STATS 214 Autumn 2021 Homework 3

SUNet ID: 06009508

Name: Brandon McKinzie

Collaborators: N/A

By turning in this assignment, I agree by the Stanford honor code and declare that all of this is my own work.

N.B. Throughout my work for problem 1, I will make use of the following:

$$\nabla \hat{L}(w) = \frac{1}{n}\sum_{i=1}^{n}\left(\sigma(w^\top x^{(i)}) - y^{(i)}\right)\sigma'(w^\top x^{(i)})x^{(i)} \tag{1}$$

$$\nabla^2 \hat{L}(w) = \frac{1}{n}\sum_{i=1}^{n}\left(\sigma''(w^\top x^{(i)})\left(\sigma(w^\top x^{(i)}) - y^{(i)}\right) + \sigma'(w^\top x^{(i)})^2\right)x^{(i)}x^{(i)\top} \tag{2}$$

$$\nabla L(w) = \mathbb{E}_{x,y}\left[\left(\sigma(w^\top x) - y\right)\sigma'(w^\top x)x\right] \tag{3}$$

$$\nabla^2 L(w) = \mathbb{E}_{x,y}\left[\left(\sigma''(w^\top x)\left(\sigma(w^\top x) - y\right) + \sigma'(w^\top x)^2\right)xx^\top\right] \tag{4}$$

Problem 1(a) Non-Convexity of the Expected Loss

Construct an example where the expected loss $L(w)$ is not a convex function of w, where

$$L(w) = \mathbb{E}_{x,y}\left[\frac{1}{2}\left(y - \sigma(w^T x)\right)^2\right] \tag{5}$$

Let $d = 1$, and let σ be the sigmoid function:

$$\sigma(x) = \frac{1}{1+e^{-x}} \tag{6}$$

$$\sigma'(x) = \sigma(x)(1-\sigma(x)) \tag{7}$$

$$\sigma''(x) = \sigma(x)(1-\sigma(x))(1-2\sigma(x)) \tag{8}$$

Note that this satisfies the constraints provided by the problem for σ. We need to find some combination of $\{P, \epsilon, w_*\}$ such that the second derivative is negative for at least some w, where

$$\frac{\partial}{\partial w}L(w) = \mathbb{E}_{x,y}\left[\left(\sigma(wx) - y\right)\sigma'(wx)x\right] \tag{9}$$

$$\frac{\partial^2}{\partial w^2}L(w) = \mathbb{E}_{x,y}\left[\left(\sigma''(wx)\left(\sigma(wx) - y\right) + \sigma'(wx)^2\right)x^2\right] \tag{10}$$

Assume P is degenerate about some point x_0, i.e. $P(x = x_0) = 1$. Furthermore, assume $\epsilon_i = 0$ $\forall i$, i.e. assume $\epsilon \sim \delta$, where δ denotes the dirac delta distribution. In this case, we just need to find some values for $\{x_0, w_*, w\}$ for which

$$\left(\sigma''(wx_0)\left(\sigma(wx_0) - \sigma(w_*x_0)\right) + \sigma'(wx_0)^2\right)x_0^2 < 0 \tag{11}$$

$$\implies \sigma''(wx_0)\left(\sigma(wx_0) - \sigma(w_*x_0)\right) + \sigma'(wx_0)^2 < 0 \tag{12}$$

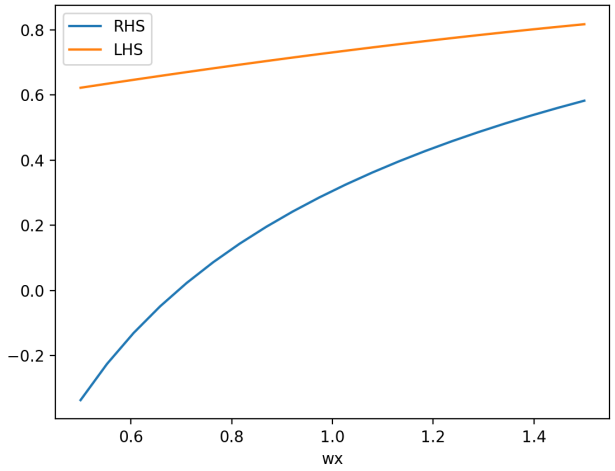
Assume that $\sigma''(wx_0) \neq 0$, then we need

$$\sigma(w_* x_0) > \sigma(wx_0) + \frac{\sigma'(wx_0)^2}{\sigma''(wx_0)} \tag{13}$$

Recall that the provided constraints ensure that $-BR \leq wx_0 \leq BR$, where $BR \leq 1$. In fact, consider the case where $w_* = R$ and $x_0 = B$, i.e. $w_* x_0 = BR$. Since σ is strictly increasing, this guarantees that $\sigma(w_* x_0) \geq \sigma(wx_0)\ \forall w$. Consider the case where $wx_0 = 1$. We can easily verify that

$$\sigma(BR \geq 1) > \sigma(1) + \frac{\sigma'(1)^2}{\sigma''(1)} \tag{14}$$

by plotting the LHS and RHS separately as a function of their input wx:



Therefore, we have shown that for $w_* = R$, $x_0 = B$, $w = \frac{1}{B}$, $\epsilon \sim \delta(x)$, $d = 1$, that the second derivative of $L(w)$ is negative. Thus, $L(w)$ is not a convex function.

Problem 1(b) All Stationary Points of the Expected Loss are Global Minima

Show that w_* is a global minimum of L.

