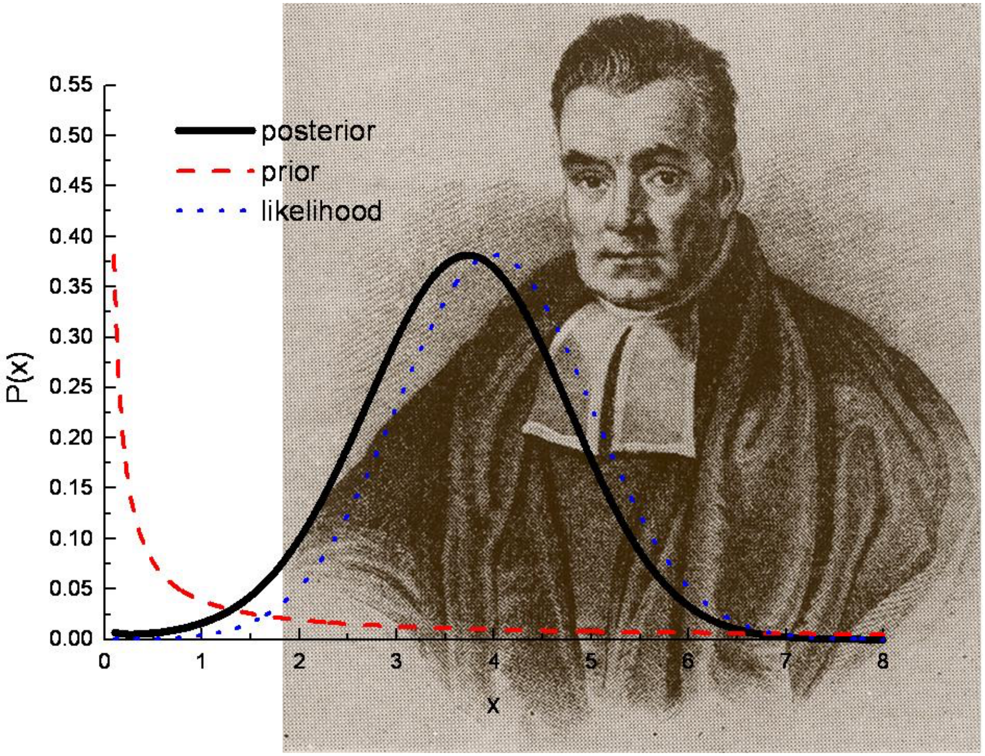
Point Estimation

Hongxin Zhang
zhx@cad.zju.edu.cn

State Key Lab of CAD&CG, ZJU
2021-03-02

What you need to know

- Point estimation: (点估计)
  - **M**aximal **L**ikelihood **E**stimation (MLE)
  - Bayesian learning
  - **M**aximize **A** **P**osterior (MAP)
- Gaussian estimation
- Regression （回归）
  - Basis function = features
  - Optimizing sum squared error
  - Relationship between regression and Gaussians
- Bias-Variance trade-off

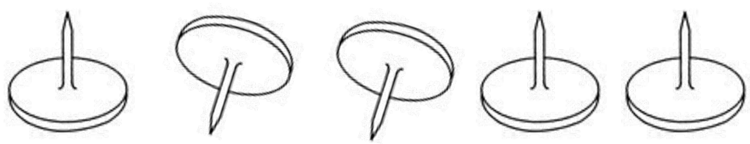


点估计是非常重要的数据计算技术，需要在入门阶段重点掌握

Point estimation

Your first consulting job

- An IT billionaire from USA asks you a question:
  - B: I have thumbtack, if I flip it, what's the probability it will fall with the nail up?
  - Y: Please flip it a few times …



  - Y: The probability is 3/5
  - B: Why???
  - Y: Because…

Binomial Distribution

- $P(\text{Heads}) = \theta$, $P(\text{Tails}) = 1\text{-}\theta$ $\quad D = \{T, H, H, T, T\}$

$$P(D \mid \theta) = (1-\theta)\theta\theta(1-\theta)(1-\theta)$$

- Flips are i.i.d. (Independent Identically distributed)
  - Independent events
  - Identically distributed according to Binomial distribution
- Sequence D of α_H Heads and α_T Tails

$$P(D \mid \theta) = \theta^{\alpha_H}(1-\theta)^{\alpha_T}$$

Maximum Likelihood Estimation

- **Data**: Observed set D of α_H Heads and α_T Tails
- **Hypothesis**: Binomial distribution
- **Learning θ is an optimization problem**
  - What's the objective function?

$$D = \{T, H, H, T, T\}$$

- **MLE**: Choose θ that maximizes the probability of observed data:

$$\begin{aligned}\hat{\theta} &= \arg\max_{\theta} P(D \mid \theta) \\ &= \arg\max_{\theta} \ln P(D \mid \theta) = \ldots\end{aligned}$$

Maximum Likelihood Estimation (cont.)

$$\begin{aligned}\hat{\theta} &= \arg\max_{\theta} P(D \mid \theta) \\ &= \arg\max_{\theta} \ln(\theta^{\alpha_H}(1-\theta)^{\alpha_T}) \\ &= \arg\max_{\theta} (\alpha_H \ln \theta + \alpha_T \ln(1-\theta))\end{aligned}$$

- Set derivative to zero:

$$\boxed{\frac{d}{d\theta} \ln P(D \mid \theta) = 0} \qquad \hat{\theta} = \frac{\alpha_H}{\alpha_H + \alpha_T} = \frac{2}{2+3}$$

How many flips do I need?

$$\hat{\theta} = \frac{\alpha_H}{\alpha_H + \alpha_T}$$

- B: I flipped 2 heads and 3 tails.
- Y: 1- θ = 3/5, I can prove it!
- B: What if I flipped 20 heads and 30 tails?
- Y: Same answer, I can prove it!
- B: What’s better?
- Y: Humm… The more the merrier???
- B: Is this why I am paying you the big bucks???

Simple bound (based on Höffding's inequality)

- For $N = \alpha_H + \alpha_T$ and $\hat{\theta} = \dfrac{\alpha_T}{\alpha_H + \alpha_T}$

http://omega.albany.edu:8008/machine-learning-dir/notes-dir/vc1/vc-l.html

- Let θ* be the true parameter, for any ε>0:

$$\boxed{P(\left|\hat{\theta} - \theta^*\right| \geq \varepsilon) \leq 2e^{-2N\varepsilon^2}} \leq \delta$$

$$N \geq \frac{1}{2\varepsilon^2}[\ln 2 - \ln \delta]$$

$$N \geq 270; (\varepsilon = 0.1, \delta = 0.01)$$

PAC Learning

- **PAC**: **Probably Approximate Correct**
- B: I want to know the thumbtack parameter θ, within ε = 0.1, with probability at least 1-δ = 0.99. How many flips?
- Y: 270, ☺

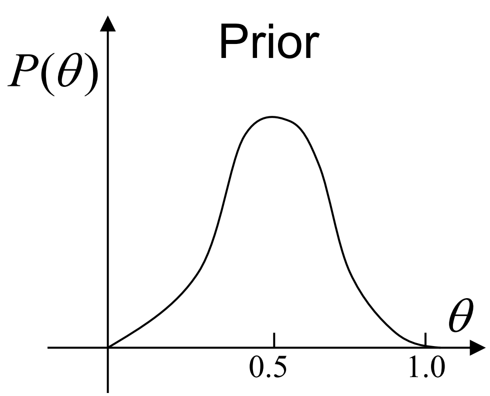
Prior: knowledge before experiments

- B: Wait, I know that the thumbtack is "close" to 50-50. What can you …?
- Y: I can learn it the Bayesian way…

- Rather than estimating a single θ, we obtain a distribution over possible values of θ

先验

Prior



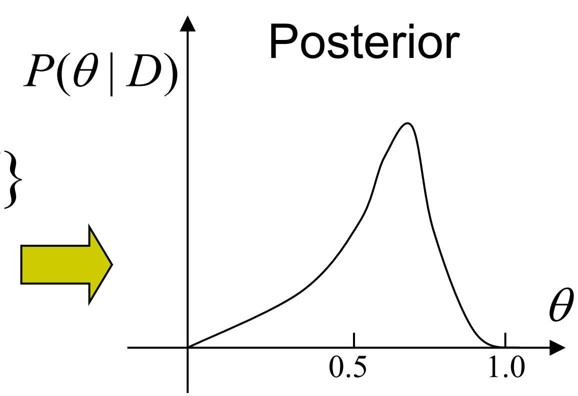
数据

Data

$$D = \{T, H, H, T, T\}$$

后验

Posterior



从一个假想案例说起

- 王某去医院作验血化验，检查他患上了X疾病的可能性，其结果居然为阳性，把他吓了一大跳，赶忙到网上查询…

- 相关资料表明：该化验有
  - 1%假阳性率
    在得病的人中做实验，有1%的人是假阳性，99%的人是真阳性
  - 1%假阴性率
    在未得病的人中做实验，有1%的人是假阴性，99%的人是真阴性
- 问王某是否病了？或者说王宏得病概率是否很大？

- 原始内容来源：http://news.sciencenet.cn/news/sub26.aspx?id=2958

从一个假想案例说起

- 王某去医院作验血化验，检查他患上了X疾病的可能性，其结果居然为阳性，把他吓了一大跳，赶忙到网上查询...
- 相关资料表明：该化验有
  - 1%假阳性率
    在得病的人中做实验，有1%的人是假阳性，99%的人是真阳性
  - 1%假阴性率
    在未得病的人中做实验，有1%的人是假阴性，99%的人是真阴性
- 关于王某得病概率，医生的回答是：9%（不要慌）
- 为什么：
  - “这种X疾病的正常比例是不大的，1000个人中只有一个人有X病”

