# ASSIGNMENT 1

An assignment report
submitted to **Prof. Ganesh Ramakrishnan**
in the subject of Optimization in Machine Learning

by

**Sandarbh Yadav**
**(22D0374)**

**INDIAN INSTITUTE OF TECHNOLOGY BOMBAY**

**2023**

Q 1. Is $f(x)$ a convex function **True/False**? Give reasons for your answers.

(a) $f(x) = (\det X)^{1/n}$ on $\operatorname{dom} f = \mathbf{S}_{++}^n$

**False**. $f(x)$ is actually concave.
Suppose $h(s) = f(Y + sU)$ such that $Y \succ 0$ and $U \in \mathbf{S}^n$.

$$\begin{aligned} h(s) &= (\det(Y + sU))^{1/n} \\ &= \left(\det Y^{1/2} \det\left(I + sY^{-1/2}UY^{-1/2}\right) \det Y^{1/2}\right)^{1/n} \\ &= (\det Y)^{1/n} \left(\prod_{i=1}^{n} (1 + s\lambda_i)\right)^{1/n} \end{aligned}$$

where $\lambda_i$ denotes the eigenvalues of $Y^{-1/2}UY^{-1/2}$. Clearly, $h$ is a concave function of $s$ on $\{s \mid Y + sU \succ 0\}$ because $\det Y > 0$ and geometric mean is concave.

(b) $f(x_1, x_2) = x_1/x_2$ on $\mathbf{R}_{++}^2$

**False**. Hessian of $f(x)$ is not positive semidefinite, so it is not convex.

$$\nabla^2 f(x) = \begin{bmatrix} 0 & -1/x_2^2 \\ -1/x_2^2 & 2x_1/x_2^3 \end{bmatrix}$$

(c) $f(x_1, x_2) = 1/(x_1 x_2)$ on $\mathbf{R}_{++}^2$

**True**. Hessian of $f(x)$ is positive semidefinite, so it is convex.

$$\nabla^2 f(x) = \frac{1}{x_1 x_2} \begin{bmatrix} 2/\left(x_1^2\right) & 1/(x_1 x_2) \\ 1/(x_1 x_2) & 2/x_2^2 \end{bmatrix} \succeq 0$$

(d) $f : \mathbf{R}^n \to \mathbf{R}$, $f(x) = \max_{i=1,2,\ldots,k} \left\|A^{(i)}x - b^{(i)}\right\|$, where $A^{(i)} \in \mathbf{R}^{m \times n}$, $b^{(i)} \in \mathbf{R}^m$, and $\|\cdot\|$ is a norm on $\mathbf{R}^m$.

**True**. The composition of a norm and affine transformation is convex. $f(x)$ calculates the pointwise maximum of such compositions, so it is also convex.

(e) Gaussian distribution function $f$

$$f(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-u^2/2} du$$

**False**. $f(x)$ is actually log-concave since $f''(x)f(x) \leq (f'(x))^2$ where $f'(x) = \frac{\exp^{-\frac{x^2}{2}}}{\sqrt{2\pi}}$ and $f''(x) = \frac{-x\exp^{-\frac{x^2}{2}}}{\sqrt{2\pi}}$.

Q 2. Consider $f$ to be a convex function, $\lambda_1 > 0, \lambda_i \leq 0$ for $i = 2, \ldots, n$ and $\sum_i \lambda_i = 1$. Let dom($f$) be affine, and for $x_1, \ldots, x_n \in \text{dom}(f)$, show that the inequality always holds:

$$f\left(\sum_{i=1}^{n} \lambda_i x_i\right) \geq \sum_{i=1}^{n} \lambda_i f(x_i)$$

**Solution:** Suppose $\gamma_1 = 1$ & $\gamma_i = -\lambda_i$ for i = 2,...,n. This means that $\gamma_i \geq 0$ for i = 1,...,n. It is given that $\lambda_1 > 0$, so we have

$$\frac{\gamma_i}{\lambda_1} \geq 0$$

It is also given that $\sum_{i=1}^{n} \lambda_i = 1$. Now, we have

$$\lambda_1 + \sum_{i=2}^{n} \lambda_i = 1$$

$$\lambda_1 = 1 - \sum_{i=2}^{n} \lambda_i$$

Since, we have defined $\gamma_i = -\lambda_i$ for i = 2,...,n & $\gamma_1 = 1$, we have

$$\lambda_1 = 1 + \sum_{i=2}^{n} \gamma_i$$

$$\lambda_1 = \sum_{i=1}^{n} \gamma_i$$

$$\sum_{i=1}^{n} \frac{\gamma_i}{\lambda_1} = 1$$

It is given that $f$ is a convex function. Using Jensen's inequality, we have

$$f\left(\sum_{i=1}^{n} \frac{\gamma_i}{\lambda_1} z_i\right) \leq \sum_{i=1}^{n} \frac{\gamma_i}{\lambda_1} f(z_i)$$

Taking $z_1 = \sum_{i=1}^{n} \lambda_i x_i$ and $z_i = x_i$, for i = 2,...,n where $x_i$ & $z_i$ belong to domain of $f$. This gives

$$f\left(\frac{\gamma_1}{\lambda_1} \sum_{i=1}^{n} \lambda_i x_i + \sum_{i=2}^{n} \frac{\gamma_i}{\lambda_1} x_i\right) \leq \frac{\gamma_1}{\lambda_1} f\left(\sum_{i=1}^{n} \lambda_i x_i\right) + \frac{\gamma_2}{\lambda_1} f(x_2) + \ldots + \frac{\gamma_n}{\lambda_1} f(x_n)$$

Substituting $\gamma_1 = 1$ & $\gamma_i = -\lambda_i$ for i = 2,...,n

$$f\left(\frac{1}{\lambda_1} \sum_{i=1}^{n} \lambda_i x_i - \frac{1}{\lambda_1} \sum_{i=2}^{n} \lambda_i x_i\right) \leq \frac{1}{\lambda_1} f\left(\sum_{i=1}^{n} \lambda_i x_i\right) - \frac{\lambda_2}{\lambda_1} f(x_2) - \ldots - \frac{\lambda_n}{\lambda_1} f(x_n)$$

$$f(x_1) \leq \frac{1}{\lambda_1} f\left(\sum_{i=1}^{n} \lambda_i x_i\right) - \frac{\lambda_2}{\lambda_1} f(x_2) - \ldots - \frac{\lambda_n}{\lambda_1} f(x_n)$$

$$\lambda_1 f(x_1) \leq f\left(\sum_{i=1}^{n} \lambda_i x_i\right) - \lambda_2 f(x_2) - \ldots - \lambda_n f(x_n)$$

$$\lambda_1 f(x_1) + \lambda_2 f(x_2) + \ldots + \lambda_n f(x_n) \leq f\left(\sum_{i=1}^{n} \lambda_i x_i\right)$$

$$\sum_{i=1}^{n} \lambda_i f(x_i) \leq f\left(\sum_{i=1}^{n} \lambda_i x_i\right)$$

Hence proved.

