BLG 527E Machine Learning

FALL 2021-2022

Assoc. Prof. Yusuf Yaslan & Assist. Prof. Ayşe Tosun

Parametric Methods

Parametric Estimation

- $\mathcal{X} = \{ x^t \}_t$ where $x^t \sim p\,(x)$
- Parametric estimation:

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

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

Maximum Likelihood Estimation

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

$$l\,(\theta|\mathrm{X}) = p\,(\mathrm{X}\,|\theta) = \prod_t p\,(x^t|\theta)$$

- Log likelihood

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

- Maximum likelihood estimator (MLE)

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

Examples: Bernoulli/Multinomial

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

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

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

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

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

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

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

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

Examples: Bernoulli (Derivation)

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

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

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

$$\frac{dL(p_0 \mid X)}{dp_0} = \sum_{t=1}^{N} x^t \frac{d}{dp_0}\log(p_0) + \sum_{t=1}^{N}(1-x^t)\frac{d}{dp_0}\log(1-p_0)$$

$$= \frac{1}{p_0}\sum_{t=1}^{N} x^t - \sum_{t=1}^{N}(1-x^t)\frac{1}{1-p_0} = 0$$

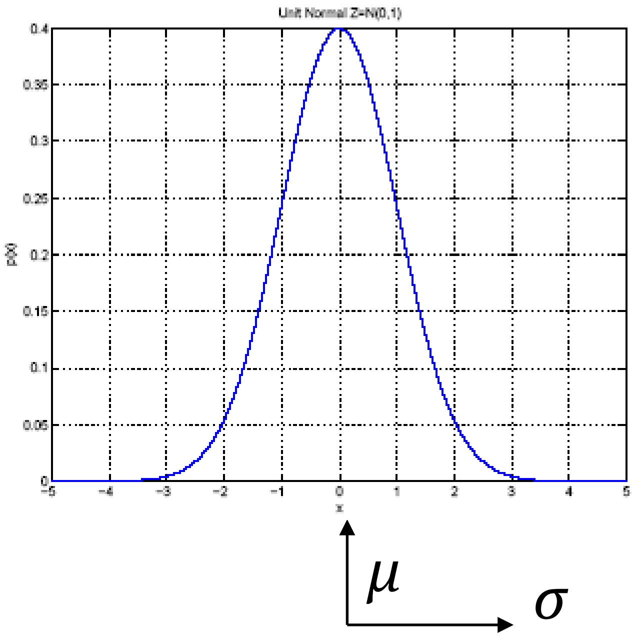
Bernoulli (Derivation)

$$= (1 - p_0)\sum_{t=1}^{N} x^t - p_0 \sum_{t=1}^{N} 1 + p_0 \sum_{t=1}^{N} x^t = 0$$

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

MLE: $p_o = \Sigma_t\, x^t / N$

Gaussian (Normal) Distribution



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

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

Gaussian (Normal) Distribution

- Given that $\mathcal{X} = \{ x^t \}_t$ with $x^t \sim \mathcal{N}(\mu, \sigma^2)$

$$L(\mu, \sigma |\ X) = -\frac{N}{2}\log(2\pi) - N log(\sigma) - \frac{\sum_{n=1}^{N}(x^t - \mu)^2}{2\sigma^2}$$

MLE for μ and σ^2:

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

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

Bias and Variance

Let X be a sample from a population specified up to a parameter θ

To evaluate the quality of this estimator we can measure how much it is different from θ
That is (d (X)-θ)^2

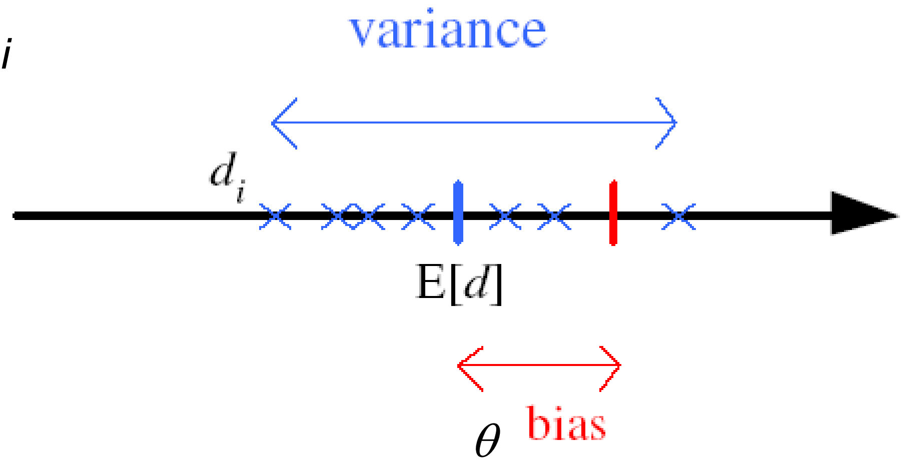
But since it is random variable (it depends on the sample) we need to average over all possible X and consider meas square error of the estimator

Remember the properties of expectation

Bias and Variance

Unknown parameter θ

Estimator $d_i = d\,(X_i)$ on sample X_i



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

Variance: $E\,[(d - E\,[d])^2]$

Mean square error:

$$r\,(d,\theta) = E\,[(d-\theta)^2] = E\,[(d - E[d] + E[d] - \theta)^2]$$

$$= (E[d]-\theta)^2 + E[(d - E[d])^2 + 2\,(d - E[d])(E[d]-\theta)]$$

Remember the properties of expectation

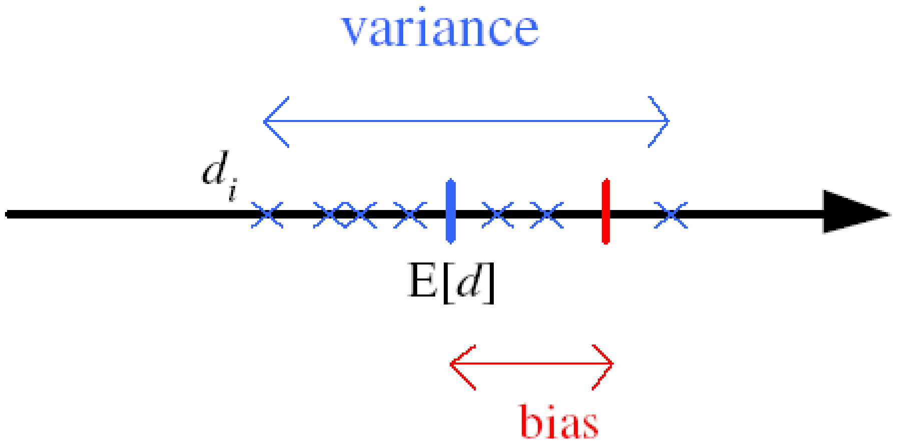
$$= E[(E[d]-\theta)^2]+E\,[(d–E\,[d])^2]\ +2\ E[(d–E[d])(E[d]-\theta)]$$

