Parametric Methods

Probability and Inference

- Result of tossing a coin is Head/1 or Tail/0
- Random variable $X \in \{1,0\}$
  Bernoulli: $P(X = 1) = p_0^X (1 - p_0)^{1-X}$
- Training set: $\mathcal{X} = \{X_t\}_{t=1}^N$
  Estimation: $p_0 = \frac{\#\ of\ heads}{\#\ of\ toesses} = \frac{\sum_{t=1}^N X_t}{N}$
- The rule for prediction of the next toss:
  Heads if $p_o > ½$,
  Tails otherwise

- Maximum likelihood estimate of p_0

Define log-likelihood function as

$$\mathcal{L}(p_0|X_1, X_2, \ldots, X_N) = logP(X_1, X_2, \ldots, X_N) = \sum_{t=1}^N logP(X_t)$$

$$= \sum_{t=1}^N X_t log p_0 + (1 - X_t) log(1 - p_0)$$

- The maximum likelihood estimate of p_0 can be obtained by solving

$$p_0 = \underset{p_0}{\operatorname{argmax}}\, \mathcal{L}(p_0|X_1, X_2, \ldots, X_N)$$

$$\frac{\partial \mathcal{L}(p_0|X_1, X_2, \ldots, X_N)}{\partial p_0} = 0$$

$$\Rightarrow \frac{\sum_{t=1}^N X_t}{p_0} - \frac{N - \sum_{t=1}^N X_t}{1 - p_0} = 0$$

$$\Rightarrow p_0 = \frac{\sum_{t=1}^N X_t}{N}$$

Parametric Estimation

- $\mathcal{X} = \{X_t\}_{t=1}^{N}$ where $X_t \sim P(X)$
- **Parametric estimation**:

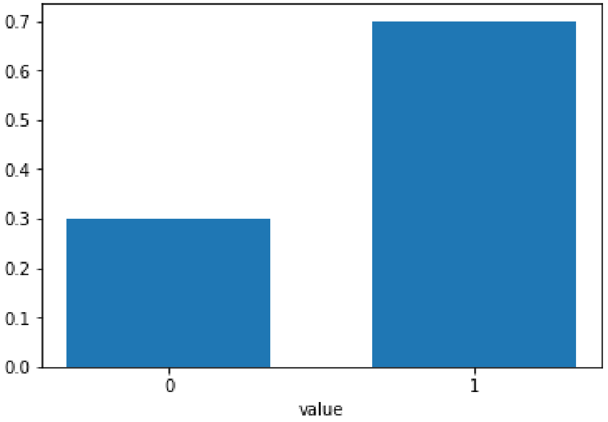
  Assume a form for $P(X|\theta)$ and estimate θ by its sufficient statistics $T(\mathcal{X})$

  e.g., Assume $X_t \sim \mathcal{N}(\mu, \sigma^2)$ and $\theta = \{\mu, \sigma^2\}$

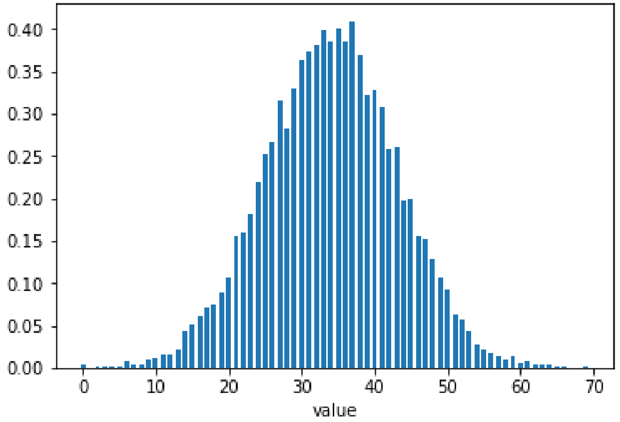
If a statistic $T(\mathcal{X})$ is a **sufficient statistic of underlying parameter** θ, we have

$$P(X = a|\theta, T(\mathcal{X})) = P(X = a|T(\mathcal{X})).$$

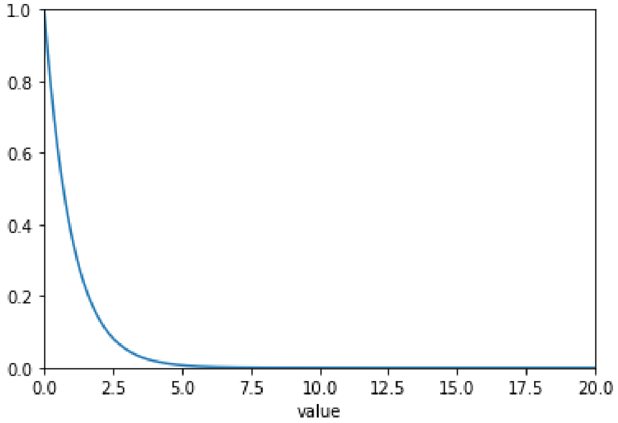
Well-Known Probability Distributions



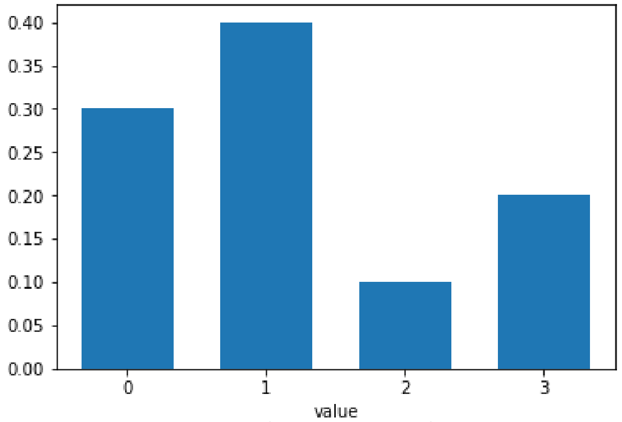
Bernoulli



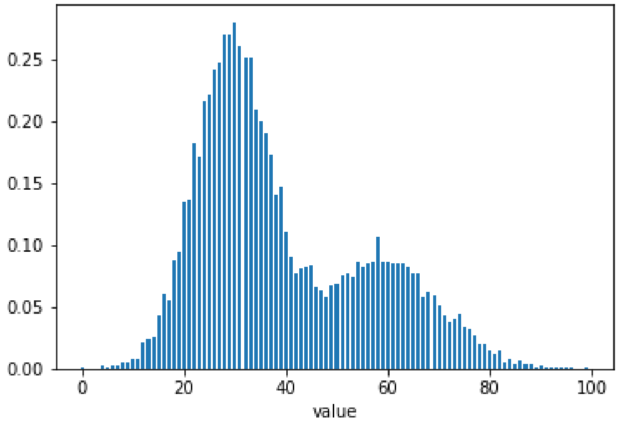
Gaussian



Exponential



Multinomial



Mixture of Gaussians

Maximum Likelihood Estimation

- Likelihood of θ given the sample $\mathcal{X} = \{X_t\}_{t=1}^{N}$

$$\ell(\theta \mid \mathcal{X}) = P(\mathcal{X} \mid \theta) = \prod_t P(X_t \mid \theta) \text{ because } X_t \text{ are i.i.d.}$$

