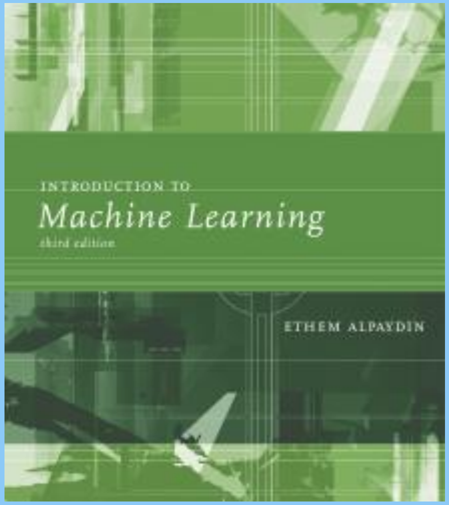


Lecture Slides for

INTRODUCTION TO MACHINE LEARNING

3RD EDITION

ETHEM ALPAYDIN

alpaydin@boun.edu.tr
http://www.cmpe.boun.edu.tr/~ethem/i2ml3e

CHAPTER 4:

PARAMETRIC METHODS

Parametric Estimation

- A statistic is any value that is calculated from a given sample.
- The advantage of the parametric approach is that the model is defined up to a small number of parameters—for example, mean, variance—the sufficient statistics of the distribution.
- $X = \{ x^t \}_t$ where $x^t \sim p(x \mid \theta)$
- Parametric estimation:

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

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

Maximum Likelihood Estimation

- Likelihood of θ given the sample X

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

- Log likelihood

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

- Maximum likelihood estimator (MLE)

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

Examples: Bernoulli

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

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

$$\begin{cases} E[X] = \sum_x x p(x) = 1.p + 0.(1-p) = p \\ Var(X) = \sum_x (x - E[x])^2 p(x) = p(1-p) \end{cases}$$

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

$$L(p \mid X) = \sum_t \left\{ x^t \log p + (1-x^t) \log(1-p) \right\}$$

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

MLE: $\dfrac{\partial L}{\partial p} = 0, \Rightarrow \hat{p} = \dfrac{1}{N} \sum_t x^t$

Examples: Multinomial

- Multinomial: $K > 2$ states, x_i in $\{0,1\}$
- Let $x_1, x_2, \ldots, x_K$ are the indicator variables where x_i is 1 if the outcome is state i and 0 otherwise.

$$P(x_1, x_2, \ldots, x_K) = \prod_i p_i^{x_i} \qquad \sum_{i=1}^{K} p_i = 1$$

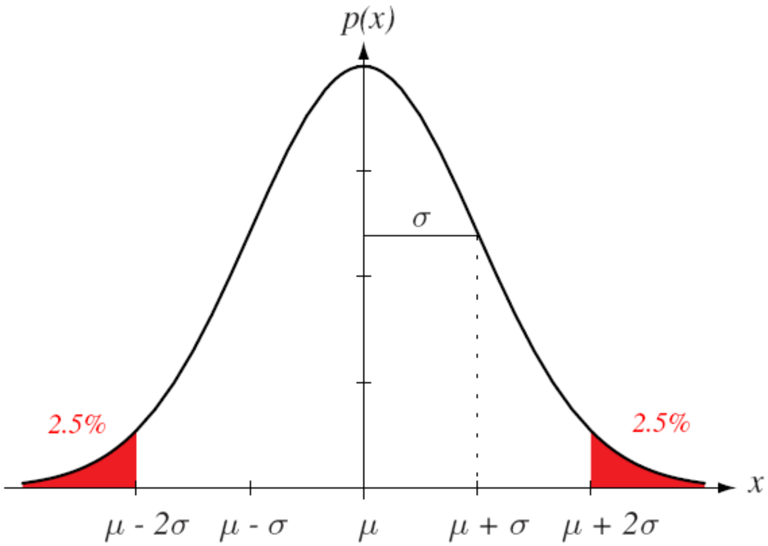
$$L(p_1, p_2, \ldots, p_K \mid X) = \log \prod_t \prod_i p_i^{x_i^t} = \log \prod_i p_i^{\sum_t x_i^t} = \log \prod_i p_i^{m^t}$$

$$= \sum_{i=1}^{K} m^t \log p_i \,, \quad m^t = \sum_t x_i^t = \text{ number of observations of } x_i^t = 1$$

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

MLE: $p_i = \frac{1}{N} \sum_t x_i^t$ (why?)

Gaussian (Normal) Distribution



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

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

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

- MLE for μ and σ^2:

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

Bias and Variance

Unknown parameter θ, Estimator $d_i = d(X_i)$ on sample X_i

The mean square error of the estimator d

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

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

If $b_\theta(d) = 0$ for all θ values, d is an unbiased estimator

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

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

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

The MLE of σ^2

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

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

Given that $Var(X) = E[X^2] - E[X]^2$, we get $E[X^2]=Var(X) + E[X]^2$, and we can write:

