# Spring 2022 CS 795 Large Scale Optimization for Machine Learning Assignment 1

Pankaj Kumar Jatav, Md Mushfiqur Rahman

Due date: February 17, 2022, 11:59:59 pm

**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_hw1.pdf`.

## 1 Operations that Preserve Convexity (45 points)

Prove that the following operations preserve convexity of the functions. The domain of the function is $\mathbb{R}^d$.

- (15 points) If $f_1, \ldots, f_n$ are convex functions, $\alpha_1, \ldots, \alpha_n$ are nonnegative scalars, show that $f := \sum_{i=1}^{n} \alpha_i f_i$ is also a convex function.

  **Proof**: To Proof the above theorem, It is enough to proof two cases.

  1. If f is convex then scaler multiply of f is also convex.

     f is convex then it implies
     $f(\lambda x + (1-\lambda)y) <= \lambda f(x) + (1-\lambda)f(y)$
     Multiple the non negative constant $\alpha$ to both side
     $\alpha f(\lambda x + (1-\lambda)y) <= \lambda \alpha f(x) + (1-\lambda)\alpha f(y)$
     **holds**
     Apply same prop to all all function we can say that $\alpha_1 f_1, \alpha_2 f_2, \ldots, \alpha_n f_k$ are also convex.

  2. If f and g both are convex then f + g are convex.

     $f + g(\lambda x + (1-\lambda)y) = f(\lambda x + (1-\lambda)y) + g(\lambda x + (1-\lambda)y)$
     $f(\lambda x + (1-\lambda)y) + g(\lambda x + (1-\lambda)y) <= \lambda f(x) + (1-\lambda)f(y) + \lambda g(x) + (1-\lambda)g(y)$
     $f(\lambda x + (1-\lambda)y) + g(\lambda x + (1-\lambda)y) <= \lambda(f(x) + g(x) + (1-\lambda)f(y) + g(x)$
     $f + g(\lambda x + (1-\lambda)y) <= \lambda f + g(x) + (1-\lambda)f + g(x)$
     **holds**
     Apply same prop to all all function we can say that $f_1 + f_2 + \ldots + f_n$ are also convex.

Combine 1 and 2, we can say $f := \sum_{i=1}^{n} \alpha_i f_i$ is also a convex function.

- (15 points) If $f_\theta$ is a convex function for all $\theta \in \Theta$, show that $f := \sup_{\theta \in \Theta} f_\theta$ is also a convex function.

  **Proof**: Pick $x, y \in f$

  $f(\lambda x + (1-\lambda)y) = f_\theta(\lambda x + (1-\lambda)y)$ ( for any $\theta \in \Theta$ )
  $f_\theta(\lambda x + (1-\lambda)y) <= \lambda f_\theta(x) + (1-\lambda) f_\theta(y)$ (because $f_\theta$ in convex )
  As we know that $f := \sup_{\theta \in \Theta} f_\theta$ then $f(x) >= f_\theta(x)$ for any value of $\theta \in \Theta$
  $f_\theta(\lambda x + (1-\lambda)y) <= \lambda f(x) + (1-\lambda) f(y)$
  $f(\lambda x + (1-\lambda)y) <= \lambda f(x) + (1-\lambda) f(y)$
  **holds**

- (15 points) If $g$ is a convex function, show that for any matrix $A \in \mathbb{R}^{d \times d}$ and vector $b \in \mathbb{R}^d$, the function $f(x) = g(Ax + b)$ is also a convex function. **Proof**:

  let h(x) = Ax + b;
  $h(\lambda x + (1-\lambda)y) = A(\lambda x + (1-\lambda)y) + b$
  $= A\lambda x + Ay - \lambda Ay + b$
  $= A\lambda x + Ay - \lambda Ay + b + \lambda b - \lambda b$
  $= \lambda(Ax + b) + (1-\lambda)(Ax + b)$
  $h(\lambda x + (1-\lambda)y) = \lambda h(x) + (1-\lambda) h(y)$ – (**1**)
  Now we have f(x) = g(h(x))
  $f(\lambda x + (1-\lambda)y) = g(h(\lambda x + (1-\lambda)y))$
  $= g(\lambda h(x) + (1-\lambda) h(y))$ (From **1**)
  As g is convex, to visualize h(x) and h(y) are equivalent to x and y in
  $g(\lambda x + (1-\lambda)y) <= \lambda g(x) + (1-\lambda) g(y)$
  $<= \lambda g(h(x)) + (1-\lambda) g(h(y))$
  $<= \lambda f(x) + (1-\lambda) f(y)$ ( f(x) = g(h(x)) )
  $f(\lambda x + (1-\lambda)y) <= \lambda f(x) + (1-\lambda) f(y)$
  **holds**

