theta} p(\mathcal{X}|\theta)$
- **Bayes' estimator** (or expected value of the posterior probability):

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

Bayes' Estimator: Example

$x^t \sim \mathcal{N}(\theta, \sigma^2)$ and $\theta \sim \mathcal{N}(\mu_0, \sigma_0^2)$, μ_0, σ_0^2, and σ^2 are known

$$p(\mathcal{X}|\theta) = \frac{1}{(2\pi)^{N/2}\sigma^N} \exp\left[-\frac{\sum_t (x^t - \theta)^2}{2\sigma^2}\right]$$

$$p(\theta) = \frac{1}{\sqrt{2\pi}\sigma_0} \exp\left[-\frac{(\theta - \mu_0)^2}{2\sigma_0^2}\right]$$

reduce the posterior density to a single pt

It can be shown that $p(\theta|\mathcal{X})$ is normal with

$$E[\theta|\mathcal{X}] = \frac{N/\sigma^2}{N/\sigma^2 + 1/\sigma_0^2} m + \frac{1/\sigma_0^2}{N/\sigma^2 + 1/\sigma_0^2} \mu_0$$

Thus the Bayes' estimator is a weighted average of the prior mean μ_0 and the sample mean m, with weights being inversely proportional to their variances. As the sample size N increases, the Bayes' estimator gets closer to the sample average, using more the information provided by the sample. When σ_0^2 is small, that is, when we have little prior uncertainty regarding the correct value of θ, or when N is small, our prior guess μ_0 has a higher effect.

Example of Bayes' Estimator

Let $(X_1, X_2, ..., X_n)$ be the random sample of a Bernoulli R.V. X with

$$P(x) = p^x(1-p)^{1-x} \qquad x = 0, 1.$$

p is unknown but it is uniformly distributed on $(0, 1)$, or

$$f(p) = 1 \qquad 0 < p < 1.$$

The posterior pdf of p is: $f(p|x_1, ..., x_n) = \dfrac{f(x_1, ..., x_n, p)}{P(x_1, .., x_n)}$

$$\begin{aligned} f(x_1, ..., x_n, p) &= f(x_1, ..., x_n|p)f(p) \\ &= p^{\sum_{i=1}^n x_i}(1-p)^{n-\sum_{i=1}^n x_i} = p^m(1-p)^{n-m} \times 1 \end{aligned}$$

where $m = \sum_{i=1}^n x_i$

Example of Bayes' Estimator

$$P(x_1, .., x_n) \quad = \quad \int_0^1 f(x_1, ..., x_n, p)dp = \int_0^1 p^m(1-p)^{n-m}dp$$

From calculus, we have: $\int_0^1 p^m(1-p)^k dp = \frac{m!k!}{(m+k+1)!}$

So: $f(p|x_1, ..., x_n) = \frac{f(x_1,..,x_n,p)}{P(x_1,..,x_n)} = \frac{(n+1)!p^m(1-p)^{n-m}}{m!(n-m)!}$

Example of Bayes' Estimator

Bayes' estimator $E[p|\mathcal{X}]$:

$$\begin{aligned} E[p|\mathcal{X}] &= \int_0^1 p f(p|x_1,..,x_n)dp \\ &= \frac{(n+1)!}{m!(n-m)!}\int_0^1 p^{(m+1)}(1-p)^{n-m}dp \\ &= \frac{(n+1)!}{m!(n-m)!}\frac{(m+1)!(n-m)!}{(n+2)!} = \frac{m+1}{n+2} \\ &= \frac{1}{n+2}\left(\sum_{i=1}^{n} x_i + 1\right) \end{aligned}$$