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

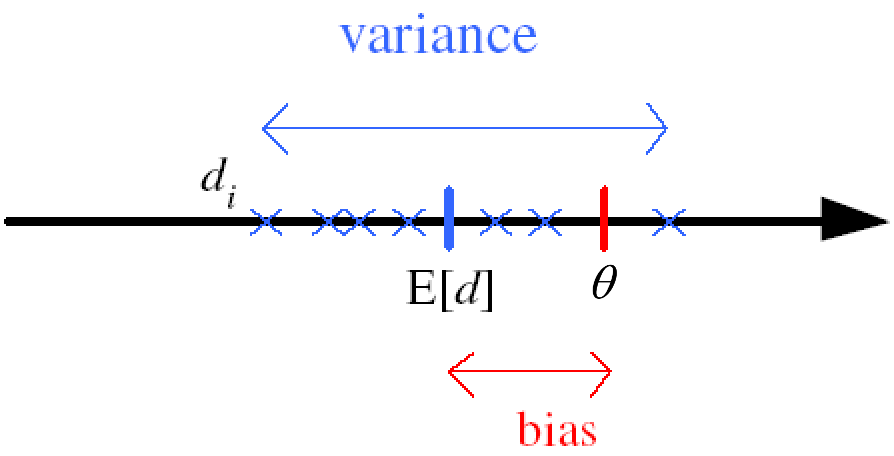
$$E[s^2] = \frac{N(\sigma^2 + \mu^2) - N(\sigma^2/N + \mu^2)}{N} = \left(\frac{N-1}{N}\right)\sigma^2 \neq \sigma^2$$

which shows that s^2 is a biased estimator of σ^2.

This is an example of an asymptotically unbiased estimator whose bias goes to 0 as N goes to infinity.

Mean square error: (Proof: Refer to textbook)

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



Bayes' Estimator

- Treat θ as a random var with prior $p(\theta)$
- Bayes' rule: $p(\theta|X) = p(X|\theta)\, p(\theta) / p(X)$
- $p(x|X) = \int p(x,\theta|X)d\theta = \int p(x|\theta, X)\, p(\theta|X)\, d\theta$
- $= \int p(x|\theta)\, p(\theta|X)\, d\theta$
- Where $p(x|\theta,X) = p(x|\theta)$ because once we know θ, the sufficient statistics, we know everything about the distribution.
- Evaluating the integrals may be quite difficult, except in cases where the posterior has a nice form.

Bayes' Estimator

- If we can assume that $p(\theta|X)$ has a narrow peak around its mode, then using the maximum a posteriori (MAP) estimate will make the calculation easier.
- Maximum a Posteriori (MAP):

$$\theta_{\mathrm{MAP}} = \mathrm{argmax}_{\theta}\, p(\theta|\mathrm{X})$$

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

- Maximum Likelihood (ML): $\theta_{\mathrm{ML}} = \mathrm{argmax}_{\theta}\, p(\mathrm{X}|\theta)$
- Bayes' Estimator: $\theta_{\mathrm{Bayes'}} = \mathrm{E}[\theta|\mathrm{X}] = \int \theta\, p(\theta|\mathrm{X})\, d\theta$

Bayes' Estimator: Example

- $x^t \sim N(\theta, \sigma^2)$ and $\theta \sim N(\mu_0, \sigma_0^2)$
- $\theta_{ML} = m$
- $\theta_{MAP} = \theta_{Bayes'} =$