$$= E[(E[d]-\theta)^2]+E\,[(d–E\,[d])^2]\ +2\ (E[d]–E[d])(E[d]-\theta)$$

$$= (E\,[d] – \theta)^2 +E\,[(d–E\,[d]^{)2}]$$

$$= (E\,[d] – \theta)^2 + E\,[(d–E\,[d])^2]$$

$$= \text{Bias}^2 + \text{Variance}$$



Bayes' Estimator

- Sometimes before looking at a sample we may have some prior information on the possible value range that a parameter , θ, may take.
- This information is quite useful especially when the sample is small.
- Treat θ as a random variable with prior $p\ (\theta)$

- Bayes' rule: $p\ (\theta|X) = p(X|\theta)\ p(\theta)\ /\ p(X)$
- Density at x: $p(x|X) = \int p(x|\theta, X)\ p(\theta|X)\ d\theta = \int p(x|\theta)\ p(\theta|X)\ d\theta$

Bayes' Estimator

- Evaluating the $p(x|\mathrm{X})$ integrals may be quite difficult except in cases where the posterior has a nice form
- Maximum a Posteriori (MAP): $\theta_{\mathrm{MAP}} = \mathrm{argmax}_\theta\, p(\theta|\mathrm{X})$
- Maximum Likelihood (ML): $\theta_{\mathrm{ML}} = \mathrm{argmax}_\theta\, p(\mathrm{X}|\theta)$
- Bayes': $\theta_{\mathrm{Bayes'}} = \mathrm{E}[\theta|\mathrm{X}] = \int \theta\, p(\theta|\mathrm{X})\, d\theta$
- If we have no prior reason to favor some values of θ then the prior density is flat and the posterior will have the same form as the likelihood $p(\mathrm{X}|\theta)$

Bayes' Estimator: Example

- $x^t \sim N(\theta, \sigma_o^2)$ and $\theta \sim N(\mu, \sigma^2)$ where are $\mu, \sigma^2, \sigma_o^2$ known
- $\theta_{\mathrm{ML}} = m$
- $p(\theta|\mathrm{X}) \propto p(\mathrm{X}|\ \theta)p(\theta)$
- Take the derivative with respect to θ

$$P(X \mid \theta) = \frac{1}{\sqrt{2\pi}\sigma_0} \exp\left[-\frac{(x-\theta)^2}{2\sigma_0^2} \right]$$

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

Likelihood:

$$l(\theta) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left[-\frac{(\theta-\mu)^2}{2\sigma^2} \right] \prod_{t=1}^{N} \frac{1}{\sqrt{2\pi}\sigma_0} \exp\left[-\frac{(x^t-\theta)^2}{2\sigma_0^2} \right]$$

Loglikelihood:

$$L(\theta) = \log\left(\frac{1}{\sqrt{2\pi}\sigma} \right) + \left[-\frac{(\theta-\mu)^2}{2\sigma^2} \right] \sum_{t=1}^{N} \left[\log\left(\frac{1}{\sqrt{2\pi}\sigma_0} \right) + \left[-\frac{(x^t-\theta)^2}{2\sigma_0^2} \right] \right]$$

$$L(\theta) = \log\left(\frac{1}{\sqrt{2\pi}\sigma}\right) + \left[-\frac{(\theta-\mu)^2}{2\sigma^2}\right] + \sum_{t=1}^{N}\left[\log\left(\frac{1}{\sqrt{2\pi}\sigma_0}\right) + \left[-\frac{(x^t-\theta)^2}{2\sigma_0^2}\right]\right]$$

$$\frac{\partial L(\theta)}{\partial\theta} = -\frac{(\theta-\mu)}{\sigma^2} + \sum_{t=1}^{N}\frac{(x^t-\theta)}{\sigma_0^2} = 0$$

$$\sum_{t=1}^{N}\frac{x^t}{\sigma_0^2} - \sum_{t=1}^{N}\frac{\theta}{\sigma_0^2} - \frac{(\theta-\mu)}{\sigma^2} = \frac{N}{\sigma_0^2}\sum_{t=1}^{N}\frac{x^t}{N} - \sum_{t=1}^{N}\frac{\theta}{\sigma_0^2} - \frac{\theta}{\sigma^2} + \frac{\mu}{\sigma^2} =$$

$$\frac{N}{\sigma_0^2}m + \frac{\mu}{\sigma^2} = \frac{N\theta}{\sigma_0^2} + \frac{\theta}{\sigma^2}$$

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

Bayes' Estimator: Example

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

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

Bayesian Learning for Coin Model

- Bayesian Learning procedure:
- Given data $x^1, x^2, \ldots, x^N$ write down expression for likelihood $p(X|\theta)$

Specify a prior $P(\theta)$

Compute posterior $p(\theta|X) = p(X|\theta)p(\theta)/p(X)$

$$p(\theta|X) \propto p(X|\theta)p(\theta)$$

- This example is obtained from Nando Freitas lecture notes

Bayesian Learning for Coin Model

- For coin model likelihood of data (i.i.d. in our case)

$$P(x^1, x^2, \ldots, x^N | \theta) = \prod_t \theta^{x^t} (1 - \theta)^{(1 - x^t)} = \theta^m (1 - \theta)^{(N - m)}$$

Where $x^t \in \{0,1\}$ and m is the number of 1's

- Specify a prior on θ. For this we need to introduce Beta distribution

Beta Distribution

$$p(\theta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \theta^{\alpha-1}(1-\theta)^{\beta-1}$$