- Log-likelihood function

$$\mathcal{L}(\theta \mid \mathcal{X}) = log \ell(\theta \mid \mathcal{X}) = \sum_t log P(X_t \mid \theta)$$

- Maximum likelihood estimator (MLE)

$$\theta^* = \underset{\theta}{\operatorname{argmax}} \mathcal{L}(\theta \mid \mathcal{X})$$

Bernoulli/Multinomial Density

- **Bernoulli**: Two states, failure/success, $X_t \in \{0,1\}$
  - $P(X) = p_0^X (1 - p_0)^{1-X}$
  - $\mathcal{L}(p_0|\mathcal{X}) = log \prod_t p_0^{X_t} (1 - p_0)^{1-X_t}$
  - MLE: $p_0 = \frac{\sum_t X_t}{N}$

- **Multinomial**: $X_t = [x_{1;t}, \dots, x_{K;t}]$, K>2, $x_{i;t} \in \{0,1\}$
  - $P(x_1, \dots, x_K) = \prod_i p_i^{x_i}$
  - $\mathcal{L}(p_1, \dots, p_K|\mathcal{X}) = log \ \prod_t \prod_i p_i^{x_{i;t}}$
  - MLE: $p_i = \frac{\sum_t x_{i;t}}{N}$

Gaussian (Normal) Distribution

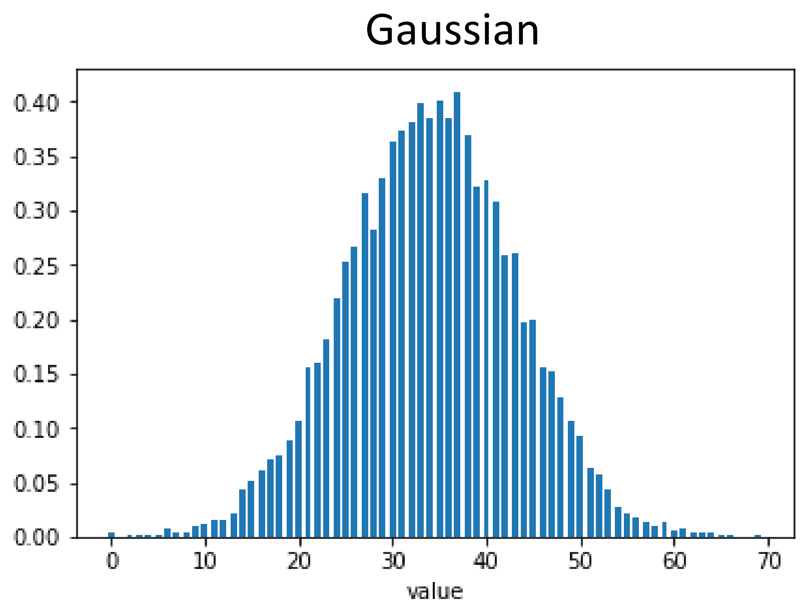
- $P(x) \sim \mathcal{N}(\mu, \sigma^2)$

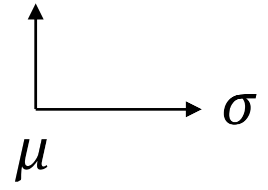
  $$P(x) = \frac{1}{\sqrt{2\pi}\sigma} exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$$

- MLE for $\boldsymbol{\mu}$ and $\boldsymbol{\sigma}^2$:

  $m = \frac{\sum_t x_t}{N}$ (sample mean)

  $s^2 = \frac{\sum_t (x_t - m)^2}{N}$ (sample covariance)





Bayes' Estimator

- Use prior information about the possible value range for the parameter
  - Useful for a small number of training examples
  - Treat θ as a random variable with prior $P(\theta)$
  - Bayes' rule: $P(\theta|\mathcal{X}) = P(\mathcal{X}|\theta)P(\theta)/P(\mathcal{X})$=$P(\mathcal{X}|\theta)P(\theta)/\int P(\mathcal{X}|\theta')P(\theta')d\theta'$

- **Full**: $P(x|\mathcal{X}) = \int P(x|\theta,\mathcal{X})P(\theta|\mathcal{X})d\theta = \int P(x|\theta)P(\theta|\mathcal{X})d\theta$

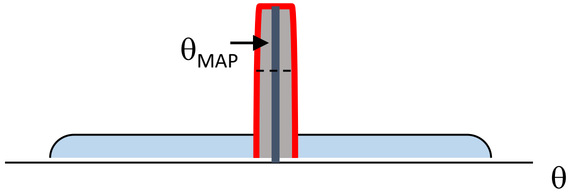
sufficient statistics

- **Bayes'**: $\theta_{Bayes} = \int \theta P(\theta|\mathcal{X})d\theta$

➢ Difficult to evaluate when $P(\theta|\mathcal{X})$ does not have a simple form

- **Maximum a Posteriori (MAP)**: $\theta_{MAP} = \underset{\theta}{\operatorname{argmax}} P(\theta|\mathcal{X})$

➢ Assume that $P(\theta|\mathcal{X})$ has a narrow peak around its mode



- **Maximum Likelihood (ML)**: $\theta_{ML} = \underset{\theta}{\operatorname{argmax}} P(\mathcal{X}|\theta)$

➢ Have no prior information about θ (i.e., $P(\theta)$ is flat)

An Example of Bayes' Estimator

- $x_t \sim \mathcal{N}(\theta, \sigma^2)$ and $\theta \sim \mathcal{N}(\mu_0, \sigma_0^2)$, where $\sigma^2, \mu_0, \sigma_0^2$ are known
- $\theta_{ML} = m$ (sample mean)
- $\boxed{\theta_{MAP} = \theta_{Bayes} = E[\theta|\mathcal{X}]} = \frac{\frac{N}{\sigma^2}}{\frac{1}{\sigma_0^2}+\frac{N}{\sigma^2}} \times \underset{\text{sample mean}}{m} + \frac{\frac{1}{\sigma_0^2}}{\frac{1}{\sigma_0^2}+\frac{N}{\sigma^2}} \times \underset{\text{prior mean}}{\mu_0} \xrightarrow[N\to\infty]{} m$

Because $P(\theta|\mathcal{X})$ is normal

Classification by Likelihood-Based Approaches (Generative Models)

- Estimate $P(x|C_i)$and $P(C_i)$ from training samples.

