Linear regression

PRML 3, MLPP 7, UML 9

Domain set $\mathcal{X} = \mathbb{R}^D$

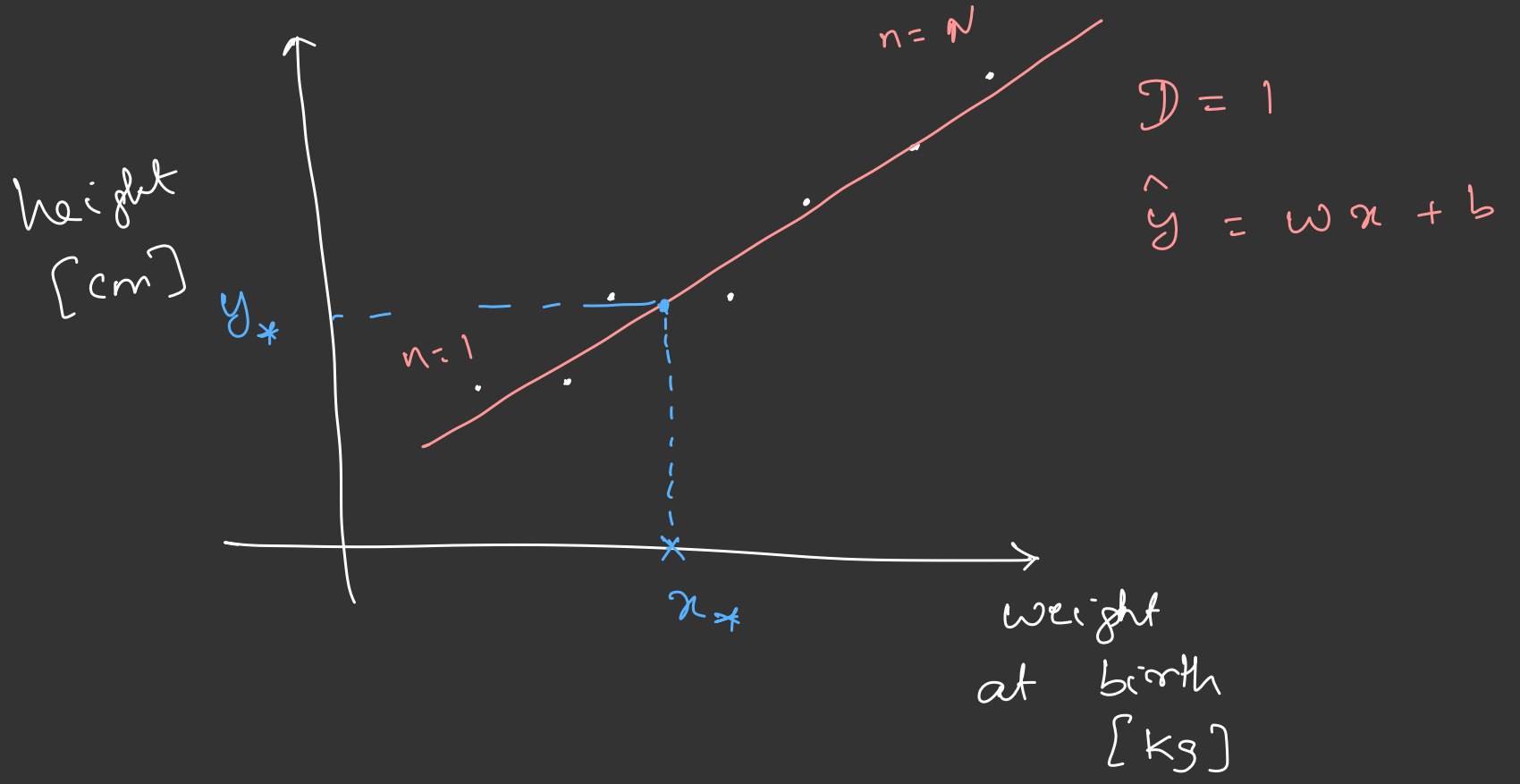
Label set $\mathcal{Y} = \mathbb{R}$

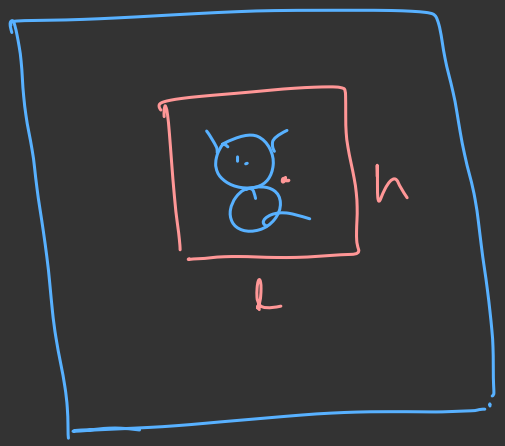
Goal: Learn a linear function $h : \mathbb{R}^D \to \mathbb{R}$ that best approximates the relationship between our variables.

