Tutorium 5

1. Februar 2023 - Solution

5.1 Stochastic Optimization

5.1.1 Setting

Let us consider a data set $\{(x_i, y_i)\}_{i=1}^n \subset \mathbb{R}^d \times \mathbb{R}$ with $x_i \neq x_j$ for $i \neq j$. We have d-dimensional vectors as input values and outputs taking real values. We now assume this given data correlates to a linear mapping. Therefore we need to find $w \in \mathbb{R}^d$ such that

$$x_i^\top w \approx y_i \text{ for } i = 1, ..., n. \tag{5.1}$$

To obtain a classifier, we want to solve the following minimization problem

$$\min_{w \in \mathbb{R}^d} f(w) = \min_{w \in \mathbb{R}^d} \frac{1}{n} \sum_{i=1}^n \frac{1}{2} (x_i^\top w - y_i)^2 + \frac{\lambda}{2} \|w\|_2^2 \tag{5.2}$$

with $\lambda > 0$ the regularization parameter. With the matrix notation

$$X = [x_1, ..., x_n] \in \mathbb{R}^{d \times n},\ y = [y_1, ..., y_n] \in \mathbb{R}^n$$

we can rewrite the objective function in 5.2 as

$$f(w) = \frac{1}{2n} \|X^\top w - y\|_2^2 + \frac{\lambda}{2} \|w\|_2^2. \tag{5.3}$$

In order to be able to use stochastic gradient descent, a suitable representation of (5.2) must be found. Let

$$A = \frac{1}{n} X X^\top + \lambda I \in \mathbb{R}^{d \times d} \text{ and } b = \frac{1}{n} X y$$

and rewrite the problem $Aw = b$ as a linear least squares problem

$$\min_{w \in \mathbb{R}^d} g(w) = \min_{w \in \mathbb{R}^d} \frac{1}{2} \|Aw - b\|_2^2 = \min_{w \in \mathbb{R}^d} \sum_{j=1}^d p_j g_j(w) \tag{5.4}$$

with $g_j(w) = \frac{1}{2p_j} (A_{j:} w - b_j)^2$, where $A_{j:}$ denotes the j-th row of A and $p_j = \frac{\|A_{j:}\|_2^2}{\|A\|_F^2}$ for $j = 1, ..., d$ with $\|A\|_F^2 = tr(A^\top A)$ denotes the Frobenius norm of A. Assume that $\bar{w}$ is a solution of (5.2) and $\hat{w}$ is a solution of (5.4). Note that $A\hat{w} = b$.

Exercise 1

Show that

$$\nabla g_j(w) = \frac{1}{p_j} A_{j:}^\top A_{j:} (w - \hat{w}) \tag{5.5}$$

and that

$$\mathbb{E}_{j\sim p}[\nabla g_j(w)] := \sum_{i=1}^{d} p_i \nabla g_i(w) = A^\top A(w - \hat{w}) \tag{5.6}$$

thus $\nabla g_j(w)$ is an unbiased estimator of the full gradient of the objective function in (5.4). This justifies applying the stochastic gradient method.

Proof. □

Exercise 2

Show that the solution $\bar{w}$ of (5.2) and the solution $\hat{w}$ of (5.4) are equal.

Proof. □

From a given $w^0 \in \mathbb{R}^d$, consider the iterates

$$w^{k+1} = w^k - \alpha_k \nabla g_j(w^k), \tag{5.7}$$

where

$$\alpha_k = \frac{1}{\|A\|_F^2} \tag{5.8}$$

with j is a random index chosen from $\{1, ..., d\}$ sampled with probability p_j. In other words, $\mathbb{P}(j = i) = p_i$ for $i = 1, ..., d$.

Exercise 3

Define $\Pi_j = \frac{A_{j:}^\top A_{j:}}{\|A_{j:}\|_2^2}$ and show that

1. $\Pi_j \Pi_j = \Pi_j$
2. $(I - \Pi_j)(I - \Pi_j) = I - \Pi_j$

hold.

Proof. □

Exercise 4

Show that the distance to the solution satisfies the following recurrence

$$\|w^{k+1} - \bar{w}\|_2^2 = \|w^k - \bar{w}\|_2^2 - \left\langle \frac{A_{j:}^\top A_{j:}}{\|A_{j:}\|_2^2}(w^k - \bar{w}), w^k - \bar{w} \right\rangle.$$

Proof. □