Note that

$$L(w_*) = \mathbb{E}_{x,y}\left[\frac{1}{2}\left(\sigma(w_*^\top x) + \epsilon - \sigma(w_*^\top x)\right)^2\right] \tag{15}$$

$$= \mathbb{E}_\epsilon\left[\frac{1}{2}\epsilon^2\right] \tag{16}$$

Next, for all $w \neq w_*$,

$$L(w) = \mathbb{E}_{x,y}\left[\frac{1}{2}\left(\sigma(w_*^\top x) + \epsilon - \sigma(w^\top x)\right)^2\right] \tag{17}$$

$$= \frac{1}{2}\mathbb{E}_{x,y}\left[\left(\sigma(w_*^\top x) - \sigma(w^\top x)\right)^2 + \epsilon^2 + 2\epsilon\left(\sigma(w_*^\top x) - \sigma(w^\top x)\right)\right] \tag{18}$$

$$= \frac{1}{2}\mathbb{E}_{x,y}\left[\left(\sigma(w_*^\top x) - \sigma(w^\top x)\right)^2 + \epsilon^2\right] \tag{19}$$

$$\geq L(w_*) \tag{20}$$

where 19 is because $\mathbb{E}[\epsilon] = 0$ and $\epsilon \perp x$. Therefore, w_* is a global minimum of $L(w)$.

In addition, show that

$$\langle \nabla L(w), w - w_* \rangle \geq \gamma^2 \lambda \cdot ||w - w_*||_2^2 \tag{21}$$

We can show this by evaluating the gradient (3) and exploiting the assumptions/constraints provided for the problem, and by utilizing the mean value theorem. First, now that we can slightly simplify 3 as follows:

$$\nabla L(w) = \mathbb{E}_{x,y}\left[\left(\sigma(w^\top x) - y\right)\sigma'(w^\top x)x\right] \tag{22}$$

$$= \mathbb{E}_{x,y}\left[\left(\sigma(w^\top x) - \sigma(w_*^\top x)\right)\sigma'(w^\top x)x\right] \quad [\mathbb{E}[\epsilon] = 0, \epsilon \perp x] \tag{23}$$

$$\implies \langle \nabla L(w), w - w_* \rangle = \mathbb{E}\left[\left(\sigma(w^\top x) - \sigma(w_*^\top x)\right)\sigma'(w^\top x)\langle w - w_*, x\rangle\right] \tag{24}$$

By the mean value theorem, we know that $\exists t \in [w^\top x, w_*^\top x]$ such that $\sigma'(t) = \frac{\sigma(w^\top x) - \sigma(w_*^\top x)}{w^\top x - w_*^\top x}$. Therefore

$$\langle \nabla L(w), w - w_* \rangle = \mathbb{E}\left[\sigma'(t)\sigma'(w^\top x)(w^\top x - w_*^\top x)^2\right] \tag{25}$$

Since both t and $w^\top x$ are $\in [-BR, BR]$, we know that $\sigma'(t)\sigma'(w^\top x) \geq \gamma^2$. Furthermore, since

$$\mathbb{E}\left[\left(w^\top x - w_*^\top x\right)^2\right] = \mathbb{E}\left[(w - w_*)^T x x^T (w - w_*)\right] \geq (w - w_*)^\top \lambda I_d (w - w_*) \tag{26}$$

$$= \lambda\,||w - w_*||_2^2 \tag{27}$$

we can plug this back in to obtain the desired result:

$$\langle \nabla L(w), w - w_* \rangle = \mathbb{E}\left[\sigma'(t)\sigma'(w^\top x)(w^\top x - w_*^\top x)^2\right] \tag{28}$$

$$\geq \gamma^2 \mathbb{E}\left[\left(w^\top x - w_*^\top x\right)^2\right] \tag{29}$$

$$\geq \gamma^2 \lambda \left|\left|w - w_*\right|\right|_2^2 \tag{30}$$

for all w such that $||w||_2 \leq R$.

Problem 1(c) Upper Bounds on the Hessians

Show that

$$\left|\left|\nabla^2 L(w)\right|\right|_{op} \leq 2B^2 \quad \textit{and} \quad \left|\left|\nabla^2 \hat{L}(w)\right|\right|_{op} \leq 3B^2 \tag{31}$$

Proceed by using the provided hint about $||\cdot||_{op}$, then exploiting some of the problem assumptions, to obtain the desired results.

$$\begin{aligned}
\left|\left|\nabla^2 L(w)\right|\right|_{op} &= \sup_{||v||_2 \leq 1} v^\top \mathbb{E}_{x,y}\left[\left(\sigma''(w^\top x)\left(\sigma(w^\top x) - y\right) + \sigma'(w^\top x)^2\right) xx^\top\right] v && (32)\\
&= \sup_{||v||_2 \leq 1} \mathbb{E}_{x,y}\left[\left(\sigma''(w^\top x)\left(\sigma(w^\top x) - y\right) + \sigma'(w^\top x)^2\right) \langle x, v\rangle^2\right] && (33)\\
\text{[Cauchy-Schwarz]} \quad &\leq \sup_{||v||_2 \leq 1} \mathbb{E}_{x,y}\left[\left(\sigma''(w^\top x)\left(\sigma(w^\top x) - y\right) + \sigma'(w^\top x)^2\right) ||x||_2^2\, ||v||_2^2\right] && (34)\\
[||x||_2 \leq B, ||v||_2 \leq 1] \quad &\leq B^2 \mathbb{E}_{x,y}\left[\sigma''(w^\top x)\left(\sigma(w^\top x) - y\right) + \sigma'(w^\top x)^2\right] && (35)\\
\sup_t \{|\sigma'(t)|, |\sigma''(t)|\} \leq 1 \quad &\leq B^2 \mathbb{E}_{x,y}\left[\left(\sigma(w^\top x) - y\right) + 1\right] && (36)\\
[\mathbb{E}[\epsilon] = 0] \quad &= B^2 \mathbb{E}_x\left[\sigma(w^\top x) - \sigma(w_*^\top x) + 1\right] && (37)\\
[\sigma(t) \in [0,1]] \quad &\leq 2B^2 && (38)
\end{aligned}$$