Q 3. We say that a function $f$ is log-convex on the real interval $\mathcal{D} = [a, b]$ if $\forall x, y \in \mathcal{D}$ and $\lambda \in [0, 1]$, the function satisfies

$$f(\lambda x + (1-\lambda)y) \leq f^{\lambda}(x) f^{1-\lambda}(y)$$

We will show that for an increasing log-convex function $f : \mathcal{D} \to \mathbb{R}$ and $0 \leq t \leq 1$,

$$f\left(\frac{a+b}{2}\right) \leq \phi(a,b) \leq \frac{1}{b-a}\int_a^b f(x)dx \leq \psi(a,b,t) \leq \mathcal{L}(f(a), f(b)) \leq \frac{f(a)+f(b)}{2}$$

where

$$\phi(a,b) = \sqrt{f\left(\frac{3a+b}{4}\right) f\left(\frac{a+3b}{4}\right)}$$

$$\mathcal{L}(a,b) = \frac{a-b}{\ln\left(\frac{a}{b}\right)}$$

$$\psi(a,b,t) = (1-t)\mathcal{L}(f(ta+(1-t)b), f(a)) + t\mathcal{L}(f(b), f(ta+(1-t)b))$$

(a) First, we prove the following inequalities -

(i) For $0 < t < 1$, the following holds

$$t^t (1-t)^{1-t} \geq \frac{1}{2}$$

**Solution:** A log-convex function satisfies

$$f(\lambda x + (1-\lambda)y) \leq f^{\lambda}(x) f^{1-\lambda}(y)$$

Let $f(z) = 1/z$. Clearly, f is log-convex because $log(1/z)$ is convex. Since $0 < t < 1$, we can take $\lambda = t$. Taking points $x = 1/t$ and $y = 1/(1-t)$, we get

$$f(t\frac{1}{t} + (1-t)\frac{1}{1-t}) \leq f^t(\frac{1}{t}) f^{1-t}(\frac{1}{1-t})$$

$$f(2) \leq t^t(1-t)^{1-t}$$

Clearly, $f(2) = 1/2$ by our choice of $f$. So, we have

$$t^t(1-t)^{1-t} \geq \frac{1}{2}$$

Hence proved.

(ii) For $0 < a < b$ and $0 \leq t \leq 1$, the following holds

$$\sqrt{ab} \geq \begin{cases} a^{1-t}b^t & t \leq \frac{1}{2} \\ a^t b^{1-t} & t > \frac{1}{2} \end{cases}$$

and

$$\sqrt{ab} \leq \frac{a^{1-t}b^t + a^t b^{1-t}}{2} \leq \frac{a+b}{2}$$

**Solution:** Suppose that $t \leq \frac{1}{2}$.

This implies

$$0 \leq \frac{1}{2} - t$$

It is given that $b > a$, so we have

$$b^{\frac{1}{2}-t} \geq a^{\frac{1}{2}-t}$$

On rearranging we get

$$a^{\frac{1}{2}}b^{\frac{1}{2}} \geq a^{1-t}b^t$$

Thus, for $t \leq \frac{1}{2}$, we have shown that $\sqrt{ab} \geq a^{1-t}b^t$

Now consider the case when $t > \frac{1}{2}$. This gives

$$t - \frac{1}{2} > 0$$

Again using the fact that $b > a$, we get

$$b^{t-\frac{1}{2}} \geq a^{t-\frac{1}{2}}$$

On rearranging we get

$$a^{\frac{1}{2}}b^{\frac{1}{2}} \geq a^t b^{1-t}$$

Thus, for $t > \frac{1}{2}$, we have shown that $\sqrt{ab} \geq a^t b^{1-t}$

Hence, first inequality is proved. Now, we will prove the second inequality.

Consider LHS of the second inequality. We will prove this by completing the square methodology.

$$\frac{a^{1-t}b^t + a^t b^{1-t}}{2} = \frac{1}{2}\left(\left(a^{\frac{1-t}{2}}b^{\frac{t}{2}}\right)^2 + \left(a^{\frac{t}{2}}b^{\frac{1-t}{2}}\right)^2\right)$$

$$= \frac{1}{2}\left(\left(a^{\frac{1-t}{2}}b^{\frac{t}{2}}\right)^2 + \left(a^{\frac{t}{2}}b^{\frac{1-t}{2}}\right)^2 - 2\sqrt{ab} + 2\sqrt{ab}\right)$$

$$= \frac{1}{2}\left(\left(a^{\frac{1-t}{2}}b^{\frac{t}{2}} - a^{\frac{t}{2}}b^{\frac{1-t}{2}}\right)^2 + 2\sqrt{ab}\right)$$

Since a squared term is always non-negative, we have

$$= \frac{1}{2}\left(a^{\frac{1-t}{2}}b^{\frac{t}{2}} - a^{\frac{t}{2}}b^{\frac{1-t}{2}}\right)^2 + \sqrt{ab} \geq \sqrt{ab}$$

Hence,

$$\sqrt{ab} \leq \frac{a^{1-t}b^t + a^t b^{1-t}}{2}$$

We have proved LHS of the second inequality. Now, consider RHS of the second inequality. From the first inequality, we have

$$\sqrt{ab} \geq \begin{cases} a^{1-t}b^t, & t \leq \frac{1}{2} \\ a^t b^{1-t}, & t > \frac{1}{2} \end{cases}$$

Adding them, we get

$$a^{1-t}b^t + a^t b^{1-t} \leq 2\sqrt{ab}$$

$$\frac{a^{1-t}b^t + a^t b^{1-t}}{2} \leq \sqrt{ab}$$

The AM-GM inequality states

$$\sqrt{ab} \leq \frac{a+b}{2}$$

Combining both, we have

$$\frac{a^{1-t}b^t + a^t b^{1-t}}{2} \leq \frac{a+b}{2}$$

Hence, we have proved RHS of the second inequality too.

