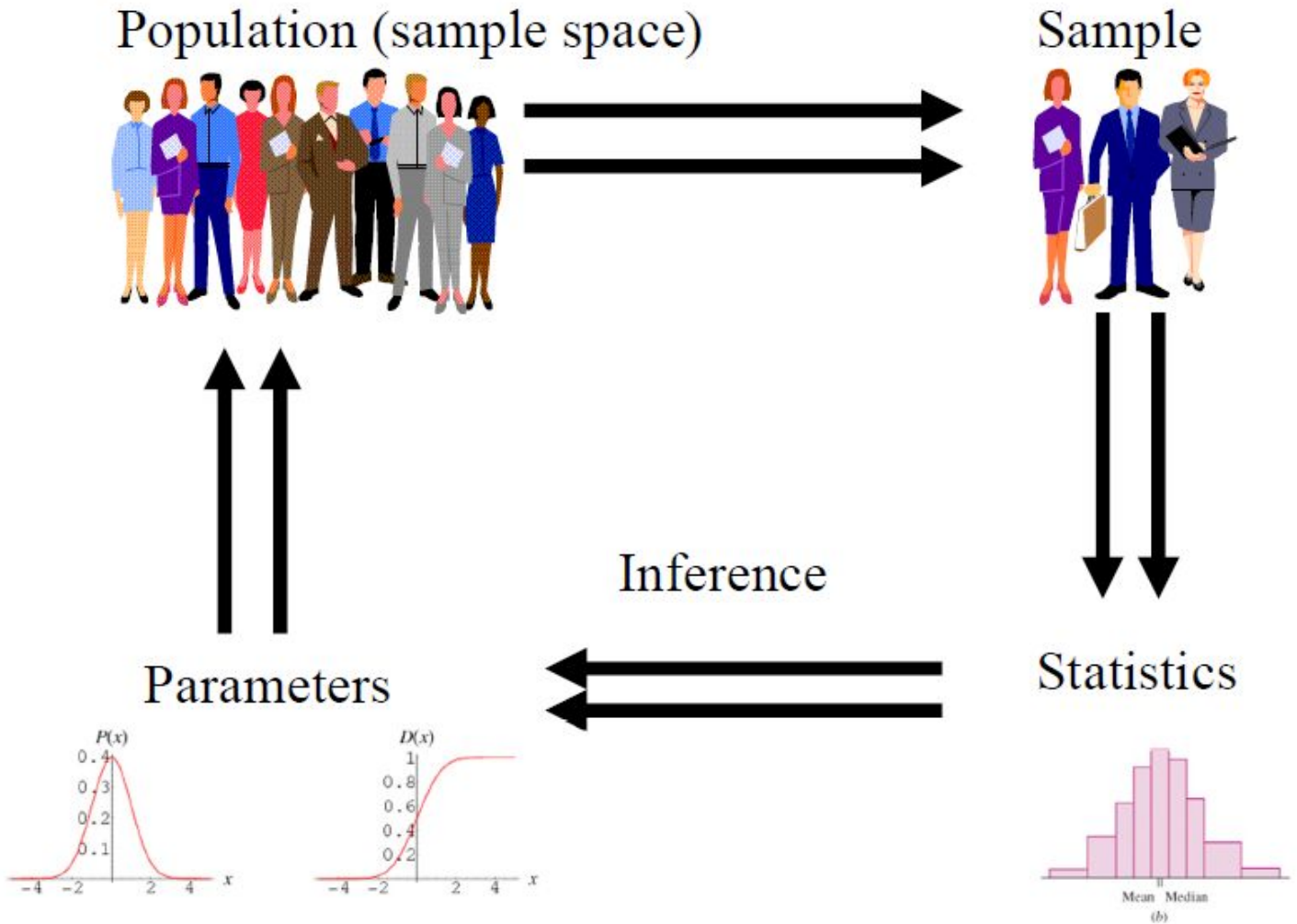
Parametric Methods (Chapter 4)

Sample Statistics and Population Parameters

- A Schematic Depiction



Examples: Bernoulli/Multinomial

- Bernoulli: Two states, failure/success, x in {0,1}

  $P(x) = p_o^x (1 - p_o)^{(1-x)}$

  $\mathrm{L}(p_o|\mathrm{X}) = \log \prod_t p_o^{xt} (1 - p_o)^{(1-xt)}$

  MLE: $p_o = \sum_t x^t / N$

- Multinomial: K>2 states, x_i in {0,1}

  $P(x_1,x_2,...,x_K) = \prod_i p_i^{xi}$

  $\mathrm{L}(p_1,p_2,...,p_K|\mathrm{X}) = \log \prod_t \prod_i p_i^{xit}$

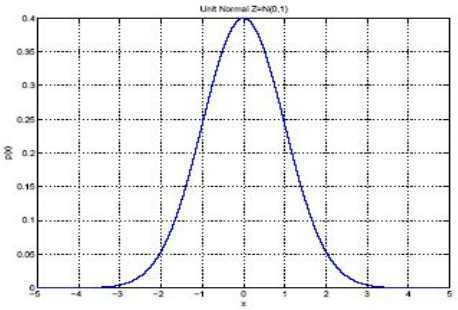
  MLE: $p_i = \sum_t x_i^t / N$

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

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

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

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



Maximum Likelihood Estimation

- Assume the instances $\mathbf{x}=\{x^1,x^2,\ldots,x^t,\ldots,x^N\}$ are independent and identically distributed (*iid*), and drawn from some known probability distribution X
  - $X^t \sim p(x^t|\theta)$
  - θ : model parameters (assumed to be fixed but unknown here)
- MLE attempts to find θ that make $\mathbf{x}$ the most likely to be drawn
  - Namely, maximize the likelihood of the instances

$$l(\theta|\mathbf{x}) = p(\mathbf{x}|\theta) = p(x^1,\cdots,x^N|\theta) = \prod_{t=1}^{N} p(x^t|\theta)$$

($x^1,\ldots,x^N$ are *iid*)

Maximum Likelihood Estimation

- Because logarithm will not change the value of θ when it take its maximum (monotonically increasing/decreasing)
  - Finding θ that maximizes the likelihood of the instances is equivalent to finding θ that maximizes the log likelihood of the samples

$$a \geq b \Rightarrow \log a \geq \log b$$

$$L(\theta|\mathbf{x}) = \log l(\theta|\mathbf{x}) = \sum_{t=1}^{N} \log p(x^t|\theta)$$

  - As we shall see, logarithmic operation can further simplify the computation when estimating the parameters of those distributions that have exponents

MLE: Bernoulli Distribution (1/3)

- Bernoulli Distribution
  - A random variable X takes either the value $x=1$ (with probability r) or the value $x=1$ (with probability $1-r$)
    - Can be thought of as X is generated form two distinct states
  - The associated probability distribution

$$P(x)=r^{x}(1-r)^{1-x} \quad ,x\in\{0,1\}$$

- The log likelihood for a set of *iid* instances $\mathbf{x}$ drawn from Bernoulli distribution

$$\mathbf{x}=\{x^1,x^2,\ldots,x^t,\ldots,x^N\}$$

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

$$=\left(\sum_{t=1}^{N}x^t\right)\log r+\left(N-\sum_{t=1}^{N}x^t\right)\log(1-r)$$

θ

MLE: Bernoulli Distribution (2/3)

- MLE of the distribution parameter r

$$\hat{r} = \frac{\sum_{t=1}^{N} x^t}{N}$$

  - The estimate for r is the ratio of the number of occurrences of the event ($x^t = 1$) to the number of experiments

- The expected value for X

$$E[X] = \sum_{x \in \{0,1\}} x \cdot P(x) = 0 \cdot (1 - r) + 1 \cdot r = r$$

- The variance value for X

$$\text{var}(X) = E\left[X^2\right] - (E[X])^2 = r - r^2 = r(1 - r)$$

MLE: Bernoulli Distribution (3/3)

$$\frac{dL\left(r\middle|X\right)}{dr}=\frac{\partial\left[\left(\sum_{t=1}^{N}x^{t}\right)\log\ r+\left(N-\sum_{t=1}^{N}x^{t}\right)\log\left(1-r\right)\right]}{dr}=0$$

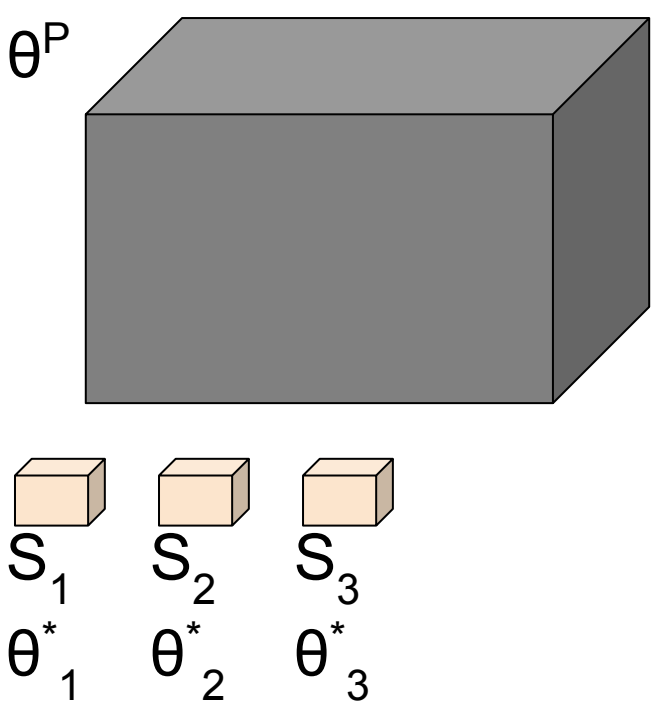
$$\Rightarrow\ \frac{\left(\sum_{t=1}^{N}x^{t}\right)}{r}-\frac{\left(N-\sum_{t=1}^{N}x^{t}\right)}{1-r}=0$$

