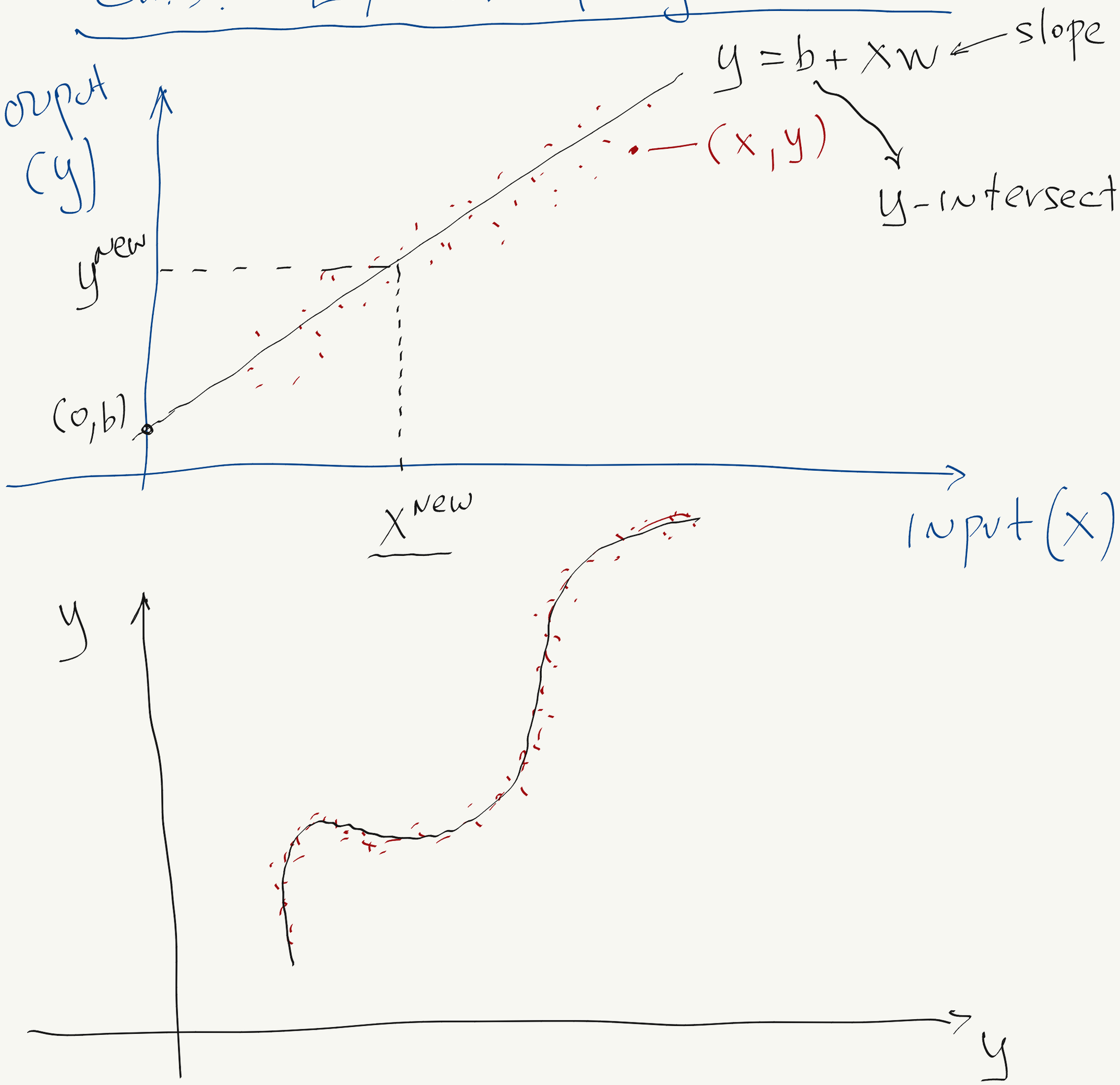
4/17/20

Ch. 5. Linear Regression.