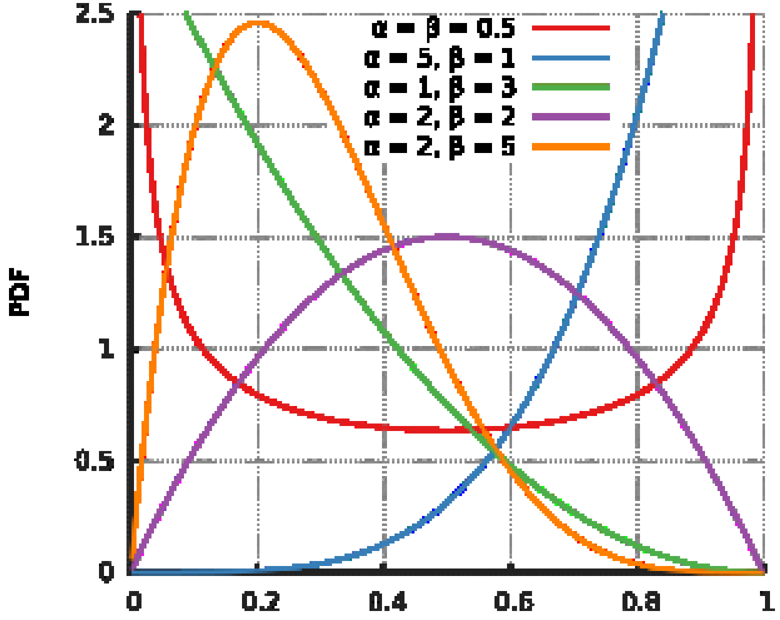
$$\Gamma(z) = \int_0^\infty e^x x^{z-1} dx$$

$$\int p(\theta) d\theta = 1$$

$$\int \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \theta^{\alpha-1}(1-\theta)^{\beta-1} d\theta = 1$$

$$\int \theta^{\alpha-1}(1-\theta)^{\beta-1} d\theta = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}$$

α, β are hyperparameters



$$E[\theta] = \frac{\alpha}{\alpha+\beta}$$

- The figure is obtained from wikipedia

Bayesian Learning for Coin Model

Compute Posterior: $p(\theta\,|X) \propto p(X|\,\theta)p(\theta)$

$P(x^1, x^2, \ldots, x^N|\,\theta\,) = \prod_t \theta^{\,x^t}(1-\theta)^{(1-x^t)} = \theta^{\,m}(1-\theta)^{(N-m)}$

$$p(\theta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\theta^{\alpha-1}(1-\theta)^{\beta-1} = \frac{1}{const}\theta^{\alpha-1}(1-\theta)^{\beta-1}$$

$$p(\theta\,|\,X) = \frac{1}{const}\theta^{\alpha-1}(1-\theta)^{\beta-1}\theta^{m}(1-\theta)^{N-m} = p(\theta\,|\,X) = \frac{\Gamma(\alpha'+\beta')}{\Gamma(\alpha')\Gamma(\beta')}\theta^{\alpha'-1}(1-\theta)^{\beta'-1}$$

$$\alpha' = m+\alpha, \beta' = N-m+\beta$$

Conjugate Prior

- Conjugate priors: A likelihood-prior pair is said to be conjugate if they result in a posterior which is of the same form as the prior.
- This enables us to compute the posterior density analytically without having to worry about computing the denominator in Bayes' rule, the marginal likelihood.

Prior	**Likelihood**
Gaussian	Gaussian
Beta	Binomial
Dirichlet	Multinomial
Gamma	Gaussian

Example

- Suppose that we observe X= {1,1,1,1,1,1} where each x^t comes from Bernoulli distribution $\theta_{ML} = 1$
- We can compute posterior and use its mean as the estimate

$$p(\theta \mid X) = \frac{1}{const}\theta^{\alpha-1}(1-\theta)^{\beta-1}\theta^{m}(1-\theta)^{N-m} = p(\theta \mid X) = \frac{\Gamma(\alpha'+\beta')}{\Gamma(\alpha')\Gamma(\beta')}\theta^{\alpha'-1}(1-\theta)^{\beta'-1} \qquad E[\theta] = \frac{\alpha'}{\alpha'+\beta'}$$

$$\alpha' = m + \alpha, \beta' = N - m + \beta$$

- Using Beta(2,2) prior $\theta_B = \frac{8}{10}$

Parametric Classification

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

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 $\mathcal{X} = \{x^t, r^t\}_{t=1}^N$

