A. Kernels

Show that the following kernels $K$ are PDS:

1. For all integers $n > 0$, $K(x, y) = \sum_{i=1}^{N} \cos(n(x_i^2 - y_i^2))$ over $\mathbb{R}^N \times \mathbb{R}^N$.
2. $K(x, y) = \min(x, y) - xy$ over $[0, 1] \times [0, 1]$.
3. $\forall \sigma > 0, K(x, y) = e^{-\frac{\|x - y\|}{\sigma}}$ over $\mathbb{R}^N \times \mathbb{R}^N$.

B. Boosting

In class, we showed that AdaBoost can be viewed as coordinate descent applied to a convex upper bound on the empirical error. Here, we consider instead an algorithm seeking to minimize the empirical margin loss. For any $0 \leq \rho < 1$, using the same notation as in class, let $\hat{R}_\rho(f) = \frac{1}{m} \sum_{i=1}^{m} 1_{y_i f(x_i) \leq \rho}$ denote the empirical margin loss of a function $f$ of the form $f = \sum_{t=1}^{T} \alpha_t h_t$ for a labeled sample $S = ((x_1, y_1), \ldots, (x_m, y_m))$.

1. Show that $\hat{R}_\rho(f)$ can be upper bounded as follows:

$$\hat{R}_\rho(f) \leq \frac{1}{m} \sum_{i=1}^{m} \exp \left( -y_i \sum_{t=1}^{T} \alpha_t h_t(x_i) + \rho \sum_{t=1}^{T} \alpha_t \right).$$

2. For any $\rho > 0$, let $G_\rho$ be the objective function defined for all $\alpha \geq 0$ by

$$G_\rho(\alpha) = \frac{1}{m} \sum_{i=1}^{m} \exp \left( -y_i \sum_{j=1}^{N} \alpha_j h_j(x_i) + \rho \sum_{j=1}^{N} \alpha_j \right),$$

with $h_j \in H$ for all $j \in [1, N]$, with the notation used in class in the boosting lecture. Show that $G_\rho$ is convex and differentiable.
3. Derive a boosting-style algorithm $A_\rho$ by applying (maximum) coordinate descent to $G_\rho$. You should justify in detail the derivation of the algorithm, in particular the choice of the base classifier selected at each round and that of the step. Compare both to their counterparts in AdaBoost.

4. What is the equivalent of the weak learning assumption for $A_\rho$ (Hint: use non-negativity of the step value)?

5. Give the full pseudocode of the algorithm $A_\rho$. What can you say about the $A_0$ algorithm?

6. Implement the $A_\rho$ algorithm. The algorithm admits two parameters: the number of rounds $T$ and $\rho$. Experiment with this algorithm to tackle the same classification problem as the one described in the second homework assignment, with the same choice of training and test sets and the same cross-validation set-up. Use a grid search to determine the best values of $\rho$ (e.g., $\rho \in \{2^{-10}, 2^{-9}, \ldots, 2^{-1}\}$ and $T \in \{100, 200, 500, 1000\}$) and report the test error obtained. Compare your result with the one obtained by running AdaBoost using the same set-up. Also, compare these results with those obtained in the second homework assignment using SVMs.

7. For both the $A_\rho$ results and AdaBoost, plot the cumulative margins after 500 rounds of boosting, that is plot the fraction of training points with margin less than or equal to $\theta$ as a function of $\theta \in [0, 1]$.

8. Bound on $\hat{R}_\rho(f)$.

   (a) Prove the upper bound $\hat{R}_\rho(f) \leq \exp \left( \sum_{t=1}^{T} \alpha_t \rho \right) \prod_{t=1}^{T} Z_t$, where the normalization factors $Z_t$ are defined as in the case of AdaBoost in class (with $\alpha_t$ the step chosen by $A_\rho$ at round $t$).

   (b) Give the expression of $Z_t$ as a function of $\rho$ and $\epsilon_t$, where $\epsilon_t$ is the weighted error of the hypothesis found by $A_\rho$ at round $t$ (defined in the same way as for AdaBoost in class). Use that to prove the following upper bound

   \[
   \hat{R}_\rho(f) \leq \exp \left( - \sum_{t=1}^{T} D \left( \frac{1 - \rho}{2} \| \epsilon_t \right) \right),
   \]

   where $D(p\|q)$ denotes the binary relative entropy of $p$ and $q$: $D(p\|q) = p \log \frac{p}{q} + (1 - p) \log \frac{1 - p}{1 - q}$, for any $p, q \in [0, 1]$.
(c) Assume that for all $t \in [1, T]$, $\frac{1 - \rho}{2} - \epsilon_t > \gamma > 0$. Use the result of the previous question to show that

$$\hat{R}_\rho(f) \leq \exp \left(-2\gamma^2 T\right).$$

(Hint: you can use Pinsker’s inequality: $D(p\|q) \geq 2(p - q)^2$ for all $p, q \in [0, 1]$). Show that for $T > \frac{\log m}{2\gamma^2}$, all points have margin at least $\rho$. 