Optimization for datascience exercises

December 9, 2024

Ex. 1 — We consider samples $x_1, \ldots, x_n \in \mathbb{R}^p$ and targets $y_1, \ldots, y_n \in \mathbb{R}$. We define the scalar function $\phi(u) = \sqrt{u^2 + 1}$ and consider the optimization problem

$$\min_\beta f(\beta) = \frac{1}{n} \sum_{i=1}^{n} \phi(\langle x_i, \beta \rangle - y_i)$$

Mark as true or false.

1. This loss function corresponds to a classification problem
2. The function ϕ is such that $\phi''(u) \in [0, 1]$ for all u.
3. The function f is convex.
4. The function f is strongly convex
5. The function f is smooth
6. The function f is $\frac{1}{n} \sum_{i=1}^{n} \|x_i\|^2-$smooth

Ex. 2 — Let A a symmetric, positive $d \times d$ matrix, b a vector of size d and define

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

We consider the iterates of gradient descent with a step size α^t, starting form $w^0 = 0$.

$$w^{t+1} = w^t - \alpha^t \nabla f(w^t)$$

1. We take α^t that minimizes $f(w^{t+1})$. What is the value of $\langle \nabla f(w^{t+1}), \nabla f(w^t) \rangle$?
2. What is the step size that minimizes $f(w^{t+1})$?
3. We take for all t the step size α^t that minimizes $f(w^{t+1})$. With that choice of step size, does gradient descent converges to the solution in a finite number of iterations ?
4. Is there a sequence of step sizes α^t such that gradient descent converges in $d + 1$ iterations ?

Ex. 3 — Consider the problem given by

$$w^* \in \arg\min_{w\in\mathbb{R}^d} f(w) = \frac{1}{n}\sum_{i=1} f_i(w), \tag{1}$$

where f_i is L–smooth for $i = 1, \dots, n,$. We assume that f is $\mu-$strongly convex, let w^* its minimizer, and suppose that we have **for all i,** $\nabla f_i(w^*) = 0$.
The iterates of the SGD (stochastic gradient descent) method with constant step size are given by

$$w^{t+1} = w^t - \alpha\nabla f_{i_t}(w^t), \tag{2}$$

where $\alpha > 0$ is the step size and $i_t \in \{1, \dots, n\}$ is chosen i.i.d with uniform probability at each iteration.

1. Show that we have for all w:

$$\|\nabla f_i(w)\| \le L\|w - w^*\|$$

2. Demonstrate

$$\mathbb{E}_{i_t}\left[\|w^{t+1} - w^*\|^2\right] \le (1 - 2\alpha\mu + \alpha^2 L^2)\|w^t - w^*\|^2$$

   where the expectation is taken with respect to the random index i_t. **Hint:** you can show that for all w, we have $\langle\nabla f(w), w - w^*\rangle \ge \mu\|w - w^*\|^2$.

3. What is the value of α that gives the fastest convergence rate ? What type of bound on $\|w^t - w^*\|^2$ do we get?

4. What convergence regime do you get ? Is this surprising considering the behavior of SGD seen in class? Comment.

Ex. 4 — We let $f : \mathbb{R}^p \to \mathbb{R}$. Coordinate descent tries to minimize f alternatively with respect to individual coordinates.
We denote w^t the iterates. At iteration t, we chose an index $i \in \{1, \dots, p\}$ and try to minimize f with respect to w_i^t without changing the other coordinates w_j^t, $j \ne i$. More formally, we define $\phi_i(x, w) = f(w_1, \dots, w_{i-1}, x, w_{i+1}, w_p)$ and set at each iteration:

$$w_i^{t+1} = \arg\min_x \phi_i(x, w^t) \;\text{ and }\; w_j^{t+1} = w_j^t \;\text{ for }\; j \ne i$$

The index i is typically chosen as cyclic : $i = 1 + (t \mod p)$. Therefore , at iteration 1, the coordinate 1 is updated, at iteration 2, the coordinate 2 is updated, ... , at iteration p the coordinate p is updated and at iteration $p + 1$ the coordinate 1 is modified again.

1 Assume that f is the quadratic function:

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

Compute the update rule to minimize ϕ_i.

2 At iteration $t+1$, we update the coordinate i. Demonstrate 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 At iteration t, the coordinate that is updated is 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, or slower? Hint: what is the link between A_{max} and $\sigma_{\max}(A)$?