The proof for $\left|\left|\nabla^2 \hat{L}(w)\right|\right|_{op} \leq 3B^2$ is the same sequence of logic[1] up until before 37, since for $\hat{L}$ we aren't taking an expectation (the data has already been sampled), meaning we have

$$\begin{aligned}
\left|\left|\nabla^2 \hat{L}(w)\right|\right|_{op} &\leq B^2 \frac{1}{n}\sum_{i=1}^n 1 + \sigma(w^\top x^{(i)}) - y^{(i)} && (39)\\
&= B^2 \frac{1}{n}\sum_{i=1}^n 1 + \sigma(w^\top x^{(i)}) - \sigma(w_*^\top x^{(i)}) - \epsilon^{(i)} && (40)\\
&\leq B^2 \frac{1}{n}\sum_{i=1}^n 3 && (41)\\
&= 3B^2 && (42)
\end{aligned}$$

where 41 follows from $|\epsilon^{(i)}| \leq 1$, and $\sigma(t) \in [0,1]$.

[1] By "logic" I'm referring to the justifications in blue on the LHS of each step.

Problem 1(d) Norm Bound via Covering

Show that $(\forall x \in \mathbb{R}^d)\ ||x||_2 \leq 2 \sup_{v \in N} \langle x, v \rangle$

First, note that for $x = 0$, this is trivially true. Next $\forall x \neq 0$,

$$||x||_2^2 = \langle x, x \rangle \tag{43}$$

$$= ||x||_2 \left\langle x, \frac{x}{||x||_2} \right\rangle \quad \text{[homogeneity]} \tag{44}$$

$$\leq ||x||_2 \sup_{||v||_2=1} \langle x, v \rangle \tag{45}$$

where the last step follows from the orthogonal decomposition of any vector $v = cx + w$, where $\langle x, w \rangle = 0$, i.e.

$$\sup_{||v||_2=1} \langle x, v \rangle = \sup_{||v||_2=1} \left\langle x, \underbrace{\frac{\langle v, x \rangle}{||x||_2^2} x}_{cx} + \underbrace{\left(v - \frac{\langle v, x \rangle}{||x||_2^2} x \right)}_{w} \right\rangle \tag{46}$$

and thus is maximized when v is a scalar multiple of x. Recapping, we now have

$$||x||_2 \leq \sup_{||v||_2=1} \langle x, v \rangle \tag{47}$$

$$\leq \sup_{||v||_2 \leq 1} \langle x, v \rangle \tag{48}$$

where the last step follows from the reasons just described (46). Since, by definition, $\exists v \in N(B(1), \frac{1}{2})$ within $\epsilon = \frac{1}{2}$ from the optimal solution, we have the final result that

$$||x||_2 \leq 2 \sup_{v \in N(B(1), \frac{1}{2})} \langle x, v \rangle \tag{49}$$

Problem 1(e) Point-Wise Concentration of Gradients

For any $w \in \mathbb{R}^d$ and $t > 0$, show that

$$Pr\left[\left|\left|\nabla \hat{L}(w) - \nabla L(w)\right|\right|_2 \geq t\right] \leq \exp\left(-\frac{nt^2}{32B^2} + d\log 5\right) \tag{50}$$

We begin by using the result from part (d), and by noting that for any $f(X)$, $g(X) \leq f(X)$ for all values of a random variable X, that $\Pr[g(X) \geq t] \leq \Pr[f(X) \geq t]$.

$$\Pr\left[\left|\left|\nabla \hat{L}(w) - \nabla L(w)\right|\right|_2 \geq t\right] \leq \Pr\left[2 \sup_{v \in N\left(\mathcal{B}(1), \frac{1}{2}\right)} \left\langle \nabla \hat{L}(w) - \nabla L(w), v \right\rangle \geq t\right] \tag{51}$$

For compactness, let $f(v) \triangleq \left\langle \nabla \hat{L}(w) - \nabla L(w), v \right\rangle$. Recall that we can interpret probabilities involving sup as

$$\Pr\left[2 \sup_{v \in N(B(1), \frac{1}{2})} f(v) \geq t\right] = \Pr\left[\exists v \in N\left(\mathcal{B}(1), \frac{1}{2}\right) \text{ s.t. } f(v) \geq \frac{t}{2}\right] \tag{52}$$

$$\leq \sum_{v \in N\left(\mathcal{B}(1), \frac{1}{2}\right)} \Pr\left[f(v) \geq \frac{t}{2}\right] \tag{53}$$

which gives us a familiar union-bound form. Furthermore, notice that we can interpret $f(v)$ as an empirical mean subtracted by their expectation:

$$f(v) \triangleq \left\langle \nabla \hat{L}(w) - \nabla L(w), v \right\rangle \tag{54}$$

$$= \frac{1}{n}\sum_{i=1}^{n} \left\langle z^{(i)}, v \right\rangle - \left\langle \mathbb{E}_{x,y}\left[\nabla \hat{L}(w)\right], v \right\rangle \tag{55}$$

$$= \frac{1}{n}\sum_{i=1}^{n} \left\langle z^{(i)}, v \right\rangle - \mathbb{E}_{x,y}\left[\frac{1}{n}\sum_{i=1}^{n} \left\langle z^{(i)}, v \right\rangle\right] \tag{56}$$

$$\text{where} \quad z^{(i)} \triangleq \left(\sigma(w^\top x^{(i)}) - y^{(i)}\right)\sigma'(w^\top x^{(i)})x^{(i)} \tag{57}$$

If we can show the $\left\langle z^{(i)}, v \right\rangle$ are bounded i.i.d. random variables, we can then use Hoeffding's inequality to obtain the desired result.

$$\left\langle z^{(i)}, v \right\rangle = \left(\sigma(w^\top x^{(i)}) - y^{(i)}\right)\sigma'(w^\top x^{(i)}) \left\langle x^{(i)}, v \right\rangle \tag{58}$$

$$\leq B\left(\sigma(w^\top x^{(i)}) - y^{(i)}\right)\sigma'(w^\top x^{(i)}) \quad [||x||_2 \leq B, ||v||_2 \leq 1] \tag{59}$$

$$\leq 2B \quad |\sigma'(t)| \leq 1, \sigma(t) \in [0,1], |\epsilon^{(i)}| \leq 1 \tag{60}$$

and a similar argument using the other side of the absolute values shows that $\left\langle z^{(i)}, v \right\rangle \geq -2B$, so we have our bounded random variable $-2B \leq \left\langle z^{(i)}, v \right\rangle \leq 2B$.

Resuming from 53 and plugging directly into Hoeffding's inequality yields:

$$\Pr\left[2 \sup_{v \in N\left(\mathcal{B}(1), \frac{1}{2}\right)} f(v) \geq t\right] \leq \sum_{v \in N\left(\mathcal{B}(1), \frac{1}{2}\right)} \Pr\left[\frac{1}{n}\sum_{i=1}^{n}\left\langle z^{(i)}, v\right\rangle - \mathbb{E}_{x,y}\left[\frac{1}{n}\sum_{i=1}^{n}\left\langle z^{(i)}, v\right\rangle\right] \geq \frac{t}{2}\right] \tag{61}$$

$$\leq \sum_{v \in N\left(\mathcal{B}(1), \frac{1}{2}\right)} \exp\left(-\frac{2n^2(t/2)^2}{\sum_{i=1}^{n}(2B-(-2B))^2}\right) \tag{62}$$

$$= \sum_{v \in N\left(\mathcal{B}(1), \frac{1}{2}\right)} \exp\left(-\frac{n^2t^2}{2\sum_{i=1}^{n}16B^2}\right) \tag{63}$$

$$= \sum_{v \in N\left(\mathcal{B}(1), \frac{1}{2}\right)} \exp\left(-\frac{nt^2}{32B^2}\right) \tag{64}$$

$$\leq 5^d \exp\left(-\frac{nt^2}{32B^2}\right) = \exp\left(-\frac{nt^2}{32B^2} + d\log 5\right) \tag{65}$$

where the last line follows from $|N\left(\mathcal{B}(1), \frac{1}{2}\right)| \leq 5^d$.

Problem 1(f) Union Bounds

Show that for any $0 < \epsilon < R$, we can take $N(\mathcal{B}(R), \epsilon)$ to be an ϵ-covering of $\mathcal{B}(R)$ and get

$$Pr\left[\sup_{w \in N(\mathcal{B}(R),\epsilon)} \left|\left|\nabla \hat{L}(w) - \nabla L(w)\right|\right|_2 \geq t\right] \leq \exp\left(-\frac{nt^2}{32B^2} + d\log\frac{15R}{\epsilon}\right) \tag{66}$$

Similar to part (e), we can rewrite the probability in terms of a union bound:

$$\Pr\left[\sup_{w \in N(\mathcal{B}(R),\epsilon)} \left|\left|\nabla \hat{L}(w) - \nabla L(w)\right|\right|_2 \geq t\right] = \Pr\left[\exists w \in N(\mathcal{B}(R), \epsilon) \text{ s.t. } \left|\left|\nabla \hat{L}(w) - \nabla L(w)\right|\right|_2 \geq t\right] \tag{67}$$

$$\leq \sum_{w \in N(\mathcal{B}(R),\epsilon)} \Pr\left[\left|\left|\nabla \hat{L}(w) - \nabla L(w)\right|\right|_2 \geq t\right] \tag{68}$$

$$\leq |N(\mathcal{B}(R), \epsilon)| \exp\left(-\frac{nt^2}{32B^2} + d\log 5\right) \tag{69}$$

where the last line follows from the result derived for part (e). From homework 2 we know that

$$|N(\mathcal{B}(R), \epsilon)| \leq \left(1 + \frac{2R}{\epsilon}\right)^d \tag{70}$$

$$\leq \left(\frac{3R}{\epsilon}\right)^d \quad [0 < \epsilon < R] \tag{71}$$

Plugging this back in yields the desired result:

$$\Pr\left[\sup_{w \in N(\mathcal{B}(R),\epsilon)} \left|\left|\nabla \hat{L}(w) - \nabla L(w)\right|\right|_2 \geq t\right] \leq \left(\frac{3R}{\epsilon}\right)^d \exp\left(-\frac{nt^2}{32B^2} + d\log 5\right) \tag{72}$$

$$= \exp\left(-\frac{nt^2}{32B^2} + d\log\frac{15R}{\epsilon}\right) \tag{73}$$

Problem 1(g) Uniform Convergence of Gradients

First Part

Assume $BR \geq 1$. Show that w.p. at least $1-\delta$

$$\sup_{w \in \mathcal{B}(R)} \left|\left|\nabla \hat{L}(w) - \nabla L(w)\right|\right|_2 \leq C_1 \cdot B\sqrt{\frac{d\left(C_2 + \log(nBR)\right) + \log\frac{1}{\delta}}{n}} \tag{74}$$

My proof will proceed largely the same way as the proof for Theorem 4.8 in the scribe notes.

1. Denote the event $E \triangleq \{\sup_{w \in N(\mathcal{B}(R),\epsilon)} \left|\left|\nabla \hat{L}(w) - \nabla L(w)\right|\right|_2 \leq \delta\}$
2. From part (f) we know that

$$\Pr[E] \geq 1 - \exp\left(-\frac{n\delta^2}{32B^2} + d\log\frac{15R}{\epsilon}\right) \tag{75}$$

$$\geq 1 - \exp\left(-\frac{n\delta^2}{32B^2} + Cd + d\log(nBR)\right) \tag{76}$$

   if we set $\epsilon := \frac{1}{B}\sqrt{d/n}$.

3. By definition of an ϵ-cover, $\forall w \in \mathcal{B}(R)$, $\exists w_0 \in N\left(\mathcal{B}(R), \epsilon\right)$ such that $||w - w_0||_2 \leq \epsilon$. Combining this with the result from part (c), we have

$$||\nabla L(w) - \nabla L(w_0)||_2 \leq 2B^2 ||w - w_0||_2 \leq 2B^2\epsilon = 2B\sqrt{d/n} \tag{77}$$

$$\left|\left|\nabla \hat{L}(w) - \nabla \hat{L}(w_0)\right|\right|_2 \leq 3B^2 ||w - w_0||_2 \leq 3B^2\epsilon = 3B\sqrt{d/n} \tag{78}$$

4. We can then get a bound on $\left|\left|\nabla \hat{L}(w) - \nabla L(w)\right|\right|_2$ by using telescoping sums. Conditional on the event E,

$$\left|\left|\nabla \hat{L}(w) - \nabla L(w)\right|\right|_2 \leq \left|\left|\nabla \hat{L}(w) - \nabla \hat{L}(w_0)\right|\right|_2 + \left|\left|\nabla \hat{L}(w_0) - \nabla L(w_0)\right|\right|_2 + ||\nabla L(w_0) - \nabla L(w)||_2 \tag{79}$$

$$\leq 3B^2\epsilon + \delta + 2B^2\epsilon \tag{80}$$

$$= 5B^2\epsilon + \delta \tag{81}$$

$$= 5B\sqrt{d/n} + \delta \tag{82}$$

   for all $w \in \mathcal{B}(R)$, where $5B^2\epsilon$ represents the error from our brute-force discretization. In other words, we currently have that, with probability at least eq 76,

$$\left|\left|\nabla \hat{L}(w) - \nabla L(w)\right|\right|_2 \leq 5B\sqrt{d/n} + \delta \tag{83}$$

To obtain the desired result, we can set

$$\delta' = \exp\left(-\frac{n\delta^2}{32B^2} + Cd + d\log(nBR)\right) \tag{84}$$

$$\implies \delta = \sqrt{32}B\sqrt{\frac{d\left(C_2 + \log(nBR)\right) + \log\frac{1}{\delta'}}{n}} \tag{85}$$

which yields with probability at least $1 - \delta'$ that

$$\left|\left|\nabla\hat{L}(w) - \nabla L(w)\right|\right|_2 \leq 5B\sqrt{d/n} + \sqrt{32}B\sqrt{\frac{d\left(C_2 + \log(nBR)\right) + \log\frac{1}{\delta'}}{n}} \tag{86}$$

Noting that the sum on the RHS has the form

$$C_1'B\sqrt{d/n} + C_2'B\sqrt{d/n}\sqrt{(C_2 + \log(nBR)) + \log\frac{1}{\delta'}} \tag{87}$$

$$= (C_1' + C_2')B\sqrt{d/n}\sqrt{(C_2 + \log(nBR)) + \log\frac{1}{\delta'}} \tag{88}$$

which is the desired result with $C_1 := C_1' + C_2'$.

