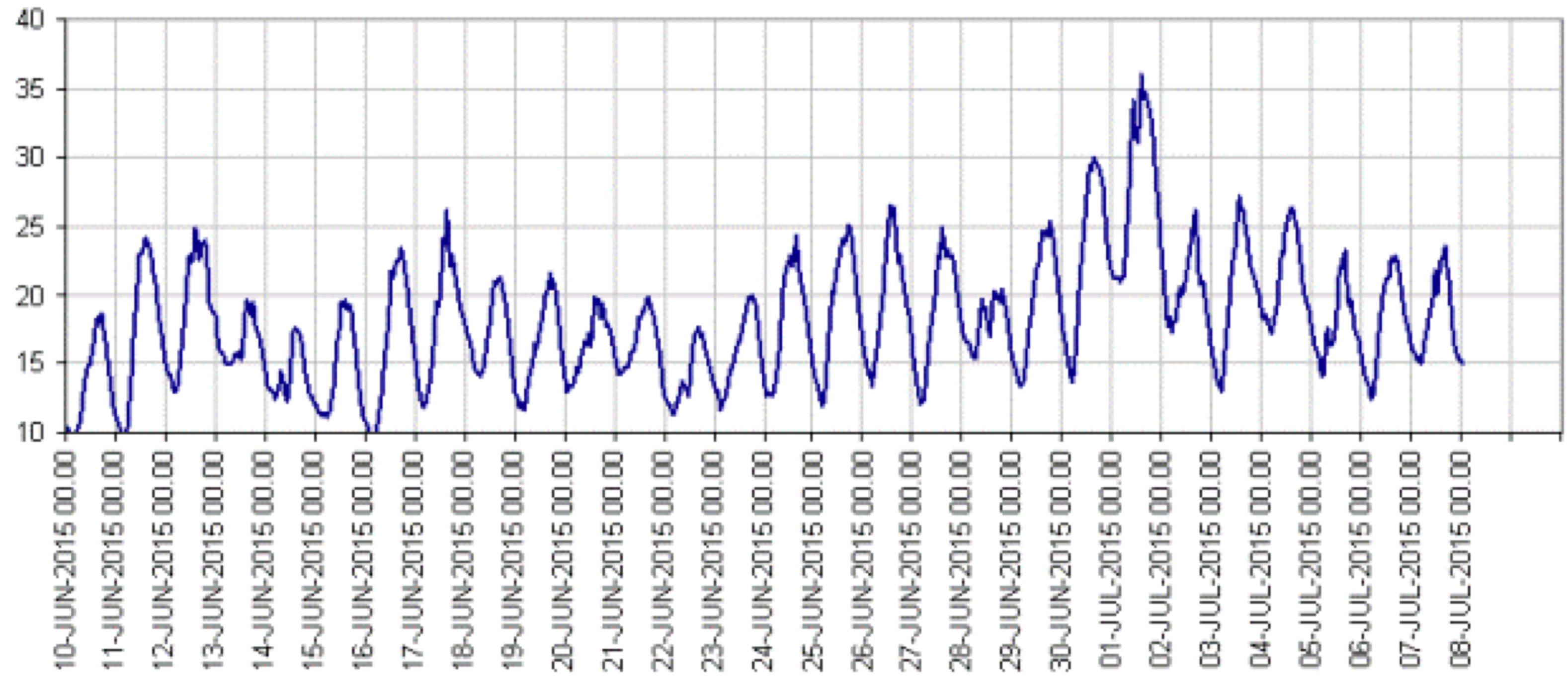
Linear regression

Goodfellow et al. 2016 (sec. 5.1.4)

