What is Statistics?

- American Heritage Dictionary® defines statistics as: "The mathematics of the collection, organization, and interpretation of numerical data, especially the analysis of population characteristics by inference from sampling."
- The Merriam-Webster's Collegiate Dictionary® definition is: "A branch of mathematics dealing with the collection, analysis, interpretation, and presentation of masses of numerical data."

- "I like to think of statistics as the science of learning from data ... . It presents exciting opportunities for those who work as professional statisticians. Statistics is essential for the proper running of government, central to decision making in industry, and a core component of modern educational curricula at all levels." by Jon Kettenring, ASA President, 1997

- Statistics is neither really a science nor a branch of mathematics. It is perhaps best considered as a meta-science (or meta-language) for dealing with data collection, analysis, and interpretation. As such its scope is enormous and it provides much guiding insight in many branches of science, business, etc. Critical statistical reasoning can be extremely useful for making sense of the ever increasing amount of information becoming available (e.g. via the web).

- The purpose of statistics is to develop and apply methodology for extracting useful knowledge from both experiments and data. In addition to its fundamental role in data analysis, statistical reasoning is also extremely useful in data collection (design of experiments and surveys) and also in guiding proper scientific inference (Fisher, 1990).

Why Statistics?

- To understand this, we must first address the nature of science and experimentation.
- A characteristic method used by Sciencetist is to study a relatively small collection of objects, say 2500 people, and a characteristic, say longevity, and through experimentation of observation, draw a conclusion appropriate for the entire class of objects (i.e. people in general).

- Inductive Reasoning: From sample to population
- Deductive Reasoning: From the general to the particular
- Statistics then becomes a bridge between the inductive uncertainty of science and the deductive certainty of mathematics. In his classic book,

- “The Design of Experiments,” by Sir Ronald A. Fisher expresses this idea beautifully: We may at once admit that any inference from the particular to the general must be attended with some degree of uncertainty, but this is not the same as to admit that such inference cannot be absolutely rigorous, for the nature and degree of the uncertainty may itself be capable of rigorous expression.

- Statistics, therefore, is the mathematical method by which the uncertainty inherent in the scientific method is rigorously quantified.

Why Probability?

- Probability is an important concept for making forecasts and risk assessments.
- "What is Probability?"
  by Saunder, Simon (2004), in Quantum Mechanics, A. Elitzur, S. Dolev, and N. Kolenda, eds., Springer-Verlag.

Chapter 4:

Parametric Methods

Parametric Estimation

- $\mathcal{X} = \{ x^t \}_t$ where $x^t \sim p(x)$
- Parametric estimation:

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

  e.g., $\mathcal{N}(\mu, \sigma^2)$ where $\theta = \{ \mu, \sigma^2 \}$

Why Parametric?

- Simplification (in both operation and interpretation)
- How to check parametric assumptions?
- No model is the true model! All models are approximations to the true model. Some models may be better than others under some criterions.

Maximum Likelihood Estimation

- Likelihood of θ given the sample $\mathcal{X}$

  $$l(\theta|\mathcal{X}) = p(\mathcal{X}|\theta) = \prod_t p(x^t|\theta)$$

- Log likelihood

  $$\mathcal{L}(\theta|\mathcal{X}) = \log l(\theta|\mathcal{X}) = \sum_t \log p(x^t|\theta)$$

- Maximum likelihood estimator (MLE)