Bayes' estimator p_B is $\frac{1}{n+2}\left(\sum_{i=1}^{n} x_i + 1\right)$

Regression

r: dependent variable
x: independent variable

unknown function *f()*

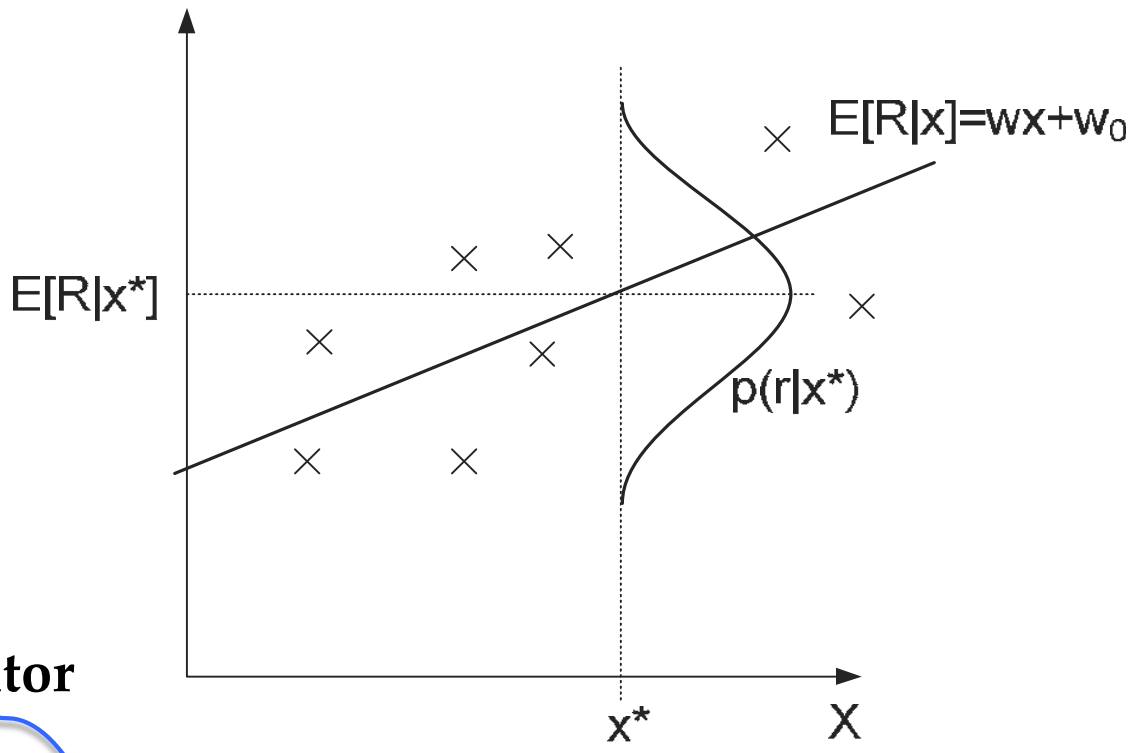
noise

$$r = f(x) + \varepsilon$$

$$\text{estimator} : g(x|\theta)$$

$$\varepsilon \sim \mathcal{N}(0, \sigma^2)$$

$$p(r|x) \sim \mathcal{N}(g(x|\theta), \sigma^2)$$



log maximum likelihood estimator

$$\mathcal{L}(\theta|\mathcal{X}) = \log \prod_{t=1}^{N} p(x^t, r^t)$$

can be ignored

$$= \log \prod_{t=1}^{N} p(r^t|x^t) + \log \prod_{t=1}^{N} p(x^t)$$

Regression: From LogL to Error

$$\begin{aligned}
\mathcal{L}(\theta|\mathcal{X}) &= \log \prod_{t=1}^{N} \frac{1}{\sqrt{2\pi}\sigma} \exp\left[-\frac{[r^t - g(x^t|\theta)]^2}{2\sigma^2}\right] \\
&= \log\left(\frac{1}{\sqrt{2\pi}\sigma}\right)^N \exp\left[-\frac{1}{2\sigma^2}\sum_{t=1}^{N}[r^t - g(x^t|\theta)]^2\right] \\
&= -N\log(\sqrt{2\pi}\sigma) - \frac{1}{2\sigma^2}\sum_{t=1}^{N}[r^t - g(x^t|\theta)]^2
\end{aligned}$$