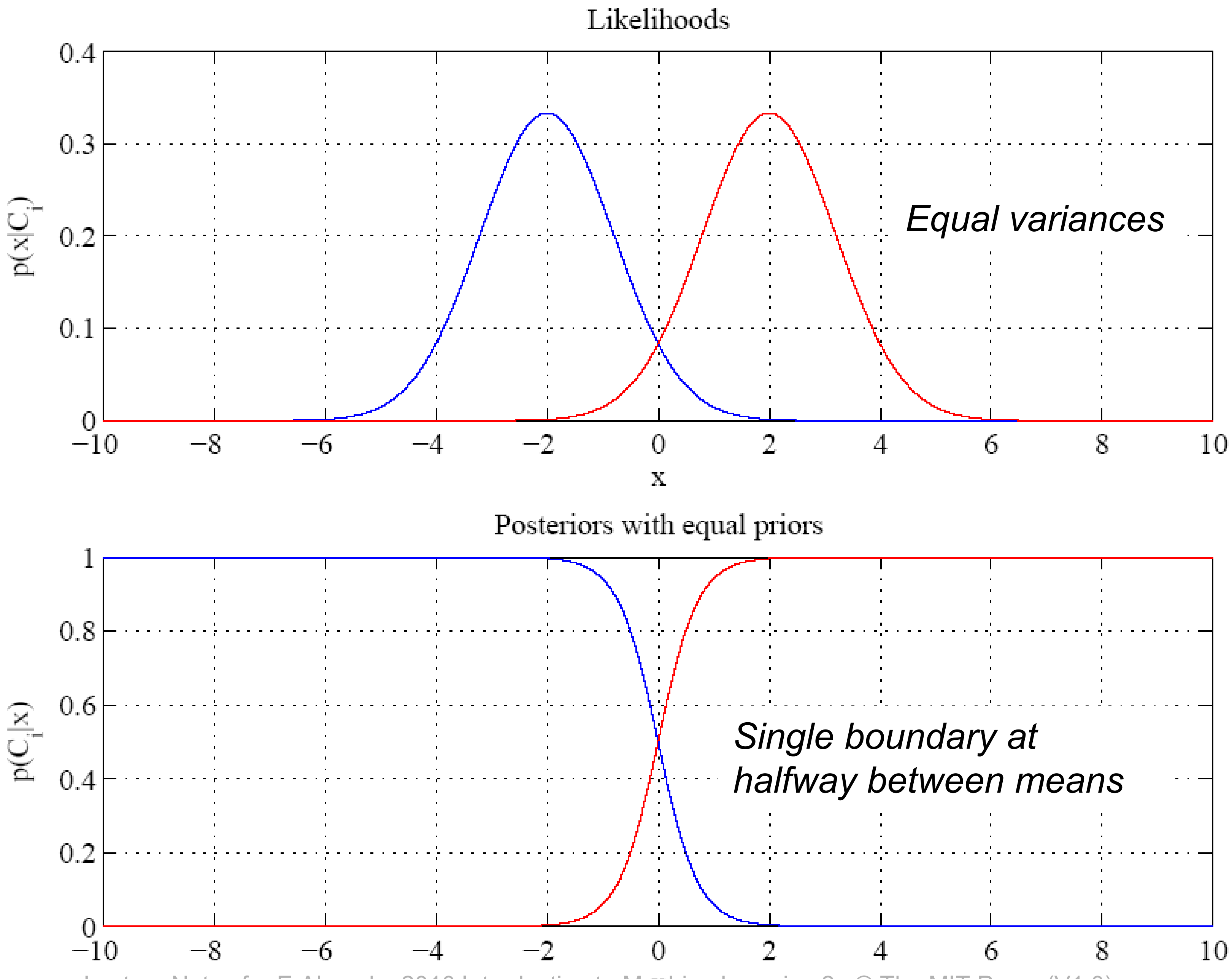
$$x \in \Re \qquad r_i^t = \begin{cases} 1 \text{ if } x^t \in C_i \\ 0 \text{ if } x^t \in C_j,\ j \neq i \end{cases}$$

