CHAPTER 4:

# PARAMETRİC METHODS SECTIONS 4.1 ~ 4.5 half of 4.6

# 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 $\theta$, its sufficient statistics, using X

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

# Maximum Likelihood Estimation



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

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

  (note the conditional independence)

- Log likelihood

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

- Maximum likelihood estimator (MLE)

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

# Examples: Bernoulli/Multinomial



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

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

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

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

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

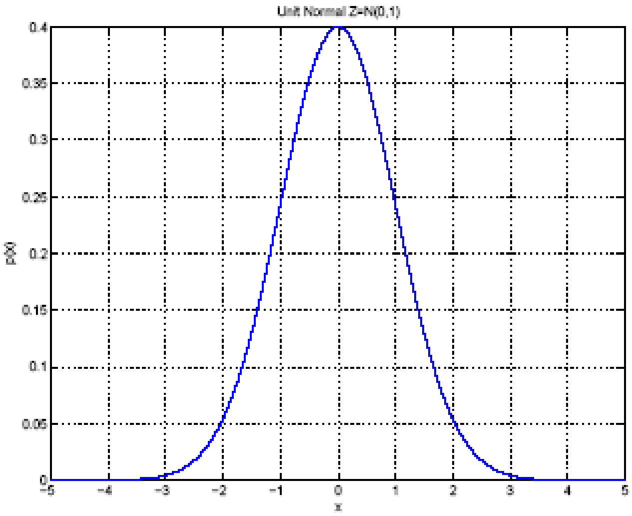
  $P(x_1, x_2, \ldots, x_K \mid p_1, p_2, \ldots, p_K) = \prod_i p_i^{x_i}$

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

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

  *See Tutorial 1 for the derivation.*

# 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 $\mu$ and $\sigma^2$:

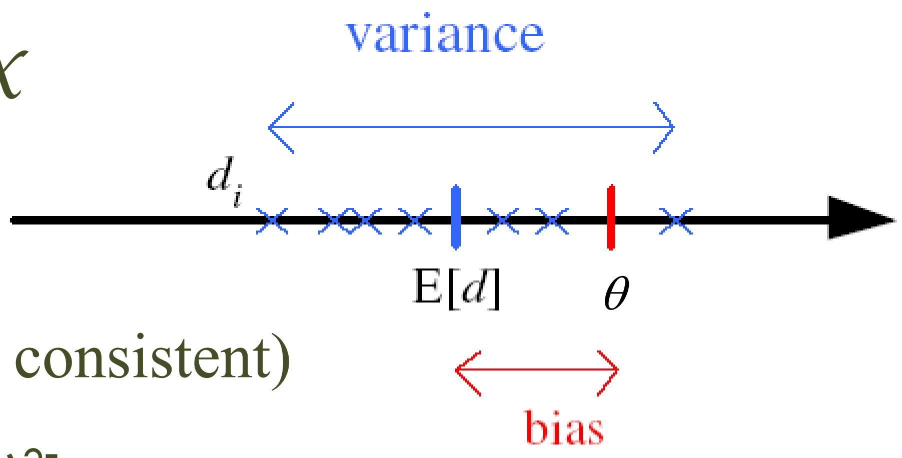
$$m = \frac{\sum_t x^t}{N}$$

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

# Evaluating an estimator: Bias and Variance



Unknown parameter $\theta$
Estimator $d = d(\mathcal{X})$ on sample $\mathcal{X}$