- Assign x to Class i
  if $P(x|C_i)P(C_i) > P\left(x|C_j\right)P\left(C_j\right), j \neq i$

- Discriminant function:

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

or

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

Classification by Gaussian Generative Models

- If $P(x|C_i)$ are Gaussian distributions:

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

discriminant functions are

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

- Given the sample: $\mathcal{X} = \{x_t, \boldsymbol{r}_t\}_{t=1}^{N}, \boldsymbol{r}_t = [r_{1;t}, \dots, r_{K;t}]$

$$x_t \in \mathcal{R}, r_{i;t} = \begin{cases} 1 & if\, x_t \in C_i \\ 0 & if\, x_t \in C_j, j \neq i \end{cases}$$

Learning

- ML estimates are

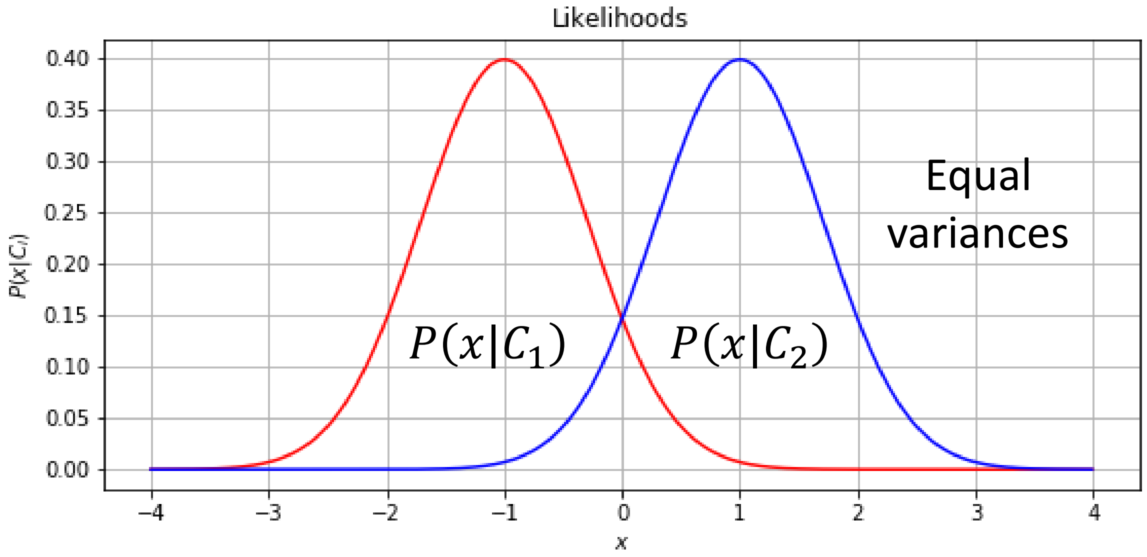
$$\hat{P}(C_i) = \frac{\sum_t r_{i;t}}{N}, m_i = \frac{\sum_t r_{i;t}\, x_t}{\sum_t r_{i;t}}, s_i^2 = \frac{\sum_t r_{i;t}(x_t - m_i)^2}{\sum_t r_{i;t}}$$