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

Parametric Classification

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

or

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

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

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

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

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

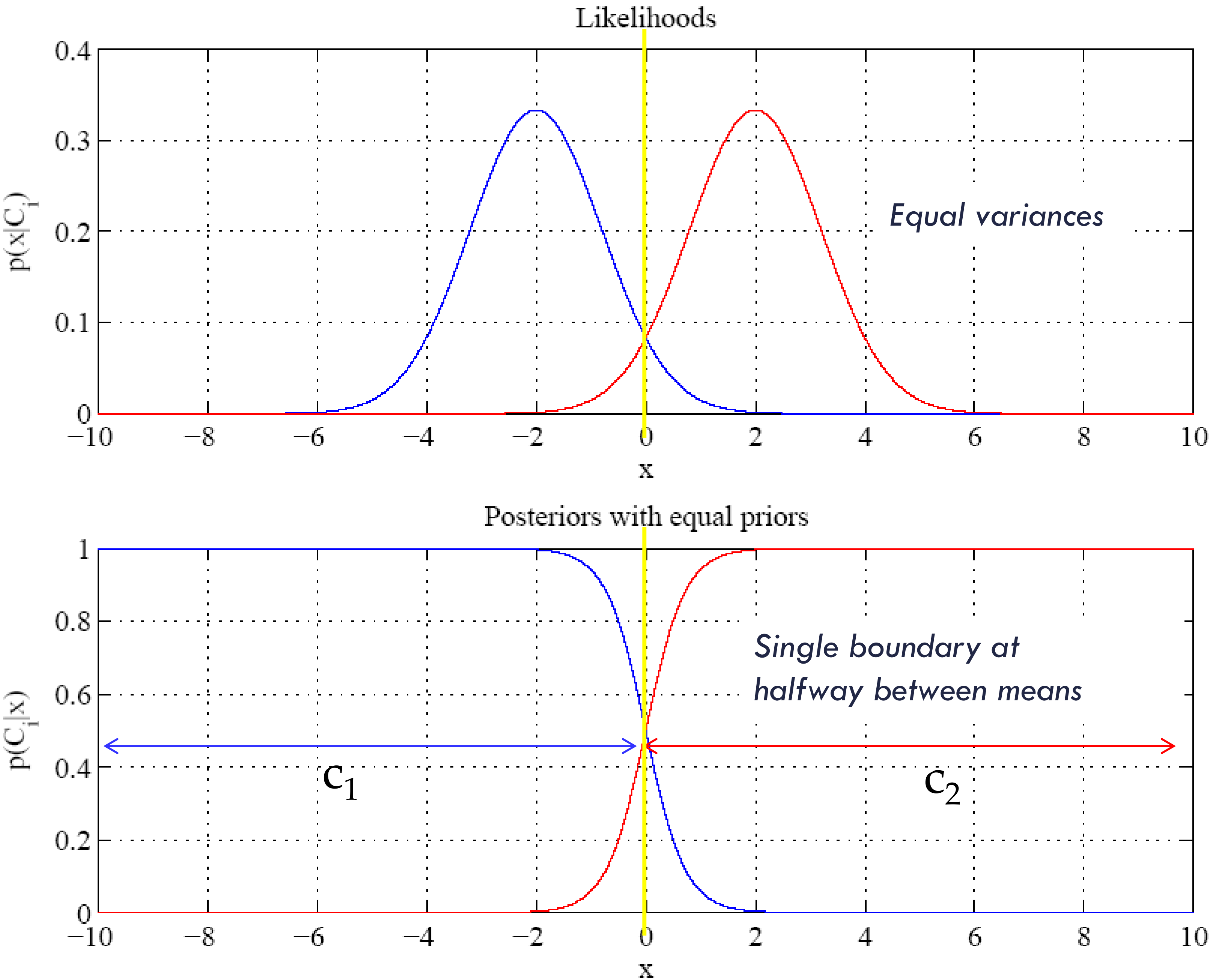