First term is not a function of θ, maximize the above term is equivalent to **minimize** the following error expression (E) :

$$E(\theta|\mathcal{X}) = \frac{1}{2}\sum_{t=1}^{N}[r^t - g(x^t|\theta)]^2$$

θ that minimize this function is the ***least square estimates***

Example: Linear Regression

$$g(x^t|w_1, w_0) = w_1 x^t + w_0$$

Using the error function in previous slide, taking derivatives for two unknowns, we have:

$$\begin{aligned} \sum_t r^t &= N w_0 + w_1 \sum_t x^t \\ \sum_t r^t x^t &= w_0 \sum_t x_t + w_1 \sum_t (x^t)^2 \end{aligned}$$

vector-matrix form as $\mathbf{A}\boldsymbol{w} = \boldsymbol{y}$

$$\mathbf{A} = \begin{bmatrix} N & \sum_t x^t \\ \sum_t x^t & \sum_t (x^t)^2 \end{bmatrix}, \quad \boldsymbol{w} = \begin{bmatrix} w_0 \\ w_1 \end{bmatrix}, \quad \boldsymbol{y} = \begin{bmatrix} \sum_t r^t \\ \sum_t r^t x^t \end{bmatrix}$$

2 x 2 matrix *2 x 1 vector* *2 x 1 vector*

solved as $\boldsymbol{w} = \mathbf{A}^{-1}\boldsymbol{y}$.

Example: Polynomial Regression of order k

$$g(x^t|w_k,\ldots,w_2,w_1,w_0) = w_k(x^t)^k + \cdots + w_2(x^t)^2 + w_1x^t + w_0$$

Taking (*k+1*) derivatives of *E*, we have (*k+1*) equations :

$\mathbf{A}\bar{\boldsymbol{w}} = \boldsymbol{y}$ where

$$\mathbf{A} = \begin{bmatrix} N & \sum_t x^t & \sum_t (x^t)^2 & \cdots & \sum_t (x^t)^k \\ \sum_t x^t & \sum_t (x^t)^2 & \sum_t (x^t)^3 & \cdots & \sum_t (x^t)^{k+1} \\ \vdots & & & & \\ \sum_t (x^t)^k & \sum_t (x^t)^{k+1} & \sum_t (x^t)^{k+2} & \cdots & \sum_t (x^t)^{2k} \end{bmatrix}$$

$$\boldsymbol{w} = \begin{bmatrix} w_0 \\ w_1 \\ w_2 \\ \vdots \\ w_k \end{bmatrix}, \quad \boldsymbol{y} = \begin{bmatrix} \sum_t r^t \\ \sum_t r^t x^t \\ \sum_t r^t (x^t)^2 \\ \vdots \\ \sum_t r^t (x^t)^k \end{bmatrix}$$

Example: Polynomial Regression of order k

We can write $\mathbf{A} = \mathbf{D}^T\mathbf{D}$ and $\boldsymbol{y} = \mathbf{D}^T\boldsymbol{r}$ where

$$\mathbf{D} = \begin{bmatrix} 1 & x^1 & (x^1)^2 & \cdots & (x^1)^k \\ 1 & x^2 & (x^2)^2 & \cdots & (x^2)^k \\ \vdots & & & & \\ 1 & x^N & (x^N)^2 & \cdots & (x^N)^k \end{bmatrix}, \boldsymbol{r} = \begin{bmatrix} r^1 \\ r^2 \\ \vdots \\ r^N \end{bmatrix}$$

solve for the parameters as

$$\boldsymbol{w} = (\mathbf{D}^T\mathbf{D})^{-1}\mathbf{D}^T\boldsymbol{r}$$

Other Error Measures

- **Square Error:**

$$E(\theta|\mathcal{X}) = \frac{\sum_{t=1}^{N}\left[\boldsymbol{r}^t - g(\boldsymbol{x}^t|\theta)\right]^2}{2}$$

- **Relative Square Error:** $E(\theta|\mathcal{X}) = \frac{\sum_{t=1}^{N}\left[\boldsymbol{r}^t - g(\boldsymbol{x}^t|\theta)\right]^2}{\sum_{t=1}^{N}(\boldsymbol{r}^t - \bar{\boldsymbol{r}})^2}$

- **Absolute Error:** $E(\theta|\mathcal{X}) = \sum_{t=1}^{N}|\boldsymbol{r}^t - g(\boldsymbol{x}^t|\theta)|$