从一个假想案例说起

- 已知条件：
  - 相关资料表明：该化验有
    - 1%假阳性率
      在得病的人中做实验，有1%的人是假阳性，99%的人是真阳性
    - 1%假阴性率

      在未得病的人中做实验，有1%的人是假阴性，99%的人是真阴性
  - 这种X疾病的正常比例是不大的，1000个人中只有一个人有X病，即0.1%的发病率

- 医生的计算方法：
  - 因为测试的误报率是1%，1000个人将有10个被报为“假阳性”，而根据X病在人口中的比例为1/1000，真阳性只有1个
  - 所以，大约11个测试为阳性的人中只有一个是真阳性（有病）的，因此，王宏被感染的几率是大约1/11，即0.09(9%)

从一个假想案例说起

张天蓉博主博文中提到的一个概率问题引发了科学网博客上关于贝叶斯定理的大讨论，精彩纷呈，妙趣横生，让观者如痴如醉。

A：普通人群中的王宏感染X病
B：阳性结果

$P(A)$ 普通人群中感染X病的概率

$P(B|A)$ 阳性结果的正确率

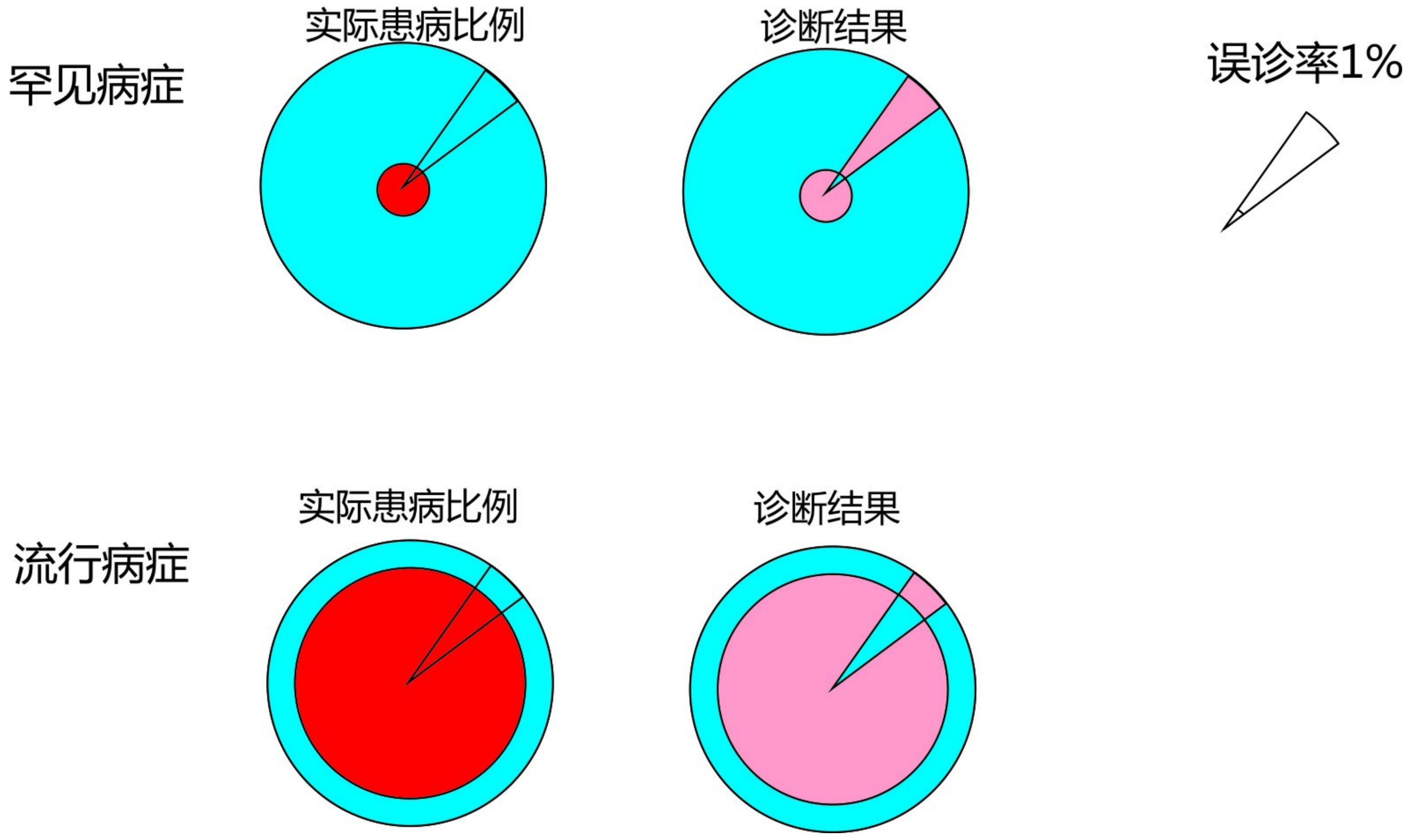
$P(A|B)$ 有了阳性结果的条件下，王宏感染X病之概率

$P(B)$ 结果为阳性的总可能性 = 检查阳性中的真阳性 + 检查阴性中的真阳性

$$\boxed{P(A|B)} = \frac{P(B|A)}{P(B)} \boxed{P(A)} = \frac{99\%}{99\% * (1/1000) + 1\% * (999/1000)} \times (1/1000)$$

$$= \frac{99}{1098} = 9\%$$

从一个假想案例说起



从一个假想案例说起

我们之前听说，好莱坞影星安吉丽娜朱莉通过《纽约时报》向大家揭露了她惨痛的经历。因携带有BRCA1/2致癌基因，她有87%的机会患上乳腺癌（我们假设5年）。为了防患于未然，她决定切除双乳(确切说是切除双侧乳腺)。美国乳腺癌的发病率为246,680/104,442,302=0.236%（每年每1000成年妇女中有2.36个人患上癌症。我们把预防癌症定位为5年不得癌症。

A:普通妇女5年癌症得病；

B：BRCA1/2致癌基因阳性；

P(A)：普通妇女人群5年患癌概率1.18%；

P(B｜A)：乳腺癌病人中BRCA1/2致癌基因阳性检测率87%；

P(A｜B)：有阳性结果的条件下，安吉丽娜朱莉5年内患癌概率；

P(B)：结果为阳性的总可能性=检查阳性中的真阳性+检查阴性中的真阳性。

通过贝叶斯公式，P(A｜B)= P(B｜A)×P(A)/ P(B)=87%×1.18%/（87%×1.18%+13%×98.2%）=7.4%。

Bayesian Learning

- Bayes rule:

Prior　　Likelihood

$$\text{Posterior} \longrightarrow P(\theta \mid D) = \frac{\overbrace{P(\theta)}^{\text{Prior}}\,\overbrace{P(D \mid \theta)}^{\text{Likelihood}}}{P(D)} \longleftarrow \text{Data distribution}$$

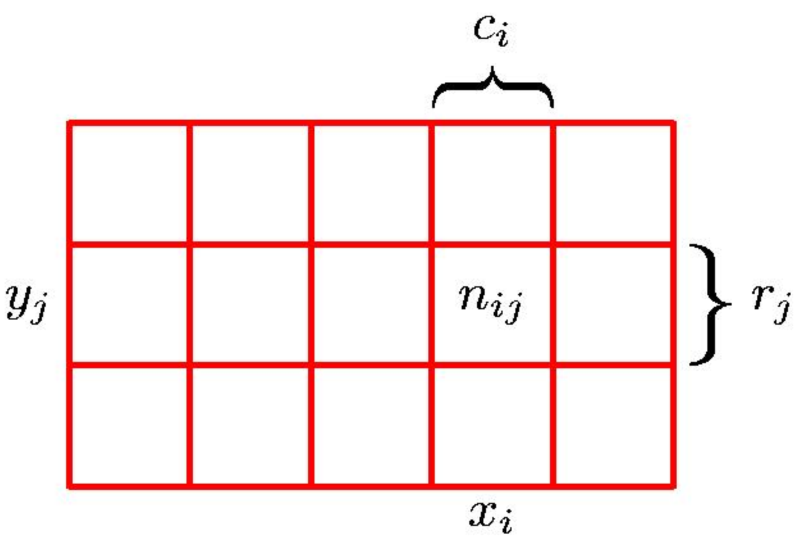
evidence

(Normalization)

- Or equivalently:

$$P(\theta \mid D) \propto P(\theta)P(D \mid \theta)$$

Probability Theory



- Sum Rule

$$p(X = x_i) = \frac{c_i}{N} = \frac{1}{N}\sum_{j=1}^{L} n_{ij}$$

$$= \sum_{j=1}^{L} p(X = x_i, Y = y_j)$$

Product Rule

$$\begin{aligned} p(X = x_i, Y = y_j) &= \frac{n_{ij}}{N} = \frac{n_{ij}}{c_i} \cdot \frac{c_i}{N} \\ &= p(Y = y_j | X = x_i) p(X = x_i) \end{aligned}$$

Probability concepts

- Random variables: x
- Probability (function): $P(X \le x)$, $P(x)$
- Density (function): $f(x)$,
- Independency: $P(x, y)=P(x)P(y)$
- Feature quantities:
  - Mean, expectation $E(x) = \int x\, f(x)\, \mathrm{d}x$
  - Covariance
    - $\mathrm{cov}(x,y)=0$, uncorrelatedness / irrelevant (统计无关)
  - Higher order moments

The Rules of Probability

- Sum Rule $$p(X) = \sum_Y p(X, Y)$$
- Product Rule $$p(X, Y) = p(Y|X)p(X)$$

Bayes' Theorem

$$p(Y|X) = \frac{p(X|Y)p(Y)}{p(X)}$$

$$p(X) = \sum_Y p(X|Y)p(Y)$$

posterior $\propto$ likelihood $\times$ prior

Bayesian Learning in our case

- Likelihood function is simply Binomial:

$$P(D \mid \theta) = \theta^{\alpha_H}(1-\theta)^{\alpha_T}$$

- What about prior?
  - Represent expert knowledge
  - Simple posterior form
- Conjugate priors: （共轭先验）
  - Closed-form representation of posterior
  - For Binomial, conjugate prior is Beta distribution