- 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 becomes

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

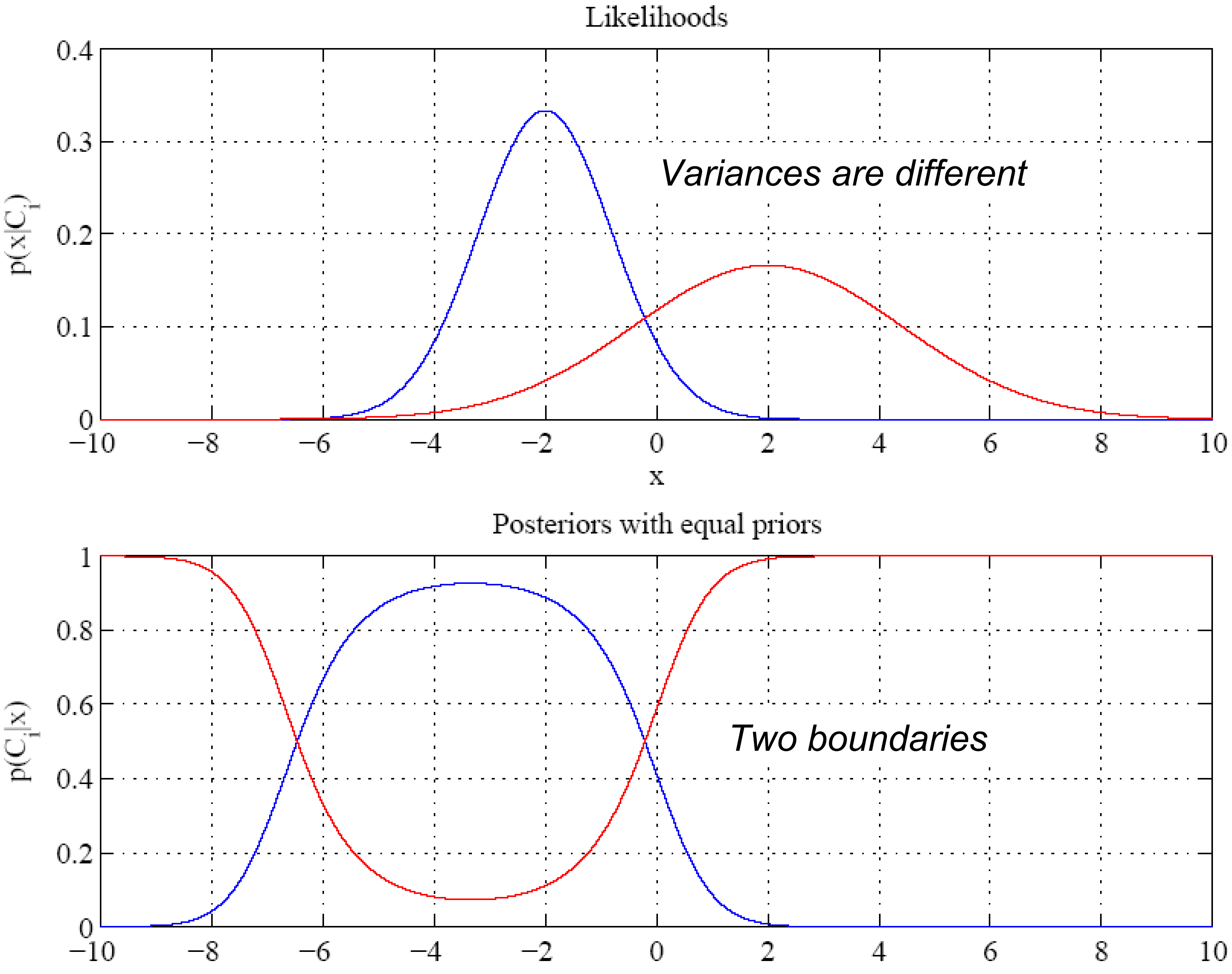


Likelihoods

Equal variances

Posteriors with equal priors

Single boundary at halfway between means



Variances are different

Two boundaries

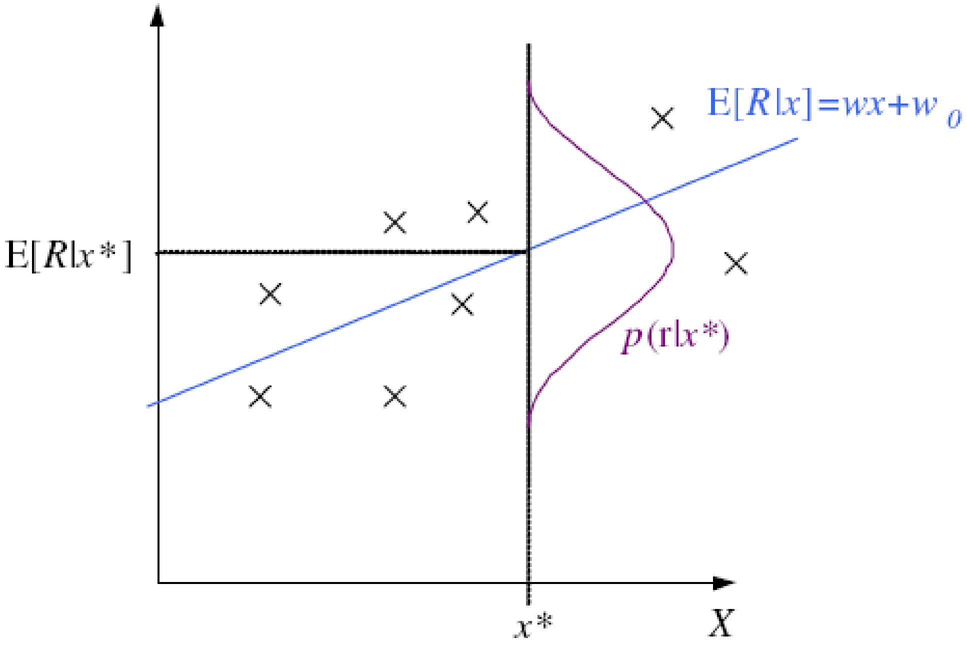
Probabilistic Interpretation of Linear Regression

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

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

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

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



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

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

Regression: From LogL to Error

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

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

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

Linear Regression

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

Take derivative of E

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

…wrto w0

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

…wrto w1

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

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

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

Polynomial Regression

$$g\left(x^t \mid w_k,\ldots,w_2,w_1,w_0\right)=w_k\left(x^t\right)^k+\cdots+w_2\left(x^t\right)^2+w_1x^t+w_0$$

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

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

Other Error Measures

- Square Error:

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

- Relative Square Error:

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

- Absolute Error: $E\ (\vartheta \mid X) = \sum_t |r^t - g(x^t \mid \vartheta)|$
- ε-sensitive Error:

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

Bias and Variance

$$E\left[(r-g(x))^2 \mid x\right] = \underset{\textit{noise}}{E\left[(r-E[r \mid x])^2 \mid x\right]} + \underset{\textit{squared error}}{(E[r \mid x]-g(x))^2}$$

$$E_X\left[(E[r \mid x]-g(x))^2 \mid x\right] = \underset{\textit{bias}}{(E[r \mid x]-E_X[g(x)])^2} + \underset{\textit{variance}}{E_X\left[(g(x)-E_X[g(x)])^2\right]}$$

Estimating Bias and Variance

- *M* samples X$_i$={x^t_i , r^t_i}, *i*=1,...,*M* are used to fit $g_i(x)$, *i* =1,...,*M* and *t*=1,…,N

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

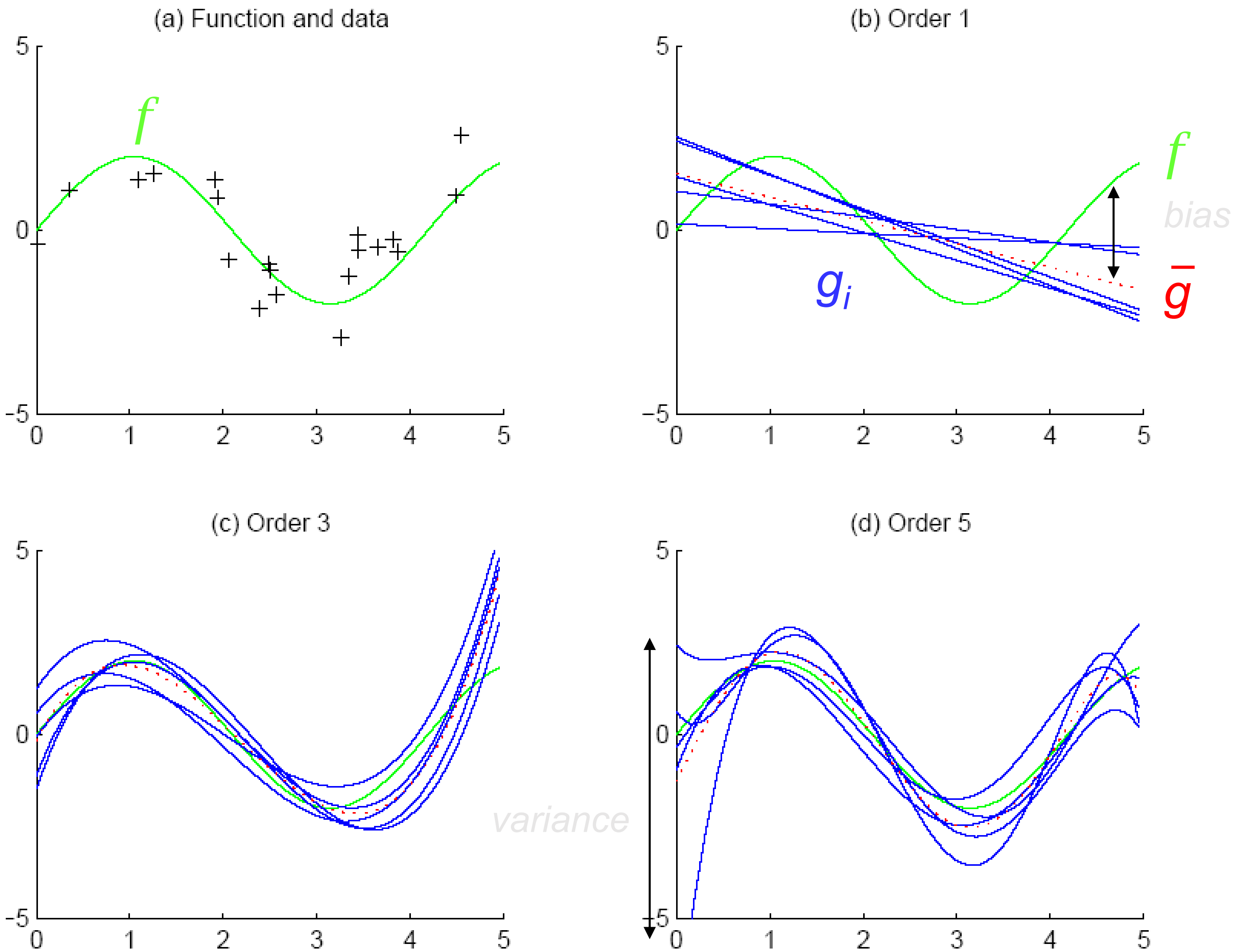
$$\text{Variance}(g) = \frac{1}{NM}\sum_t \sum_i \left[g_i(x^t) - \overline{g}(x^t)\right]^2$$