Second Part

Conditioned on the event above, show that $\forall w \in \mathcal{B}(R)$*, we have*

$$\left\langle \nabla \hat{L}(w) - \nabla L(w), w - w_* \right\rangle \leq C_1 \cdot B \sqrt{\frac{d\left(C_2 + \log(nBR)\right) + \log \frac{1}{\delta}}{n}} \cdot ||w - w_*||_2 \tag{89}$$

This is just a direct application of the Cauchy-Schwarz inequality and the previous result:

$$\begin{aligned}
\left\langle \nabla \hat{L}(w) - \nabla L(w), w - w_* \right\rangle &\leq \left|\left| \nabla \hat{L}(w) - \nabla L(w) \right|\right|_2 ||w - w_*||_2 && (90)\\
&\leq \left(\sup_{w \in \mathcal{B}(R)} \left|\left| \nabla \hat{L}(w) - \nabla L(w) \right|\right|_2 \right) ||w - w_*||_2 && (91)\\
&\leq C_1 \cdot B \sqrt{\frac{d\left(C_2 + \log(nBR)\right) + \log \frac{1}{\delta}}{n}} \cdot ||w - w_*||_2 && (92)
\end{aligned}$$

Problem 1(h) The Training Loss Has No Bad Local Minima

Show that w.p. at least $1-\delta$, $\forall \|w\|_2 \leq R$, *if* $\nabla \hat{L}(w) = 0$, *then*

$$\|w - w_*\|_2 \leq \frac{C_1 B}{\gamma^2 \lambda} \sqrt{\frac{d\left(C_2 + \log(nBR)\right) + \log \frac{1}{\delta}}{n}} \tag{93}$$

Recall the result from part (g):

$$\begin{aligned}
\left\langle \nabla \hat{L}(w) - \nabla L(w), w - w_* \right\rangle &\leq \left\| \nabla \hat{L}(w) - \nabla L(w) \right\|_2 \|w - w_*\|_2 && (94)\\
&= \|\nabla L(w)\|_2 \|w - w_*\|_2 && (95)\\
&\leq C_1 \cdot B \sqrt{\frac{d\left(C_2 + \log(nBR)\right) + \log \frac{1}{\delta}}{n}} \cdot \|w - w_*\|_2 && (96)
\end{aligned}$$

From part (b) we also know that

$$\langle \nabla L(w), w - w_* \rangle \geq \gamma^2 \lambda \|w - w_*\|_2^2 \tag{97}$$

Combining these two, while exploiting that $\nabla \hat{L}(w) = 0$, yields

$$\begin{aligned}
\gamma^2 \lambda \|w - w_*\|_2^2 &\leq \langle \nabla L(w), w - w_* \rangle && (98)\\
&= \left\langle \nabla \hat{L}(w) - \nabla L(w), w - w_* \right\rangle && (99)\\
&\leq C_1 \cdot B \sqrt{\frac{d\left(C_2 + \log(nBR)\right) + \log \frac{1}{\delta}}{n}} \cdot \|w - w_*\|_2 && (100)
\end{aligned}$$

which implies the desired result:

$$\begin{aligned}
\gamma^2 \lambda \|w - w_*\|_2 &\leq C_1 \cdot B \sqrt{\frac{d\left(C_2 + \log(nBR)\right) + \log \frac{1}{\delta}}{n}} && (101)\\
\implies \|w - w_*\|_2 &\leq \frac{C_1 B}{\gamma^2 \lambda} \sqrt{\frac{d\left(C_2 + \log(nBR)\right) + \log \frac{1}{\delta}}{n}} && (102)
\end{aligned}$$

Problem 2 Part I Generalization Bounds of the ℓ_1-SVM

For $x, y \sim P$, let $x \sim Unif\{-1, +1\}^d$, and

$$y = \begin{cases} 1 & \text{if } x^T e_1 = 1 \\ -1 & \text{if } x^T e_1 = -1 \end{cases} \tag{103}$$

Part (a)

Show that $\gamma_{\ell_1} \geq 1$.

First, note that $y_i x_i^T e_1 = 1 \; \forall i$. In fact, since we can express α the standard basis as a linear combination of the standard basis vector e_i, we have

$$\alpha = \sum_{i=1}^{d} \alpha_i e_i \tag{104}$$

$$y_i x_i^T \alpha = \alpha_1 + y \sum_{i=2}^{d} \alpha_i x_i^T e_i \tag{105}$$

In other words, we could just set $\alpha_i = 0$ for all $2 \leq i \leq d$, and maximize α_1 (i.e. set it to 1). This would achieve $\gamma_{\ell_1} = 1$. Since the optimization problem is to maximize α, this means that $\alpha_{\ell_1} \geq 1$.

Part (b)

Let $\mathcal{H}^1 \triangleq \{x \mapsto x^T \alpha : ||\alpha||_1 \leq 1\}$. Show that $R_n(\mathcal{H}^1) \leq C\sqrt{\log(2d)/n}$.

We can simply apply Theorem 5.7 from the notes. Here, $B = 1$ since $||\alpha||_1 \leq 1$, and $C = 1$[2] since $||x||_\infty \leq 1$. A direct application of the theorem then yields

$$R_S(\mathcal{H}^1) \leq \sqrt{2}\sqrt{\frac{\log(2d)}{n}} \tag{106}$$

$$\text{Therefore} \qquad R_n(\mathcal{H}^1) \triangleq \mathbb{E}\left[R_S(\mathcal{H}^1)\right] \tag{107}$$

$$\leq C\sqrt{\frac{\log(2d)}{n}} \tag{108}$$

for some universal constant $C > 0$ (for this derivation, at least, it was $\sqrt{2}$).

[2]Here I am referencing the "C" from Theorem 5.7, not this homework problem.

Part (c)

Show that $\gamma_{\ell_2} \geq 1$.