(b) Show that if $f$ is a positive log-convex function on $[a, b]$, then

$$f\left(\frac{a+b}{2}\right) \leq \frac{1}{b-a}\int_a^b f(x)dx \leq \frac{f(a)+f(b)}{2}$$

**Solution:** Consider LHS of the inequality first. We break the integral in two intervals from $a$ to $\frac{a+b}{2}$ and from $\frac{a+b}{2}$ to $b$.

$$\frac{1}{b-a}\int_a^b f(x)dx = \frac{1}{b-a}\left[\int_a^{(a+b)/2} f(x)dx + \int_{(a+b)/2}^b f(x)dx\right]$$

In the first integral from $a$ to $(a+b)/2$, we take

$$u = \frac{2x-(a+b)}{-(b-a)}$$

$$dx = \frac{-(b-a)}{2}du$$

In the second integral from $(a+b)/2$ to $b$, we take

$$u = \frac{2x-(a+b)}{(b-a)}$$

$$dx = \frac{b-a}{2}du$$

Applying the change of variable, we get

$$= \int_0^1 \frac{1}{2}[f((a+b)/2 - u(b-a)/2) + f((a+b)/2 + u(b-a)/2)du]$$

$$\geq \int_0^1 f((a+b)/2)du = f\left(\frac{a+b}{2}\right)$$

Hence,

$$f\left(\frac{a+b}{2}\right) \leq \frac{1}{b-a}\int_a^b f(x)dx$$

We have proved LHS of the inequality. Now, consider RHS of the inequality. It is given that $f$ is a log-convex function. We know that a log-convex function is convex too. So, we have

$$f(x) \leq f(a) + \frac{f(b) - f(a)}{b - a}(x - a)$$

Integrating on both sides from $a$ to $b$, we get

$$\int_a^b f(x)dx \leq f(a)(b - a) + \frac{f(b) - f(a)}{b - a}\left[\int_a^b xdx - a(b - a)\right]$$

$$= f(a)(b - a) + \frac{f(b) - f(a)}{b - a}\left[\frac{(b^2 - a^2)}{2} - a(b - a)\right]$$

$$= f(a)(b - a) + \frac{f(b) - f(a)}{2}(b - a)$$

$$= (b - a)\frac{f(a) + f(b)}{2}$$

Finally,

$$\int_a^b f(x)dx \leq (b - a)\frac{f(a) + f(b)}{2}$$

$$\frac{1}{b - a}\int_a^b f(x)dx \leq \frac{f(a) + f(b)}{2}$$

Hence, we have proved RHS of the inequality too.

(c) Finally, show that $\sqrt{ab} \leq \mathcal{L}(a, b) \leq \frac{a+b}{2}$ and prove the required inequality.

**Solution:** Consider LHS of the inequality first. From Cauchy-Schwarz inequality, we have

$$\left(\int_a^b f(t)g(t)dt\right)^2 \leq \left(\int_a^b f^2(t)dt\right)\left(\int_a^b g^2(t)dt\right)$$

Suppose, $f(t) = 1/t$ and $g(t) = 1$.

$$\left(\int_a^b (1/t)dt\right)^2 \leq \int_a^b (1/t)^2 dt \int_a^b 1dt$$

$$(\ln b - \ln a)^2 \leq \left(\frac{1}{a} - \frac{1}{b}\right)(b - a) = \frac{(b - a)^2}{ab}$$

It is given that b $\geq$ a as they form a closed interval. So, we have

$$\ln\left(\frac{b}{a}\right) \leq \frac{(b-a)}{\sqrt{ab}}$$

$$\sqrt{ab} \leq \frac{a-b}{\ln\frac{a}{b}}$$

$$\sqrt{ab} \leq \mathcal{L}(a,b)$$

We have proved LHS of the inequality. Now, consider RHS of the inequality. Suppose, $f(t) = \frac{1}{\sqrt{t}}$ and $g(t) = \sqrt{t}$ in Cauchy-Schwarz inequality.

$$\left(\int_a^b dt\right)^2 \leq \int_a^b (1/t)dt \int_a^b tdt$$

$$(b-a)^2 \leq (\ln b - \ln a)\left(\frac{b^2-a^2}{2}\right)$$

$$\frac{a-b}{\ln\frac{a}{b}} \leq \frac{(a+b)}{2}$$

$$\mathcal{L}(a,b) \leq \frac{a+b}{2}$$

Hence, we have proved RHS of the inequality too. Now, we will prove the required inequality for increasing log-convex function.

$$f\left(\frac{a+b}{2}\right) \leq \phi(a,b) \leq \frac{1}{b-a}\int_a^b f(x)dx \leq \mathcal{L}(f(a),f(b)) \leq \frac{f(a)+f(b)}{2}$$

First consider the leftmost inequality.

$$f\left(\frac{a+b}{2}\right) \leq \phi(a,b)$$

Starting from LHS, we have

$$f\left(\frac{a+b}{2}\right) = f\left(\frac{4a+4b}{8}\right)$$

$$= f\left(\frac{3a+b}{8} + \frac{a+3b}{8}\right)$$

$$= f\left(\frac{1}{2}\left(\frac{3a+b}{4}\right) + \frac{1}{2}\left(\frac{a+3b}{4}\right)\right)$$

It is given that $f$ is log-convex. So, we have

$$f\left(\frac{1}{2}\left(\frac{3a+b}{4}\right)+\frac{1}{2}\left(\frac{a+3b}{4}\right)\right) \leq f\left(\frac{3a+b}{4}\right)^{\frac{1}{2}} f\left(\frac{a+3b}{4}\right)^{\frac{1}{2}}$$

Finally,

$$f\left(\frac{a+b}{2}\right) \leq \sqrt{f\left(\frac{3a+b}{4}\right) f\left(\frac{a+3b}{4}\right)}$$

$$f\left(\frac{a+b}{2}\right) \leq \phi(a,b)$$

Hence proved.

Now, consider the rightmost inequality.

$$\mathcal{L}(f(a), f(b)) \leq \frac{f(a)+f(b)}{2}$$

We have already shown above that

$$\mathcal{L}(a,b) \leq \frac{a+b}{2}$$

It is given that $f$ is an increasing function. Hence, the inequality holds true on substituting $f(a)$ and $f(b)$ in place of $a$ and $b$.

$$\mathcal{L}(f(a), f(b)) \leq \frac{f(a)+f(b)}{2}$$

Hence proved.

