A. Boosting-type Algorithm

1. Show that for all $u \in \mathbb{R}$ and integer $p > 1$, $1_{u \leq 0} \leq \Phi_p(-u)$ where $\Phi_p(u) = \max((1 + u)^p, 0)$. Show that $\Phi_p$ is convex and differentiable.

Solution: We first show that $1_{u \leq 0} \leq \Phi_p(-u)$ for all $u$. Observe that for $u > 0$, $1_{u > 0} = 0 \leq \Phi_p(-u)$ by definition of $\Phi_p$. For $u \leq 0$, $-u \geq 0$ and $\Phi_p(-u) = (1 - u)^p \geq 1 = 1_{u \leq 0}$, which proves the desired statement.

Now we show that $\Phi_p(u)$ is differentiable and convex. We consider two cases: $p$ is even and $p$ is odd. If $p$ is even then $\Phi_p(u) = (1 + u)^p$ for all $u$ since $(1 + u)^p \geq 0$ for all $u$. Therefore, $\Phi'_p(u) = p(1 + u)^{p-1}$ and $\Phi''_p(u) = p(p - 1)(1 + u)^{p-2}$. Moreover, $p(p - 1)(1 + u)^{p-2} > 0$ since $p - 2$ is even and $p - 1 > 0$. This shows that $\Phi_p(u)$ is differentiable and convex in this case.

Now if $p$ is odd then $\Phi_p(u) = 0$ for $u \in (-\infty, -1]$ and $\Phi_p(u) = (1 + u)^p$ for $u \in (-1, \infty)$. Therefore, $\Phi'_p(u) = 0$ on $(-\infty, -1)$ and $\Phi'_p(u) = p(1 + u)^{p-1}$ for $u \in (-1, \infty)$. To show that $\Phi_p$ is differentiable at $-1$ we consider left and right derivatives:

$$\lim_{u \uparrow -1} \frac{\Phi_p(u) - \Phi_p(-1)}{u + 1} = \lim_{u \uparrow -1} \frac{0}{u + 1} = 0$$

$$\lim_{u \downarrow -1} \frac{\Phi_p(u) - \Phi_p(-1)}{u + 1} = \lim_{u \downarrow -1} \frac{(1 + u)^p}{1 + u} = \lim_{u \downarrow -1} (1 + u)^{p-1} = 0$$

Similarly, we observe that $\Phi''_p(u) = 0$ on $(-\infty, -1)$ and $\Phi''_p(u) = p(1 + u)^{p-2}$ for $u \in (-1, \infty)$. Using the same arguments as for the first derivative and the fact that $p \geq 3$ since $p$ is odd we get

$$\lim_{u \uparrow -1} \frac{\Phi'_p(u) - \Phi'_p(-1)}{u + 1} = \lim_{u \uparrow -1} \frac{0}{u + 1} = 0$$

$$\lim_{u \downarrow -1} \frac{\Phi'_p(u) - \Phi'_p(-1)}{u + 1} = \lim_{u \downarrow -1} \frac{p(1 + u)^{p-1}}{1 + u} = \lim_{u \downarrow -1} p(1 + u)^{p-2} = 0$$
and hence $\Phi''_p(-1) = 0$. It follows that $\Phi''_p(u) \geq 0$ for all $u$ and $\Phi_p$ is convex.

2. Use $\Phi_p$ to derive a boosting-type algorithm using coordinate descent. You should give a full description of your algorithm, including the pseudocode, details for the choice of the step and direction, as well as a generalization bound.

Solution: We assume that we have access to $N$ weak learners $h_1, \ldots, h_N$ and the goal is to learn an ensemble hypothesis $g = \sum_{j=1}^N \alpha_j h_j$ and predict according to $\text{sgn}(g)$. Observe that

$$\frac{1}{m} \sum_{i=1}^m 1_{\text{sgn}(g(x_i)) \neq y_i} = \frac{1}{m} \sum_{i=1}^m 1_{y_i g(x_i) \leq 0} \leq \frac{1}{m} \sum_{i=1}^m \Phi_p(-y_i g(x_i))$$

by previous part of this question and our boosting-type algorithm that we provide consists of applying coordinate descent to this convex and differentiable objective. If $F(\alpha_t) = \frac{1}{m} \sum_{i=1}^m \Phi_p(-y_i f_t(x_i))$ and $f_t = \sum_{s=1}^t \alpha_s h_s$ is the solution after $t$ iterations, then at iteration $t + 1$ our algorithm picks the direction

$$h_k = \arg\min_{h_j \in \{j=1, \ldots, N\}} F'(\alpha_t + \eta e_j)\big|_{\eta=0}.$$ 

That is,

$$h_k = \arg\max_{h_j \in \{j=1, \ldots, N\}} \sum_{i=1}^m y_i h_j(x_i) (1 - y_i f_t(x_i))^{p-1}. \quad (1)$$

Once the direction is determined, the step size $\alpha_{t+1}$ is set by solving

$$F'(\alpha_t + \eta e_k) = 0$$

for $\eta$. This solution can be found using line search. The pseudocode for this algorithm is given in Algorithm 1. Note that $M_t(i)$s are used to avoid computing $y_i f_t(x_i)$ from scratch at every iteration $t$.