Beta prior distribution – P(θ)

- Prior: Beta distribution

Gamma function

$$\Gamma(x+1) = x\Gamma(x), \Gamma(1) = 1$$

$$P(\theta) = \frac{\theta^{\beta_H - 1}(1-\theta)^{\beta_T - 1}}{B(\beta_H, \beta_T)} \sim Beta(\theta \mid \beta_H, \beta_T) = \frac{\Gamma(\beta)}{\Gamma(\beta_H)\Gamma(\beta_T)} \theta^{\beta_H - 1}(1-\theta)^{\beta_T - 1}$$

- Likelihood: Binomial distribution

$$P(D \mid \theta) = \theta^{\alpha_H}(1-\theta)^{\alpha_T}$$

- Posterior:

$$\begin{aligned} P(\theta \mid D) &\propto P(\theta)P(D \mid \theta) \\ &\propto \theta^{\alpha_H}(1-\theta)^{\alpha_T}\theta^{\beta_H - 1}(1-\theta)^{\beta_T - 1} \\ &\sim Beta(\alpha_H + \beta_H, \alpha_T + \beta_T) \end{aligned}$$

Using Bayesian posterior

- Posterior distribution:

$$P(\theta \mid D) \sim Beta(\alpha_H + \beta_H, \alpha_T + \beta_T)$$

- Bayesian inference:
  - No longer single parameter:

$$E[f(\theta)] \sim \int_0^1 f(\theta) P(\theta \mid D) d\theta$$

  - Integral, ☹

Expectation (数学期望)

- Random variable: θ
- Random function: $f(\theta)$
- Expectation:

$$E[f(\theta)] \sim \int_0^1 f(\theta) P(\theta \mid D) d\theta$$

MAP: Maximum a posteriori approximation

$$P(\theta \mid D) \sim Beta(\alpha_H + \beta_H, \alpha_T + \beta_T)$$

$$E[f(\theta)] = \int_0^1 f(\theta) P(\theta \mid D) d\theta$$

approximation

- MAP: use most likely parameter

$$\widehat{\theta} = \arg\max_{\theta} P(\theta \mid D) \qquad E[f(\theta)] \approx f(\widehat{\theta})$$

MAP for Beta distribution

$$P(\theta \mid D) \sim Beta(\alpha_H + \beta_H, \alpha_T + \beta_T)$$

- MAP: use most likely parameter

$$\hat{\theta} = \arg\max_{\theta} P(\theta|D) = \frac{\alpha_H + \beta_H - 1}{\alpha_H + \beta_H + \alpha_T + \beta_T - 2}$$

- Beta prior equivalent to extra thumbtack flips
- As $N = \alpha_T + \alpha_H \to \infty$, prior is "forgotten"
- But, for small sample size, prior is important!

More …

- B: Can we handle more complex cases?
- Y: Yes, :-D

- Prior: a mixture of beta distribution
  - $P(\theta) \sim 0.4 Beta(20,1) + 0.4 Beta(1,20) + 0.2 Beta(2,2)$

Multinomial distribution

- B: Now if I give you a dice (骰子), then …
- Y: I can solve this problem in a similar way.
- Likelihood:

$$P(X = x^k \mid \boldsymbol{\theta}) = \theta_k,\ k = 1, 2, \ldots, r,$$

$$\boldsymbol{\theta} = \{\theta_1, \ldots, \theta_r\},\ \theta_1 + \ldots + \theta_r = 1$$

$$D = \{X_1 = x_1, \ldots, X_N = x_N\} => \{N_1, \ldots, N_r\}$$

$$\boxed{P(D \mid \boldsymbol{\theta}) = \prod_{i=1}^{r} \theta_i^{N_i}}$$

Multinomial distribution

- Conjugate prior (Dirichlet distribution):

$$P(\boldsymbol{\theta}) = \mathrm{Dir}(\boldsymbol{\theta} \mid \alpha_1, \ldots, \alpha_r) = \frac{\Gamma(\alpha)}{\prod_{k=1}^{r} \Gamma(\alpha_k)} \prod_{k=1}^{r} \theta_k^{\alpha_k - 1}, \quad \alpha = \sum_{k=1}^{r} \alpha_k$$

- Solution:

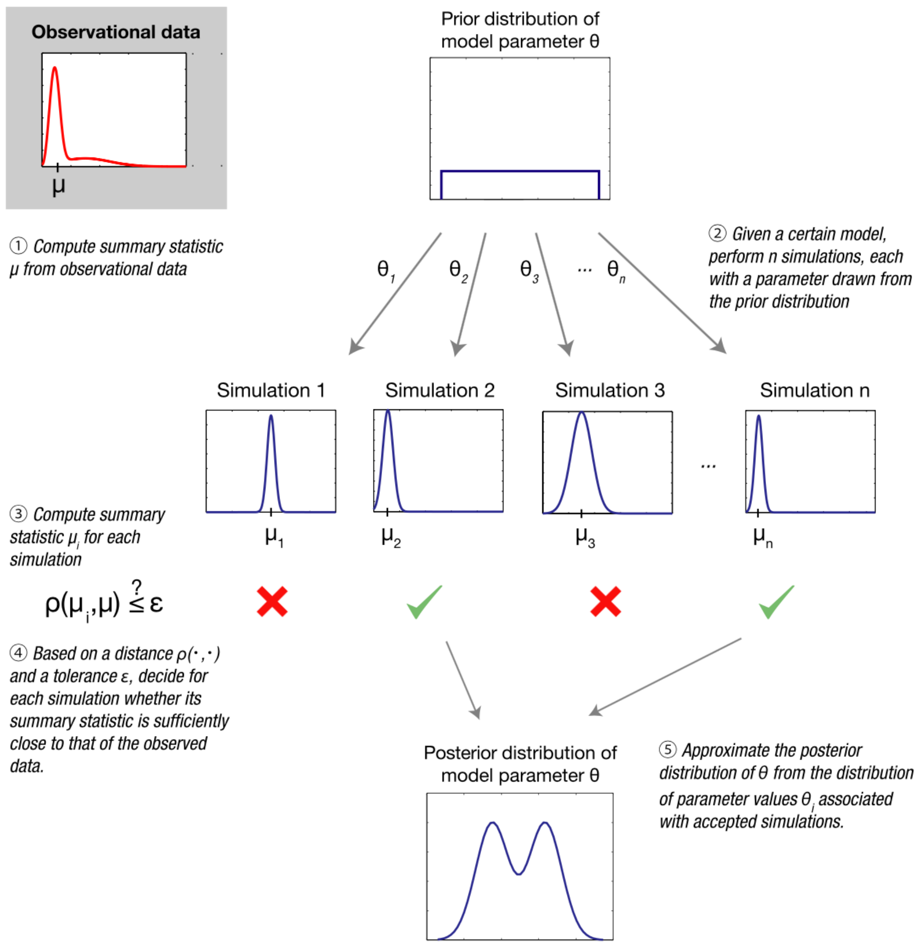
$$P(X_{N+1} = x^k \mid D) = \int \theta_k \mathrm{Dir}(\boldsymbol{\theta} \mid \alpha_1 + N_1, \ldots, \alpha_r + N_r) d\boldsymbol{\theta} = \frac{\alpha_k + N_k}{\alpha + N}$$

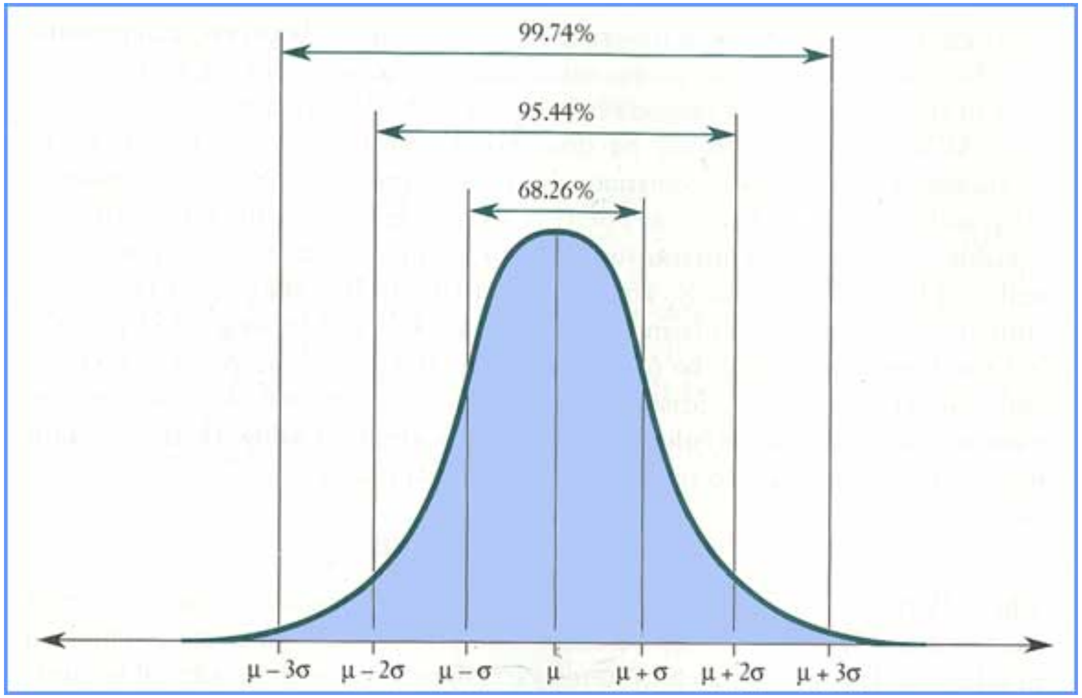
- Important fact:

$$P(D) = \frac{\Gamma(\alpha)}{\Gamma(\alpha + N)} \prod_{k=1}^{r} \frac{\Gamma(\alpha_k + N_k)}{\Gamma(\alpha_k)}$$

Beyond MAP:
Approximate Bayesian Computation

- Approximate Bayesian Computational methods [ARXIV]
- Approximate Bayesian Computation (ABC) in practice [PDF]





在数据处理中，高斯分布无处不在

Gaussian distribution

Gaussian distribution

均值

Continuous random variable:

mean

$$P(x \mid \mu, \delta) \sim \frac{1}{\boxed{\delta}\sqrt{2\pi}} e^{-\frac{(x-\boxed{\mu})^2}{2\delta^2}}$$

variance

standard deviation

Normalization

方差

标准差

Consider the difference between continuous and discrete variables?