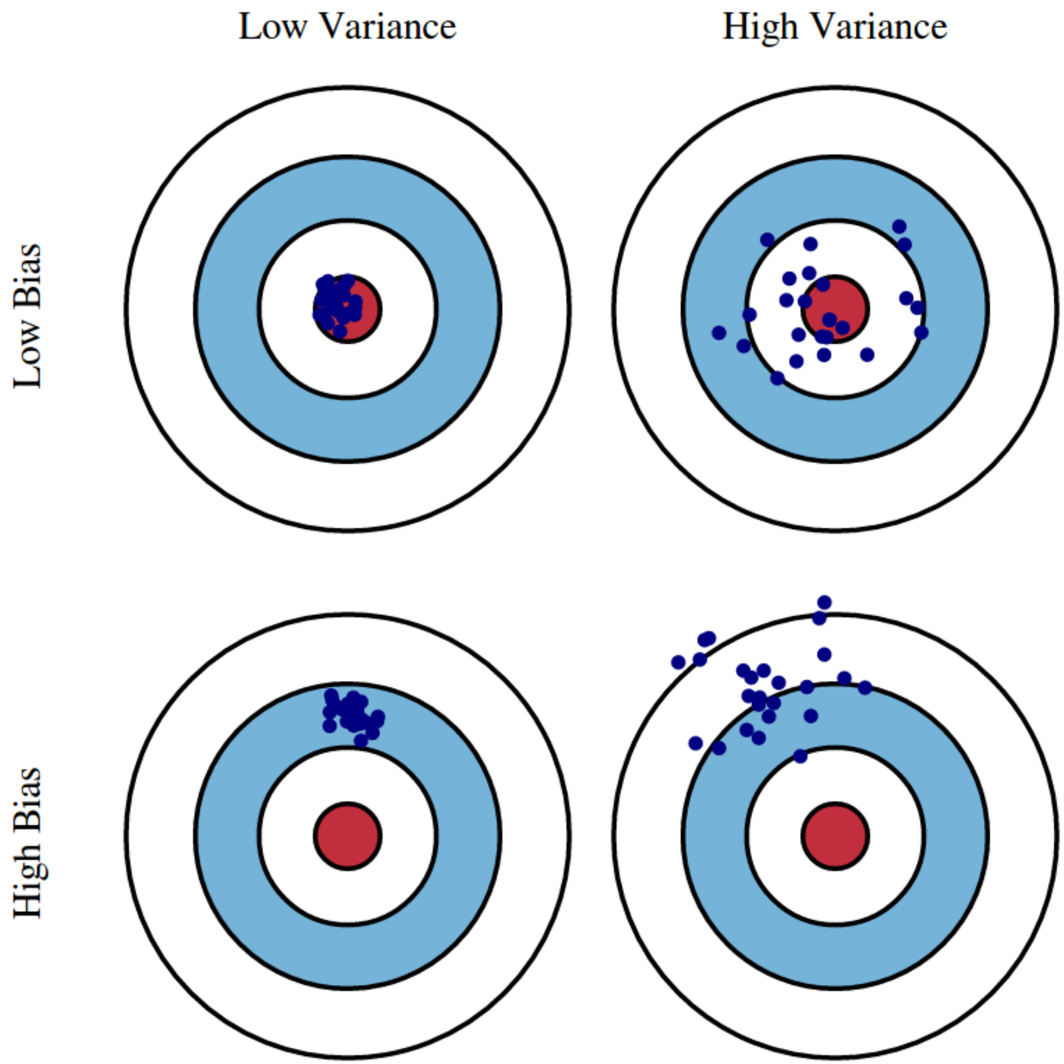

Bias: $b_\theta(d) = E_\mathcal{X}[d(\mathcal{X})] - \theta$
(0 means unbiased estimator,
e.g. sample mean, which is also consistent)

Variance: $E_\mathcal{X}[(d(\mathcal{X}) - E_\mathcal{X}[d(\mathcal{X})])^2]$

Mean square error:
$$r(d,\theta) = E[(d-\theta)^2]$$
$$= (E[d] - \theta)^2 + E[(d-E[d])^2]$$
$$= \text{Bias}^2 + \text{Variance}$$
(see textbook page 70 for a detailed derivation)

# Evaluating an estimator: Bias and Variance





# Bayes' Estimator



- Treat $\vartheta$ as a random variable with prior $p\,(\vartheta)$
- Bayes' rule: $p\,(\vartheta|\mathcal{X}) = p(\mathcal{X}|\vartheta)\ p(\vartheta)\ /\ p(\mathcal{X})$ (posterior)
- Maximum a Posteriori (MAP):

$$\vartheta_{MAP} = \text{argmax}_{\vartheta}\ p(\vartheta|\mathcal{X}) = \text{argmax}_{\vartheta}\ p(\mathcal{X}|\vartheta)\ p(\vartheta)$$