This can be shown with the exact same technique as part (a). Namely, It is still true that $\forall i$,

$$y_i x_i^T \alpha = \alpha_1 + y \sum_{i=2}^{d} \alpha_i x_i^T e_i \tag{109}$$

and since setting $\alpha_1 = 1$ and $\alpha_{i \neq 1} = 0$ still satisfies $||\alpha||_2 \leq 1$, the same argument from part (a) applies:

> *This would achieve $\gamma_{\ell_2} = 1$. Since the optimization problem is to maximize α, this means that $\alpha_{\ell_2} \geq 1$.*

Part (d)

Let $\mathcal{H}^2 \triangleq \{x \mapsto x^T \alpha : ||\alpha||_2 \leq 1\}$. Show that $R_n(\mathcal{H}^2) \leq C\sqrt{d/n}$.

In much the same way as part (b), we can apply a Theorem from the notes and make minor adjustments based on the specifics of this problem. Using the notation from Theorem 5.5, here we have $B = 1$ since $||\alpha||_2 \leq 1$, and $C = \sqrt{d}$ since $\mathbb{E}_{x \sim \{-1,+1\}^d}\left[||x||_2^2\right] = d$. Applying the theorem directly yields

$$R_n(\mathcal{H}) \leq C\sqrt{\frac{d}{n}} \tag{110}$$

for some universal constant $C > 0$ (for this derivation, at least, it was 1).

Part (e)

Show that $\gamma_{\ell_2} \leq \max_{||\alpha||_2 \leq 1} \frac{1}{n} \sum_{i=1}^{n} y_i x_i^T \alpha$.

For any given fixed value of α, define $i_{min} \triangleq \arg\min_{i \in [n]} y_i x_i^T \alpha$. Then

$$\frac{1}{n} \sum_{i=1}^{n} y_i x_i^T \alpha = \frac{1}{n} y_{i_{min}} x_{i_{min}}^T \alpha + \frac{1}{n} \sum_{\substack{i \in [n] \\ i \neq i_{min}}} y_i x_i^T \alpha \tag{111}$$

If we choose $\alpha := \alpha_{\ell_2}$, this becomes

$$\frac{1}{n} \sum_{i=1}^{n} y_i x_i^T \alpha_{\ell_2} = \frac{1}{n} \gamma_{\ell_2} + \frac{1}{n} \sum_{\substack{i \in [n] \\ i \neq i_{min}}} y_i x_i^T \alpha_{\ell_2} \tag{112}$$

By definition of γ_{ℓ_2}, every term in the summation on the RHS of 112 is greater than or equal to γ_{ℓ_2}. Therefore, when $\alpha := \alpha_{\ell_2}$, we have

$$\frac{1}{n} \sum_{i=1}^{n} y_i x_i^T \alpha_{\ell_2} \geq \frac{1}{n} \gamma_{\ell_2} + \frac{1}{n} \sum_{\substack{i \in [n] \\ i \neq i_{min}}} \gamma_{\ell_2} \tag{113}$$

$$= \gamma_{\ell_2} \tag{114}$$

Lastly, since

$$\max_{||\alpha||_2 \leq 1} \frac{1}{n} \sum_{i=1}^{n} y_i x_i^T \alpha \geq \frac{1}{n} \sum_{i=1}^{n} y_i x_i^T \alpha_{\ell_2} \geq \gamma_{\ell_2} \tag{115}$$

we have the desired result.

Part (f)

Prove that $\gamma_{\ell_2} \leq \frac{1}{n} \left|\left|\sum_{i=1}^{n} y_i x_i\right|\right|_2$.

Using the bound we obtained from part (e), if we can show the following, then we'll have achieved the desire result:

$$\max_{||\alpha||_2 \leq 1} \sum_{i=1}^{n} y_i x_i^T \alpha \leq \left|\left| \sum_{i=1}^{n} y_i x_i \right|\right|_2 \tag{116}$$

My approach will be to upper bound the LHS by an intermediate quantity, and show that the RHS must be greater than or equal to this intermediate quantity.

First, let's show that the LHS can be no larger than n, and that this bound is maximally tight, in that setting $\alpha := e_1$ achieves it. Recall that

$$\sum_{i=1}^{n} y_i x_i^T \alpha = \sum_{i=1}^{n} \left(\alpha_1 + \sum_{j=2}^{d} y_i (x_i)_j \alpha_j \right) \tag{117}$$

Since we are finding the α that maximizes this quantity, constrained such that $||\alpha||_2 \leq 1$, and since each coefficient $y_i(x_i)_j$ is at most 1, an optimal choice for α is e_1. In other words, we have

$$\max_{||\alpha||_2 \leq 1} \sum_{i=1}^{n} y_i x_i^T \alpha \leq n \tag{118}$$

Similarly, since we know that the first coefficient $y_i(x_i)_1 = 1$ $(\forall i)$ from the problem definition, it must be the case that

$$\left(\sum_{i=1}^{n} y_i x_i \right)_1 = \sum_{i=1}^{n} y_i (x_i)_1 = n \tag{119}$$

Therefore

$$\left|\left| \sum_{i=1}^{n} y_i x_i \right|\right|_2 = \sqrt{n^2 + \sum_{j=2}^{d} \left(\sum_{i=1}^{n} y_i (x_i)_j \right)^2} \tag{120}$$

$$\geq n \tag{121}$$

Combining part (e) with 118 and 121 yields the desired result:

$$\gamma_{\ell_2} \leq \max_{||\alpha||_2 \leq 1} \frac{1}{n} \sum_{i=1}^{n} y_i x_i^T \alpha \leq 1 \leq \frac{1}{n} \left|\left| \sum_{i=1}^{n} y_i x_i \right|\right|_2 \tag{122}$$