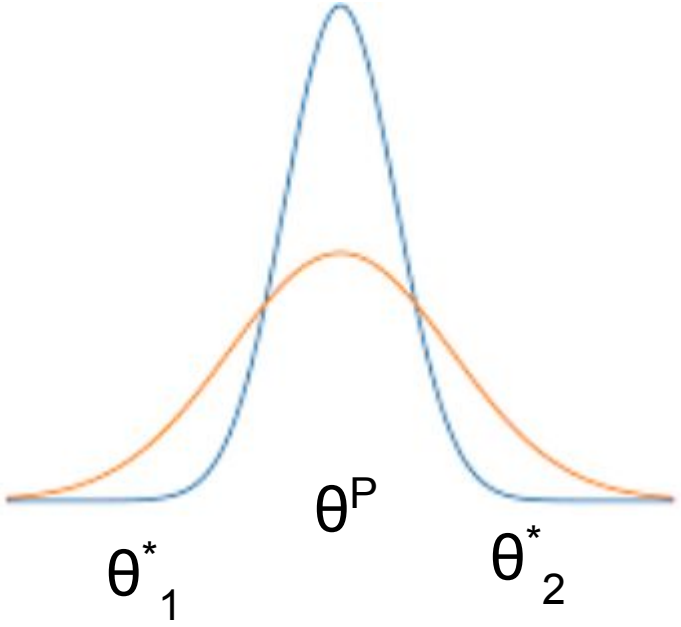
$$\Rightarrow\ \hat{r}=\frac{\sum_{t=1}^{N}x^{t}}{N}$$

$$\frac{d\log\ y}{dy}=\frac{1}{y}$$

The maximum likelihood estimate of the mean is the sample average

Properties of estimator





$E(\hat{\theta}) = \theta^P$

A statistic is said to be an **unbiased estimate** of a given parameter when the mean of the sampling distribution of that statistic can be shown to be equal to the parameter being estimated.

Consistency of an estimator means that as the sample size gets large the estimate gets closer and closer to the true value of the parameter

MLE: Multinomial Distribution (1/3)

- Multinomial Distribution
  - A generalization of Bernoulli distribution
  - The value of a random variable X can be one of *K* mutually exclusive and exhaustive states $x \in \{s_1, s_2, \cdots, s_K\}$ with probabilities $r_1, r_2, \cdots, r_K$, respectively
  - The associated probability distribution

$$p(x) = \prod_{i=1}^{K} r_i^{s_i}, \qquad \sum_{i=1}^{K} r_i = 1$$

$$s_i = \begin{cases} 1 & \text{if } X \text{ choose state } s_i \\ 0 & \text{otherwise} \end{cases}$$

- The log likelihood for a set of *iid* instances $\mathbf{x}$ drawn from a multinomial distribution X

$$L(\mathbf{r}|\mathbf{x}) = \log \prod_{t=1}^{N} \prod_{i=1}^{K} r_i^{s_i^t} \qquad \mathbf{x} = \{x^1, x^2, \ldots, x^t, \ldots, x^N\}$$

MLE: Multinomial Distribution (2/3)

- MLE of the distribution parameter r_i

$$\hat{r}_i = \frac{\sum_{t=1}^{N} s_i^t}{N}$$

  - The estimate for r_i is the ratio of the number of experiments with outcome of state i ($s_i^t = 1$) to the number of experiments

MLE: Multinomial Distribution (3/3)

$$L(\mathbf{r}|\mathbf{x}) = \log \prod_{t=1}^{N} \prod_{i=1}^{K} r_i^{s_i^t} = \sum_{t=1}^{N} \sum_{i=1}^{K} \log\, r_i^{s_i^t}, \quad \text{with constraint} : \sum_{i=1}^{K} r_i = 1$$

$$\frac{\partial \overline{L}(\mathbf{r}|\mathbf{x})}{\partial r_i} = \frac{\partial \left[\sum_{t=1}^{N} \sum_{i=1}^{K} s_i^t \cdot \log\, r_i + \lambda \left(\sum_{i=1}^{K} r_i - 1 \right) \right]}{\partial r_i} = 0$$

Lagrange Multiplier

$$\Rightarrow \sum_{t=1}^{N} s_i^t \cdot \frac{1}{r_i} + \lambda = 0$$

$$\Rightarrow r_i = -\frac{1}{\lambda} \sum_{t=1}^{N} s_i^t$$

$$\Rightarrow \sum_{i=1}^{K} r_i = 1 = -\frac{1}{\lambda} \sum_{t=1}^{N} \left(\sum_{i=1}^{K} s_i^t \right)$$

=1

$$\Rightarrow \lambda = -N$$

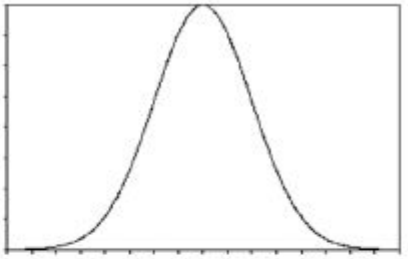
$$\Rightarrow \hat{r}_i = \frac{\sum_{t=1}^{N} s_i^t}{N}$$

Lagrange Multiplier: http://www.slimy.com/~steuard/teaching/tutorials/Lagrange.html

MLE: Gaussian Distribution (1/3)

- Also called Normal Distribution
  - Characterized with mean μ and variance σ^2

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



  - Recall that mean and variance are sufficient statistics for Gaussian

- The log likelihood for a set of *iid* instances drawn from Gaussian distribution X

$$L(\mu, \sigma | \mathbf{x}) = \log \prod_{t=1}^{N} \frac{1}{\sqrt{2\pi}\sigma} e^{-\left(\frac{(x^t-\mu)^2}{2\sigma^2}\right)} \qquad \mathbf{x} = \{x^1, x^2, \ldots, x^t, \ldots, x^N\}$$

$$= -\frac{N}{2}\log(2\pi) - N\log\sigma - \frac{\sum_{t=1}^{N}(x^t-\mu)^2}{2\sigma^2}$$

MLE: Gaussian Distribution (2/3)

- MLE of the distribution parameters μ and σ^2

$$m = \hat{\mu} = \frac{\sum_{t=1}^{N} x^t}{N}$$ sample average

$$s^2 = \hat{\sigma}^2 = \frac{\sum_{t=1}^{N} \left(x^t - m\right)^2}{N}$$ sample variance

- Remind that μ and σ^2 are still fixed but unknown

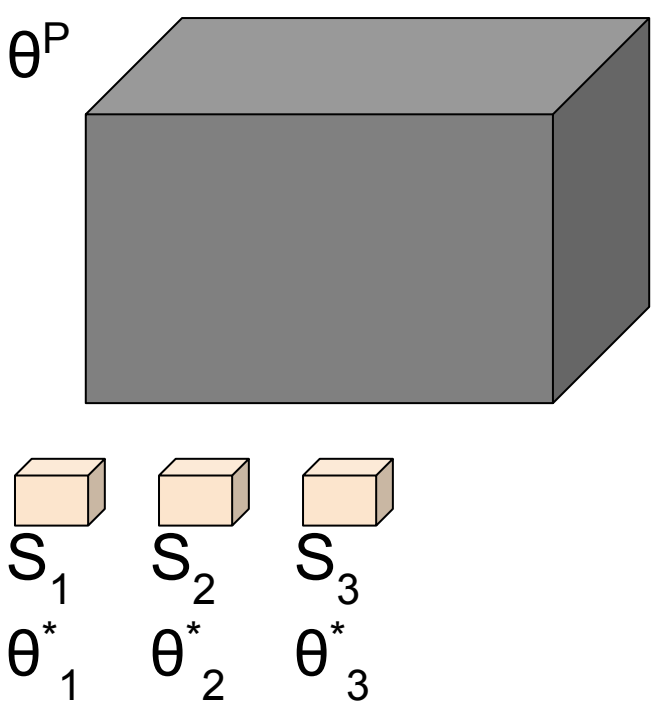
MLE: Gaussian Distribution (3/3)

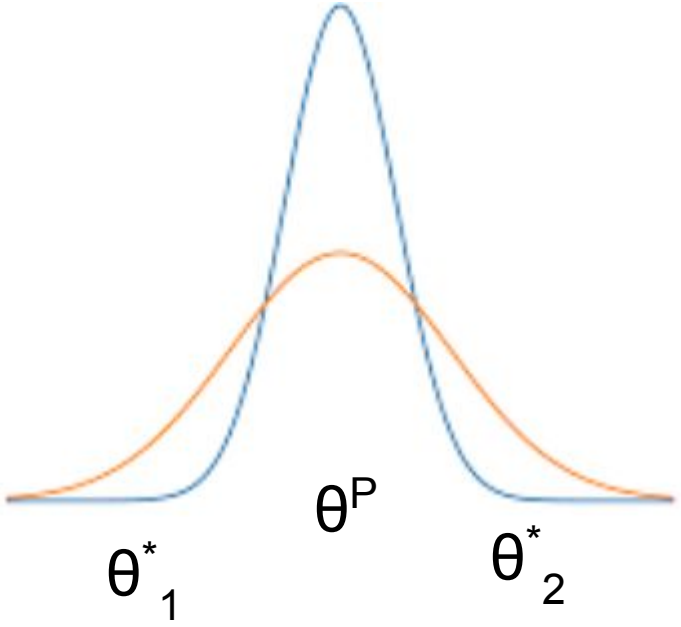
$$L(\mu,\sigma|\mathbf{x}) = -\frac{N}{2}\log(2\pi) - \frac{N}{2}\log\sigma^2 - \frac{\sum_{t=1}^{N}(x^t-\mu)^2}{2\sigma^2}$$

$$\frac{\partial L(\mu,\sigma|\mathbf{x})}{\partial \mu} = 0 \Rightarrow \frac{1}{\sigma^2}\sum_{t=1}^{N}(x^t-\mu)^2 = 0 \Rightarrow \hat{\mu} = \frac{\sum_{t=1}^{N} x^t}{N}$$

$$\frac{\partial L(\mu,\sigma|\mathbf{x})}{\partial \sigma^2} = 0 \Rightarrow -N + \frac{1}{\sigma^2}\sum_{t=1}^{N}(x^t-\mu)^2 = 0 \Rightarrow \hat{\sigma}^2 = \frac{\sum_{t=1}^{N}(x^t-\hat{\mu})^2}{N}$$

Properties of estimator





$E(\hat{\theta}) = \theta^P$

A statistic is said to be an **unbiased estimate** of a given parameter when the mean of the sampling distribution of that statistic can be shown to be equal to the parameter being estimated.