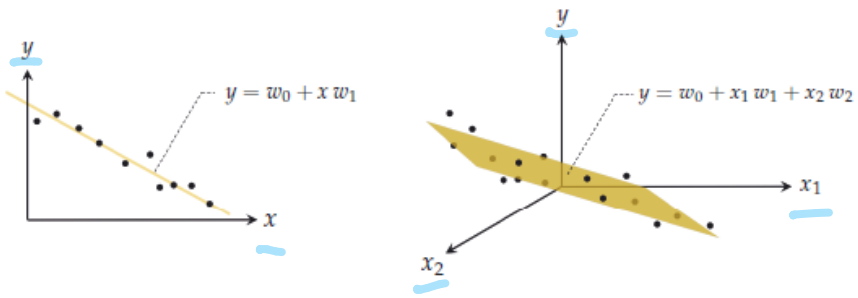


Figure 5.1 (left panel) A simulated dataset in two dimensions along with a well-fitting line. A line in two dimensions is defined as $w_0 + xw_1 = y$, where w_0 is referred to as the bias and w_1 the slope, and a point $\left(x_p, y_p\right)$ lies close to it if $w_0 + x_p w_1 \approx y_p$. (right panel) A simulated three-dimensional dataset along with a well-fitting hyperplane. A hyperplane in general is defined as $w_0 + x_1 w_1 + x_2 w_2 + \cdots + x_N w_N = y$, where again w_0 is the bias and $w_1, w_2, \ldots, w_N$ the hyperplane's coordinate-wise slopes, and a point $\left(\mathbf{x}_p, y_p\right)$ lies close to it if $w_0 + x_{1,p} w_1 + x_{2,p} w_2 + \cdots + x_{N,p} w_N \approx y_p$. Here $N = 2$.

Regression data : P input/output pairs

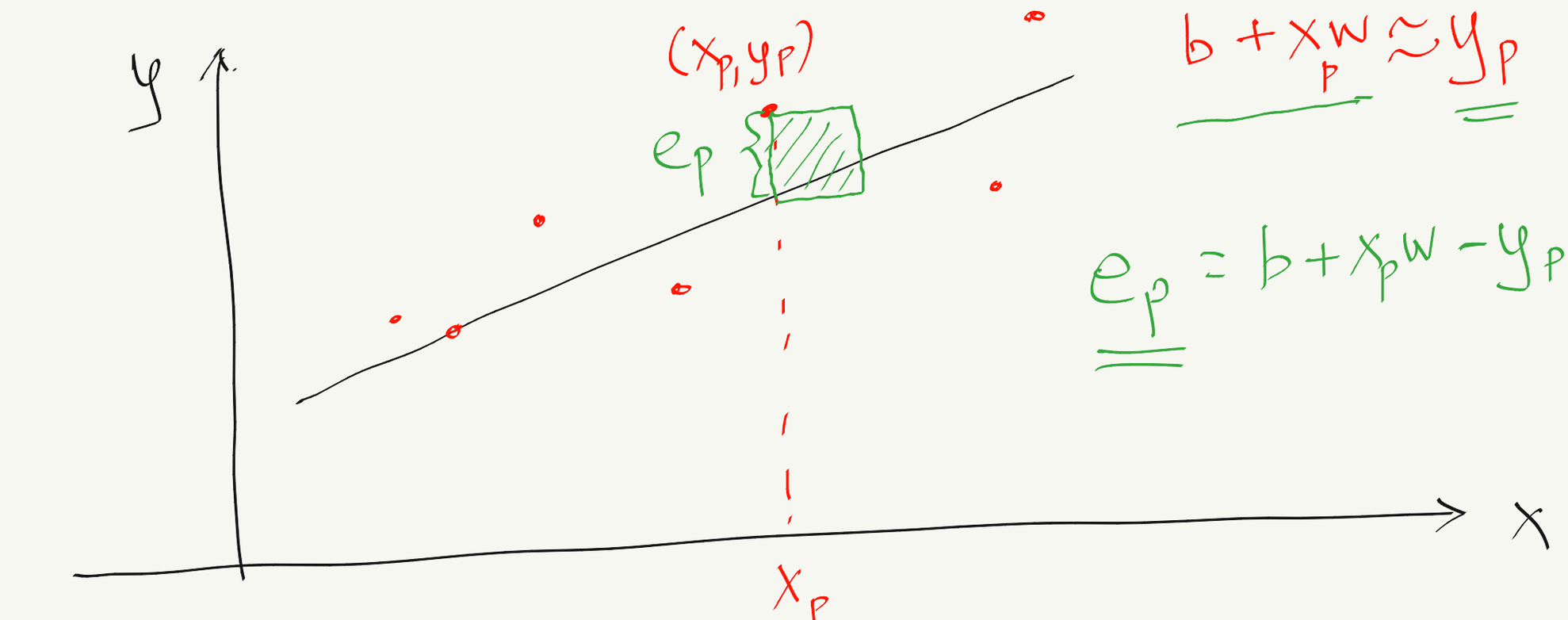
$$\{(\bar{x}_1, y_1), (\bar{x}_2, y_2) \dots, (\bar{x}_P, y_P)\}$$