- **ε-sensitive Error:**

$$E(\theta|\mathcal{X}) = \sum_{t=1}^{N} 1\left(|\boldsymbol{r}^t - g(\boldsymbol{x}^t|\theta)| > \epsilon\right)\left(|\boldsymbol{r}^t - g(\boldsymbol{x}^t|\theta)| - \epsilon\right)$$

Bias and Variance

The expected square error at the point x is:

$$E[(r-g(x))^2|x] = \underbrace{E[(r-E[r|x])^2|x]}_{noise} + \underbrace{(E[r|x]-g(x))^2}_{squared\ error}$$

variance *bias²*

The expected value (average over samples X, all of size N and drawn from the same joint density $p(r,x)$ is):

$$E_X[(E[r|x]-g(x))^2|x] = \underbrace{(E[r|x]-E_X[g(x)])^2}_{bias^2} + \underbrace{E_X[(g(x)-E_X[g(x)])^2]}_{variance}$$

Illustrating Bias and Variance

- M samples $X_i = \{x_i^t, r_i^t\}, i = 1, \ldots, M$ of known function $f()$ with added noise. Use each data sets to fit estimator $g_i(x)$, $i = 1,\ldots,M$
- $E[g(x)]$ is estimated by average over M of $g_i()$

$$\bar{g}(x) = \frac{1}{M}\sum_{i=1}^{M} g_i(x)$$

$$\text{Bias}^2(g) = \frac{1}{N}\sum_t [\bar{g}(x^t) - f(x^t)]^2$$

$$\text{Variance}(g) = \frac{1}{NM}\sum_t \sum_i [g_i(x^t) - \bar{g}(x^t)]^2$$

Bias/Variance Dilemma (or tradeoff)

- If $g_i(x) = 2$: It has no variance and high bias
- If $g_i(x) = \sum_t \frac{r_i^t}{N}$: It has *lower bias, higher variance*
- As we increase complexity,

  bias decreases (a better fit to data) and

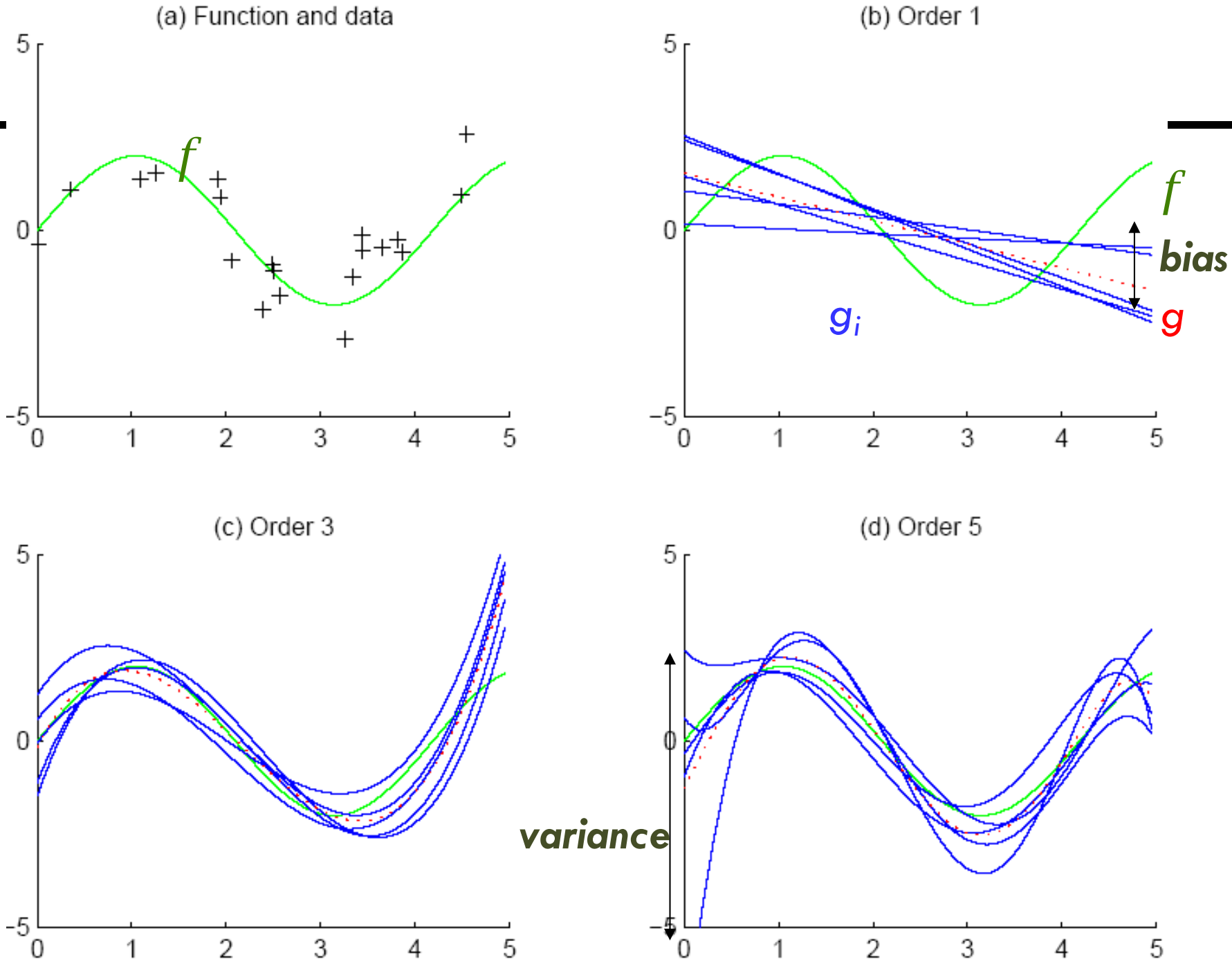
  variance increases (fit varies more with data)
- Bias/Variance dilemma: (Geman et al., 1992)

