Solutions
Optimization for Data Science
June 4, 2018

1. (7 POINTS) Describe the gradient method and prove that it converges at a linear rate in the strongly convex case.

   Solution 1. See Notes Section 4.3.1.

2. (6 POINTS) Describe in depth the stochastic gradient approaches that exploit variance reduction and explain the reasons why those methods work better than the classic stochastic gradient.

   Solution 2. See Notes Section 4.7.5.

3. (7 POINTS) Consider the problem

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

   with f continuously differentiable and convex function and

$$\Delta = \{x \in \mathbb{R}^n : e^\top x = 1,\ x \geq 0\}.$$

   Calculate the maximum stepsize that can be taken at a point x_k along the away-step direction (i.e., d_k^{AS}) and the pairwise Frank Wolfe direction (i.e., $d_k^{PWFW} = d_k^{FW} + d_k^{AS}$).

   Solution 3. We need to choose in both case a stepsize

$$\bar{\alpha} = \max_{\beta}\{x_k + \beta d_k \in \Delta\}$$

   . In the away-step case the direction

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

   Since the feasible set is Δ, we have that the vertices are $e_i,\ i = 1, \ldots, n$, hence

$$d_k^{AS} = x_k - e_{\hat{\imath}},$$

   where $\hat{x}_k^{AS} = e_{\hat{\imath}}$. Thus we get:

$$x_{k+1} = x_k + \bar{\alpha}(x_k - e_{\hat{\imath}}) = (1 + \bar{\alpha})x_k - \bar{\alpha}e_{\hat{\imath}}.$$

   Now, we need to choose $\bar{\alpha}$ that guarantees feasibility of the iterate, that is

$$x_{k+1} \in \Delta.$$

   First we notice that

$$e^\top x_{k+1} = e^\top\left[(1 + \bar{\alpha})x_k - \bar{\alpha}e_i\right] = 1.$$

   Thus we need to check that

$$x_{k+1} \geq 0.$$

   Due to the way x_{k+1} is defined, the only component that might become negative is $\hat{\imath}$. We then determine the maximum stepsize by solving this simple equation:

$$(1 + \bar{\alpha})x_k^{\hat{\imath}} - \bar{\alpha} = 0,$$

where $x_k^{\hat{\imath}}$ is the $\hat{\imath}$-th component of x_k. Hence we can write

$$\bar{\alpha} = \frac{x_k^{\hat{\imath}}}{1 - x_k^{\hat{\imath}}}.$$

Since the pairwise Frank Wolfe direction is given as $d_k^{PWFW} = d_k^{FW} + d_k^{AS}$, by taking into account the fact that the feasible set is Δ, we get

$$d_k^{PWFW} = d_k^{FW} + d_k^{AS} = e_{\hat{\jmath}} - e_{\hat{\imath}}.$$

Following the same reasoning as before, we get

$$\bar{\alpha} = x_k^{\hat{\imath}}.$$

4. (8 POINTS) Consider the mean-risk problem:

$$\min_{x \in \Delta} \quad \gamma\sqrt{x^\top M x} - c^\top x$$

where $M \in \mathbb{R}^{n \times n}$ is a symmetric positive definite matrix, $c \in \mathbb{R}^n$, $\gamma > 0$ is the risk-aversion parameter and

$$\Delta = \{x \in \mathbb{R}^n : e^\top x = 1,\ x \geq 0\}.$$

Analyze in depth its properties. Then describe a method for finding an optimal solution and properly justify the choice.

Solution 4. We first need to analyze the properties of the problem. The feasible set is a polytope (it is hence convex and compact). The objective function is given by the sum of the term $f_1(x) = \gamma\sqrt{x^\top M x}$ and $f_2(x) = -c^\top x$. First term is the square root of a quadratic form. This function is continuous, but non-smooth when $x = 0$. Anyway, thanks to the shape of our feasible set, f_1 is continuously differentiable over Δ. It is easy to see that f_1 is convex. The second term is linear, thus continuously differentiable. We hence want to minimize a convex (and continuously differentiable over Δ) function over a unit simplex. We can easily use a first order method to solve the problem. A good choice might be using any Frank-Wolfe variant (e.g. Pairwise Frank-Wolfe). In this case, thanks to the sparsity of the direction, the cost per iteration would be quite small (i.e., $\mathcal{O}(n)$).

5. (8 POINTS) Consider the problem

$$\min_{x \in \mathbb{R}^n} f(x) = \frac{1}{2} x^\top Q x + c^\top x,$$

with $Q \in \mathbb{R}^{n \times n}$ symmetric positive definite matrix and $c \in \mathbb{R}^n$. Let $\gamma_1, \ldots, \gamma_n$ be the eigenvalues of the matrix Q. Prove that the gradient method defined as follows:

$$x_{k+1} = x_k - \frac{1}{\gamma_k} \nabla f(x_k),$$

with $k \geq 1$, converges in at most n iterations to the solution of the problem.

Solution 5. Taking into account the way we calculate the gradient, we get

$$\nabla f(x_{k+1}) = Q x_{k+1} + c = Q x_k + c - Q \frac{1}{\gamma_k} \nabla f(x_k) = (I - \frac{1}{\gamma_k} Q) \nabla f(x_k).$$

By using induction, we get

$$\nabla f(x_k) = \left(\prod_{j=1}^{k-1} (I - \frac{1}{\gamma_j} Q) \right) \nabla f(x_1).$$

Now, we consider n linearly independent eigenvectors $v_i, i = 1, \ldots, n$ of Q, such that

$$Qv_i = \gamma_i v_i.$$

Those vectors represent a basis in $\mathbb{R}^n$, thus we can write

$$\nabla f(x_1) = \sum_{l=1}^{n} \xi_l v_l.$$

For each $k \geq 2$ we then have

$$\nabla f(x_k) = \left(\prod_{j=1}^{k-1} (I - \frac{1}{\gamma_j} Q) \right) \sum_{l=1}^{n} \xi_l v_l,$$

When $k = n + 1$, using the property of eigenvectors, we finally get

$$\nabla f(x_{n+1}) = \sum_{l=1}^{n} \xi_l \left(\prod_{j=1}^{n} (1 - \frac{1}{\gamma_j} \gamma_l) \right) v_l = 0.$$