As an ensemble method this algorithm enjoys generalization bound of Corollary 6.1 from the textbook.
Algorithm 1 Boosting-type Algorithm.

**Inputs:** sample \(((x_1, y_1), \ldots, (x_m, y_m))\).

for \(i = 1\) to \(m\) do
\[
M_1(i) \leftarrow 0
\]
end for

for \(t = 1\) to \(T\) do
\[
h_t \leftarrow \text{solution of (1)}
\]
\[
\alpha_t \leftarrow \text{solution of (2)}
\]
for \(i = 1\) to \(m\) do
\[
M_{t+1}(i) \leftarrow M_t(i) + y_i \alpha_t h_t(x_i)
\]
end for
end for

\[
g \leftarrow \sum_{t=1}^{T} \alpha_t h_t
\]
return: \(h = \text{sgn}(g)\).

B. \(L_2\)-Regularized Maxent

This problem studies \(L_2\)-regularized Maxent. We will use the notation introduced in class and will denote by \(J_S\) the dual objective function minimized given a sample \(S\):

\[
J_S(w) = \frac{\lambda}{2} \|w\|_2^2 + \mathbb{E}_{x \sim S} \left[ -\log p_w(x) \right],
\]

where \(\lambda > 0\) is a regularization parameter. We will assume that the feature vector is bounded: \(\|\Phi(x)\|_2 \leq r\) for all \(x \in X\), for some \(r > 0\).

1. Use McDiarmid’s inequality to prove that for any \(\delta > 0\), with probability at least \(1 - \delta\), the following inequality holds:

\[
\mathbb{E}_{x \sim \mathcal{D}}[\Phi(x)] - \mathbb{E}_{x \sim S}[\Phi(x)]_2 \leq \sqrt{\frac{2r^2}{m} \left( 1 + \log \sqrt{\frac{1}{\delta}} \right)}.
\]

**Solution:**

For any sample \(S\), define \(\Gamma(S) = \|E_{x \sim \mathcal{D}}[\Phi(x)] - E_{x \sim S}[\Phi(x)]\|_2\). Let \(S'\) be a sample differing from \(S\) by one point, say \(x_m\) in \(S\) and \(x'_m\) in \(S'\), then, we can write

\[
|\Gamma(S') - \Gamma(S)| \leq \left\| E_{x \sim S'}[\Phi(x)] - E_{x \sim S}[\Phi(x)] \right\|_2 \leq \frac{1}{m} \left\| \Phi(x'_m) - \Phi(x_m) \right\|_2 \leq \frac{2r}{m}.
\]
Thus, for any $\delta > 0$, with probability at least $1 - \delta$,

$$\Gamma(S) \leq E_{S \sim D^m}[\Gamma(S)] + \sqrt{\frac{2r^2}{m} \log \frac{1}{\delta}}.$$  

Recall that $E_{x \sim S}[\Phi(x)] = \frac{1}{m} \sum_{i=1}^{m} \Phi(x_i)$, and denote $X_i = \frac{1}{m} \left[ E_{x \in D} \Phi(x) - \Phi(x_i) \right]$, so that $\sum_{i=1}^{m} X_i = E_{x \sim D}[\Phi(x)] - E_{x \sim S}[\Phi(x)]$.

Then, by Jensen’s inequality,

$$E_{S \sim D^m}[^{\Gamma(S)}] = E_{S \sim D^m}[^{\left\| E_{x \sim D^m}^{\Phi(x)} - E_{x \sim S^m}^{\Phi(x)} \right\|^2}] \leq \sqrt{E_{S \sim D^m}[^{\left\| \frac{1}{m^2} \sum_{i,j=1}^{m} \Phi(x_i) \cdot \Phi(x_j) \right\|^2}]} = \sqrt{\frac{1}{m} E_{i,1}[^{\left\| \Phi(x_1) \right\|^2}]} \text{ (for } i \neq j, \ E_{i,1}[\Phi(x_i) \cdot \Phi(x_j)] = E_{i,1}[\Phi(x_i)] \cdot E_{i,1}[\Phi(x_j)] = 0)$$