Part (g)

Fix $j \in \{2, \ldots, d\}$ and $0 < \delta < 1$. Show that w.p. at least $1 - \delta$

$$\left(\sum_{i=1}^{n} y_i x_i^T e_j \right)^2 \leq Cn \log \frac{2}{\delta} \tag{123}$$

We can treat each x_i and y_i as bounded random variables in $[-1, 1]$. Therefore, we can denote each summation term $z_{i,j} \triangleq y_i x_i^T e_j$ (for $2 \leq j \leq d$) as a bounded random variable also in $[-1, 1]$. Also, since $y_i \perp (x_i)_{j \neq 1}$, we have that $\mathbb{E}[z_{i,j}] = 0$:

$$\mathbb{E}[z_{i,j}] = \mathbb{E}_{x,y}\left[y x^T e_{j \neq 1} \right] = \mathbb{E}[y]\, \mathbb{E}\left[x^T e_{j \neq 1} \right] = 0 \tag{124}$$

Then, from previous homeworks, we know that $z_{i,j}$ is sub-Gaussian with parameter $\sigma^2 = (1 - (-1))^2/4 = 1$. Similarly, since $z_{i,j} \perp z_{i',j}$ for $i \neq i' \in [n]$, the sum $\sum_{i=1}^{n} z_{i,j}$ is also sub-Gaussian with parameter $\sigma^2 = \sum_{i=1}^{n} 1 = n$. Therefore

$$\Pr\left[\left(\sum_{i=1}^{n} y_i x_i^T e_j \right)^2 \leq \epsilon \right] = \Pr\left[\left(\sum_{i=1}^{n} z_{i,j} \right)^2 \leq \epsilon \right] \tag{125}$$

$$= \Pr\left[\left| \sum_{i=1}^{n} z_{i,j} \right| \leq \sqrt{\epsilon} \right] \tag{126}$$

$$= \Pr\left[\left| \sum_{i=1}^{n} z_{i,j} - \mathbb{E}\left[\sum_{i'=1}^{n} z_{i',j} \right] \right| \leq \sqrt{\epsilon} \right] \tag{127}$$

$$\geq 1 - 2 \exp\left(-\frac{\epsilon}{2n} \right) \tag{128}$$

It is straightforward to verify that if $\epsilon := Cn \log \frac{2}{\delta}$, that

$$\exp\left(\frac{\epsilon}{Cn} \right) = \frac{2}{\delta} \implies \delta = 2 \exp\left(\frac{-\epsilon}{Cn} \right) \tag{129}$$

which gives us the desired result with $C = 2$.

Part (h)

Show that w.p. at least $\frac{1}{2}$,

$$\gamma_{\ell_2} \leq \frac{1}{n}\left|\left|\sum_{i=1}^{n} y_i x_i\right|\right|_2 \leq C\sqrt{\frac{n + d\log 2d}{n}} \tag{130}$$

We already proved the first part of the inequality with probability 1 in part (f). For the second part, begin by expanding

$$\Pr\left[\frac{1}{n}\left|\left|\sum_{i=1}^{n} y_i x_i\right|\right|_2 \leq C\sqrt{\frac{n + d\log 2d}{n}}\right] = \Pr\left[\left|\left|\sum_{i=1}^{n} y_i x_i\right|\right|_2^2 \leq n^2 C^2 \frac{n + d\log 2d}{n}\right] \tag{131}$$

$$= \Pr\left[\left|\left|\sum_{i=1}^{n} y_i x_i\right|\right|_2^2 \leq nC^2(n + d\log 2d)\right] \tag{132}$$

Next, we can decompose the LHS as follows to utilize our result from part (g):

$$\left|\left|\sum_{i=1}^{n} y_i x_i\right|\right|_2^2 = \left(\sum_{i=1}^{n} y_i x_i\right)_1^2 + \sum_{j=2}^{d}\left(\sum_{i=1}^{n} y_i x_i\right)_j^2 \tag{133}$$

$$= n^2 + \sum_{j=2}^{d}\left(\sum_{i=1}^{n} y_i x_i\right)_j^2 \tag{134}$$

$$= n^2 + \sum_{j=2}^{d}\underbrace{\left(\sum_{i=1}^{n} y_i x_i^\top e_j\right)^2}_{\triangleq g_j} \tag{135}$$

where 134 follows from the definition for y_i as a function of $(x_i)_1 \equiv x_i^\top e_1$.Therefore, noting the definition of g_j in 135, in order to obtain the desired result, we just need to show that

$$\Pr\left[n^2 + \sum_{j=2}^{d} g_j \leq n^2 C^2 + nC^2 d\log 2d\right] \geq \frac{1}{2} \tag{136}$$

From part (g), we know that

$$\Pr\left[g_j \leq C' n \log\frac{2}{\delta}\right] \geq 1 - \delta \tag{137}$$

where I'm distinguishing C' as the constant associated with part (g). If we set $\delta := \frac{1}{2d}$, note that $1 - \delta \geq \frac{1}{2}$. Therefore, with probability at least $\frac{1}{2}$, we have that

$$g_j \leq C' n \log 4d \tag{138}$$

$$\implies \sum_{j=2}^{d} g_j \leq (d-1)C' n \log 4d \tag{139}$$

$$\implies n^2 + \sum_{j=2}^{d} g_j \leq n^2 + (d-1)C' n \log 4d \leq n^2 + nC' d \log 2d \tag{140}$$

which is the desired result.