Example: weather prediction



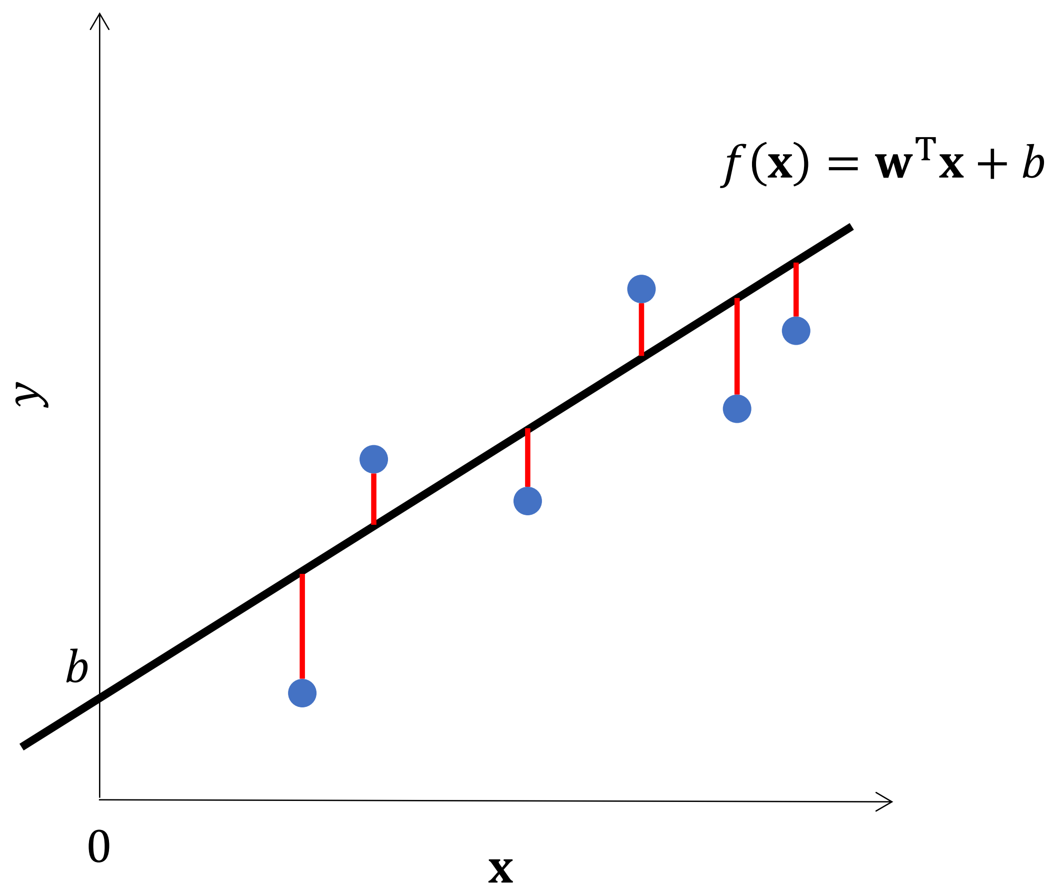
Forecast air temperature in London

Linear model

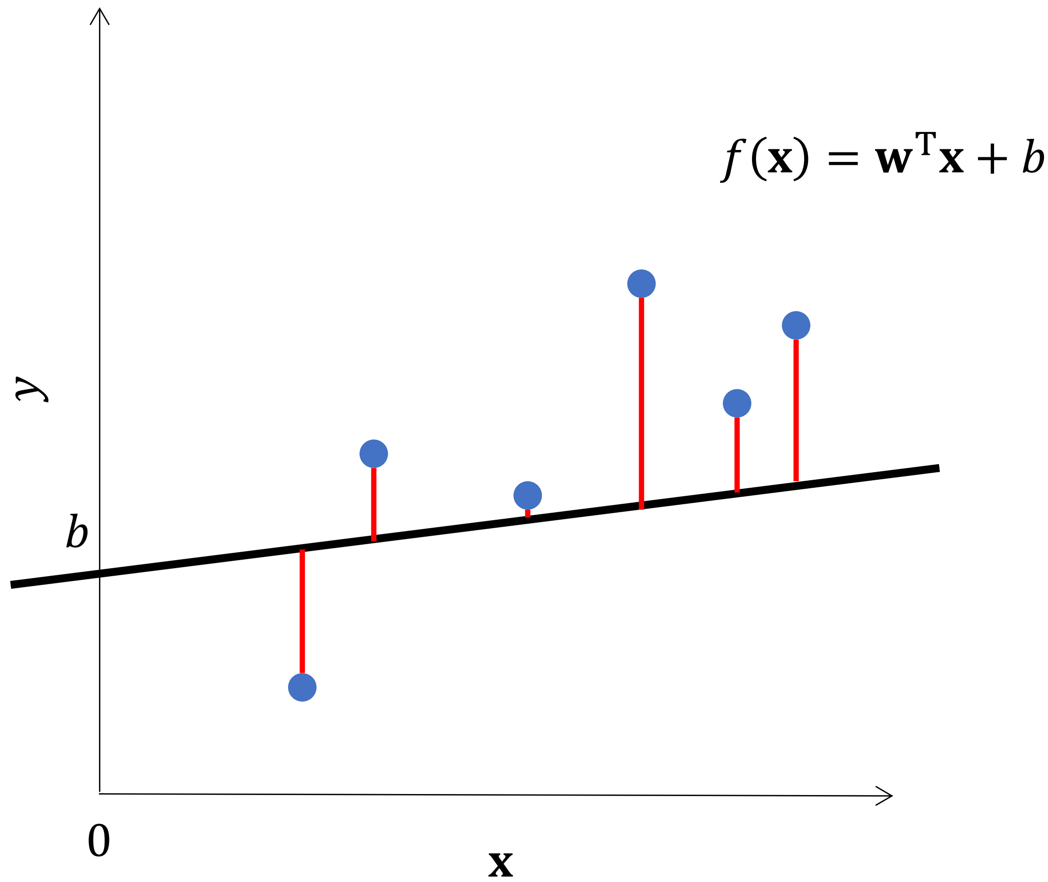
- Given training data $\{(\mathbf{x}_i, y_i) \sim p \ \text{ i. i. d.}\}_{i=1}^{n}$
- Assuming linear model $\hat{y} = f_{\mathbf{w},b}(\mathbf{x}) = \mathbf{w}^{\mathrm{T}}\mathbf{x} + b$
- Find optimal parameters $\mathbf{w}, b$ by minimizing empirical loss

$$\hat{L}(\mathbf{w}, b) = \frac{1}{n} \sum_{i=1}^{n} \left(\mathbf{w}^{\mathrm{T}}\mathbf{x}_i + b - y_i\right)^2$$

Linear model

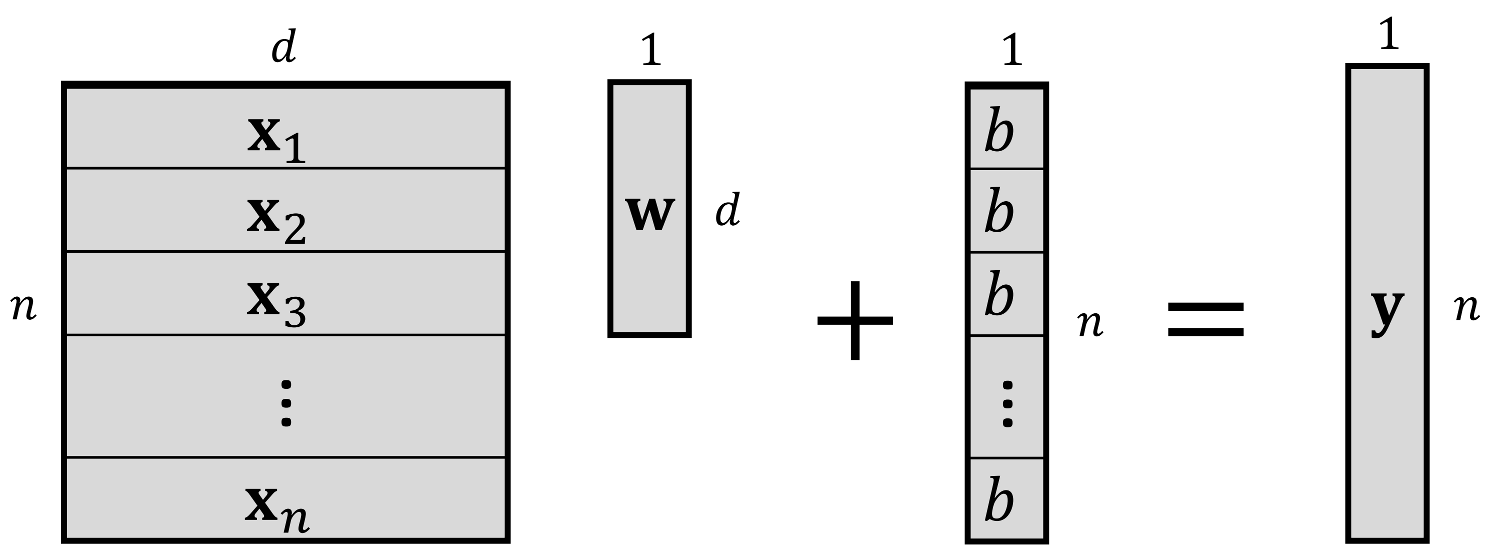


Linear model



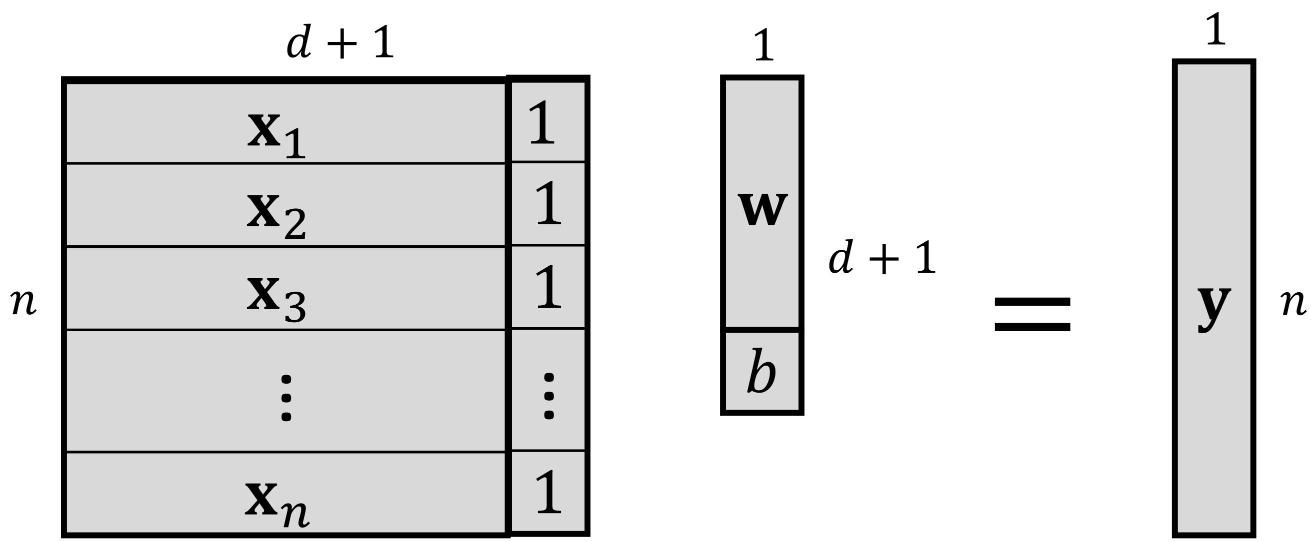
Finding optimal parameters

- Matrix-vector notation $\mathbf{X}\mathbf{w} + b\mathbf{1} = \mathbf{y}$



Finding optimal parameters

- Matrix-vector notation: absorb the bias into the weight vector
  $\mathbf{X}\mathbf{w} = \mathbf{y}$



