Linear Regression

Cornell CS 3/5780 · Spring 2026

1. Setup and Assumptions

- **Data Assumption**: $y_i \in \mathbb{R}$ (real-valued labels) and $\mathbf{x}_i \in \mathbb{R}^d$
- **Model Assumption**: data looks like a "line" through the origin, with small errors

$$y_i = \mathbf{w}^\top \mathbf{x}_i + \epsilon_i \quad \text{where } \epsilon_i \text{ is small}$$

- **Ordinary Least Squares**: minimize the sum of squared errors

$$\min_{\mathbf{w}} \frac{1}{n} \sum_{i=1}^{n} (\mathbf{x}_i^\top \mathbf{w} - y_i)^2$$

- **MLE connection**: This learning objective is equivalent to MLE (maximizing $P(D|\mathbf{w})$) under the probability model assumption:

$$y_i | \mathbf{x}_i \sim N(\mathbf{w}^\top \mathbf{x}_i, \sigma^2) \quad \Rightarrow \quad P(y_i | \mathbf{x}_i, \mathbf{w}) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(\mathbf{x}_i^\top \mathbf{w} - y_i)^2}{2\sigma^2}}$$

2. Solving the minimization

- **Gradient**: the gradient of the learning objective

$$\nabla_{\mathbf{w}} L(\mathbf{w}) \; = \frac{2}{n} \sum_{i=1}^{n} \mathbf{x}_i (\mathbf{x}_i^\top \mathbf{w} - y_i)$$

- **At optimum**: gradient is equal to zero

$$\frac{2}{n} \sum_{i=1}^{n} \mathbf{x}_i (\mathbf{x}_i^\top \mathbf{w} - y_i) = \mathbf{0} \iff \sum_{i=1}^{n} \mathbf{x}_i \mathbf{x}_i^\top \mathbf{w} = \sum_{i=1}^{n} \mathbf{x}_i y_i$$

- This is a system of linear equations that we need to solve for $\mathbf{w}$
- **Matrix notation**: $\mathbf{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_n] \in \mathbb{R}^{d \times n}$, $\mathbf{y} = [y_1, \ldots, y_n] \in \mathbb{R}^{1 \times n}$

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

3. Solving system of linear equations

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

- If $\mathrm{rank}(\mathbf{X}) = d$ (full rank): then $\mathbf{X}\mathbf{X}^\top$ is invertible
  - Occurs when $n \geq d$ and data spans feature space
  - **Unique solution**: the standard closed-form solution

$$\mathbf{w}^* = (\mathbf{X}\mathbf{X}^\top)^{-1} \mathbf{X}\mathbf{y}$$

- If $\mathrm{rank}(\mathbf{X}) < d$: then $\mathbf{X}\mathbf{X}^\top$ is not invertible
  - Fewer data points than features ($n < d$) OR data lies in lower-dimensional subspace
  - **Infinitely many solutions** achieve the same minimum loss

4. Decomposing the Weight Vector

- **Any weight vector** can be decomposed: $\mathbf{w} = \mathbf{w}_\parallel + \mathbf{w}_\perp$
  - $\mathbf{w}_\parallel$ lies in **column space** of $\mathbf{X}$: $\mathbf{w}_\parallel = \mathbf{X}\mathbf{v}$ for some $\mathbf{v}$
  - $\mathbf{w}_\perp$ lies in **null space** of $\mathbf{X}^\top$: $\mathbf{X}^\top \mathbf{w}_\perp = \mathbf{0}$
  - They are orthogonal: $\mathbf{w}_\parallel^\top \mathbf{w}_\perp = 0$
- **Key observation**: prediction depends only on $\mathbf{w}_\parallel$!

$$\mathbf{x}_i^\top \mathbf{w} = \mathbf{x}_i^\top(\mathbf{w}_\parallel + \mathbf{w}_\perp) = \mathbf{x}_i^\top \mathbf{w}_\parallel + \underbrace{\mathbf{x}_i^\top \mathbf{w}_\perp}_{=0} = \mathbf{x}_i^\top \mathbf{w}_\parallel$$

  Thus, any two weight vectors with the same $\mathbf{w}_\parallel$ achieve the same loss.
- **Solution set** (affine subspace):

$$\{\mathbf{w}_\parallel^* + \mathbf{w}_\perp : \mathbf{w}_\perp \in \mathrm{null}(\mathbf{X}^\top)\}$$

- **Minimum-norm solution**: Unique solution where $\mathbf{w}_\perp = \mathbf{0}$ defined by pseudoinverse

$$\mathbf{w}^*_{\text{min-norm}} = \mathbf{X}^\top(\mathbf{X}\mathbf{X}^\top)^\dagger \mathbf{y}$$

5. (Stochastic) Gradient Descent

- **Gradient descent**: Starting from $\mathbf{w}^{(0)}$, with learning rate $\eta > 0$:

$$\mathbf{w}^{(t+1)} = \mathbf{w}^{(t)} - \eta \frac{2}{n} \sum_{i=1}^{n} \mathbf{x}_i(\mathbf{x}_i^\top \mathbf{w}^{(t)} - y_i)$$

- **SGD**: Approximate gradient using single random data point i

$$\mathbf{w}^{(t+1)} = \mathbf{w}^{(t)} - \eta \cdot 2\mathbf{x}_i(\mathbf{x}_i^\top \mathbf{w}^{(t)} - y_i)$$

- **Key observation**: gradient is in the column space because it is a linear combination of $\mathbf{x}_i$ (columns of the data matrix $\mathbf{X}$)
- (S)GD updates only change $\mathbf{w}_\parallel$, never $\mathbf{w}_\perp$
- (S)GD will **converge to min-norm solution** if initialized at $\mathbf{w}^{(0)} = \mathbf{0}$.

6. Ridge Regression

- **Data & Model Assumption**: Same setup $y_i = \mathbf{w}^\top \mathbf{x}_i + \epsilon_i$, but now we also assume weights are small.
- **Ridge Objective**: minimize squared errors plus a penalty on weight size, where $\lambda > 0$:

$$\min_{\mathbf{w}} \frac{1}{n} \sum_{i=1}^{n} (\mathbf{x}_i^\top \mathbf{w} - y_i)^2 + \lambda \|\mathbf{w}\|_2^2$$

- **MAP connection**: This objective is equivalent to MAP estimation, maximizing $P(\mathbf{w}|D) \propto P(D|\mathbf{w})P(\mathbf{w})$, with the Gaussian prior:

$$P(\mathbf{w}) = \frac{1}{\sqrt{2\pi\tau^2}} e^{-\frac{\mathbf{w}^\top \mathbf{w}}{2\tau^2}} \quad \Rightarrow \quad \lambda = \frac{\sigma^2}{n\tau^2}$$

- **Closed-form solution**: Set gradient to zero and solve (always unique):

$$\mathbf{w}^* = (\mathbf{X}\mathbf{X}^\top + n\lambda \mathbf{I})^{-1} \mathbf{X}\mathbf{y}$$

7. Summary

Ordinary Least Squares (unregularized):

- From MLE with Gaussian noise assumption
- Closed form: $\mathbf{w} = (\mathbf{X}\mathbf{X}^\top)^\dagger \mathbf{X}\mathbf{y}$
- GD from origin $\implies$ minimum-norm solution (implicit regularization!)

Ridge Regression (regularized):

- From MAP with Gaussian prior on weights
- Closed form: $\mathbf{w} = (\mathbf{X}\mathbf{X}^\top + n\lambda \mathbf{I})^{-1} \mathbf{X}\mathbf{y}$
- Always has unique solution, more stable