COMS 4995-004: Optimization for Machine Learning
Homework 2 **(Corrected version)**.

HW2 is due Saturday, Oct 12 by 1:00 pm. No late assignments will be accepted[1]**. Please refer to** `https://www.satyenkale.com/optml-f19/` **for instructions on how to submit homework assignments.**

Question 1. The gradient methods we studied in class for minimizing β-smooth functions and β-smooth & α-strongly convex functions have a desirable *anytime* guarantee on the iterates: we can stop the method at any time step t and are guaranteed that the iterate x_t has a guaranteed suboptimality. Specifically, the analysis we saw in class immediately yields the following statements:

- For β-smooth functions f, the gradient method run with step-size $\eta = \frac{1}{\beta}$ guarantees that at any time step t, we have $f(x_t) - f(x^*) \leq \frac{\beta\|x_0 - x^*\|^2}{2t}$.
- For β-smooth & α-strongly convex functions the gradient method run with step-size $\eta = \frac{1}{\beta}$ guarantees that at any time step t, we have $f(x_t) - f(x^*) \leq (1 - \frac{\alpha}{\beta})^t \frac{\beta\|x_0 - x^*\|^2}{2}$.

Unfortunately, the analysis we saw for L-Lipschitz convex functions with a constant step-size η *does not* have such a guarantee. In this question, we will derive a tweak to the gradient method that enjoys an anytime guarantee using *decreasing* step-sizes. Suppose we run the *projected* gradient method for minimizing an L-Lipschitz convex function over a convex set K which has diameter bounded by D, i.e. for any $x, x' \in K$, we have $\|x - x'\|_2 \leq D$. Consider running the method with decreasing step-sizes $\eta_1 \geq \eta_2 \geq \eta_3 \cdots$. I.e. we start at an arbitrary point x_0, and at time step t we set $x_{t+1} = \prod_K (x_t - \eta_t \nabla f(x_t))$.

1. **(1 point)** Show that

$$f(x_t) - f(x^*) \leq \frac{1}{2\eta_t}(\|x_t - x^*\|^2 - \|x_{t+1} - x^*\|^2) + \frac{\eta_t}{2}\|\nabla f(x_t)\|^2.$$

2. **(4 points)** Using the above inequality, show that

$$\sum_{i=0}^{t} f(x_i) - f(x^*) \leq \frac{D^2}{2\eta_t} + \frac{L^2}{2}\sum_{i=0}^{t}\eta_i.$$

3. **(4 points)** Suppose we set $\eta_i = \frac{D}{L\sqrt{i+1}}$. Let $\bar{x}_t = \frac{1}{t+1}\sum_{i=0}^{t} x_i$. Then show that

$$f(\bar{x}_t) - f(x^*) \leq \frac{2DL}{\sqrt{t+1}}.$$

[1] Unless you have an emergency; in that case please write to Satyen as soon as possible.

Question 2. Consider the following training set: $S = \{(x_i, y_i) \in \mathbb{R}^3 \times \mathbb{R} \mid i = 1, 2, 3\}$, where

$$(x_1, y_1) = ((2, 0, 0), 1)$$
$$(x_2, y_2) = ((0, 1, 0), -1)$$
$$(x_3, y_3) = ((0, 0, 0.5), 1).$$

Suppose we want to train a linear predictor $f_w = \langle w, x \rangle$ for some weight vector $w \in K = \{w \in \mathbb{R}^3 \mid \|w\|_2 \leq 10\}$. Consider training the predictor using the following three loss functions and regularization functions:

(i) (Square loss with no regularization) loss function $\ell(\hat{y}, y) = (\hat{y} - y)^2$, no regularization.

(ii) (Square loss with ℓ_1 regularization) loss function $\ell(\hat{y}, y) = (\hat{y} - y)^2$, regularization $R(w) = \|w\|_1$, regularization constant $\lambda = 1$.

(iii) (Logistic loss with no regularization) loss function $\ell(\hat{y}, y) = \log(1 + \exp(-\hat{y}y))$, no regularization.

(iv) (Logistic loss with ℓ_2 regularization) loss function $\ell(\hat{y}, y) = \log(1 + \exp(-\hat{y}y))$, regularization $R(w) = \frac{1}{2}\|w\|_2^2$, regularization constant $\lambda = 1$.

Suppose we want to minimize the training loss function in each of the above cases up to a sub-optimality gap of $\epsilon = 0.01$ using a gradient method starting from $w_0 = 0$. Describe which version of gradient descent taught in class will require the minimum number of iterations T to achieve the sub-optimality gap of ϵ. For each case, specify numerical values for the step-size η you will use in the algorithm, and the number of iterations T that will be necessary to achieve the suboptimality gap of ϵ. **(4 points per training loss function)**

Note: the setup is (almost) exactly the same as problem 2 in HW1 – the only difference is that w is chosen from a bounded set K rather than all of $\mathbb{R}^3$. You may reuse all the calculations from HW1 from your own solution or the one posted online. In doing the calculations, it is OK to make somewhat crude numerical approximations.