Machine Learning - II
Assignment 2

V S Siva Kumar Lakkoju (CS2139)

Problem 1

Given f a real valued differentiable convex function on the convex set C which is L-lipschitz. Firstly note that,

$$f(y) - f(x) = \int_0^1 \langle \nabla f(x + t(y-x)), (y-x) \rangle \, dt$$

this is a direct consequence of the fundamental theorem of calculus applied component-wise. From this we can readily see that,

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

from this we conclude,

$$\begin{aligned} \left| f(y) - f(x) - \langle \nabla f(x), (y-x) \rangle \right| &\leq \left| \int_0^1 \langle \nabla f(x + t(y-x)) - \nabla f(x), (y-x) \rangle \, dt \right| \\ &\leq \int_0^1 \left| \langle \nabla f(x + t(y-x)) - \nabla f(x), (y-x) \rangle \right| \, dt \\ &\leq \int_0^1 \left\| \nabla f(x + t(y-x)) - \nabla f(x) \right\| \cdot \|y - x\| \, dt \\ &\leq L \int_0^1 t \|y - x\|^2 \, dt \\ &= \frac{L}{2} \|y - x\|^2 \end{aligned}$$

where we have used integral inequality, cauhy-schwarz and the fact that f is L-lipschitz. Now for the second half, we define,

$$\phi(y) = f(y) - \langle \nabla f(x), y \rangle$$

then we obtain the following,

$$\nabla \phi(y) = \nabla f(y) - \nabla f(x)$$

noting that ϕ a real valued function, we note that, ϕ is minimized at $y = x$, from which it follows that,

$$\phi(x) \leq \phi(y - \frac{1}{L} \nabla \phi(y))$$

Further note that,

$$\begin{aligned} \left\| \nabla \phi(b) - \nabla \phi(a) \right\| &= \left\| \nabla f(b) - \nabla f(x) - \nabla f(a) + \nabla f(x) \right\| \\ &= \left\| \nabla f(b) - \nabla f(a) \right\| \\ &\leq L \|b - a\| \end{aligned}$$

So ϕ is L-lipschitz since f is L-lipschitz. It then follows that,

$$\begin{aligned} \phi(y - \frac{1}{L} \nabla \phi(y)) &\leq \phi(y) + \langle \nabla \phi(y), -\frac{1}{L} \phi(y) \rangle + \frac{1}{2} \left\| -\frac{1}{L} \nabla \phi(y) \right\|^2 \\ &= \phi(y) - \frac{1}{L} \left\| \nabla \phi(y) \right\|^2 + \frac{1}{2L} \left\| \nabla \phi(y) \right\|^2 \end{aligned}$$

Substituting the previous inequality ($\phi(x) \leq \phi(y - \frac{1}{L}\nabla\phi(y))$) implies,

$$\begin{aligned}\phi(x) &\leq \phi(y) - \frac{1}{L}\left\|\nabla\phi(y)\right\|^2 + \frac{1}{2L}\left\|\nabla\phi(y)\right\|^2 \\ &= \phi(y) - \frac{1}{2L}\left\|\nabla\phi(y)\right\|^2\end{aligned}$$

from this we obtain $\frac{1}{2L}\left\|\nabla\phi(y)\right\|^2 \leq \phi(y) - \phi(x)$ Finally this gives us the desired inequality,

$$\frac{1}{2L}\left\|\nabla f(y) - \nabla f(x)\right\|^2 \leq f(y) - f(x) - \langle\nabla f(x), (y - x)\rangle$$

Problem 2,3

We have,

$$x_{k+1} = x_k - \frac{1}{L}\nabla f(x_k)$$

We will first show that this minimizing sequence $\{x_k\}_0^\infty$ converges to a minimizer. Since f is an L-lipschitz, convex differentiable function we have,

$$f(x_{k+1}) \leq f(x_k) + \langle\nabla f(x_k), x_{k+1} - x_k\rangle + \frac{1}{2}\|x_{k+1} - x_k\|^2$$

This implies,

$$\begin{aligned}f(x_{k+1}) - f(x_k) &\leq \langle\nabla f(x_k), -\frac{1}{L}\nabla f(x_k)\rangle + \frac{1}{2}\|x_{k+1} - x_k\|^2 \\ &\leq -\frac{1}{L}\left\|\nabla f(x_k)\right\|^2 + \frac{1}{2}\|x_{k+1} - x_k\|^2 \\ &\leq \frac{1}{2}\|x_{k+1} - x_k\|^2\end{aligned}$$