Illustration

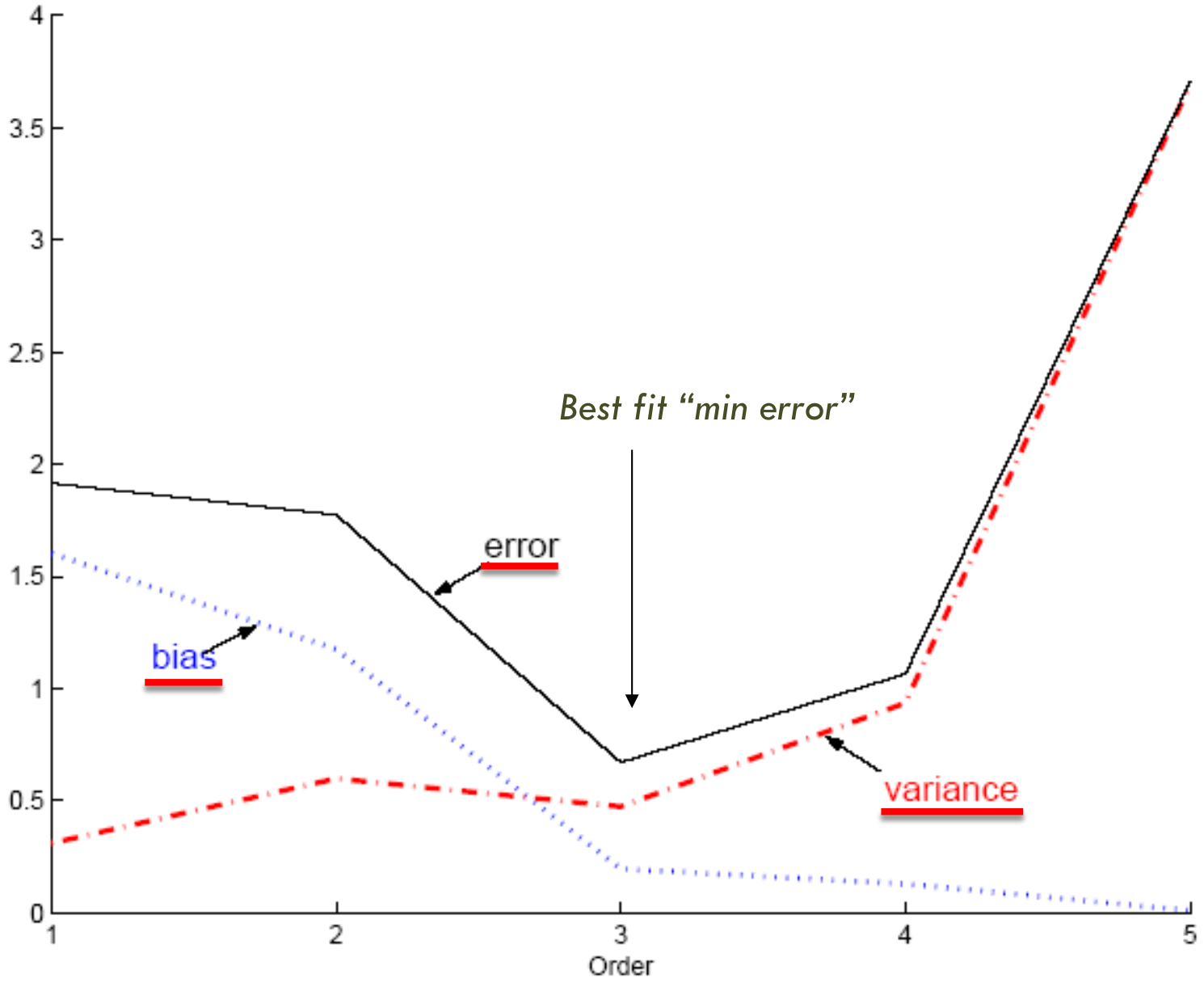
$$r = f(x) + \varepsilon$$

- **Function**: $f(x) = 2\sin(1.5x)$
- **Noise:** $(\mathcal{N}(0,1))$
- **Samples:** *M=5 samples are taken, each contains N=20 points*