$\bar{x}_1 \in R^N$ $y_1 \in R^1$

$$\bar{x}_p = \begin{bmatrix} x_{1p} & \cdots & x_{Np} \end{bmatrix}^T$$

x_{1p}: 1st input feature

x_{Np}: N^{th} input feature



$$b + x_p w \approx y_p$$

$$e_p = b + x_p w - y_p$$

error for p-th data point

$$e_p = (b + x_p w - y_p)$$

Least-Squares Regression.

$$g(b,w) = \frac{1}{P} \sum_{p=1}^{P} e_p^2$$

$$= \frac{1}{P} \sum_{p=1}^{P} (b + x_p w - y_p)^2$$

$N=1$

N-dim

$$g(b, \bar{w}) = \frac{1}{P} \sum_{p=1}^{P} (b + \bar{x}_p^T \bar{w} - y_p)^2$$

$\Rightarrow$ $\Leftarrow$

b, $\bar{w}$: feature touching weights

b: unknown

$\bar{x}_p^T$: $1 \times N$

$\bar{w}$: $N \times 1$ unknowns

N-dim : $b + \bar{x}^T w = y$ equation of hyperplane

Optimization Problem

Find $b, \bar{w}$ so that $g(b, \bar{w})$ is as small as possible

---

Modify Notation

$$\tilde{x}_p = \begin{bmatrix} 1 \\ \bar{x}_p \end{bmatrix} \qquad \tilde{w} = \begin{bmatrix} b \\ \bar{w} \end{bmatrix}$$

$(N+1) \times 1$ $\qquad$ $(N+1) \times 1$

$$\bar{x}_p^T \cdot \tilde{w} = \begin{bmatrix} 1 & \bar{x}_p^T \end{bmatrix} \begin{bmatrix} b \\ \bar{w} \end{bmatrix} = b + \bar{x}_p^T \bar{w}$$

Rewrite cost function

$$g(\tilde{w}) = \frac{1}{P} \sum_{p=1}^{P} \left(\underbrace{\tilde{x}_p^T \tilde{w} - y_p}_{\text{error}} \right)^2$$

$(x^T B x + x^T b)$

{ Optimization problem

$$\min_{\tilde{w}} g(\tilde{w}) \Rightarrow \tilde{w}^* = \arg \min_{\tilde{w}} g(\tilde{w})$$

1st order optimality condition

$\Rightarrow$ stationary points

$$\nabla_{\tilde{w}} g(\tilde{w}) = 0$$

$$\therefore \quad \nabla \frac{1}{P} \sum_{p=1}^{P} \left(\tilde{x}_p^T \tilde{w} - y_p \right)^2 = 0$$

$$\Rightarrow \frac{1}{P} \sum_{p=1}^{P} \nabla \left(\tilde{x}_p^T \tilde{w} - y_p \right)^2 = 0$$

$$\Rightarrow \frac{1}{P} \sum_{p=1}^{P} 2 \left(\tilde{x}_p^T \tilde{w} - y_p \right) \cdot \underbrace{\nabla_{\tilde{w}} \left(\tilde{x}_p^T \tilde{w} - y_p \right)}_{\nabla_{\tilde{w}} (\tilde{x}_p^T \tilde{w}) - \nabla_{\tilde{w}} y_p \,\to\, 0} = 0$$

$$\nabla_{\tilde{w}} (\tilde{x}_p^T \tilde{w}) = \nabla_{\tilde{w}} (\tilde{w}^T \tilde{x}_p) = \tilde{x}_p$$