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

Ex.: Bernoulli Multinomial

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

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

  $\mathcal{L}(p_o|\mathcal{X}) = \log \prod_t p_o^{x^t} (1 - p_o)^{(1-x^t)}$

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

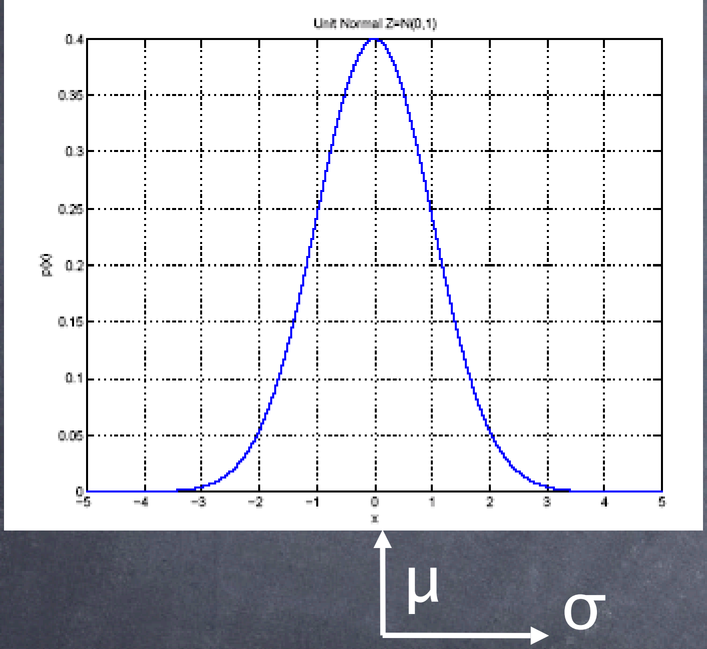
- Multinomial: K>2 states, x_i in {0,1}

  $P(x_1,x_2,\dots,x_K) = \prod_i p_i^{x_i}$

  $\mathcal{L}(p_1,p_2,\dots,p_K|\mathcal{X}) = \log \prod_t \prod_i p_i^{x_i^t}$

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

Gaussian (Normal) Distribution



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

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

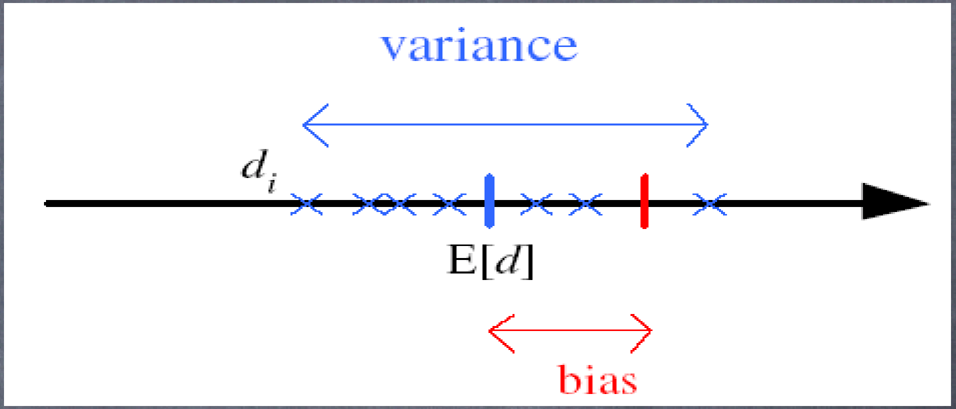
- MLE for μ and σ^2:

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

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

Bias and Variance

Unknown parameter θ
Estimator $d_i = d(\mathcal{X}_i)$ on sample X_i



Bias: $b_\theta(d) = E[d] - \theta$
Variance: $E[(d - E[d])^2]$

Mean square error:
$r(d,\theta) = E[(d-\theta)^2]$
$= (E[d] - \theta)^2 + E[(d-E[d])^2]$
= Bias^2 + Variance

Bayes' Estimator

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

Bayes' Estimator: Example

- $x^t \sim \mathcal{N}(\theta, \sigma_o^2)$ and $\theta \sim \mathcal{N}(\mu, \sigma^2)$
- $\theta_{ML} = m$
- $\theta_{MAP} = \theta_{Bayes'} =$

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

Parametric Classification

$p_i(x) = p(x|C_i)P(C_i)$ or, equivalently

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

Ex:

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

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

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

$$X \in R \qquad r_i^t = \begin{cases} 1 & \text{if } x^t \in C_i \\ 0 & \text{if } x^t \in C_j,\ j \neq i \end{cases}$$

