**EE621 ADAPTATION AND LEARNING** **Instructor: Ali H. Sayed**
**Homework Assignment #1 (Background and Fundamentals)** **Due: March 4, 2019**

---

1) Let $\|X\|_{\rm F}$ denote the Frobenius norm of matrix $X$. Show that $\nabla_X \|X\|_{\rm F}^2 = 2X$.
2) Let $f(X)$ denote a scalar real-valued function of a real-valued $M \times N$ matrix argument $X$. Let $X_{mn}$ denote the $(m,n)-$th entry of $X$. The gradient of $f(\cdot)$ relative to $X$ is defined as the following matrix of partial derivatives:

$$\nabla_X f(X) \triangleq \begin{bmatrix} \frac{\partial f(X)}{\partial X_{11}} & \frac{\partial f(X)}{\partial X_{12}} & \cdots & \frac{\partial f(X)}{\partial X_{1N}} \\ \frac{\partial f(X)}{\partial X_{21}} & \frac{\partial f(X)}{\partial X_{22}} & \cdots & \frac{\partial f(X)}{\partial X_{2N}} \\ \vdots & & \ddots & \vdots \\ \frac{\partial f(X)}{\partial X_{M1}} & \frac{\partial f(X)}{\partial X_{M2}} & \cdots & \frac{\partial f(X)}{\partial X_{MN}} \end{bmatrix}$$

   Show that $\nabla_X \det(X) = \det(X) X^{-\mathsf{T}}$.
3) Consider a $\nu-$strongly convex function $g(z) : \mathbb{R}^M \to \mathbb{R}$ with $\mathrm{dom}(g) = \mathbb{R}^M$. Assume $g(z)$ has $\delta-$Lipschitz gradients, i.e., $\|\nabla_z g(z_2) - \nabla_z g(z_1)\| \leq \delta \|z_2 - z_1\|$ for any $z_1, z_2 \in \mathrm{dom}(g)$. Establish the co-coercivity property:

$$\Big(\nabla_{z^{\mathsf{T}}} g(z_2) - \nabla_{z^{\mathsf{T}}} g(z_1)\Big)^{\mathsf{T}} (z_2 - z_1) \geq \frac{\nu\delta}{\nu+\delta} \|z_2 - z_1\|^2 + \frac{1}{\nu+\delta} \|\nabla_z g(z_2) - \nabla_z g(z_1)\|^2$$

4) Let $g(z) = \sum_{n=1}^N |\gamma(n) - h_n^{\mathsf{T}} z|$, where $z, h_n \in \mathbb{R}^M$ and $\gamma(n) \in \mathbb{R}$. Show that a subgradient for $g(z)$ is given by

$$\sum_{n=1}^N -h_n \mathrm{sign}\left(\gamma(n) - h_n^{\mathsf{T}} z\right) \in \partial_{z^{\mathsf{T}}} g(z)$$

   where $\mathrm{sign}(x) = +1$ if $x \geq 0$ and $\mathrm{sign}(x) = -1$ if $x < 0$.
5) Let $h(w) : \mathbb{R}^M \to \mathbb{R}$ denote a convex function. Establish the following properties for the proximal operator of $h(w)$, for any vectors $a, b \in \mathbb{R}^M$:
   (a) $\|\mathrm{prox}_h(a) - \mathrm{prox}_h(b)\| \leq \|a - b\|$.
   (b) $\|\mathrm{prox}_h(a) - \mathrm{prox}_h(b)\|^2 \leq (a-b)^{\mathsf{T}} (\mathrm{prox}_h(a) - \mathrm{prox}_h(b))$.
   (c) $\|\mathrm{prox}_h(a) - \mathrm{prox}_h(b)\|^2 + \|(a - \mathrm{prox}_h(a)) - (b - \mathrm{prox}_h(b))\|^2 \leq \|a-b\|^2$.
   (d) $\|a-b\| = \|\mathrm{prox}_h(a) - \mathrm{prox}_h(b)\| \iff a - b = \mathrm{prox}_h(a) - \mathrm{prox}_h(b)$.
6) Let $P(w)$ be a real-valued first-order differentiable risk function whose gradient vector satisfies a $\delta-$Lipschitz condition. The risk $P(w)$ is *not* assumed convex. Instead, we assume that it is lower-bounded, namely, $P(w) \geq L$ for all $w$ and for some finite value $L$. Consider the gradient-descent algorithm. Show that if the step-size $\mu$ satisfies $\mu < 2/\delta$, then the sequence of iterates $\{w_n\}$ satisfies the following two properties:
   (a) $P(w_n) \leq P(w_{n-1})$.
   (b) $\lim_{n\to\infty} \nabla_w P(w_n) = 0$.
7) The group LASSO problem involves solving a regularized least-squares problem of the following form. We partition each observation vector into $K$ sub-vectors, say, $h_m = \mathrm{col}\{h_{mk}\}$ for $k = 1, 2, \ldots, K$. We similarly partition the weight vector into $K$ sub-vectors, $w = \mathrm{col}\{w_k\}$, of similar dimensions to $h_{mk}$. Now consider the problem:

$$\min_w \left\{ \alpha \sum_{k=1}^K \|w_k\| \; + \; \frac{1}{N} \sum_{m=0}^{N-1} \left( \gamma(m) - \sum_{k=1}^K h_{mk}^{\mathsf{T}} w_k \right)^2 \right\}$$

   (a) Derive a subgradient batch solution for this problem.
   (b) Derive a proximal batch solution for the same problem.

**Computer project**. The statement of the computer projects will be left purposefully vague to give you freedom to explore and choose the settings for your simulations. You are required to submit plots and your code, along with a description of your solution and commentary on the results.

(a) Simulate the Douglas-Rachford algorithm from the chapter on Proximal Operator. Generate learning curves illustrating the convergence behavior of the iterates $w_n$ towards the global minimizer. Describe the data model you used, along with your choices for $q(w)$ and $E(w)$.
(b) Generate a plot that illustrates the $O(\lambda^{n/2})$ vs $O(\lambda^n)$ convergence rate of the subgradient learning algorithm with and without smoothing.