Finding optimal parameters

- Loss $\hat{L}(\mathbf{w}) = \frac{1}{n}||\mathbf{X}\mathbf{w} - \mathbf{y}||^2$

- Set the gradient to zero to get the minimizer:

$$0 = \nabla_{\mathbf{w}}\hat{L}(\mathbf{w}) = \frac{1}{n}\nabla_{\mathbf{w}}||\mathbf{X}\mathbf{w} - \mathbf{y}||^2$$

$$0 = \nabla_{\mathbf{w}}[(\mathbf{X}\mathbf{w} - \mathbf{y})^{\mathrm{T}}(\mathbf{X}\mathbf{w} - \mathbf{y})]$$

$$0 = \nabla_{\mathbf{w}}[\mathbf{w}^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}\mathbf{X}\mathbf{w} - 2\mathbf{w}^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}\mathbf{y} + \mathbf{y}^{\mathrm{T}}\mathbf{y}]$$

$$0 = 2\mathbf{X}^{\mathrm{T}}\mathbf{X}\mathbf{w} - 2\mathbf{X}^{\mathrm{T}}\mathbf{y}$$

$$\boxed{\mathbf{w} = \left(\mathbf{X}^{\mathrm{T}}\mathbf{X}\right)^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{y}}$$

Matrix cookbook

Regularization

- Inverse well-defined only when $\mathrm{rank}\left(\mathbf{X}^{\mathrm{T}}\mathbf{X}\right) = d$ implying $n \geq d$
  = (over-)determined system = more equations than variables

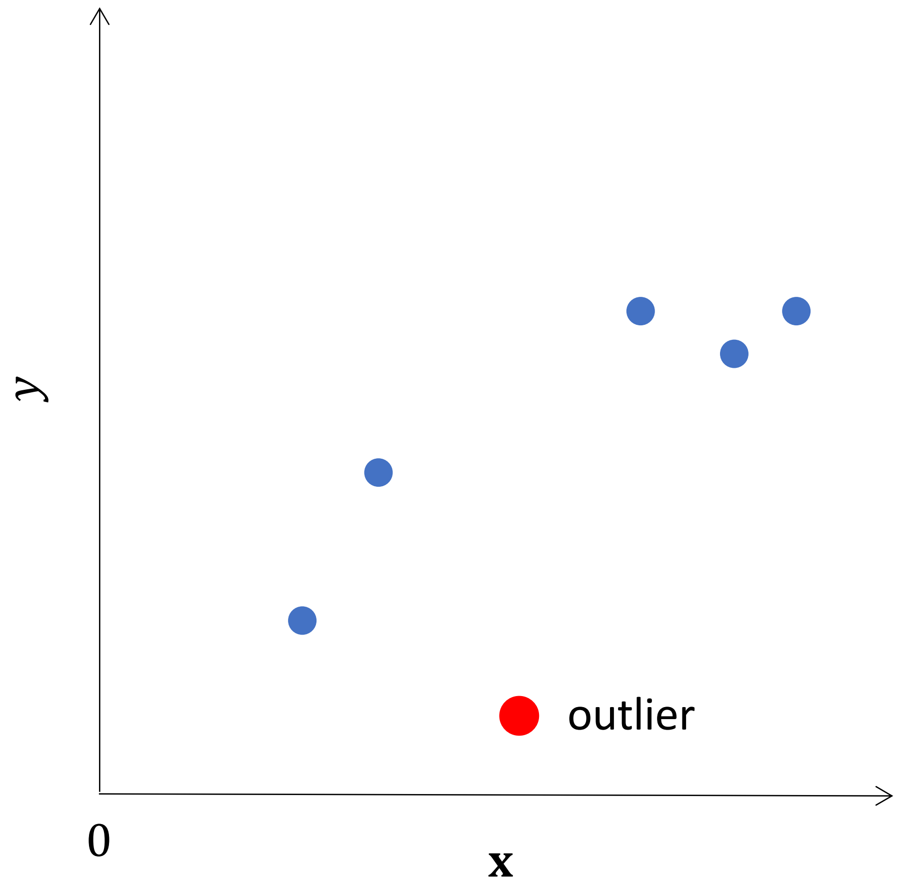
- Otherwise, add **regularization term** (**weight decay**)

$$\hat{L}(\mathbf{w}) = \frac{1}{n}\|\mathbf{X}\mathbf{w} - \mathbf{y}\|^2 + \alpha\|\mathbf{w}\|^2$$

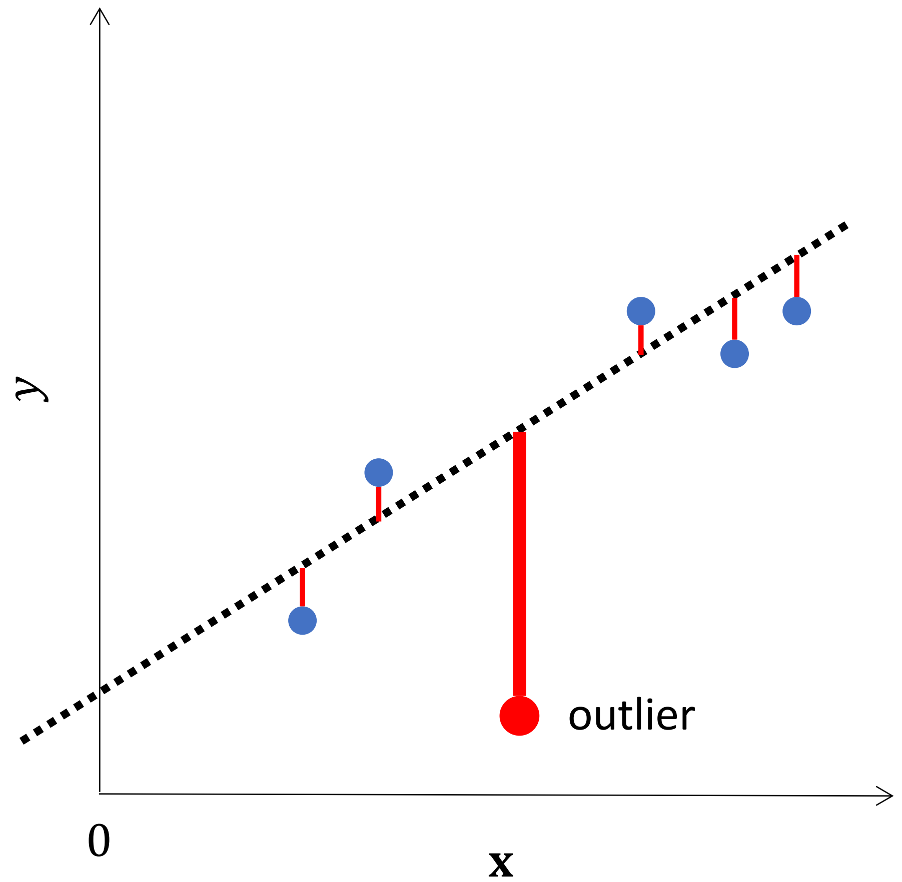
$$0 = \nabla_{\mathbf{w}}\hat{L}(\mathbf{w}) = \frac{2}{n}\mathbf{X}^{\mathrm{T}}\mathbf{X}\mathbf{w} - \frac{2}{n}\mathbf{X}^{\mathrm{T}}\mathbf{y} + 2\alpha\mathbf{w}$$

$$\boxed{\mathbf{w} = \left(\frac{1}{n}\mathbf{X}^{\mathrm{T}}\mathbf{X} + \alpha\mathbf{I}\right)^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{y}}$$