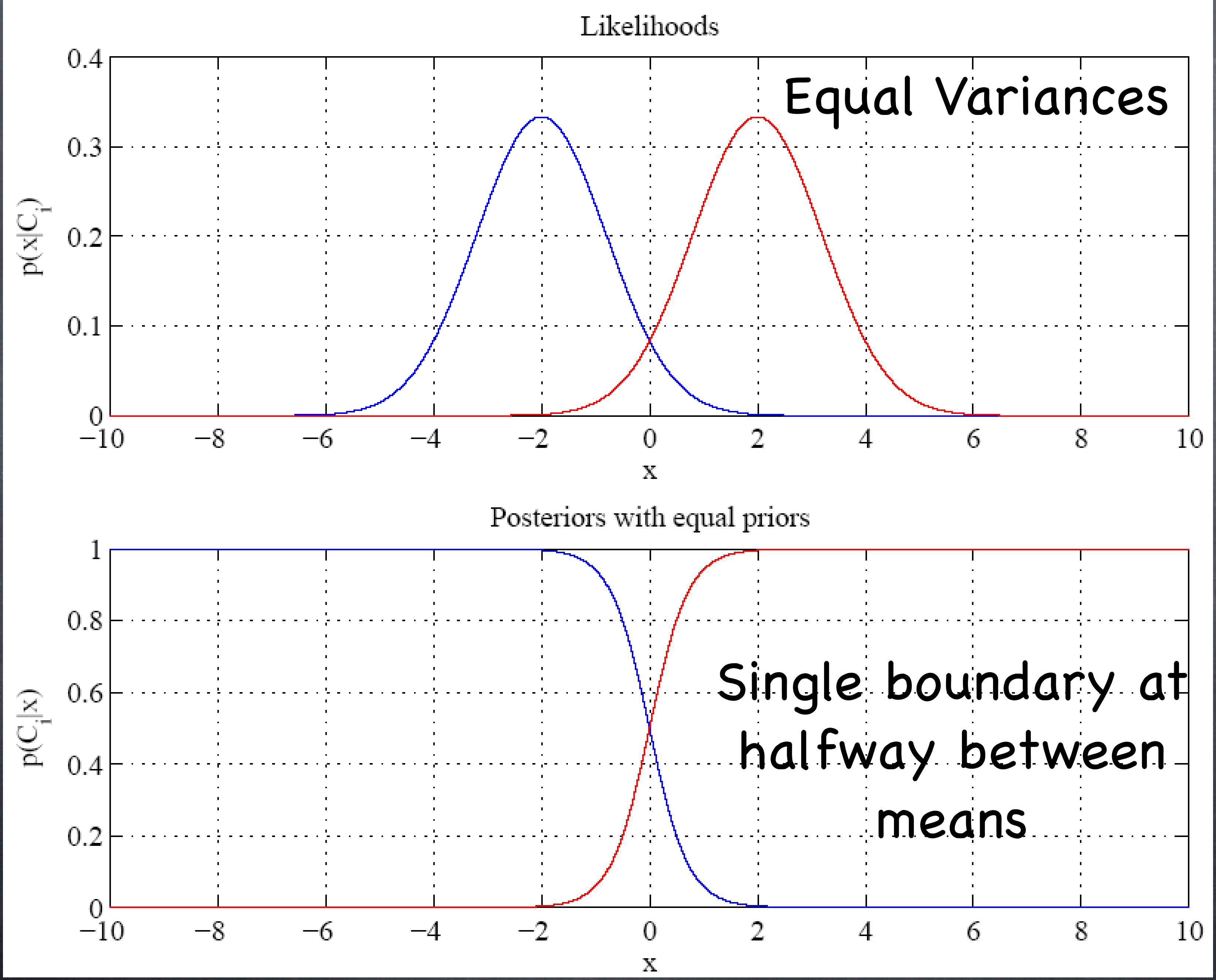
- ML estimates are

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

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

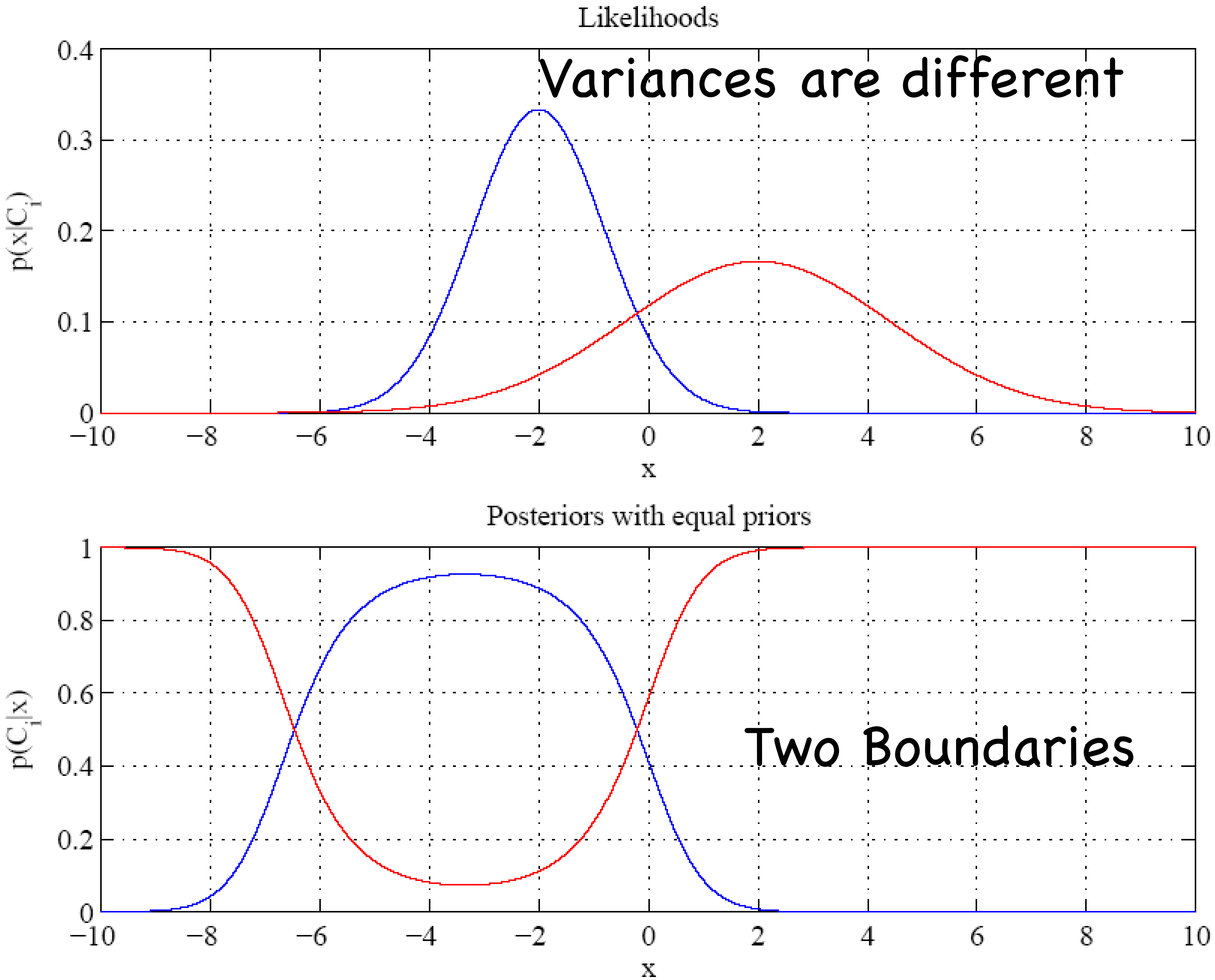
- Discriminant becomes

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



Equal Variances

Single boundary at halfway between means



Likelihoods

Variances are different

Posteriors with equal priors

Two Boundaries

Regression

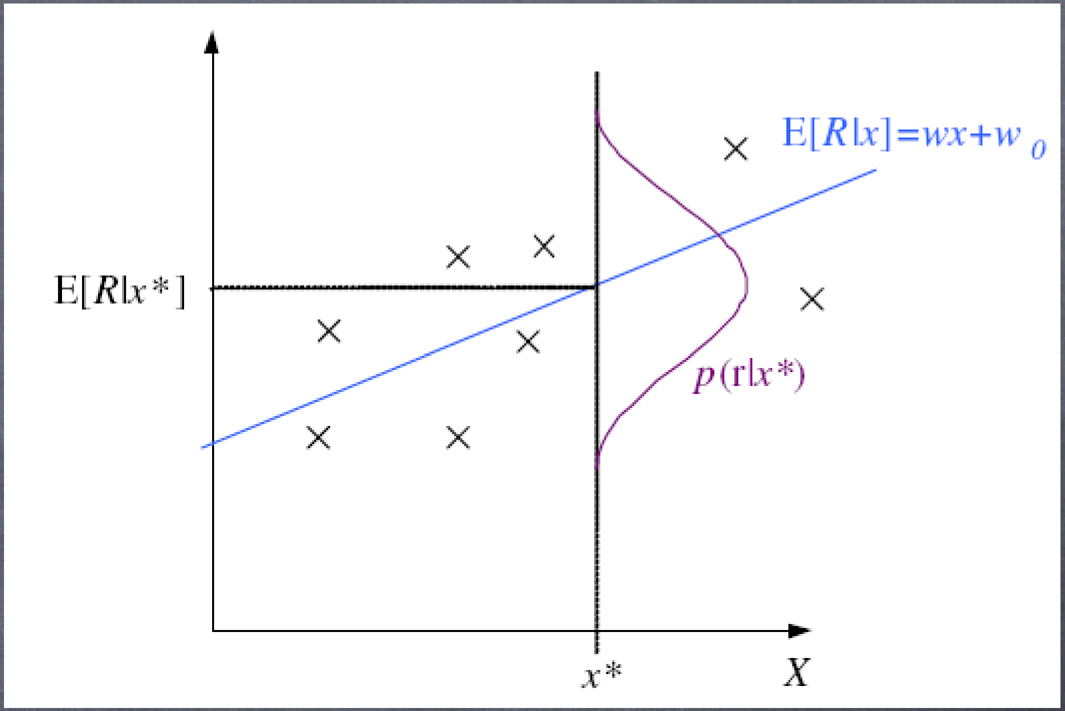
$r = f(x) + \epsilon$

$f(x)$ is the true model

estimator: $g(x|\theta)$

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

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



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

