# Parametric Methods

Introduction to Machine Learning – GIF-7015

Professor: Christian Gagné

Week 2

UNIVERSITÉ LAVAL

## 2.3 Parametric estimation

# Parametric estimation

- Dataset $\mathcal{X} = \{x^t\}_{t=1}^N$ where $x^t \sim p(x)$
  - Independent and identically distributed variable (iid)
- Parametric estimation
  - Family of probability densities $p(x|\theta)$
  - Estimate $\theta$ : sufficient density statistics
  - With a normal distribution $\mathcal{N}(\mu,\sigma^2)$, $\theta = \{\mu,\sigma\}$
- Estimate of $\theta$ from $\mathcal{X}$

## Likelihood of an estimate

- Likelihood of an estimate parameterized by $\theta$

$$l(\theta|\mathcal{X}) \equiv p(\mathcal{X}|\theta) = \prod_{t=1}^{N} p(x^t|\theta)$$

  - $p(x|\theta)$ is equivalent to the likelihood that a sample $x^t$ is obtained given $\theta$
  - Since the $x^t$ are iid, we do a product of likelihoods

# Maximum likelihood

- Log-likelihood function

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

  - $\log(ab) = \log(a) + \log(b)$
  - $\log(a^n) = n\log(a)$
- Log allows to simplify the equations for some densities (e.g. normal distribution)
- Maximum likelihood estimate: find the value of $\theta$ making the sampling $\mathcal{X}$ the most probable

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

# Example: Bernoulli's law

- Bernoulli's law: $P(x) = p^x(1-p)^{1-x}$, $x \in \{0, 1\}$
- Log-likelihood function:

$$
\begin{aligned}
L(p|\mathcal{X}) &= \log \prod_{t=1}^{N} p^{(x^t)}(1-p)^{(1-x^t)} \\
&= \sum_{t=1}^{N} x^t \log p + \left(N - \sum_{t=1}^{N} x^t\right) \log(1-p)
\end{aligned}
$$

- Maximum likelihood estimate:

$$
\frac{dL(p|\mathcal{X})}{dp} = 0 \quad \Rightarrow \quad \hat{p} = \frac{\sum_{t=1}^{N} x^t}{N}
$$

## Example: categorical law

- Categorical law: Bernoulli generalization to $K$ mutually exclusive states
  - State $\mathbf{x} = (x_1, x_2, \dots, x_K)$, variables $x_i \in \{0, 1\}$ and $\sum_i x_i = 1$
  - Each variable $x_i$ has a probability $p_i$, with $\sum_i p_i = 1$
  - State probability: $p(\mathbf{x}) = \prod_{i=1}^{K} p_i^{x_i}$
  - Independent experiments: $\mathcal{X} = \{\mathbf{x}^t\}_{t=1}^{N}$
- Maximum likelihood estimate:

$$\frac{\partial L(p|\mathcal{X})}{\partial p_i} = 0 \quad \Rightarrow \quad \hat{p}_i = \frac{\sum_t x_i^t}{N},\ i = 1, \dots, K$$

# Example: normal distribution

- Normal distribution: distribution parameterized by a mean $\mu$ and a standard deviation $\sigma$

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

- Likelihood according to a sampling $\mathcal{X} = \{x^t\}_{t=1}^N$ with $x^t \sim \mathcal{N}(\mu, \sigma^2)$

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

  - Maximum likelihood with $\frac{\partial L(\mu,\sigma|\mathcal{X})}{\partial \mu} = 0$ and $\frac{\partial L(\mu,\sigma|\mathcal{X})}{\partial \sigma} = 0$

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

# Bias of an estimator

- $d(\mathcal{X})$, estimate of $\theta$ with $\mathcal{X}$
- Quality of the estimate of $d(\mathcal{X})$: $(d(\mathcal{X}) - \theta)^2$
- Quality of estimator $d$:

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

  - Evaluation of $d$ on all possible samples $\mathcal{X}$
- Bias of the estimator

