Problem sheet 4: Stochastic gradient descent

Analisi dei Dati – CdS Matematica – 2024/25

When suitable, please provide summarising and explanatory pictures.

Exercise 4.1. Consider the optimization of linear regression through the mean square error,

$$f(\theta) = |X\theta - Y|^2,$$

in the homogeneous formulation, where X is the input dataset, and Y is the output. Let

$$\begin{cases} \theta_{t+1} = \theta_t - \eta\nabla f(\theta_t), \\ \theta_0 = a_0, \end{cases}$$

be the gradient descent iterations. Prove that

- $(f(\theta_t))_{t\geqslant 0}$ is non-increasing for $\eta > 0$ small enough,
- deduce that $|\theta_{t+1} - \theta_t| \longrightarrow 0$,
- deduce that $X^T(X\theta_t - Y) \longrightarrow 0$,
- if X^TX is invertible, conclude that $\theta_t \longrightarrow (X^TX)^{-1}X^TY$.

Exercise 4.2. (overparametrised case) In the same framework of previous exercise, consider the overparametrised case (namely, the number of features is larger than the sample size). In particular, $A = X^TX$ cannot be invertible. Denote by $A^\dagger$ the *Moore-Penrose pseudo-inverse* of A.

- If $\bar\theta = A^\dagger X^TY$, prove that $A\bar\theta = X^TY$.
- Prove that $(I - AA^\dagger)\theta_t$ is constant in t.
- Conclude that $\theta_t \longrightarrow (I - AA^\dagger)a_0 + \bar\theta$.

Exercise 4.3. Consider a linear regression problem with squared loss function. Implement stochastic gradient descent with

- a mini-batch of size 1,
- a larger and larger mini-batch size,
- the whole sample.

Evaluate the result in terms of effectiveness and number of iterations.

Exercise 4.4. Repeat problem 4.3 with the additional condition on vanishing learning rate (that is $\eta \to 0$).