Honours Math for Machine Learning HW2

Saina Koukpari - McGill University

Problem 1.2

Since $(\ell_{class}, c, C_{class})$ is an upper bound for the error, we have that for all $h \in \mathbb{R}, y \in \mathcal{Y}_{\pm}$,

$$\begin{aligned} &\ell_{class}(h,y) \geq C_{class}\ell_{0-1}(c(h),y) \\ \implies &\frac{1}{C_{class}}\ell_{class}(h,y) \geq \ell_{0-1}(c(h),y) \end{aligned}$$

Then we have,

$$\begin{aligned} \hat{L}_{0-1}(c(h)) &= \frac{1}{m}\sum_{i=1}^{m}\ell_{0-1}(h_i,y_i) \\ &\leq \frac{1}{C_{class}}\frac{1}{m}\sum_{i=1}^{m}\ell_{class}(h_i,y_i) \\ &= \frac{1}{C_{class}}\hat{L}_{class}(c(h)) \quad \blacksquare \end{aligned}$$

Problem 1.3

i) The quadratic loss is an upper bound for the zero-one loss since we can find a constant $C_{class} > 0$ such that $\ell_2(h,y) \geq C_{class}\ell_{0-1}(c(h),y)$. Consider the case for $y = 1$, then

$$\ell_2(h,1) = (h-1)^2 = \begin{cases} 0 & h = 1 \\ > 0 & h \neq 1 \end{cases} \geq \begin{cases} 0 & h = 1 \\ C & h \neq 1 \end{cases} = C\ell_{0-1}(c(h),1)$$

For $0 < C \leq 1$, we can choose $C = 1$ to set $(h-1)^2$ as the bound, as seen in the following plot,

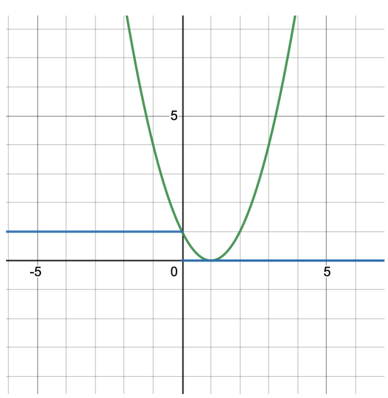

Figure 1: ℓ_{0-1} bounded by $(h-1)^2$

Now consider $y = -1$, then

$$\ell_2(h,1) = (h+1)^2 = \begin{cases} 0 & h = -1 \\ > 0 & h \neq -1 \end{cases} \geq \begin{cases} 0 & h = -1 \\ C & h \neq -1 \end{cases} = C\ell_{0-1}(c(h),1)$$

For $0 < C \leq 1$. Choosing $C = 1$ we have the following plot,

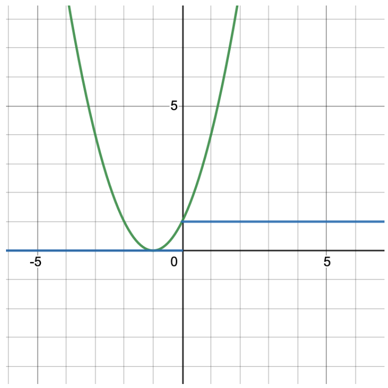

Figure 2: ℓ_{0-1} bounded by $(h+1)^2$

ii) For any constant $C_{class} > 0$, $C_{class}\ell(h,y) = C_{class}|h+y|$ where $y = \pm 1$, changes the function in terms of horizontal stretch or compression and hence it is never the case that $\ell(h,y) \geq C_{class}\ell_{0-1}(c(h),y)$ as there are points on $|h+y|$ which lie below 1 (the upper bound for zero one loss). We can see this visually with the following plots,