Loss choice

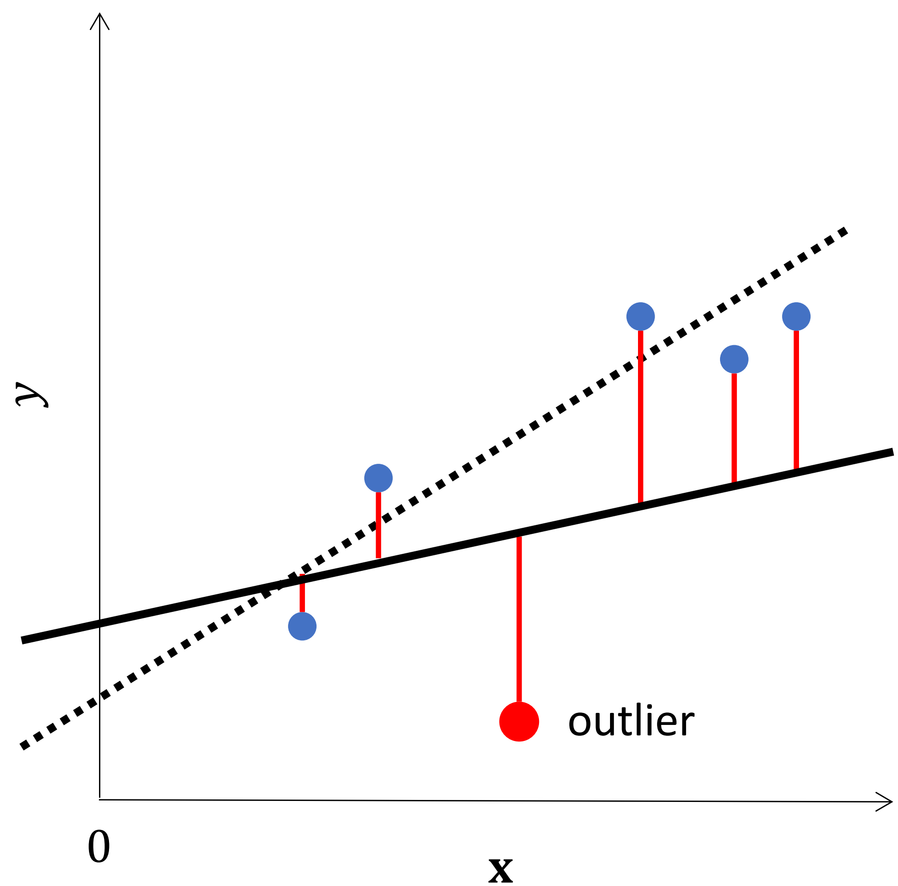


Loss choice



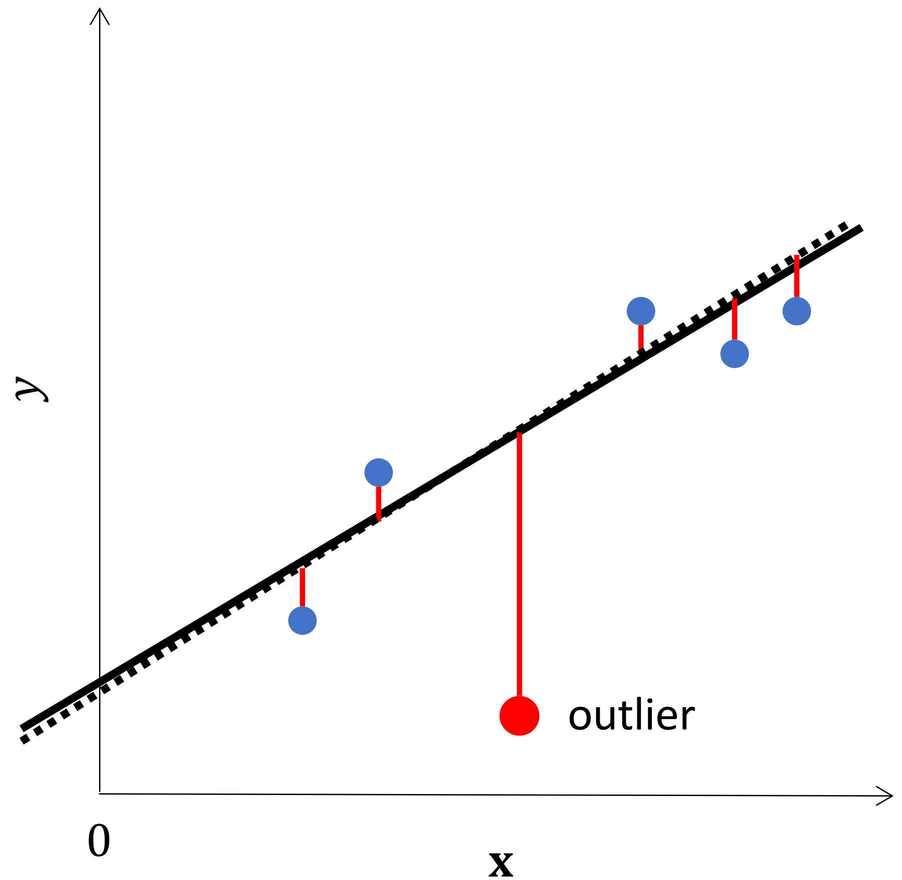
L_2 loss $\quad \|\mathbf{X}\mathbf{w}-\mathbf{y}\|_2^2 = \frac{1}{n}\sum_{i=1}^{n}\left(\mathbf{w}^{\mathrm{T}}\mathbf{x}_i + b - y_i\right)^2$

