COMS 4995-004: Optimization for Machine Learning
Homework 2.

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

Solution.
1. Let $y_{t+1} = x_t - \eta_t \nabla f(x_t)$. We have
$$\|y_{t+1} - x^*\|^2 = \|x_t - \eta_t \nabla f(x_t) - x^*\|^2 = \|x_t - x^*\|^2 + \eta_t^2\|\nabla f(x_t)\|^2 - 2\eta_t \nabla f(x_t)^\top (x_t - x^*).$$

Using the properties of projection on a convex set, we have $\|y_{t+1} - x^*\|^2 \geq \|x_{t+1} - x^*\|^2$. Putting these two inequalities together, and simplifying, we get

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

Using the convexity of f, we get $\nabla f(x_t)^\top (x_t - x^*) \geq f(x_t) - f(x^*)$. Putting these bounds together, we get the required inequality.

2. Summing up the inequality in part 1, we get

$$\begin{aligned}
\sum_{i=0}^{t} f(x_i) - f(x^*) &\leq \sum_{i=0}^{t} \frac{1}{2\eta_i}(\|x_i - x^*\|^2 - \|x_{i+1} - x^*\|^2) + \sum_{i=0}^{t} \frac{\eta_i}{2}\|\nabla f(x_i)\|^2 \\
&= \frac{1}{2\eta_0}\|x_0 - x^*\|^2 - \frac{1}{2\eta_t}\|x_{t+1} - x^*\|^2 + \sum_{i=1}^{t}(\frac{1}{2\eta_i} - \frac{1}{2\eta_{i-1}})\|x_i - x^*\|^2 + \sum_{i=0}^{t} \frac{\eta_i}{2}\|\nabla f(x_i)\|^2 \\
&\leq \frac{1}{2\eta_0}D^2 + \sum_{i=1}^{t}\left(\frac{1}{2\eta_i} - \frac{1}{2\eta_{i-1}}\right) D^2 + \sum_{i=0}^{t} \frac{\eta_i}{2}L^2 \\
&= \frac{1}{2\eta_t}D^2 + \frac{L^2}{2}\sum_{i=0}^{t}\eta_i.
\end{aligned}$$

The second inequality uses the following facts:

1. $\left(\frac{1}{2\eta_i} - \frac{1}{2\eta_{i-1}}\right) \geq 0$ since $\eta_i \leq \eta_{i-1}$.
2. $\|x_i - x^*\|^2 \leq D^2$ for all i.
3. The term $-\frac{1}{2\eta_t}\|x_{t+1} - x^*\|^2$ can be dropped since it is non-positive.

3. If we set $\eta_i = \frac{D}{L\sqrt{i+1}}$, then $\sum_{i=0}^{t}\eta_i = \sum_{i=0}^{t}\frac{D}{L\sqrt{i+1}}$. We have

$$\sum_{i=0}^{t}\frac{1}{\sqrt{i+1}} \leq 1 + \int_{x=1}^{t+1}\frac{1}{\sqrt{x}} = 2\sqrt{t+1} - 1 \leq 2\sqrt{t+1}.$$

Plugging this bound in the inequality of part 2, we get

$$\sum_{i=0}^{t} f(x_i) - f(x^*) \leq \frac{1}{2\eta_t}D^2 + \frac{L^2}{2}\sum_{i=0}^{t}\eta_i \leq \frac{1}{2}DL\sqrt{t+1} + DL\sqrt{t+1} \leq 2DL\sqrt{t+1}.$$

Dividing by $t+1$, and using Jensen's inequality to the convex function f, we get

$$f(\bar{x}_t) - f(x^*) \leq \frac{1}{t+1}\sum_{i=0}^{t} f(x_i) - f(x^*) \leq \frac{2DL}{\sqrt{t+1}}.$$

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

$$\begin{aligned}
(x_1, y_1) &= ((2, 0, 0), 1) \\
(x_2, y_2) &= ((0, 1, 0), -1) \\
(x_3, y_3) &= ((0, 0, 0.5), 1).
\end{aligned}$$

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.

Solution.

Since $w^* \in K$, we have $\|w_0 - w^*\| \leq 10$. We will use this bound $D = 10$ in the analysis.

