TD 7 – Coordinate descent

Mathematics of data

20/11/24

Exercise 1. Let $f\colon \mathbb{R}^p \to \mathbb{R}$. Coordinate descent is an optimization method which tries to minimize f alternatively with respect to individual coordinates. We denote w^t the iterates. At iteration t, we choose an index $i \in \{1, \dots, p\}$ and we try to minimize f with respect to its i-th coordinate without changing the other coordinates w_j^t, $j \neq i$. More formally, we define $\phi_i(x, w) = f(w_1, \dots, w_{i-1}, x, w_{i+1}, \dots, w_p)$ and set at each iteration:

$$w_i^{t+1} = \operatorname{argmin}_x \phi_i(x, w^t) \quad \text{and} \quad w_j^{t+1} = w_j^t \text{ for } j \neq i.$$

The index i is typically chosen as cyclic: $i = 1 + (t \mod p)$.

The aim of this exercise is to prove a convergence rate for coordinate descent on a quadratic function

$$f(w) = \frac{1}{2}\langle w, Aw\rangle - \langle b, w\rangle,$$

where $A \in \mathbb{R}^{p\times p}$ is a positive definite symmetric matrix.

1. Assume that we optimize the coordinate i at step $t+1$. Compute the update rule of w_i^t.
2. Show that
$$f(w^{t+1}) - f(w^t) = -\frac{(Aw^t - b)_i^2}{2A_{ii}} \leq -\frac{(Aw^t - b)_i^2}{2A_{\max}},$$
where $A_{\max} = \max_i A_{ii}$.
3. Assume that we do a *greedy* coordinate descent, which means that at iteration $t+1$, we update the coordinate i such that $(Aw^t - b)_i^2$ is maximal. Show that
$$f(w^{t+1}) - f(w^t) \leq -\frac{\|Aw^t - b\|^2}{2pA_{\max}}.$$
4. Let $w^* = A^{-1}b$. Demonstrate that
$$\|Aw - b\|^2 \geq 2\sigma_{\min}(A)(f(w) - f(w^*)).$$
Provide a convergence rate for the coordinate descent method. What is the difference with gradient descent? When is it faster? Slower?

Exercise 2.

1. Given a convex, differentiable map $f\colon \mathbb{R}^p \to \mathbb{R}$, if we are at a point $x = (x_1, \dots, x_p)$ such that $f(x)$ is minimized along each coordinate axis, i.e.
$$f(x) = \min_{z\in\mathbb{R}} f(x_1, \dots, x_{i-1}, z, x_{i+1}, \dots, x_p),$$
for $i = 1, \dots, p$, have we found a *global minimizer*?
2. Same question, but for f convex (not necessarily differentiable)?