$$\mathbf{r} \in \{0,1\}^K$$

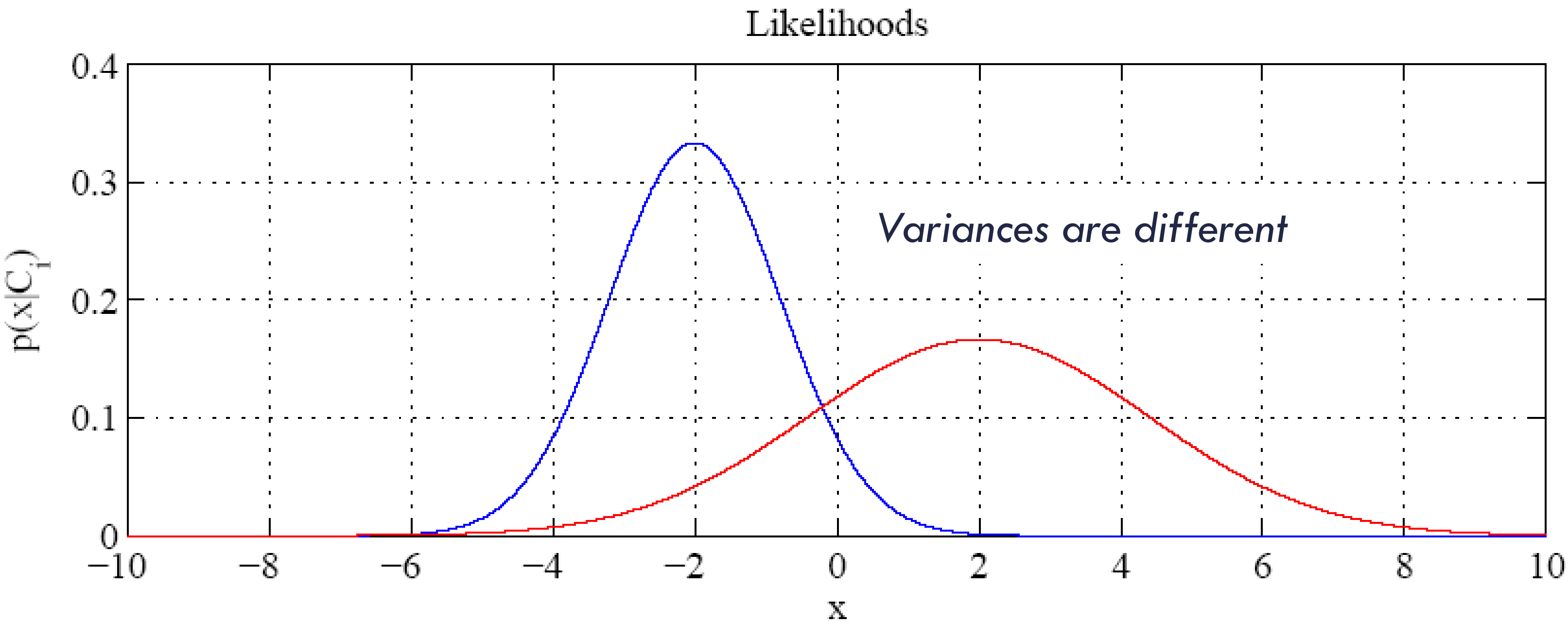
- ML estimates are

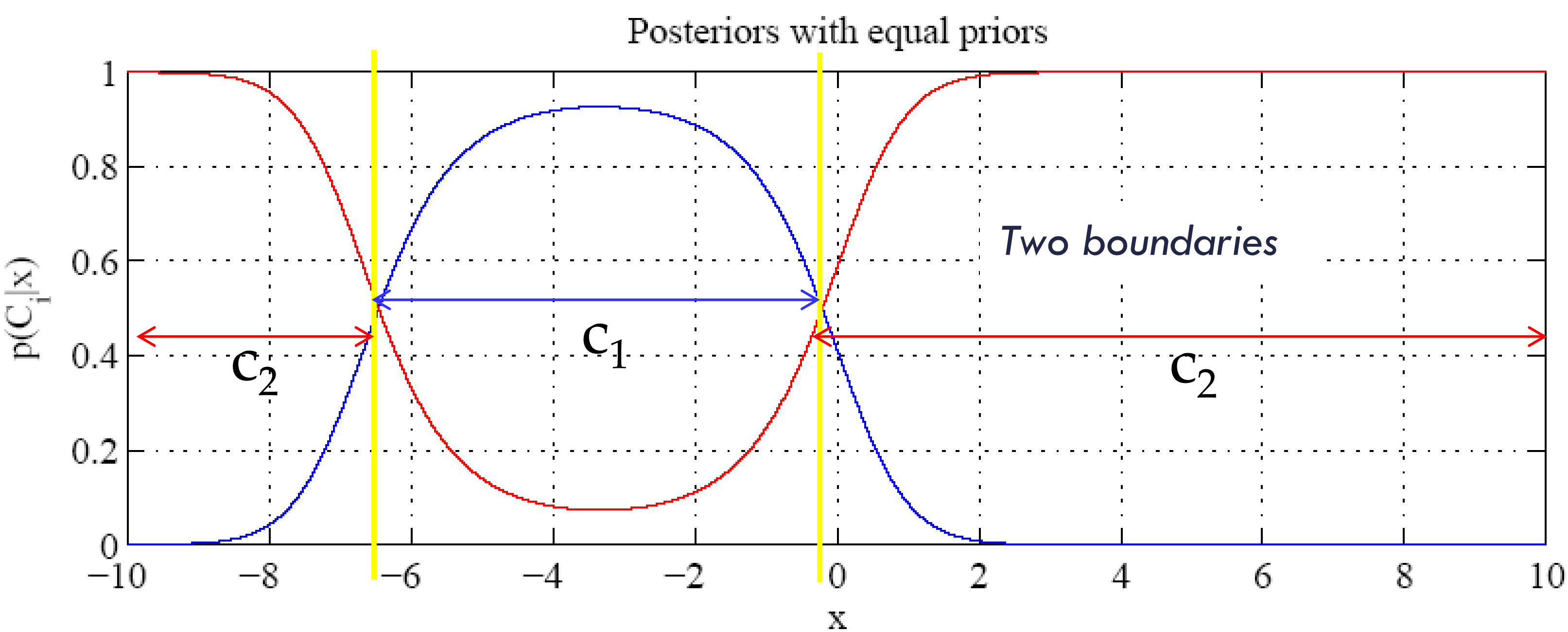
$$m_i = \frac{\sum_t x^t r_i^t}{\sum_t r_i^t}, \quad s_i^2 = \frac{\sum_t \left(x^t - m_i\right)^2 r_i^t}{\sum_t r_i^t}, \quad \hat{P}\left(C_i\right) = \frac{\sum_t r_i^t}{N}$$

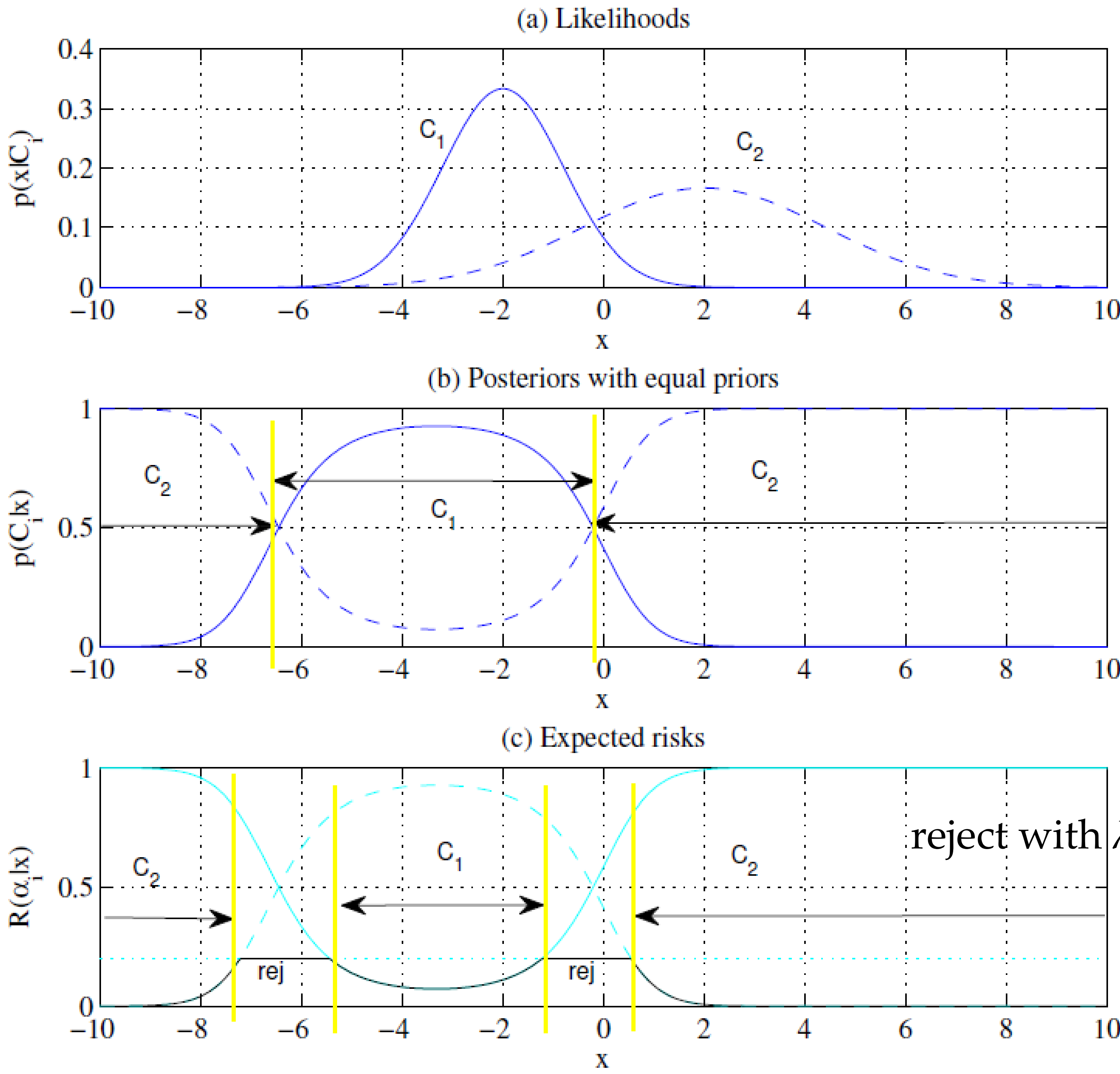
- Plugging these estimates into equation, we get the discriminant function