$$\mathcal{D} = \{(\underline{x}_i, y_i)\}_{i=1}^{N}$$

Hypothesis class of linear regression predictors is simply a linear function:

$$H_{reg} = L_D = \{ \underline{x} \longmapsto \langle \underline{w}, \underline{x} \rangle + b : \underline{w} \in \mathbb{R}^D, b \in \mathbb{R} \}$$





$$\underline{y} = \begin{bmatrix} L \\ h \\ p_x \\ p_y \end{bmatrix} \in \mathbb{R}^4$$

$$\{\underline{x}_i, \underline{y}_i\}$$

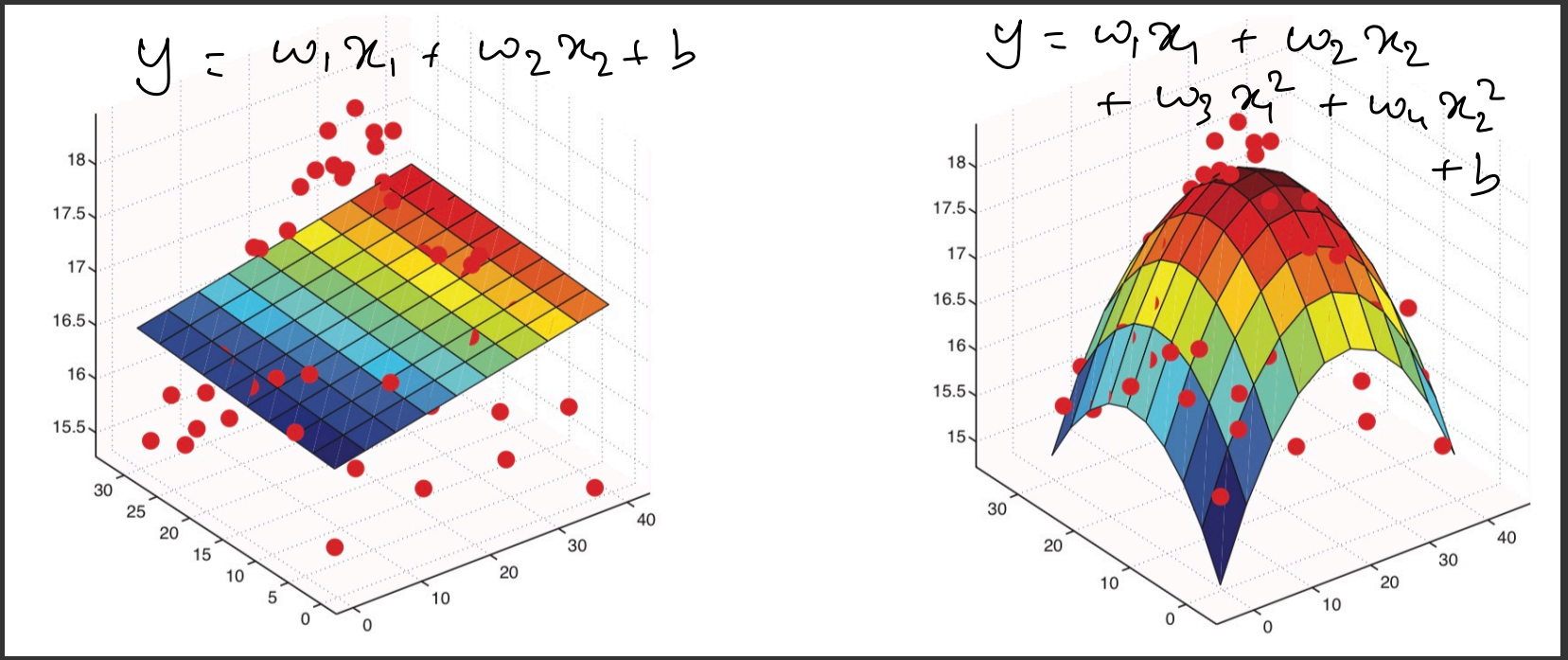
$$\underline{y}_i \approx h(x_i)$$

Linear regression:

$$h(\underline{x}) = \underline{w}^T \underline{x}$$

$$= w_1 x_1 + w_2 x_2 + \cdots + w_D x_D$$

Basis function models:



Linear regression can be used to model non-linear relationships by replacing $\underline{x}$ with $\phi(\underline{x})$

$$h(\underline{x}) = \underline{w}^T \phi(\underline{x})$$

$$\phi(\underline{x}) = [1, x, x^2, \ldots x^D]$$

Loss functions :-

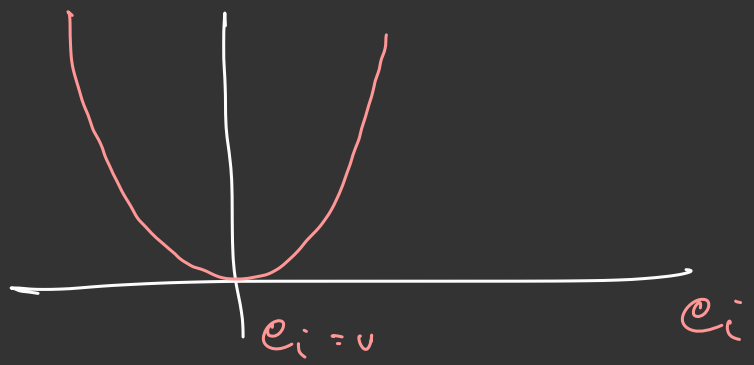
Discrepancy between $h(\underline{x})$ and y

$$l(h(\underline{x}), y) = h(x) - y$$

ERM:

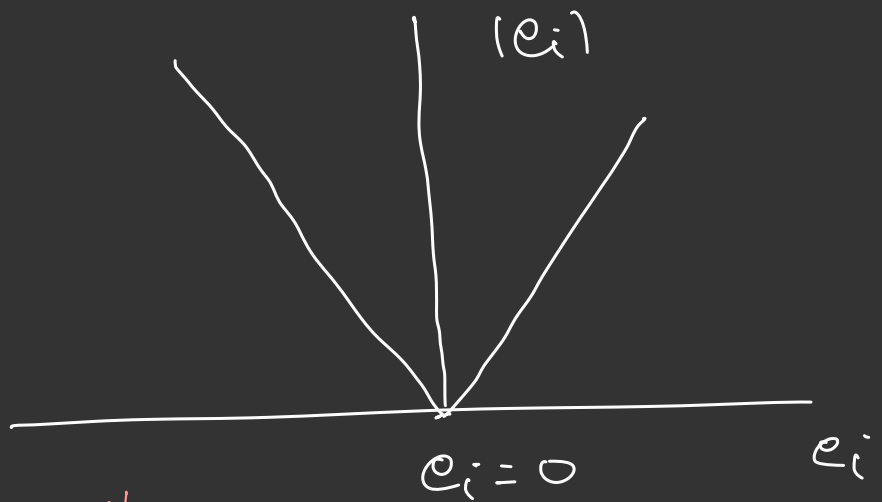
(A) Mean squared error:

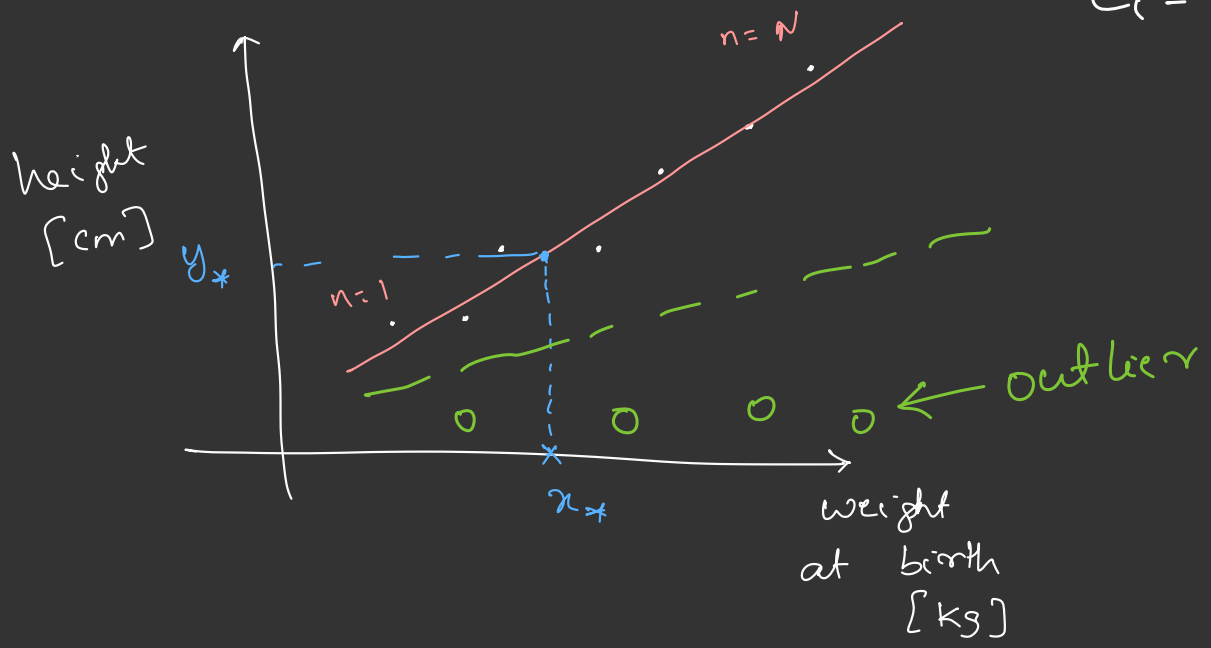
$$L_D(\underline{w}) = \frac{1}{N} \sum_{i=1}^{N} \underbrace{(h(\underline{x}_i) - y_i)}_{e_i}^2$$