Loss choice



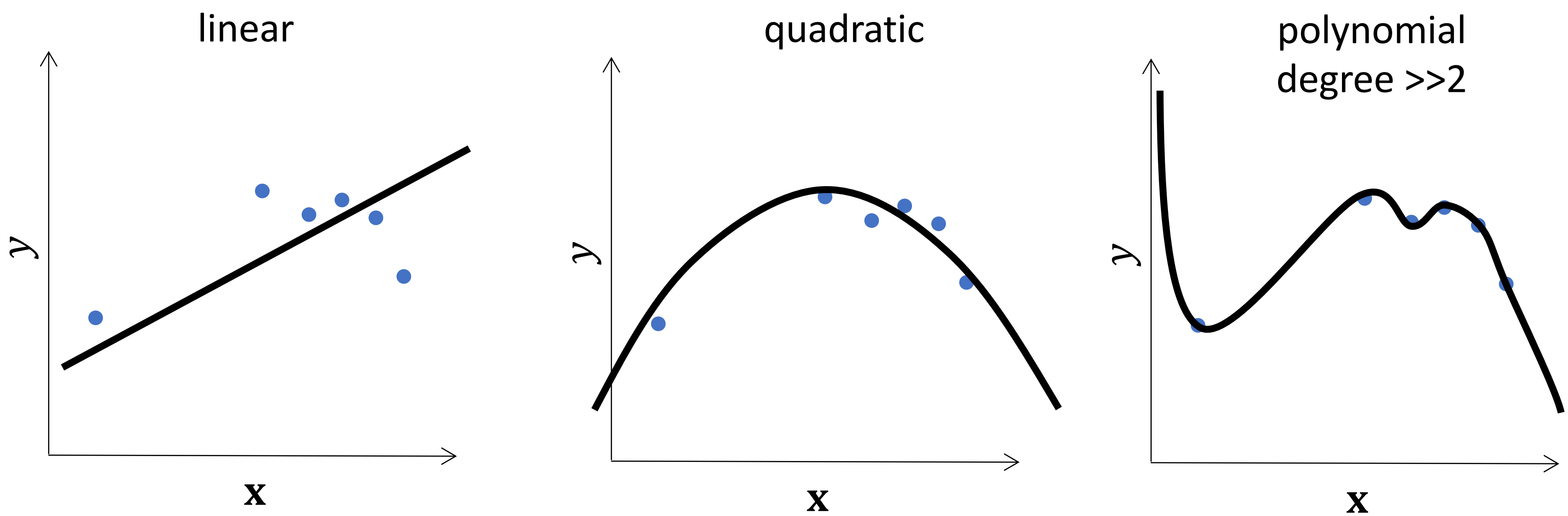
L_2 loss $\quad \|\mathbf{X}\mathbf{w}-\mathbf{y}\|_2^2 = \frac{1}{n}\sum_{i=1}^{n}\left(\mathbf{w}^{\mathrm{T}}\mathbf{x}_i + b - y_i\right)^2$

Loss choice

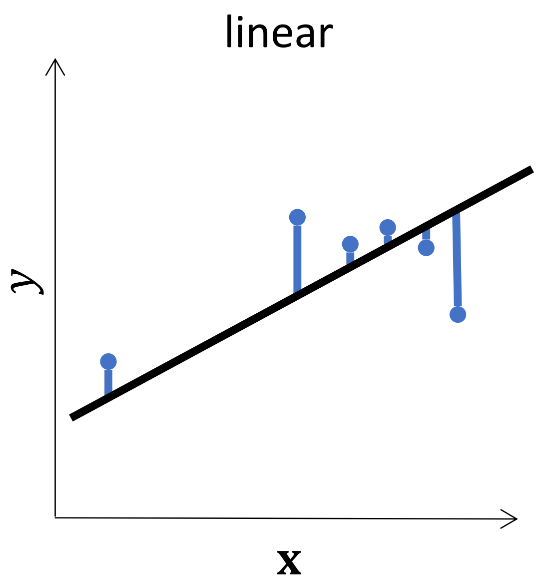


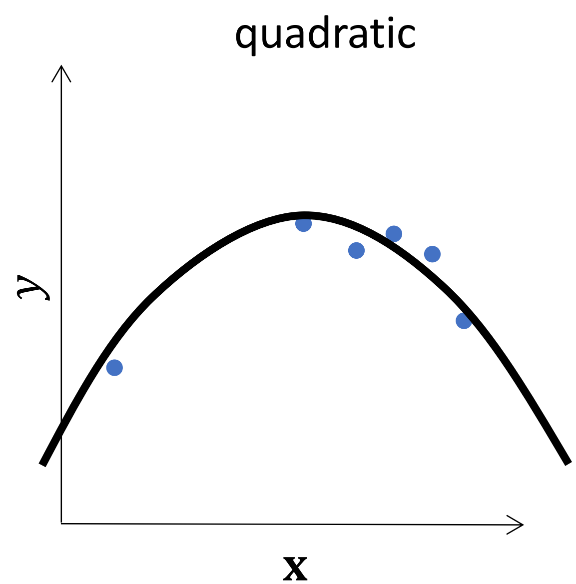
L_1 loss $\quad \|\mathbf{X}\mathbf{w} - \mathbf{y}\|_1 \; = \dfrac{1}{n}\displaystyle\sum_{i=1}^{n} \left|\mathbf{w}^{\mathrm{T}}\mathbf{x}_i + b - y_i\right|$