$$= \sqrt{\frac{1}{m} E_{i,1}[\left\| \Phi(x_1) \right\|^2]} \text{ (} x_i \text{ drawn i.i.d.)}$$

$$= \sqrt{\frac{E_{i,1}[\left\| \Phi(x_1) \right\|^2] + E_{i,1}[\left\| \Phi(x_2) \right\|^2]}{2m}}$$

$$= \sqrt{\frac{E_{i,1}[\left\| \Phi(x_1) - \Phi(x_2) \right\|^2]}{2m} \text{ (} \Phi(x_1) \cdot \Phi(x_2) = E_{i,1}[\Phi(x_1)] \cdot E_{i,1}[\Phi(x_2)] = 0)$$

$$= \sqrt{\frac{(2r)^2}{2m}} = \sqrt{\frac{2r^2}{m}}$$

2. Let $\hat{w}$ be the $L_2$-regularized maxent solution for a sample $S$ and $w_D$ the solution for an infinite sample:

$$\hat{w} = \arg\min_{w \in \mathbb{R}^N} J_S(w) \text{ and } w_D = \arg\min_{w \in \mathbb{R}^N} J_D(w).$$

where $J_D(w) = \frac{1}{2} \|w\|_2^2 + E_{x \sim D}[ - \log p_w(x) ]$. Use the definition of $\hat{w}$ and $w_D$ as minimizers (use gradients) to prove that the following inequality holds:

$$\|\hat{w} - w_D\|_2 \leq \frac{\left\| E_{x \sim S}[\Phi(x)] - E_{x \sim D}[\Phi(x)] \right\|_2}{\lambda}.$$
Solution:

Define function $Q$ for all $w$ by $Q(w) = \log Z = \log \left( \sum_x \exp(w \cdot \Phi(x)) \right)$. $Q$ is convex as a composition of the log-sum function with an affine function and we can write for any $w$:

$$J_S(w) = \frac{\lambda}{2} \|w\|^2 - w \cdot \mathbb{E}_{x \sim S} [\Phi(x)] + Q(w)$$

$$J_D(w) = \frac{\lambda}{2} \|w\|^2 - w \cdot \mathbb{E}_{x \sim D} [\Phi(x)] + Q(w).$$

Since the gradient of the objective function is zero at the minimum, we can write

$$\nabla J_S(\hat{w}) = 0 = \lambda \hat{w} - \mathbb{E}_{x \sim S} [\Phi(x)] + \nabla Q(\hat{w})$$

$$\nabla J_D(w_D) = 0 = \lambda w_D - \mathbb{E}_{x \sim D} [\Phi(x)] + \nabla Q(w_D).$$

Taking the difference yields:

$$\lambda (w_D - \hat{w}) = \mathbb{E}_{x \sim D} [\Phi(x)] - \mathbb{E}_{x \sim S} [\Phi(x)] + \nabla Q(\hat{w}) - \nabla Q(w_D).$$

Taking the inner product of each side with $w_D - \hat{w}$ gives:

$$\lambda \|w_D - \hat{w}\|^2 = (w_D - \hat{w}) \cdot \left[ \mathbb{E}_{x \sim D} [\Phi(x)] - \mathbb{E}_{x \sim S} [\Phi(x)] \right]$$

$$+ (\nabla Q(\hat{w}) - \nabla Q(w_D)) \cdot (w_D - \hat{w})$$

$$\leq (w_D - \hat{w}) \cdot \left[ \mathbb{E}_{x \sim D} [\Phi(x)] - \mathbb{E}_{x \sim S} [\Phi(x)] \right]$$

$$\leq \|w_D - \hat{w}\|_2 \left\| \mathbb{E}_{x \sim D} [\Phi(x)] - \mathbb{E}_{x \sim S} [\Phi(x)] \right\|_2.$$ (Cauchy-Schwarz ineq.)

where we used $(\nabla Q(\hat{w}) - \nabla Q(w_D)) \cdot (w_D - \hat{w}) \leq 0$, which holds by the convexity of $Q$.