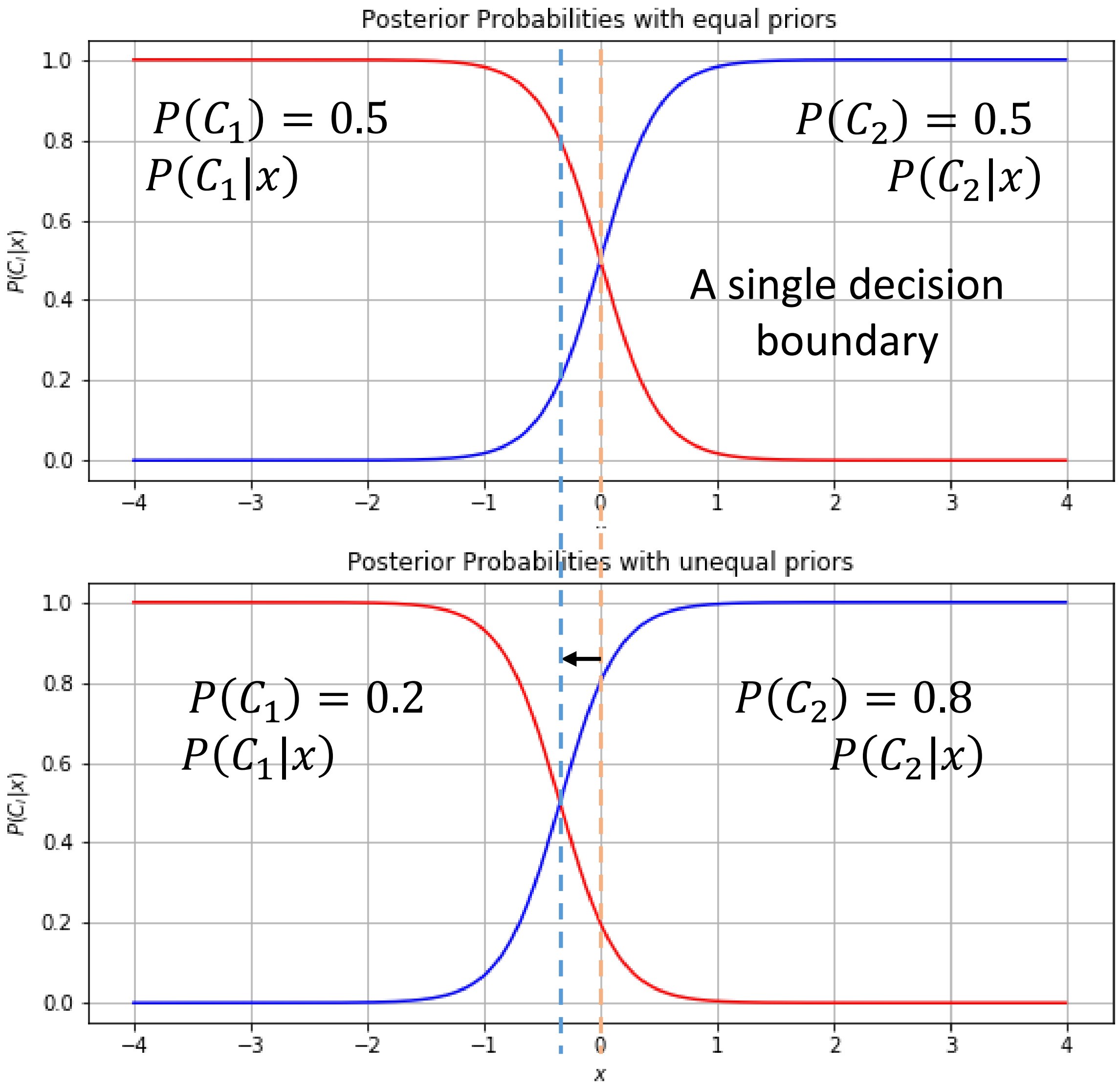
- Discriminant functions are

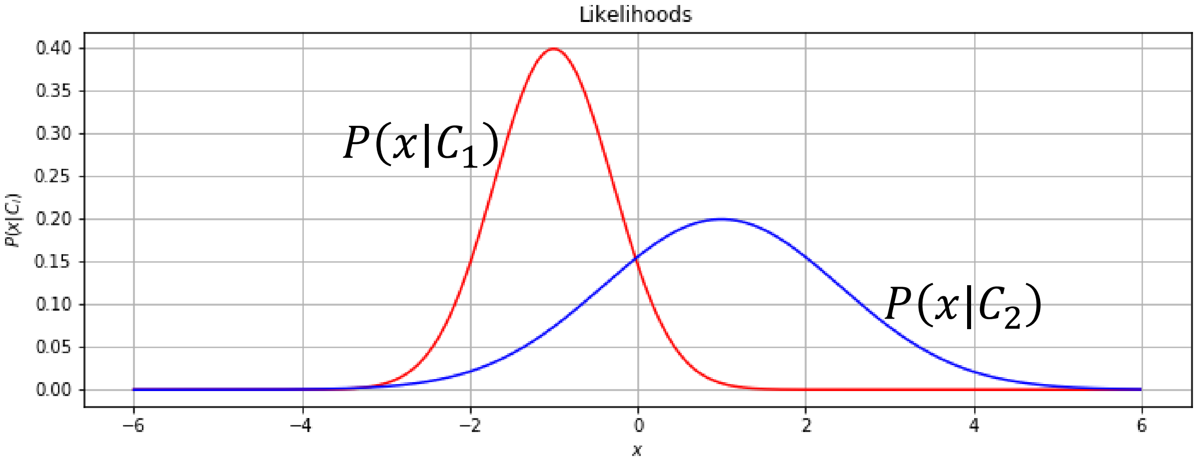
$$g_i(x) = -\frac{1}{2}\log(2\pi) - \log(s_i) - \frac{(x - m_i)^2}{2s_i^2} + log\hat{P}(C_i)$$

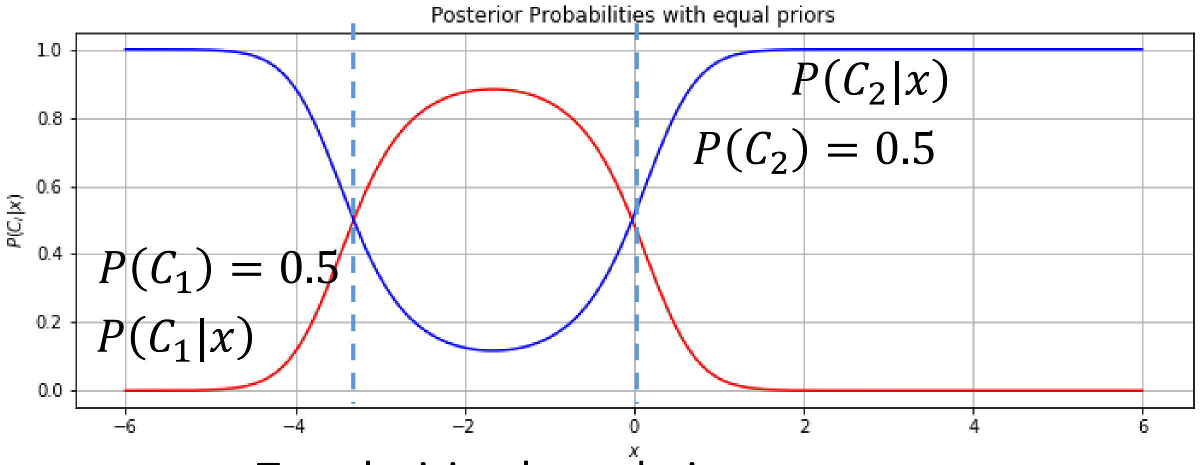
or

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

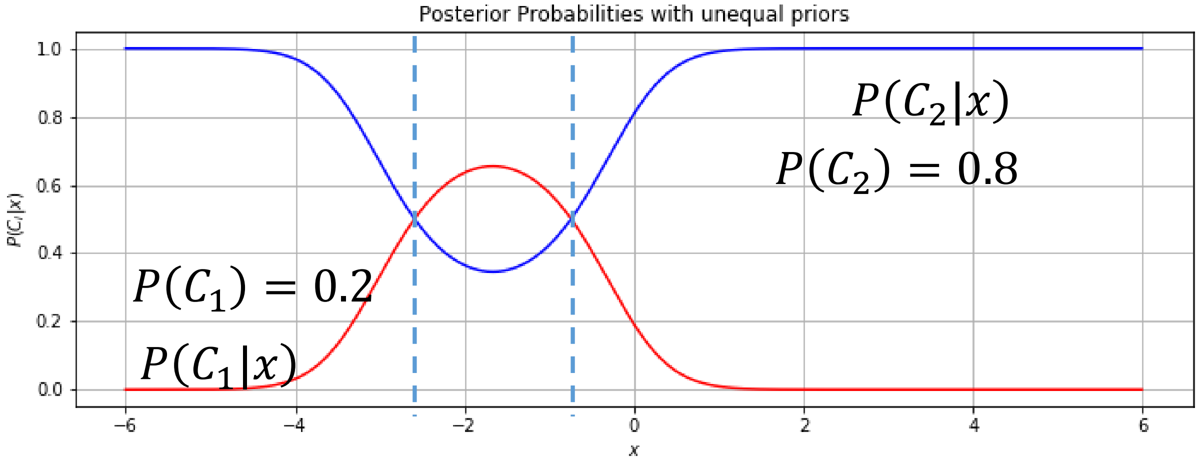

Likelihoods
Equal variances
$P(x|C_1)$
$P(x|C_2)$
$P(x|C_i)$
x


Posterior Probabilities with equal priors
$P(C_1) = 0.5$
$P(C_1|x)$
$P(C_2) = 0.5$
$P(C_2|x)$
A single decision boundary
Posterior Probabilities with unequal priors
$P(C_1) = 0.2$
$P(C_1|x)$
$P(C_2) = 0.8$
$P(C_2|x)$
$P(C_i|x)$
x



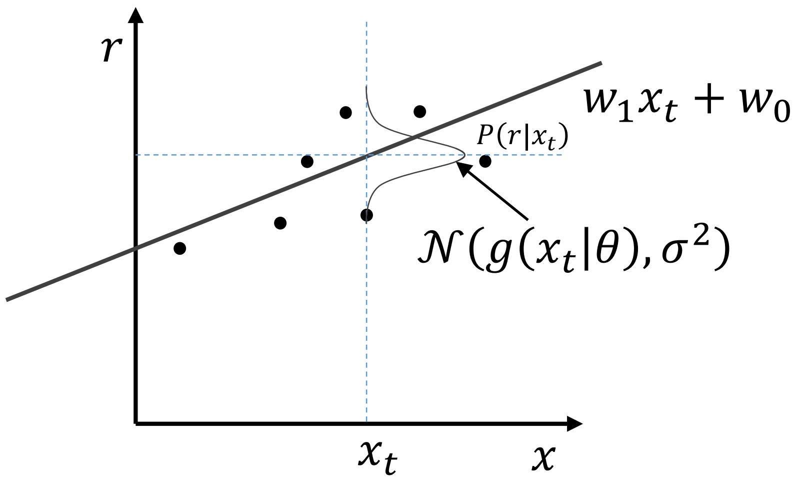


Two decision boundaries



Regression

$$r = f(x) + \epsilon,\ \epsilon \sim \mathcal{N}(0, \sigma^2)$$



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

- Define the log-likelihood function as $\mathcal{L}(\theta|\mathcal{X}) = \sum_t \log\big(P(r_t|x_t)\big)$