- Maximum Likelihood (ML): $\vartheta_{ML} = \text{argmax}_{\vartheta}\ p(\mathcal{X}|\vartheta)$
- Bayes': $\vartheta_{Bayes'} = E[\vartheta|\mathcal{X}] = \int \vartheta\ p(\vartheta|\mathcal{X})\ d\vartheta$

- Full: $p(x\,|\,\mathcal{X}) = \int p(x\,|\,\vartheta)\ p(\vartheta\,|\,\mathcal{X})\ d\vartheta$

# Bayes' Estimator: Example



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

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

Derivation:

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

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

# Parametric Classification

**Discriminant function**

$$g_i(x) = p(x \mid C_i)P(C_i) \propto P(C_i|x)$$

or

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

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

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

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

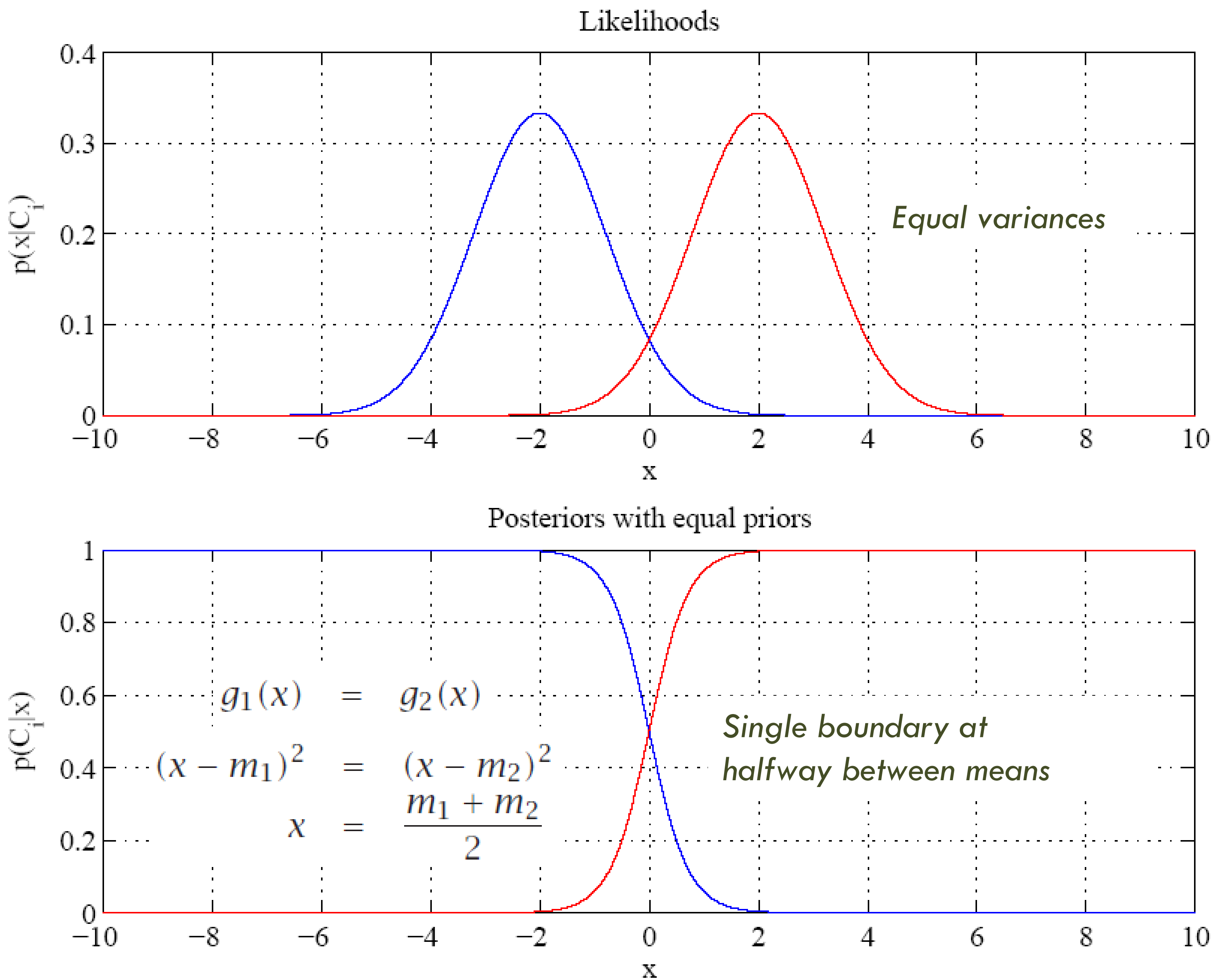
$$x \in \Re \qquad r_i^t = \begin{cases} 1 \text{ if } x^t \in C_i \\ 0 \text{ if } x^t \in C_j, j \neq i \end{cases}$$