(i) (Square loss with no regularization) loss function $\ell(\hat{y}, y) = (\hat{y} - y)^2$, no regularization. From HW1, we have

$$\nabla L(w) = \frac{2}{3}\sum_{i=1}^{3}(\langle w, x_i \rangle - y_i)x_i$$

and

$$\nabla L(w) = \frac{2}{3}\sum_{i=1}^{3} x_i x_i^\top = \frac{2}{3}\begin{bmatrix} 4 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 0.25 \end{bmatrix}.$$

We bound the gradient norm as follows:

$$\|\frac{2}{3}\sum_{i=1}^{3}(\langle w, x_i \rangle - y_i)x_i\| \leq \frac{2}{3}\sum_{i=1}^{3}(\|w\|\|x_i\| + |y_i|)\|x_i\|.$$

Since $\|w\| \leq 10$, the above can be bounded as

$$\frac{2}{3}\sum_{i=1}^{3}(\|w\|\|x_i\| + |y_i|)\|x_i\| \leq \frac{2}{3}((10 \cdot 2 + 1) \cdot 2 + (10 \cdot 1 + 1) \cdot 1 + (10 \cdot 0.5 + 1) \cdot 0.5) \leq 38.$$

In HW1 we showed that the training loss function is $\frac{1}{6}$-strongly convex and $\frac{8}{3}$-smooth. We can now bound the number of iterations for various flavors of gradient descent as follows.

(a) **GD for L-Lipschitz functions:** the number of iterations is bounded by

$$\frac{D^2L^2}{\epsilon^2} \leq \frac{10^2 \cdot 38^2}{0.01^2} = 1.444 \times 10^9.$$

(b) **GD for β-smooth functions:** the number of iterations is bounded by

$$\frac{\beta D^2}{2\epsilon} = \frac{\frac{8}{3} \cdot 10^2}{2 \cdot 0.01} \approx 13,334.$$

(c) **GD for α-strongly convex and β-smooth functions:** after T steps, the sub-optimality gap is bounded by

$$L \cdot \left(1 - \frac{\alpha}{\beta}\right)^{T/2} D + \frac{\beta}{2}\left(1 - \frac{\alpha}{\beta}\right)^{T} D^2 = (1-\tfrac{1}{16})^{T/2} \cdot 38 \cdot 10 + (1-\tfrac{1}{16})^{T} \cdot \frac{4}{3} \cdot 10^2 \leq (1-\tfrac{1}{16})^{T/2} \cdot 600.$$

To make this smaller than 0.01, we need

$$2 \cdot \frac{\log(\frac{0.01}{600})}{\log(1 - \frac{1}{16})} \approx 341$$

iterations.

(d) **GD for L-Lipschitz and α-strongly convex functions:** after T steps, the sub-optimality gap is bounded by

$$\frac{L^2 \ln(T)}{2\alpha T} = \frac{38^2 \ln(T)}{2 \cdot \frac{1}{6} T} = \frac{4332 \ln(T)}{T}.$$

To make this smaller than 0.01, we need T to be around 7×10^6.

So, the best method is the one for α-strongly convex and β-smooth functions, which requires 341 iterations. The step size required for this is $\frac{1}{\beta} = \frac{3}{8}$.

(ii) (Square loss with ℓ_1 regularization) loss function $\ell(\hat{y}, y) = (\hat{y} - y)^2$, regularization $R(w) = \|w\|_1$, regularization constant $\lambda = 1$.
Here, the training loss function is non-smooth but it is $\frac{1}{6}$ strongly convex. The gradient is $\frac{2}{3}\sum_{i=1}^{3}(\langle w, x_i\rangle - y_i)x_i + \langle \text{sign}(w_1), \text{sign}(w_2), \text{sign}(w_3)\rangle$. We have $\|\langle \text{sign}(w_1), \text{sign}(w_2), \text{sign}(w_3)\rangle\| = \sqrt{3} \leq 2$. Thus the Lipschitz constant can be bounded by $38 + 2 = 40$. We can now bound the number of iterations for various flavors of gradient descent as follows.

(a) **GD for L-Lipschitz functions:** the number of iterations is bounded by

$$\frac{D^2L^2}{\epsilon^2} \leq \frac{10^2 \cdot 40^2}{0.01^2} = 1.6 \times 10^9.$$

(b) **GD for L-Lipschitz and α-strongly convex functions:** after T steps, the sub-optimality gap is bounded by

$$\frac{L^2 \ln(T)}{2\alpha T} = \frac{40^2 \ln(T)}{2 \cdot \frac{1}{6} T} = \frac{4800 \ln(T)}{T}.$$

To make this smaller than 0.01, we need T to be around 7×10^6.

So, the best method is the one for α-strongly convex Lipschitz functions, which requires 7×10^6 iterations. We need to use decreasing step sizes for this, viz. $\eta_t = \frac{1}{\alpha(t+1)} = \frac{6}{t+1}$.

