Spring 2022 CS 795 Large Scale Optimization for Machine Learning
Assignment 2

mpavlak - pjatav

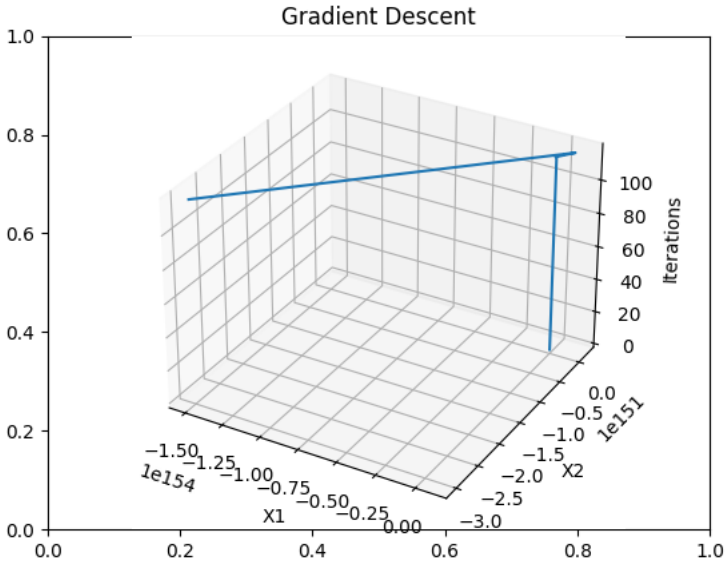
Announcement Please use LaTeX to type your solution. Only PDF document generated by LaTeX is accepted. In each team, at the very beginning of the PDF document, please include you and your teammate's name and GMU NetID. The submitted file name should be `netid1_netid2_hw2.pdf`.

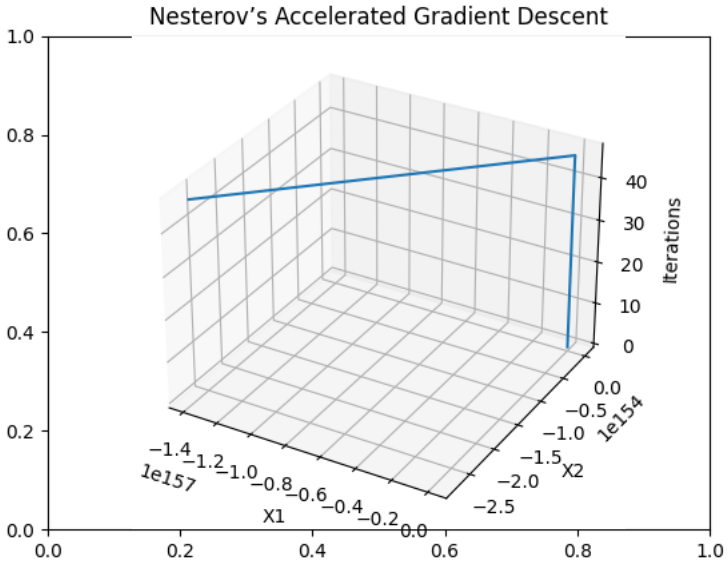
1 Experiments: Nesterov's Accelerated Gradient Descent (30 points)

Consider the quadratic function $f(x) = \frac{1}{2}x^\top Ax$, where $x \in \mathbb{R}^d$, $A \in \mathbb{R}^{d\times d}$ with $\lambda_{\min}(A) = \alpha > 0$ and $\lambda_{\max}(A) = L > 0$. Run Nesterov's Accelerated Gradient Descent (1) with $\eta = \frac{1}{L}$ and $\gamma = \frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1}$ where $\kappa = L/\alpha$.

- (15 points) Define $B = \begin{bmatrix} (1+\gamma)(1-\eta\lambda_i) & -\gamma(1-\eta\lambda_i) \\ 1 & 0 \end{bmatrix}$, where λ_i is the i-th larget eigenvalue of A. Show that the magnitude of all eigenvalues of B is at most $1 - \Theta(1/\sqrt{\kappa})$.
- (15 points) Let $A = \begin{bmatrix} 1000 & 2 \\ 5 & 1 \end{bmatrix}$, and the starting point of the algorithm is $x_0 = [100; 100]$. Run both gradient descent and Nesterov's AGD and plot them on the same figure (with vertical axis being function value and horizontal axis being number of iterations). Explain your observation.

Proof. We implement the algorithm in python for both gradient descent and Nesterov's AGD and plot the graph how fast both reach to x_* form the $x_0 = [100; 100]$.
Below are the plots.