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


Likelihoods
0.4
0.3
0.2
0.1
0
p(x|C_i)
Equal variances
−10
−8
−6
−4
−2
0
2
4
6
8
10
x
Posteriors with equal priors
1
0.8
0.6
0.4
0.2
0
p(C_i|x)
Single boundary at
halfway between means
C1
C2
−10
−8
−6
−4
−2
0
2
4
6
8
10
x


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


Posteriors with equal priors
p(C_i|x)
1
0.8
0.6
0.4
0.2
0
−10 −8 −6 −4 −2 0 2 4 6 8 10
x
Two boundaries
C2
C1
C2



reject with $\lambda = 0.2$

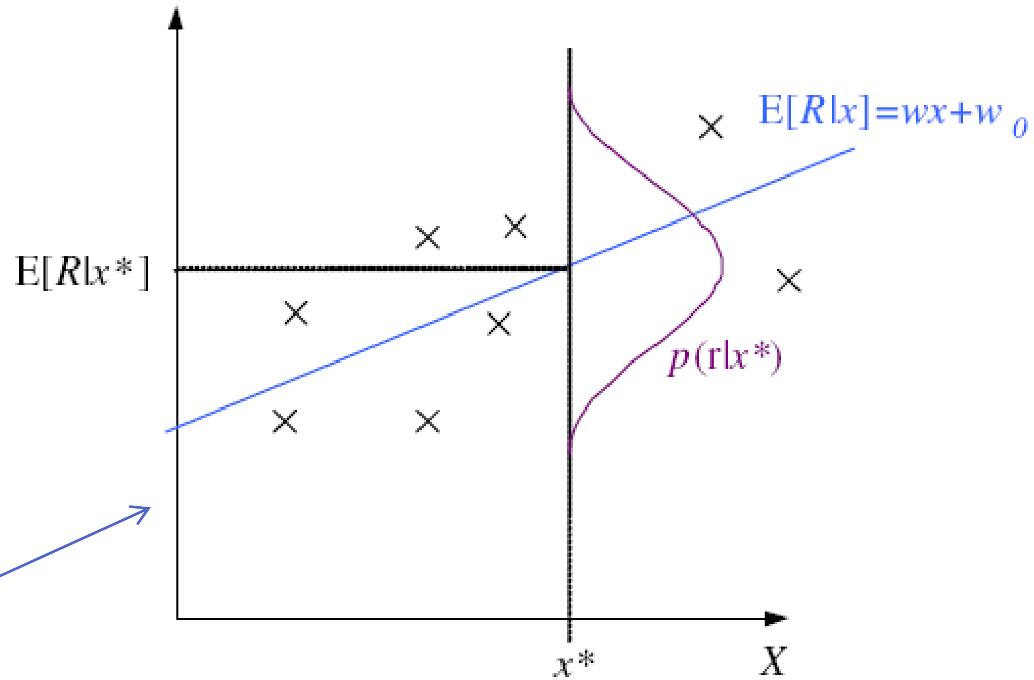
Regression

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

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

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

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



f(x) is the unknown function, which we would like to approximate by our estimator, g(x|θ), defined up to a set of parameters θ.

$$p(r,x) = p(r|x)\,p(x)$$

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

Regression: From LogL to Error

Ignoring the 2nd term since it does not depend on our estimator

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

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

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

the sum of squared errors called the least squares estimates

Linear Regression

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

taking the derivative of the sum of squared errors with respect to w_1 and w_0

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

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

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

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

Polynomial Regression

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

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

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

$$\mathbf{A} = \left(\mathbf{D}^T\mathbf{D}\right), \qquad \mathbf{y} = \mathbf{D}^T\mathbf{r}$$

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

Assuming Gaussian distributed error and maximizing likelihood corresponds to minimizing the sum of squared errors. Another measure is the relative square error (RSE).

Other Error Measures

- Square Error: $$E(\theta|\mathcal{X}) = \frac{1}{2}\sum_{t=1}^{N}\left[r^t - g\left(x^t|\theta\right)\right]^2$$
- Relative Square Error: $$E(\theta|\mathcal{X}) = \frac{\sum_{t=1}^{N}\left[r^t - g\left(x^t|\theta\right)\right]^2}{\sum_{t=1}^{N}\left[r^t - \overline{r}\right]^2}$$
- If E_{RSE} is close to 1, then our prediction is as good as predicting by the average; as it gets closer to 0, we have better fit. If E_{RSE} is close to 1, this means that using a model based on input x does not work better than using the average which would be our estimator if there were no x; if E_{RSE} is close to 0, input x helps.

- A measure to check the goodness of fit by regression is the coefficient of determination that is

$$R^2=1- E_{\mathrm{RSE}}$$

  and for regression to be considered useful, we require R^2 to be close to 1.
- Absolute Error: $E(\theta \mid \mathrm{X}) = \sum_t |r^t - g(x^t|\theta)|$
- ε-sensitive Error:

$$E(\theta \mid \mathrm{X}) = \sum_t \mathbf{1}(|r^t - g(x^t|\theta)|>\varepsilon)\,(|r^t - g(x^t|\theta)| - \varepsilon)$$

Tuning Model Complexity: Bias and Variance

Why?