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

  - Unbiased estimator: $b_\theta(d) = 0$ for all the possible values of $\theta$

# Reminder: expectation

- Expectation of a continuous random variable $X$ having a density $f_X(x)$:

$$\mathbb{E}(X) = \int_{\mathbb{R}} x\, f_X(x)\, dx$$

  - The transfer theorem applies for measurable functions $g(X)$ of the random variable $X$:

$$\mathbb{E}(g(X)) = \int_{\mathbb{R}} g(x)\, f_X(x)\, dx$$

- So for a constant $a$, the expectation of $g(X) = a\,X$ is:

$$\mathbb{E}(a\,X) = \int_{\mathbb{R}} a\,x\, f_X(x)\, dx = a \int_{\mathbb{R}} x\, f_X(x)\, dx = a\,\mathbb{E}(X)$$

- And for the sum of two functions of $X$, $g(X) = m(X) + n(X)$:

$$\mathbb{E}(m(X) + n(X)) = \int_{\mathbb{R}} (m(x) + n(X))\, f_X(x)\, dx = \mathbb{E}(m(X)) + \mathbb{E}(n(X))$$

# Bias of the estimator $m$

- Suppose samples with a density of mean $\mu$
  - $m$ is an unbiased estimator of $\mu$

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

- Variance of the estimator

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

  - Reminder: $\mathrm{Var}(x) = \mathbb{E}[(x - \mathbb{E}[x])^2] = \mathbb{E}[x^2] - \mathbb{E}[x]^2$
  - Efficient estimator: $\lim_{N\to\infty} \mathrm{Var}_{\mathcal{X}}(m) = 0$
- Convergent estimator: $\lim_{N\to\infty} m = \mu$
  - Strong law of large numbers

## Bias of the estimator $s^2$

- Standard deviation $\sigma$ of a normal distribution $\mathcal{N}(\mu,\sigma^2)$
  - $s^2$ is a maximum likelihood estimator of $\sigma^2$

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

  - Quality of the estimator $s^2$

$$\begin{aligned}
\mathbb{E}_{\mathcal{X}}[(x^t)^2] &= \sigma^2 + \mu^2 \\
\mathbb{E}_{\mathcal{X}}[m^2] &= \sigma^2/N + \mu^2 \\
\mathbb{E}_{\mathcal{X}}[s^2] &= \frac{\sum_t \mathbb{E}_{\mathcal{X}}[(x^t)^2] - N\,\mathbb{E}_{\mathcal{X}}[m^2]}{N} \\
&= \frac{N(\sigma^2+\mu^2) - N(\sigma^2/N+\mu^2)}{N} = \frac{N-1}{N}\sigma^2 \neq \sigma^2
\end{aligned}$$

  - Estimator $s^2$ is biased!

## 2.4 Bayesian classification

# Bayesian classification

- Bayes rule for classification

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

- Corresponding discriminant function ($p(x)$ the same $\forall C_i$)

$$\begin{aligned} \mathrm{h}_i(x) &= p(x|C_i)P(C_i) \\ &\equiv \log p(x|C_i) + \log P(C_i) \end{aligned}$$

- With $p(x|C_i)$ following a normal distribution $\mathcal{N}(\mu_i,\sigma_i^2)$

$$\begin{aligned} p(x|C_i) &= \frac{1}{\sqrt{2\pi}\sigma_i} \exp\left[-\frac{(x-\mu_i)^2}{2\sigma_i^2}\right] \\ \mathrm{h}_i(x) &= -\frac{1}{2}\log 2\pi - \log \sigma_i - \frac{(x-\mu_i)^2}{2\sigma_i^2} + \log P(C_i) \end{aligned}$$