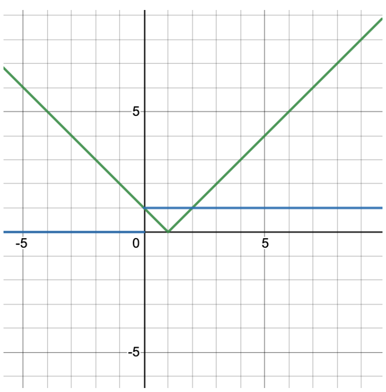

Figure 3: $\ell(h,y) = |h+y|$ for $y = -1$

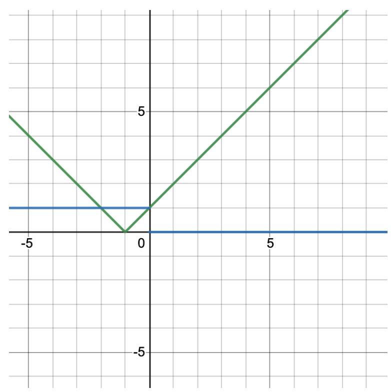

Figure 4: $\ell(h,y) = |h+y|$ for $y = 1$

iii) Since $h < 0$ and $y = 1$, $h \neq y$, so the zero-one loss $\ell_{0-1}(h,1) = 1$ but the given loss function $\ell(h,1) = 0$, i.e. we have a counterexample where the loss function lies below the zero-one loss(hence does not bound it from above). Further, for any constant $C > 0$, $C\ell(h,1) = 0 \implies \ell(h,1) = 0 \leq 1/C \cdot 1 = 1/C\ell_{0-1}(h,1)$ which contradicts the definition of an upper bound for error.

vi) A simple converse is $\ell(h,1) = 1$ for $h < 0$ since then, $\ell_{0-1}(h,1) = 1$ but now the case that our function always bounds the zero-one loss since we can multiply by some constant $C > 0$ so that $\ell(h,1) = 1 \geq 1/C \cdot 1 = 1/C\ell_{0-1}(h,1)$.

Problem 4.1

$$\hat{L}_{abs}(s_w) = \frac{1}{m}\sum_{i=1}^{m} \ell_{abs}(s_w(x_i), y_i) = \frac{1}{m}\sum_{i=1}^{m} \ell_{abs}(x_i - w, y_i) = \frac{1}{m}\sum_{i=1}^{m} \begin{cases} \max(x_i - w, 0) & y_i = -1 \\ \max(w - x_i, 0) & y_i = +1 \end{cases}$$

Equivalently, we can write the above in terms of correct and incorrect results as follows,

$$\hat{L}_{abs}(s_w) = \frac{1}{m}\sum_{i=1}^{m} \begin{cases} 0 & sgn(s) = y \\ |x_i - x| & o.w. \end{cases} = \frac{1}{m}\sum_{i=1}^{m} \max(-(x_i - w)y, 0) = \frac{1}{m}\sum_{i=1}^{m} \max((x_i - w)y, 0)$$

Note that,

$$\frac{d}{dh}\ell_{abs}(s, y) = \begin{cases} 0 & \ell = 0 \\ -y & \ell > 0 \end{cases} = -y\mathbf{1}_{\{(x_i - w)y < 0\}} \quad \text{and} \quad \frac{d}{dw}s_w(x) = -1$$

So setting the derivative equal to zero, the minimizer is such that the following holds,

$$\begin{aligned} \hat{L}'_{abs}(s_w) = \frac{1}{m}\sum_{i=1}^{m} y\mathbf{1}_{\{(x_i - w)y < 0\}} &= 0 \\ \implies \sum_{i=1}^{m} y\mathbf{1}_{\{(x_i - w)y < 0\}} &= 0 \\ \implies \sum_{y_i=1}^{m} \mathbf{1}_{\{(x_i - w) < 0\}} + \sum_{y_i=-1}^{m} -\mathbf{1}_{\{(x_i - w) > 0\}} &= 0 \\ \implies \sum_{y_i=1}^{m} \mathbf{1}_{\{(x_i - w) < 0\}} &= \sum_{y_i=-1}^{m} \mathbf{1}_{\{(x_i - w) > 0\}} \end{aligned}$$

The above using the majority classifier $c_{maj}(s) = sgn(s - 1/2)$ gives,

$$\hat{L}_{abs}(s_w) = \frac{1}{m}\sum_{i=1}^{m} \begin{cases} 0 & sgn(s - 1/2) = y \\ |x_i - x| & o.w. \end{cases} = \frac{1}{m}\sum_{i=1}^{m} \max((x_i - w)y, 0)$$

Then the minimizer must satisfy the condition that,

$$\begin{aligned} \hat{L}'_{abs}(s_w) = \frac{1}{m}\sum_{i=1}^{m} y\mathbf{1}_{\{(x_i - w - 1/2)y < 0\}} &= 0 \\ \implies \sum_{i=1}^{m} y\mathbf{1}_{\{(x_i - w - 1/2)y < 0\}} &= 0 \\ \implies \sum_{y_i=1}^{m} \mathbf{1}_{\{(x_i - w - 1/2) < 0\}} + \sum_{y_i=-1}^{m} -\mathbf{1}_{\{(x_i - w - 1/2) > 0\}} &= 0 \\ \implies \sum_{y_i=1}^{m} \mathbf{1}_{\{(x_i - w - 1/2) < 0\}} &= \sum_{y_i=-1}^{m} \mathbf{1}_{\{(x_i - w - 1/2) > 0\}} \end{aligned}$$

The condition for a minimizer is that $E_n = E_p$ where E_n is the number of false or marginal negatives and E_p is the number of false or marginal positives.

