A. Boosting-type Algorithm

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

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.

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 $S$:

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

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

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

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

$$\tilde{w} = \arg\min_{w \in \mathbb{R}^N} J_S(w) \quad \text{and} \quad w_D = \arg\min_{w \in \mathbb{R}^N} J_D(w).$$
where \( J_D(w) = \frac{\lambda}{2} \|w\|_2^2 + E_{x \sim D} \left[ - \log \hat{p}_w(x) \right] \). Use the definition of \( \hat{w} \) and \( w_D \) as minimizers (use gradients) to prove that the following inequality holds:

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

3. For any \( w \) and any distribution \( Q \) define \( L_Q(w) \) by \( L_Q(w) = E_{x \sim Q}[-\log \hat{p}_w(x)] \). Show that

\[
L_D(\hat{w}) - L_D(w_D) \leq (\hat{w} - w_D) \left[ \frac{\lambda}{2} \|w_D\|_2^2 - \frac{\lambda}{2} \|\hat{w}\|_2^2 \right].
\]

4. Use that to show that the following inequality holds for any \( w \):

\[
L_D(\hat{w}) \leq \frac{1}{\lambda} \left\| E_{x \sim S}[\Phi(x)] - E_{x \sim D}[\Phi(x)] \right\|_2^2 + L_D(w) + \frac{\lambda}{2} \|w\|_2^2.
\]

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

\[
L_D(\hat{w}) \leq \inf_{w \in \mathbb{R}^N} L_D(w) + \frac{\lambda}{2} \|w\|_2^2 + \frac{2 \sigma^2}{\lambda m} \left( 1 + \sqrt{\log \frac{1}{\delta}} \right)^2.
\]

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 Having 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

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

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 $\mu_t = 1_{y_t \neq \tilde{y}_t}$, prove that for all $t \geq 1$, $E[\mu_t] \leq \frac{\Phi_t - \Phi_{t+1}}{2}$.

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

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