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 $i \in \{1, \ldots, p\}$ 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 $\phi_i(x, w) = f(w_1, \ldots, w_{i-1}, x, w_{i+1}, \ldots, w_p)$ 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 $x = (x_1, \ldots, x_p)$ such that $f(x)$ is minimized along each coordinate axis, i.e.
$$f(x) = \min_{z\in\mathbb{R}} f(x_1, \ldots, x_{i-1}, z, x_{i+1}, \ldots, x_p),$$
for $i = 1, \ldots, p$, have we found a *global minimizer*?

2. Same question, but for f convex (not necessarily differentiable)?

Solutions

Exercise 1.

1. As $\phi_i(\cdot, w^t)$ is convex and coercive, it is minimized at x such that $\partial_x \phi(x, w^t) = 0$. We have

$$\partial_x \phi(x, w^t) = (\nabla f(w_1^t, \dots, x, \dots, w_p^t))_i = \sum_{j \neq i} A_{ij} w_j^t + A_{ii} x - b_i,$$

so that

$$w_i^{t+1} = \frac{1}{A_{ii}}(b_i - \sum_{j \neq i} A_{ij} w_j^t) = w_i^t + \frac{1}{A_{ii}}(b_i - \sum_{j=1}^p A_{ij} w_j^t).$$

2. We have $f(w^{t+1}) - f(w^t) = \phi_i(w_i^{t+1}, w^t) - \phi_i(w_i^t, w^t)$. The map $\phi(\cdot, w^t)$ is a quadratic function in dimension 1, so we can write it

$$\phi(x, w^t) = \alpha x^2 + \beta x + \gamma = \alpha(x - \frac{\beta}{2\alpha})^2 - \frac{\beta^2}{2\alpha}.$$

By definition, we have $w_i^{t+1} = \frac{\beta}{2\alpha}$ and $\phi_i(w_i^{t+1}, w^t) = -\frac{\beta^2}{2\alpha}$. Moreover, $\alpha = A_{ii}/2$, so that

$$f(w^{t+1}) - f(w^t) = -\frac{A_{ii}}{2}(w_i^{t+1} - w_i^t)^2,$$

which leads to the desired result. The bound then comes from $A_{ii} \leq A_{\max}$.

3. It is simply the bound $\max_j (Aw^t - b)_j^2 \geq \|Aw^t - b\|^2 / p$.

4. We have $Aw - b = A(w - w^*)$. Then

$$\|Aw - b\|^2 = \langle A(w - w^*), A(w - w^*) \rangle \geq \sigma_{\min}(A) \langle A(w - w^*), w - w^* \rangle = 2\sigma_{\min}(A)(f(w) - f(w^*)).$$

This leads to

$$f(w^{t+1}) - f(w^t) \leq -\frac{\sigma_{\min}(A)}{pA_{\max}}(f(w^t) - f(w^*)).$$

Subtracting $f(w^*)$ to both sides of the inequality gives

$$f(w^{t+1}) - f(w^*) \leq (1 - \frac{\sigma_{\min}(A)}{pA_{\max}})(f(w^t) - f(w^*)),$$

and finally

$$f(w^t) - f(w^*) \leq (1 - \frac{\sigma_{\min}(A)}{pA_{\max}})^t (f(w^0) - f(w^*)).$$

We can compare this to the convergence rate of gradient descent:

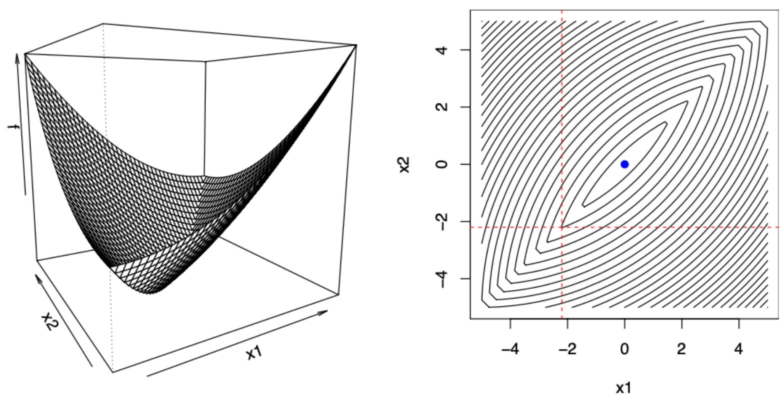
$$f(w^t) - f(w^*) \leq (1 - \frac{\sigma_{\min}(A)}{\sigma_{\max}(A)})^{2t} (f(w^0) - f(w^*))$$

(see for example Francis Bach's lecture notes `https://www.di.ens.fr/~fbach/learning_theory_class/lecture4.pdf`).

Exercise 2.

1. Yes, because $\nabla f(x) = (\partial_{x_i} f(x))_i = 0$.

2. No. Take for instance $f(x) = \|x\|_\infty$ with $x = (1, \dots, 1)$. Another example is given by the figure below (from `https://www.stat.cmu.edu/~ryantibs/convexopt/lectures/coord-desc.pdf`).



f
x2
x1
4
2
0
−2
−4
−4
−2
0
2
4
x1
x2