Now, consider the inner left inequality.

$$\phi(a,b) \leq \frac{1}{b-a}\int_a^b f(x)dx$$

In part (b), we have already shown that

$$f\left(\frac{a+b}{2}\right) \leq \frac{1}{b-a}\int_a^b f(x)dx$$

Now, we use this inequality on intervals $\left[a, \frac{a+b}{2}\right]$ & $\left[\frac{a+b}{2}, b\right]$ to get

$$f\left(\frac{3a+b}{4}\right) \leq \frac{2}{b-a}\int_a^{\frac{(a+b)}{2}} f(x)dx \text{ and } f\left(\frac{a+3b}{4}\right) \leq \frac{2}{b-a}\int_{\frac{(a+b)}{2}}^b f(x)dx$$

On adding these two inequalities together, we get

$$\frac{1}{2}\left[f\left(\frac{3a+b}{4}\right)+f\left(\frac{a+3b}{4}\right)\right] \leq \frac{1}{b-a}\int_a^b f(x)dx$$

Using AM-GM inequality on LHS, we get

$$\sqrt{f\left(\frac{3a+b}{4}\right)f\left(\frac{a+3b}{4}\right)} \leq \frac{1}{2}\left[f\left(\frac{3a+b}{4}\right)+f\left(\frac{a+3b}{4}\right)\right] \leq \frac{1}{b-a}\int_a^b f(x)dx$$

Finally,

$$\phi(a,b) \leq \frac{1}{b-a}\int_a^b f(x)dx$$

Hence proved.

Q 4. Let $\mathcal{D}=[a,b]$ and $f:\mathcal{D}\to\mathbb{R}$ be a convex or concave $\mathcal{C}^2$ class function. Show that if $|f'(x)|\geq\zeta$ for all $x\in\mathcal{D}$ and $\zeta>0$, then

$$\left|\int_a^b e^{\iota f(x)}dx\right| \leq \frac{2}{\zeta}$$

where $\iota=\sqrt{-1}$

**Solution:** Consider LHS of the inequality.

$$\left|\int_a^b e^{\iota f(x)}dx\right|$$

We know that $e^{\iota\phi}=\cos\phi+\iota\sin\phi$. Using this, we can write

$$\left|\int_a^b \cos f(x)+\iota\sin f(x)dx\right|$$

Substitute $f(x)=y$ such that $\frac{df(x)}{dx}=\frac{dy}{dx}$ i.e., $dx=\frac{dy}{f'(x)}$.

$$\left|\int_{f(a)}^{f(b)}\frac{\cos y+\iota\sin y}{f'(x)}dy\right|$$

It is given that $|f'(x)|\geq\zeta$. So, $\frac{1}{|f'(x)|}\leq\frac{1}{\zeta}$. Using this inequality, we have

$$\left|\int_{f(a)}^{f(b)}\frac{\cos y+\iota\sin y}{f'(x)}dy\right| \leq \frac{1}{\zeta}\left|\int_{f(a)}^{f(b)}(\cos y+\iota\sin y)dy\right|$$

$$= \frac{1}{\zeta} |\cos f(b) + \iota \sin f(b) - \cos f(a) - \iota \sin f(a)|$$

$$= \frac{1}{\zeta} |\cos f(b) - \cos f(a) + \iota(\sin f(b) - \sin f(a))|$$

$$= \frac{1}{\zeta} \left[ (\cos f(b) - \cos f(a))^2 + (\sin f(b) - \sin f(a))^2) \right]^{1/2}$$

$$= \frac{1}{\zeta} \left( \left( \cos^2 f(b) - 2\cos f(b) \cos f(a) + \cos^2 f(a) \right) + \left( \sin^2 f(b) - 2\sin f(b) \sin f(a) + \sin^2 f(a) \right) \right)^{1/2}$$

$$= \frac{1}{\zeta} [2 - 2(\cos f(b) \cos f(a) + \sin f(b) \sin f(a))]^{1/2}$$

$$= \frac{1}{\zeta} [2 - 2\cos(f(b) - f(a))]^{1/2}$$

$$= \frac{1}{\zeta} [2(1 - \cos(f(b) - f(a)))]^{1/2}$$

Now, since the minimum value of cosine function is -1, the maximum value of term inside square bracket is 4. Using this, we have

$$\frac{1}{\zeta} [2(1 - \cos(f(b) - f(a)))]^{1/2} \leq \frac{4^{1/2}}{\zeta} = \frac{2}{\zeta}$$

Finally, we have

$$\left| \int_a^b e^{\iota f(x)} dx \right| \leq \frac{2}{\zeta}$$

Hence proved.

Q 5. The basic idea behind many reinforcement learning algorithms is to estimate the action-value function $Q^*(s, a)$ by using the Bellman equation as an iterative update,

$$Q_{i+1}(s, a) = \mathbb{E}_{s'} \left[ r + \gamma \max_{a'} Q_i (s', a') \mid s, a \right]$$

where $\{a\}$ are the actions, $\{s\}$ are the states, $r$ is the reward and $\gamma$ is a discounting factor. In practice, such iterative methods converge to the optimal value function as $i \to \infty$.

It is seen that, this is infeasible and a neural network $Q(s, a, \theta)$ is used as an approximator to estimate this optimal action-value function as $Q(s, a; \theta) \approx Q^*(s, a)$. During training, we minimize the mean-squared error in the Bellman equation, and the loss function of such a network is given as

$$L_i(\theta_i) = \mathbb{E}_{(s,a,r,s')\sim U(D)}\left[\left(r + \gamma \max_{a'} Q\left(s', a'; \theta_i^-\right) - Q\left(s, a; \theta_i\right)\right)^2\right]$$

where $\mathbf{e} = (s, a, r, s')$ are the experiences forming the dataset $D$. It is known that $\theta_i^-$ is fixed. Find the gradient of the above loss function w.r.t $\theta_i$.

**Solution:** The gradient of above loss function w.r.t $\theta_i$ is

$$\nabla_{\theta_i} L(\theta_i) = \nabla_{\theta_i}\mathbb{E}_{(s,a,r,s')\sim U(D)}\left[\left(r + \gamma \max_{a'} Q\left(s', a'; \theta_i^-\right) - Q\left(s, a; \theta_i\right)\right)^2\right]$$

The derivative of expectation of a function is expectation of derivative of the function.

$$= \mathbb{E}_{(s,a,r,s')\sim U(D)}\left[\nabla_{\theta_i}\left(r + \gamma \max_{a'} Q\left(s', a'; \theta_i^-\right) - Q\left(s, a; \theta_i\right)\right)^2\right]$$

Using chain rule, we get

$$= \mathbb{E}_{(s,a,r,s')\sim U(D)}\left[2\left(r + \gamma \max_{a'} Q\left(s', a'; \theta_i^-\right) - Q\left(s, a; \theta_i\right)\right)\nabla_{\theta_i}\left(r + \gamma \max_{a'} Q\left(s', a'; \theta_i^-\right) - Q\left(s, a; \theta_i\right)\right)\right]$$

The term $r + \gamma \max_{a'} Q\left(s', a'; \theta_i^-\right)$ is independent of $\theta_i$. So we get

$$= -2\mathbb{E}_{(s,a,r,s')\sim U(D)}\left[\left(r + \gamma \max_{a'} Q\left(s', a'; \theta_i^-\right) - Q\left(s, a; \theta_i\right)\right)\nabla_{\theta_i} Q\left(s, a; \theta_i\right)\right]$$

This is the required gradient. One thing to note is that the coefficient -2 is often merged with the learning rate. So, quite often, gradient is mentioned without the coefficient.