# Example of Bayesian classification

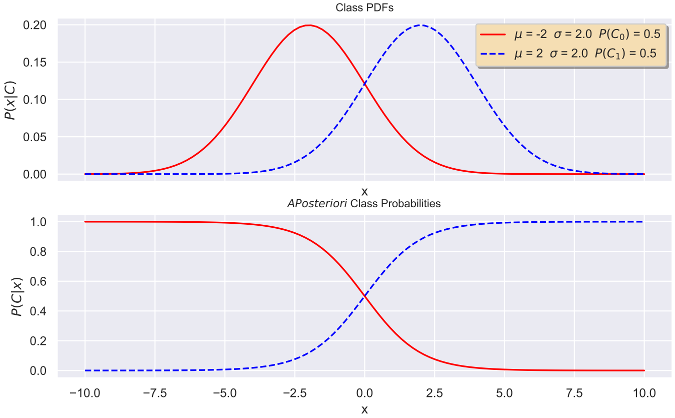
- Suppose dataset $\mathcal{X} = \{x^t, \mathbf{r}^t\}_{t=1}^N$ where $r_i^t = 1$ if $x^t \in C_i$ and $r_i^t = 0$ otherwise
  - Estimation of a priori probabilities: $\hat{P}(C_i) = \frac{\sum_t r_i^t}{N}$
  - Estimation of means: $m_i = \frac{\sum_t x^t r_i^t}{\sum_t r_i^t}$
  - Estimation of standard deviations: $s_i^2 = \frac{\sum_t (x^t - m_i)^2 r_i^t}{\sum_t r_i^t}$
- Corresponding discriminant function

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

- Simplifications
  1. $-\frac{1}{2}\log 2\pi$ is a constant
  2. Assume an equal variance, $\sigma_i = \sigma_j, \forall i, j$
  3. Suppose the same a priori probability, $\hat{P}(C_i) = \hat{P}(C_j), \forall i, j$
- We then do a classification based on the closest mean

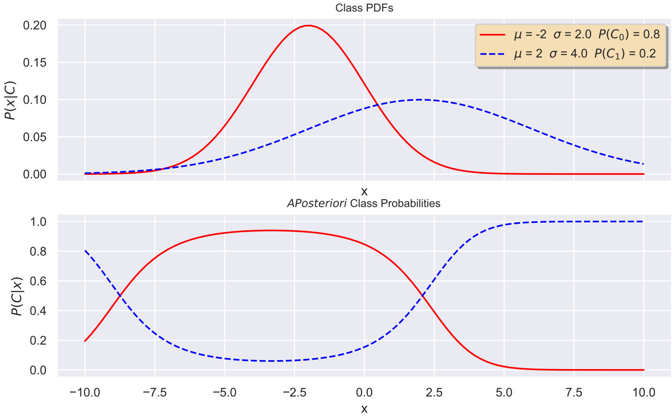
$$\mathrm{h}_i(x) = -(x - m_i)^2 \quad \Rightarrow \quad C_i = \underset{C_k}{\operatorname{argmin}} |x - m_k|$$

# Likelihoods with two classes, same variance




Boundaries: $\mathrm{h}_1(x) = \mathrm{h}_2(x) \quad \Rightarrow \quad (x - m_1)^2 = (x - m_2)^2 \quad \Rightarrow \quad x = \frac{m_1 + m_2}{2}$

# Likelihoods with two classes, different variance



## 2.5 Regression

# Regression

- Regression of a function $\mathrm{f}(x)$
  - $r = \mathrm{f}(x) + \epsilon$
  - $x$: independent variable
  - $\mathrm{f}(x)$: dependent variable
  - $\epsilon$: noise
- Approximation of $\mathrm{f}(x)$ using the estimator (hypothesis) $\mathrm{h}(x|\theta)$
  - We can assume $\epsilon \sim \mathcal{N}(0,\sigma^2)$, a Gaussian white noise centred at zero (mean $= 0$) and with constant variance $\sigma^2$

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

# Maximum likelihood estimate

- Log-likelihood with dataset $\mathcal{X} = \{x^t, r^t\}_{t=1}^N$ iid

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

- As $p(x^t)$ is independent of $\theta$ and $p(r|x) \sim \mathcal{N}(\mathrm{h}(x|\theta), \sigma^2)$

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

# Least squares estimate

- Least squares estimate

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

- Maximize likelihood

$$L(\theta|\mathcal{X}) = -N\log\left(\sqrt{2\pi}\sigma\right) - \frac{1}{2\sigma^2}\sum_{t=1}^{N}(r^t - \mathrm{h}(x^t|\theta))^2$$