Consistency of an estimator means that as the sample size gets large the estimate gets closer and closer to the true value of the parameter

Evaluating an Estimator : Bias and Variance (2/6)

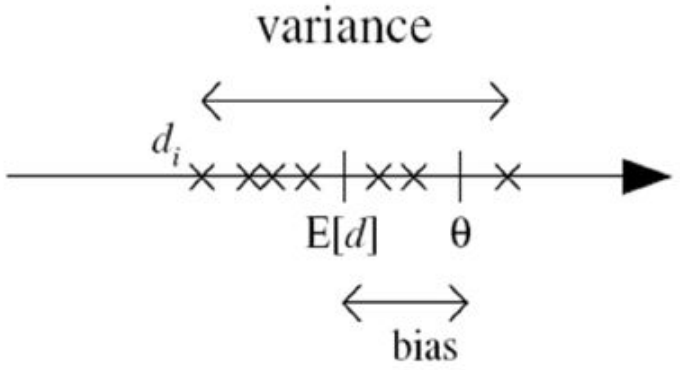


Figure 4.1: θ is the parameter to be estimated. d_i are several estimates (denoted by '$\times$') over different samples. Bias is the difference between the expected value of d and θ. Variance is how much d_i are scattered around the expected value. We would like both to be small.

Evaluating an Estimator

Let $\mathcal{X}$ be a sample from a population specified up to a parameter θ, and let $d = d(\mathcal{X})$ be an estimator of θ. To evaluate the quality of this estimator, we can measure how much it is different from θ, that is, $(d(\mathcal{X}) - \theta)^2$.

$$r(d, \theta) = E[(d(\mathcal{X}) - \theta)^2]$$

The *bias* of an estimator is given as

$$b_\theta(d) = E[d(\mathcal{X})] - \theta$$

sample average, m, is an unbiased estimator of the mean, μ, because

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

Evaluating an Estimator

sample average, m, is an unbiased estimator of the mean, μ, because

$E[m] =$

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

$\text{Var}(m) =$

Evaluating an Estimator

sample average, m, is an unbiased estimator of the mean, μ, because

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

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

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

Evaluating an Estimator : Bias and Variance (1/6)

- The mean square error of the estimator d can be further decomposed into two parts respectively composed of bias and variance

$$r(d,\theta) = E\left[(d-\theta)^2\right]$$

Evaluating an Estimator : Bias and Variance (3/6)

- Example 1: **sample average and sample variance**
  - Assume samples $\mathbf{x}=\{x^1,x^2,\ldots,x^t,\ldots,x^N\}$ are independent and identically distributed (*iid*), and drawn from some known probability distribution X with mean μ and variance σ^2
    - Mean $\mu = E[X] = \sum_x x \cdot p(x)$
    - Variance $\sigma^2 = E\left[(X-\mu)^2\right] = E[X^2] - (E[X])^2$
    - Sample average (mean) for the observed samples $m = \frac{1}{N}\sum_{t=1}^{N} x^t$
    - Sample variance for the observed samples $s^2 = \frac{1}{N}\sum_{t=1}^{N}\left(x^t - m\right)^2$

      or $s^2 = \frac{1}{N-1}\sum_{t=1}^{N}\left(x^t - m\right)^2$?

Evaluating an Estimator : Bias and Variance (4/6)

- Example 1 (count.)
  - Sample average m is an unbiased estimator of the mean μ

$$E[m] = E\left[\frac{1}{N}\sum_{t=1}^{N} X^t\right] = \frac{1}{N}\sum_{t=1}^{N} E[X] = \frac{N\cdot\mu}{N} = \mu$$

$$\therefore E[m] - \mu = 0$$

  - m is also a consistent estimator: $\mathrm{Var}(m) \to 0$ as $N \to \infty$

$$\mathrm{Var}(m) = \mathrm{Var}\left(\frac{1}{N}\sum_{t=1}^{N} X^t\right) = \frac{1}{N^2}\sum_{t=1}^{N}\mathrm{Var}(X) = \frac{N\cdot\sigma^2}{N^2} = \frac{\sigma^2}{N} \xrightarrow{N=\infty} 0$$

$$\mathrm{Var}(aX+b) = a^2\cdot\mathrm{Var}(X)$$
$$\mathrm{Var}(X+Y) = \mathrm{Var}(X) + \mathrm{Var}(Y)$$

Evaluating an Estimator : Bias and Variance (5/6)

- Example 1 (count.)
  - Sample variance s^2 is an asymptotically unbiased estimator of the variance σ^2

$$s^2 = \frac{1}{N}\sum_{t=1}^{N}\left(x^t - m\right)^2$$

$$\sum_{t=1}^{N} x^t = N \cdot m$$

$$E\left[s^2\right] = E\left[\frac{1}{N}\sum_{t=1}^{N}\left(X^t - m\right)^2\right]$$

$$= E\left[\frac{1}{N}\sum_{t=1}^{N}\left(X - m\right)^2\right] \quad \left(X^t\text{'s are i.i.d.}\right)$$

$$= E\left[\frac{1}{N}\sum_{t=1}^{N}\left(X^2 - 2X \cdot m + m^2\right)\right]$$

$$= E\left[\frac{N \cdot X^2 - 2N \cdot m^2 + Nm^2}{N}\right]$$

$$= E\left[\frac{N \cdot X^2 - N \cdot m^2}{N}\right] = \frac{N \cdot E\left[X^2\right] - N \cdot E\left[m^2\right]}{N}$$

Evaluating an Estimator : Bias and Variance (6/6)

- Example 1 (count.)
  - Sample variance s^2 is an asymptotically unbiased estimator of the variance σ^2

$$\mathrm{Var}(m) = \frac{\sigma^2}{N} = E[m^2] - (E[m])^2$$

$$\Rightarrow E[m^2] = \frac{\sigma^2}{N} + (E[m])^2 = \frac{\sigma^2}{N} + \mu^2$$

$$E[s^2] = \frac{N \cdot E[X^2] - N \cdot E[m^2]}{N}$$

$$= \frac{N(\sigma^2 + \mu^2) - N\left(\frac{\sigma^2}{N} + \mu^2\right)}{N}$$

$$\mathrm{Var}(X) = \sigma^2 = E[X^2] - (E[X])^2$$

$$\Rightarrow E[X^2] = \sigma^2 + (E[X])^2 = \sigma^2 + \mu^2$$

$$= \frac{(N-1)}{N}\sigma^2 \xrightarrow{N=\infty} \sigma^2$$

The size of the observed sample set

Exercise

Show that $\bar{x}_1 - \bar{x}_2$ is an unbiased estimator for $\mu_1 - \mu_2$.
Also show that the variance of this estimator is $\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}$

Bias and Variance: Example 2

different samples for an unknown population

$X \rightarrow (x, y)$

$y = F(x)$

$$E\left[(d - E[d])^2\right] + (E[d] - \theta)^2$$

variance bias²

$$y' = F(x) + \varepsilon$$

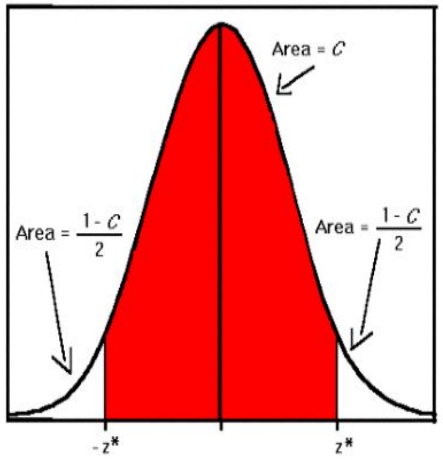
As we increase complexity,
bias decreases (a better fit to data)
variance increases (fit varies more with data)

Bayes' Estimator

- Some expert says that with 90% confidence, θ lies between 5 and 9, symmetrically around 7.
- $p(\theta) \sim N(7, (2/1.64)^2)$.

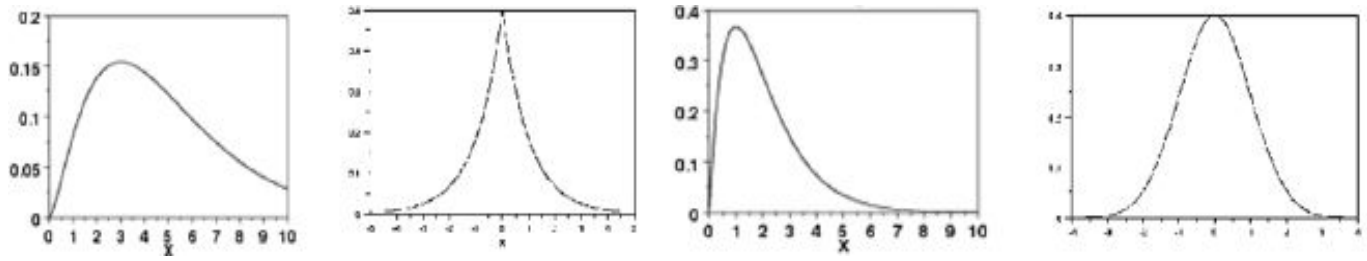
Confidence Level	Confidence Coefficient, $1-\alpha$	z value, $Z_{\alpha/2}$
80%	.80	1.28
90%	.90	1.645
95%	.95	1.96
98%	.98	2.33
99%	.99	2.57
99.8%	.998	3.08
99.9%	.999	3.27

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



Assume known prior density of the parameter

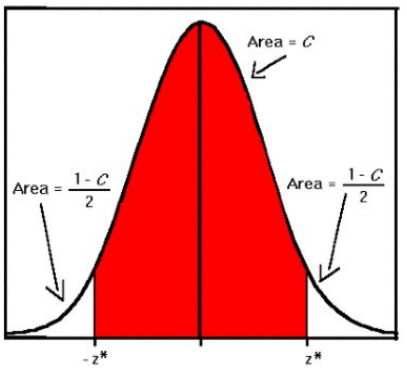
- Assume that examples are drawn from some distribution that obeys a known model



Assume prior density *p(θ)*

e.g. *θ* is approximately normal and with 90% confidence, *θ* lies between 5 and 9, symmetrically around 7.

p(θ) ~ $\mathcal{N}(7, (2/1.64)^2)$.