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

Regression: From LogL to Error

Maximizing

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

$$= -N \log \sqrt{2\pi}\sigma - \frac{1}{2\sigma^2} \sum_{t=1}^{N} [r^t - g(x^t|\theta)]^2$$

equivalent to minimizing

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

Skip

Linear Regression

$$g\left(x^t \mid w_1, w_0\right) = w_1 x^t + w_0$$

$$\sum_t r^t = N w_0 + w_1 \sum_t x^t$$

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

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

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

Skip

Polynomial Regression

$$g\left(x^t \mid w_k,\ldots,w_2,w_1,w_0\right) = w_k\left(x^t\right)^k + \cdots + w_2\left(x^t\right)^2 + w_1 x^t + w_0$$

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

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

Other Error Measures

- Square Error:

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

- Relative Square Error:

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

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

- ϵ-sensitive Error: (Robust)

1: indicator function

$$E(\theta|\mathcal{X}) = \sum_t 1(|r^t - g(x^t|\theta)|>\epsilon)\ (|r^t - g(x^t|\theta)| - \epsilon)$$

Bias and Variance

$$E[(r - g(x))^2|x] = E[(r - E[r|x])^2|x)] \quad + \quad (E[r|x] - g(x))^2$$

noise square error

$$\begin{aligned} E_x[(E[r|x] - g(x))^2|x] =& (E[r|x] - E_x[g(x)])^2 \; \text{Bias} \\ &+ E_x[(g(x) - E_x[g(x)])^2]\text{Variance} \end{aligned}$$