Plot 2. Plot

We have observed the below results:

Nesterov's AGD reaches X* with less iteration compared to gradient descent.

Nesterov's AGD does not always reach to X* with less iteration, to do that we have found the right hyperparameter values.

□

2 Some Properties of Nesterov's Accelerated Gradient Descent (45 points)

Assume f is a convex and L-smooth function. In class, we showed that gradient descent satisfies the descent lemma, i.e., $f(x_{t+1}) \leq f(x_t) - \frac{\eta}{2}\|\nabla f(x_t)\|^2$, where $\eta \leq \frac{1}{L}$ is the learning rate. In this problem, we try to study the behavior of Nesterov's Accelerated Gradient Descent (AGD):

$$\begin{aligned} y_t &= x_t + \gamma(x_t - x_{t-1}) \\ x_{t+1} &= y_t - \eta \nabla f(y_t), \end{aligned} \tag{1}$$

where $\eta \leq \frac{1}{L}$ is the learning rate, $\gamma \in [0, 1]$.

- (15 points) Show that AGD is not necessarily monotonically decreasing, i.e., there exists a function f such that $f(x_{t+1}) \leq f(x_t)$ does not hold for every t by running AGD.

  Proof. By example, we can show running NAGD on $f(x) = x^2$ is not monotonically decreasing We can choose $x0 = y0 = 5$, $\gamma = 0.5$, and $\eta = \frac{1}{10}$, with $L = 10$

i	x	y
0	5.0	5.0
1	4.0	5.0
2	2.8	3.5
3	1.76	2.2
4	1	1.24
5	0.4864	0.608
6	0.18688	0.2336
7	0.0297	0.03712
8	-0.0391168	-0.048896
9	-0.05881856	-0.0735232
10	-0.054935552	-0.06866944

  $f(x_{t+1} \leq f(x_t)$ does not hold in this example for $t = 9$
  $f(x_{10}) \leq f(x_9)$ implies $-0.054935552 \leq -0.05881856$, which is false.

  Idea: NAGD has 3 parts;
  1. the normal G.D. step between x_t and x_{t+1},
  2. a 'look ahead' term that considers the gradient evaulated at the x_{t+1} value,
  3. and the momentum of the previous steps, which is independent of x_t and x_{t+1} at any given iteration ste

NAGD knows it has reached a minimum when the sign of the gradient at the look-ahead value changes dire
The gradient at the look ahead reduces the amount by which the minimum is over shoot,
but because of the momentum of the previous terms, which is not dependant on the gradient at look-ahead
there is no gaurantee that the magnitude of the momentum is ALWAYS less than the sum of the G.D. step
Thus, there is no gaurantee that NAGD will not overshoot the minimum multiple times as NAGD starts to
NAGD is gauranteed to eventually converge to the minumum, but it may start to osscolate around the true
until the momentum of the previous terms also starts to converge. □

Proof.
Let $V_t = x_t + \gamma(x_t - x_{t-1})$
Assume $f(x_{t+1}) \leq f(x_t)$, thus $f(x_t) - f(x_{t-1}) \leq 0$
By running NAGD:
$f(V_t - \eta\nabla f(V_t)) - f(V_{t-1} - \eta\nabla f(V_{t-1})) \leq 0$
Plug in definition of V_t:

$f(x_t) - f(x_{t-1}) + \gamma f(x_t) - 2\gamma f(x_{t-1}) + \gamma f(x_{t-2}) - \eta\nabla f(x_t) - \eta\nabla f(x_{t-1}) - \eta\gamma\nabla f(x_t) + \eta\gamma\nabla f(x_{t-2}) \leq 0$

By assumption: $((1+\gamma)(f(x_t) - f(x_{t-1}))) \leq 0$
By assumption: $\gamma f(x_{t-1}) + \gamma f(x_{t-2}) \geq 0$
By assumption: $-\eta[(1+\gamma)\nabla f(x_t) + \nabla f(x_{t-1})] \leq 0$
By assumption: $\eta\gamma\nabla f(x_{t-2}) \geq 0$

This shows that there exist some $f(x)$ where the momentum of the x_{t-2} and x_{t-1} terms can cause contradi
depending on the magitude of momentum relative to the noraml G.D. step and look ahead gradient evaulat
Thus, NAGD is not gauranteed to be monotonically decreasing.