Given training sample, estimate density of an input

.

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

$$= $$

$$= $$

Given training sample, estimate density of an input

.

$$
\begin{aligned}
p(x|\mathcal{X}) &= \int p(x,\theta|\mathcal{X})d\theta \\
&= \int p(x|\theta,\mathcal{X})p(\theta|\mathcal{X})d\theta \\
&= \int p(x|\theta)p(\theta|\mathcal{X})d\theta
\end{aligned}
$$

Sufficient statistics

Given training sample, estimate density of an input

.

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

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

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

If we are doing prediction as in regression g(.)

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

Difficult to compute integral, assume $p(\theta|X)$ has narrow peak

Given training sample, estimate density of an input

- *assume p(θ|X)* has narrow peak around its mode
- use the maximum a posteriori (MAP) to make the calculation easier

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

$$\approx p(x|\theta)_{\text{MAP}}$$

If we are doing prediction as in regression g(.)

$$y = g(x|\theta)_{\text{MAP}}$$

Difficult to compute integral, assume *p(θ|X)* has narrow peak

Assume $p(\theta|X)$ has narrow peak

Using *maximum a posteriori (MAP)* estimate:

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

Estimate density of (predict output of) an input:

$$p(x|\mathcal{X}) = p(x|\theta_{MAP})$$
$$y_{MAP} = g(x|\theta_{MAP})$$

Assume $p(\theta|X)$ has narrow peak and $p(\theta)$ is flat

Instead of *Maximum a posteriori (MAP)* estimate:

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

Use *maximum likelihood (ML)* estimate:

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

the MAP estimate will be equivalent to the maximum likelihood estimate

Example

As an example, let us suppose $x^t \sim \mathcal{N}(\theta, \sigma_0^2)$ and $\theta \sim \mathcal{N}(\mu, \sigma^2)$, where μ, σ, and σ_0^2 are known:

What is MAP and ML estimate of θ?

$$p(\mathcal{X}|\theta) = \frac{1}{(2\pi)^{N/2}\sigma_0^N} \exp\left[-\frac{\sum_t (x^t - \theta)^2}{2\sigma_0^2}\right]$$

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

What is MAP and ML estimate of θ? $\frac{\sum_t x^t}{N}$

$$\frac{N/\sigma_0^2}{N/\sigma_0^2 + 1/\sigma^2} m + \frac{1/\sigma^2}{N/\sigma_0^2 + 1/\sigma^2} \mu$$

Given training sample, estimate density of an input

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

Bayes' estimator: instead, find expected value of the posterior density:

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

The best estimate of a random variable is its mean

Example:

Suppose x^t and θ are both from Normal distribution

$$x^t \sim \mathcal{N}(\theta, \sigma^2) \text{ and } \theta \sim \mathcal{N}(\mu_0, \sigma_0^2)$$

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

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

Bayes' estimator of θ

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

N increases?

Parametric Classification

Posterior probability of class C_i

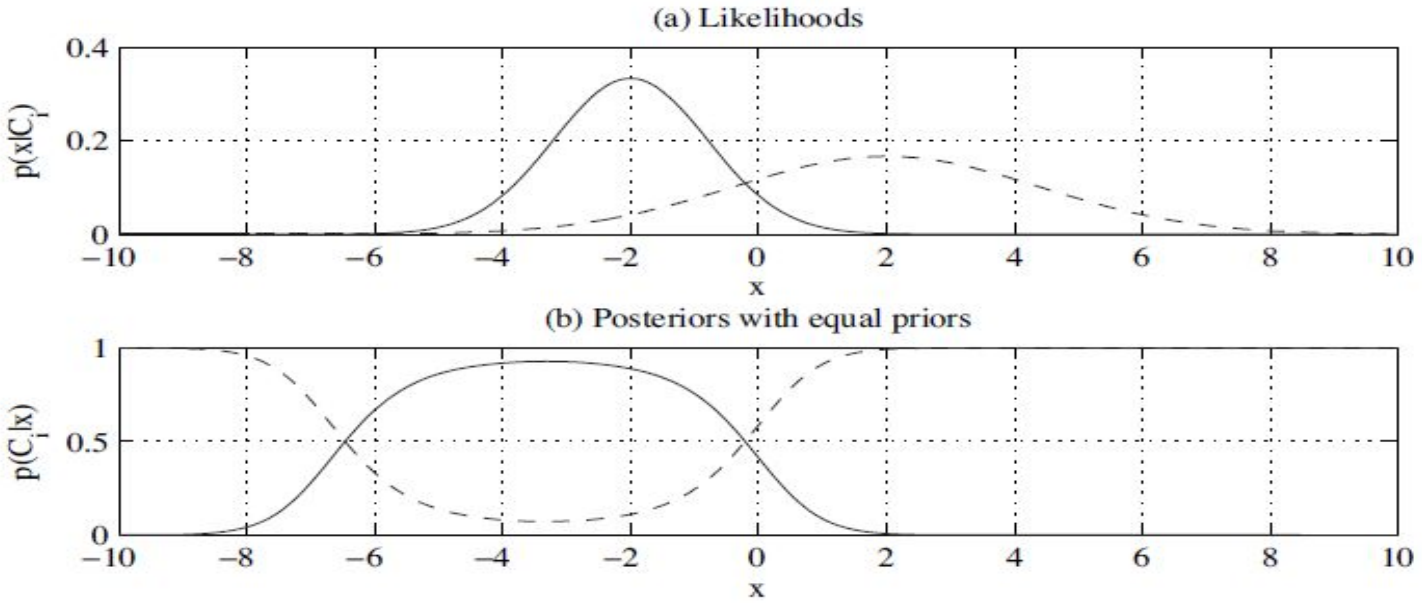
$$P(C_i|x) = \frac{p(x|C_i)P(C_i)}{p(x)}$$

Discriminant function

$$g_i(x) = p(x|C_i)P(C_i)$$

If Gaussian distribution:

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



Parametric Classification

probability of class C_i as

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

and use the discriminant function

$$g_i(x) = p(x|C_i)P(C_i)$$

or equivalently

$$g_i(x) = \log p(x|C_i) + \log P(C_i)$$

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

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

Parametric Classification: Example

- Assume we are a car company selling K different cars, and for simplicity, let us say that the sole factor that affects a customer's choice is his or her yearly income, which we denote by x.
- $P(C_i)$ is the proportion of customers who buy car type i.
- If the yearly income distributions of such customers can be approximated with a Gaussian:
  - $p(x|C_i)$: the probability that a customer who bought car type i has income x:
  - $N(\mu_i, \sigma_i^2)$
  - μ_i:? meaning

$$\mathcal{X} = \{x^t, \mathbf{r}^t\}_{t=1}^N$$

$$r_i^t = \begin{cases} 1 & \text{if } \mathbf{x}^t \in C_i \\ 0 & \text{if } \mathbf{x}^t \in C_k, k \neq i \end{cases}$$

$$m_i = \frac{\sum_t x^t r_i^t}{\sum_t r_i^t}$$

$$s_i^2 = \frac{\sum_t (x^t - m_i)^2 r_i^t}{\sum_t r_i^t}$$

discriminant function?

choose C_i if?

Parametric Classification: Example

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

The priors are equal, the last term can also be dropped. If we can further assume that variances are equal, we can write

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

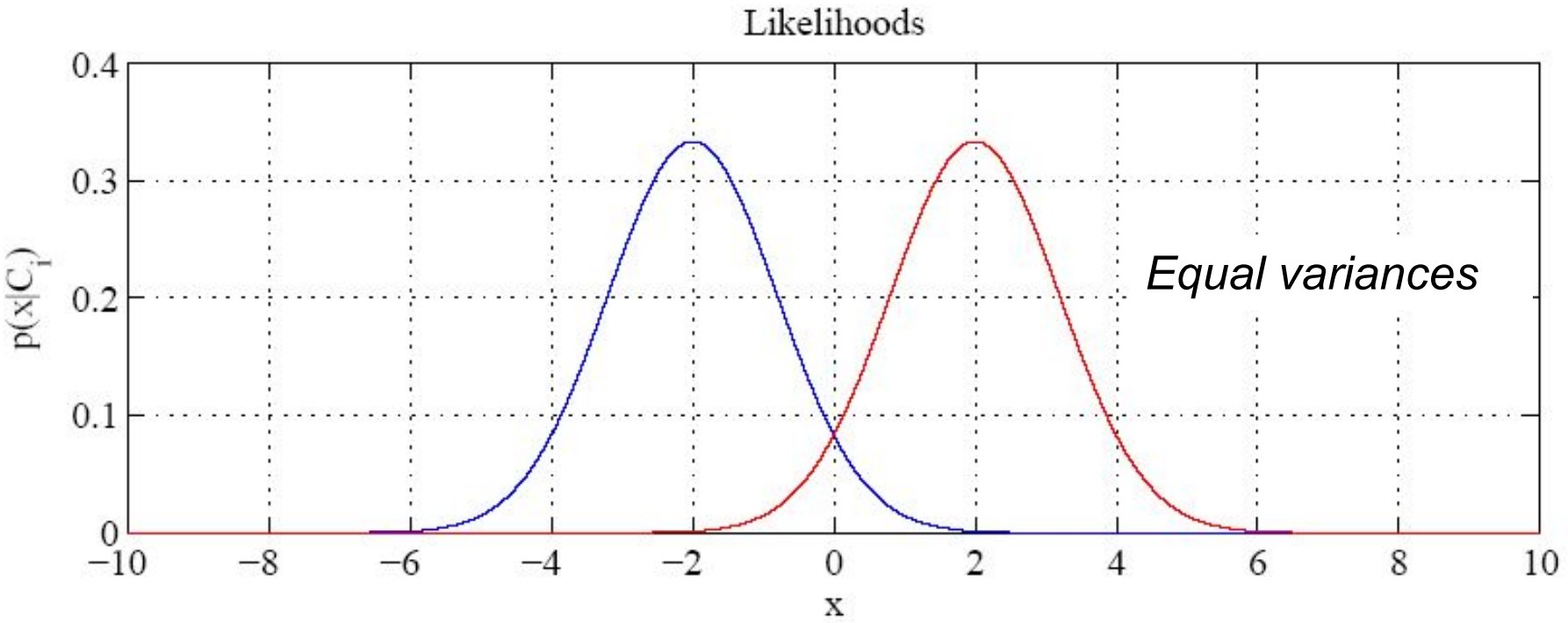
$$\text{Choose } C_i \text{ if } |x - m_i| = \min_k |x - m_k|$$