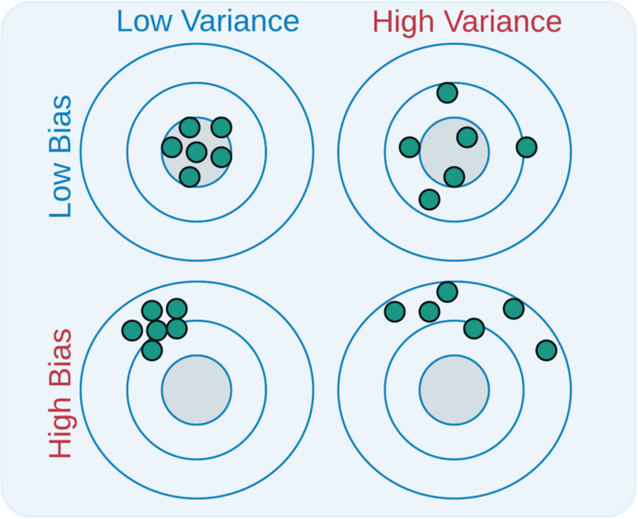
$$\overline{g}(x) = \frac{1}{M}\sum_i g_i(x)$$

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

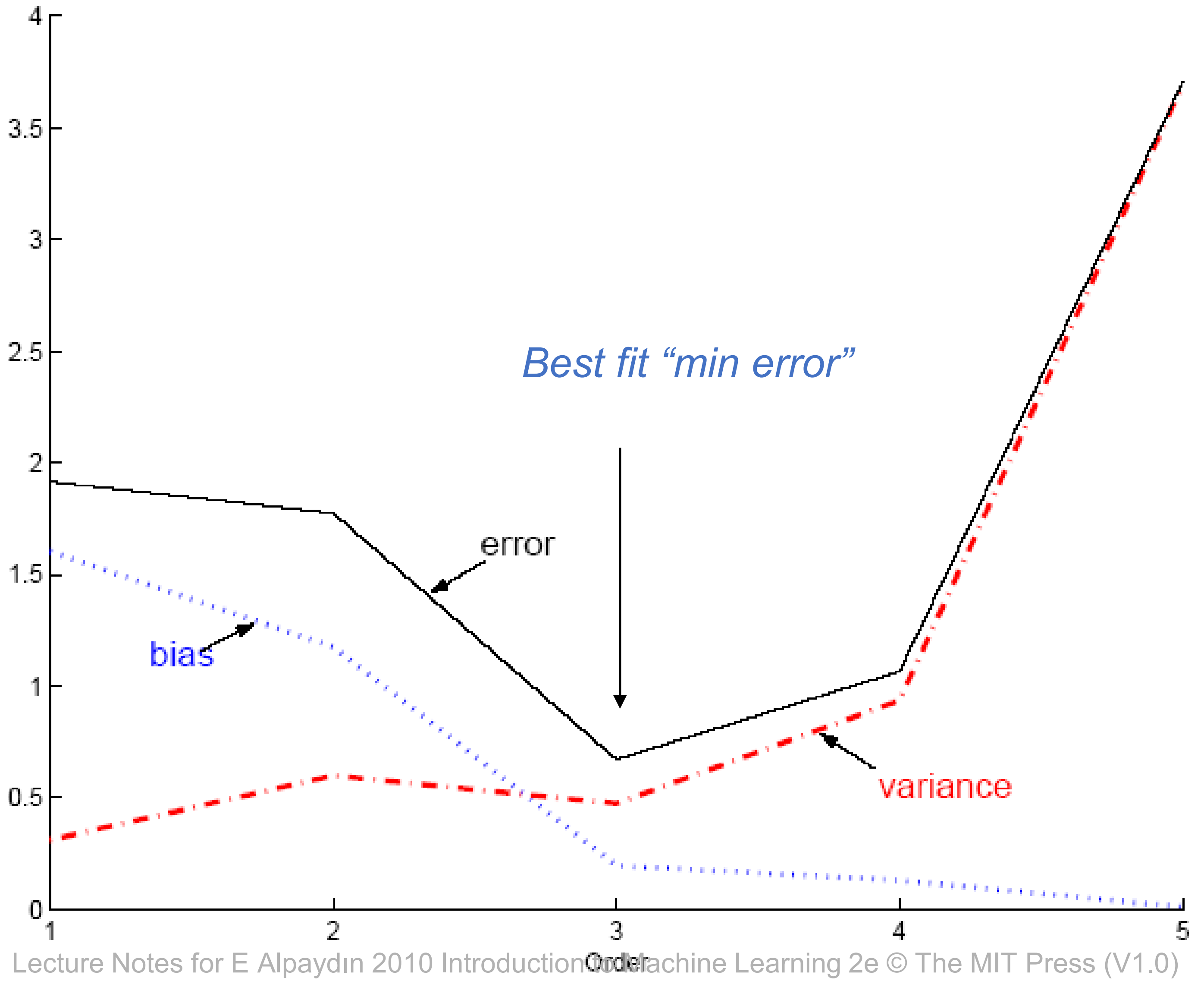
- As we increase complexity,
  bias decreases (a better fit to data) and
  variance increases (fit varies more with data)
- Bias/Variance dilemma: (Geman et al., 1992)