(B) Mean absolute error :

$$L_D(\underline{w}) := \frac{1}{N} \sum_{i=1}^{N} |h(\underline{x}_i) - y_i|$$





Least Squares:

$$\arg\min_{\underline{w}} \quad L_D(\underline{w}) = \frac{1}{N} \sum_{i=1}^{N} (\underline{w}^T \underline{x}_i - y_i)^2$$

$$\frac{\partial L_D}{\partial \underline{w}} = 0$$

$$\frac{2}{N} \sum_{i=1}^{N} \underline{x}_i (\underline{w}^T \underline{x}_i - y_i) = 0$$

$$\left(\sum_{i=1}^{N} \underline{x}_i \underline{x}_i^T \right) \underline{w} = \sum_{i=1}^{N} y_i \underline{x}_i$$

$$w = \left(\sum_{i=1}^{N} \underline{x}_i \underline{x}_i^T \right)^{-1} \sum_{i=1}^{N} y_i \underline{x}_i$$

$$\underline{w} = (X^T X)^{-1} X^T \underline{y}$$

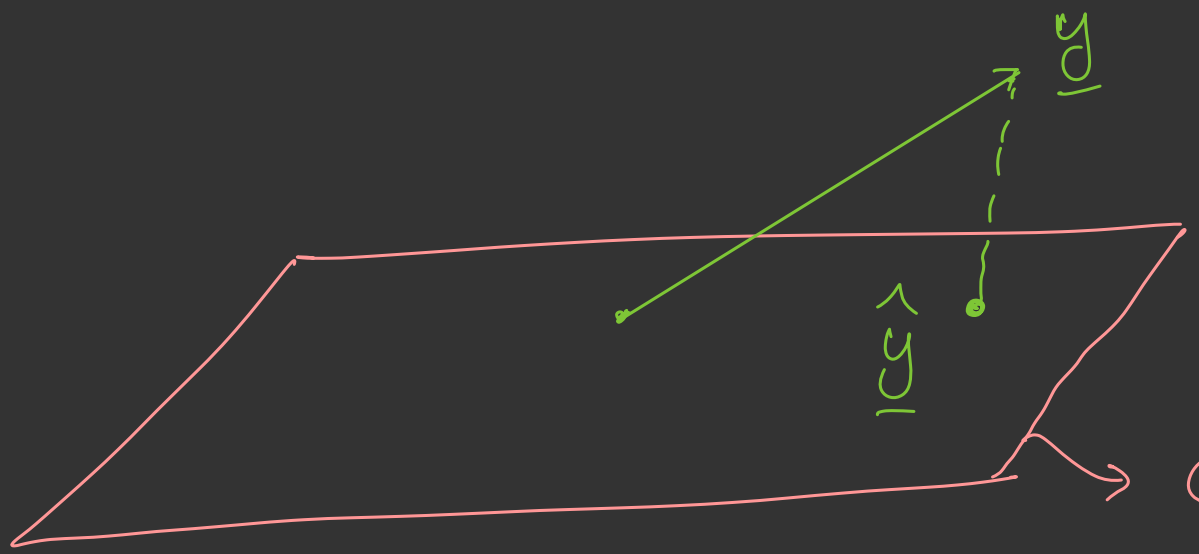
$$X^T = [\underline{x}_1, \underline{x}_2, \ldots, \underline{x}_N] \quad : D \times N$$

$$\underline{y} = [y_1, y_2, \ldots, y_N]^T \quad : N \times 1$$

$$X^\dagger = (X^T X)^{-1} X^T$$

$$X^\dagger X = (X^T X)^{-1} (X^T X) = I$$

$$X X^\dagger = P_X$$



$$\hat{\underline{y}} = P_X \underline{y}$$
$$= X(X^TX)^{-1}X^T\underline{y}$$

$$\underline{e} = \underline{y} - \hat{\underline{y}} = [I - XX^+]\underline{y}$$

$X \perp \underline{e}$

Probabilistic discriminative model:

$$y = h(x) + e$$

$$e \sim N(0, \sigma^2)$$

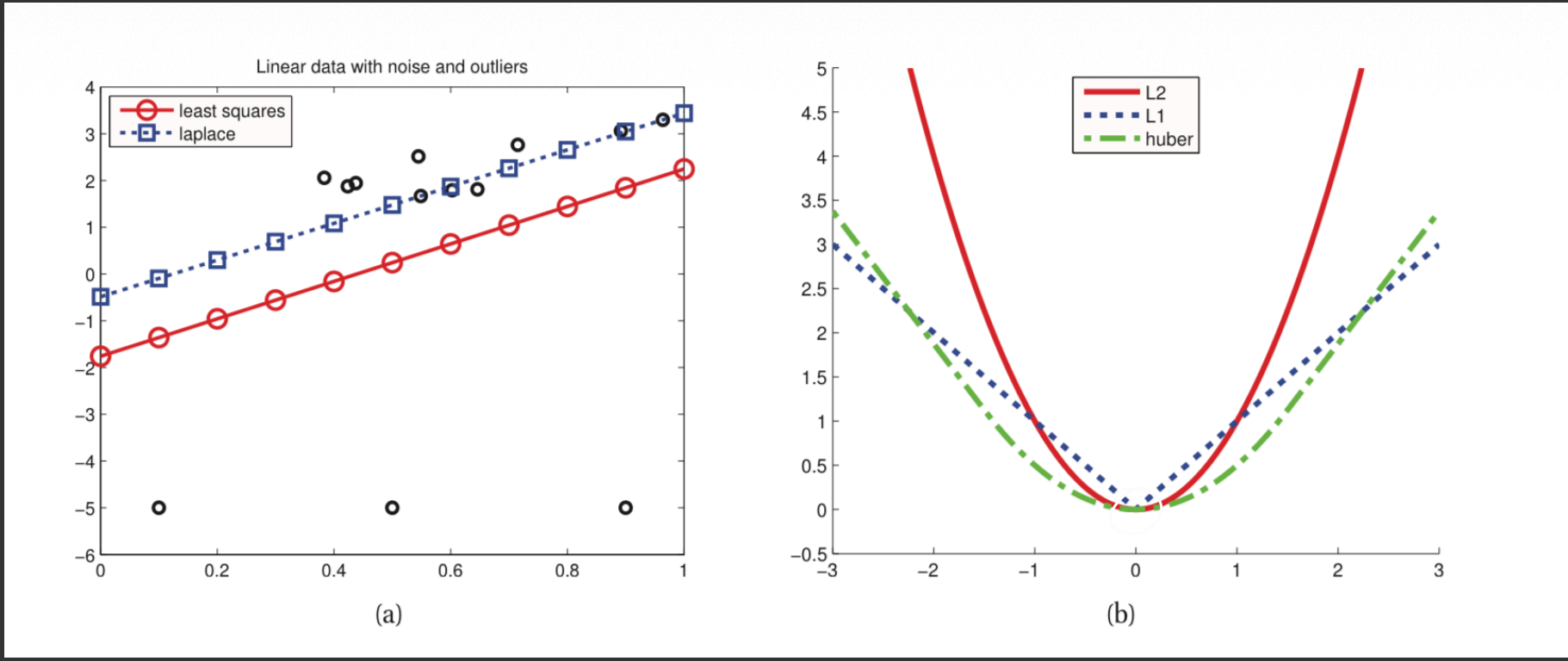
$$p(\underline{y} \mid \underline{x}; \underline{w}) = N(\underline{w}^T \underline{x}, \sigma^2)$$

Log-likelihood function:

$$l(\underline{w}) = \sum_{i=1}^{N} \log \left(\frac{1}{2\pi\sigma^2} \right)^{\frac{1}{2}} \exp\left(\frac{-1}{2\sigma^2} (y_i - \underline{w}^T x_i)^2 \right)$$

$$= -\frac{1}{2\sigma^2} \sum_{i=1}^{N} (y_i - \underline{w}^T x_i)^2 - \frac{N}{2} \log(2\pi\sigma^2)$$

Robust linear regression :



To account for outliers use distributions with heavy tail :

$$p(\underline{y} \mid \underline{x}; \underline{w}, b) = \text{Lap}(\underline{y} \mid \underline{w}^T \underline{x}, b)$$

$$\propto \exp\left(\frac{-1}{b} \mid y - \underline{w}^T \underline{x} \mid\right)$$

Assume b is fixed.

$$(P): \min_{\underline{\omega}} \quad L(\omega) = \sum_{i=1}^{N} |r_i(\underline{\omega})|$$

$$r_i = r_i^+ - r_i^- \qquad r_i^+ \geq 0, \quad r_i^- \geq 0$$

$$|r_i| = r_i^+ + r_i^-$$

$$\underline{\omega}^T x_i + r_i^+ - r_i^- = y_i$$

$\equiv (P)$

Linear program:

$$\min_{\underline{\omega}, \underline{r}^+, \underline{r}^-} \quad \sum_{i=1}^{N} r_i^+ + r_i^-$$