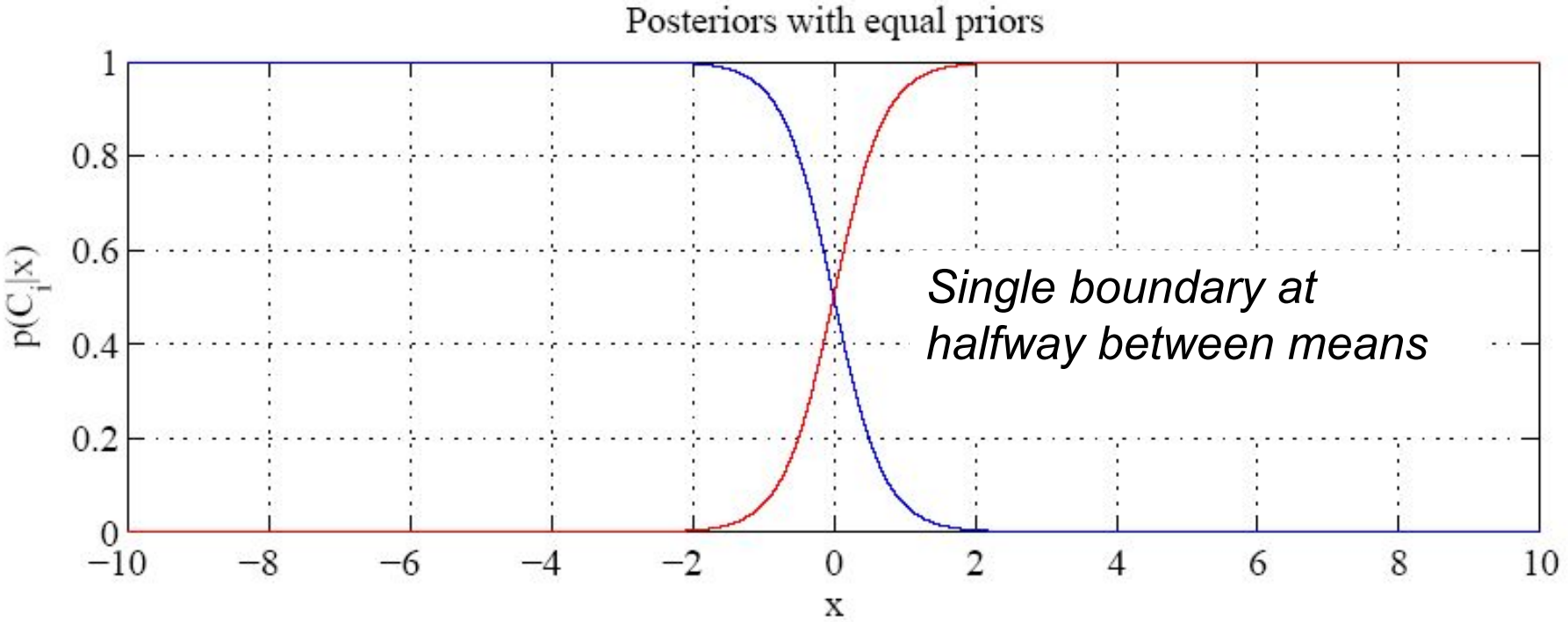
threshold of decision

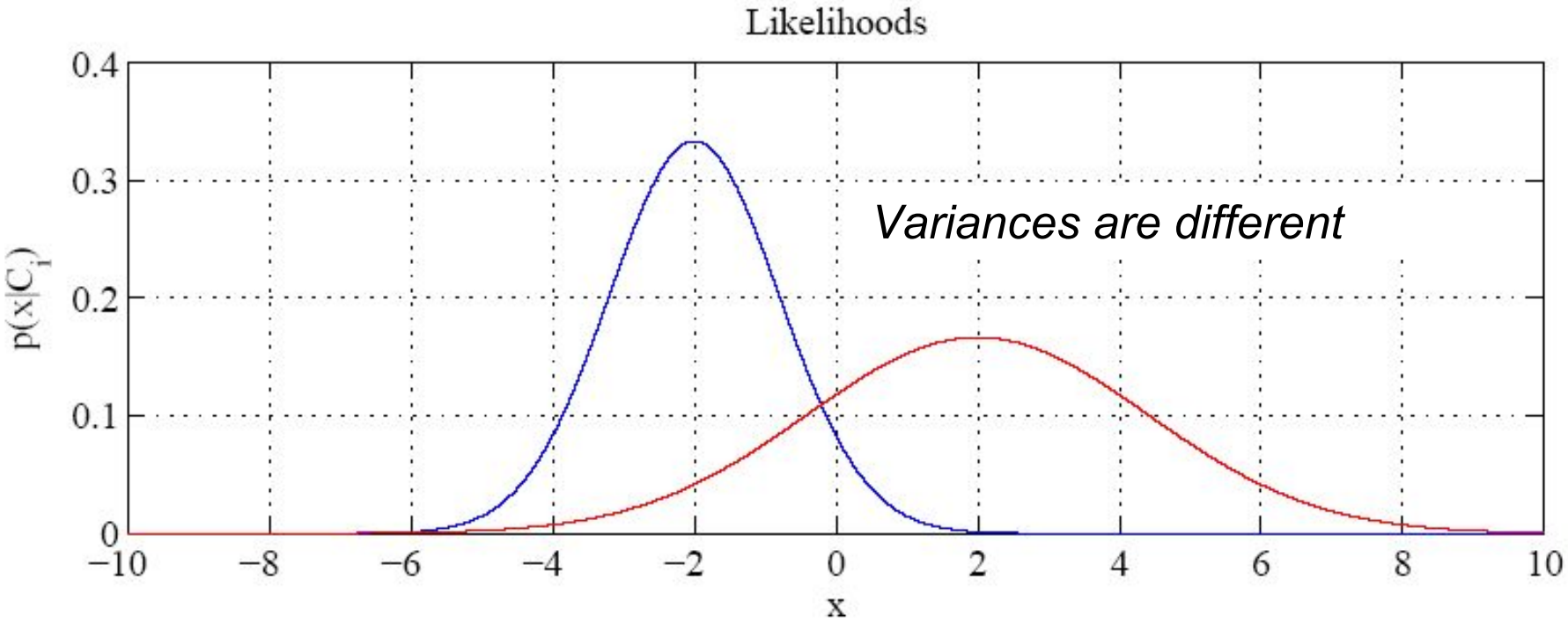
$$\begin{aligned} g_1(x) &= g_2(x) \\ (x - m_1)^2 &= (x - m_2)^2 \\ x &= \frac{m_1 + m_2}{2} \end{aligned}$$

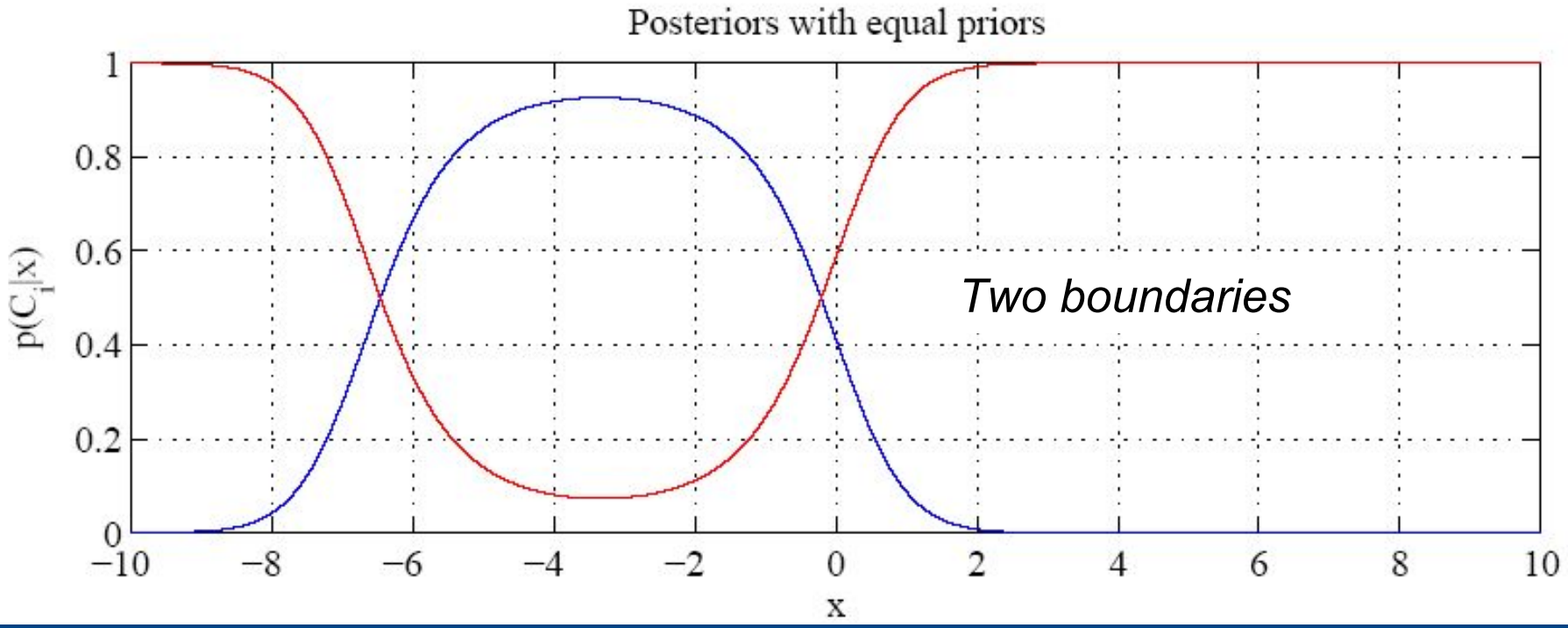
Parametric Classification: 2-class example

$$
\begin{aligned}
g_1(x) &= g_2(x) \\
(x - m_1)^2 &= (x - m_2)^2 \\
x &= \frac{m_1 + m_2}{2}
\end{aligned}
$$