Problem 4.2

By theorem 4.2, the loss is incorrect when $c(s) \neq y$, marginal when $c(s) = y; |s| \leq 1$, and confident when $c(s) = y; |s| \geq 1$.

If $c(s) \neq y$, then when $\text{sgn}(s)$ is negative, $\ell_{margin}(s,y) = \max(0, 1-s) = \max(0, 1-(-s)) = \max(0, 1+s)$ and when $\text{sgn}(s)$ is positive, $\ell_{margin}(s,y) = \max(0, 1+s)$. We see that $1+s \to \infty$ as $s \to \infty$ and $\ell_{margin}(s,y) = [1, \infty)$ for the incorrect pair.

If $c(s) = y; |s| \leq 1$, then when $\text{sgn}(s)$ is negative, $\ell_{margin}(s,y) = \max(0, 1+s) \in [\max(0,0), \max(0,2)] = [0,1]$ since the marginal loss is defined for correct values between 0 and 1. When $\text{sgn}(s)$ is positive, $\ell_{margin}(s,y) = \max(0, 1-s) \in [\max(0,2), \max(0,0)] = [1,0]$. So $\ell_{margin}(s,y) \in [0,1]$ for the marginal pair.

If $c(s) = y; |s| \geq 1$, then when $\text{sgn}(s)$ is negative, $\ell_{margin}(s,y) = \max(0, 1+s) \in [\max(0, -\infty), \max(0,0)] = (-\infty, 0]$ and when $\text{sgn}(s)$ is positive, $\ell_{margin}(s,y) = \max(0, 1-s) \in [\max(0,0), \max(0,\infty)] = [0, \infty)$. The intersect of both cases is 0 and so $\ell_{margin}(s,y) = 0$ for the confident pair.

Problem 4.3

For $\ell_{margin-t}$ with $C_{class} = 1$ and c =sgn, we have

$$\ell_{margin-t} = \begin{cases} \max(0, 1-s/t) & y = 1 \\ \max(0, 1+s/t) & y = -1 \end{cases} \geq \begin{cases} 0 & c(s) = 1 \\ 1 & c(s) \neq 1 \end{cases} = 1 \cdot \ell_{0-1}$$

Furthermore, we know that $\hat{L}_{margin}(s) \geq \hat{L}_{0-1}(c_{sgn}(s))$, since

$$\begin{aligned}
\hat{L}_{margin}(s) &= \frac{1}{m} \sum_{i=1}^{m} \ell_{margin}(s_i, y_i) \\
&= \begin{cases} \frac{1}{m} \sum_{i=1}^{m} \max(0, 1-s/t) & y = 1 \\ \frac{1}{m} \sum_{i=1}^{m} \max(0, 1+s/t) & y = -1 \end{cases} \\
&= \begin{cases} \frac{1}{m} \sum_{i=1}^{m} \max(0, 1-s/t) & y = 1 \\ \frac{1}{m} \sum_{i=1}^{m} \max(0, 1+s/t) & y = -1 \end{cases} \\
&\geq \begin{cases} \frac{1}{m} \sum_{i=1}^{m} 0 & c(s) = y \\ \frac{1}{m} \sum_{i=1}^{m} 1 & c(s) \neq y \end{cases} \\
&= \frac{1}{m} \sum_{i=1}^{m} \ell_{0-1}(c_{sgn}(s_i), y_i) \\
&= \hat{L}_{0-1}(c_{sgn}(s))
\end{aligned}$$

Problem 4.4

i) Setting $t = 1$ we have,

$$\ell_{margin,1}(s,y) = \begin{cases} 0 & sy \geq 1 \\ |s-1| & 0 \leq sy \leq 1 \\ 1 + |s| & sy \leq 0 \end{cases}$$

If $y = 1$, then

$$\ell_{margin,1}(s, y) = \begin{cases} 0 & s \geq 1 \\ |s-1| & 0 \leq s \leq 1 \\ 1+|s| & s \leq 0 \end{cases}$$

and if $y = -1$, then

$$\ell_{margin,1}(s, y) = \begin{cases} 0 & -s \geq 1 \\ |s-1| & 0 \leq -s \leq 1 \\ 1+|s| & -s \leq 0 \end{cases} = \begin{cases} 0 & s \leq -1 \\ |s+1| & 0 \geq s \geq -1 \\ 1+|s| & s \geq 0 \end{cases}$$