$$\Rightarrow \left(\frac{2}{P} \right) \sum_{p=1}^{P} \underbrace{\left(\tilde{x}_p^T \tilde{w} - y_p \right)}_{\text{scalar}} \tilde{x}_p = 0$$

$$\Rightarrow \sum_{p=1}^{P} \tilde{x}_p \cdot \left(\tilde{x}_p^T \tilde{w} - y_p \right) = 0$$

$$\Rightarrow \boxed{\sum_{p=1}^{P} \tilde{x}_p \cdot \tilde{x}_p^T \tilde{w} = \sum_{p=1}^{P} \tilde{x}_p \cdot y_p}$$

Normal Equations

$$\underbrace{\sum_{p=1}^{P} \left(\tilde{x}_p \, \tilde{x}_p^T \right)}_{A} \tilde{w} = \underbrace{\sum_{p=1}^{P} \tilde{x}_p \cdot y_p}_{\bar{b}}$$

outer product: $[(N+1)\times 1]\,[1\times(N+1)]$ $\rightarrow$ $(N+1)\times(N+1)$

$\tilde{w}$: $(N+1)\times 1$

$\bar{b}$: $(N+1)\times 1$

$$\boxed{A\tilde{w} = \bar{b}} \leftarrow$$

If A is invertible: $\tilde{w}^* = A^{-1}\bar{b}$

if A is non-invertible: $\tilde{w}^* = A^{+}\bar{b}$

A^{+}: generalized inverse

Iterative Optimization

- Newton's method: in one step we obtain the solution, since Hessian $= A$

$\Rightarrow$ solve normal equations in one-step

- GD

$$\tilde{w}^{k+1} = \tilde{w}^k - \alpha \nabla g(\tilde{w}^k)$$

$$= \tilde{w}^k - \alpha \left(A\tilde{w}^k - \bar{b} \right)$$

$$= \underbrace{\left(I - \alpha A \right)}_{B} \tilde{w}^k + \underbrace{\alpha \bar{b}}_{\bar{c}}$$

need to be computed once

$$\Rightarrow \boxed{\tilde{w}^{k+1} = B\tilde{w}^k + \bar{c}}$$

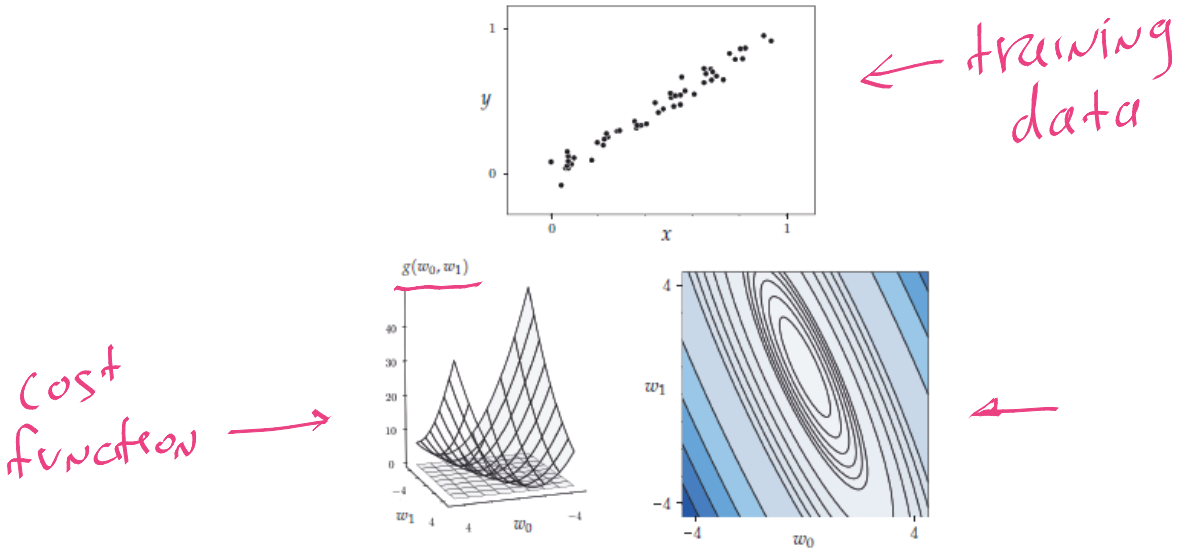


Figure 5.3 Figure associated with Example 5.1. See text for details.