MLE for Gaussian

- Prob. of i.i.d. samples $D = \{x_1, x_2, \ldots, x_N\}$

likelihood $\quad P(D \mid \mu, \sigma) = \left(\frac{1}{\sigma\sqrt{2\pi}}\right)^N \prod_{i=1}^{N} e^{\frac{-(x_i-\mu)^2}{2\sigma^2}}$

- The magic of log (to log-likelihood)

$$\begin{aligned} \ln P(D \mid \mu, \sigma) &= \ln\left(\frac{1}{\sigma\sqrt{2\pi}}\right)^N \prod_{i=1}^{N} e^{\frac{-(x_i-\mu)^2}{2\sigma^2}} \\ &= -N\ln(\sigma\sqrt{2\pi}) - \sum_{i=1}^{N} \frac{(x_i-\mu)^2}{2\sigma^2} \end{aligned}$$

MLE for mean of a Gaussian

$$
\begin{aligned}
\frac{\partial}{\partial \mu} \ln P(D \mid \mu, \sigma) &= \frac{\partial}{\partial \mu} \ln \left(\frac{1}{\sigma \sqrt{2\pi}} \right)^N \prod_{i=1}^{N} e^{\frac{-(x_i - \mu)^2}{2\sigma^2}} \\
&= \frac{\partial}{\partial \mu} - \sum_{i=1}^{N} \frac{(x_i - \mu)^2}{2\sigma^2} \\
&= \sum_{i=1}^{N} \frac{(x_i - \mu)}{\sigma^2} = 0
\end{aligned}
$$

$$
\mu = \frac{1}{N} \sum_{i} x_i
$$

MLE for variance of a Gaussian

$$\begin{aligned}
\frac{\partial}{\partial\sigma}\ln P(D\mid\mu,\sigma) &= \frac{\partial}{\partial\sigma}\ln\left(\frac{1}{\sigma\sqrt{2\pi}}\right)^N \prod_{i=1}^{N} e^{\frac{-(x_i-\mu)^2}{2\sigma^2}} \\
&= \frac{\partial}{\partial\sigma}[-N\ln\sigma\sqrt{2\pi}]-\sum_{i=1}^{N}\frac{\partial}{\partial\sigma}[\frac{(x_i-\mu)^2}{2\sigma^2}] \\
&= -\frac{N}{\sigma}+\sum_{i=1}^{N}\frac{(x-\mu)^2}{\sigma^3} \quad = 0
\end{aligned}$$

$$\sigma^2=\frac{1}{N}\sum_{i}(x_i-\mu)^2$$

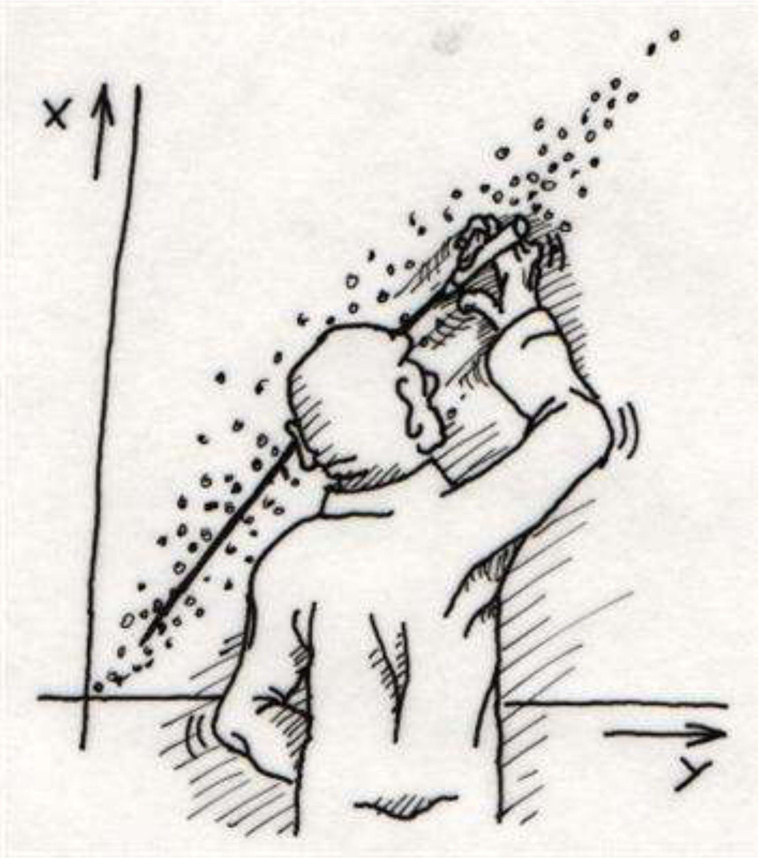
Gaussian parameters learning

- MLE

$$\hat{\mu} = \frac{1}{N}\sum_i x_i$$

$$\hat{\sigma}^2 = \frac{1}{N}\sum_i (x_i - \mu)^2$$

- Bayesian learning: prior?
- Conjugate priors:
  - Mean: Gaussian priors
  - Variance: Wishart Distribution

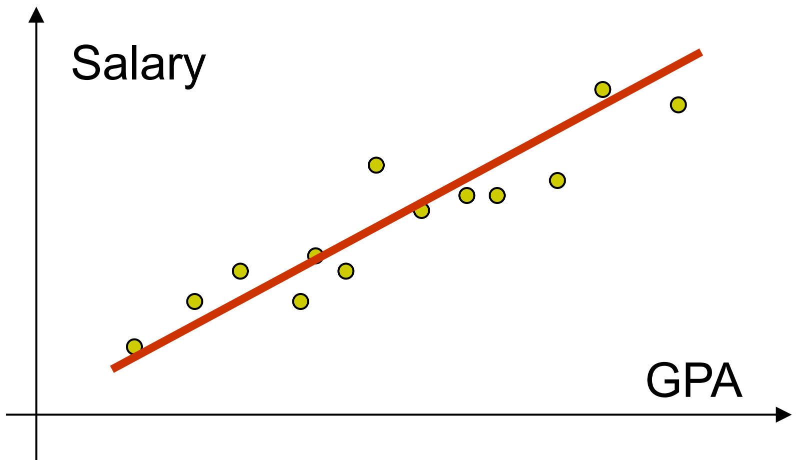


回归的思想，对于理解数据本身的特点与规律，是一种利器

Regression problems

Prediction of continuous variable

- B: Wait, that's not what I meant!
- Y: Chill out, dude.
- B: I want to predict a continuous variable for continuous inputs: I want to predict salaries from GPA.
- Y: I can regress that…



The regression problem

- **Instances:** $<\mathbf{x}_i, t_i>$
- **Learn:** mapping from $\mathbf{x}$ to $t(\mathbf{x})$.
- **Hypothesis space:** $t(\mathbf{x}) \approx \hat{f}(x) = \sum_{i=1}^{k} w_i h_i$
  - Given, basis functions $H = \{h_1, \ldots, h_k\}$
  - Find coefficients $\mathbf{w} = \{w_1, \ldots, w_k\}$
- **Problem formulation:**

$$\mathbf{w}^* = \arg\min_{\mathbf{w}} \sum_j [t(\mathbf{x}_j) - \sum_{i=1}^{k} w_i h_i(x)]^2$$

But, why sum squared error?

- Model:

$$P(t \mid \mathbf{x}, \mathbf{w}, \sigma) = \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{-[t - \sum_i w_i h_i(x)]^2}{2\sigma^2}}$$