Then we have the two plots,

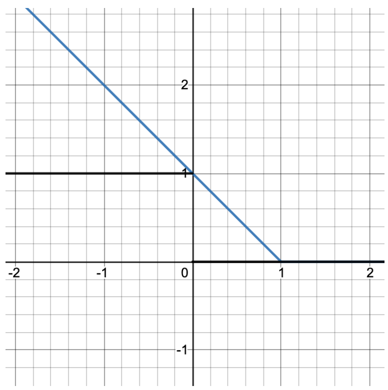

Figure 5: ℓ_{margin} for $y = 1$

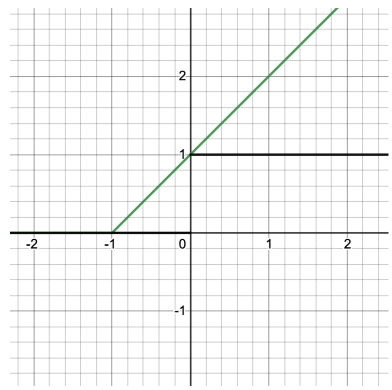

Figure 6: ℓ_{margin} for $y = -1$

which combined, recovers the standard marginal loss,

$$\ell_{margin,1}(s, y) = \begin{cases} \max(0, 1-s) & y = 1 \\ \max(0, 1+s) & y = -1 \end{cases} \quad \square$$

ii) Generalizing the results of theorem 4.2, for threshold $t \geq 0$, we say that a pair (y, s) where $y \in \mathcal{Y}_{\pm}$ and $s \in \mathbb{R}$ is

- incorrect: $y \neq c(s)$
- false positive: $y = -1, s > 0$
- marginal positive: $y = 1, 0 \leq s \leq t$
- false negative: $y = 1, s < 0$
- marginal negative: $y = -1, -t \leq s \leq 0$
- marginal: $y = c(s)$ and $|s| \leq t$
- confident: $y = c(s)$ and $|s| \geq t$

Problem 4.5

For $y = 1$ and $t > 1$ we have $\ell_{margin,t}(s, 1) = 1 - s/t$ which gives us the following plot where t changes the function in terms of slope steepness(gets less and less steep as $t \to \infty$) ,

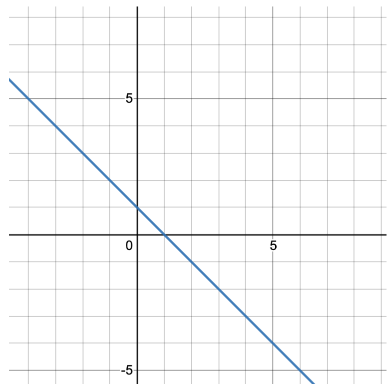

Figure 7: $\ell_{margin,1}(s, 1) = 1 - s/t\ ;\ t > 0$

To show symmetry, we have $\ell_{margin,t}(-s, -y) = \ell_{margin,t}(-s, -1) = 1 + (-s)/t = 1 - s/t = \ell_{margin,t}(s, y)$. So for $y = -1$ and $t > 0$ (modifying the slope steepness going towards less steep as $t \to \infty$) we find the following symmetric plot,

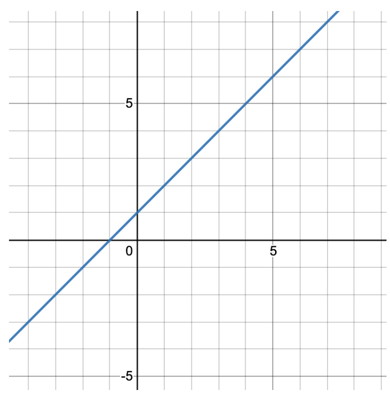

Figure 8: $\ell_{margin,1}(s, -1) = 1 + s/t\ ;\ t > 0$

Problem 4.6

For $y = 1$,

$$\frac{d}{dw}\ell_{margin,t}(s, y) = \begin{cases} 0 & sgn(s) = y \\ -1/t & o.w. \end{cases}$$

Note that since $s = x - w$, $ds/dw = -1$ so,

$$\begin{aligned} \frac{d}{dw}\hat{L}(s_w) &= \frac{1}{m}\sum_{i=1}^{m}\ell_{margin,t}(s, y)\frac{ds}{dw} \\ &= \frac{1}{m}\sum_{i=1}^{m}\begin{cases} 0 & sgn(s) = y \\ 1/t & o.w. \end{cases} \end{aligned}$$

and for $y = -1$,

$$\begin{aligned}\frac{d}{dw}\hat{L}(s_w) &= \frac{1}{m}\sum_{i=1}^{m}\ell_{margin,t}(s,y)\frac{ds}{dw}\\ &= \frac{1}{m}\sum_{i=1}^{m}\begin{cases}0 & sgn(s) = y\\ -1/t & o.w.\end{cases}\end{aligned}$$