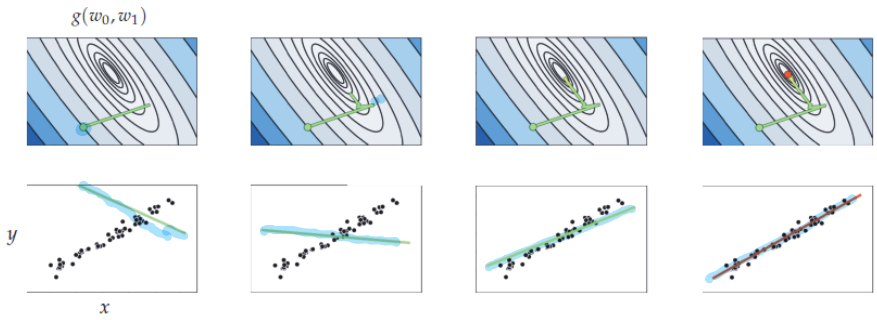
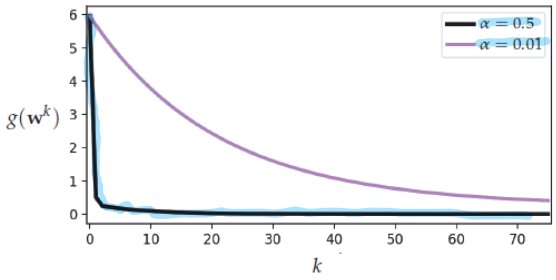


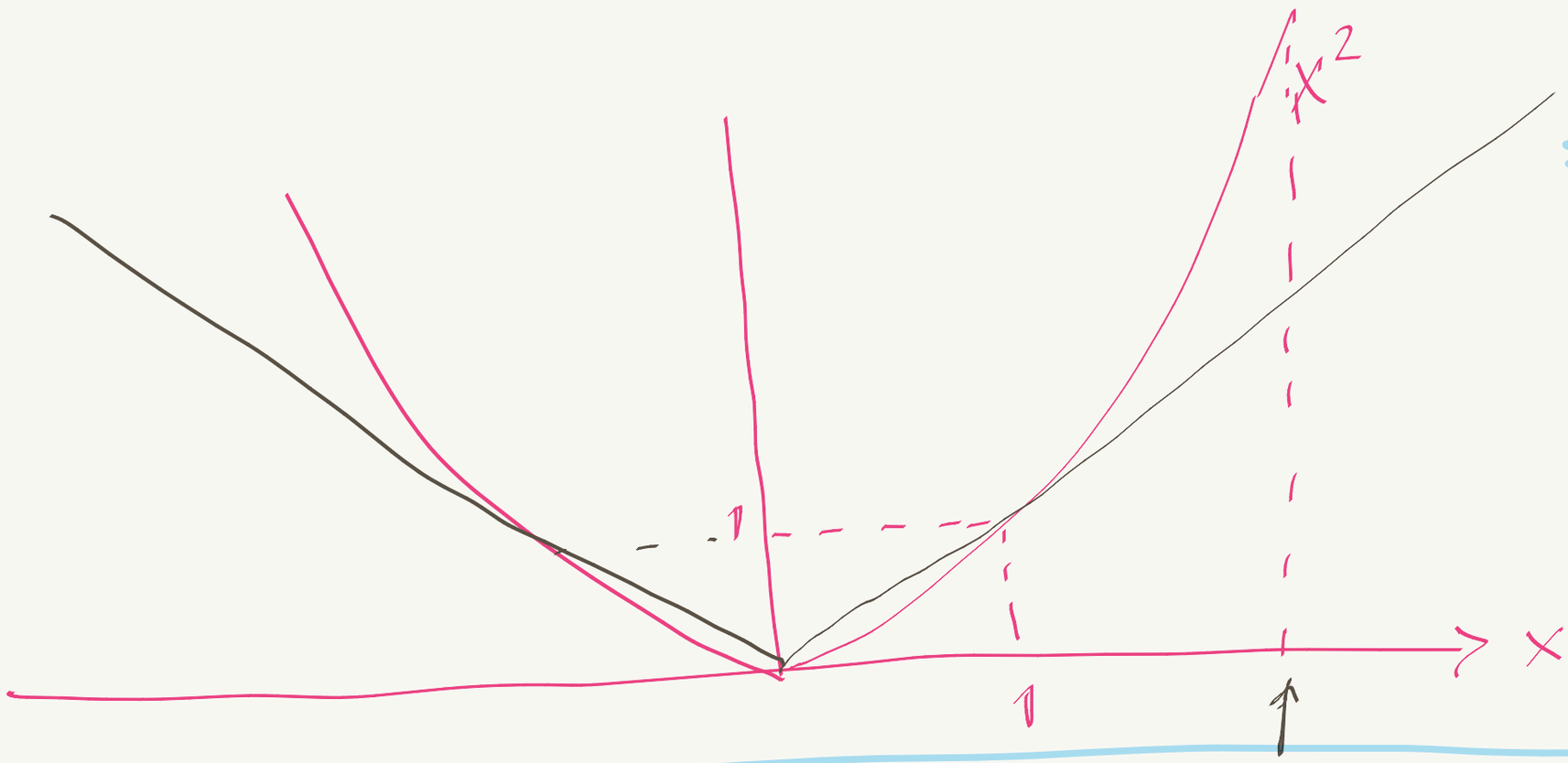
Figure 5.4 Figure associated with Example 5.2. See text for details.