- Learn **w** using MLE

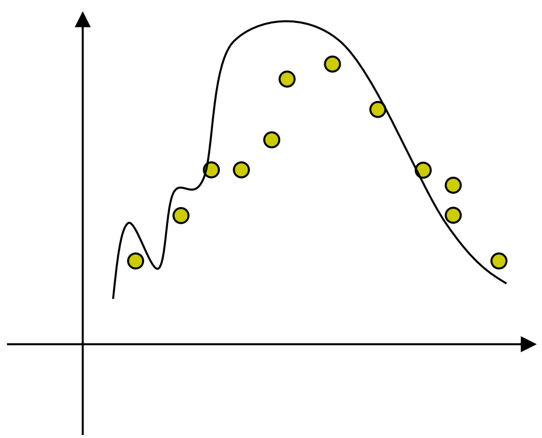
Maximizing log-likelihood

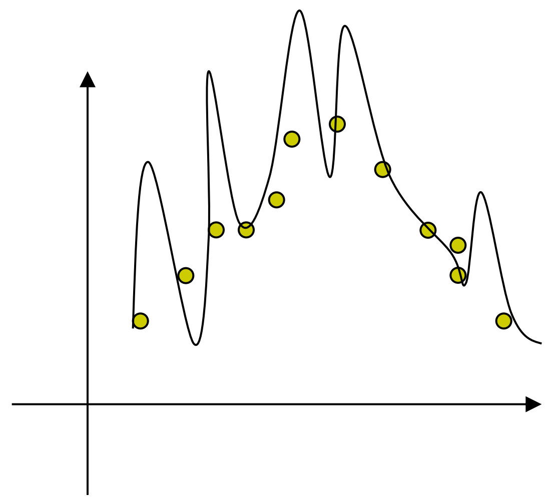
$$\ln P(D \mid \mathbf{w}, \sigma) = \ln \prod_j \left(\frac{1}{\sigma\sqrt{2\pi}} e^{\frac{-[t_j - \sum_i w_i h_i(x_j)]^2}{2\sigma^2}} \right)$$

$$\Rightarrow \min \sum_j \frac{-[t_j - \sum_i w_i h_i(x_j)]^2}{2\sigma^2}$$

Bias-Variance Tradeoff

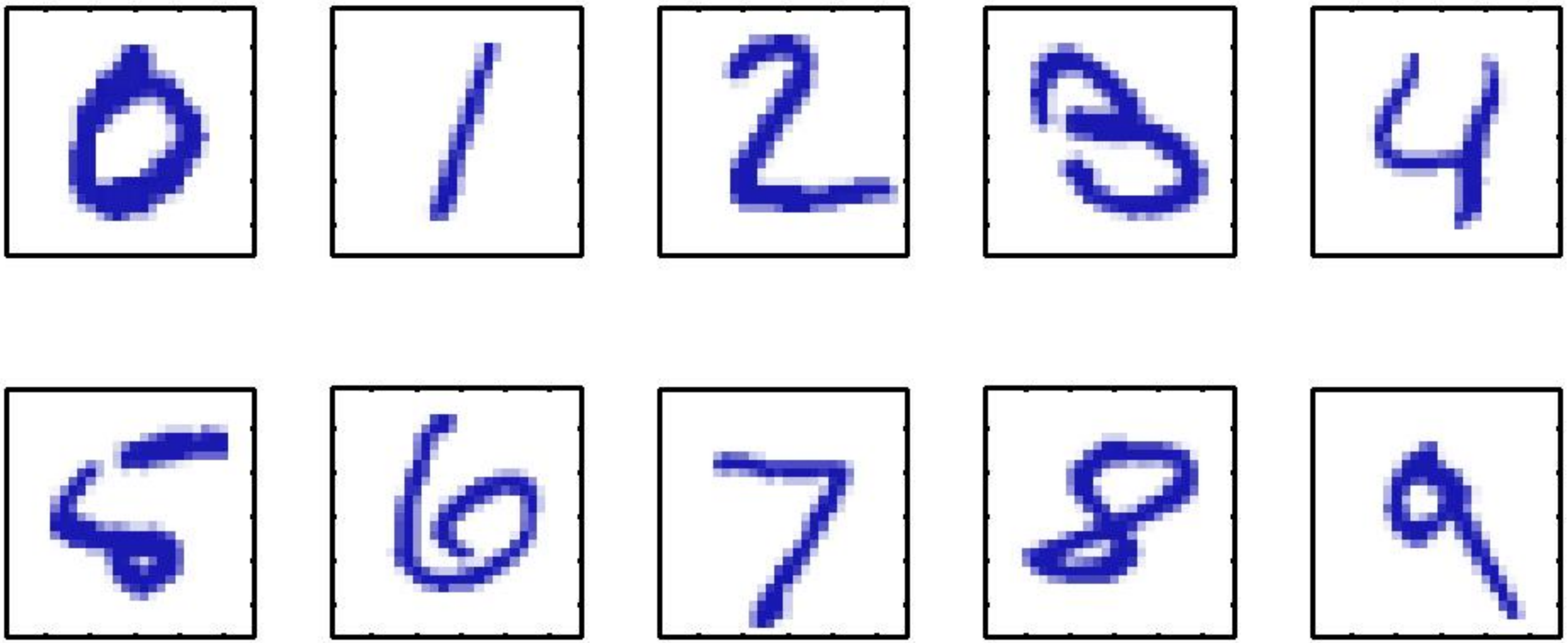
- Choice of hypothesis basis introduce learning bias:
  - More complex basis:
    - Less bias
    - More variance (over-fitting)