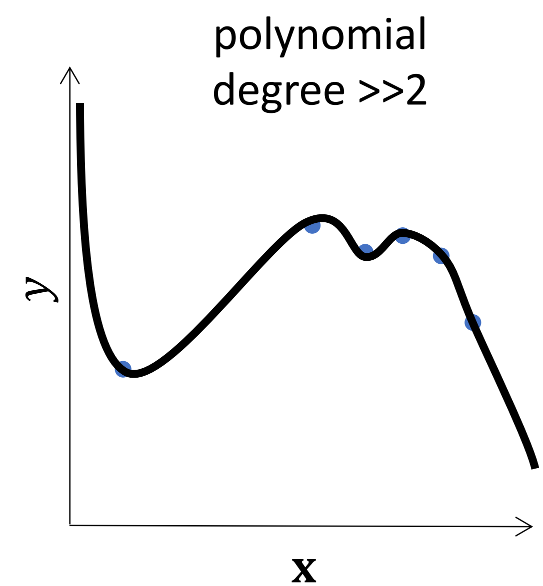
Model capacity



Model capacity







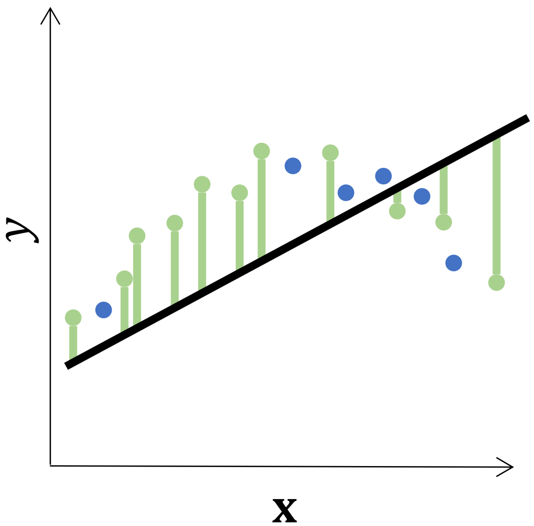
Large training error

Small training error

Model capacity

Underfitting

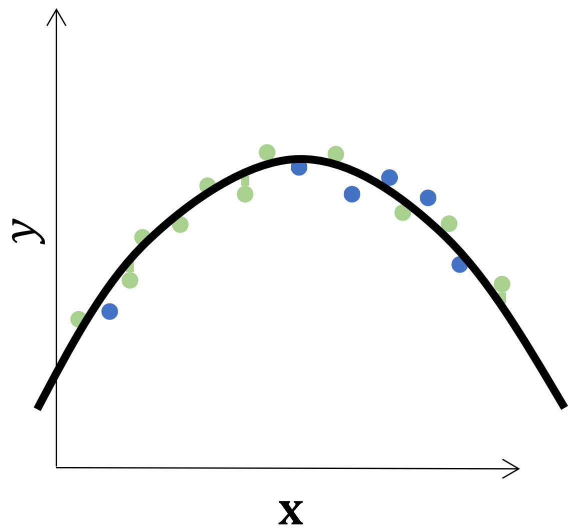
linear



Large training error
Large generalization error

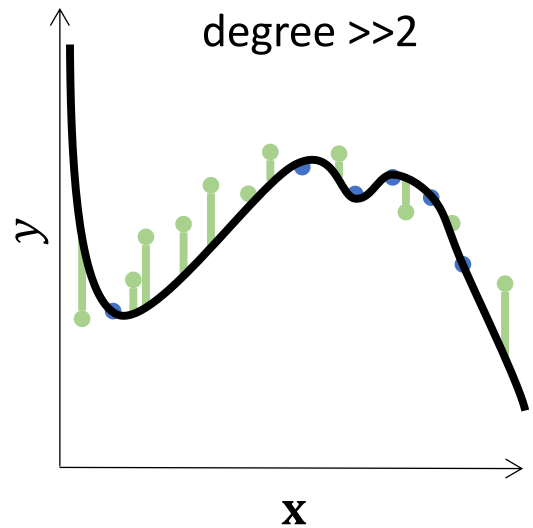
Appropriate capacity

quadratic



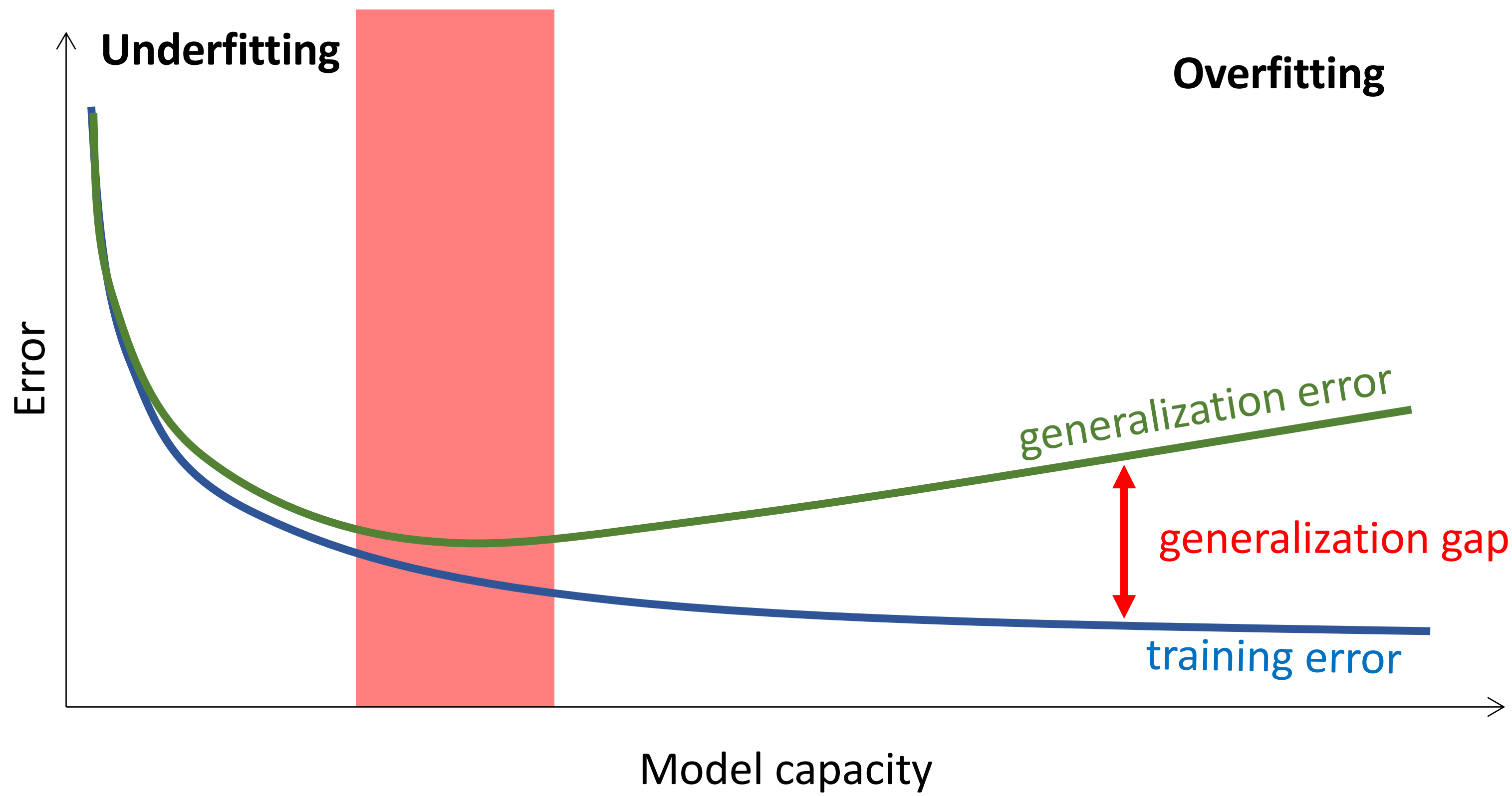
Overfitting

polynomial
degree >>2



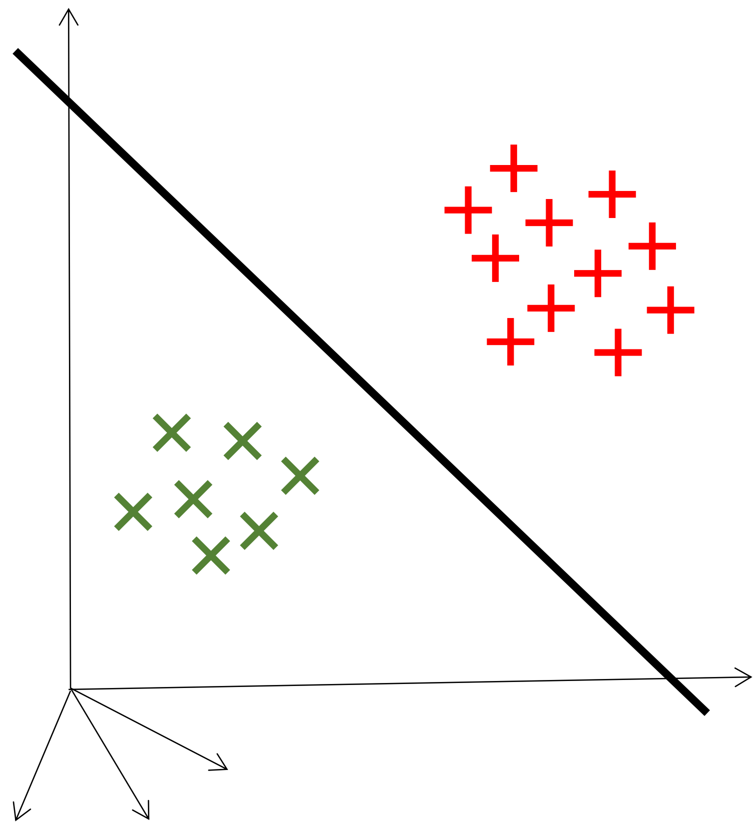
Small training error
Large generalization error