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

$$\nabla L(w) = \frac{1}{3} \sum_{i=1}^{3} \frac{-\exp(-\langle w, x_i \rangle y_i) y_i}{1 + \exp(-\langle w, x_i \rangle y_i)} x_i$$

and

$$\nabla^2 L(w) = \operatorname{diag}\left(\frac{4\exp(-2w_1)}{3(1+\exp(-2w_1))^2}, \frac{\exp(w_2)}{3(1+\exp(w_2))^2}, \frac{0.25\exp(-0.5w_3)}{3(1+\exp(-0.5w_3))^2)} \right).$$

$$\|\nabla L(w)\| = \left\| \frac{1}{3} \sum_{i=1}^{3} \frac{-\exp(-\langle w, x_i \rangle y_i) y_i}{1 + \exp(-\langle w, x_i \rangle y_i)} x_i \right\| \leq \frac{1}{3} \sum_{i=1}^{3} \|x_i\| = \frac{3.5}{3} \leq 2.$$

As worked out in HW1, the training loss function is $\frac{1}{3}$ smooth. As for strong convexity, over the set K, it is easy to check that the eigenvalues of the Hessian are at least $\frac{1}{20}\exp(-20) \approx 10^{-10}$. We can now bound the number of iterations for various flavors of gradient descent as follows.

(a) **GD for L-Lipschitz functions:** the number of iterations is bounded by

$$\frac{D^2 L^2}{\epsilon^2} \leq \frac{10^2 \cdot 2^2}{0.01^2} = 4 \times 10^6.$$

(b) **GD for β-smooth functions:** the number of iterations is bounded by

$$\frac{\beta D^2}{2\epsilon} = \frac{\frac{1}{3} \cdot 10^2}{2 \cdot 0.01} \approx 1,667.$$

(c) **GD for α-strongly convex and β-smooth functions:** since $\alpha \approx 10^{-10}$, and the bound on the number of iterations depends on $\frac{1}{\alpha}$, we need at least 10^{10} iterations, so this bound is definitely worse than the bound only using smoothness and not strong convexity.

(d) **GD for L-Lipschitz and α-strongly convex functions:** this bound also scales with $\frac{1}{\alpha} \geq 10^{10}$.

So, the best method is the one for β-smooth functions, which requires 1,667 iterations. The step size needed for this is $\frac{1}{\beta} = 3$.

(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 $\lambda = 1$.
We haev $\nabla R(w) = w$, and $\|w\| \leq 10$, so the Lipschitz constant becomes $2 + 10 = 12$. As worked out in HW1, the training loss function is $\frac{4}{3}$ smooth and 1-strongly convex. As for strong convexity, over the set K, it is easy to check that the eigenvalues of the Hessian are at least $\frac{1}{20}\exp(-20) \approx 10^{-10}$. We can now bound the number of iterations for various flavors of gradient descent as follows.

(a) **GD for L-Lipschitz functions:** the number of iterations is bounded by

$$\frac{D^2L^2}{\epsilon^2} \leq \frac{10^2 \cdot 12^2}{0.01^2} = 1.44 \times 10^8.$$

(b) **GD for β-smooth functions:** the number of iterations is bounded by

$$\frac{\beta D^2}{2\epsilon} = \frac{\frac{4}{3} \cdot 10^2}{2 \cdot 0.01} \approx 6,667.$$

(c) **GD for α-strongly convex and β-smooth functions:** after T steps, the sub-optimality gap is bounded by

$$L \cdot \left(1 - \frac{\alpha}{\beta}\right)^{T/2} D + \frac{\beta}{2}\left(1 - \frac{\alpha}{\beta}\right)^{T} D^2 = (1-\tfrac{3}{4})^{T/2} \cdot 12 \cdot 10 + (1-\tfrac{3}{4})^{T} \cdot \frac{2}{3} \cdot 10^2 \leq (1-\tfrac{1}{16})^{T/2} \cdot 200.$$

To make this smaller than 0.01, we need

$$2 \cdot \frac{\log(\frac{0.01}{200})}{\log(1 - \frac{3}{4})} \approx 15$$

iterations.

(d) **GD for L-Lipschitz and α-strongly convex functions:** after T steps, the sub-optimality gap is bounded by

$$\frac{L^2 \ln(T)}{2\alpha T} = \frac{12^2 \ln(T)}{2T} = \frac{72 \ln(T)}{T}.$$

To make this smaller than 0.01, we need T to be around $80,000$.

So, the best method is the one for α-strongly convex and β-smooth functions, which requires 15 iterations. The step size needed for this is $\frac{1}{\beta} = \frac{3}{4}$.