CIS5200: Machine Learning	Spring 2025

HW2 Solutions

1 Written Questions

Q1

1. We define $g(x) = \log(1 + e^{-x})$:

$$g'(x) = \frac{-e^{-x}}{1 + e^{-x}}$$
$$g''(x) = \frac{e^{-x}}{(1 + e^{-x})^2}$$

   We notice that $g''(x)$ is clearly < 1 and > 0 and taking the limits to ∞ and $-\infty$, $g''(x) \to 0$. Now, we take $h(w) = \log\left(1 + e^{-y_i x_i^T w}\right)$. To show that this is strongly convex we have to show that there is some positive μ s.t.

$$\nabla^2 h(w) \succeq \mu I$$

   for all w.

$$\nabla^2 h_i(w) = g''(y_i x_i^T w) x_i x_i^T$$

   So we can immediately derive that for w consisting of elements that go to ∞, the $g''(y_i x_i^T w)$ goes to 0, so it is not strongly convex.

2. As far as the 1-smoothness, we need to show that:

$$\nabla^2 h_i(w) \preceq I$$

   Since, we have seen that $g''(x) < 1$ for all x, we only to show that $x_i x_i^T \preceq I$. To show this, it is enough to show that

$$v^T x_i x_i^T v \leq v^T v$$

   for all v.But,

$$v^T x_i x_i^T v = (x_i^T v)^2 \leq \|x_i\|^2 \|v\|^2 \leq \|v\|^2$$

   since $\|x_i\| \leq 1$.After the breakdown of each $h_i(w)$, we need to add them all together and compute the hessian of their sum (which is the sum of their Hessians) and since we divide by m we conclude that $H(w)$ (the Hessian of the objective) is :

$$I \succeq H(w) \succeq 0$$

3. For a L-smooth function, suppose the learning rate is η, then by using the definition of smoothness, with $w' = w - \eta \nabla F(w)$, we have

$$
\begin{aligned}
F(w') &\leq F(w) + \nabla F(w)^\top (w' - w) + \frac{L}{2}\|w' - w\|_2^2 \\
&= F(w) + \frac{1}{\eta}(w - w')^\top (w' - w) + \frac{L}{2}\|w' - w\|_2^2 \\
&= F(w) - \frac{1}{\eta}\|w' - w\|_2^2 + \frac{L}{2}\|w' - w\|_2^2 \\
&= F(w) + \left(\frac{L}{2} - \frac{1}{\eta}\right)\|w' - w\|_2^2.
\end{aligned}
$$

So as long as $L/2 \leq 1/\eta$, the function will be non-decreasing. This gives us that $\eta \leq 2/L = 2$ to always ensure non-increasing behavior of iterates.

4. Since this is the convex smooth setting, the convergence guarantee we have is:

$$
F(w_{T+1}) - F(w_*) \leq \frac{L\|w_1 - w_*\|_2^2}{2T}.
$$

Substituting $L = 1$, we get,

$$
\frac{\|w_1 - w_*\|_2^2}{2T} \leq \epsilon \implies T \geq \frac{\|w_1 - w_*\|_2^2}{2\epsilon}.
$$

5. We have $\nabla \mathcal{R}(w) = \begin{bmatrix} 2\lambda_1 w_1 \\ 2\lambda_2 w_2 \\ \vdots \\ 2\lambda_d w_d \end{bmatrix}$, so therefore, $\nabla^2 \mathcal{R}(w) = \begin{bmatrix} 2\lambda_1 & 0 & \dots & 0 \\ 0 & 2\lambda_2 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & 2\lambda_d \end{bmatrix}$. Thus in order for the hessian of $\mathcal{R}(w)$ to be μ-strongly convex, we need $2\lambda_j \geq \mu$ for all j (since the diagonal entries are the eigenvalues), or in other words, $2\min_{j\in[d]} \lambda_j \geq \mu$. Therefore, $\mathcal{R}(w)$ is $2\min_{j\in[d]} \lambda_j$-strongly convex. Now since the objective function without $\mathcal{R}(w)$ is not strongly convex, this implies that the objective function with $\mathcal{R}(w)$ is strongly convex with $0 + 2\min_{j\in[d]} \lambda_j = \boxed{2\min_{j\in[d]} \lambda_j = \mu}$ (since they are additive).

   By similar reasoning, we see that $\mathcal{R}(w)$ is $2\max_{j\in[d]} \lambda_j$-smooth since we need $2\lambda_j \leq L$ for all j. Since the objective function without $\mathcal{R}(w)$ is 1-smooth, we see that these added together is L-smooth where $\boxed{L = 1 + 2\max_{j\in[d]} \lambda_j}$ (since they are additive).