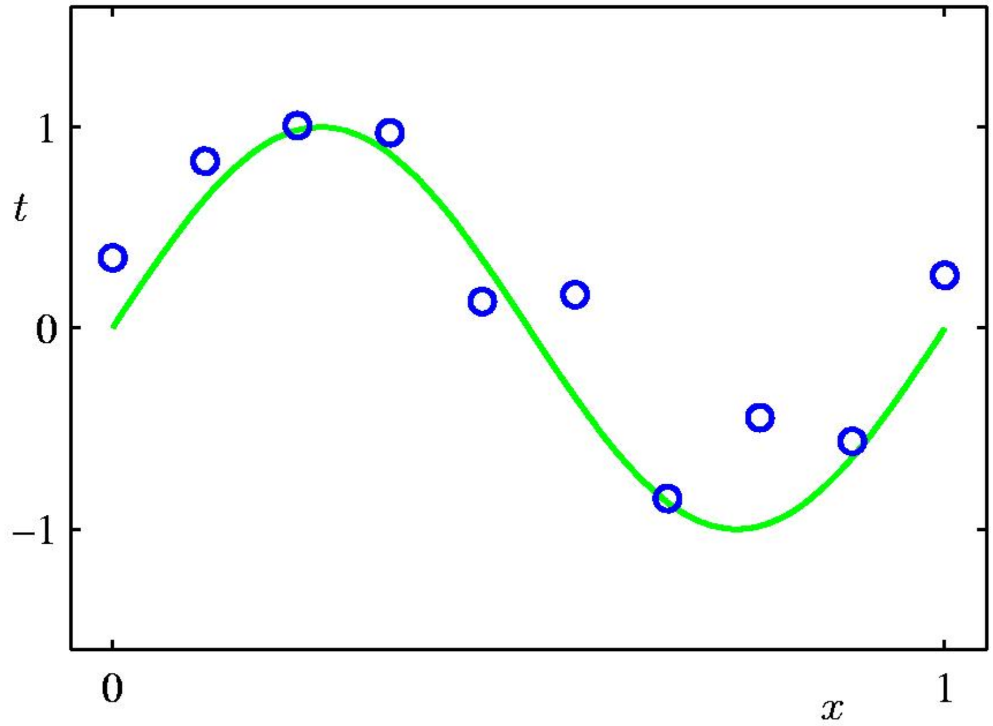


Example

Handwritten Digit Recognition

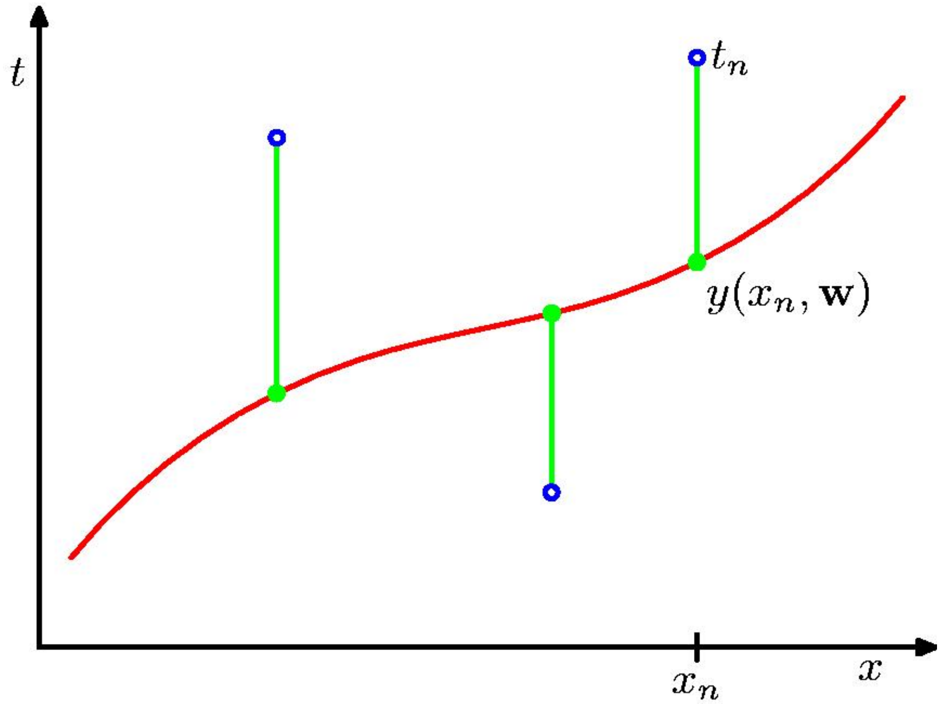


Polynomial Curve Fitting



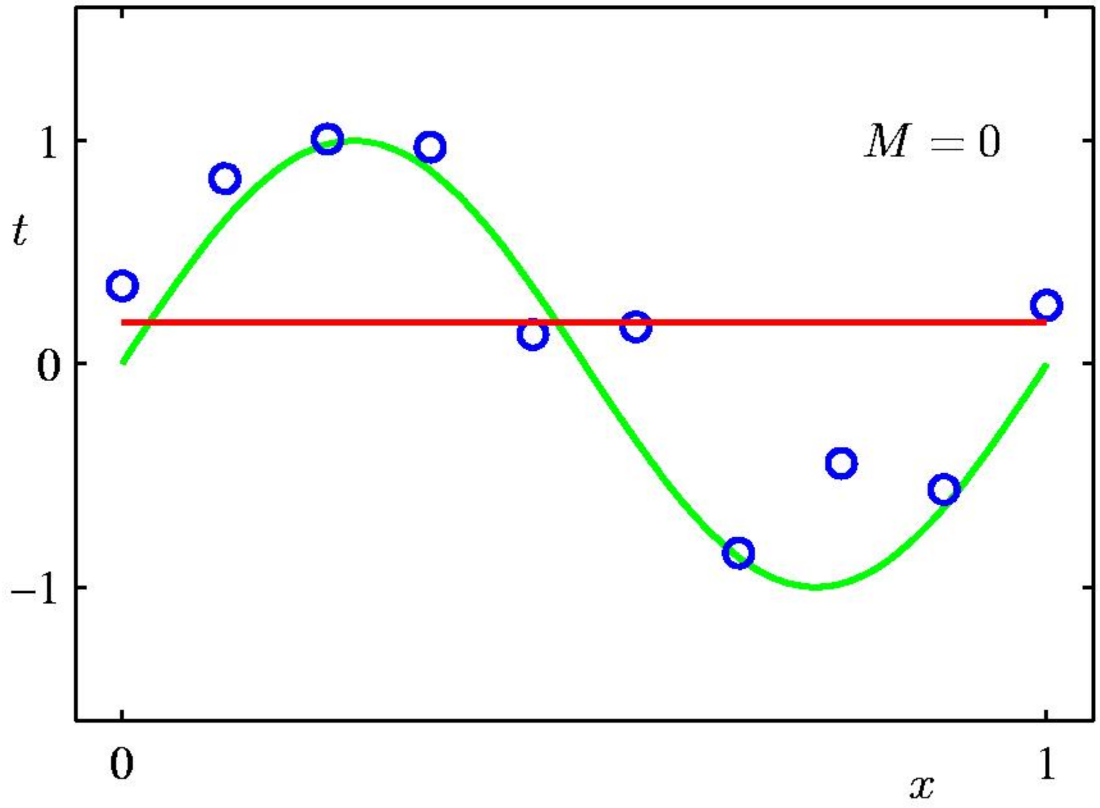
$$y(x, \mathbf{w}) = w_0 + w_1 x + w_2 x^2 + \ldots + w_M x^M = \sum_{j=0}^{M} w_j x^j$$