  - $\mathcal{L}(\theta|\mathcal{X}) = log \prod_t P(x_t, r_t) = log \prod_t P(r_t|x_t) P(x_t) = log \prod_t P(r_t|x_t) + \underbrace{\boxed{log \prod_t P(x_t)}}_{\text{ignore}}$

Regression: From Log-Likelihood to Error

Estimating θ by maximization of

$$\mathcal{L}(\theta|\mathcal{X}) = log \prod_t \frac{1}{\sqrt{2\pi}\sigma} exp\left(-\frac{\left(r_t - g(x_t|\theta)\right)^2}{2\sigma^2}\right)$$

$$= -N log\left(\sqrt{2\pi}\sigma\right) \boxed{- \frac{1}{2\sigma^2} \sum_t \left(r_t - g(x_t|\theta)\right)^2}$$

is equivalent to estimating θ by minimization of

$$E[\theta|\mathcal{X}] = \frac{1}{2} \sum_t \left(r_t - g(x_t|\theta)\right)^2$$

$\theta^* = \underset{\theta}{\operatorname{argmin}}\, E[\theta|\mathcal{X}]$ are called least squares estimates

Linear Regression

- $g(x_t|w_1, w_0) = w_1 x_t + w_0$

$$E[\theta|\mathcal{X}] = \frac{1}{2}\sum_t \big(r_t - g(x_t|\theta)\big)^2$$

$$\boxed{\begin{aligned}
&\frac{\partial E[\theta|\mathcal{X}]}{\partial w_1} = 0 \Rightarrow \sum_t x_t(r_t - w_1 x_t - w_0) = 0 \\
&\frac{\partial E[\theta|\mathcal{X}]}{\partial w_0} = 0 \Rightarrow \sum_t (r_t - w_1 x_t - w_0) = 0 \\
&\Rightarrow \begin{bmatrix} \Sigma_t x_t^2 & \Sigma_t x_t \\ \Sigma_t x_t & N \end{bmatrix} \begin{bmatrix} w_1 \\ w_0 \end{bmatrix} = \begin{bmatrix} \Sigma_t x_t r_t \\ \Sigma_t r_t \end{bmatrix} \\
&\Rightarrow \begin{bmatrix} w_1 \\ w_0 \end{bmatrix} = \begin{bmatrix} \Sigma_t x_t^2 & \Sigma_t x_t \\ \Sigma_t x_t & N \end{bmatrix}^{-1} \begin{bmatrix} \Sigma_t x_t r_t \\ \Sigma_t r_t \end{bmatrix}
\end{aligned}}$$

Polynomial Regression

- $g(x_t|w_k, \dots, w_1, w_0) = \sum_{i=0}^{k} w_i x_t^i$

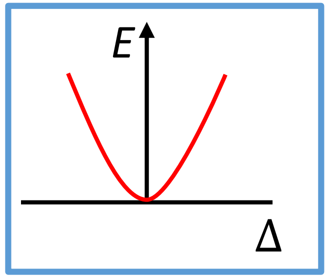
$$\begin{bmatrix} x_1^k & \cdots & x_1 & 1 \\ x_2^k & \cdots & x_2 & 1 \\ \vdots & & \vdots & \vdots \\ x_N^k & \cdots & x_N & 1 \end{bmatrix} \begin{bmatrix} w_k \\ \vdots \\ w_1 \\ w_0 \end{bmatrix} = \begin{bmatrix} r_1 \\ r_2 \\ \vdots \\ r_N \end{bmatrix}$$

$$\Rightarrow \mathbf{Dw} = \mathbf{r}$$

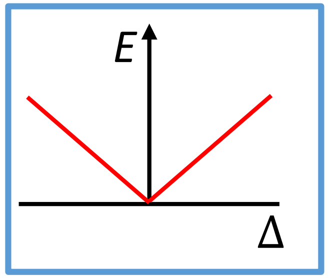
$$\Rightarrow \mathbf{w} = (\mathbf{D}^T\mathbf{D})^{-1}\mathbf{D}^T\mathbf{r}$$

Error Measures

- **Square Error**: $E[\theta|\mathcal{X}] = \frac{1}{2}\sum_t \big(\underbrace{r_t - g(x_t|\theta)}_{\Delta}\big)^2$

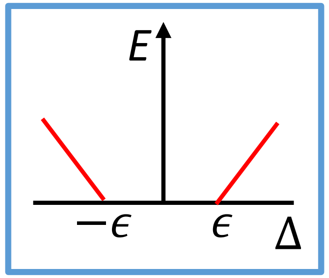


- **Relative Square Error**: $E[\theta|\mathcal{X}] = \frac{\sum_t \big(r_t - g(x_t|\theta)\big)^2}{\sum_t (r_t - \bar{r})^2}$, where $\bar{r} = \frac{1}{N}\sum_t r_t$



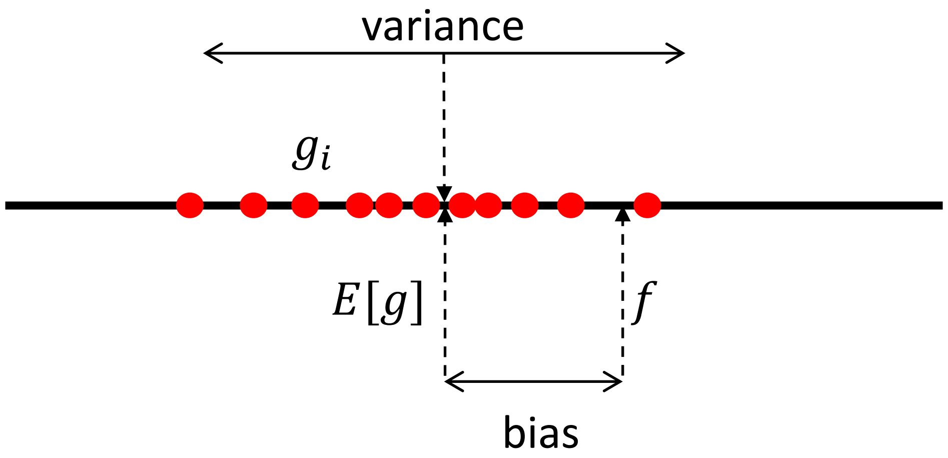
- **Absolute Error**: $E[\theta|\mathcal{X}] = \sum_t |\underbrace{r_t - g(x_t|\theta)}_{\Delta}|$