We see that each term in the combined derivative is either zero or $\pm 1/t$ so we can separate the sum and minimize as follows,

$$\begin{aligned}&\sum_{y_i=1}^{m} 1/t + \sum_{y_i=-1}^{m} -1/t = 0\\ \implies &\sum_{y_i=1}^{m} 1/t = \sum_{y_i=-1}^{m} 1/t\\ \implies &\frac{1}{t}\sum_{y_i=1}^{m} 1 = \frac{1}{t}\sum_{y_i=-1}^{m} 1\\ \implies &\frac{1}{t}E_n = \frac{1}{t}E_p\\ \implies &E_n = E_p\end{aligned}$$

So the w^* is the threshold which $E_n = E_p$. ■

Problem 5.1

i) Since $p(r) = r/(r+1)$, we have

$$p(e^x) = \frac{e^x}{e^x+1} = \frac{e^x}{e^x+1}\cdot\frac{e^{-x}}{e^{-x}} = \frac{1}{1+e^{-x}} = \sigma(x) \quad \square$$

Now, since $r(p)$ is the inverse of $p(r)$, we take $r(p) = p/(p+1)$ and solve for p to define $r(p)$ as $p/(1-p)$. Note that

$$p(r(p)) = \frac{r(p)}{r(p)+1} = \frac{\frac{p}{1-p}}{\frac{p}{1-p}+1} = \frac{p}{1-p}\cdot\frac{1}{\frac{1}{1-p}} = \frac{p}{1-p}\frac{p-1}{1} = p.$$

which confirms our conclusions for the inverse. Then we have,

$$\log(r(p)) = \log\left(\frac{1}{1-p}\right) = \text{logit}(p).$$

ii) σ and logit are inverses if $\sigma \circ \text{logit}(p) = p$ and $\text{logit} \circ \sigma(x) = x$. From part (i), we have

$$\sigma(\text{logit}(p)) = \sigma(\log(r(p))) = \frac{1}{1+e^{-\log(\frac{1}{1-p})}} = \frac{1}{1+\frac{1-p}{p}} = \frac{p}{p+1-p} = p$$

and

$$\text{logit}(\sigma(x)) = \text{logit}(p(e^x)) = \log\Big(r(p(e^x))\Big) = \log(e^x) = x.$$

since $e^{\log(x)} = x$, $\log(e(x)) = x$, $p(r(p)) = p$ and $r(p(r)) = r$. ■

Problem 5.2

$$
\begin{aligned}
2\sigma(x) &= \frac{2}{1+e^{-x}} \\
&= \frac{2e^x}{e^x+1} \\
&= \frac{2e^x}{e^x+1} - \frac{e^x-1}{e^x+1} + \frac{e^x-1}{e^x+1} \\
&= \frac{2e^x}{e^x+1} - \frac{e^x-1}{e^x+1} + \tanh(x/2) \\
&= \frac{2e^x - e^x + 1}{e^x+1} + \tanh(x/2) \\
&= 1 + \tanh(x/2) \quad \square
\end{aligned}
$$

$$
\begin{aligned}
1-\sigma(x) &= 1 - \frac{1}{1+e^{-x}} \\
&= \frac{1+e^{-x}}{1+e^{-x}} - \frac{1}{1+e^{-x}} \\
&= \frac{e^{-x}}{1+e^{-x}} \\
&= \frac{1}{e^x+1} \\
&= \sigma(-x) \quad \square
\end{aligned}
$$

$$
\begin{aligned}
\sigma'(x) &= \frac{d}{dx}\frac{1}{1+e^{-x}} \\
&= \frac{-(-1\cdot e^{-x})}{(1+e^{-x})^2} \\
&= \frac{1\cdot e^{-x}}{(1+e^{-x})^2} \\
&= \left(\frac{1}{1+e^{-x}}\right)\left(\frac{e^{-x}}{1+e^{-x}}\right) \\
&= \left(\frac{1}{1+e^{-x}}\right)\left(\frac{e^{-x}+1-1}{1+e^{-x}}\right) \\
&= \left(\frac{1}{1+e^{-x}}\right)\left(\frac{e^{-x}+1}{1+e^{-x}} - \frac{1}{1+e^{-x}}\right) \\
&= \sigma(x)(1-\sigma(x)) \quad \square
\end{aligned}
$$