Model capacity



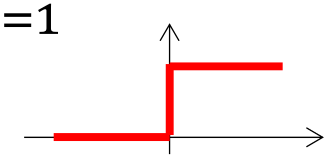
Linear classification

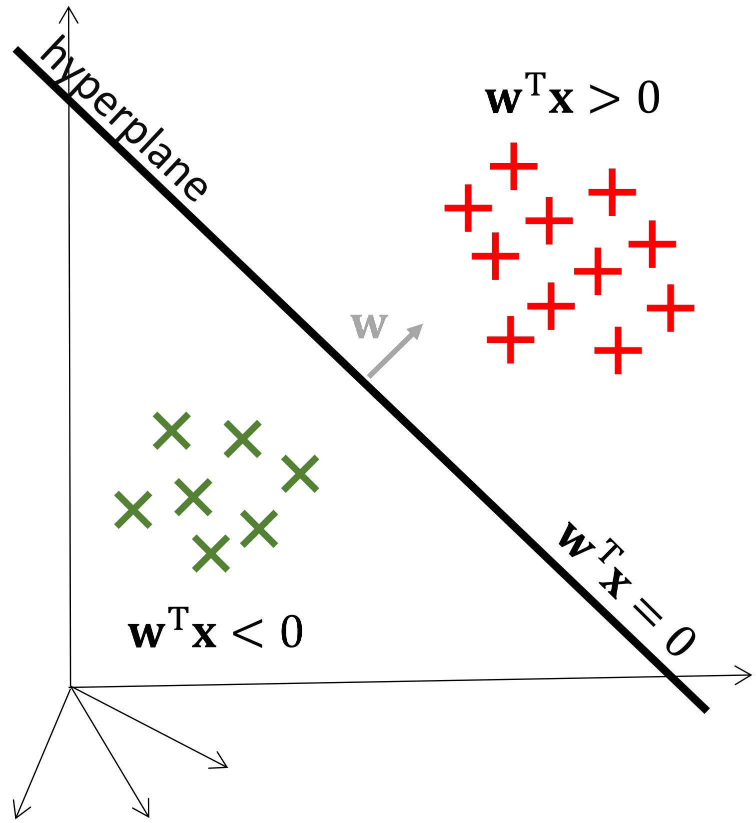
Linear classification



Linear classification

- Training data
$$\{(\mathbf{x}_i, y_i \in \{0,1\})\}_{i=1}^{n}$$
- Linear model $f_{\mathbf{w}}(\mathbf{x}) = \mathbf{w}^{\mathrm{T}}\mathbf{x}$
- 0-1 loss
$$\hat{L}(\mathbf{w}) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{I}_{\mathrm{step}(\mathbf{w}^{\mathrm{T}}\mathbf{x}_i)\neq y_i}$$



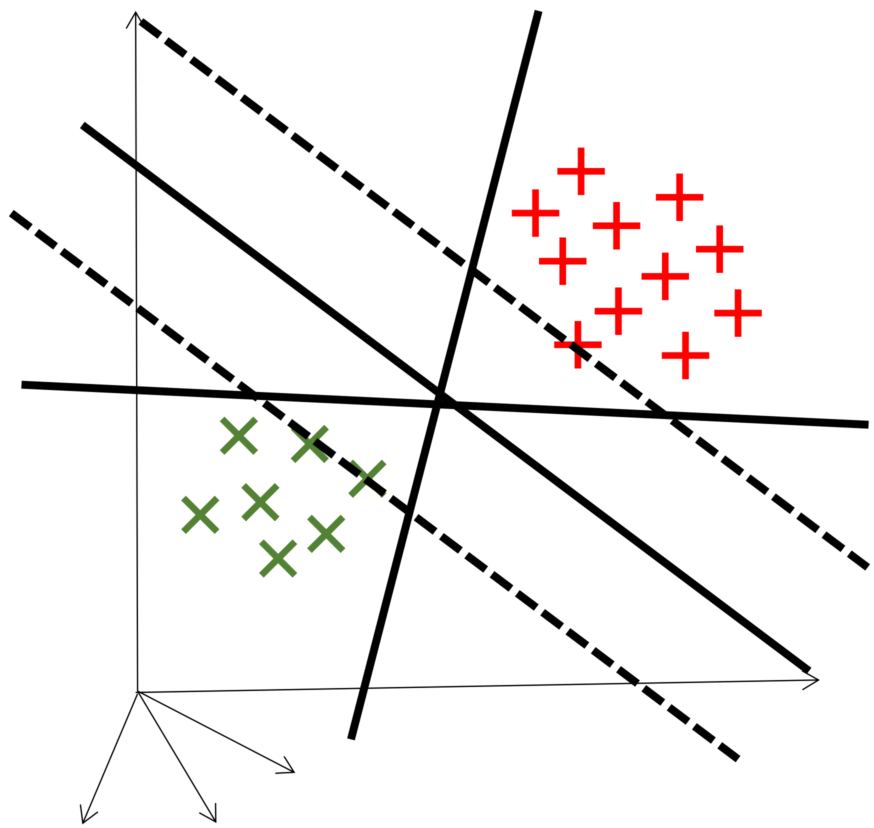


Hard to optimize!

* Often class labels in binary classification problems are denoted by +1 and -1

Linear classification

- Training data
$$\{(\mathbf{x}_i, y_i \in \{0,1\})\}_{i=1}^{n}$$
- Linear model $f_{\mathbf{w}}(\mathbf{x}) = \mathbf{w}^{\mathrm{T}}\mathbf{x}$
- 0-1 loss
$$\hat{L}(\mathbf{w}) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{I}_{\mathrm{step}(\mathbf{w}^{\mathrm{T}}\mathbf{x}_i) \neq y_i}$$