Figure 5.5 Figure associated with Example 5.2. See text for details.



Squaring the error increases the importance of large error

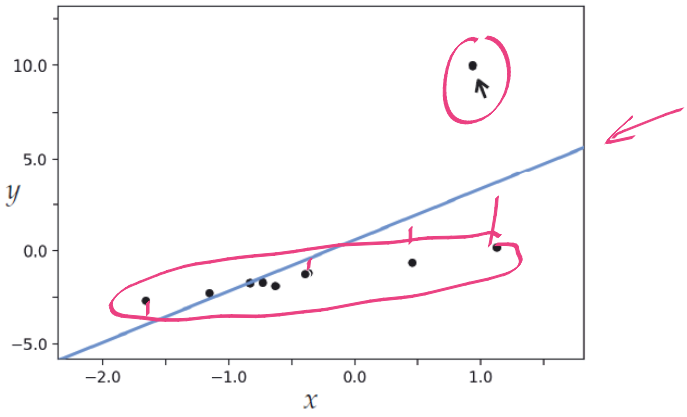
$\Longrightarrow$ in trying to minimize them we overfit to outlier



Least Absolute Deviations Regressor

$$g(\tilde{w}) = \frac{1}{P} \sum_{p=1}^{P} \underbrace{|\tilde{x}^T \tilde{w} - y_p|}_{\text{error}}$$

Figure 5.6 Figure associated with Example 5.3. See text for details.



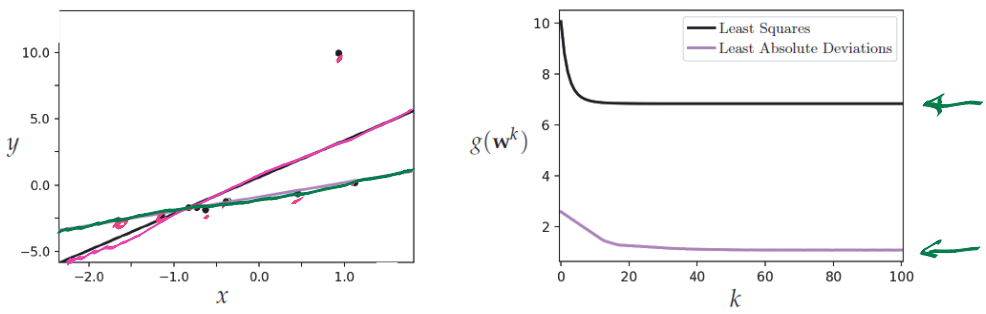


Figure 5.7 Figure associated with Example 5.4. See text for details.