Q 6. Let $x_1, \ldots, x_n$ be non-negative points, and $p_1, \ldots, p_n$ be positive numbers such that $\sum_i p_i = 1$. Define a non-decreasing convex function $f : \mathrm{conv}\{x_1, \ldots, x_n\} \to \mathbb{R}$. Then show that

(a)

$$\prod_{i=1}^{n} x_i^{p_i} \leq \sum_{i=1}^{n} p_i x_i \leq \sum_{i=1}^{n} x_i - (n-1)\prod_{i=1}^{n} x_i^{\frac{1-p_i}{n-1}}$$

**Solution:** Consider LHS of the inequality first. We know that logarithm is concave. Using Jensen's inequality, we have

$$\sum_{i=1}^{n} p_i \log(x_i) \leq \log\left(\sum_{i=1}^{n} p_i x_i\right)$$

Using properties of log, we get

$$\sum_{i=1}^{n} \log\left(x_i^{p_i}\right) \leq \log\left(\sum_{i=1}^{n} p_i x_i\right)$$

$$\log \prod_{i=1}^{n} x_i^{p_i} \leq \log \left( \sum_{i=1}^{n} p_i x_i \right)$$

We also know that log is a strictly increasing function, so we have

$$\prod_{i=1}^{n} x_i^{p_i} \leq \sum_{i=1}^{n} p_i x_i$$

We have proved LHS of the inequality. Now, consider RHS of the inequality. Using LHS of the inequality, we have

$$\prod_{i=1}^{n} x_i^{\lambda_i} \leq \sum_{i=1}^{n} \lambda_i x_i$$

Putting $\lambda_i = \frac{1-p_i}{n-1}$ as $\sum_{i=1}^{n} \lambda_i = 1$.

$$\prod_{i=1}^{n} x_i^{\frac{1-p_i}{n-1}} \leq \sum_{i=1}^{n} \frac{1-p_i}{n-1} x_i$$

$$\left( x_1^{\frac{1-p_1}{n-1}} \ldots x_n^{\frac{1-p_n}{n-1}} \right) \leq \frac{1-p_1}{n-1} x_1 + \frac{1-p_2}{n-1} x_2 + \frac{1-p_3}{n-1} x_3 + \ldots + \frac{1-p_n}{n-1} x_n$$

$$\left( x_1^{\frac{1-p_1}{n-1}} \ldots x_n^{\frac{1-p_n}{n-1}} \right) \leq \frac{x_1 (1-p_1) + \ldots + x_n (1-p_n)}{n-1}$$

$$\left( x_1^{\frac{1-p_1}{n-1}} \ldots x_n^{\frac{1-p_n}{n-1}} \right) (n-1) \leq x_1 (1-p_1) + \ldots + x_n (1-p_n)$$

$$\left( x_1^{\frac{1-p_1}{n-1}} \ldots x_n^{\frac{1-p_n}{n-1}} \right) (n-1) + p_1 x_1 + p_2 x_2 + \ldots + p_n x_n \leq x_1 + x_2 + \cdots + x_n$$

$$\sum_{i=1}^{n} p_i x_i \leq \sum_{i=1}^{n} x_i - (n-1) \prod_{i=1}^{n} x_i^{\frac{1-p_i}{n-1}}$$

Hence, we have proved RHS of the inequality too.

(b)

$$f\left( \prod_{i=1}^{n} x_i^{p_i} \right) \leq \sum_{i=1}^{n} p_i f(x_i) \leq \sum_{i=1}^{n} f(x_i) - (n-1) f\left( \prod_{i=1}^{n} x_i^{\frac{1-p_i}{n-1}} \right)$$

**Solution:** It is given that $f$ is a non-decreasing function. This means that $f(y) \geq f(x)$ for $y > x$.

Consider LHS of the inequality first. We have already shown that

$$\prod_{i=1}^{n} x_i^{p_i} \leq \sum_{i=1}^{n} p_i x_i$$

Using non-decreasing function $f$ preserves the inequality. So, we have

$$f\left(\prod_{i=1}^{n} x_i^{p_i}\right) \leq f\left(\sum_{i=1}^{n} p_i x_i\right)$$

It is also given that $f$ is convex. Using Jensen's inequality, we get

$$f\left(\prod_{i=1}^{n} x_i^{p_i}\right) \leq \sum_{i=1}^{n} p_i f\left(x_i\right)$$

We have proved LHS of the inequality. Now, consider RHS of the inequality. From part (a), since $\sum_{i=1}^{n} \lambda_i = 1$, we have

$$\prod_{i=1}^{n} x_i^{\lambda_i} \leq \sum_{i=1}^{n} \lambda_i x_i$$

Using non-decreasing function $f$ preserves the inequality. So, we have

$$f\left(\prod_{i=1}^{n} x_i^{\lambda_i}\right) \leq f\left(\sum_{i=1}^{n} \lambda_i x_i\right)$$

Again, putting $\lambda_i = \frac{1-p_i}{n-1}$ as $\sum_{i=1}^{n} \lambda_i = 1$.

$$f\left(\prod_{i=1}^{n} x_i^{\frac{1-p_i}{n-1}}\right) \leq f\left(\sum_{i=1}^{n} \frac{1-p_i}{n-1} x_i\right)$$

Since $f$ is convex, it satisfies Jensen's inequality.

$$f\left(\sum_{i=1}^{n} \lambda_i x_i\right) \leq \sum_{i=1}^{n} \lambda_i f\left(x_i\right)$$

Putting $\lambda_i = \frac{1-p_i}{n-1}$ as $\sum_{i=1}^{n} \lambda_i = 1$.

$$f\left(\sum_{i=1}^{n} \frac{1-p_i}{n-1} x_i\right) \leq \sum_{i=1}^{n} \frac{1-p_i}{n-1} f\left(x_i\right)$$

Combining the above 2 inequalities, we get

$$f\left(\prod_{i=1}^{n} x_i^{\frac{1-p_i}{n-1}}\right) \leq \sum_{i=1}^{n} \frac{1-p_i}{n-1} f\left(x_i\right)$$

After rearranging, we get

$$\sum_{i=1}^{n} p_i f\left(x_i\right) \leq \sum_{i=1}^{n} f\left(x_i\right)-(n-1) f\left(\prod_{i=1}^{n} x_i^{\frac{1-p_i}{n-1}}\right)$$

Hence, we have proved RHS of the inequality too.