- **Five polynomial fit:**

  $g_i(x)$, *i=1,…,5 of order 1, 3, and 5*


(a) Function and data
f
(b) Order 1
f
bias
g
g_i
(c) Order 3
(d) Order 5
variance

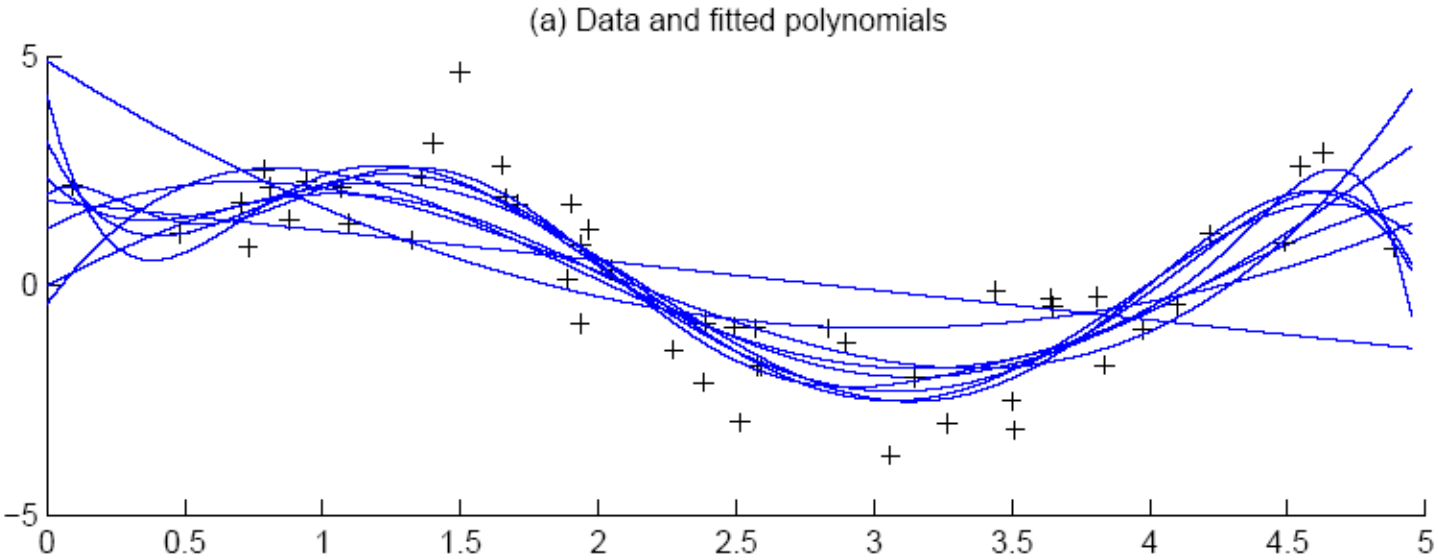
Polynomial Regression (*M=100*)