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

variance _ *noise* *squared error*

1st: The variance of r given x; it does not depend on $g(\cdot)$ or X. It is the variance of noise added, σ^2 . This is the part of error that can never be removed, no matter what estimator we use.
2nd: Deviation from the regression function, $E[r \mid x]$. This does depend on the estimator and the training set.

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

bias *variance*

Estimating Bias and Variance

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

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

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

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

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,

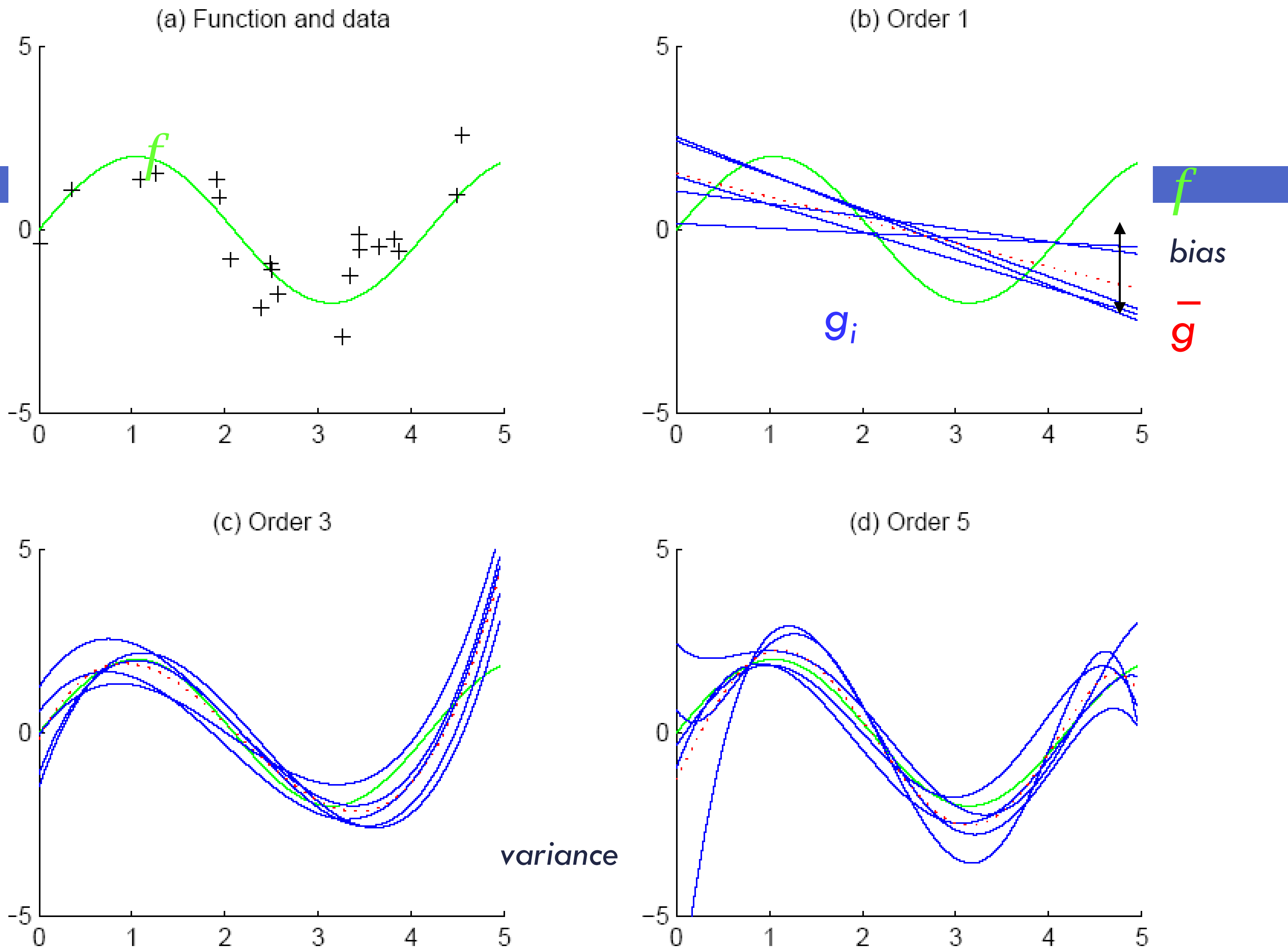
  bias decreases (a better fit to data) and

  variance increases (fit varies more with data)

