Optimization for Data Science
September 20, 2019

1. (6 POINTS) Describe in depth the stochastic variance reduced gradient methods.

   Solution 1. See Notes Section 4.7.

2. (7 POINTS) Describe in depth the Frank-Wolfe method and its main variants.

   Solution 2. See Notes Section 5.3.

3. (8 POINTS) Given the following problem:

$$\min_{x\in\Delta} \frac{1}{2}x^\top Qx + c^\top x,$$

   with $\Delta = \{x \in \mathbb{R}^n : e^\top x = 1,\ x \geq 0\}$ and $Q \in \mathbb{R}^{n\times n}$ positive definite matrix. Calculate the computational cost of performing the exact line search at a point $x_k \in \Delta$ when using the Pairwise Frank-Wolfe direction (assume that the gradient is given).

   Solution 3. We notice that, in this specific case, the Pairwise Frank-Wolfe direction is

$$\hat{x}_k = e_{i_k} - e_{j_k},$$

   with $i_k = \arg\min_i \nabla_i f(x_k)$ and $j_k = \arg\max_{i\in S^k} \nabla_i f(x_k)$, and $S^k = \{i : \ x_i^k > 0\}$. The exact line search is given as:

$$\min_{\alpha\in(0,1]} \phi(\alpha) = f(x_k + \alpha d_k).$$

   Taking into account its properties, we only need to solve the following equation

$$\nabla f(x_k + \alpha d_k)^\top d_k = 0.$$

   Now, since $\nabla f(x_k) = Qx_k + c$, we hence obtain,

$$[Q(x_k + \alpha d_k) + c]^\top d_k = 0.$$

   If we suitably rewrite last equation, we get

$$\alpha = -\frac{(Qx_k + c)^\top d_k}{d_k^\top Q d_k} = -\frac{\nabla f(x_k)^\top d_k}{d_k^\top Q d_k}.$$

   Finally, replacing $d_k = e_{i_k} - e_{j_k}$, we obtain

$$\alpha = -\frac{\nabla_{i_k} f(x_k) - \nabla_{j_k} f(x_k)}{Q_{i_k i_k} - 2Q_{i_k j_k} + Q_{j_k j_k}},$$

   and we get

$$\alpha_{ex} = \min\{\alpha, 1\}.$$

   This has $\mathcal{O}(1)$ cost.

4. (7 POINTS) Consider the Boosting problem and explain PROs and CONs of using the Gradient method for solving it and calculate the gradient at a point w_k. Suggest an alternative solution method (please motivate the answer).

   Solution of 4. We can write the boosting problem as follows:

$$\min_{w \in R^l} f(w) = \sum_{i=1}^{m} \frac{e^{-\rho(Aw)^i y^i}}{m}, \tag{1}$$

   where m is the number of samples in our training set and l is the number of classifiers. The gradient at a point w^k is

$$\nabla f(w^k) = -\rho \sum_{i=1}^{m} y^i \frac{e^{-\rho(Aw_k)^i y^i}}{m} A^i, \tag{2}$$

   where $A^i \in \mathbb{R}^l$ is the i-th row of the matrix A. Thanks to the properties of the problem (See Chapter 4 for further details), we have that the gradient method has a high rate of convergence in this case. When number of points in the training and/or number of classifiers are huge, then calculating the gradient might be costly. In this case it is possible to use some method that tries to exploit the structure of the problem. We can use a BCGD method when number of classifiers is large. While we might consider a stochastic-like gradient in case we have a significant number of training samples.

5. (4 POINTS) Let $f : \mathbb{R}^n \to \mathbb{R}$ be continuously differentiable on $\mathbb{R}^n$ and bounded from below. Let $\{x_k\}$ be an infinite sequence such that for all k we have

   (a)
$$\nabla f(x_k)^\top d_k < 0;$$

   (b)
$$f(x_{k+1}) = f(x_k + \alpha_k d_k) \le f(x_k) + \gamma \alpha_k \nabla f(x_k)^\top d_k$$
   with $\gamma \in (0, 1)$;

   (c) there exists a value $\mu > 0$ such that
$$\alpha_k \ge \mu \frac{|\nabla f(x_k)^\top d_k|}{\|d_k\|^2}.$$

   Prove that

$$\sum_{k=0}^{\infty} \left(\frac{\nabla f(x_k)^\top d_k}{\|d_k\|} \right)^2 < \infty.$$

   Solution 5. Taking into account the conditions given above, we can write

$$f(x_k) - f(x_{k+1}) \ge \gamma \mu \frac{|\nabla f(x_k)^\top d_k|^2}{\|d_k\|^2}.$$

   By applying the inequality multiple times, we have

$$\sum_{k=0}^{m} (f(x_k) - f(x_{k+1})) \ge \gamma \mu \sum_{k=0}^{m} \frac{|\nabla f(x_k)^\top d_k|^2}{\|d_k\|^2}.$$

Since f is bounded from below, we can write

$$\sum_{k=0}^{m}(f(x_k) - f(x_{k+1})) = f(x_0) - f(x_{m+1}) \leq M.$$

Hence, when $m \to \infty$ we have

$$\sum_{k=0}^{\infty} \frac{|\nabla f(x_k)^\top d_k|^2}{\|d_k\|^2} \leq \frac{M}{\gamma\mu} < \infty,$$

which completes the proof.