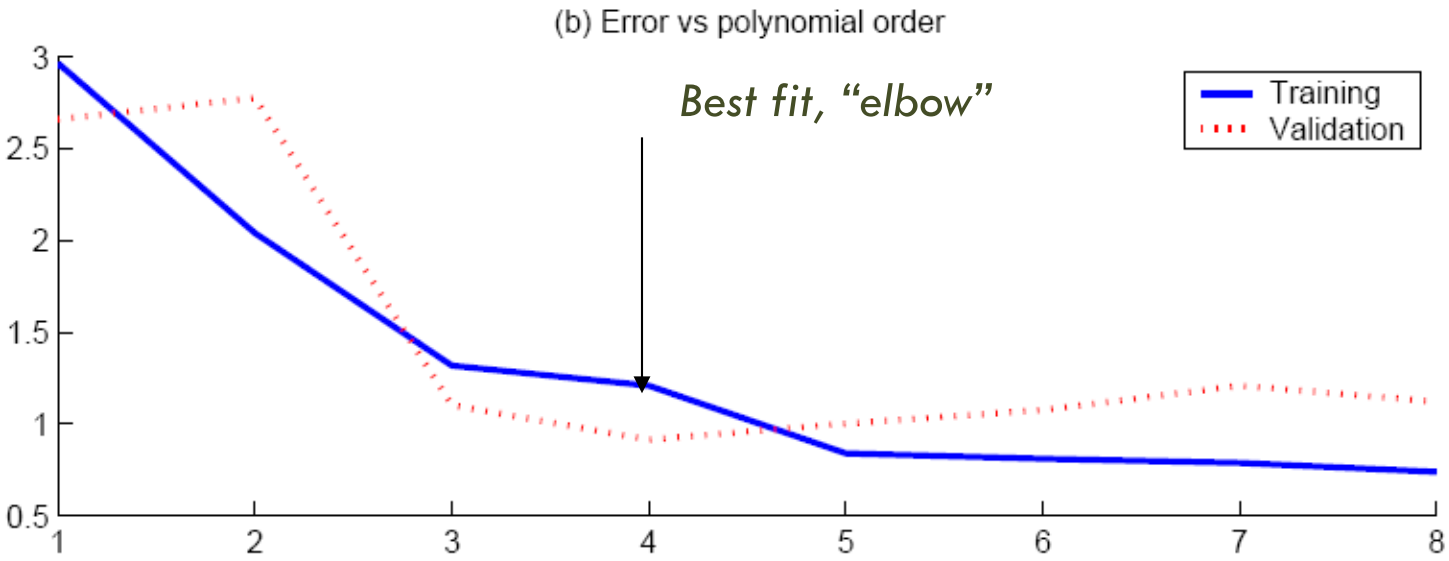


Underfitting vs. Overfitting

- High bias implies "*underfintting*"
- High variance implies "*overfitting*"
- Tradeoff via *cross-validation*

M=50


(a) Data and fitted polynomials
5
0
−5
0
0.5
1
1.5
2
2.5
3
3.5
4
4.5
5


(b) Error vs polynomial order
Best fit, "elbow"
Training
Validation
3
2.5
2
1.5
1
0.5
1
2
3
4
5
6
7
8

Model Selection

- **Cross-validation:** Measure generalization accuracy by testing on data unused during training
- **Regularization:** Penalize complex models

  E'=error on data + λ model complexity

  Akaike's information criterion (AIC), Bayesian information criterion (BIC)
- **Minimum description length (MDL):** Kolmogorov complexity, shortest description of data
- **Structural risk minimization (SRM)**