Sum-of-Squares Error Function

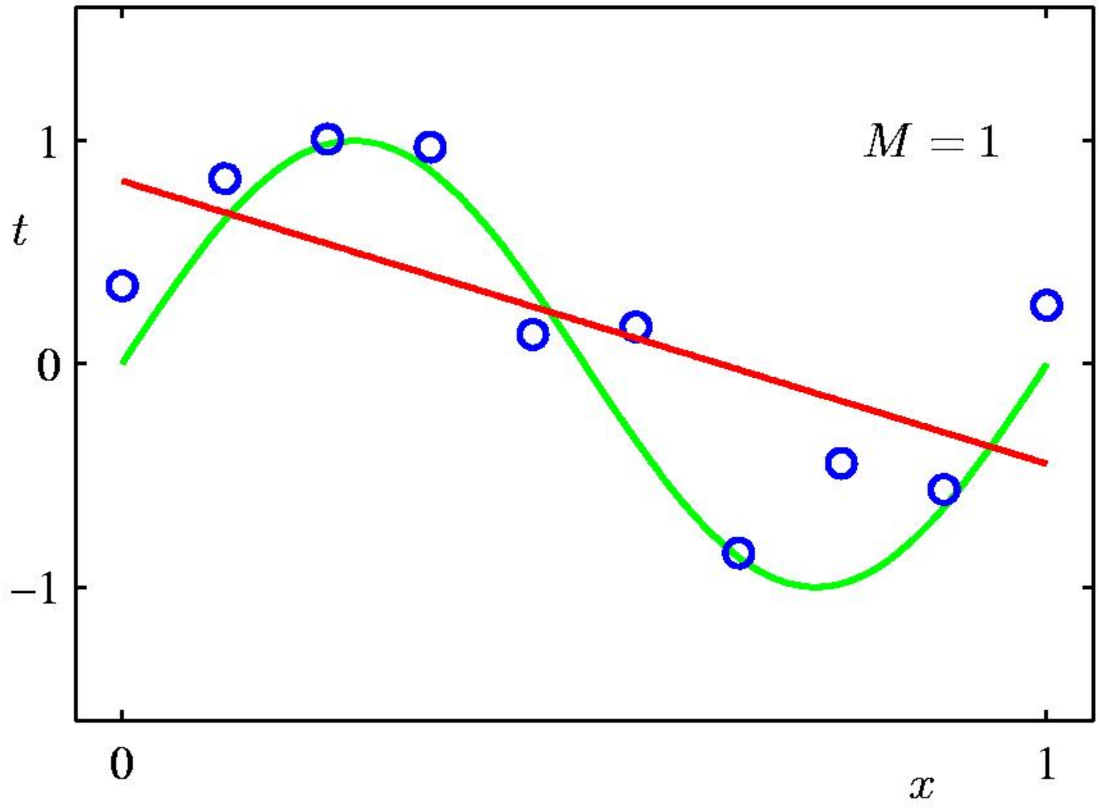


$$E(\mathbf{w}) = \frac{1}{2} \sum_{n=1}^{N} \{y(x_n, \mathbf{w}) - t_n\}^2$$

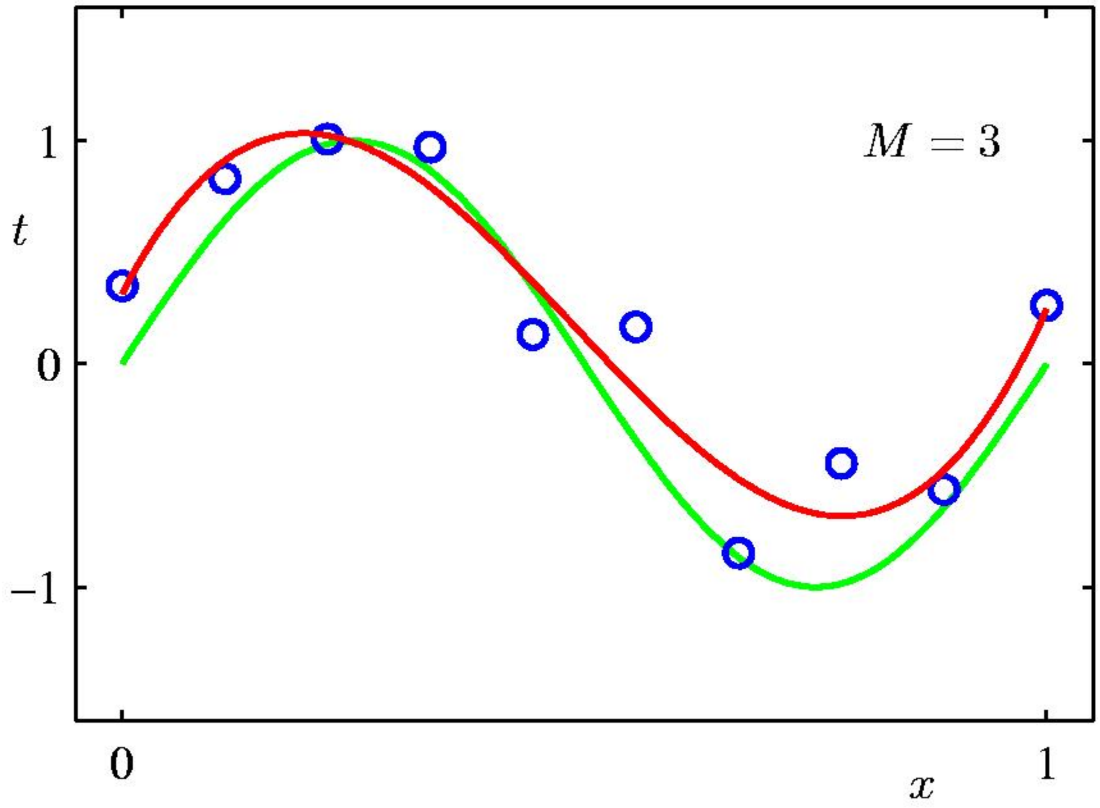
0^{th} Order Polynomial



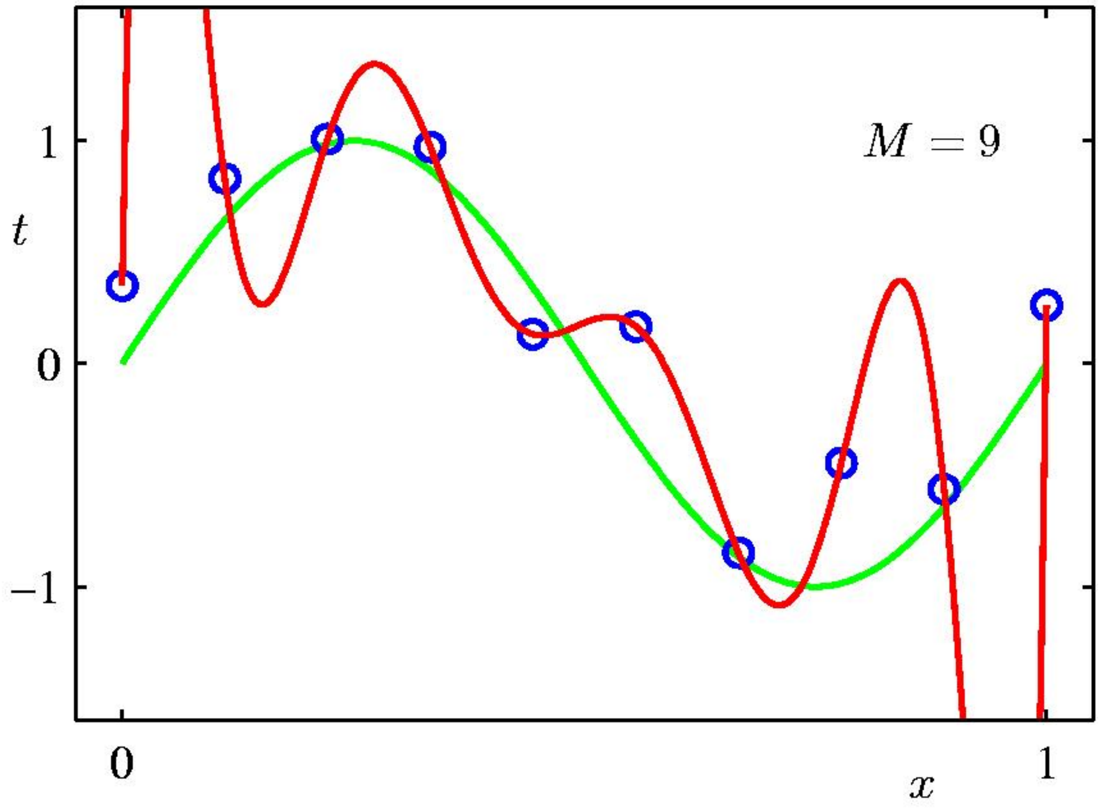
1st Order Polynomial



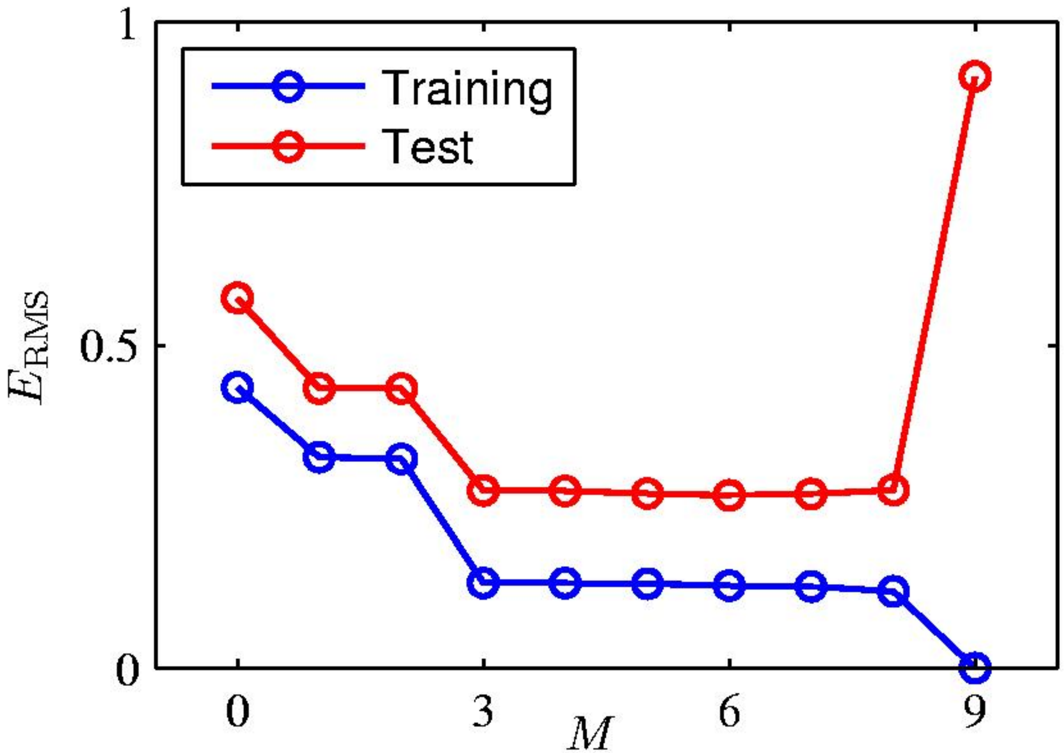
3rd Order Polynomial



9^{th} Order Polynomial



Over-fitting



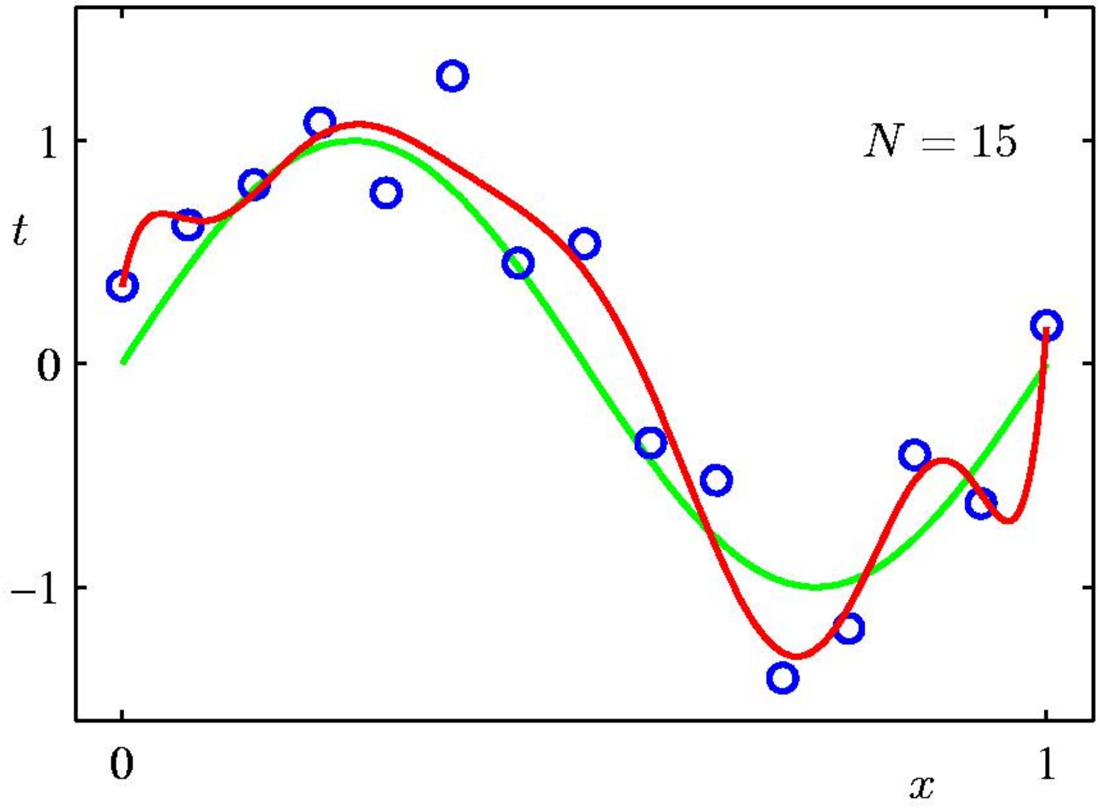
Root-Mean-Square (RMS) Error: $E_{\mathrm{RMS}} = \sqrt{2E(\mathbf{w}^\star)/N}$

Polynomial Coefficients

	$M = 0$	$M = 1$	$M = 3$	$M = 9$
$w_0^\star$	0.19	0.82	0.31	0.35
$w_1^\star$		-1.27	7.99	232.37
$w_2^\star$			-25.43	-5321.83
$w_3^\star$			17.37	48568.31
$w_4^\star$				-231639.30
$w_5^\star$				640042.26
$w_6^\star$				-1061800.52
$w_7^\star$				1042400.18
$w_8^\star$				-557682.99
$w_9^\star$				125201.43