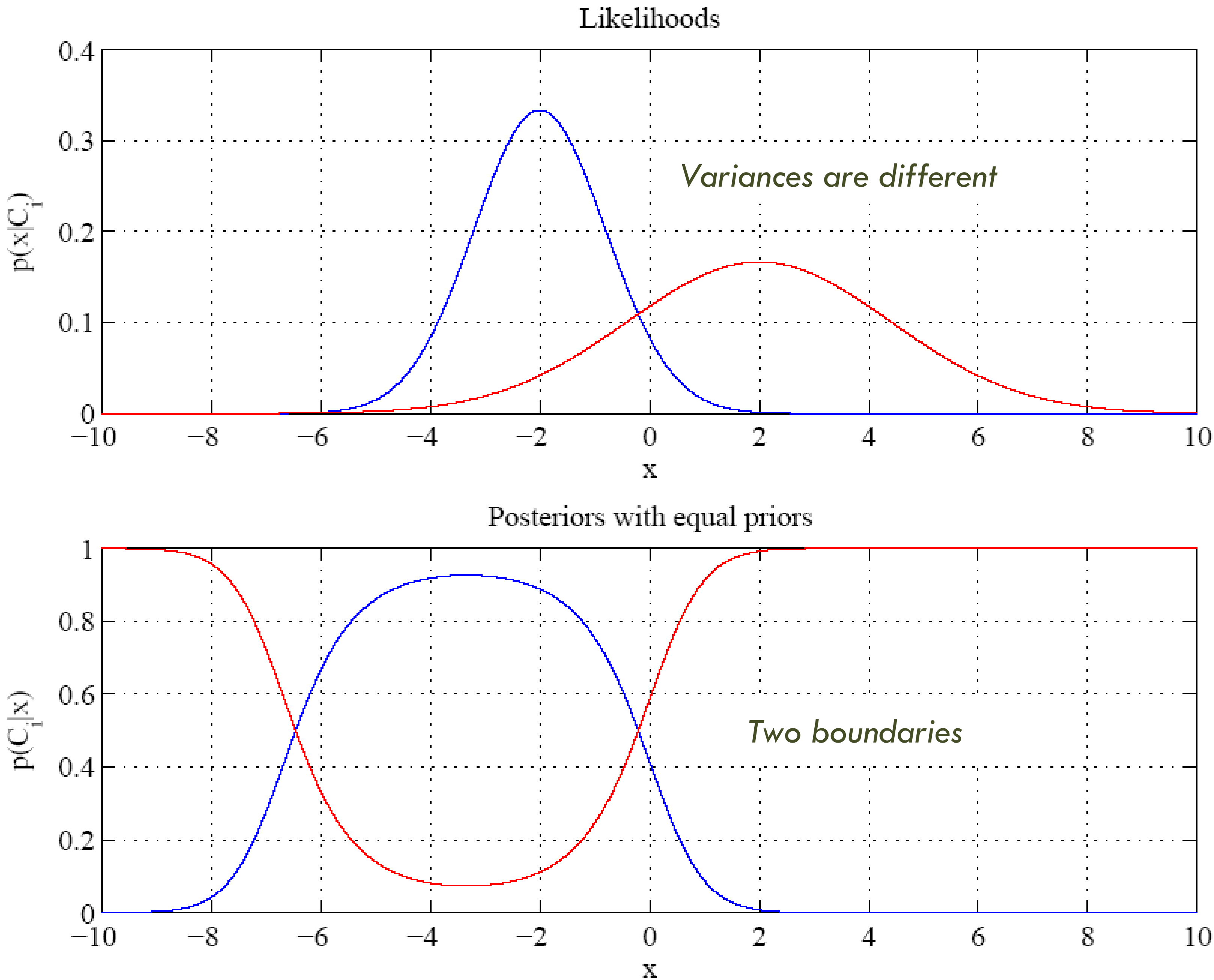
- Assume $p(x|C_i)$ are Gaussian
- ML estimates are

