This homework is due at **0am, January 9, 2019**.

Problem 1

(20 points) Let $x_1, \dots, x_p \in \mathbb{R}^n$, and let $X \in \mathbb{R}^{n \times p}$ be the matrix whose j-th column is given by x_j. Let $y \in \mathbb{R}^n$. Suppose that we would like to find a sparse vector $\hat{\beta} \in \mathbb{R}^p$, such that $y \approx X\hat{\beta}$. A modified *forward stagewise regression method* iterates as follows: Let $\beta_0 \in \mathbb{R}^p$ be the all-zero vector, and $r_0 = y$. For every $t \in \{0\} \cup \mathbb{N}$, compute

$$j_t \in \operatorname*{argmax}_j \left\{ |\langle r_t, x_j \rangle| \;\middle|\; j \in \{1, \dots, p\} \right\},$$

$$r_{t+1} \leftarrow r_t - \tau_t \left[\operatorname{sign}(\langle r_t, x_{j_t} \rangle) x_j + \frac{1}{C}(r_t - y) \right],$$

$$\beta_{t+1}^{(j)} \leftarrow \begin{cases} \left(1 - \frac{\tau_t}{C}\right) \beta_t^{(j)} + \tau_t \operatorname{sign}(\langle r_t, x_{j_t} \rangle) & \text{, if } j = j_t, \\ \left(1 - \frac{\tau_t}{C}\right) \beta_t^{(j)} & \text{, otherwise,} \end{cases}$$

for some $\tau_t \in]0, C[$ and $C > 0$. When there are multiple maximizers in the optimization problem defining j_t, we arbitrarily choose only one of them.

The vectors r_t are called the *residual*. The method is perhaps most understandable when we set $C \to +\infty$: Notice that if $y = X\beta$ exactly for some $\beta \in \mathbb{R}^p$, then y is a linear combination of $x_1, \dots, x_p$. In each iteration, the method finds the component x_{j_t} that contributes the most to the residual, removes the effect of x_{j_t} to the residual, and then adds x_{j_t} to β_{t+1} with a proper scaling.

Show that the iteration rule of the modified forward stagewise regression method is equivalent to that of the Frank-Wolfe algorithm applied to compute the lasso. The lasso is given by

$$\beta^\star \in \operatorname*{argmin}_\beta \left\{ \frac{1}{2} \| y - X\beta \|_2^2 \;\middle|\; \beta \in \mathbb{R}^p, \|\beta\|_1 \le C \right\},$$

where C is the same as in the modified forward stagewise regression method.

Problem 2

(20 points) Consider the optimization problem

$$f^\star = \min_x \left\{ f(x) \;\middle|\; x \in \mathbb{R}^{d \times d}, x \ge 0, \operatorname{Tr}(x) = 1 \right\},$$

for some smooth (with respect to the Frobenius norm) convex function f. Let $\varepsilon > 0$. Suppose that we would like to find an ε-approximate solution x_ε, such that

$$f(x_\varepsilon) - f^\star \le \varepsilon.$$

In [1], it was proved that any such x_ε must satisfy

$$\operatorname{rank}(x_\varepsilon) = \Omega\left(\frac{1}{\varepsilon}\right).$$

Show that this lower bound is tight. That is, show that there exists an ε-approximate solution of rank $O(1/\varepsilon)$.

Problem 3

Consider the stochastic optimization problem

$$x^\star \in \operatorname*{argmin}_x \left\{ f(x) \;\middle|\; x \in \mathcal{X} \right\},$$

Problem 3 continued on the next page...

for some bounded closed convex set $\mathscr{X} \subset \mathbb{R}^p$, where the objective function is given by

$$f(x) := \mathsf{E}\left[F(x;\xi)\right],$$

for some function F and random variable ξ. Suppose that we have access to a stochastic first-order oracle, which for any request $x \in \mathscr{X}$, returns some $g(x;\xi) \in \mathbb{R}^p$, such that

$$\mathsf{E}\left[g(x;\xi)\right] \in \partial f(x).$$

Assume that f is convex and continuous, and

$$\mathsf{E}\left[\|g(x;\xi)\|_2^2\right] \leq L^2, \quad \forall x \in \mathscr{X},$$

for some $L > 0$.

Let $\xi_1, \xi_2, \ldots$ be independent and identically distributed random variables following the probability distribution of ξ. Let $x_1 \in \mathscr{X}$. Consider the *stochastic gradient method*, which iterates as

$$x_{t+1} \leftarrow \Pi_{\mathscr{X}}\left(x_t - \eta_t g(x_t;\xi_t)\right), \quad \forall t \in \mathbb{N}.$$

1. (20 points) **Show that**

$$\eta_t \mathsf{E}\left[f(x_t) - f(x^\star)\right] \leq \mathsf{E}\left[\frac{1}{2}\|x_t - x^\star\|_2^2\right] - \mathsf{E}\left[\frac{1}{2}\|x_{t+1} - x^\star\|_2^2\right] + \frac{1}{2}\eta_t^2 L^2.$$

2. (20 points) Define

$$\bar{x}_{t_1:t_2} := \frac{\sum_{\tau=t_1}^{t_2} \eta_\tau x_\tau}{\sum_{\tau=t_1}^{t_2} \eta_\tau}, \quad R := \max_x \{\, \|x - x_1\|_2 \mid x \in \mathscr{X} \,\}.$$

   Show that

$$\mathsf{E}\left[f(\bar{x}_{t_1:t_2}) - f(x^\star)\right] \leq \frac{4R^2 + L^2 \sum_{\tau=t_1}^{t_2} \eta_\tau^2}{2\sum_{\tau=t_1}^{t_2} \eta_\tau}, \quad \forall 1 \leq t_1 \leq t_2.$$

3. (20 points) Set

$$\eta_t = \frac{R}{L\sqrt{t}}, \quad \forall t \in \mathbb{N}.$$

   Show that then,

$$\mathsf{E}\left[f(\bar{x}_{t_1:t_2}) - f(x^\star)\right] \leq \frac{RL}{\sqrt{t_2}}\left[2\left(\frac{t_2}{t_2 - t_1 + 1}\right) + \frac{1}{2}\sqrt{\frac{t_2}{t_1}}\right], \quad \forall 1 \leq t_1 \leq t_2.$$

   Notice that therefore, if we choose $t_1 = \alpha t_2$ for some $\alpha \in]0,1[$, then

$$\mathsf{E}\left[f(\bar{x}_{t_1:t_2}) - f(x^\star)\right] = O\left(\frac{RL}{\sqrt{t_2}}\right).$$

References

[1] Clarkson, K. L. Coresets, sparse greedy approximation, and the Frank-Wolfe algorithm. *ACM Trans. Algorithms 6*, 4 (2010).