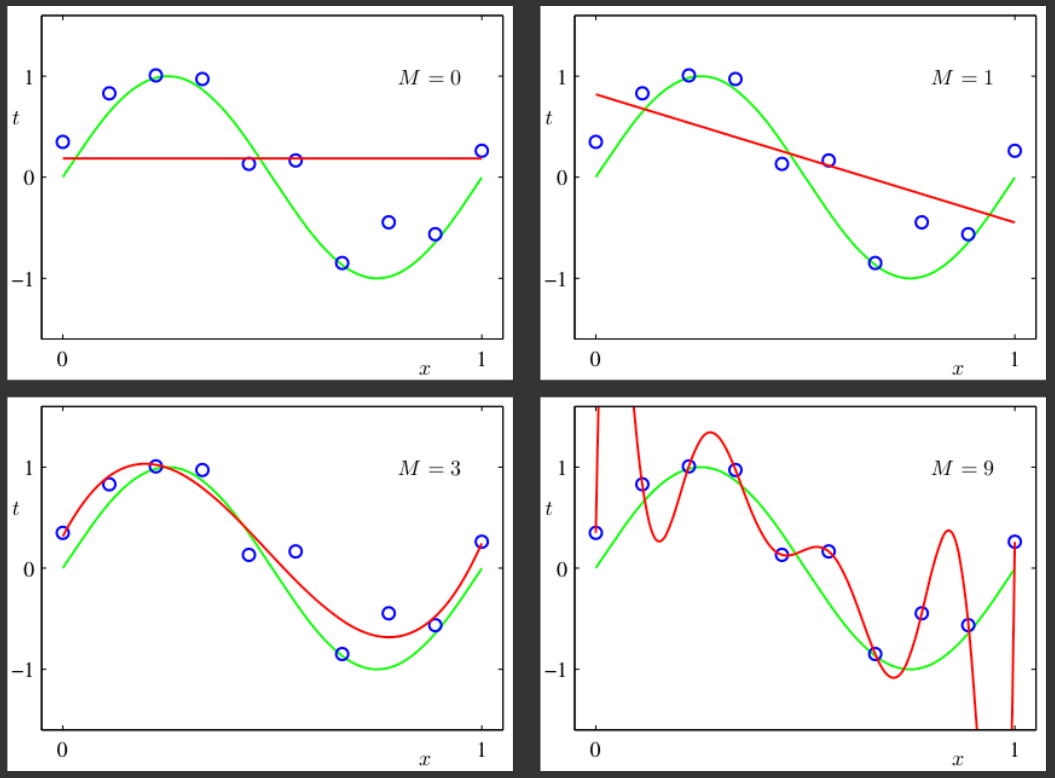
$$\text{s.t.} \quad r_i^+ \geq 0$$

$$r_i^- \geq 0$$

$$\underline{\omega}^T \underline{x}_i + r_i^+ - r_i^- = y_i \quad , \; i = 1, \ldots, N$$

Ridge regression:

$$h_w(x) = \sum_{j=0}^{M} w_j x^j$$



Add a regularizer to the error to control over-fitting:

$$E_D(\underline{w}) + \lambda E_w(\underline{w})$$

$E_D(\underline{w})$: data-dependent

$E_w(\underline{w})$: regularizer

l_2 - norm regularizer:

$$E_w(\underline{w}) = \frac{1}{2}\,\underline{w}^T\underline{w} = \frac{1}{2}\,\|\underline{w}\|_2^2$$

Loss:

$$L(\underline{w}) = \frac{1}{2}\sum_{i=1}^{N}(y_i - \underline{w}^T\underline{x}_i)^2 + \frac{\lambda}{2}\,\underline{w}^T\underline{w}$$

$$\text{minimize}_{\underline{w}} \quad L(w)$$

Ridge regression
or
weight decay
or
Parameter shrinkage

$\equiv$

$$\text{minimize}_{\underline{w}} \quad E_D(\underline{w})$$

$$\text{s.t.} \quad \|\underline{w}\|_2^2 \leq \alpha$$

Sol:

$$\frac{\partial L}{\partial \underline{w}} = 0$$

$$\frac{1}{2} \cdot 2 \sum_{i=1}^{N} (-\underline{x}_i)(y_i - \underline{w}^T \underline{x}_i) + \frac{\lambda}{2} \cdot 2\underline{w} = 0$$

$$\left(\sum_{i=1}^{N} \underline{x}_i \, \underline{x}_i^T + \lambda I \right) \underline{w} = \sum_{i=1}^{N} y_i \, \underline{x}_i$$

$$\boxed{\underline{w}_{ridge} = (\lambda I + X^T X)^{-1} X^T \underline{y}}$$

Bayesian view (Maximum aposteriori Estimate)

$$P(\underline{w} \mid \underline{x}, \underline{y}) \propto P(\underline{y} \mid \underline{x}, \underline{w}) \, P(\underline{w})$$

Suppose

$$P(\underline{y} \mid \underline{x}, \underline{w}, \beta) = \prod_{n=1}^{N} \mathcal{N}(y_n \mid \underline{w}^T \underline{x}_n, \beta^{-1})$$

$$P(\underline{w} \mid \alpha) = \mathcal{N}(\underline{w} \mid 0, \alpha^{-1} I)$$

$$= \left(\frac{\alpha}{2\pi}\right)^{\frac{(D+1)}{2}} \exp\left(-\frac{\alpha}{2} \underline{w}^T \underline{w}\right)$$

MAP: $\max_{\underline{w}} \; P(\underline{w} \mid \underline{x}, \underline{y}, \alpha, \beta)$

$$= \min_{\underline{w}} \; \frac{\beta}{2} \sum_{i=1}^{N} (y_i - \underline{w}^T x_i)^2 + \frac{\alpha}{2} \underline{w}^T \underline{w}$$

$$\lambda = \alpha/\beta$$

Bias – variance trade-off:

(model complexity)

How to determine λ ?

