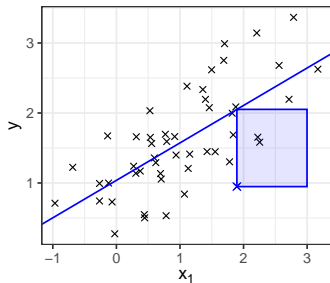# Introduction to Machine Learning

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

$$\hat{\boldsymbol{\theta}} \in \arg\min_{\boldsymbol{\theta}} \mathcal{R}_{\text{emp}}(\boldsymbol{\theta}) = \arg\min_{\boldsymbol{\theta}} \sum_{i=1}^{n} \left(y^{(i)} - \boldsymbol{\theta}^{\top} \mathbf{x}^{(i)} \right)^2$$

$$= \arg\min_{\boldsymbol{\theta}} \| \mathbf{y} - \mathbf{X}\boldsymbol{\theta} \|_2^2$$

- Find via **normal equations**

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

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

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

ANALYTICAL OPTIMIZATION – PROOF

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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