## 2 Equivalent Characterizations for $L$-Lipschitz Functions (20 points)

Let $f$ be a convex function and $\mathcal{X}$ be a convex set. $f$ is $L$-Lipschitz if for any $x, y \in \mathcal{X}$ we have $|f(x) - f(y)| \leq L\|x - y\|$. Show that

- (10 points) If for all $x \in \mathcal{X}$ and all subgradients $g \in \partial f(x)$, $\|g\| \leq L$, then $f$ is $L$-Lipschitz.

  **Proof:** By the dual norm

  $$||\mathcal{X}||_* = max\{w, x : ||w|| \leq 1\} \tag{1}$$

The above dual norm implies the inequality such that

$$\langle w, x \rangle \leq ||w||.||x||_*$$

We say that $g_x \in \mathbb{R}^d$ is a subgradient of $f$ at $x \in \mathbb{X}$

$$f(y) \geq f(x) + g_x^\top (y - x) \text{for any } y \in \mathbb{X}$$

$$f(y) - f(x) \geq \langle g, y - x \rangle$$

Multiplying by -1,

$$f(x) - f(y) \leq \langle g, y - x \rangle$$

Using equation 1,

$$f(x) - f(y) \leq ||g||.||x - y||$$

$$\frac{f(x) - f(y)}{||x - y||} \leq ||g|| \tag{2}$$

As it is mentioned that $||g|| \leq L$ which means by equation 2,

$$\frac{f(x) - f(y)}{||x - y||} \leq L$$

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

Here, $f$ is $\mathcal{L}$-Lipschitz.

- (10 points) If $f$ is $L$-Lipschitz, then for all $x \in \text{int}(\mathcal{X})$ and all $g \in \partial f(x)$, $\|g\| \leq L$, where $\text{int}(\cdot)$ denotes the interior of a set [1].

  **Proof**: $L$-Lipschitz is defined by

$$f(x) - f(y) \leq L||x - y|| \tag{3}$$

  By the gradient,

$$f(x) - f(y) \geq \langle g, y - x \rangle \tag{4}$$

  By combining 1 & 2,

$$\langle g, y - x \rangle \leq f(x) - f(y) \leq L||x - y||$$

  By the dual norm,

$$\langle g, y - x \rangle = ||g||.||y - x||$$

$$||g||.||y - x|| \leq f(x) - f(y) \leq L||x - y||$$

  from above, we can say,

$$||g|| \leq L$$

[1] https://en.wikipedia.org/wiki/Interior_(topology)

# 3 Last Iterate Guarantee of Gradient Descent for Convex Lipschitz Functions (35 points)

In this section, our goal is to prove the following Theorem.

**Theorem 1.** *Suppose that $\mathcal{X}$ is a convex such that $\sup_{x,x'\in\mathcal{X}} \|x - x'\| \le R$, the function $f$ is convex and $L$-Lipschitz. Then running projected subgradient descent with learning rate $\eta_s = \frac{R}{L\sqrt{s}}$ for all $s \in \mathbb{N}$. Then we have for any $t \ge 1$:*

$$f(x_t) - \min_{x\in\mathcal{X}} f(x) \le O\left(\frac{RL\log t}{\sqrt{t}}\right).$$

- (15 points) Using similar techniques in the lecture, show that for any $t, k \in \mathbb{N}$, $k \le t$, we have

$$\sum_{s=t-k}^{t} [f(x_s) - f(x_{t-k})] \le O\left(RL\left(\sqrt{t} - \sqrt{t-k-1}\right)\right). \tag{5}$$

