CS446 / ECE 449: Machine Learning, Fall 2020, Homework 1

Name: Saud Alrasheed (Sauda2)

Worked individually

Problem (3)

Solution:

Convexity 1:
We will show that for any convex function $f : \mathbb{R}^n \to \mathbb{R}$, matrix $A \in \mathbb{R}^{n\times m}$ and $b \in \mathbb{R}^n$, we have that $f(Ax + b)$ is convex.

Let $g(x) = f(Ax + b)$, then for all $x, y \in \mathbb{R}^m$ we have that,

$$g(\alpha x + (1-\alpha)y) = f(A(\alpha x + (1-\alpha)y) + b)$$

We have that $b = \alpha b + (1-\alpha)b$. Therefore, we obtain,

$$g(\alpha x+(1-\alpha)y) = f(A(\alpha x+(1-\alpha)y)+b) = f(A(\alpha x+(1-\alpha)y)+\alpha b+(1-\alpha)b) = f(\alpha(Ax+b)+(1-\alpha)(Ay+b))$$

Now, by the convexity of f, we know that $f(\alpha v_1 + (1-\alpha)v_2) \leq \alpha f(v_1) + (1-\alpha)f(v_2)$, for all v_1, $v_2 \in \mathbb{R}^n$. Therefore,

$$g(\alpha x+(1-\alpha)y) = f(\alpha(Ax+b)+(1-\alpha)(Ay+b)) \leq \alpha f(Ax+b)+(1-\alpha)f(Ay+b) = \alpha g(x)+(1-\alpha)g(y)$$

Thus, $g(x) = f(Ax + b)$ is convex.

Convexity 2:
Let $h(x) = f(x) + g(x)$. Therefore, for all x, y, we have that,

$$h(\alpha x + (1-\alpha)y) = f(\alpha x + (1-\alpha)y) + g(\alpha x + (1-\alpha)y)$$

By the λ-strong convexity of f, g we have that,

$$f(\alpha x+(1-\alpha)y)+g(\alpha x+(1-\alpha)y) \leq \alpha f(x)+(1-\alpha)f(y)+\alpha g(x)+(1-\alpha)g(y)-\lambda\alpha(1-\alpha)||y-x||^2$$

But we know that,

$$\alpha f(x)+(1-\alpha)f(y)+\alpha g(x)+(1-\alpha)g(y)-\lambda\alpha(1-\alpha)||y-x||^2 = \alpha(f(x)+g(x))+(1-\alpha)(f(y)+g(y))-\lambda\alpha(1-\alpha)||y-x||^2$$

And,

$$(f(x)+g(x))+(1-\alpha)(f(y)+g(y))-\lambda\alpha(1-\alpha)||y-x||^2 \leq (f(x)+g(x))+(1-\alpha)(f(y)+g(y))-\frac{\lambda\alpha(1-\alpha)}{2}||y-x||^2$$

Thus,

$$h(\alpha x+(1-\alpha)y) \leq \lambda h(x)+(1-\alpha)h(x)-\frac{\lambda\alpha(1-\alpha)}{2}||y-x||^2$$

Thus function h is λ-strongly convex.

Lipschitz continuity and smoothness 1:

We have that for all $w_1, w_2 \in \mathbb{R}^d$,

$$||f(w_1)-f(w_2)|| = ||-yw_1^Tx+yw_2^Tx|| = ||-yx^T(w_1-w_2)||$$

Now, by the Cauchy-Bunyakovsky-Schwarz inequality, we obtain,

$$||f(w_1)-f(w_2)|| = ||-yx^T(w_1-w_2)|| \leq ||-yx^T||\;||(w_1-w_2)|| = \rho||(w_1-w_2)||$$

where $\rho = ||-yx^T||$. Thus, the function is $\rho - Lipschitz$.

Lipschitz continuity and smoothness 2:

Since f, g are Lipschitz functions, we have that for all $x_1, x_2 \in \mathbb{R}$,

$$||f(x_1)-f(x_2)|| \leq \rho_1||(x_1-x_2)|| \quad and \quad ||g(x_1)-g(x_2)|| \leq \rho_2||(x_1-x_2)||$$

Now, let $h(x) = f(x)+g(x)$, then we get,

$$||h(x_1)-h(x_2)|| = ||f(x_1)+g(x_1)-(f(x_2)+g(x_2))|| = ||f(x_1)-f(x_2)+g(x_1)-g(x_2)||$$

Using the triangle inequality, we obtain,

$$||h(x_1)-h(x_2)|| = ||f(x_1)-f(x_2)+g(x_1)-g(x_2)|| \leq ||f(x_1)-f(x_2)||+||g(x_1)-g(x_2)||$$

Thus, since both f, g are Lipschitz, we have,

$$||h(x_1)-h(x_2)|| = \rho_1||(x_1-x_2)||+\rho_2||(x_1-x_2)|| = (\rho_1+\rho_2)||(x_1-x_2)||$$

Thus, the function h is Lipschitz.