Problem 5.3

The loss $\ell_{score,log}(h,y)$ is given by

$$
\begin{aligned}
-\ell_{log}(\sigma(h),y) &= \begin{cases} \log(\sigma(h)) & y=1 \\ \log(1-\sigma(h)) & y=-1 \end{cases} \\
&= \begin{cases} \log\left(\frac{1}{1+e^{-h}}\right) & y=1 \\ \log\left(1-\frac{1}{1+e^{-h}}\right) & y=-1 \end{cases}
\end{aligned}
$$

This is an upper bound for the zero-one loss since we can find a constant C_{class} such that $\ell_{score,log}(h,y) \geq C_{class}\ell_{0-1}(c(h),y)$. Consider the case for $y = 1$, then choosing $C = 1$ we get the following plot,

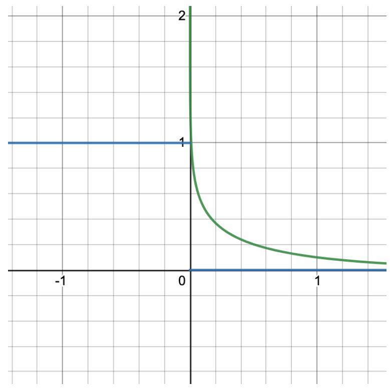

Figure 9: ℓ_{0-1} bounded by $\log(1/(1-e^{-h}))$

Now consider $y = -1$, choosing $C = 1$, we have the following plot,

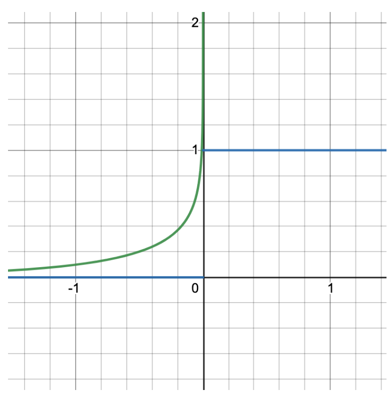

Figure 10: ℓ_{0-1} bounded by $\log(1-1/(1-e^{-h}))$

It follows from theorem 1.5 that for some $C > 0$ we have

$$\hat{L}_{0-1}(round(\sigma(h))) \leq \frac{1}{C}\hat{L}_{score,log}(h)$$

Problem 5.4

Differentiating we have,

$$\begin{aligned}
\frac{d}{dw}\hat{L}_{\log}(p_w) &= \frac{d}{dw}\left(\frac{1}{m}\sum_{j\in J^+} -\log(\sigma(x_j-w))\frac{d\sigma}{dw} + \frac{1}{m}\sum_{j\in J^-} -\log(1-\sigma(x_j-w))\frac{d\sigma}{dw}\right) \\
&= \frac{1}{m}\sum_{j\in J^+}\frac{\sigma(x_j-w)(1-\sigma(x_j-w))}{\sigma(x_j-w)} + \frac{1}{m}\sum_{j\in J^-} -\frac{\sigma(x_j-w)(1-\sigma(x_j-w))}{1-\sigma(x_j-w)} \\
&= \frac{1}{m}\sum_{j\in J^+} 1-\sigma(x_j-w) - \frac{1}{m}\sum_{j\in J^-}\sigma(x_j-w)
\end{aligned}$$

Then at a minimizer,

$$\begin{aligned}
&\sum_{j\in J^+}(1-\sigma(x_j-w)) = \sum_{j\in J^-}\sigma(x_j-w)\\
\implies &\sum_{j\in J^+}(1-p(x)) = \sum_{j\in J^-}p(x)\\
\implies &\sum_{j\in J^+}e(p,1) = \sum_{j\in J^-}e(p,-1) \quad \blacksquare
\end{aligned}$$

Problem 6.1

i) A function is convex iff $f(tx_1+(1-t)x_2) \leq tf(x_1)+(1-t)f(x_2)$ for $0 \leq t \leq 1$.

