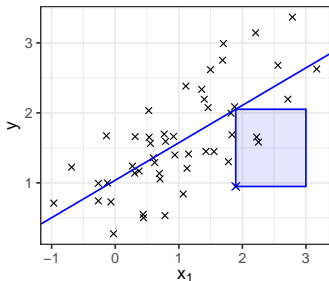


# Introduction to Machine Learning

## Supervised Regression

## Deep Dive: Proof OLS Regression



### Learning goals

- Understand analytical derivation of OLS estimator for LM

# ANALYTICAL OPTIMIZATION

- Special property of LM with  $L2$  loss: **analytical solution** available

$$\begin{aligned}\hat{\theta} \in \arg \min_{\theta} \mathcal{R}_{\text{emp}}(\theta) &= \arg \min_{\theta} \sum_{i=1}^n \left( y^{(i)} - \theta^{\top} \mathbf{x}^{(i)} \right)^2 \\ &= \arg \min_{\theta} \|\mathbf{y} - \mathbf{X}\theta\|_2^2\end{aligned}$$

- Find via **normal equations**

$$\frac{\partial \mathcal{R}_{\text{emp}}(\theta)}{\partial \theta} = 0$$

- Solution: **ordinary-least-squares (OLS)** estimator

$$\hat{\theta} = (\mathbf{X}^{\top} \mathbf{X})^{-1} \mathbf{X}^{\top} \mathbf{y}$$



# ANALYTICAL OPTIMIZATION – PROOF

$$\mathcal{R}_{\text{emp}}(\theta) = \sum_{i=1}^n \underbrace{(y^{(i)} - \theta^\top \mathbf{x}^{(i)})^2}_{=: \epsilon_i} = \underbrace{\|\mathbf{y} - \mathbf{X}\theta\|_2^2}_{=: \epsilon}; \quad \theta \in \mathbb{R}^{\tilde{p}} \text{ with } \tilde{p} := p + 1$$

$$0 = \frac{\partial \mathcal{R}_{\text{emp}}(\theta)}{\partial \theta} \quad (\text{sum notation})$$

$$0 = \frac{\partial}{\partial \theta} \sum_{i=1}^n \epsilon_i^2 \quad \Big| \quad \text{sum \& chain rule}$$

$$0 = \sum_{i=1}^n \frac{\partial \epsilon_i^2}{\partial \epsilon_i} \frac{\partial \epsilon_i}{\partial \theta}$$

$$0 = \sum_{i=1}^n 2\epsilon_i (-1) (\mathbf{x}^{(i)})^\top$$

$$0 = \sum_{i=1}^n (y^{(i)} - \theta^\top \mathbf{x}^{(i)}) (\mathbf{x}^{(i)})^\top$$

$$\theta^\top \sum_{i=1}^n \mathbf{x}^{(i)} (\mathbf{x}^{(i)})^\top = \sum_{i=1}^n y^{(i)} (\mathbf{x}^{(i)})^\top \quad \Big| \quad \text{transpose}$$

$$\left( \sum_{i=1}^n \underbrace{\mathbf{x}^{(i)} (\mathbf{x}^{(i)})^\top}_{\tilde{p} \times \tilde{p}} \right) \theta = \sum_{i=1}^n \mathbf{x}^{(i)} y^{(i)}$$

$$(\mathbf{X}^\top \mathbf{X}) \theta = \mathbf{X}^\top \mathbf{y}$$

$$\text{NB: } \sum_{i=1}^n \mathbf{x}^{(i)} (\mathbf{x}^{(i)})^\top = \mathbf{X}^\top \mathbf{X} \quad \text{is easy to show (try it!) – and good to remember (this is basically the estimation of Cov(X))}$$

$$0 = \frac{\partial \mathcal{R}_{\text{emp}}(\theta)}{\partial \theta} \quad (\text{matrix notation})$$

$$0 = \frac{\partial \|\epsilon\|_2^2}{\partial \theta}$$

$$0 = \frac{\partial \epsilon^\top \epsilon}{\partial \theta} \quad \Big| \quad \text{chain rule}$$

$$0 = \frac{\partial \epsilon^\top \epsilon}{\partial \epsilon} \cdot \frac{\partial \epsilon}{\partial \theta}$$

$$0 = 2\epsilon^\top \cdot (-1 \cdot \mathbf{X})$$

$$0 = (\mathbf{y} - \mathbf{X}\theta)^\top \mathbf{X}$$

$$0 = \mathbf{y}^\top \mathbf{X} - \theta^\top \mathbf{X}^\top \mathbf{X}$$

$$\theta^\top \mathbf{X}^\top \mathbf{X} = \mathbf{y}^\top \mathbf{X} \quad \Big| \quad \text{transpose}$$

$$\mathbf{X}^\top \mathbf{X} \theta = \mathbf{X}^\top \mathbf{y}$$

$$\theta = \underbrace{(\underbrace{\mathbf{X}^\top \mathbf{X}}_{\tilde{p} \times \tilde{p}})^{-1} \underbrace{\mathbf{X}^\top \mathbf{y}}_{\tilde{p} \times n \quad n \times 1}}_{\tilde{p} \times 1}$$