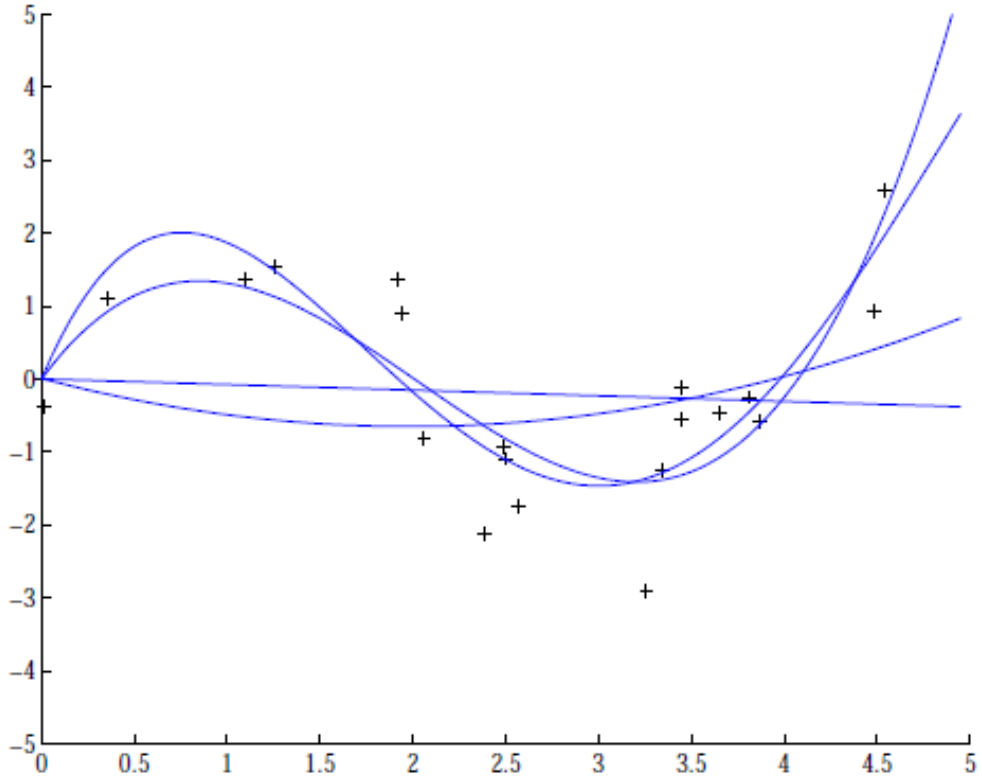
Bayesian Model Selection

- Prior on models, $p(\text{model})$

$$p(\text{model}|\text{data}) = \frac{p(\text{data}|\text{model})p(\text{model})}{p(\text{data})}$$

- Regularization, when prior favors simpler models
- Bayes, MAP of the posterior, $p(\text{model}|\text{data})$
- Average over a number of models with high posterior (voting, ensembles: Chapter 17)

Regression example



Coefficients increase in magnitude as order increases:

1: [-0.0769, 0.0016]

2: [0.1682, -0.6657, 0.0080]

3: [0.4238, -2.5778, 3.4675, -0.0002

4: [-0.1093, 1.4356, -5.5007, 6.0454, -0.0019]

Regularization (L2): $E = \sum_t [r^t - g(x^t|\boldsymbol{w})]^2 + \lambda \sum_i w_i^2$

Conclusion

- Simple Bayesian (or Bayes' rule) has some shortcomings, in particular, when $\boldsymbol{x}$ is not in the database/training data. So we have problem getting the likelihood (or evidence)
- In parametric methods, we assume $\boldsymbol{x}$ follows some probability distribution. So if we can have an "accurate" estimate of the parameters of the distribution, we can use the probability density for the point $\boldsymbol{x}$, this can overcome the problem in the Bayesian method.
- We discuss how to obtain the MLE for parameters in Bernoulli, multinomial, Gaussian probability distribution
- We discuss how to do regression using parametric method
- We discuss Unbiased estimation
- We discuss mean square *error* $= bias^2 +$ *variance*
- We discuss their tradeoffs.