Likelihoods
p(x|C_i)
x
Equal variances


Posteriors with equal priors
p(C_i|x)
x
Single boundary at
halfway between means


Likelihoods
Variances are different
p(x|C_i)
x
0.4
0.3
0.2
0.1
0
−10
−8
−6
−4
−2
0
2
4
6
8
10


Posteriors with equal priors
Two boundaries
p(C_i|x)
x
1
0.8
0.6
0.4
0.2
0
−10
−8
−6
−4
−2
0
2
4
6
8
10

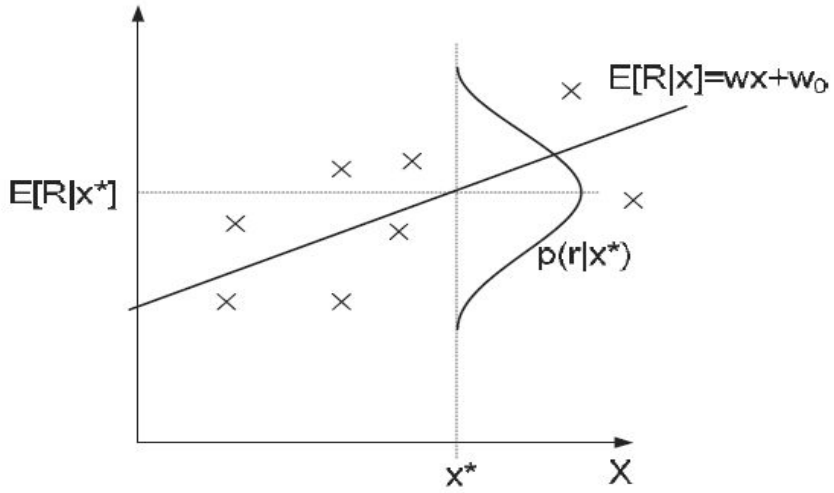
Regression

Output: Deterministic function of input with random noise:

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

Estimate with $g(x|\theta)$

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



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

Use maximum likelihood to estimate θ

Regression

Sample $\mathcal{X} = \{x^t, r^t\}_{t=1}^N$

Log likelihood:

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

Maximize (minimize):

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

Least squares estimate

Linear regression

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

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

Derivative wrt w_1 and w_2

$$\begin{aligned}\sum_t r^t &= Nw_0 + w_1\sum_t x^t \\ \sum_t r^t x^t &= w_0\sum_t x_t + w_1\sum_t (x^t)^2\end{aligned}$$

Re-write in vector-matrix form

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

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

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

Linear regression, higher order polynomial

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

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

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

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

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

and we can then solve for the parameters as

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

Other error measures

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

Relative square error $$E_{RSE} = \frac{\sum_t [r^t - g(x^t|\theta)]^2}{\sum_t (r^t - \bar{r})^2}$$

Coefficient of determination $$R^2 = 1 - E_{RSE}$$

Model selection

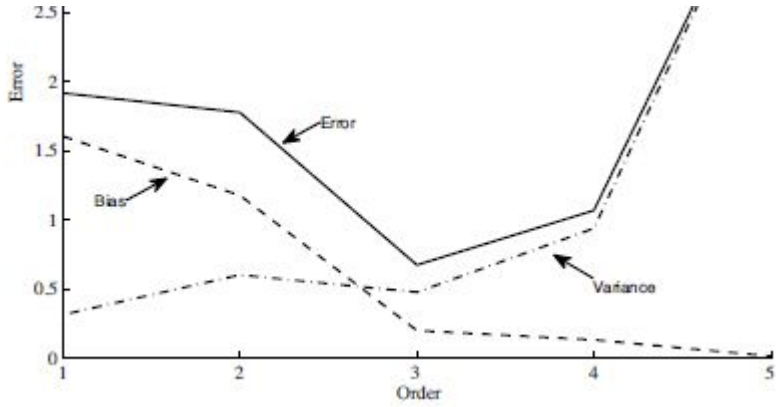
- Last lecture: evaluate models with ?

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

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

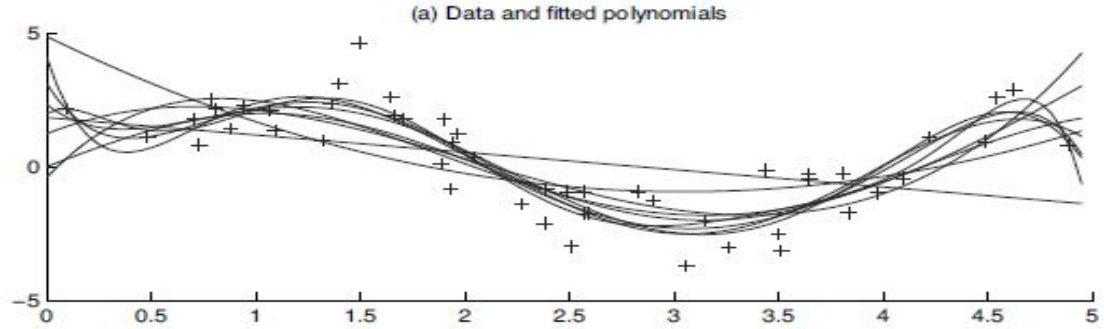
$$\text{Variance}(g) = \frac{1}{NM}\sum_t \sum_i [g_i(x^t) - \bar{g}(x^t)]^2$$

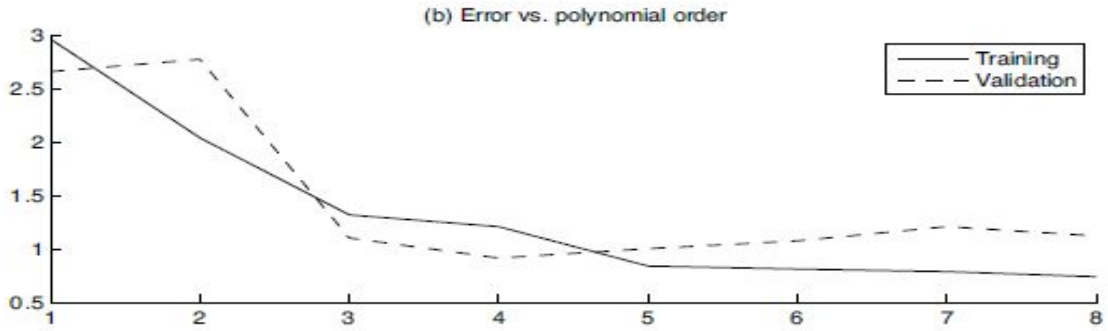
- In practice, use *cross-validation*
- Divide into training and validation sets

