Gradient Descent for Linear Regression

Semih Cayci

Mathematical Foundations of Deep Learning

RWTH Aachen

Gradient Descent for Linear Regression

Linear Regression Problem

Let $\Phi : \mathbb{X} \to \mathbb{R}^d$ be a given feature mapping. We aim to solve

$$\min_{\theta \in \mathbb{R}^d} \frac{1}{2n} \sum_{j=1}^{n} (\theta^\top \Phi(x_j) - y_j)^2.$$

Assuming realizability: $y_j = \Phi^\top(x_j)\theta^\star$ for some $\theta^\star \in B_2(0, \alpha)$.

Algorithm 1: Projected Gradient Descent

1: $\theta(0) = 0$ — Initialization
2: **for** $t = 0, 1, \ldots, T-1$ **do**
3: $\quad \tilde{\theta}(t+1) = \theta(t) - \eta \cdot \frac{1}{n} \sum_{j=1}^{n} \left(\theta^\top(t)\Phi(x_j) - y_j\right) \cdot \Phi(x_j)$
4: $\quad \theta(t+1) = \Pi_{B_2(0,\rho)}\{\tilde{\theta}(t+1)\}$
5: **end for**

Convergence Analysis

Lyapunov function $\mathcal{L}(\theta) = \|\theta - \theta^\star\|_2^2$. Then, we have:

$$\mathcal{L}(\theta(t+1)) = \|\Pi_{B_2(0,\rho)}\tilde{\theta}(t+1) - \Pi_{B_2(0,\rho)}\theta^\star\|_2^2 \leq \|\tilde{\theta}(t+1) - \theta^\star\|_2^2,$$

$$= \|\theta(t) - \theta^\star\|_2^2 - 2\eta\nabla_\theta g(\theta(t))\Big(\theta(t) - \theta^\star\Big) + \eta^2\|\nabla_\theta g(\theta(t))\|_2^2.$$

Thus, the Lyapunov drift becomes:

$$\mathcal{L}(\theta(t+1)) - \mathcal{L}(\theta(t)) \leq -2\eta\nabla_\theta g(\theta(t))\Big(\theta(t) - \theta^\star\Big) + \eta^2\|\nabla_\theta g(\theta(t))\|_2^2.$$

Convexity: $-2\eta\nabla_\theta g(\theta(t))\Big(\theta(t) - \theta^\star\Big) \leq -2 \cdot \eta \cdot g(\theta(t))$.
Lipschitz continuity:

$$\|\nabla g(\theta(t))\|_2 \leq \frac{1}{n}\Big(\|\theta(t)\|_2\|\Phi(x_j)\|_2 + \alpha\|\Phi(x_j)\|_2\Big)\|\Phi(x_j)\|_2 \leq (\alpha + \rho).$$

Convergence Analysis

Then,

$$\mathcal{L}(\theta(t+1)) - \mathcal{L}(\theta(t)) \leq -2\eta g(\theta(t)) + \eta^2(\alpha+\rho)^2.$$

By telescoping sum over $t = 0, 1, \dots, T-1$:

$$\mathcal{L}(\theta(T)) - \mathcal{L}(\theta(0)) \leq -2\eta \sum_{t<T} g(\theta(t)) + \eta^2 T(\alpha+\rho)^2.$$

Rearranging terms:

$$\min_{0\leq t<T} g(\theta(t)) \leq \frac{1}{T}\sum_{t<T} g(\theta(t)) \leq \frac{\mathcal{L}(\theta(0))}{2\eta T} + \frac{\eta(\alpha+\rho)^2}{2}.$$

$\mathcal{L}(\theta(0)) = \|\theta(0) - \theta^\star\|_2^2 \leq \alpha^2$. Then, choosing $\eta = 1/\sqrt{T}$,

$$\min_{0\leq t<T} g(\theta(t)) \leq \frac{\alpha^2 + (\alpha+\rho)^2}{2\sqrt{T}}.$$