□

- (15 points) Define $E_t := f(x_t) + \frac{1}{2\eta}\|x_t - x_{t-1}\|^2$, show that

$$E_{t+1} \leq E_t - \frac{1-\gamma^2}{2\eta}\|x_t - x_{t-1}\|^2. \tag{2}$$

Proof.
By applying NAGD and using the L-smoothness of f, we know:
$f(x - \frac{1}{L}\nabla f(x)) - f(y) \leq \frac{-1}{2L}||\nabla f(x)||^2 + \nabla f(x)^T(x - y)$
Let $x = y_{t+1}$ and $y = y_t$ in the above inequality, resulting in:
$f(y_{s+1}) - f(y_s) = f(x_s - \frac{1}{L}\nabla f(x_s)) - f(y_s) \leq \frac{-1}{2L}||\nabla f(x_s)||^2 + \nabla f(x_s)^T(x_s - y_s)$
By smoothness of f:
$\leq \frac{-L}{2}||y_{s+1} - x_s||^2 - L(y_{s+1} - x_s)^T(x_s - y_s)$
By definition of y_t and y_{t+1}:

$\leq \frac{-L}{2}||x_{t+1} + \gamma(x_{t+1} - x_t) - x_t||^2 - L(x_{t+1} + \gamma(x_{t+1} - x_t) - x_t)^T(x_t - x_t + \gamma(x_t - x_{t-1}))$
By simplifying terms and moving $f(y_s)$ to the other side of the inequality:
$f(y_{s+1}) \leq f(y_s) - \frac{L}{2}||x_{t+1} - x_t||^2 - L(x_{t+1} - x_t)(1 - \gamma^2)$
By applying the definition of E_t and $\eta = \frac{1}{L}$:
$E_{t+1} \leq E_t - \frac{1}{2\eta}||x_t - x_{t-1}||^2(1 - \gamma^2)$

□

- (15 points) Built upon (2), choose the value of γ to show that AGD satisfies

$$f(x_t) - \min_x f(x) \leq O(1/t^2).$$

Proof.
By definition:
$f(y_{s+1}) - f(x_\star) \leq -\frac{L}{2}||y_{s+1} - x_s||^2 - L(y_{s+1} - x_s)^T(x_s - x_\star)$
Let $\delta_s = f(y_s) - f(x_\star)$
By multiplying both sides of the inequality by $(\gamma_s - 1)$ and convexity:
$\gamma_s\delta_{s+1} - (\gamma_s - 1)\delta_s \leq -\frac{L}{2}\gamma_s||y_{s+1} - x_s||^2 - L(y_{s+1} - x_s)^T(\gamma_s x_s - (\gamma_s - 1)y_s - x_\star)$
We know $\gamma_{s-1}^2 \leq \gamma_s^2 - \gamma_s$; by multiplying both sides by γ_s:
$\gamma_s^2\delta_{s+1} - \gamma_{s-1}^2\delta_s \leq -\frac{L}{2}(||\gamma_s(y_{s+1} - x_s)||^2 + 2\gamma_s(y_{s+1} - x_s)^T(\gamma_s x_s - (\gamma_s - 1)y_s - x_\star))$
$\leq -\frac{L}{2}(||\gamma_s y_{s+1} - (\gamma_s - 1)y_s - x_\star||^2 - ||\gamma_s x_s - (\gamma_s - 1)y_s - x_\star||^2)$
Let $Q_s = \gamma_s x_s - (\gamma_s - 1)y_s - x_\star$:
$\gamma_s^2\delta_{s+1}^2 - \gamma_{s+1}^2\delta_s^2 \leq \frac{L}{2}(||Q_s||^2 - ||Q_{s+1}||^2)$
By summing from $s = 1$ to $t - 1$, the middle terms cancel out and we are left with:
$\delta_t \leq \frac{L}{2\gamma_{t-1}^2}||Q_1||^2$
We know $\gamma_{t-1}^2 \leq (\frac{t}{2})^2$, thus:
$f(x_t) - f(x_\star) \leq \frac{L}{2(\frac{t}{2})^2}||\gamma_1 x_1 - (\gamma_1 - 1)y_1 - x_\star||^2$
$\leq \frac{2L}{t^2}||x_1 - x\star||^2 \leq O(1/t^2)$

□

3 Recover Strongly Convex Rate by Convex Results (20 points)

Suppose we have an algorithm $\mathcal{A}$ (which is not necessarily gradient descent). The algorithm takes an initial point x_1 and an integer $t \in \mathbb{N}$ as input, and has the following theoretical guarantees: for any L-smooth, convex function f, after quering the gradient oracle for t times, the output x_t satisfies:

$$f(x_t) - f(x_*) \leq \frac{L\|x_1 - x_*\|^2}{t}.$$

Prove that, for any L-smooth, α-strongly convex function f, to find a point $\hat{x}$ such that $f(\hat{x}) - \min_x f(x) \leq \epsilon$, it suffices to query the gradient oracle $\widetilde{O}(L/\alpha)$ [1] times, by smart uses of the algorithm

[1] The notation $\mathcal{O}(\cdot)$ hides polylogarithmic terms in terms of $1/\epsilon$.