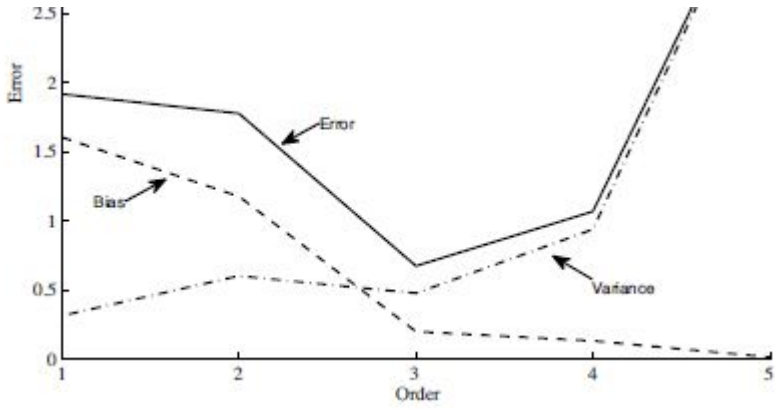


Model selection

- In practice, use *cross-validation*
- Divide into training and validation sets



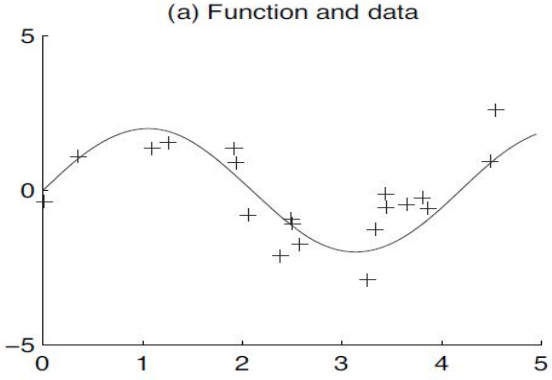


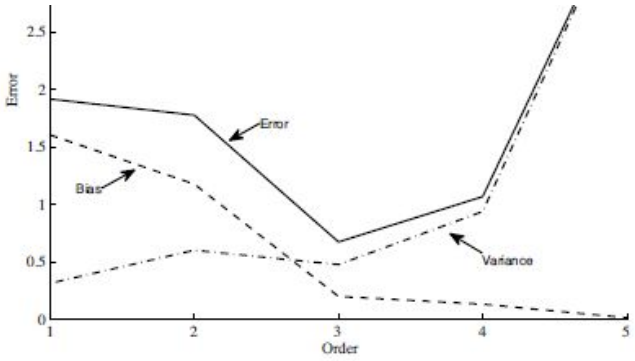


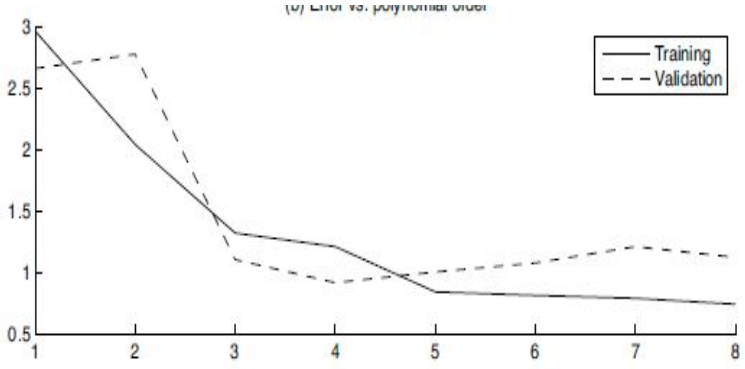
Homework 1

Function, $f(x) = 2\sin(1.5x)$, and one noisy $(\mathcal{N}(0,1))$ dataset

- 100 samples each containing 20 instances
  - Fit 100 models of order 1, order 3, order 5 polynomials
  - Plot bias, variance and error of polynomials
- 10 samples, each containing 100 instances
  - Split each sample to training and validation
  - Plot mean training and validation error for each polynomial
- Use a real dataset (Iris dataset) for training and validation error







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)**

Model selection

- Prior on models, p(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(model|data)

$$\log p(\text{model}|\text{data}) = \log p(\text{data}|\text{model}) + \log p(\text{model}) - c$$

- Exercise: Find MAP for regression and use prior $p(\boldsymbol{w}) \sim \mathcal{N}(0, 1/\lambda)$

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

- Average over a number of models with high posterior (voting, ensembles: Chapter 15)

Bayes' Estimator

- Treat θ as a random var with prior $p(\theta)$
- Bayes' rule: $p(\theta|\mathrm{X}) = p(\mathrm{X}|\theta)\, p(\theta) / p(\mathrm{X})$
- Full: $p(x|\mathrm{X}) = \int p(x|\theta)\, p(\theta|\mathrm{X})\, d\theta$
- Maximum a Posteriori (MAP): $\theta_{\mathrm{MAP}} = \mathrm{argmax}_{\theta}\, p(\theta|\mathrm{X})$
- Maximum Likelihood (ML): $\theta_{\mathrm{ML}} = \mathrm{argmax}_{\theta}\, p(\mathrm{X}|\theta)$
- Bayes': $\theta_{\mathrm{Bayes'}} = \mathrm{E}[\theta|\mathrm{X}] = \int \theta\, p(\theta|\mathrm{X})\, d\theta$

Bayes' Estimator: Example

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

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

- θ_{ML} = *?*
- $\theta_{MAP} = \theta_{Bayes'}$ =

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

Exercise

- Given two normal distributions $p(x|C_1) \sim N(\mu_1, \sigma_1^2)$ and $p(x|C_2) \sim N(\mu_2, \sigma_2^2)$ and P(C1) and P(C2), calculate the Bayes' discriminant points analytically.

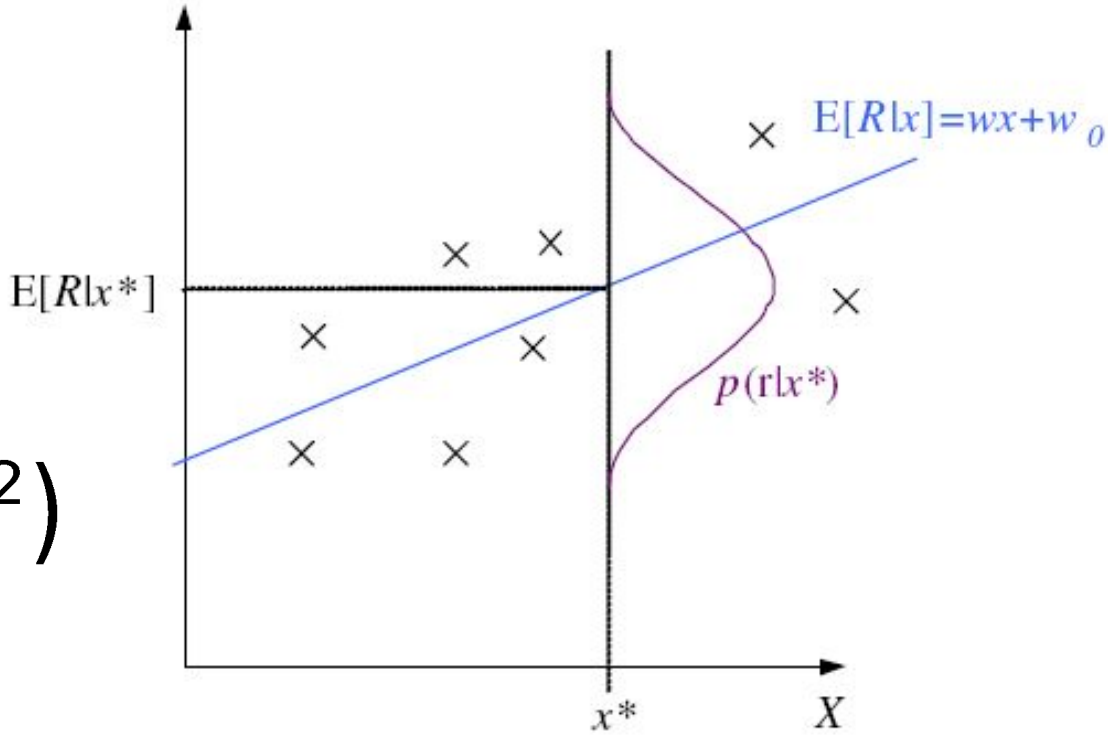
Regression

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

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

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

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



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

Regression: From LogL to Error

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

Maximizing this is equivalent to minimizing

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

Linear Regression

$$g(x^t|w_1,w_0)=w_1x^t+w_0$$

$$\sum_t r^t=Nw_0+w_1\sum_t x^t$$

$$\sum_t r^t x^t=w_0\sum_t x^t+w_1\sum_t (x^t)^2$$

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

$$w=A^{-1}y$$

Polynomial Regression

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

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

$$w=(D^TD)^{-1}D^Tr$$

Other Error Measures

- Square Error:

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

- Relative Square Error:

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

- Absolute Error: $E(\theta|X) = \sum_t |r^t-g(x^t|\theta)|$

Tuning model complexity: Bias/Variance

$$E\left[(r-g(x))^2 \mid x\right] = E\left[(r-E[r \mid x])^2 \mid x\right] + (E[r \mid x]-g(x))^2$$

noise *squared error*

$$E_X\left[(E[r \mid x]-g(x))^2 \mid x\right] = (E[r \mid x]-E_X[g(x)])^2 + E_X\left[(g(x)-E_X[g(x)])^2\right]$$

bias *variance*

$$\underbrace{E\left[(d-E[d])^2\right]}_{\text{variance}} + \underbrace{(E[d]-\theta)^2}_{\text{bias}^2}$$

Estimating Bias and Variance

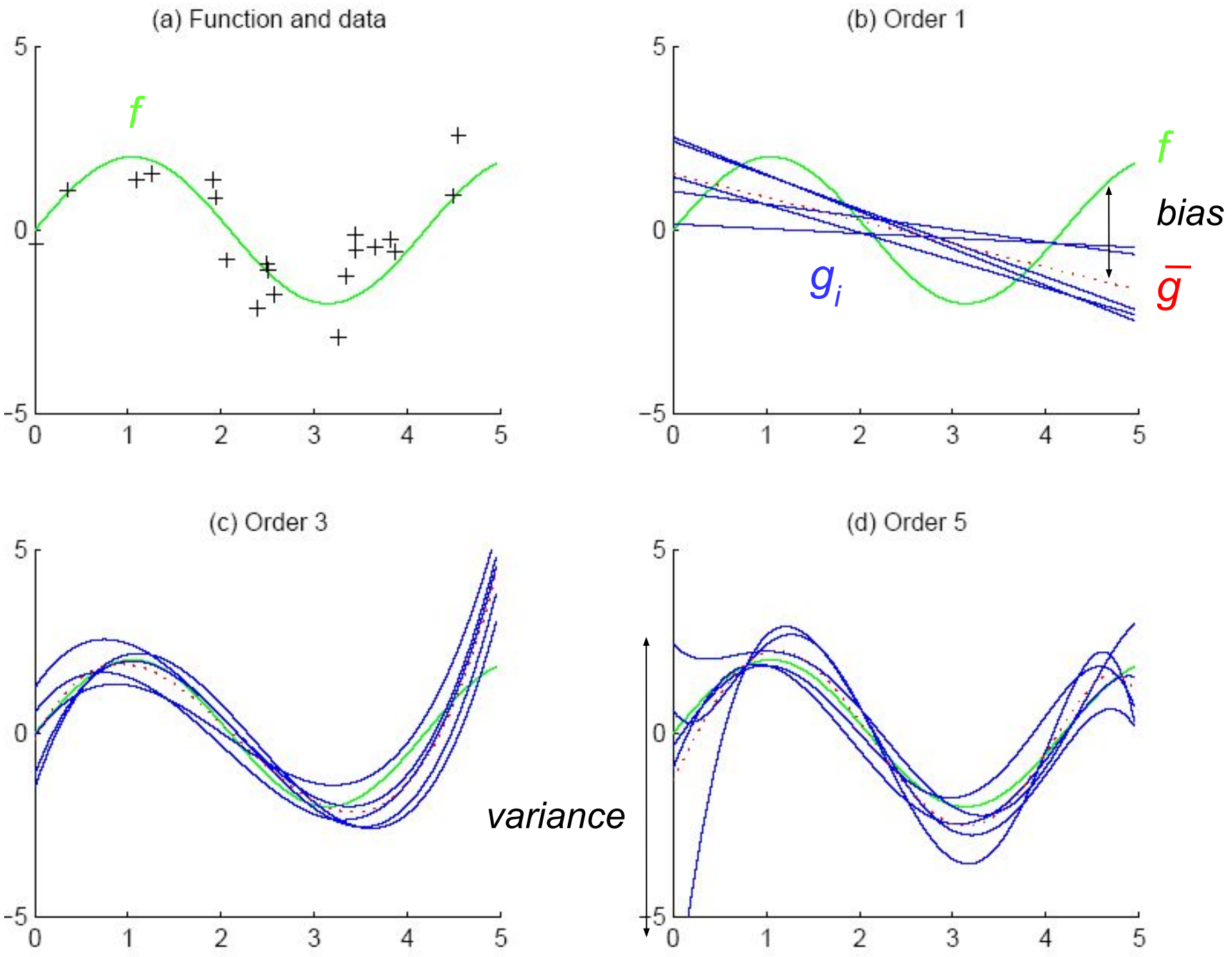
- M samples $\mathcal{X}_i=\{x^t_i, r^t_i\}$, $i=1,\ldots,M$
  are used to fit $g_i(x)$, $i=1,\ldots,M$