- $-N\log\left(\sqrt{2\pi}\sigma\right)$ and $1/\sigma^2$ are independent of $\theta$
  - $L(\theta|\mathcal{X}) = -\frac{1}{2}\sum_{t=1}^{N}(r^t - \mathrm{h}(x^t|\theta))^2$
  - $E(\theta|\mathcal{X}) = \frac{1}{2}\sum_{t=1}^{N}(r^t - \mathrm{h}(x^t|\theta))^2$ is the quadratic error
  - Minimizing $E(\theta|\mathcal{X})$ gives a least squares estimate of $\theta$
  - $\theta^*_{MV} = \underset{\forall\theta}{\operatorname{argmax}}\, L(\theta|\mathcal{X})$ is equivalent to $\theta^*_{MC} = \underset{\forall\theta}{\operatorname{argmin}}\, E(\theta|\mathcal{X})$

# Linear regression

- Linear model of $\mathrm{h}(x|\theta)$

$$\mathrm{h}(x^t|w_1,w_0) = w_1 x^t + w_0$$

- Estimate of $w_1$ and $w_0$ according to $E(w_1,w_0|\mathcal{X})$

$$\begin{aligned}
\frac{\partial E(w_1,w_0|\mathcal{X})}{\partial w_0} &= \sum_{t=1}^{N}\left(-r^t + w_1 x^t + w_0\right) = 0 \\
\Rightarrow \sum_{t=1}^{N} r^t &= N w_0 + w_1 \sum_{t=1}^{N} x^t \\
\frac{\partial E(w_1,w_0|\mathcal{X})}{\partial w_1} &= \sum_{t=1}^{N}\left(-r^t x^t + w_1 (x^t)^2 + w_0 x^t\right) = 0 \\
\Rightarrow \sum_{t=1}^{N} r^t x^t &= w_0 \sum_{t=1}^{N} x^t + w_1 \sum_{t=1}^{N} (x^t)^2
\end{aligned}$$

# Matrix formulation (ordre 1)

- Matrix formulation of the estimate of $w_1$ and $w_0$ according to $E(w_1, w_0|\mathcal{X})$

$$\mathbf{Aw} = \mathbf{y}$$

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

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

# Matrix formulation (order $k$)

- A polynomial of order $k$

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

- Solving the equation $\mathbf{Aw} = \mathbf{y}$

$$\mathbf{A} = \begin{bmatrix} N & \sum_t x^t & \sum_t(x^t)^2 & \cdots & \sum_t(x^t)^k \\ \sum_t x^t & \sum_t(x^t)^2 & \sum_t(x^t)^3 & \cdots & \sum_t(x^t)^{k+1} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \sum_t(x^t)^k & \sum_t(x^t)^{k+1} & \sum_t(x^t)^{k+2} & \cdots & \sum_t(x^t)^{2k} \end{bmatrix}, \mathbf{w} = \begin{bmatrix} w_0 \\ w_1 \\ w_2 \\ \vdots \\ w_k \end{bmatrix}, \mathbf{y} = \begin{bmatrix} \sum_t r^t \\ \sum_t r^t x^t \\ \sum_t r^t(x^t)^2 \\ \vdots \\ \sum_t r^t(x^t)^k \end{bmatrix}$$

- By using $\mathbf{A} = \mathbf{D}^\top\mathbf{D}$ and $\mathbf{y} = \mathbf{D}^\top\mathbf{r}$

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

- We can now solve $\mathbf{w} = (\mathbf{D}^\top\mathbf{D})^{-1}\mathbf{D}^\top\mathbf{r}$

# Other types of errors

- Quadratic error

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

- Relative quadratic error

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

- Absolute error

$$E(\theta|\mathcal{X}) = \sum_{t=1}^{N}\left|r^t - \mathrm{h}(x^t|\theta)\right|$$