The quadratic loss, $\ell(p_i,y_i) = \frac{1}{2}(p_i-y_i)^2$, is convex since taking $\mathbf{x}_1 = (p_1,y_1)$ and $\mathbf{x_2} = (p_2,y_2)$, we have

$$\begin{aligned}
\ell(t(p_1,y_1)+(1-t)(p_2,y_2)) &= \ell(tp_1+(1-t)p_2, ty_1+(1-t)y_2)\\
&= \frac{1}{2}\left(tp_1+(1-t)p_2-ty_1-(1-t)y_2\right)^2\\
&= \frac{1}{2}\left(t(p_1-y_1)+(1-t)(p_2-y_2)\right)^2\\
&\leq \frac{1}{2}(t(p_1-y_1))^2+((1-t)(p_2-y_2))^2 \quad \text{(triangle inequality)}\\
&\leq \frac{1}{2}t((p_1-y_1))^2+(1-t)((p_2-y_2))^2 \quad \text{(because for } 0\leq t\leq 1,\ t^2\leq t)\\
&= t\ell(p_1,y_1)+(1-t)\ell(p_2,y_2) \quad \square
\end{aligned}$$

The logarithmic loss, $\ell(p_i,y_i) = -\log(p_i)$ for $y_i = 1$ and $\ell(p_i,y_i) = -\log(1-p_i)$ for $y_i = -1$, is convex since taking $\mathbf{x}_1 = (p_1,y_1)$ and $\mathbf{x_2} = (p_2,y_2)$ and using the fact that the log function is convex, we have
Case 1:

$$\begin{aligned}
\ell(t(p_1,y_1)+(1-t)(p_2,y_2)) &= \ell(tp_1+(1-t)p_2, ty_1+(1-t)y_2)\\
&= -\log(tp_1+(1-t)p_2)\\
&\leq -t\log(p_1)-(1-t)\log(p_2) \quad \text{(since log is convex)}\\
&= t\ell(p_1,y_1)+(1-t)\ell(p_2,y_2) \quad \square
\end{aligned}$$

Case 2:

$$\begin{aligned}
\ell(t(p_1,y_1)+(1-t)(p_2,y_2)) &= \ell(tp_1+(1-t)p_2, ty_1+(1-t)y_2)\\
&= -\log(1-(tp_1+(1-t)p_2))\\
&\leq -t\log(1-p_1)-(1-t)\log(1-p_2) \quad \text{(since log is convex)}\\
&= t\ell(p_1,y_1)+(1-t)\ell(p_2,y_2) \quad \square
\end{aligned}$$

ii) Consider $\hat{L}(p) = \frac{1}{m}\sum_{i=1}^m \ell(p,y_i) = \frac{1}{m}\sum_{y_i=1}^m \ell(p,1)+\frac{1}{m}\sum_{y_i=-1}^m \ell(p,-1) = q\ell(p,1)+(1-q)\ell(p,-1)$.

To minimize the quadratic loss, $\min_p q(1-p)^2+(1-q)p^2$, we have,

$$
\begin{aligned}
&\frac{d}{dp}q(1-p)^2 + (1-q)p^2 = 0 \\
&-2q(1-p) + 2p(1-q) = 0 \\
\implies \quad & q(1-p) = p(1-q) \\
\implies \quad & \frac{q}{1-q} = \frac{p}{1-p}
\end{aligned}
$$

which gives that p must equal q for the equality to hold and we have that the loss is proper. □

To minimize the logarithmic loss, $\min_p q\log(p) + (1-q)\log(1-p)$, we have,

$$
\begin{aligned}
&\frac{d}{dp}pq\log(p) + (1-q)\log(1-p) = 0 \\
&q(\frac{1}{p}) + (1-q)(\frac{-1}{1-p}) = 0 \\
\implies \quad & \frac{q}{p} = \frac{1-q}{1-p} \\
\implies \quad & \frac{q}{1-q} = \frac{p}{1-p}
\end{aligned}
$$

which gives $p = q$ and that the loss is proper. □

Problem 6.2

i) The spherical loss for $y = 1$, $\ell(p_i, y_i) = p_i/(p_i^2 + (1-p_i^2))^{1/2}$, is convex since taking $\mathbf{x_1} = (p_1, y_1)$ and $\mathbf{x_2} = (p_2, y_2)$, we have

$$
\begin{aligned}
\ell(t(p_1,y_1) + (1-t)(p_2,y_2)) &= \ell(tp_1 + (1-t)p_2, ty_1 + (1-t)y_2) \\
&= \frac{tp_1 + (1-t)p_2}{\sqrt{(tp_1 + (1-t)p_2)^2 + (1-(tp_1+(1-t)p_2)^2)}} \\
&= \frac{tp_1 + (1-t)p_2}{1} \\
&\left(\text{*since p is a probability, } p \leq 1, \text{ so } p^2 \leq 1 \text{ and } \sqrt{p_2^2 + (1-p_2^2)} \leq 1\right) \\
&\leq t\frac{p_1}{\sqrt{p_1^2 + (1-p_1^2)}} + (1-t)\frac{p_2}{\sqrt{p_2^2 + (1-p_2^2)}} \\
&= t\ell(p_1,y_1) + (1-t)\ell(p_2,y_2) \quad \square
\end{aligned}
$$