**Proof:** Based on the definition of subgradient, we have,

$$f(x_s) - f(x_{t-k}) \le g_s^\top (x_s - x_{t-k})$$

As we know,

$$g_s = \frac{1}{\eta_s}(x_s - y_{s+1})$$

$$f(x_s) - f(x_{t-k}) \le \frac{1}{\eta_s}(x_s - y_{s+1})^\top (x_s - x_{t-k})$$

By the triangle inequality,

$$\begin{aligned} f(x_s) - f(x_{t-k}) \le& \frac{1}{2\eta_s}\left[||x_s - y_{s+1}||^2 + ||x_s - x_{t-k}||^2 - ||y_{s+1} - x_{t-k}||^2\right] \\ \le& \frac{1}{2\eta_s}\left[||x_s - y_{s+1}||^2 - ||y_{s+1} - x_{t-k}||^2\right] + \frac{\eta_s}{2}||g_s||^2 \end{aligned}$$

By the projection lemma, $||y_{s+1} - x_*|| \ge ||x_{s+1} - x_*||$ can be written as $||y_{s+1} - x_{t-k}|| \ge ||x_{s+1} - x_{t-k}||$

$$\begin{aligned} ||g|| \le& L \\ \le& \frac{1}{2\eta_s}\left[||x_s - x_{t-k}||^2 - ||x_{s+1} - x_{t-k}||^2\right] + \frac{\eta_s L^2}{2} \\ \le& \frac{L\sqrt{s}}{2R}\left[||x_s - x_{t-k}||^2 - ||x_{s+1} - x_{t-k}||^2\right] + \frac{RL}{2\sqrt{s}} \end{aligned}$$

Telescoping,

$$\sum_{s=t-k}^{t}[f(x_s)-f(x_s-k)] \leq \frac{L}{2R}\left[||x_{t-k}-x_{t-k}||^2-||x_{t+1}-x_{t-k}||^2\right] sum_{s=t-k}^{t}\sqrt{s}+\frac{RL}{2}\sum_{s=t-k}^{t}\frac{1}{\sqrt{s}}$$

$$\sum_{s=t-k}^{t}[f(x_s)-f(x_s-k)] \leq \frac{L}{2R}\left[-||x_{t+1}-x_{t-k}||^2\right]\sqrt{(t-k-1)}+\frac{RL}{2}\sqrt{t}$$

$$\sum_{s=t-k}^{t}[f(x_s)-f(x_s-k)] \leq -\frac{LR}{2R}\sqrt{(t-k-1)}+\frac{RL}{2}\sqrt{t}$$

$$\sum_{s=t-k}^{t}[f(x_s)-f(x_s-k)] \leq \frac{LR}{2R}(\sqrt{t}-\sqrt{(t-k-1)})$$

**Hence Proved**

- (15 points) For a fixed $t \in \mathbb{N}$, define $S_k = \frac{1}{k+1}\sum_{s=t-k}^{t} f(x_s)$, prove the following using (5):

$$S_{k-1} \leq S_k + O\left(\frac{RL}{k\sqrt{t}}\right). \tag{6}$$

**Proof:**

$$\sum_{s=t-k}^{t} f(x_s) \leq \sum_{s=t-k}^{t} f_{t-k} + RL\left(\sqrt{t}-\sqrt{t-k-1}\right) \tag{7}$$

Dividing by k,

$$S_{k-1} = \frac{1}{k}\sum_{s=t-k}^{t} f(x_s) \leq \frac{1}{k}\sum_{s=t-k}^{t} f_{t-k} + \frac{RL}{k}\left(\sqrt{t}-\sqrt{t-k-1}\right) \tag{8}$$

$$S_k = \frac{1}{k+1}\sum_{s=t-k}^{t} f(x_s) \leq \frac{1}{k+1}\sum_{s=t-k}^{t} f_{t-k} + \frac{RL}{k+1}\left(\sqrt{t}-\sqrt{t-k-1}\right) \tag{9}$$