- **ε-sensitive Error**: $E[\theta|\mathcal{X}] = \sum_t 1(|r_t - g(x_t|\theta)| > \epsilon)(|\underbrace{r_t - g(x_t|\theta)}_{\Delta}| - \epsilon)$



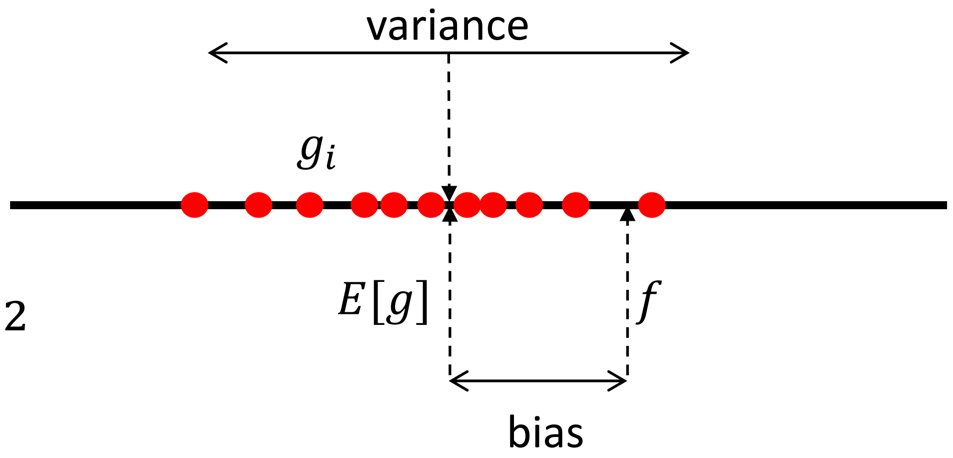
Bias and Variance

- **Bias**: The difference between the expectation of the approximating function and the target function.
- **Variance:** The average squared error between the output on a given particular training set and the average of all training patterns used.



Estimating Bias and Variance

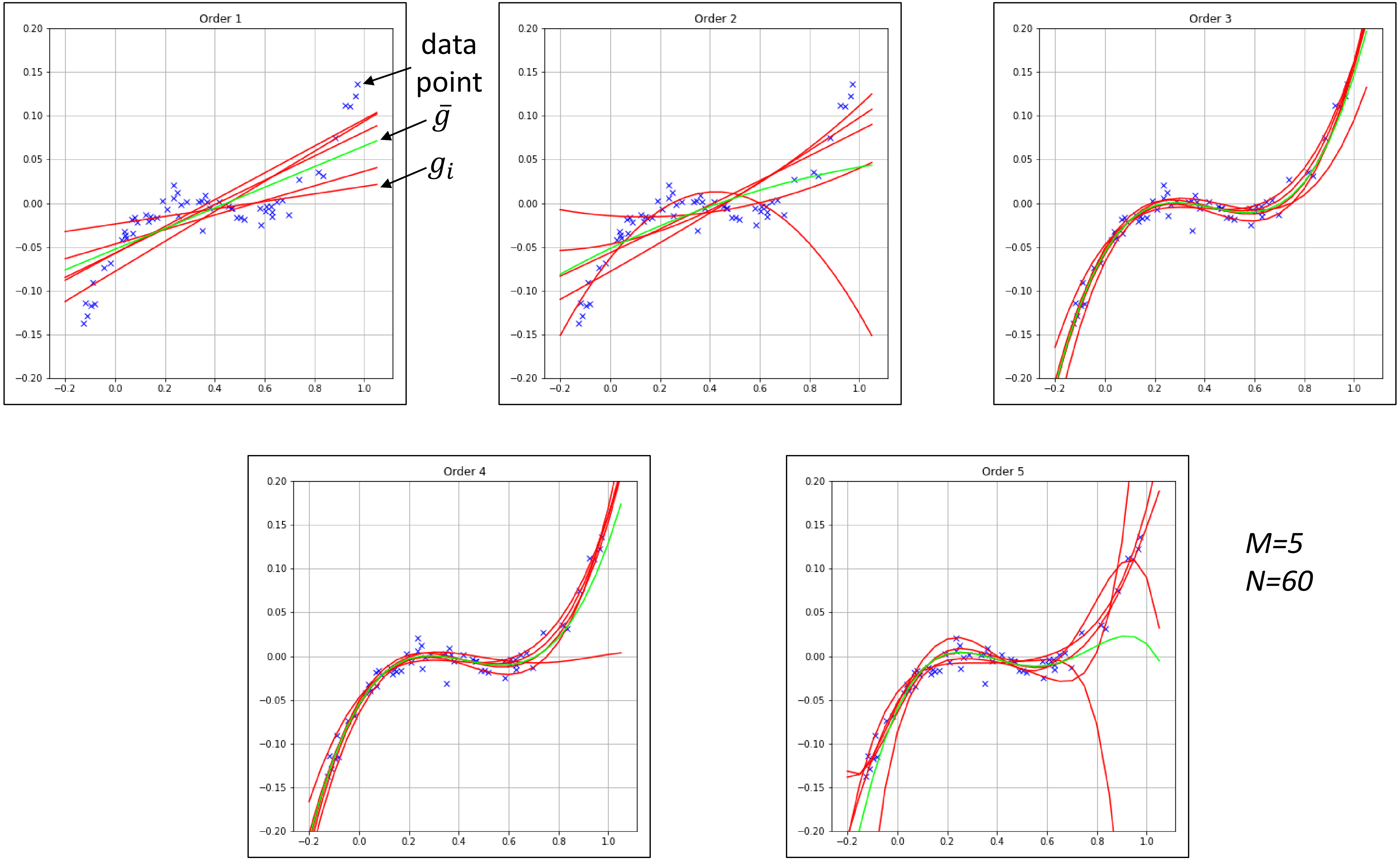
- A training set $\mathcal{X} = \{x_t, r_t\}_{t=1}^{N}$ is partitioned in to M sample sets $\mathcal{X}_i, i = 1, \dots, M,$ to fit $g_i(x), i = 1, \dots, M,$ respectively
  - $f(x)$: the target function
  - $\bar{g}(x) = \frac{1}{M}\sum_i g(x)$
  - $Bias^2(g) = \frac{1}{N}\sum_t(\bar{g}(x_t) - f(x_t))^2$
  - $Variance(g) = \frac{1}{NM}\sum_t\sum_i\left(g_i(x_t) - \bar{g}(x_t)\right)^2$