## Q 7.

(a) Show that the following definitions are equivalent:

A function $f$ is $L$-smooth with Lipschitz constant $L > 0$, if

- $\forall \mathbf{x}, \boldsymbol{y} \in \operatorname{dom}(f), \|\nabla f(\mathbf{x})-\nabla f(\boldsymbol{y})\| \leq L\|\mathbf{x}-\boldsymbol{y}\|$ (i.e, $\nabla f$ is $L$-Lipschitz continuous)
- a quadratic function upper bounds $f$, i.e, $|f(\boldsymbol{y})-f(\mathbf{x})-\langle\nabla f(\mathbf{x}), \boldsymbol{y}-\mathbf{x}\rangle| \leq \frac{L}{2}\|\mathbf{x}-\boldsymbol{y}\|_2^2$

[Hint: Try to express $f(\boldsymbol{y})-f(\mathbf{x})$ as an integral.]

**Solution:** Using the hint, let's express $f(\boldsymbol{y})-f(\mathbf{x})$ as a definite integral from 0 to 1 as

$$\begin{aligned} f(\boldsymbol{y})-f(\boldsymbol{x}) &= \int_0^1\langle\nabla f(\boldsymbol{x}+\theta(\boldsymbol{y}-\boldsymbol{x})), \boldsymbol{y}-\boldsymbol{x}\rangle d\theta \\ &= \int_0^1\langle\nabla f(\boldsymbol{x}+\theta(\boldsymbol{y}-\boldsymbol{x}))-\nabla f(\boldsymbol{x})+\nabla f(\boldsymbol{x}), \boldsymbol{y}-\boldsymbol{x}\rangle d\theta \end{aligned}$$

Taking $\nabla f(\boldsymbol{x})$ out of integral since it is independent of $\theta$, we get

$$\begin{aligned} f(\boldsymbol{y})-f(\boldsymbol{x}) &= \langle\nabla f(\boldsymbol{x}), \boldsymbol{y}-\boldsymbol{x}\rangle+\int_0^1\langle\nabla f(\boldsymbol{x}+\theta(\boldsymbol{y}-\boldsymbol{x}))-\nabla f(\boldsymbol{x}), \boldsymbol{y}-\boldsymbol{x}\rangle d\theta \\ f(\boldsymbol{y})-f(\boldsymbol{x})-\langle\nabla f(\boldsymbol{x}), \boldsymbol{y}-\boldsymbol{x}\rangle &= \int_0^1\langle\nabla f(\boldsymbol{x}+\theta(\boldsymbol{y}-\boldsymbol{x}))-\nabla f(\boldsymbol{x}), \boldsymbol{y}-\boldsymbol{x}\rangle d\theta \end{aligned}$$

Taking absolute value on both sides, we get

$$|f(\boldsymbol{y})-f(\boldsymbol{x})-\langle\nabla f(\boldsymbol{x}), \boldsymbol{y}-\boldsymbol{x}\rangle| = \left|\int_0^1\langle\nabla f(\boldsymbol{x}+\theta(\boldsymbol{y}-\boldsymbol{x}))-\nabla f(\boldsymbol{x}), \boldsymbol{y}-\boldsymbol{x}\rangle d\theta\right|$$

Now, since absolute value of a summation is lesser than or equal to the sum of absolute values of individual components, we have

$$|f(\boldsymbol{y})-f(\boldsymbol{x})-\langle\nabla f(\boldsymbol{x}), \boldsymbol{y}-\boldsymbol{x}\rangle| \leq \int_0^1|\langle\nabla f(\boldsymbol{x}+\theta(\boldsymbol{y}-\boldsymbol{x}))-\nabla f(\boldsymbol{x}), \boldsymbol{y}-\boldsymbol{x}\rangle| d\theta$$

Using Cauchy-Schwarz inequality, we get

$$|f(\boldsymbol{y})-f(\boldsymbol{x})-\langle\nabla f(\boldsymbol{x}), \boldsymbol{y}-\boldsymbol{x}\rangle| \leq \int_0^1\|\nabla f(\boldsymbol{x}+\theta(\boldsymbol{y}-\boldsymbol{x}))-\nabla f(\boldsymbol{x})\| \cdot\|\boldsymbol{y}-\boldsymbol{x}\| d\theta$$

Using Lipschitz gradient inequality in first component of integral, we have

$$\|\nabla f(\boldsymbol{x}+\theta(\boldsymbol{y}-\boldsymbol{x}))-\nabla f(\boldsymbol{x})\| \leq L\|\theta(\boldsymbol{y}-\boldsymbol{x})\| \leq L|\theta|\|\boldsymbol{y}-\boldsymbol{x}\| = L\theta\|\boldsymbol{y}-\boldsymbol{x}\|$$

Since the integral is in 0 to 1, we have removed the absolute sign from $\theta$ in last step. Finally,

$$|f(\boldsymbol{y})-f(\boldsymbol{x})-\langle\nabla f(\boldsymbol{x}),\boldsymbol{y}-\boldsymbol{x}\rangle| \leq \int_0^1 L\theta d\theta \cdot \|\boldsymbol{y}-\boldsymbol{x}\|^2 = \frac{L}{2}\|\boldsymbol{y}-\boldsymbol{x}\|^2$$

Thus, for $L>0$, we showed by using $\|\nabla f(\boldsymbol{x})-\nabla f(\boldsymbol{y})\| \leq L\|\boldsymbol{x}-\boldsymbol{y}\|$ that

$$|f(\boldsymbol{y})-f(\boldsymbol{x})-\langle\nabla f(\boldsymbol{x}),\boldsymbol{y}-\boldsymbol{x}\rangle| \leq \frac{L}{2}\|\boldsymbol{y}-\boldsymbol{x}\|_2^2$$

Hence, both the definitions are equivalent.

(b) Let $f:\mathbb{R}^n\to\mathbb{R}$ be such that:

- $f$ is a convex function
- $\nabla f$ is Lipschitz-continuous with Lipschitz constant $2\mu$

Show that, for all $\mathbf{x},\boldsymbol{y}\in\mathbb{R}^n$

$$\frac{1}{4\mu}\|\nabla f(\mathbf{x})-\nabla f(\boldsymbol{y})\|_2^2 \leq \left|f(\boldsymbol{y})-f(\mathbf{x})-\nabla f(\mathbf{x})^\top(\boldsymbol{y}-\mathbf{x})\right| \leq \mu\|\boldsymbol{y}-\mathbf{x}\|^2$$

What can you comment about $f$ in this case?