$\lambda = 0$ over fitting

$\lambda \uparrow$ under fit.

$$y = f(\underline{x}) + \varepsilon$$

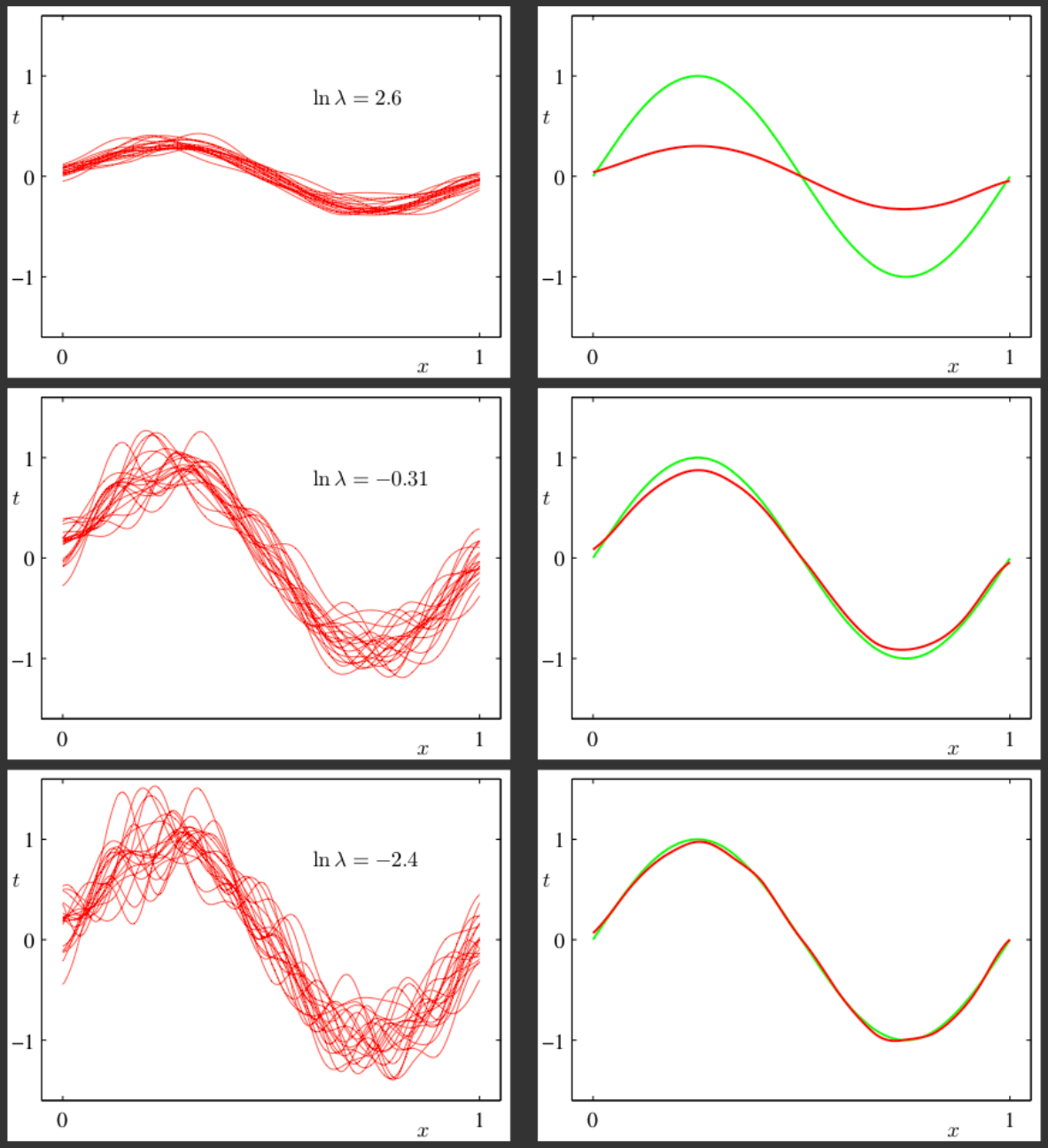
$$\text{var}(y) = \sigma^2$$

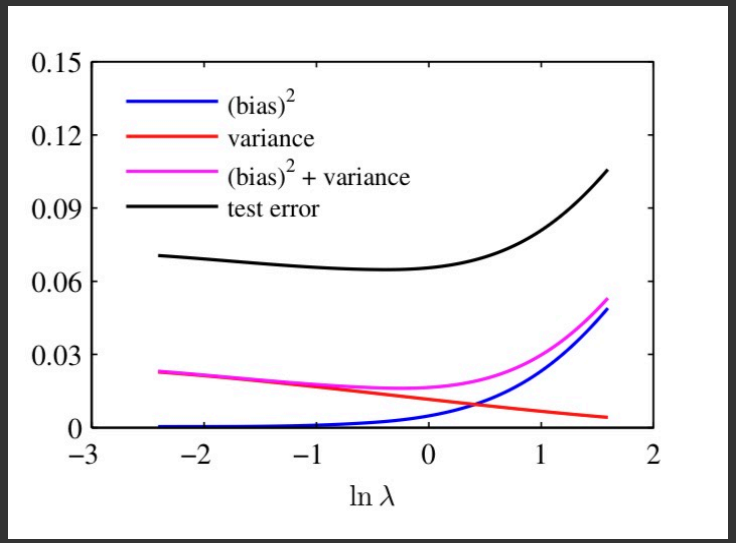
$$E\left[(y - \hat{f})^2\right] = E\left[(f(x) + \varepsilon - \hat{f})^2\right]$$

$$= E[y^2] + E[\hat{f}^2] - E[2y\hat{f}]$$

$$= \text{var}[y] + E[y]^2 + \text{var}[\hat{f}] + E[\hat{f}]^2 - 2fE[\hat{f}]$$

$$= \underbrace{\sigma^2}_{\text{irreducible error}} + \underbrace{\text{var}(\hat{f})}_{\text{variance}} + \underbrace{(f - E[\hat{f}])^2}_{(\text{bias})^2}$$

