Homework 2 **AY2019/2020 (Sem 2)** **Due Apr 6th 5pm**

Instructions.
For submission of hardcopy: you can submit a handwritten version of solution for Q1, Q2 and Q5 and attach a printout of your codes for Q3 and Q4. The hardcopy can be submitted to the locker space next to LT34 (under the pigeonhole "DSA3102").

For submission of electronic copy: please upload a **single pdf** to LumiNUS with all your codes attached to that single pdf. **Rename your filename with your name and student ID**, e.g. SweeOngChua-A0xxxxxxx-HW2.

1. Consider $f : \mathbb{R}^n \to \mathbb{R}$ defined by $f(\mathbf{x}) = \frac{1}{p}\sum_{i=1}^n |x_i|^p$, for $1 < p < \infty$. Find f^*, the conjugate of f.

2. Let $g : \mathbb{R}^n \to (-\infty, \infty]$ be a proper function. Let $\lambda \neq 0$ and $\mathbf{a} \in \mathbb{R}^n$. Define $f(\mathbf{x}) = g(\lambda\mathbf{x} + \mathbf{a})$. Show that the proximal mapping of f can be written as:

$$Prox_f(\mathbf{x}) = \frac{1}{\lambda}[Prox_{\lambda^2 g}(\lambda\mathbf{x} + \mathbf{a}) - \mathbf{a}].$$

3. Consider the following l_1-regularized logistic regression problem

$$\min_{\mathbf{w}\in\mathbb{R}^p} \quad l(\mathbf{w}) + \beta\|\mathbf{w}\|_1,$$

where $l(\mathbf{w})$ is the negative log-likelihood function defined by

$$l(\mathbf{w}) := \sum_{i=1}^n \log(1 + e^{-y_i\mathbf{w}^T\mathbf{x}_i}), \quad \mathbf{y} \in \mathbb{R}^n, \ \mathbf{x}_i \in \mathbb{R}^p.$$

Solve the above problem for $\mathbf{w}$ with the proximal gradient and accelerated proximal gradient (APG) method with $\beta = 0.05$ and accuracy $tol = 0.001$. Use the same data in Q5 of HW1.

Note that you can write the two algorithms in the same script since the proximal gradient method is a special case of APG.

(a) Show that $\nabla l(\mathbf{w})$ is Lipschitz continuous with Lipschitz constant $\frac{1}{2}\|X\|_F^2$, where X is the matrix whose rows are $\mathbf{x}_i$, and $\|X\|_F$ is the Frobenius norm of X.
(Hint: Given $c \in \mathbb{R}$ and $\mathbf{x} \in \mathbb{R}^p$, the function $g(\mathbf{w}) := \frac{1}{1+e^{c\mathbf{w}^T\mathbf{x}}}$ is Lipschitz continuous with Lipschitz constant $\frac{1}{2}|c|\|\mathbf{x}\|$.)

(b) Discuss how you could solve the subproblem of proximal gradient method/APG, i.e. write down the formula you used to update the new iterate $\mathbf{w}^{k+1}$.

(c) Discuss some practical strategies you used to speed up the algorithm. (You can refer to remark 8.3 in Chapter 8 for such a strategy).

(d) Discuss the condition you used to terminate your algorithms, i.e. given the tolerance $\epsilon > 0$ and new iterate $\mathbf{w}^{k+1}$, how do you determine if $\mathbf{w}^{k+1}$ is the optimal solution for the l_1-regularized logistic regression problem?

(e) How many nonzero entries are there in the optimal solution $\mathbf{w}^*$ using training data X_{train} and $\mathbf{y}_{train}$? Compare $\mathbf{w}^*$ with the one you get from the unregularized version in Q4 or HW1 Q5. Use the same accuracy for a fair comparison.

(f) How does the performance of APG compared to that of the proximal gradient method? Plot the objective function as a function of number of iterations for the two methods in the **same** graph.

(g) Report the misclassification rate for training and test data.

Deliverables: Show all the codes you used in your pdf file (you don't need to submit the program files separately).

4. Solve Q5 of HW1 with the stochastic gradient algorithm. Compare the stochastic gradient algorithm with the steepest descent algorithm you had coded for HW 1 in terms of number of iterations and computational time to reach to the same accuracy. (Use $tol \in \{0.1, 0.01, 0.001\}$). You don't need to try many parameters for the algorithms, one initial solution and one step size strategy is enough. Use the same strategy for both algorithms for a fair comparison.

   Deliverables: Show all the codes you used in your pdf file (you don't need to submit the program files separately) and write a short discussion about the comparison of the algorithms.

5. Given $\mathbf{b} \in \mathbb{R}^n$ and positive definite $Q \in \mathbb{R}^{n \times n}$, consider the following box-constrained quadratic program:

$$\begin{aligned} \text{minimize} \quad & \tfrac{1}{2}\mathbf{x}^T Q\mathbf{x} - \mathbf{b}\mathbf{x} \\ \text{subject to} \quad & \mathbf{l} \leq \mathbf{x} \leq \mathbf{u}. \end{aligned}$$

   We want to apply a coordinate descent algorithm to solve this problem. Assume that the current solution is $\hat{\mathbf{x}}$ and the next coordinate to use in the algorithm is i. Solve the corresponding 1-dimensional problem and give an expression for the next iterate.