Estimating Bias and Variance

- M samples $\mathcal{X}_i=\{x^t_i, r^t_i\}$, $i=1,\ldots,M$

  are used to fit $g_i(x)$, $i=1,\ldots,M$

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

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

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

Bias/Variance Dilemma

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

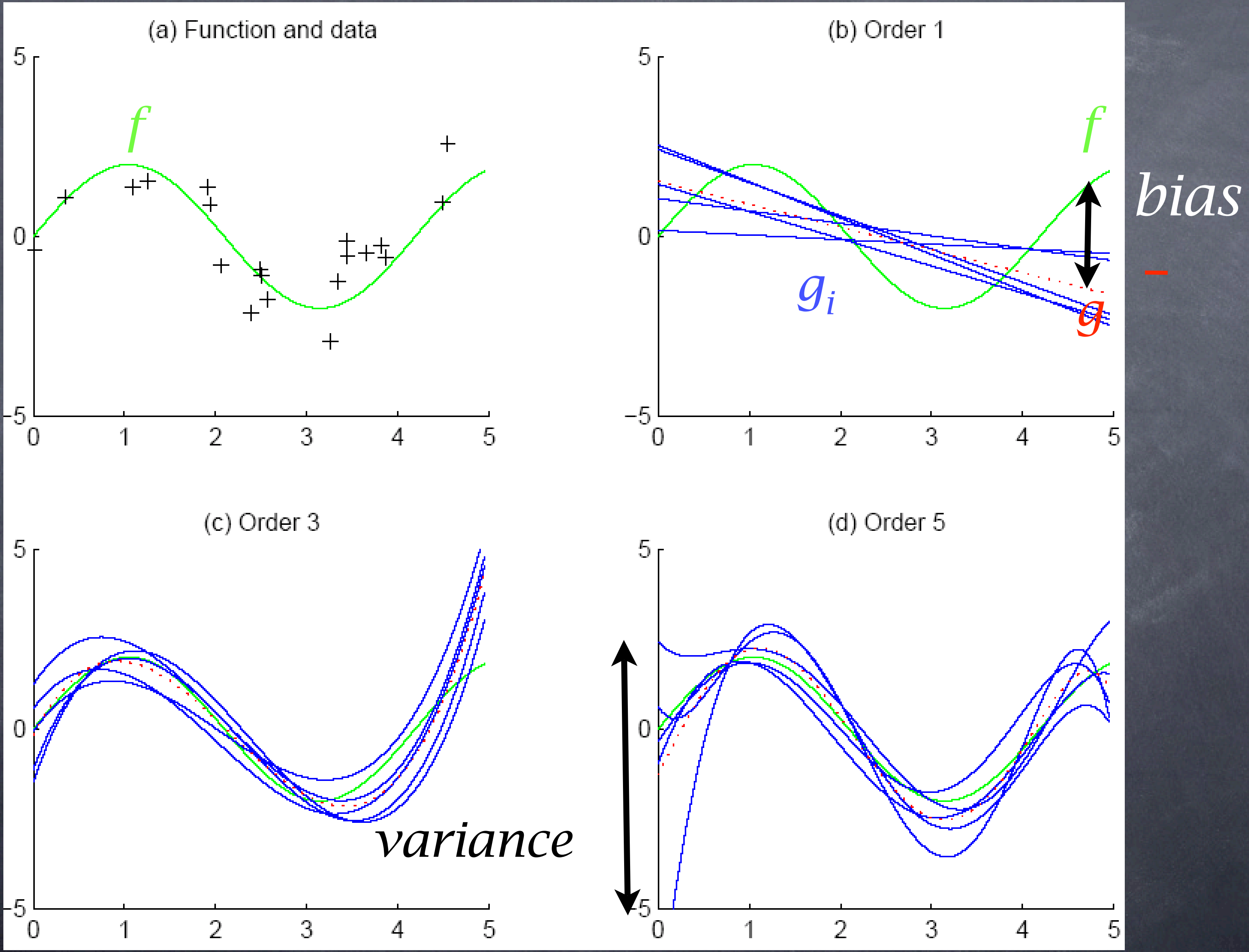
  $g_i(x)$= $\sum_t r^t{}_i$/N has lower bias with variance