Firstly this proves that $\{f(x_k)\}_0^\infty$ is monotonically decreasing, convergence then follows from convergence of $\{x_k\}_0^\infty$, and then since its decreasing it follows that we obtain the minima. Now we obtain the convergence rate bounds, let x^* be the minimizer, now since f is convex, we have,

$$\begin{aligned}f(x_k) &\leq f(x^*) + \langle x_{k-x^*}, \nabla f(x_k)\rangle \\ f(x_k) - f(x^*) &\leq \|x_k - x^*\| \cdot \left\|\nabla f(x_k)\right\| \\ &\leq L\|x_k - x^*\| \cdot \|x_{k+1} - x_k\|\end{aligned}$$

where at the end we used the fact that f, is L-lipschitz. Let $a_k = f(x_k) - f(x^*)$,

$$\begin{aligned}f(x_k) - f(x_{k+1}) &= f(x_k) - f(x^*) - f(x_{k+1}) + f(x^*) \\ &= a_k - a_{k+1} \\ &\geq \frac{L}{2}\|x_k - x_{k+1}\|^2 \\ &\geq \frac{L}{2}\left(\frac{a_k}{L\|x_k - x_{k+1}\|}\right)^2 \\ &\geq \frac{a_{k+1}^2}{L\|x_k - x_{k+1}\|^2}\end{aligned}$$

where we used the fact that $a_{k+1} \leq a_k \leq \frac{L}{2}\|x_k - x_{k+1}\|^2$. Further we can rewrite the inequality obtained as follows,

$$\begin{aligned}a_k &\geq a_{k+1} + \frac{a_{k+1}^2}{L\|x_k - x_{k+1}\|^2} \\ &= a_{k+1}\left(\frac{a_{k+1}}{L\|x_k - x_{k+1}\|^2}\right)\end{aligned}$$

This implies,

$$a_k^{-1} \leq a_{k+1}^{-1} \left(1 + \frac{a_{k+1}}{2L\|x_k - x_{k+1}\|^2} \right)^{-1}$$

Now $a_{k+1} \leq a_k \leq \frac{L}{2}\|x_k - x_{k+1}\|^2$ implies

$$\frac{a_{k+1}}{2L\|x_k - x^*\|^2} \leq \frac{1}{4}$$

Let $t = \frac{a_{k+1}}{2L\|x_k - x^*\|^2}$, then since

$$(1+t)^{-1} \leq (1 - \frac{4t}{5}) \text{ for } 0 \leq t \leq \frac{1}{4}$$

so we get,

$$\begin{aligned} a_k^{-1} &\leq a_{k+1}^{-1}(1+t)^{-1} \\ &\leq a_{k+1}^{-1}(1 - \frac{4t}{5}) \end{aligned}$$

so we finally get, $0 \leq a_k^{-1} - a_{k+1}^{-1} \leq \frac{4}{5} a_{k+1}^{-1} t$

$$0 \leq a_k^{-1} - a_{k+1}^{-1} \leq \frac{2}{5L\|x_k - x_{k+1}\|^2}$$

adding these up, for $k = 0$ to m, gives us the desired inequality,

$$f(x_k) - f(x^*) \leq \frac{5L}{2\sum_{i=0}^{k-1} \|x_i - x^*\|^{-2}}$$

Problem 6

Given a differentiable strictly increasing and convex function such that ∇f is L -lipschitz, $L(x) = \frac{1}{2}\|x - x^*\|^2$. Let, $x_{k+1} = x_k - \frac{1}{2}\nabla f(x)$. So we obtain,

$$\begin{aligned} L(x_{k+1}) &= \frac{1}{2}\|x_{k+1} - x^*\| \\ &= \frac{1}{2}\left\| x_k - \frac{1}{2}\nabla f(x) - x^* \right\|^2 \\ &= \frac{1}{2}\left(\|x_k - x^*\|^2 + \frac{\|\nabla f(x_k)\|^2}{L^2} - \frac{2}{L}\langle x_k - x^*, \nabla f(x_k) - \nabla f(x^*)\rangle \right) \\ &\leq \frac{1}{2}\left(\|x_k - x^*\|^2 + \frac{\|\nabla f(x_k)\|^2}{L^2} - \frac{2}{L^2}\|f(x_k)^2\| \right) \\ &= \frac{1}{2}\left(\|x_k - x^*\|^2 - \frac{\|\nabla f(x)\|^2}{L^2} \right) \\ &\leq \frac{1}{2}\|x_k - x^*\|^2 = L(x_k) \end{aligned}$$