3. For any $w$ and any distribution $Q$ define $\mathcal{L}_Q(w)$ by $\mathcal{L}_Q(w) = \mathbb{E}_{x \sim Q} [-\log p_w(x)]$. Show that

$$\mathcal{L}_D(\hat{w}) - \mathcal{L}_D(w_D) \leq (\hat{w} - w_D) \cdot \left[ \mathbb{E}_{x \sim S} [\Phi(x)] - \mathbb{E}_{x \sim D} [\Phi(x)] \right] + \frac{\lambda}{2} \|w_D\|^2 - \frac{\lambda}{2} \|\hat{w}\|^2.$$

Solution:
5. Conclude by proving that for any $\delta > 0$, with probability at least $1 - \delta$, the following inequality holds:

$$L_D(\hat{w}) \leq \inf_{w \in \mathbb{R}^N} L_D(w) + \frac{\lambda}{2} \|w\|^2 + \frac{2\tau^2}{\lambda m} \left(1 + \sqrt{\log \frac{1}{\delta}}\right)^2.$$ 

Solution:

This follows immediately the application of the inequality derived in 1).
C. Randomized Halving

In class, we showed that, in the realizable scenario (at least one expert is always correct), the number of mistakes made by Halving is upper bounded by $\log_2 N$. Here, we consider for the same realizable scenario a randomized version of Halving defined as follows.

As for Halving, let $H_t$ denote the set of remaining experts at the beginning of round $t$, with $H_1 = H$ the full set of $N$ experts. At each round, let $r_t$ be the fraction of experts in $H_t$ predicting 1. Then, the prediction $\hat{y}_t$ made by the algorithm is 1 with probability

$$p_t = \left[\frac{1}{2} \log_2 \frac{1}{1 - r_t}\right]_{1 - r_t \leq \frac{3}{4}} + 1_{r_t > \frac{3}{4}},$$

0 with probability $1 - p_t$. The true label $y_t$ is then received and $H_{t+1}$ is derived from $H_t$ by removing all experts who made a mistake.

1. Write the pseudocode of the algorithm.

2. Define the potential function $\Phi_t = \log_2 |H_t|$. Let $\mu_t = 1_{y_t \neq \hat{y}_t}$, prove that for all $t \geq 1$, $E[\mu_t] \leq \frac{\Phi_t - \Phi_{t+1}}{2}$.

Solution:

If $y_t = 0$, then $E[\mu_t] = p_t$. In round $t$, $r_t|H_t|$ experts make a mistake and are removed. Thus, $|H_{t+1}| = (1 - r_t)|H_t|$ and we can write

$$\frac{1}{2}(\Phi_t - \Phi_{t+1}) = \frac{1}{2} \log_2 \frac{|H_t|}{|H_{t+1}|} = \frac{1}{2} \log_2 \frac{1}{1 - r_t} \geq \min \left(\frac{1}{2} \log_2 \frac{1}{1 - r_t}, 1\right).$$

Observe that for $r_t > \frac{3}{4}$, $\frac{1}{2} \log_2 \frac{1}{1 - r_t} > \frac{1}{2} \log_2 \frac{1}{3/4} = \frac{1}{2} \log_2 4 = 1$.

Thus, $\min \left(\frac{1}{2} \log_2 \frac{1}{1 - r_t}, 1\right) = p_t$ and $\frac{1}{2}(\Phi_t - \Phi_{t+1}) \geq p_t$.

If $y_t = 1$, then $E[\mu_t] = 1 - p_t$. In round $t$, $(1 - r_t)|H_t|$ experts make a mistake and are removed. Thus, $|H_{t+1}| = r_t|H_t|$ and we can write

$$\frac{1}{2}(\Phi_t - \Phi_{t+1}) = \frac{1}{2} \log_2 \frac{1}{r_t} = -\frac{1}{2} \log_2 r_t = 1 - \frac{1}{2} \log_2 (4r_t).$$

Thus, for $r_t > \frac{3}{4}$, $\frac{1}{2}(\Phi_t - \Phi_{t+1}) > 1 - \frac{1}{2} \log_2 (3) > 0 = 1 - 1 = 1 - r_t$.

For $r_t \leq \frac{3}{4}$, using the fact that $x(x - 1) \leq 1/4$ for all $x \in [0, 1]$, we can
write

\[ 1 - \frac{1}{2} \log_2(4r_t) = 1 - \frac{1}{2} \log_2 \left( \frac{4r_t(1-r_t)}{1-r_t} \right) \]
\[ \geq 1 - \frac{1}{2} \log_2 \left( \frac{1}{1-r_t} \right) \]
\[ = 1 - p_t. \]

3. Show that the expected number of mistakes made by randomized Halving is at most \( \frac{1}{2} \log_2 N \).

\textit{Solution:}

In view of the previous questions,

\[ \sum_{t \geq 1} \mathbb{E}[\mu_t] \leq \frac{1}{2} \sum_{t \geq 1} \Phi_t - \Phi_{t+1} \leq \frac{1}{2} \Phi_1 = \frac{1}{2} \log_2 N. \]

4. (Bonus question) Prove that no randomized algorithm makes fewer than \( \lfloor \frac{1}{2} \log_2 N \rfloor \) mistakes, in expectation.