6. As long as $2\min_{j\in[d]} \lambda_j$ is lower bounded by a constant, this is the strongly convex setting, and we have the following guarantee:

$$
\|w_{T+1} - w_*\|_2^2 \leq \left(1 - \frac{\mu}{L}\right)^T \|w_1 - w_*\|_2^2
$$

with $\mu = 2\min_{j\in[d]} \lambda_j$ and $L = 1 + 2\max_{j\in[d]} \lambda_j$. Using the fact that $1 - x \leq \exp(-x)$, we have

$$
\left(1 - \frac{\mu}{L}\right)^T \leq \exp\left(-\frac{\mu T}{L}\right)
$$

which gives us

$$
\exp\left(-\frac{\mu T}{L}\right)\|w_1 - w_*\|_2^2 \leq \epsilon \implies T \geq \frac{L}{\mu}\log\left(\frac{\|w_1 - w_*\|_2^2}{\epsilon}\right) = \frac{(1 + 2\max_{j\in[d]} \lambda_j)}{2\min_{j\in[d]} \lambda_j}\log\left(\frac{\|w_1 - w_*\|_2^2}{\epsilon}\right)
$$

7. Recal from 1.6 that

$$T \geq \frac{L}{\mu} \log\left(\frac{\|w - w_\star\|^2}{\epsilon}\right)$$

Assuming that all eigenvalues are the same:

$$\frac{L}{\mu} = \frac{1 + 2\lambda}{2\lambda}$$

Which is monotonically decreasing for $\lambda > 0$ (you can verify this by taking the derivative). That means for fixed > 0, then:

$$\frac{1 + 2\lambda}{2\lambda} \geq \frac{1 + 2(\lambda + \epsilon)}{\lambda + \epsilon}$$

Which gives:

$$\frac{1 + 2\lambda}{2\lambda} \log\left(\frac{\|w - w_\star\|^2}{\epsilon}\right) \geq \frac{1 + 2(\lambda + \epsilon)}{\lambda + \epsilon} \log\left(\frac{\|w - w_\star\|^2}{\epsilon}\right)$$

Suggesting that as the regularization parameter λ increases, the bound on T loosens, and hence, the number of iterations required to reach ϵ difference in solution. In other words, we converge faster.

However, choosing the largest λ is not always desirable, recall the objective function:

$$\min \hat{R}(w) + \sum_{j=1}^{d} \lambda_j w_j^2$$

For large λ, the gradient descent can "cheat" and minimize the objective by choosing $w_j \approx 0$. You can confirm this by working out the GD update rule and observing the λw term which will be large.

Hence, w fails to actually fit the data at all for big λ

Q2

1. We have $y = w^T x + \epsilon$, where $\epsilon \sim \mathcal{N}(0, \sigma^2)$. This means that $y \sim \mathcal{N}(w^T x, \sigma^2)$ (the mean of y is $w^T x$) by the additive property of Gaussian random variables. Thus by the definition of a normal distribution's density function, $P(y|x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{1}{2}\left(\frac{y - w^T x}{\sigma}\right)^2\right) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(y - w^T x)^2}{2\sigma^2}\right)$.

2. We see that

$$
\begin{aligned}
R(f) &= \mathbb{E}_{x,y}[(y - f(x))^2] \\
&= \mathbb{E}_{x,y}[(y - \mathbb{E}[y|x])^2] \\
&= \mathbb{E}_x[\ \mathbb{E}_y[(y - \mathbb{E}[y \mid x])^2 \mid x]\ \] && \text{by Adam's Law} \\
&= \mathbb{E}_x[\ \mathrm{Var}(y \mid x)\ \] && \text{by definition of Conditional Variance} \\
&= \mathbb{E}_x[\sigma^2] \\
&= \sigma^2
\end{aligned}
$$

3. We have

$$\hat{L}(w, \sigma) = P(y_1, \ldots, y_m | x_1, \ldots, x_m) = \prod_{i=1}^{m} P(y_i | x_i)$$

Thus the log conditional likelihood is

$$
\begin{aligned}
\log \hat{L}(w, \sigma) = \log(P(y_1, \ldots, y_m | x_1, \ldots, x_m)) &= \log\left(\prod_{i=1}^{m} P(y_i|x_i)\right) \\
&= \sum_{i=1}^{m} \log P(y_i|x_i) \\
&= \sum_{i=1}^{m} \log\left(\frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(y_i - w^T x_i)^2}{2\sigma^2}\right)\right) \\
&= \sum_{i=1}^{m} -\log\left(\sqrt{2\pi}\sigma\right) - \frac{(y_i - w^T x_i)^2}{2\sigma^2} \\
&= \boxed{-m \log\left(\sqrt{2\pi}\sigma\right) - \frac{1}{2\sigma^2} \sum_{i=1}^{m} (y_i - w^T x_i)^2}
\end{aligned}
$$

4. We want to find the value of w that maximizes $\log \hat{L}(w, \sigma)$, or in other words, $\text{argmax}_w(\log \hat{L}(w, \sigma))$.

$$
\begin{aligned}
\text{argmax}_w(\log \hat{L}(w, \sigma)) &= \text{argmax}_w \left(-m \log\left(\sqrt{2\pi}\sigma\right) - \frac{1}{2\sigma^2} \sum_{i=1}^{m} (y_i - w^T x_i)^2 \right) \\
&= \text{argmin}_w \left(m \log\left(\sqrt{2\pi}\sigma\right) + \frac{1}{2\sigma^2} \sum_{i=1}^{m} (y_i - w^T x_i)^2 \right) \\
&= \text{argmin}_w \left(\frac{1}{2\sigma^2} \sum_{i=1}^{m} (y_i - w^T x_i)^2 \right) \\
&= \text{argmin}_w \left(\sum_{i=1}^{m} (y_i - w^T x_i)^2 \right) \\
&= \text{argmin}_w \left(\frac{1}{m} \sum_{i=1}^{m} (y_i - w^T x_i)^2 \right) \\
&= \text{argmin}_w \ \hat{R}(w)
\end{aligned}
$$