**Solution:** Consider LHS of the inequality first. As $f$ is convex, we have

$$f(\boldsymbol{y}) \geq f(\boldsymbol{x})+\nabla f(\boldsymbol{x})^T(\boldsymbol{y}-\boldsymbol{x})$$

Adding and subtracting $f(c)$, we get

$$\begin{aligned}
f(\boldsymbol{y})-f(\boldsymbol{x}) &= f(\boldsymbol{c})-f(\boldsymbol{x})-[f(\boldsymbol{c})-f(\boldsymbol{y})]\\
&\geq \nabla f(\boldsymbol{x})^T(\boldsymbol{c}-\boldsymbol{x})-\left[\nabla f(\boldsymbol{y})^T(\boldsymbol{c}-\boldsymbol{y})+\frac{L}{2}\|\boldsymbol{c}-\boldsymbol{y}\|_2^2\right]\\
&= \nabla f(\boldsymbol{x})^T(\boldsymbol{c}-\boldsymbol{x})-\nabla f(\boldsymbol{y})^T(\boldsymbol{c}-\boldsymbol{y})-\frac{L}{2}\|\boldsymbol{c}-\boldsymbol{y}\|_2^2\\
&= \nabla f(\boldsymbol{x})^T(\boldsymbol{y}-\boldsymbol{x}+\boldsymbol{c}-\boldsymbol{y})-\nabla f(\boldsymbol{y})^T(\boldsymbol{c}-\boldsymbol{y})-\frac{L}{2}\|\boldsymbol{c}-\boldsymbol{y}\|_2^2\\
&= \nabla f(\boldsymbol{x})^T(\boldsymbol{y}-\boldsymbol{x})+\nabla f(\boldsymbol{x})^T(\boldsymbol{c}-\boldsymbol{y})-\nabla f(\boldsymbol{y})^T(\boldsymbol{c}-\boldsymbol{y})-\frac{L}{2}\|\boldsymbol{c}-\boldsymbol{y}\|_2^2
\end{aligned}$$

Now, we have

$$f(\boldsymbol{y}) - f(\boldsymbol{x}) \geq \nabla f(\boldsymbol{x})^T(\boldsymbol{y}-\boldsymbol{x}) + \nabla f(\boldsymbol{x})^T(\boldsymbol{c}-\boldsymbol{y}) - \nabla f(\boldsymbol{y})^T(\boldsymbol{c}-\boldsymbol{y}) - \frac{L}{2}\|\boldsymbol{c}-\boldsymbol{y}\|_2^2$$

On rearranging, we get

$$(\nabla f(\boldsymbol{x}) - \nabla f(\boldsymbol{y}))^T(\boldsymbol{c}-\boldsymbol{y}) - \frac{L}{2}\|\boldsymbol{c}-\boldsymbol{y}\|^2 \leq f(\boldsymbol{y}) - f(\boldsymbol{x}) - \nabla f(\boldsymbol{x})^T(\boldsymbol{y}-\boldsymbol{x})$$

$$(\nabla f(\boldsymbol{x}) - \nabla f(\boldsymbol{y}))^T(\boldsymbol{c}-\boldsymbol{y}) - \frac{L}{2}\|\boldsymbol{c}-\boldsymbol{y}\|^2 \leq |f(\boldsymbol{y}) - f(\boldsymbol{x}) - \nabla f(\boldsymbol{x})^T(\boldsymbol{y}-\boldsymbol{x})|$$

Now, let's consider the following value of c

$$\boldsymbol{c} = \boldsymbol{y} - \frac{1}{L}(\nabla f(\boldsymbol{y}) - \nabla f(\boldsymbol{x}))$$
$$\boldsymbol{c} - \boldsymbol{y} = \frac{1}{L}(\nabla f(\boldsymbol{x}) - \nabla f(\boldsymbol{y}))$$

Multiplying both sides by $(\nabla f(\boldsymbol{x}) - \nabla f(\boldsymbol{y}))^T$, we get

$$(\nabla f(\boldsymbol{x}) - \nabla f(\boldsymbol{y}))^T(\boldsymbol{c}-\boldsymbol{y}) = \frac{1}{L}\|\nabla f(\boldsymbol{x}) - \nabla f(\boldsymbol{y})\|_2^2$$

We also have

$$\frac{L}{2}\|\boldsymbol{c}-\boldsymbol{y}\|_2^2 = \frac{1}{2L}\|\nabla f(\boldsymbol{x}) - \nabla f(\boldsymbol{y})\|_2^2$$

On subtracting we get,

$$(\nabla f(\boldsymbol{x}) - \nabla f(\boldsymbol{y}))^T(\boldsymbol{c}-\boldsymbol{y}) - \frac{L}{2}\|\boldsymbol{c}-\boldsymbol{y}\|_2^2 = \frac{1}{2L}\|\nabla f(\boldsymbol{x}) - \nabla f(\boldsymbol{y}\|_2^2$$

Using this in our rearranged inequality, we get

$$\frac{1}{2L}\|\nabla f(\boldsymbol{x}) - \nabla f(\boldsymbol{y})\|_2^2 \leq |f(\boldsymbol{y}) - f(\boldsymbol{x}) - \nabla f(\boldsymbol{x})^\top(\boldsymbol{y}-\boldsymbol{x})|$$

It is given that $L = 2\mu$. On substituting, we get

$$\frac{1}{4\mu}\|\nabla f(\boldsymbol{x}) - \nabla f(\boldsymbol{y})\|_2^2 \leq |f(\boldsymbol{y}) - f(\boldsymbol{x}) - \nabla f(\boldsymbol{x})^\top(\boldsymbol{y}-\boldsymbol{x})|$$

We have proved LHS of the inequality. Now, consider RHS of the inequality. From part (a), we have

$$|f(\boldsymbol{y}) - f(\boldsymbol{x}) - \langle \nabla f(\boldsymbol{x}), \boldsymbol{y}-\boldsymbol{x}\rangle| \leq \frac{L}{2}\|\boldsymbol{x}-\boldsymbol{y}\|_2^2$$

As $L = 2\mu$,

$$\left|f(\boldsymbol{y}) - f(\boldsymbol{x}) - \nabla f(\boldsymbol{x})^\top (\boldsymbol{y} - \boldsymbol{x})\right| \leq \mu \|\boldsymbol{y} - \mathbf{x}\|_2^2$$

Hence, we have proved RHS of the inequality too.

We can say that $f$ is $L$-smooth function. The interpretation is that the error in first order Taylor approximation is bounded.

Q 8. Implement numerically correct versions of the following functions:

1. Logistic Loss: $L(w) = \sum_{i=1}^{n} \log\left(1 + \exp\left(-y_i w^T x_i\right)\right)$
2. Hinge Loss/SVMs: $L(w) = \sum_{i=1}^{n} \max\left\{0, 1 - y_i w^T x_i\right\}$. Here $y_i \in \{-1, +1\}$.
3. Least Squares Loss: $L(w) = \sum_{i=1}^{n} \left(y_i - w^T x_i\right)^2$. Here $y_i \in R$.