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

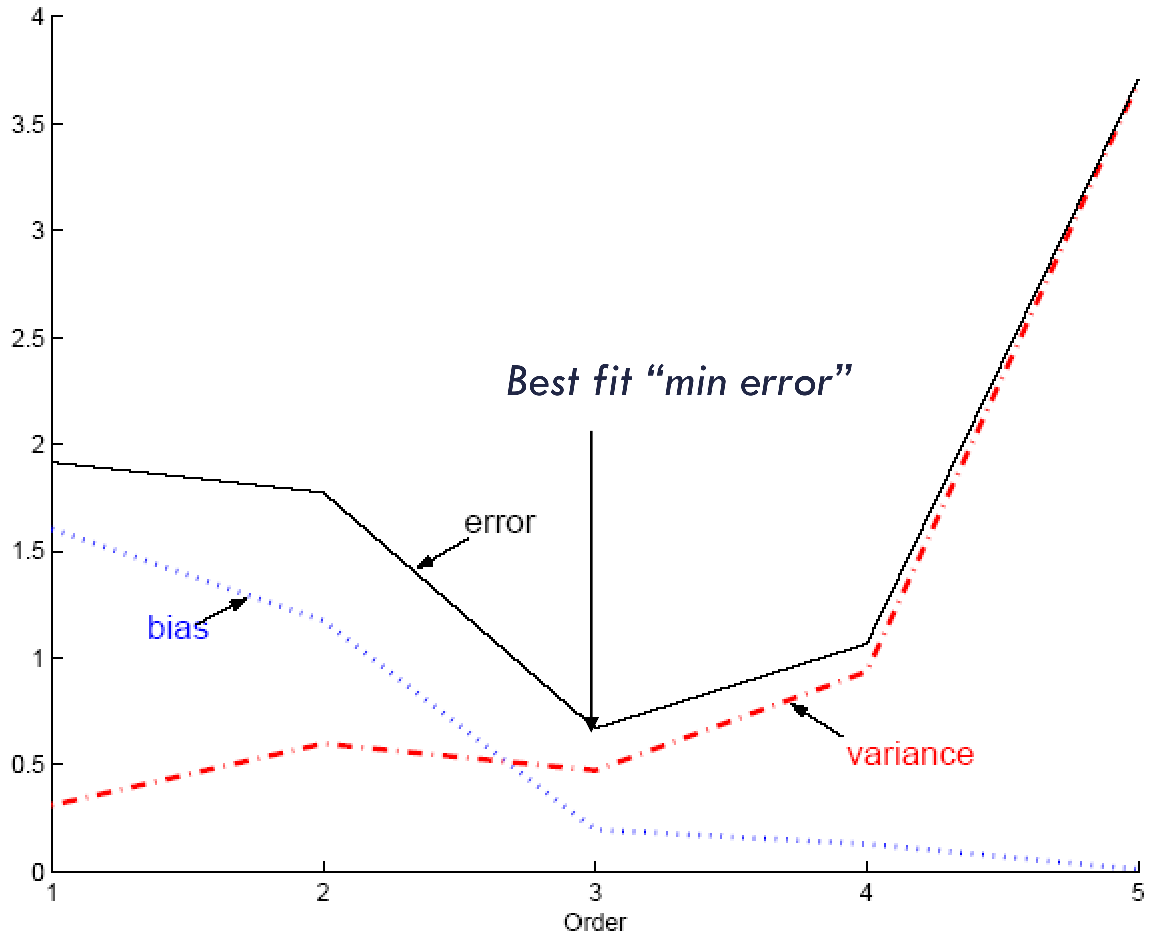
Underfitting and overfitting

- If there is bias, this indicates that our model class does not contain the solution; this is underfitting.
- If there is variance, the model class is too general and also learns the noise; this is overfitting.
- If $g(\cdot)$ is of the same hypothesis class with $f(\cdot)$, we have an unbiased estimator, and estimated bias decreases as the number of models increases.
- This shows the error-reducing effect of choosing the right model, which we called inductive bias.