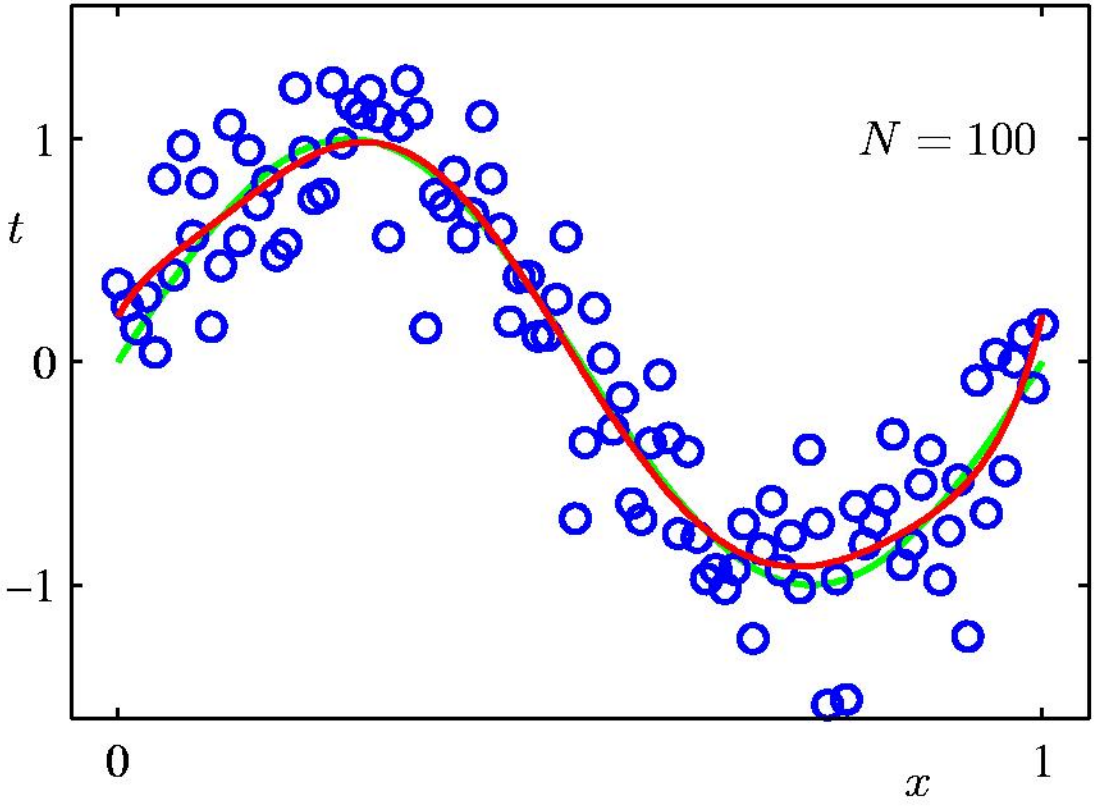
Data Set Size: $N = 15$

9^{th} Order Polynomial



Data Set Size: $N = 100$

9^{th} Order Polynomial

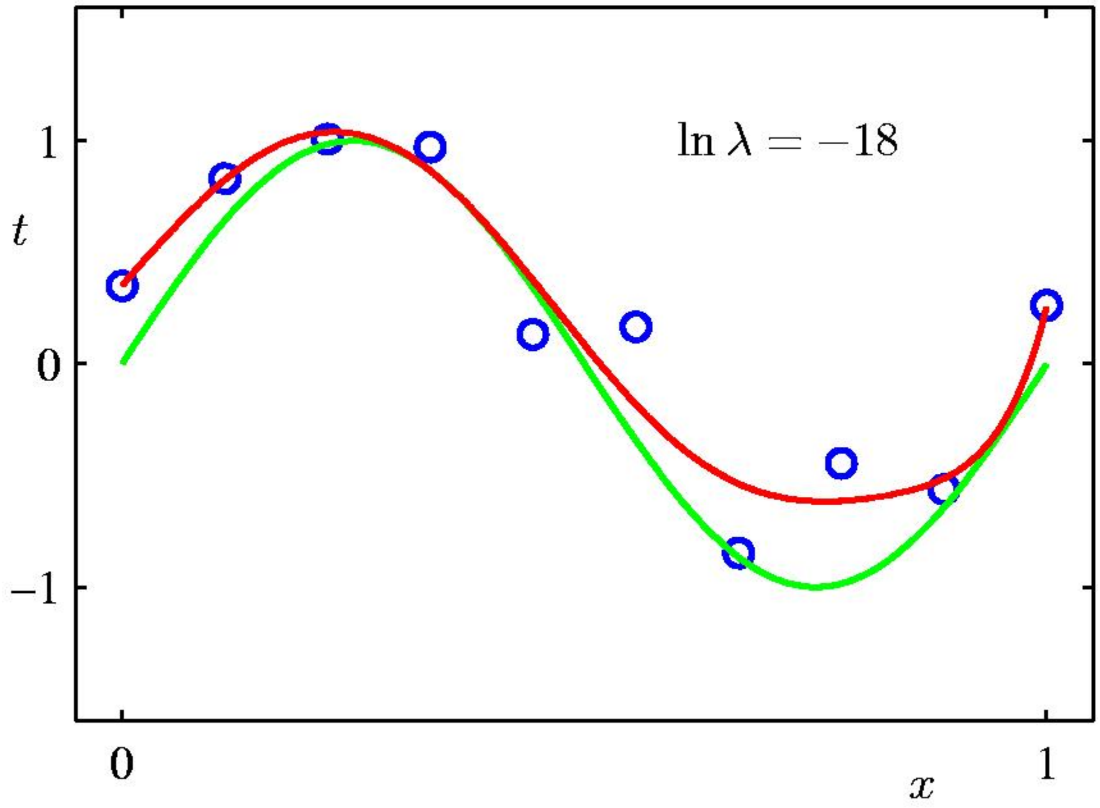


Regularization

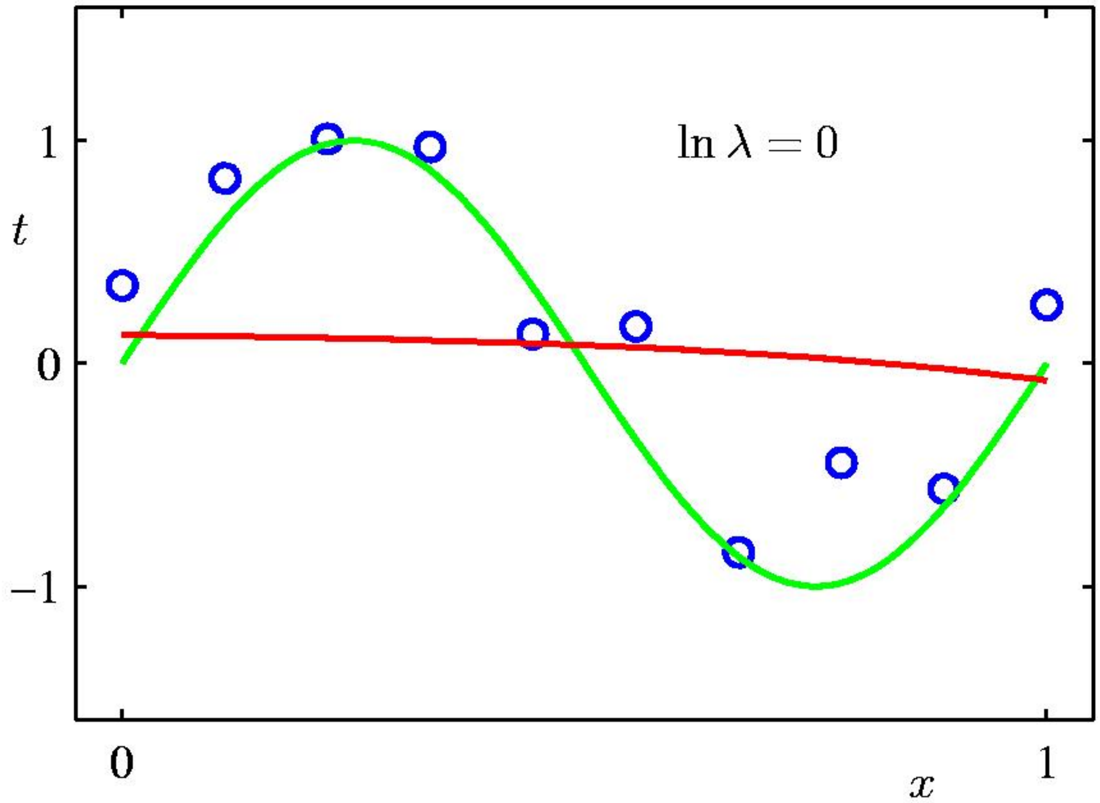
- Penalize large coefficient values

$$\widetilde{E}(\mathbf{w}) = \frac{1}{2}\sum_{n=1}^{N}\{y(x_n,\mathbf{w}) - t_n\}^2 + \frac{\lambda}{2}\|\mathbf{w}\|^2$$

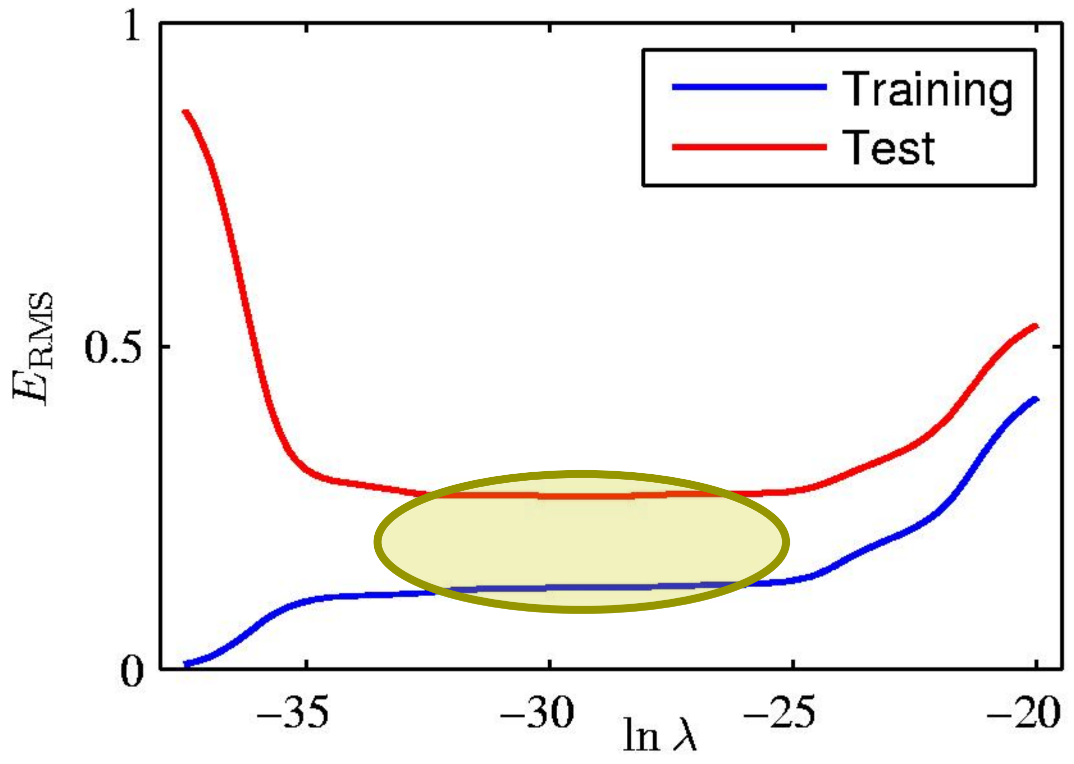
Regularization: $\ln \lambda = -18$



Regularization: $\ln \lambda = 0$



Regularization: E_{RMS} vs. $\ln\lambda$



What you need to know

- Point estimation:
  - Maximal Likelihood Estimation
  - Bayesian learning
  - Maximal a Posterior
- Gaussian estimation
- Regression
  - Basis function = features
  - Optimizing sum squared error
  - Relationship between regression and Gaussians
- Bias-Variance trade-off

Homework-01

- Python programming
  - 1-D regression
  - Learn TensorFlow

Question for Review

- Finish the “Gaussian parameters learning”
  - Please use google, ^_*

The End

新浪微博：@浙大张宏鑫

微信公众号：