For $y = -1$, $\ell(p_i, y_i) = (1-p_i)/(p_i^2 + (1-p_i)^2)^{1/2}$, and we have,

$$
\begin{aligned}
\ell(t(p_1,y_1) + (1-t)(p_2,y_2)) &= \ell(tp_1 + (1-t)p_2, ty_1 + (1-t)y_2) \\
&= \frac{1 - tp_1 + (1-t)p_2}{\sqrt{(tp_1 + (1-t)p_2)^2 + (1-(tp_1+(1-t)p_2)^2)}} \\
&= \frac{1 - tp_1 + (1-t)p_2}{1} \\
&\left(\text{*since p is a probability, } \sqrt{p_2^2 + (1-p_2^2)} \leq 1\right) \\
&\leq t\frac{1-p_1}{\sqrt{p_1^2 + (1-p_1^2)}} + (1-t)\frac{1-p_2}{\sqrt{p_2^2 + (1-p_2^2)}} \\
&= t\ell(p_1,y_1) + (1-t)\ell(p_2,y_2) \quad \square
\end{aligned}
$$

ii) To minimize spherical loss $\min_p qp/(p^2 + (1-p^2))^{1/2} + (1-q)(1-p)/(p^2 + (1-p^2))^{1/2}$, we have,

$$
\begin{aligned}
& \frac{d}{dp} qp/(p^2 + (1-p^2))^{1/2} + (1-q)(1-p)/(p^2 + (1-p^2))^{1/2} = 0 \\
& \frac{q\sqrt{p^2 + (1-p)^2} - \frac{qp}{\sqrt{p^2+(1-p)^2}}}{p^2 + (1-p)^2} + \frac{(q-1)\sqrt{p^2 + (1-p)^2} - \frac{(1-q)(1-p)}{\sqrt{p^2+(1-p)^2}}}{p^2 + (1-p)^2} = 0 \\
& \frac{q(p^2 + (1-p)^2) - qp}{(p^2 + (1-p)^2)^2} + \frac{(q-1)(p^2 + (1-p)^2) - (1-q)(1-p)}{(p^2 + (1-p)^2)^2} = 0 \\
& q(p^2 + (1-p)^2 - p) + (q-1)(p^2 + (1-p)^2 - (1-p)) = 0 \\
\implies \quad & q(p^2 + (1-p)^2 - p) = -(q-1)(p^2 + (1-p)^2 - (1-p)) \\
\implies \quad & \frac{q}{1-q} = \frac{p^2 + (1-p)^2 - (1-p)}{p^2 + (1-p)^2 - p} \\
\implies \quad & \frac{q}{1-q} = \frac{2p^2 - p}{2p^2 - 3p + 1} \\
\implies \quad & \frac{q}{1-q} = \frac{p(2p-1)}{(2p-1)(p-1)} \\
\implies \quad & \frac{q}{1-q} = \frac{p}{p-1}
\end{aligned}
$$

which gives that $p = q$ for $p = 1$, $p = 0$ so the loss is proper. ■

Problem 6.3

i) The linear loss, $\ell(p_i, y_i) = |p_i - y_i|$, is convex since taking $\mathbf{x_1} = (p_1, y_1)$ and $\mathbf{x_2} = (p_2, y_2)$, we have

$$
\begin{aligned}
\ell(t(p_1, y_1) + (1-t)(p_2, y_2)) &= \ell(tp_1 + (1-t)p_2, ty_1 + (1-t)y_2) \\
&= |tp_1 + (1-t)p_2 - ty_1 - (1-t)y_2| \\
&= |t(p_1 - y_1) + (1-t)(p_2 - y_2)| \\
&\le t|p_1 - y_1| + (1-t)|p_2 - y_2| \quad \text{(triangle inequality)} \\
&= t\ell(p_1, y_1) + (1-t)\ell(p_2, y_2) \quad \square
\end{aligned}
$$

ii) To minimize linear loss, $\min_p q(1-p) + (1-q)p$, we have,

$$
\begin{aligned}
& \frac{d}{dp} q(1-p) + (1-q)p = 0 \\
& -q + 1 - q = 0 \\
\implies \quad & 2q = 1 \\
\implies \quad & q = 1/2
\end{aligned}
$$

But p must be 0 or 1, so $p \neq q$, and the loss is not proper. ■