8 - 9,

$$\begin{aligned}
S_{k-1} - S_k \leq & RL\left(\sqrt{t} - \sqrt{t-k-1}\right)\left(\frac{1}{k} - \frac{1}{k+1}\right) \\
\leq & RL\left(\sqrt{t} - \sqrt{t-k-1}\right)\left(\frac{1}{k(k+1)}\right) \\
\leq & \frac{RL}{k(k+1)}\left(\sqrt{t} - \sqrt{t-k-1}\right)\frac{\sqrt{t} + \sqrt{t-k-1}}{\sqrt{t} + \sqrt{t-k-1}} \\
\leq & \frac{RL}{k(k+1)}\frac{t - (t-k-1)}{\sqrt{t} + \sqrt{t-k-1}} \\
\leq & \frac{RL(k+1)}{k(k+1)(\sqrt{t} + \sqrt{t-k-1})} \\
\leq & \frac{RL}{k(\sqrt{t} + \sqrt{t-k-1})} \\
\leq & \frac{RL}{k\sqrt{t} + k\sqrt{t-k-1}}
\end{aligned}$$

As we know that $t \geq k$, so $k\sqrt{t} > k.\sqrt{t-k-1}$

$$S_{k-1} - S_k \leq \mathcal{O}\left(\frac{RL}{k\sqrt{t}}\right)$$

hence proved

- (5 points) Use above two results to prove Theorem 1.

# 4 Neural Networks are Nonconvex Functions (bonus point: 15 points)

In the course, we introduce convex functions. Now we aim to show that some functions such as 2-layer neural networks are not convex. Suppose we are given data $\{x_i, y_i\}_{i=1}^n$, where $x_i \in \mathbb{R}^d$, $y_i \in \mathbb{R}$, $W_2 \in \mathbb{R}^{d\times h}$ and $w_1 \in \mathbb{R}^h$ and $\sigma : \mathbb{R}^h \to \mathbb{R}^h$ is the sigmoid function [2]. Show that

$$f(w_1, W_2) = \frac{1}{n}\sum_{i=1}^{n}\left(y_i - w_1^\top \sigma(W_2^\top x_i)\right)^2 \tag{10}$$

is nonconvex in $(w_1, W_2)$.

**Proof:**

$$f(w_1, W_2) = \frac{1}{n}\sum_{i=1}^{n}\left(y_i - w_1^\top \sigma(W_2^\top x_i)\right)^2$$

$$\begin{aligned}
\frac{\partial f}{\partial \sigma(W_2^\top x)} = & \frac{2}{n}\sum_{i=1}^{n}\left(y_i - w_1^\top \sigma(W_2^\top x_i)\right)W_2^\top & (11) \\
= & \frac{2}{n}\sum_{i=1}^{n}\left(y_i W_2^\top - w_1^\top . w_1^\top \sigma(W_2^\top x_i)\right) & (12)
\end{aligned}$$

[2] https://en.wikipedia.org/wiki/Sigmoid_function

$$\frac{\partial^2 f}{\partial \left(\sigma(W_2^\top x)\right)^2} = \frac{2}{n}\sum_{i=1}^{n}\left(0 - W_i^\top . w_i^\top)\right)$$

$$\frac{\partial^2 f}{\partial \left(\sigma(W_2^\top x)\right)^2} = -\frac{2}{n}\sum_{i=1}^{n}\left(W_i^\top . w_i^\top)\right) \tag{13}$$

$$g = \sigma(W_2^\top . x)$$

$$\frac{\partial g}{\partial x} = \sigma(W_2^\top . x)(1 - \sigma(W_2^\top . x))(-W_2^\top) \tag{14}$$

$$\frac{\partial^2 g}{\partial x^2} = \sigma(W_2^\top . x)(1 - \sigma(W_2^\top . x))(-W_2^\top)\left(1 - 2\sigma(W_2^\top . x)\right)(-W_2^\top) \tag{15}$$

As we know,

$$\frac{d^2 y}{dx^2} = \frac{d^2 y}{dv^2}.\left(\frac{dv}{dx}\right)^2 + \frac{dy}{dv}.\frac{d^2 v}{dx^2}$$

From equations 12,13,14, and 15,

$\frac{\partial f}{\partial \sigma(W_2^\top x)}$ is positive

$\frac{\partial^2 f}{\partial \left(\sigma(W_2^\top x)\right)^2}$ is negative

$\frac{\partial g}{\partial x}$ is negative

$\frac{\partial^2 g}{\partial x^2}$ is negative

So, from all these, we can see that final output will be negative i.e.

$$\frac{d^2 f}{d(W_2^\top)^2} < 0$$

As per the second order rule, $f$ is a non-convex function