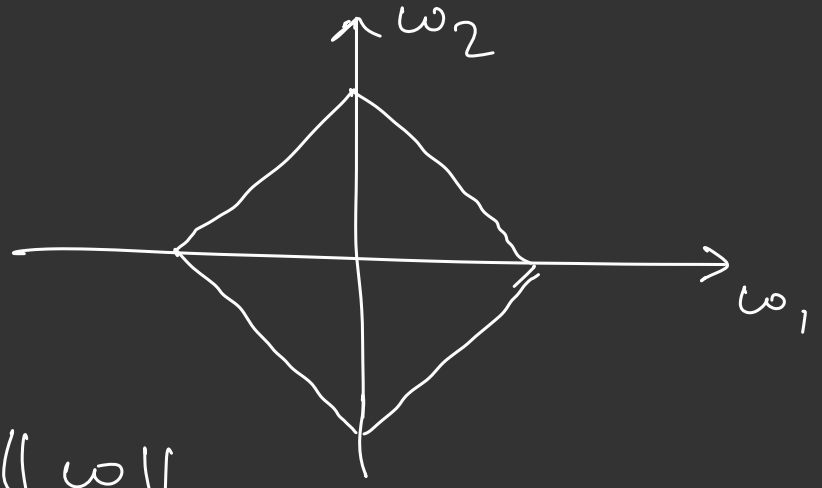

ln λ = 2.6
ln λ = −0.31
ln λ = −2.4
t
x
0
1
−1


0.15
0.12
0.09
0.06
0.03
0
−3
−2
−1
0
1
2
ln λ
(bias)²
variance
(bias)² + variance
test error

Other regularizers:

$$E_w(\underline{w}) = \frac{\lambda}{2} \sum_{j=1}^{D} |w_j|^q$$

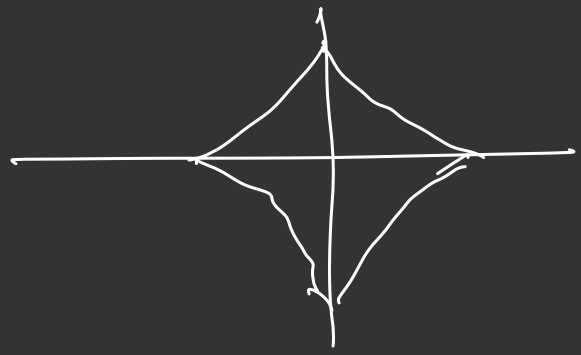
$q = 1$

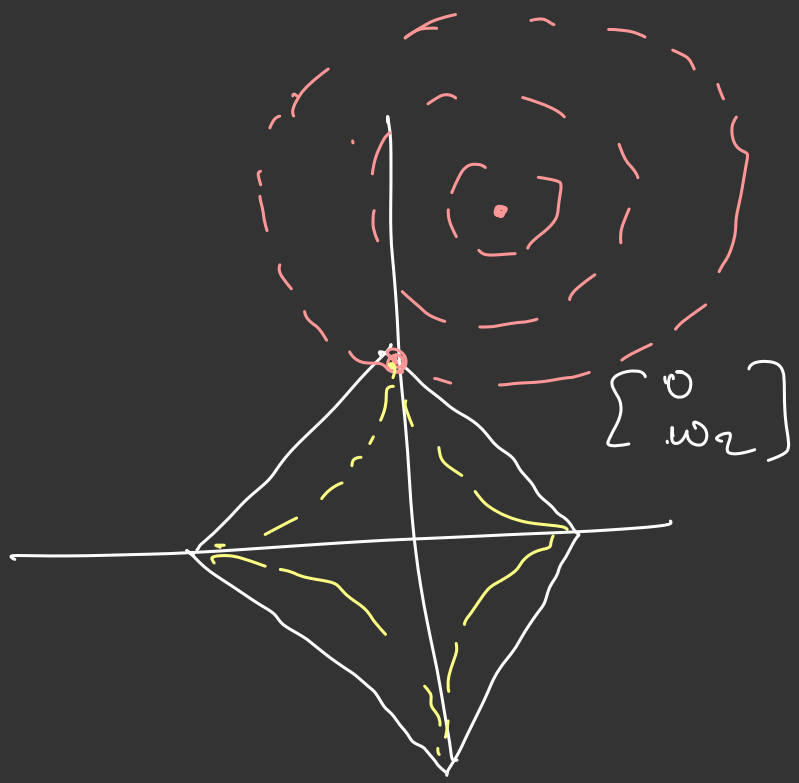


LASSO:

$$L(\underline{w}) = E_D(w) + \lambda \|\underline{w}\|_1$$

$q = 0.5$




$\begin{bmatrix} 0 \\ \omega_2 \end{bmatrix}$