$\mathcal{A}$.

Proof.

From Smoothness and Convexity of f:
$f(x_t) - f(x_\star) \leq \frac{L||x_1 - x_\star||^2}{t}$

Let $\tilde{f}(x) = f(x) - \frac{\alpha}{2}||x||^2$
Thus: $f(x) = \tilde{f}(x) + \frac{\alpha}{2}||x||^2$
We know $\frac{\alpha}{2}||x||^2 \geq 0$, so $f(x) - \frac{\alpha}{2}||x||^2 \leq f(x)$
Using the initial inequality:
$f(x_t) - f(x_{x+1}) \leq f(x_t) - f(x_{t+1}) - \frac{\alpha}{2}||x||^2 \leq \frac{L||x_1 - x_\star||^2}{t}$
For any t iteration:
$f(x_t) - f(x_{t+1}) \leq f(x_t) - f(x_{t+1} - \frac{\alpha}{2}||x_t||^2$
$f(x_{t+1}) - f(x_{t+2}) \leq f(x_{t+1}) - f(x_{t+2} - \frac{\alpha}{2}||x_{t+1}||^2$

By adding the two iterations for t and t+1:
$f(x_t) - f(x_{t+1}) + f(x_{t+1}) - f(x_{t+2}) \leq f(x_t) - f(x_{t+1} + f(x_{t+1} - f(x_{t+2}) - \frac{\alpha}{2}||x_t||^2 - \frac{\alpha}{2}||x_{t+1}||^2$
We know $\frac{\alpha}{2}||x_t||^2 \leq \frac{\alpha}{2}||x_{t+1}||^2$
So by repeating this process for t=1 to t:
$f(x_1) - f(x_\star) \leq f(x_0) - f(x_\star) - (\frac{\alpha}{2}||x_0||^2 - \frac{\alpha}{2}||x_\star||^2)t$
Using the initial inequality:
$\tilde{f}(x_1) - f(x_\star) \leq \frac{L||x_1 - x_\star||^2}{t} - \frac{2}{\alpha t}||x_1 - x_\star||^2$
Thus: $\tilde{f}(x) - f(x_\star)$ converges with $\tilde{O}(L/\alpha)$

□

4 Relaxation of Global Smoothness (35 points, including 30 bonus points)

Gradient Descent (GD) typically requires that the function $f : \mathbb{R}^d \to \mathbb{R}$ is convex and *globally* L-smooth: $\|\nabla f(x) - \nabla f(y)\| \leq L\|x - y\|$, and then the stepsize of GD can be chosen to be $\eta = 1/L$ and then we have the convergence rate: $f(x_t) - \min_x f(x) \leq \frac{L\|x_0 - x_*\|^2}{k}$, where x_* is one of the global optimal solutions. However, in some cases the function f may not be globally L-smooth.

- (5 points) Show that $f(x) = \exp(x)$ is convex and not globally L-smooth.

  Proof. Show that $f(x) = exp(x)$ is convex and not globally L-smooth

Showing $f(x)$ is convex:
Let $f(x) = e^x$

$f(\lambda x + (1-\lambda)y)$
$= e^{(\lambda x + (1-\lambda)y)}$
$= e^{\lambda x} + e^{(1-\lambda)y}$
$= e^{\lambda} e^x + e^{1-\lambda} e^y$
since $ce \leq e^c$ for $c \geq 0$
$\leq \lambda e^x + (1-y)e^y$
By definition of f(x):
$\leq \lambda f(x) + (1 - \lambda f(y)$
This shows $f(x)$ is convex

If we can choose some x and y where $||\nabla f(x) - \nabla f(y)|| \leq \frac{L}{2}||x-y||$, then e^x is not globally smooth
Let $x = 1$ and $y = 2$:
We know $\nabla f(x) = e^x$
Thus by assumption: $||e^1 - e^2|| \leq \frac{L}{2}||1-2||$
$4.67 \leq \frac{L}{2}||1||$, and $L \geq 0$, thus $f(x)$ is not globally L-smooth
This shows $f(x)$ convex but not globally L-smooth

□

- (bonus points: 30 points) Suppose that the function is locally smooth over any bounded set, i.e., there exists a constant $L(C) > 0$ where C is a bounded set, we have $\|\nabla f(x) - \nabla f(y)\| \leq L(C)\|x - y\|$ for any $x, y \in C$. Design a new gradient-based algorithm which can converge to the optimal solution with $O(1/t)$ convergence rate.