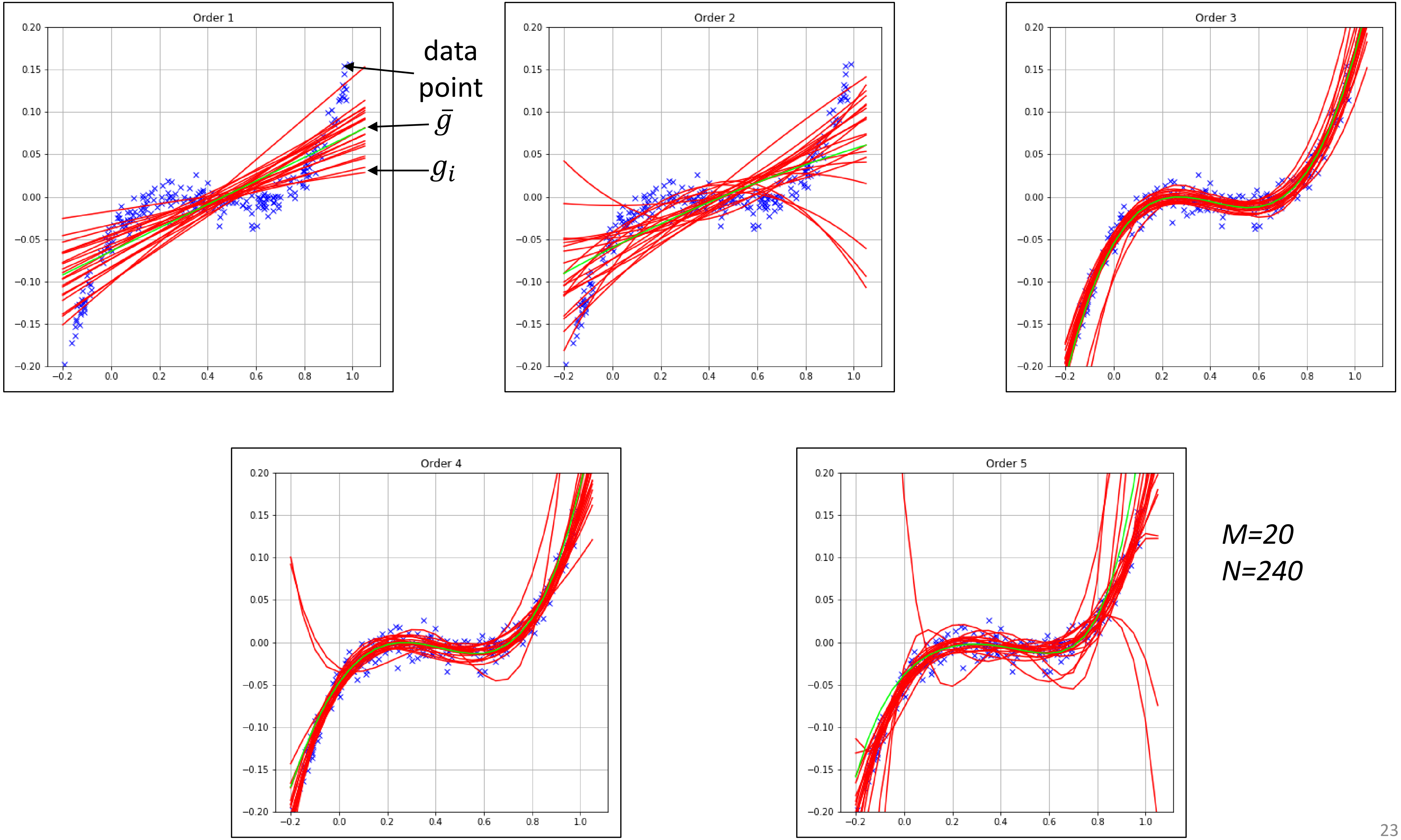
Example:

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

$g_i(x) =$ the average of the r in the ith sample set has lower bias with variance



M=5
N=60



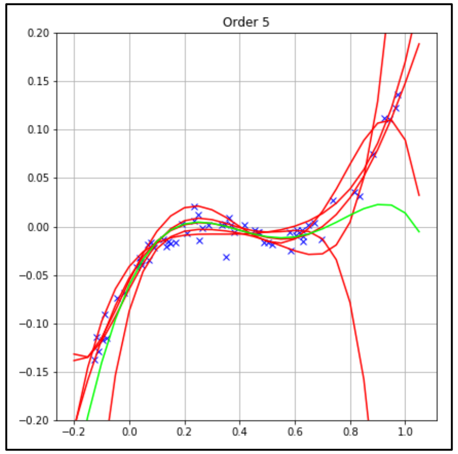
M=20
N=240

Bias/Variance Dilemma

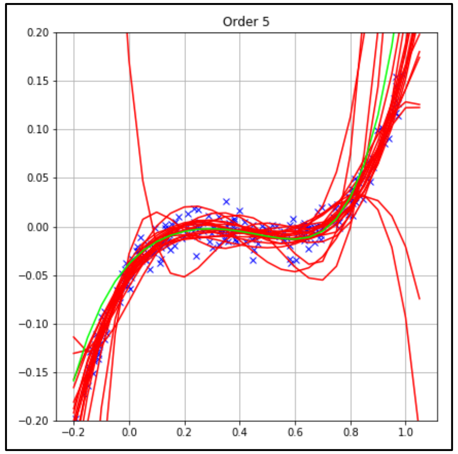
- As the model complexity increases, bias decreases and variance increases

➢The variance and bias are dependent.

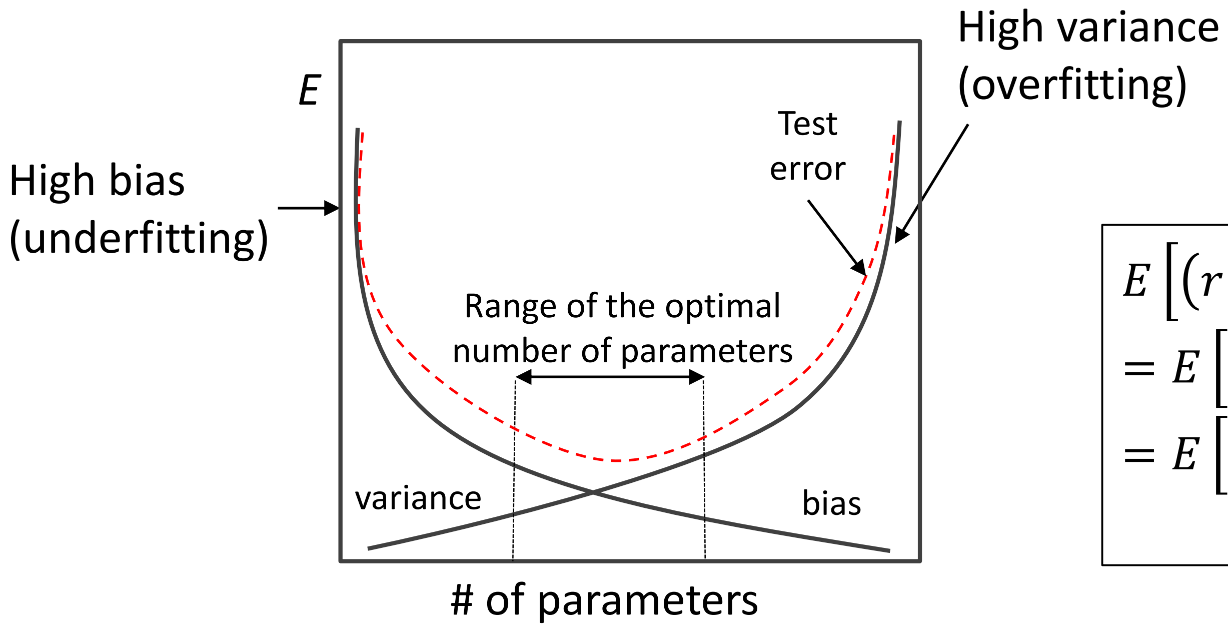
➢Increase the number of training examples.