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

- Discriminant

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


Likelihoods
p(x|C_i)
x
Equal variances
Posteriors with equal priors
p(C_i|x)
$$g_1(x) = g_2(x)$$
$$(x - m_1)^2 = (x - m_2)^2$$
$$x = \frac{m_1 + m_2}{2}$$
Single boundary at
halfway between means


Likelihoods
Variances are different
p(x|C_i)
x
Posteriors with equal priors
Two boundaries
p(C_i|x)
x

# Regression

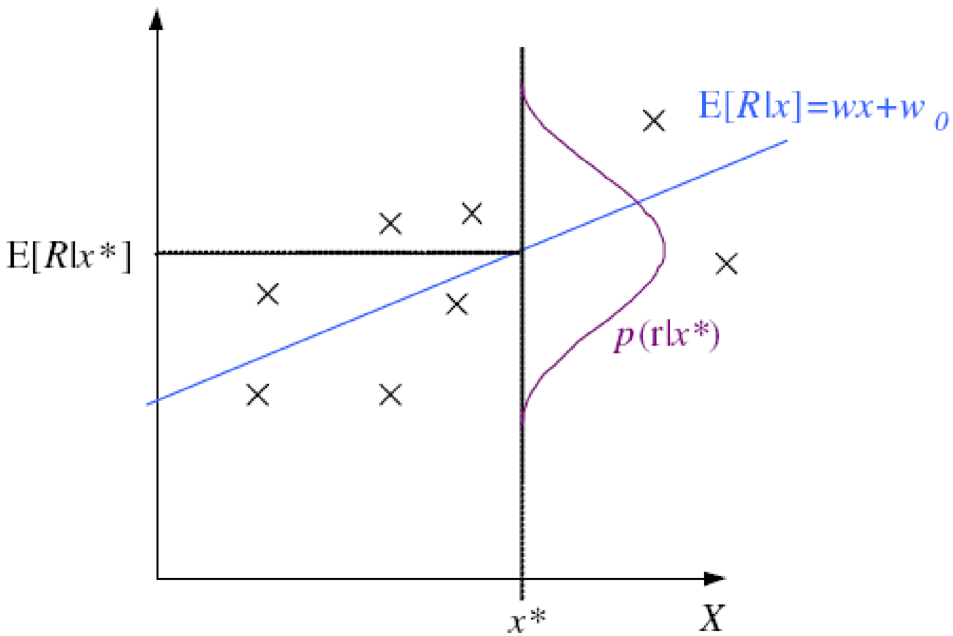


$$r = f(x) + \varepsilon$$

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

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

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




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

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

doesn't depend on $g$

# Regression: From LogL to Error



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

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

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

# 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: $E(\theta|\mathcal{X})=\frac{\sum_{t=1}^{N}\left[r^t-g\left(x^t|\theta\right)\right]^2}{\sum_{t=1}^{N}\left[r^t-\bar{r}\right]^2}$
- Absolute Error: $E(\vartheta|X)=\sum_t |r^t-g(x^t|\vartheta)|$
- ε-sensitive Error:

$$E(\vartheta|X)=\sum_t 1(|r^t-g(x^t|\vartheta)|>\varepsilon)\,(|r^t-g(x^t|\vartheta)|-\varepsilon)$$