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

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

$$\bar{g}(x)=\frac{1}{M}\sum_t g_i(x)$$

$$\underbrace{E\left[(d-E[d])^2\right]}_{\text{variance}}+\underbrace{(E[d]-\theta)^2}_{\text{bias}^2}$$


(a) Function and data
f
(b) Order 1
f
bias
g_i
$\bar{g}$
(c) Order 3
variance
(d) Order 5

Bias/Variance Dilemma

- Example: $g_i(x)=2$ has no variance and high bias

  $g_i(x)= \sum_t r_i^t/\mathrm{N}$ has lower bias with variance

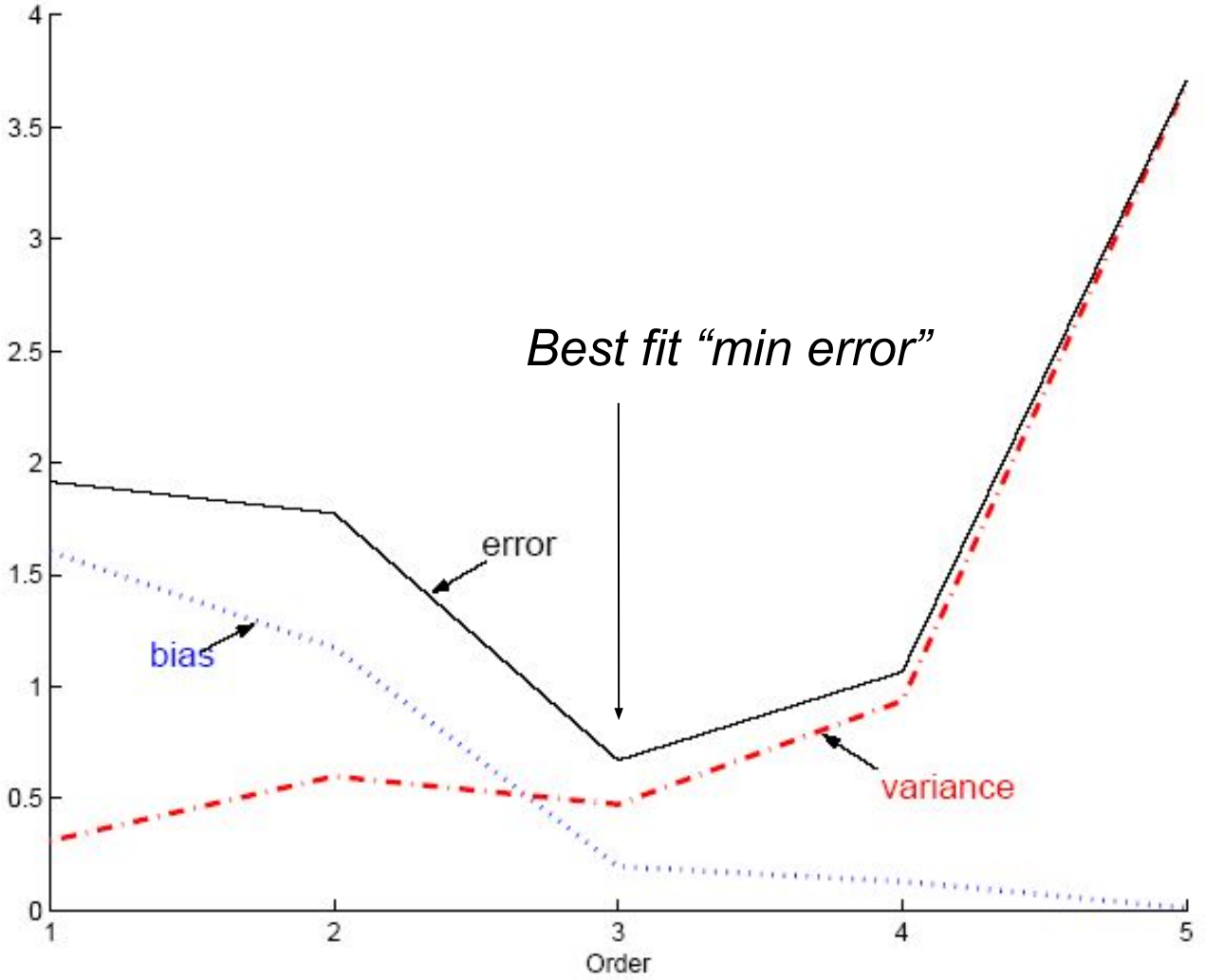
- As we increase complexity,

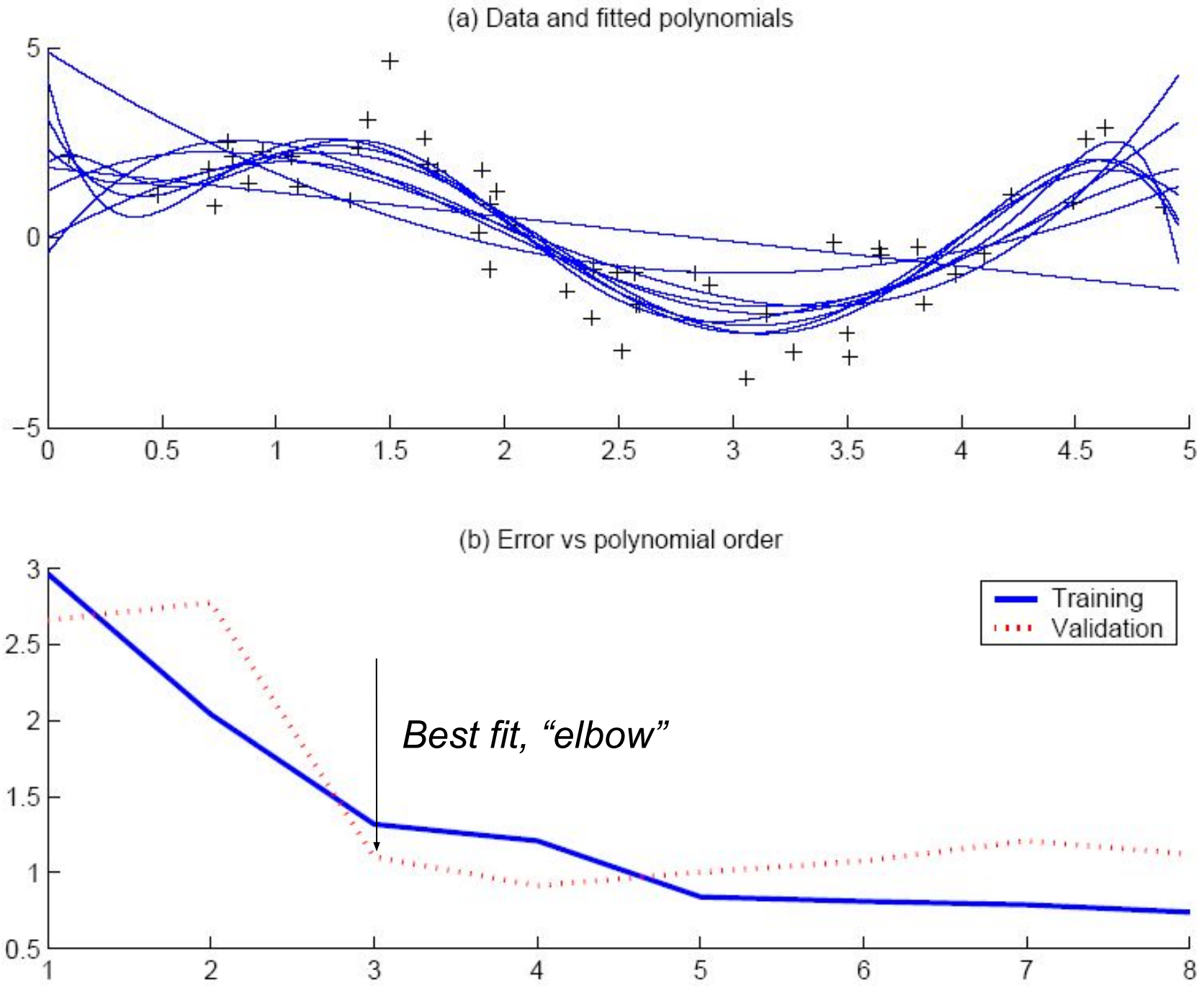
  bias decreases (a better fit to data) and

  variance increases (fit varies more with data)

- Bias/Variance dilemma: (Geman et al., 1992)

Polynomial Regression




(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
3
2.5
2
1.5
1
0.5
1
2
3
4
5
6
7
8
Training
Validation
Best fit, “elbow”

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 15)