Optimization for Data Science
June 4, 2019

1. (6 POINTS) Describe in depth the stochastic gradient approach and explain why it is widely used in data science.

   Solution 1. See Notes Section 4.7.

2. (7 POINTS) Describe in depth block coordinate gradient descent with Gauss-Southwell rule and randomized block coordinate gradient descent method, highlighting the differences between the two methods.

   Solution 2. See Notes Section 4.5.

3. (7 POINTS) Consider the ridge regression problem:

$$\min_{w \in \mathbb{R}^d} f(w) = \frac{1}{2n}\|X^\top w - y\|_2^2 + \frac{\lambda}{2}\|w\|_2^2,$$

   Describe the gradient of the function and prove that there exists a matrix C s.t. $\nabla f(w) = C(w - w^*)$, with w^* solution of the problem.

   Solution 3. We first calculate the gradient of f, that is:

$$\nabla f(w) = \frac{1}{n} X(X^\top w - y) + \lambda I w.$$

   Since $\nabla f(w^*) = 0$, we can write

$$\frac{1}{n} X(X^\top w^* - y) + \lambda I w^* = 0,$$

   thus getting

$$(\frac{1}{n} X X^\top + \lambda I) w^* = \frac{1}{n} X y.$$

   If we call $C = (\frac{1}{n} X X^\top + \lambda I)$, we have

$$\nabla f(w) = \frac{1}{n} X(X^\top w - y) + \lambda I w = Cw - \frac{1}{n} Xy = Cw - Cw^* = C(w - w^*).$$

4. (8 POINTS)Consider the two sets:

   - $D_1 = \{x \in \mathbb{R}^n : \ a \leq x \leq b, \}$, with $a, b \in \mathbb{R}^n$ and $a < b$;
   - $D_2 = \{x \in \mathbb{R}^n : \|x\|_1 \leq \tau\}$, with $\tau > 0$.

   Describe how to calculate the Frank-Wolfe direction for the problem

$$\min_{x \in D_1} f(x).$$

   Furthermore give the computational cost. Finally, consider problem

$$\min_{x \in D_2} f(x),$$

   and the origin point $x_k = 0$. Calculate the Away-Step direction in x_k choosing the minimal representation w.r.t. S_k (i.e., the one with the smallest number of nonzeroes) and the related maximum stepsize.

Solution 4. At each iteration, the Frank-Wolfe direction is

$$d_k = \hat{x}_k - x_k.$$

Hence, in order to get $\hat{x}_k$, we need to solve problem:

$$\min_{x \in D_1} \nabla f(x_k)^\top x.$$

- $D_1 = \{x \in \mathbb{R}^n : \ a \le x \le b, \}$, with $a \in \mathbb{R}^n, \ b \in \mathbb{R}^n, \ a < b$. In this case we have a simple box. Taking into account the fundamental theorem of linear programming, we can focus on the vertices of the polytope D_1 and choose the best one according to the following rule:

  $$(\hat{x}_k)_i = \begin{cases} b_i & \text{if } \nabla_i f(x_k) < 0 \\ a_i & \text{otherwise.} \end{cases}$$

  So, the solution of the Frank-Wolfe problem has a cost $\mathcal{O}(n)$.

- In this case, we need to build the Away-Step direction. Hence, we consider direction

  $$d_k = x_k - \hat{x}_k.$$

  At each iteration k of the Away-Step Frank-Wolfe method, in order to get $\hat{x}_k$, we need to solve problem:

  $$\max_{x \in S_k} \nabla f(x_k)^\top x,$$

  with S_k set of vertices whose convex combination gives $x_k = 0$. Keeping in mind that $D_2 = conv\{\pm\tau e_i, \ i = 1, \dots, n\}$, and that we need to use the minimal representation, we can choose any index $i = 1, \dots, n$ and assume that

  $$0 = x_k = \frac{\tau}{2}(e_i - e_i).$$

  The direction will thus be

  $$d_k = x_k - \hat{x}_k = -(sign(\nabla_i f(x_k)))\tau e_i,$$

  and the maximum stepsize is $\alpha_{max} = 1$.

5. (4 POINTS) Calculate the condition number[1] of the matrix C obtained at exercise 3 and explain what happens when $\lambda \to 0$ and $\lambda \to \infty$.

   Solution 5. It is easy to see that $\sigma_{max}(C) = \sigma_{max}(X)^2 + \lambda$ and $\sigma_{min}(C) = \sigma_{min}(X)^2 + \lambda$. Hence,

   $$\lim_{\lambda \to \infty} \kappa(C) = 1$$

   and

   $$\lim_{\lambda \to 0} \kappa(C) = \kappa(X)^2.$$

[1] Keep in mind that the condition number is the ratio $\kappa(C) = \frac{\sigma_{max}(C)}{\sigma_{min}(C)}$ and that

$$\sigma_{max}(C) = \|C\|_2 = \max_{x \neq 0} \frac{\|Cx\|_2}{\|x\|_2} \ \text{and} \ \sigma_{min}(C) = \min_{x \neq 0} \frac{\|Cx\|_2}{\|x\|_2}.$$