## 2.6 Bias-variance tradeoff

# Bias-variance tradeoff

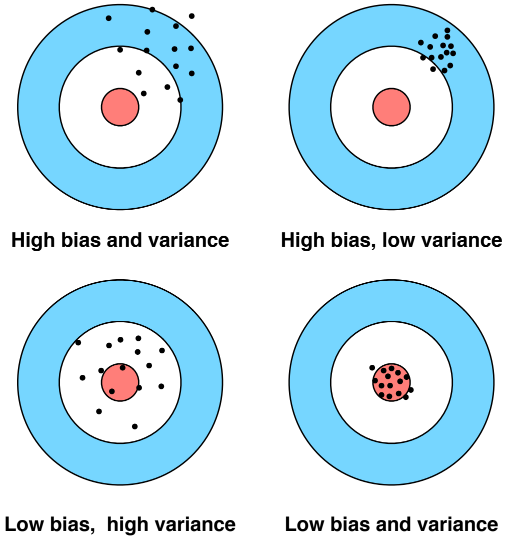
- Expected quadratic error

$$\begin{aligned}
\mathbb{E}\left[(\theta - c)^2\right] &= \mathbb{E}\left[(\theta - \mathbb{E}[\theta])^2\right] + (\mathbb{E}[\theta] - c)^2 \\
\mathbb{E}\left[(r - \mathrm{h}(x))^2|x\right] &= \underbrace{\mathbb{E}\left[(r - \mathbb{E}[r|x])^2|x\right]}_{\text{noise}} + \underbrace{(\mathbb{E}[r|x] - \mathrm{h}(x))^2}_{\text{quadratic error}}
\end{aligned}$$

  - Noise: does not depend on $\mathrm{h}(\cdot)$ or $\mathcal{X} \Rightarrow$ cannot be removed
  - Quadratic error: level of deviation of $\mathrm{h}(\cdot)$ relative to $\mathbb{E}[r|x]$
- Average of $\mathrm{h}(\cdot)$ over all possible $\mathcal{X} \sim p(r,x)$

$$\mathbb{E}_{\mathcal{X}}\left[(\mathbb{E}[r|x] - \mathrm{h}(x))^2\,|x\right] = \underbrace{(\mathbb{E}[r|x] - \mathbb{E}_{\mathcal{X}}[\mathrm{h}(x)])^2}_{\text{bias}^2} + \underbrace{\mathbb{E}_{\mathcal{X}}\left[(\mathrm{h}(x) - \mathbb{E}_{\mathcal{X}}[\mathrm{h}(x)])^2\right]}_{\text{variance}}$$

# Bias and variance



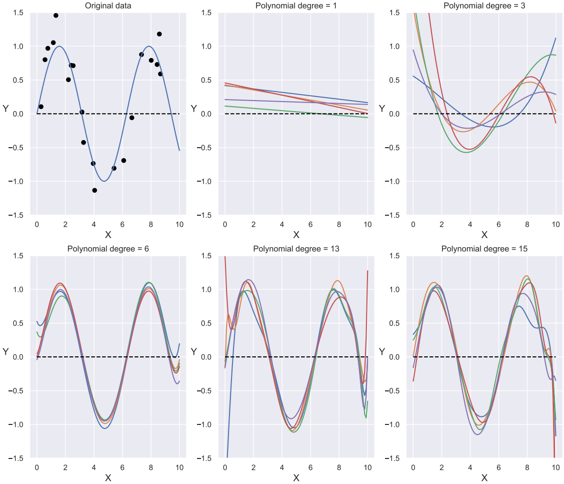
# Example of bias-variance trade-off

- Suppose different data sets $\mathcal{X}_i = \{x_i^t, r_i^t\}$, $i = 1, \ldots, M$, from a noisy function $\mathrm{f}(\cdot) + \epsilon$
  - In practice, we don't know $\mathrm{f}(\cdot)$
  - $\mathrm{h}_i(x)$ generated by training on $\mathcal{X}_i$
  - $\mathbb{E}[\mathrm{h}(x)] = \frac{1}{M}\sum_{i=1}^{M} \mathrm{h}_i(x)$
- Associated bias and variance