Note: Write your codes in the given notebook: Assignment1.ipynb with your implementations of 1), 2), 3), respectively. Do not modify the arguments.

1. Implement the following loss functions using simple loop code in Python.

**Solution:** Implementations of given loss functions using simple loop code are given below:

```
def LogisticLossNaive(w, X, y, lam):
    # Computes the cost function for all the training samples
    # where f is the function value and g is the gradient
    f = 0.0
    m = X.shape[1]
    g = np.zeros(m)
    for i in range(X.shape[0]):
      ycap = 0.0
      for j in range(X.shape[1]):
        ycap += (X[i][j]*w[j][0])
      f += np.log(1+np.exp(-(y[i]*ycap)))
      gtemp = ((-y[i])/(1+np.exp(y[i]*ycap)))*X[i]
      g = np.add(g, gtemp)
    return [f, g]
```

```
def HingeLossNaive(w, X, y, lam):
    # Computes the cost function for all the training samples
    # where f is the function value and g is the gradient
    f = 0.0
    m = X.shape[1]
    g = np.zeros(m)
    for i in range(X.shape[0]):
      ycap = 0.0
      for j in range(X.shape[1]):
        ycap += (X[i][j]*w[j][0])
      f += max(0,(1-(y[i]*ycap)))
      if y[i]*ycap >= 1:
        gtemp = np.zeros(m)
      else:
        gtemp = -y[i]*X[i]
      g = np.add(g, gtemp)
    return [f, g]
```

```
def LeastSquaresNaive(w, X, y, lam):
    # Computes the cost function for all the training samples
    # where f is the function value and g is the gradient
    f = 0.0
    m = X.shape[1]
    g = np.zeros(m)
    for i in range(X.shape[0]):
      ycap = 0.0
      for j in range(X.shape[1]):
        ycap += (X[i][j]*w[j][0])
      f += ((y[i]-ycap)**2)
      gtemp = -2*(y[i]-ycap)*X[i]
      g = np.add(g, gtemp)
    return [f, g]
```

2. Implement these functions using vectorized code and compare the result with the previous simple loop code. Also, implement these functions in CVXPY.

**Solution:** Implementations of given loss functions using vectorized code are given below:

```
def LogisticLossVec(w, X, y, lam):
    # Computes the cost function for all the training samples
    # where f is the function value and g is the gradient
    n = X.shape[0]
    f = np.sum(np.log(1+np.exp(-np.multiply(y,((X@w).reshape(n))))))
    g = X.T@((-y)/(1+np.exp(np.multiply(y,((X@w).reshape(n))))))
    return [f, g]
```

```
def HingeLossVec(w, X, y, lam):
    # Computes the cost function for all the training samples
    # where f is the function value and g is the gradient
    n = X.shape[0]
    f = np.sum(np.maximum(0,(1-np.multiply(y,((X@w).reshape(n))))))
    g = -X.T@(np.where(np.multiply(y,((X@w).reshape(n))) < 1, y, 0))
    return [f, g]
```

```
def LeastSquaresVec(w, X, y, lam):
    # Computes the cost function for all the training samples
    # where f is the function value and g is the gradient
    n = X.shape[0]
    f = np.sum(np.square(y-((X@w).reshape(n))))
    g = X.T@(-2*(y-((X@w).reshape(n))))
    return [f, g]
```

Result for Logistic Loss using simple loop code:

```
Time Taken = 0.005006074905395508
Function value = 170.91989496959212
Printing Gradient:
[23.54348171 23.14037054 26.08964159 25.67977349 20.06928227 27.38862014
 23.19305525 22.13292476 23.29513674 22.96141006]
```

Result for Logistic Loss using vectorized code:

```
Time Taken = 0.0004165172576904297
Function value = 170.91989496959226
Printing Gradient:
[23.54348171 23.14037054 26.08964159 25.67977349 20.06928227 27.38862014
 23.19305525 22.13292476 23.29513674 22.96141006]
```

Result for Hinge Loss using simple loop code:

```
Time Taken = 0.002773284912109375
Function value = 216.54613728086406
Printing Gradient:
[25.36832949 24.99684269 28.20362255 27.71198876 21.55730833 29.48552382
 25.05132393 24.13707411 25.24311254 25.02848932]
```

Result for Hinge Loss using vectorized code:

```
Time Taken = 0.0014541149139404297
Function value = 216.546137280864
Printing Gradient:
[25.36832949 24.99684269 28.20362255 27.71198876 21.55730833 29.48552382
 25.05132393 24.13707411 25.24311254 25.02848932]
```

Result for Least Squares Loss using simple loop code:

```
Time Taken = 0.004108190536499023
Function value = 1214.4187731838874
Printing Gradient:
[333.12358545 326.96683543 371.96712451 353.64384893 300.38749149
 365.40524058 342.06229321 320.16018962 344.85055403 337.93004332]
```

Result for Least Squares Loss using vectorized code:

```
Time Taken = 0.0016448497772216797
Function value = 1214.418773183888
Printing Gradient:
[333.12358545 326.96683543 371.96712451 353.64384893 300.38749149
 365.40524058 342.06229321 320.16018962 344.85055403 337.93004332]
```

Clearly, vectorized code takes significantly less time compared to simple loop code for all 3 loss functions.

Implementations of given loss functions using CVXPY are given below:

```python
def LogisticLossCVXPY(w, X, y, lam):
    # Computes the cost function for all the training samples
    # where f is the function value and g is the gradient
    n = X.shape[0]
    f = cp.sum(cp.logistic(-cp.multiply(y,((X@w).reshape(n)))))
    g = X.T@((-y)/(cp.logistic(-cp.multiply(y,((X@w).reshape(n))))))
    return [f, g]
```

```python
def HingeLossCVXPY(w, X, y, lam):
    # Computes the cost function for all the training samples
    # where f is the function value and g is the gradient
    n = X.shape[0]
    f = cp.sum(cp.maximum(0,(1-cp.multiply(y,((X@w).reshape(n))))))
    g = -X.T@(np.where(cp.multiply(y,((X@w).reshape(n))) <= 1, y, 0))
    return [f, g]
```

```python
def LeastSquaresCVXPY(w, X, y, lam):
    # Computes the cost function for all the training samples
    # where f is the function value and g is the gradient
    n = X.shape[0]
    m = X.shape[1]
    f = cp.sum_squares(y-((X@w).reshape(n)))
    g = X.T@(-2*(y-((X@w).reshape(n))))
    return [f, g]
```