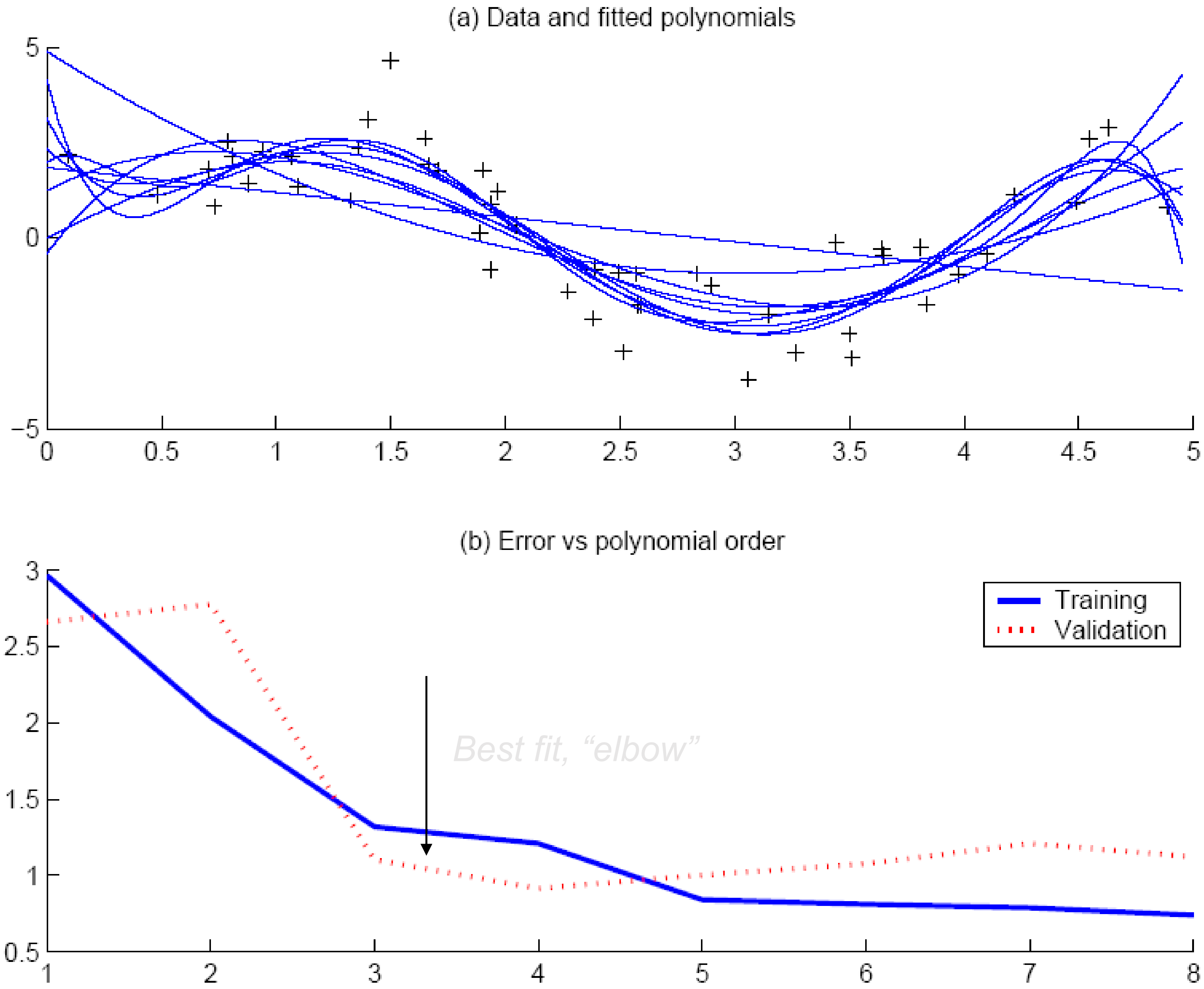


Bias Variance Image is obtained from: https://www.researchgate.net/figure/Visualizing-bias-and-variance-tradeoff-using-a-bulls-eye-diagram_fig3_318432363

Polynomial Regression



Best fit "min error"

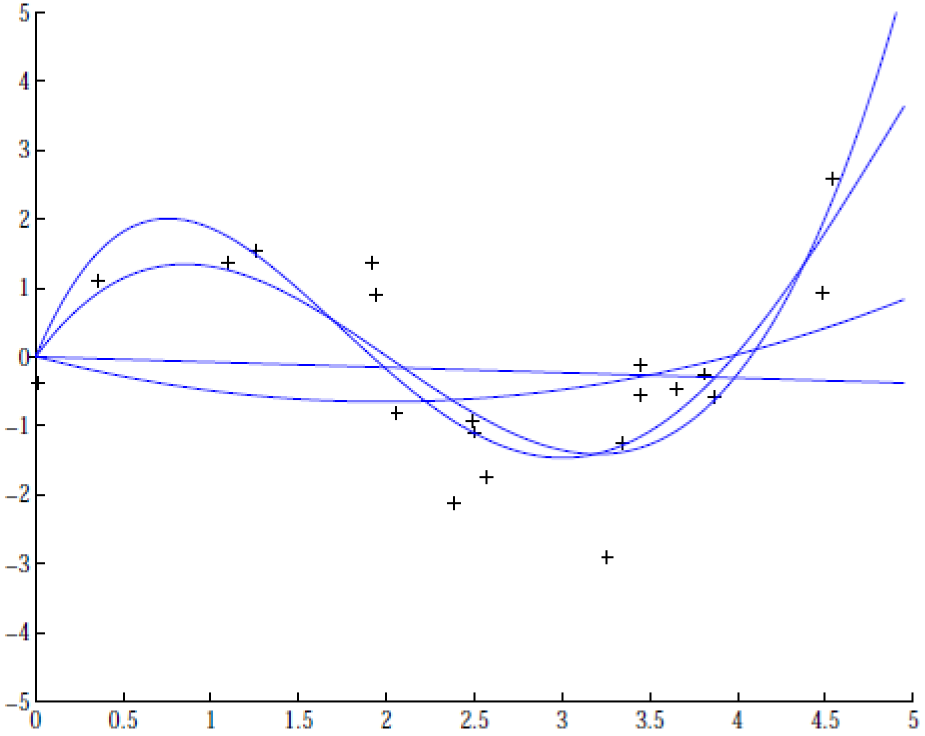


(a) Data and fitted polynomials

(b) Error vs polynomial order

Best fit, “elbow”