- As we increase complexity,

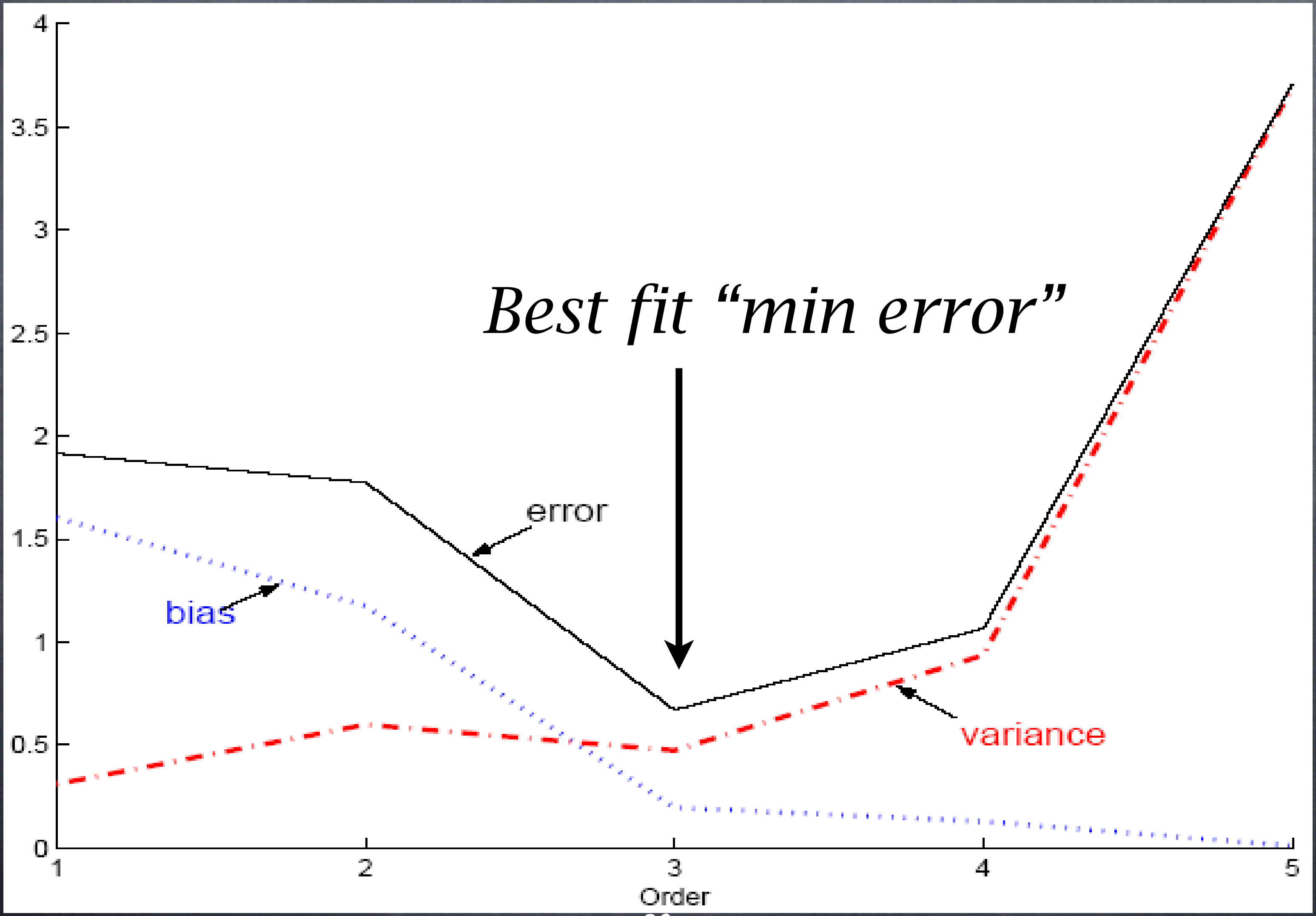
  bias decreases (a better fit to data) and

  variance increases (fit varies more with data)