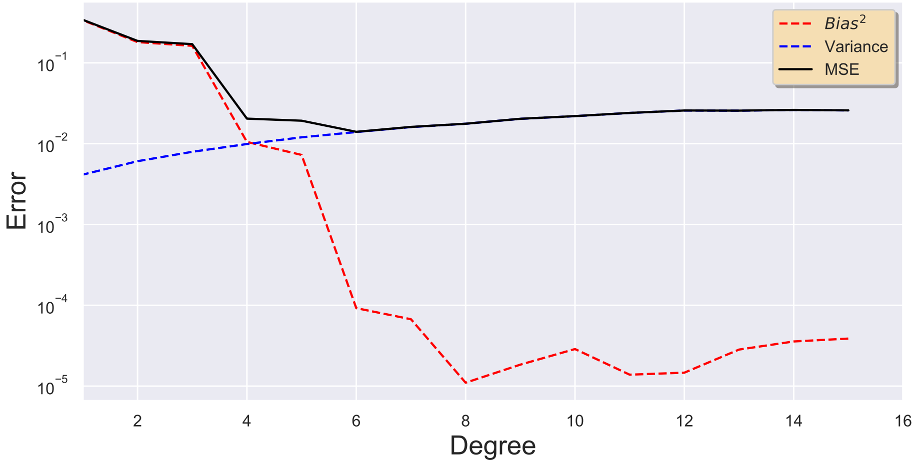
$$\begin{aligned}
\mathrm{bias}^2(\mathrm{h}) &= \frac{1}{N}\sum_{t=1}^{N}\left[\mathbb{E}[\mathrm{h}(x^t)] - \mathrm{f}(x^t)\right]^2 \\
\mathrm{variance}(\mathrm{h}) &= \frac{1}{NM}\sum_{t=1}^{N}\sum_{i=1}^{M}\left[\mathrm{h}_i(x^t) - \mathbb{E}[\mathrm{h}(x^t)]\right]^2
\end{aligned}$$

  - $\mathrm{h}_i(x^t) = c \;\Rightarrow\;$ constant bias, zero variance (underfitting)
  - $\mathrm{h}_i(x^t) = \sum_i r_i^t / N \;\Rightarrow\; \downarrow$ bias, $\uparrow$ variance
  - Low or no bias, high variance: overfitting

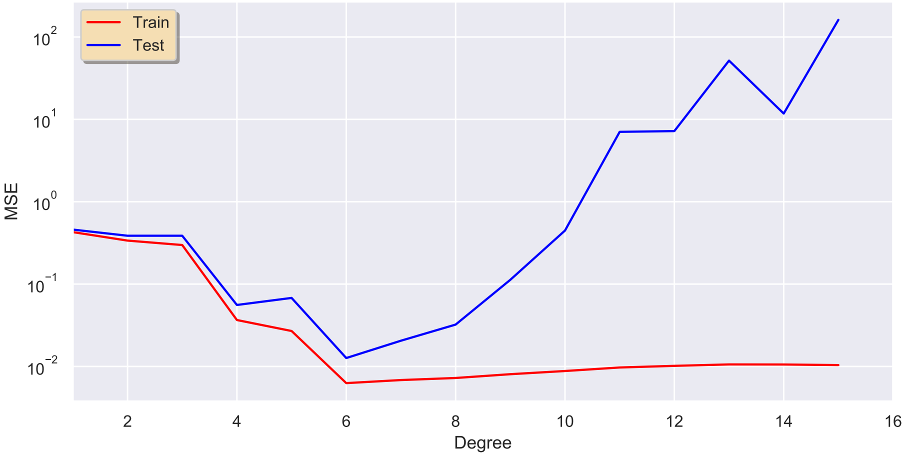
# Complexity and bias-variance trade-offs



# Error vs bias-variance trade-off



# Error vs complexity



# Model selection

- In practice, one cannot calculate the bias and variance of a model
  - Cross-validation provides an empirical measure of total error
- Regularization: integrating a measure of complexity in the optimization

$$E' = (\text{empirical error}) + \lambda\ (\text{model complexity})$$

  - $\lambda$ controls complexity penalty
  - $\lambda$ usually adjusted by cross-validation
- Measures of complexity
  - Vapnik-Chervonenkis Dimension (VC-dim)
  - *Minimum description length*: description of the minimum size of data