**COMP5212: Machine Learning** **Fall 2022**

Homework 1: Due Friday Oct. 7, 11:59 PM

**Instructions**: upload a PDF report using LaTeX containing your answers to Canvas (remember to include your name and ID number).

# Problem 1. Smoothness

A differential function $f$ is said to be $L$-smooth if its gradietns are Lipschitz continuous, that is

$$\|\nabla f(x) - \nabla f(y)\| \leq L\|x - y\|$$

let $f : \mathbb{R}^d \to \mathbb{R}$ be a twice differentiable function. If $f$ is $L$-smooth then prove the following inequality:

- (15 pt) Prove $\langle \nabla^2 f(x)v, v\rangle \leq L\|v\|_2^2, \quad \forall x, v \in \mathbb{R}^d$
- (15 pt) Prove $f(y) \leq f(x) + \langle \nabla f(x), y - x\rangle + \frac{L}{2}\|y - x\|_2^2$

# Problem 2. Gradient descent rate with line search in strongly convex function

Suppose the function $f : \mathbb{R}^n \to \mathbb{R}$ is strongly convex and twice differentiable, i.e $\nabla^2 f(x) \succeq lI$ with constant $l > 0$. Also, its gradient is Lipschitz continuous with constant $L > 0$, i.e. we have that $\|\nabla f(x) - \nabla f(y)\| \leq L\|x - y\|$ for any $x, y$.

- (5 pt) Prove $f(y) \geq f(x) + \langle \nabla f(x), y - x\rangle + \frac{l}{2}\|y - x\|^2$
- (10 pt) Prove $f(y) \geq f(x) - \frac{1}{2l}\|\nabla f(x)\|_2$
  (hint: $f(y) \geq \min_y f(y)$ )
- (15 pt) Then if we run gradient descent for $t$ iterations with step size $\alpha = \frac{1}{L}$ by using exact line search, prove it will give a linear convergence rate, i.e.

$$f(x^{t+1}) - f^* \leq (1 - \frac{l}{L})(f(x) - f^*)$$

# Problem 3. Proximal Gradient Descent

Consider solving the following problem

$$\min_{\boldsymbol{w}} \|X\boldsymbol{w} - \boldsymbol{y}\|_2^2 + \lambda\|\boldsymbol{w}\|_1,$$

where $X \in R^{n\times d}$ is the feature matrix (each row is a feature vector), $\boldsymbol{y} \in R^n$ is the label vector, $\|\boldsymbol{w}\|_1 := \sum_i |w_i|$ and $\lambda > 0$ is a constant to balance loss and regularization. This is known as the Lasso regression problem and our goal is to derive the "proximal gradient method" for solving this.

- (10 pt) The gradient descent algorithm cannot be directly applied since the objective function is non-differentiable. Discuss why the objective function is non-differentiable.
- (30 pt) In the class we showed that gradient descent is based on the idea of function approximation. To form an approximation for non-differentiable function, we split the differentiable part and non-differentiable part. Let $g(\boldsymbol{w}) = \|X\boldsymbol{w} - \boldsymbol{y}\|_2^2$, as discussed in the gradient descent lecture we approximate $g(\boldsymbol{w})$ by

$$g(\boldsymbol{w}) \approx \hat{g}(\boldsymbol{w}) := g(\boldsymbol{w}_t) + \nabla g(\boldsymbol{w}_t)^T(\boldsymbol{w} - \boldsymbol{w}_t) + \frac{\eta}{2}\|\boldsymbol{w} - \boldsymbol{w}_t\|^2.$$

  In each iteration of proximal gradient descent, we obtain the next iterate ($\boldsymbol{w}_{t+1}$) by minimizing the following approximation function:

$$\boldsymbol{w}_{t+1} = \arg\min_{\boldsymbol{w}} \hat{g}(\boldsymbol{w}) + \lambda\|\boldsymbol{w}\|_1.$$

  Derive the close form solution of $\boldsymbol{w}_{t+1}$ given $\boldsymbol{w}_t, \nabla g(\boldsymbol{w}_t), \eta, \lambda$. What's the time complexity for one proximal gradient descent iteration?