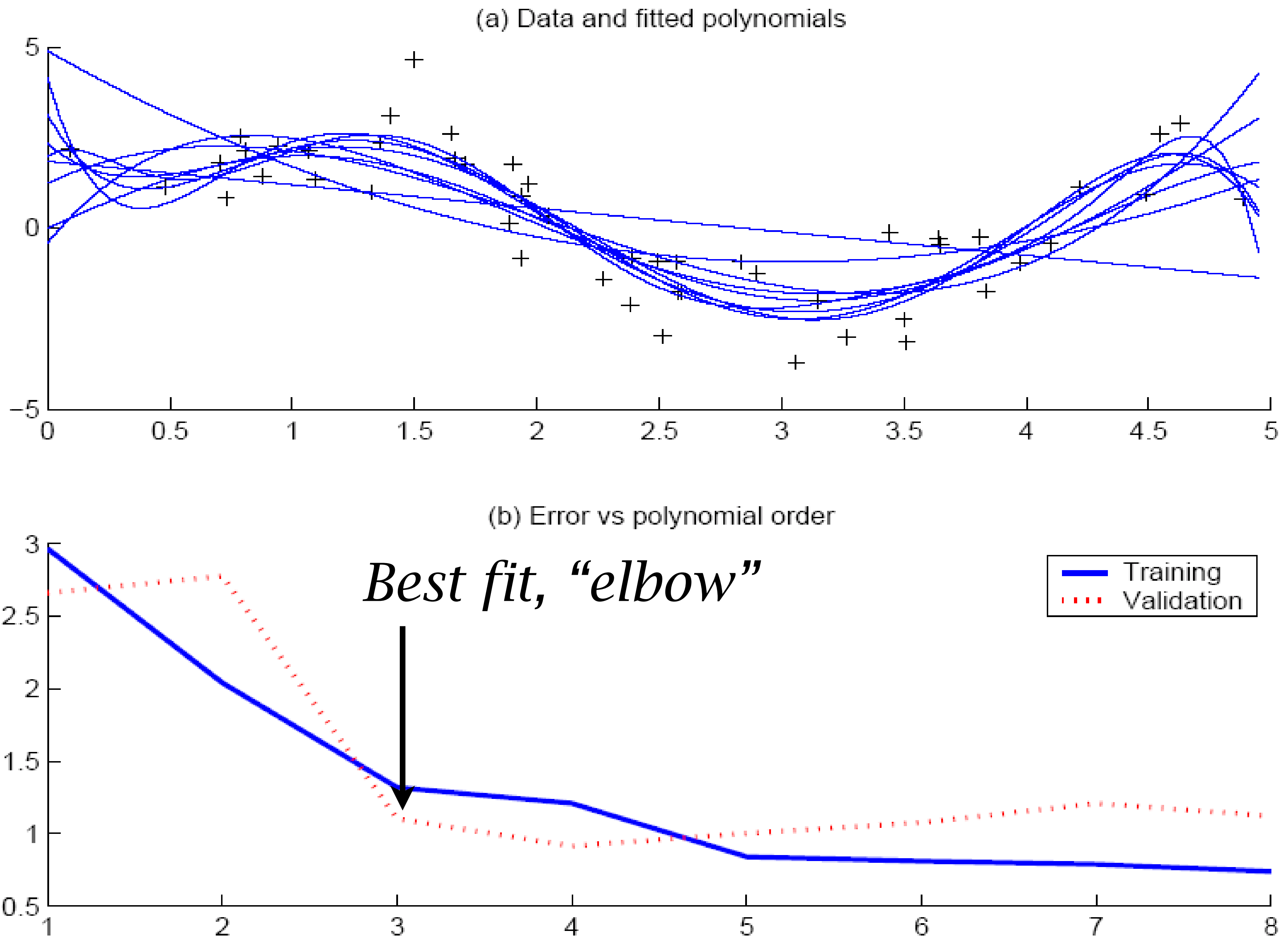
- Bias/Variance dilemma: (Geman et al., 1992)


(a) Function and data
(b) Order 1
(c) Order 3
(d) Order 5
f
f
bias
g_i
g
variance
5
0
−5
0
1
2
3
4
5

Polynomial Regression



Best fit "min error"


(a) Data and fitted polynomials

(b) Error vs polynomial order

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(model)

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

- Regularization, when prior favors simpler models
- Bayes, MAP of the posterior, p(model|data)
- Average over a number of models with high posterior (voting, ensembles: Chapter 15)