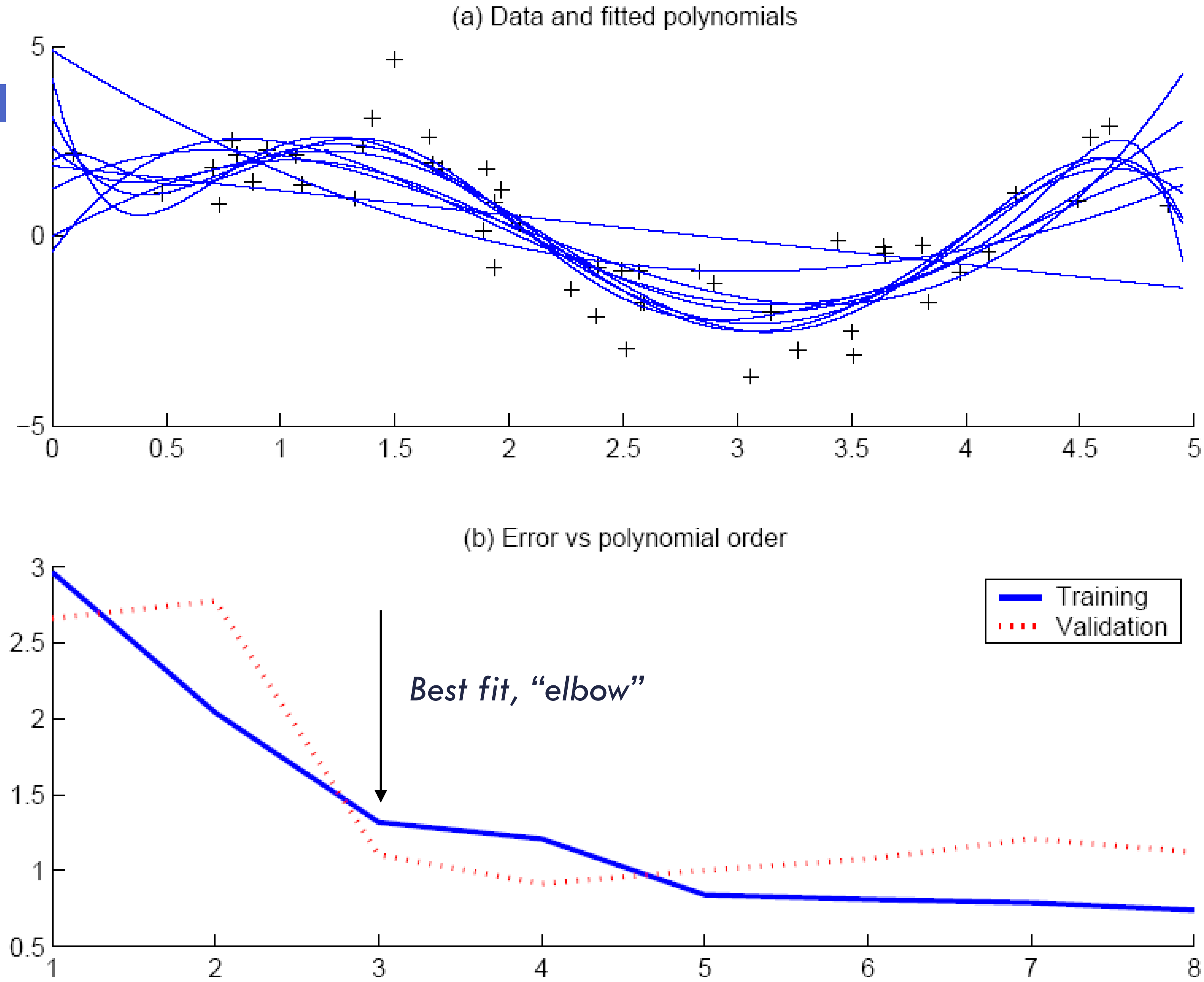


Function, $f\,(x) = 2\sin(1.5x)$, and one noisy ($N(0,\,1)$) dataset sampled from the function.

Model Selection



Cross-validation



Model Selection

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

$$E' = \textit{error on data} + \lambda \cdot \textit{model complexity}.$$

- The 2nd term that penalizes complex models with large variance, where λ gives the weight of this penalty.
- If λ is taken too large, only very simple models are allowed and we risk introducing bias. λ is optimized using cross-validation.
- Also we can consider E' as the error on new test data.

Model Selection

- The 1st term on the right is the training error and the 2nd is an optimism term estimating the discrepancy between training and test error.
- Akaike's information criterion (AIC) and Bayesian information criterion (BIC) work by estimating this optimism and adding it to the training error to estimate test error, without any need for validation.

- Structural Risk Minimization (SRM): Uses a set of models ordered in terms of their complexities. Finding the model simplest in terms of order and best in terms of empirical error on the data
- Minimum Description Length (MDL): Kolmogorov complexity, shortest description of data

Bayesian Model Selection

Regularization ($L2$):

- Prior on models, p (model)

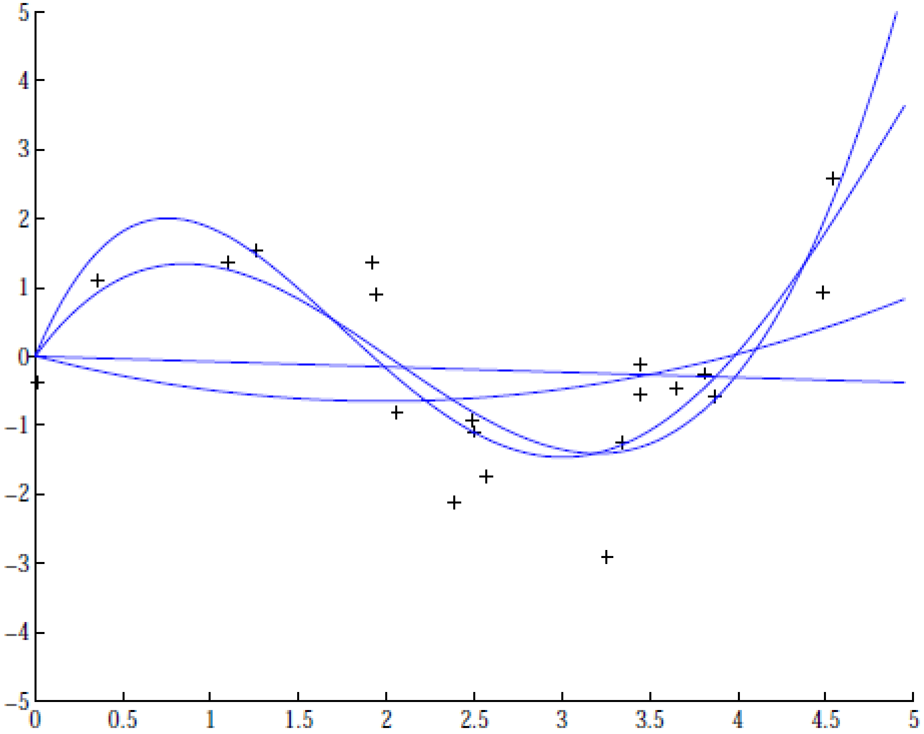
$$p\left(\text{model}|\text{data}\right)=\frac{p\left(\text{data}|\text{model}\right)\;p\left(\text{model}\right)}{p\left(\text{data}\right)}$$

- Regularization, when prior favors simpler models
- Bayes, MAP of the posterior, p (model|data)
- Average over a number of models with high posterior. If we have a regression model and use the prior $p(\mathbf{w}) \sim N(0, 1/\lambda)$, we minimize

$$E\left(\mathbf{w}|\mathrm{X}\right)=\frac{1}{2}\sum_{t=1}^{N}\left[r^t-g\left(x^t|\mathbf{w}\right)\right]^2+\lambda\sum_i w_i^2$$

- w_i are close to 0, to have smoother fitted polynomial.

Regression example



Coefficients increase in magnitude as order increases:

1: $[-0.0769, 0.0016]^T$

2: $[0.1682, -0.6657, 0.0080]^T$

3: $[0.4238, -2.5778, 3.4675, -0.0002]^T$

4: $[-0.1093, 1.4356, -5.5007, 6.0454, -0.0019]^T$