Regression example



Coefficients increase in magnitude as order increases:

1: [-0.0769, 0.0016]
2: [0.1682, -0.6657, 0.0080]
3: [0.4238, -2.5778, 3.4675, -0.0002
4: [-0.1093, 1.4356, -5.5007, 6.0454, -0.0019]

Idea: Penalize large coefficients

Regularization

- New Cost Function $E(\mathbf{w} \mid \mathcal{X}) = \frac{1}{2}\sum_{t=1}^{N}\left[y^t - g\left(x^t \mid \mathbf{w}\right)\right]^2 + \lambda \sum_i w_i^2$

- Ridge Regression $R(w) = \|w\|^2 = \sum_i w_i^2$

- LASSO: $R(w) = \|w\|_1 = \sum_i |w_i|$

$$\mathcal{L}(W) = \frac{1}{2}\sum_{i=1}^{N}(y - Xw)^2 + \lambda \sum_i w_i^2 \Rightarrow \frac{1}{2}(y - X\mathrm{w})^T(y - X\mathrm{w}) + \lambda w^T w$$

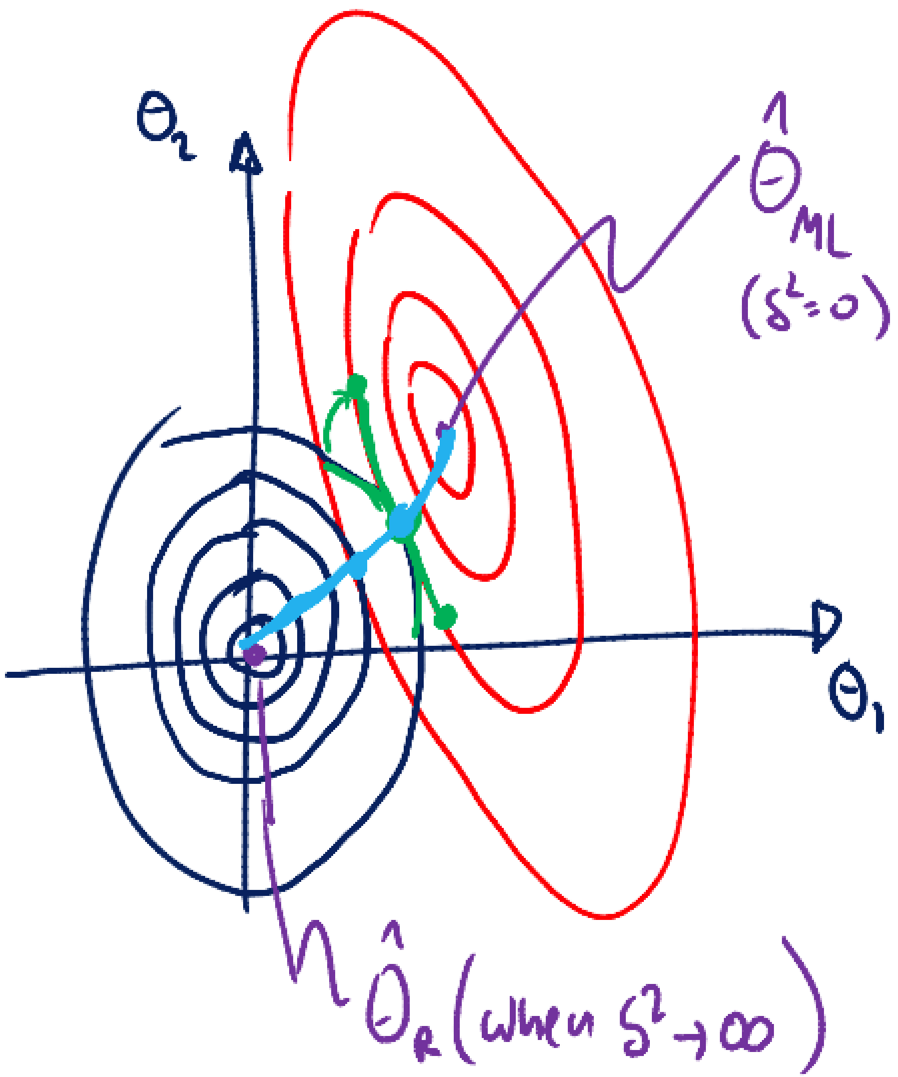
- $\nabla\ \mathcal{L}$ = - $\frac{2}{2}\ X^T(y - Xw)$+λw

- - $\frac{2}{2}\ X^T(y - Xw)$ +λw = 0 → $X^Ty = X^TX$w +λw →

- $X^Ty = (X^TX$ +λI) w

- $\widehat{w}$= $(X^TX$+λI$)^{-1}X^Ty$

$$J(\boldsymbol{\theta}) = \overset{\text{ellipses}}{\underline{(\mathbf{y} - \mathbf{X}\boldsymbol{\theta})^T(\mathbf{y} - \mathbf{X}\boldsymbol{\theta})}} + \delta^2\underline{\boldsymbol{\theta}^T\boldsymbol{\theta}}$$


θ_2
θ_1
$\hat{\theta}_{ML}$ $(\delta^2 = 0)$
$\hat{\theta}_R$ (when $\delta^2 \to \infty$)

- Image is obtained from Nando Freitas’ lecture notes