M=5



M=20



$$
\begin{aligned}
&E\left[\left(r-g(x)\right)^2\right] \\
&= E\left[\left(r-f(x)+f(x)-\bar{g}(x)+\bar{g}(x)-g(x)\right)^2\right] \\
&= \underset{noise:\,\sigma^2}{E\left[\left(r-f(x)\right)^2\right]} + \underset{bias^{\,2}}{E\left[\left(f(x)-\bar{g}(x)\right)^2\right]} + \underset{variance}{E\left[\left(\bar{g}(x)-g(x)\right)^2\right]}
\end{aligned}
$$

Model Selection Procedures

- **Cross-validation**: Measure generalization accuracy by testing on the example in the validation set (use this method if there is a large enough validation dataset)
- **Regularization**: Penalize complex models

$$E = error\ on\ data + \lambda \times model\ complexity \downarrow$$

- **Criteria for model selection**
  - **Akaike's information criterion (AIC)**,

    $$\mathrm{AIC} = k - \log(\mathcal{L}), \downarrow$$

  - **Bayesian information criterion (BIC)**

    $$\mathrm{BIC} = klogN - 2\log(\mathcal{L}), \downarrow$$

    where $\mathcal{L}$ is the largest likelihood of the model, k is the number of parameters in the model and *N* is the number of training examples
  - **Structural risk minimization (SRM)**
  - **Minimum description length (MDL)**: Kolmogorov complexity, shortest description of data

Prefer simpler models

Bayesian Model Selection

- Bayesian model selection is used when there is prior knowledge on models, $P(Model)$

$$P(Model|Data) = \frac{P(Data|Model)P(Model)}{P(Data)}$$

constant

$$\Rightarrow \log\big(P(Model|Data)\big) = \boxed{\log\big(P(Data|Model)\big) + \log\big(P(Model)\big)} - \log\,(P(Data))$$

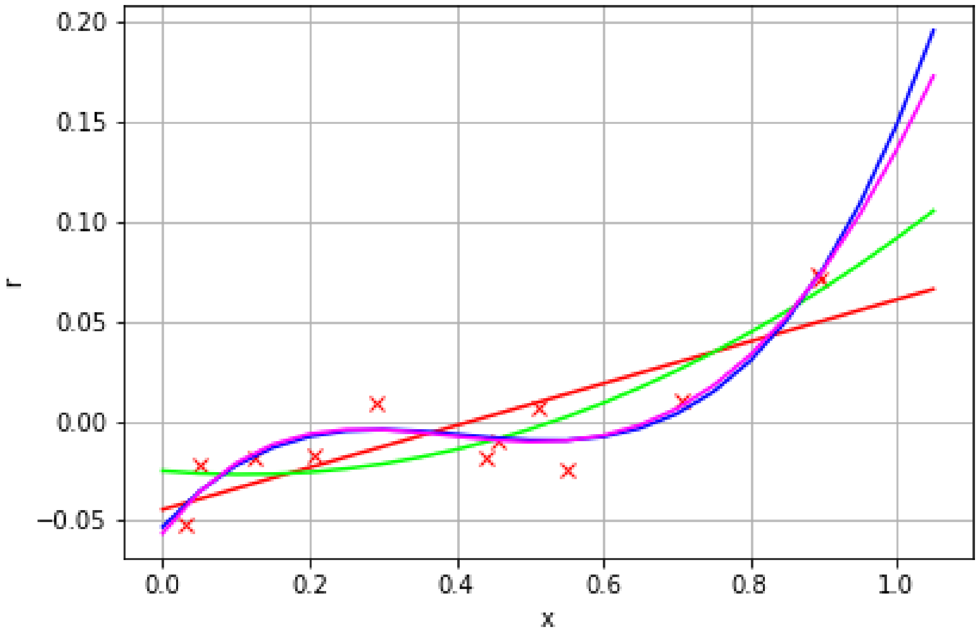
$$error\ on\ data + \lambda \times model\ complexity$$

if simpler models are favored

- Regularization≈the Bayesian approach, when simpler models are favored

- Average over a number of models with high posterior
  - Bayesian optimal classifier (most probable classification),
  - Voting, Ensembles (Chapter 17)

Regularization Example



Magnitudes may increase as polynomial order increases

Order	$\boldsymbol{w}$	$\|\boldsymbol{w}\|_2^2$
1	[0.105,-0.044]	0.114
2	[0.140,-0.033,-0.025]	0.156
3	[0.867,-1.072,0.406,-0.053]	1.438
4	[-0.387,1.565,-1.470,0.484,-0.056]	2.235

A smoother and flatter fit is desired

$L2\ regularization\text{:}\ \|\boldsymbol{w}\|_2^2$

$$E[\boldsymbol{w}|\mathcal{X}] = \frac{1}{2}\sum_t (r_t - g(x_t \boldsymbol{w}|))^2 + \lambda \boxed{\sum_i w_i^2}$$

Prior: $P(\boldsymbol{w}) \sim \mathcal{N}(0, 1/\lambda)$

$L0\ regularization\text{:}\ \|\boldsymbol{w}\|_0^2$

$L1\ regularization\text{:}\ \|\boldsymbol{w}\|_1^2$

Bayes Optimal Classifier

- Bayes optimal classifier
  - $\underset{y}{\operatorname{argmax}} \sum_{h_i \in H} P(y|x, h_i) P(h_i|Data)$
- Example,

□ $P(h_1|Data) = 0.4, P(h_2|Data) = 0.25, P(h_3|Data) = 0.35$

- h_1 is the MAP hypothesis.

□ For an input x, suppose

$$P(y = +1|x, h_1) = 1, P(y = -1|x, h_2) = 1, P(y = -1|x, h_3) = 1,$$

where $y \in \{-1, +1\}$

- The MAP classification of x is +1
- The most probable classification of x is -1
  - $\sum_{h_i \in H} P(+1|x, h_i) P(h_i|Data) = 0.4$
  - ✓ $\sum_{h_i \in H} P(-1|x, h_i) P(h_i|Data) = 0.6$