Lipschitz continuity and smoothness 3:

First, for any x, y, we have that,

$$f(y)-f(x) \leq \nabla f(y)^T(y-x)-\frac{1}{2\beta}||\nabla f(y)-\nabla f(x)||^2 \tag{1}$$

By the definition of β-smoothness, for any x, y, we have that,

$$f(x) - f(y) \leq \nabla f(y)^T(x-y) + \frac{\beta}{2}||y-x||^2 = -\nabla f(y)^T(y-x) + \frac{\beta}{2}||y-x||^2 \tag{2}$$

Adding (1) to (2), we obtain,

$$\frac{1}{2\beta}||\nabla f(y) - \nabla f(x)||^2 \leq \frac{\beta}{2}||y-x||^2 \quad \Longrightarrow \quad ||\nabla f(y) - \nabla f(x)|| \leq \beta||y-x||$$

Thus, the Lipschitz constant of the gradient is $\rho = \beta$

Convex optimization 1:

Since we are using gradient decent, one can assume that $f(w_k) \leq f(w_{k-1})$, where $w_k = w_{k-1} - \alpha\nabla f(w_{k-1})$. Thus, we obtain,

$$f(w_T) \leq f(w_i) \quad where \;\; 0 \leq i \leq T.$$

Therefore, by averaging all the inequalities we obtain,

$$Tf(w_T) \leq \sum_{i=0}^{T} f(w_i) \implies f(w_T) \leq \frac{1}{T}\sum_{i=0}^{T} f(w_i).$$

Now, by adding $f(u)$ to both sides, we get the following,

$$f(w_T) - f(u) \leq \frac{1}{T}\sum_{i=0}^{T}(f(w_i)) - f(u) = \frac{1}{T}\sum_{i=0}^{T}(f(w_i) - f(u)).$$

Thus, we are done.

Convex optimization 2:

First, we know that $w_t = w_{t-1} - \alpha\nabla f(w_{t-1})$, where $\alpha = \frac{1}{\beta}$. Also, by the definition of β-smoothness, we have that,

$$f(w_t) \leq f(w_{t-1}) + \nabla f(w_{t-1})^T(w_t - w_{t-1}) + \frac{\beta}{2}||w_t - w_{t-1}||^2$$

Thus, by combining both results, we obtain,

$$\begin{aligned}
f(w_t) &\leq f(w_{t-1}) + \nabla f(w_{t-1})^T(-\alpha\nabla f(w_{t-1})) + \frac{\beta}{2}||-\alpha\nabla f(w_{t-1})||^2 \\
&= f(w_{t-1}) - \frac{1}{\beta}\nabla f(w_{t-1})^T\nabla f(w_{t-1}) + \frac{1}{2\beta}||\nabla f(w_{t-1})||^2 \\
&= f(w_{t-1}) - \frac{1}{\beta}||\nabla f(w_{t-1})||^2 + \frac{1}{2\beta}||\nabla f(w_{t-1})||^2 \\
&= f(w_{t-1}) - \frac{1}{2\beta}||\nabla f(w_{t-1})||^2
\end{aligned}$$

Thus,

$$f(w_t) \leq f(w_{t-1}) - \frac{1}{2\beta}||\nabla f(w_{t-1})||^2 \implies ||\nabla f(w_{t-1})||^2 \leq 2\beta(f(w_{t-1}) - f(w_t)) \quad \square$$

Convex optimization 3:

We will start by setting a bound on $||w_t - u||^2$,

$$\begin{aligned} ||w_t - u||^2 &= ||w_{t-1} - \alpha\nabla f(w_{t-1}) - u||^2 \\ &\leq ||w_{t-1} - u||^2 + ||\alpha\nabla f(w_{t-1})||^2 \quad (Triangle\ inequality) \\ &= ||w_{t-1} - u||^2 + \frac{1}{\beta^2}||\nabla f(w_{t-1})||^2 \\ &\leq ||w_{t-1} - u||^2 + \frac{2}{\beta}(f(w_{t-1}) - f(w_t)) \quad (From\ ConvexOptimization.2) \end{aligned}$$

Thus, rearranging the inequality gives,

$$f(w_t) - f(w_{t-1}) \leq \frac{\beta}{2}(||w_{t-1} - u||^2 - ||w_t - u||^2)$$

Also, from 1, one can obtain that,

$$f(w_t) - f(u) \leq \frac{\beta}{2}(||w_{t-1} - u||^2 - ||w_t - u||^2)$$

Now, since we know that for each k, $1 \leq k \leq T$, we have,

$$f(w_k) - f(u) \leq \frac{\beta}{2}(||w_{k-1} - u||^2 - ||w_k - u||^2)$$

Then, by summing and averaging all possible inequalities for each k, we obtain,

$$\frac{1}{T}\sum_{i=1}^{T}(f(w_i) - f(u)) \leq \frac{\beta}{2T}(||w_0 - u||^2 - ||w_T - u||^2)$$