Problem 7

Given $P_1 \sim N(Q_1, \Sigma)$ and $P_2 \sim N(Q_2, \Sigma)$,
so, $D_{kl}(P_1||P_2) = E_{P_1}(\log \frac{P_1}{P_2})$, so for

$$P_1 = \frac{1}{\sqrt{(2\pi)^n|\Sigma|}} exp(-\frac{1}{2}(x - Q_1)^T\Sigma^{-1}(x - Q_1))$$

$$P_2 = \frac{1}{\sqrt{(2\pi)^n|\Sigma|}} exp(-\frac{1}{2}(x - Q_2)^T\Sigma^{-1}(x - Q_2))$$

so we have,

$$\log(\frac{P_1}{P_2}) = \log(P_1) - \log(P_2) = -\frac{1}{2}(x-Q_1)^T\Sigma^{-1}(x-Q_1) + \frac{1}{2}(x-Q_2)^T\Sigma^{-1}(x-Q_2)$$

Now we have,

$$\begin{aligned}D_{KL}(P_1||P_2) &= E_{p_1}[\frac{1}{2}(x-Q_2)^T\Sigma^{-1}(x-Q_2) - \frac{1}{2}(x-Q_1)^T\Sigma^{-1}(x-Q_1)] \\ &= \frac{1}{2}E_{p_1}\left[\frac{1}{2}(x-Q_2)^T\Sigma^{-1}(x-Q_2)\right] - \frac{1}{2}E_{p_1}\left[(x-Q_1)^T\Sigma^{-1}(x-Q_1)\right]\end{aligned}$$

Further we have,

$$\begin{aligned}E_{p_1}\left[(x-Q_1)^T\Sigma^{-1}(x-Q_1)\right] &= E_{p_1}\left[tr\left((x-Q_1)^T\Sigma^{-1}(x-Q_1)\right)\right] \\ &= tr\left(E_{p_1}\left[\Sigma^{-1}\Sigma\right]\right) \\ &= tr\left(I\right) = n\end{aligned}$$

Now for the other half,

$$\begin{aligned}E_{p_1}\left[(x-Q_2^T\Sigma^{-1}(x-Q_2))\right] &= E_{p_1}\left[(x-Q_1+Q_1-Q_2)^T\Sigma^{-1}(x-Q_1+Q_1+Q_2)\right] \\ &= E_{p_1}\left[(x-Q_1)^T\Sigma^{-1}(x-Q_1)\right] + E_{p_1}\left[(Q_1-Q_2)^T\Sigma^{-1}(Q_1-Q_2)\right] \\ &= n + (Q_1-Q_2)^T\Sigma^{-1}(Q_1-Q_2)\end{aligned}$$

From this it follows,

$$D_{KL}\left[P_1||P_2\right] = (Q_1-Q_2)^T\Sigma^{-1}(Q_1-Q_2)$$

Problem 8

a)

We have $D_{KL}(P^n||P_{\theta,\Sigma})$,

$$\log\left(\frac{P^n}{P_{\theta,\Sigma}}\right) = \log\left(\frac{\lambda^n 2^{\frac{n}{2}}\pi^{\frac{n}{2}}|\Sigma|^{\frac{1}{2}}}{2^n}\right) + \frac{1}{2}(x-\theta)^T\Sigma^{-1}(x-\theta) - \sum_{i=1}^{n}\lambda|x_i|$$

We have,

$$\begin{aligned}D_{KL}(P^n||P_{\theta,\Sigma}) &= E_{P^n}\left[\log\left(\frac{P^n}{P_{\theta,\Sigma}}\right)\right] \\ &= |\Sigma|^{\frac{1}{2}}\lambda^n(\frac{\pi}{2})^n + \frac{n\lambda^2}{2} + \frac{1}{2}\theta^T\Sigma^{-1}\theta\end{aligned}$$

b)

We can readily see from the positive definiteness that, $D_{KL}(P^n||P_{\theta,\Sigma})$ is minimised when $\theta = 0$.

Problem 9

Let us define,

$$S_a = \sum_{i=1}^{n} a_i,\ S_b = \sum_{i=1}^{n} b_i$$

then using Jensen's inequality we have,

$$\begin{aligned}
S_a f\left(\frac{S_b}{S_a}\right) &= S_a f\left(\frac{\sum_{i=1}^{n} b_i}{S_a}\right) \\
&= S_a f\left(\sum_{i=1}^{n} \frac{a_i}{S_a}\frac{b_i}{a_i}\right) \\
&\leq S_a \sum_{i=0}^{n} \frac{a_i}{S_a} f(\frac{b_i}{a_i}) \\
&\leq \sum_{i=0}^{n} a_i f(\frac{b_i}{a_i})
\end{aligned}$$