Many possible solutions

Support vector machine

Linear classification

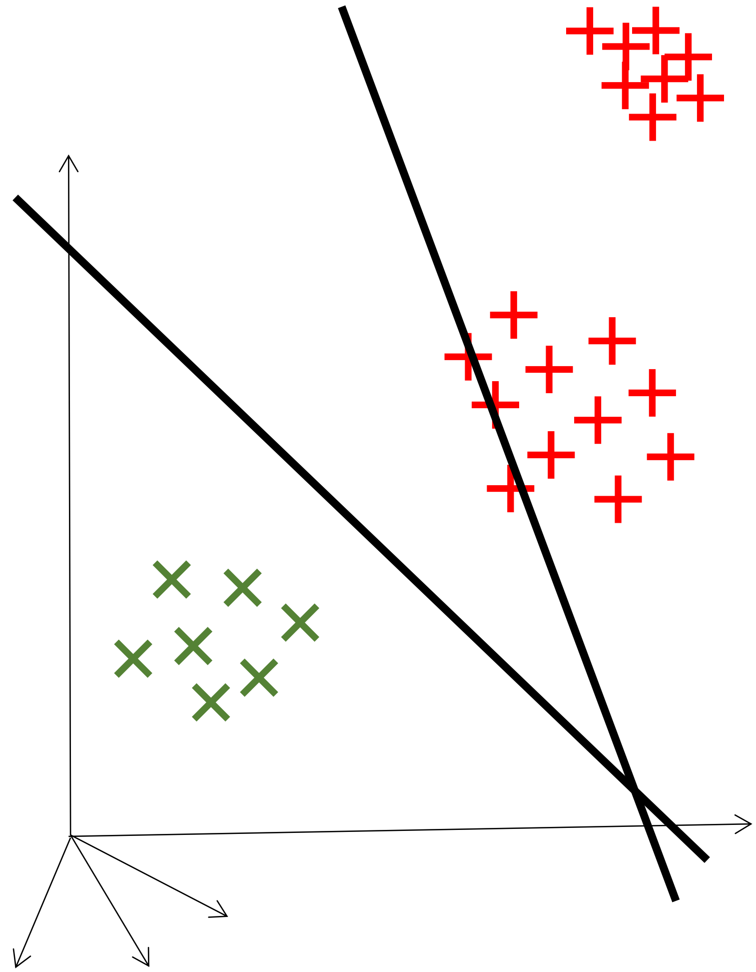
- Training data

$$\{(\mathbf{x}_i, y_i \in \{0,1\})\}_{i=1}^{n}$$

- Linear model $f_{\mathbf{w}}(\mathbf{x}) = \mathbf{w}^{\mathrm{T}}\mathbf{x}$

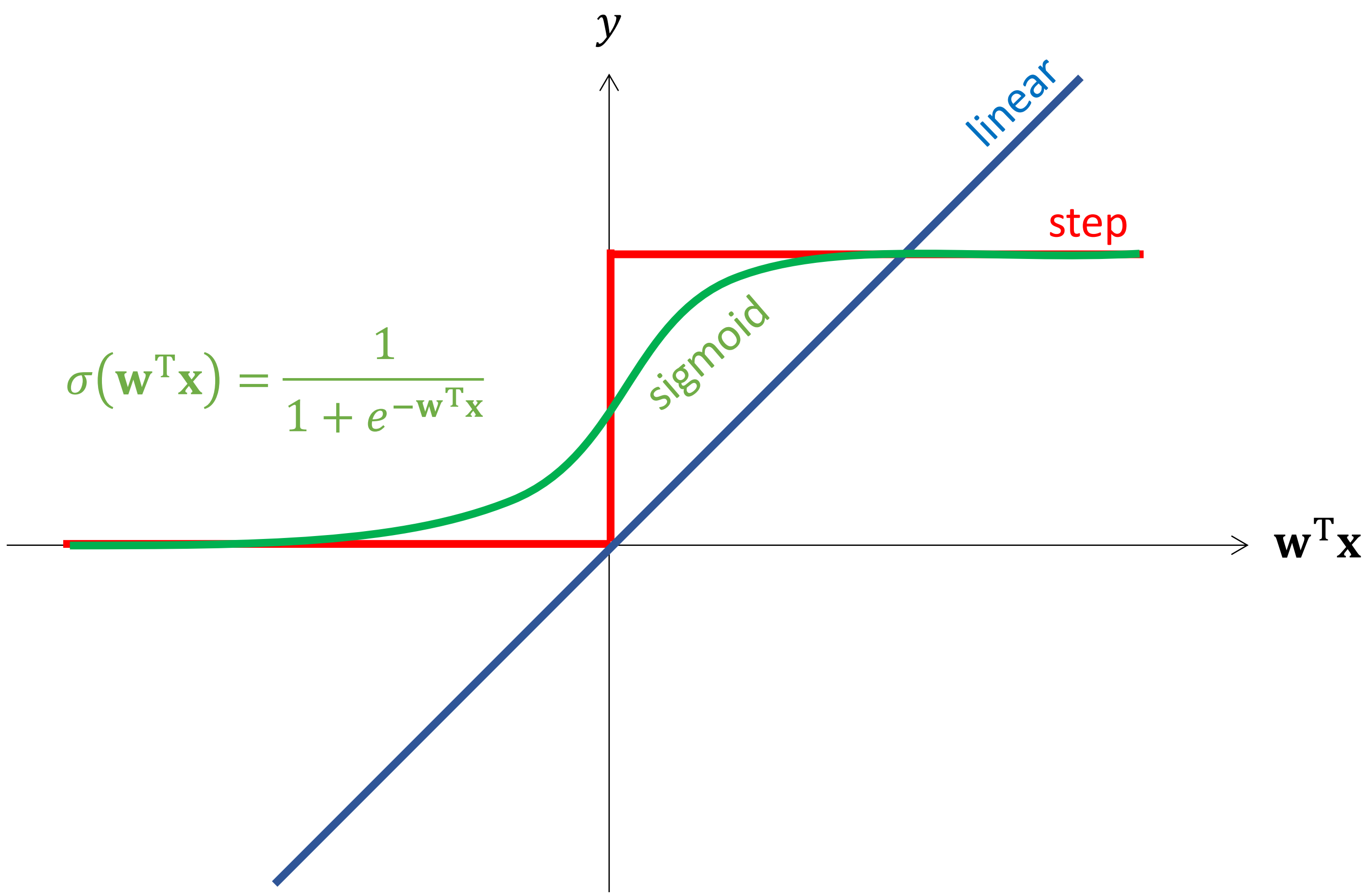
- L_2-loss (ignore binary nature of y)

$$\hat{L}(\mathbf{w}) = \frac{1}{n}\sum_{i=1}^{n}\left(\mathbf{w}^{\mathrm{T}}\mathbf{x}_i - y_i\right)^2$$



Sensitive to outliers!

Logistic function a.k.a. sigmoid



MLE interpretation of logistic regression

- Assume $p_{\mathbf{w}}(y|\mathbf{x}) = \mathrm{Bernoulli}(p)$, i.e. given $\mathbf{x}$
  $y = 1$ with probability p $\qquad$ $y = 0$ with probability $1 - p$

- Linear model applied to **log-odds** (logit)

$$\log\left(\frac{p}{1-p}\right) = \mathbf{w}^{\mathrm{T}}\mathbf{x}$$

$$\frac{p}{1-p} = e^{\mathbf{w}^{\mathrm{T}}\mathbf{x}}$$

$$p = \frac{e^{\mathbf{w}^{\mathrm{T}}\mathbf{x}}}{1 + e^{\mathbf{w}^{\mathrm{T}}\mathbf{x}}} = \frac{1}{1 + e^{-\mathbf{w}^{\mathrm{T}}\mathbf{x}}} = \sigma(\mathbf{w}^{\mathrm{T}}\mathbf{x}) = p_{\mathbf{w}}(y = 1|\mathbf{x})$$

and $p_{\mathbf{w}}(y = 0|\mathbf{x}) = 1 - \sigma(\mathbf{w}^{\mathrm{T}}\mathbf{x})$

MLE interpretation of logistic regression

- Assume $p_{\mathbf{w}}(y|\mathbf{x}) = \mathrm{Bernoulli}(p)$, i.e. given $\mathbf{x}$
  $y = 1$ with probability p $\qquad$ $y = 0$ with probability $1 - p$

- Log-likelihood

$$\hat{L}(\mathbf{w}) = -\frac{1}{n}\sum_{i=1}^{n} \log p_{\mathbf{w}}(y|\mathbf{x}_i)$$

$$= -\frac{1}{n}\sum_{i:y_i=1} \log \sigma\left(\mathbf{w}^{\mathrm{T}}\mathbf{x}_i\right) - \frac{1}{n}\sum_{i:y_i=0} \log\left(1 - \sigma\left(\mathbf{w}^{\mathrm{T}}\mathbf{x}_i\right)\right)$$

$$= -\frac{1}{n}\sum_{i=1}^{n} y_i \log \sigma\left(\mathbf{w}^{\mathrm{T}}\mathbf{x}_i\right) + (1 - y_i)\log\left(1 - \sigma\left(\mathbf{w}^{\mathrm{T}}\mathbf{x}_i\right)\right)$$

MLE interpretation of logistic regression

- Assume $p_{\mathbf{w}}(y|\mathbf{x}) = \text{Bernoulli}(p)$, i.e. given $\mathbf{x}$
  $y = 1$ with probability p $\qquad$ $y = 0$ with probability $1 - p$

- Vectorized version of log-likelihood

$$\hat{L}(\mathbf{w}) = -\frac{1}{n}\left[\mathbf{y}^{\mathrm{T}}\log\sigma(\mathbf{X}\mathbf{w}) + (\mathbf{1} - \mathbf{y})^{\mathrm{T}}\log\big(\mathbf{1} - \sigma(\mathbf{X}\mathbf{w})\big)\right]$$