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 $\mathbb{R}^N \times \mathbb{R}^N$.

   Solution: Since $\cos(x - y) = \cos x \cos y + \sin x \sin y$, $K(x, y)$ can be written as the dot product of the vectors
   
   $\Phi(x) = \begin{bmatrix} \cos x \\ \sin x \end{bmatrix}$ and $\Phi(y) = \begin{bmatrix} \cos y \\ \sin y \end{bmatrix}$;
   thus it is PDS.

   This is a consequence of the fact that the kernel $K(x, y) = \cos(x - y)$ is PDS since $\sum_{ij} c_i c_j \cos(x_i^2 - x_j^2) = \sum_{ij} c_i c_j \cos(x_i' - x_j')$, with $x_i' = x_i^2$ for all $i$.

2. $K(x, y) = \min(x, y) - xy$ over $[0, 1] \times [0, 1]$.

   Solution: Let $\langle \cdot, \cdot \rangle$ denote the inner product over the set of all measurable functions on $[0, 1]$:
   
   $\langle f, g \rangle = \int_{0}^{1} f(t)g(t)dt$.

   Note that $k_1(x, y) = \int_{0}^{1} 1_{t \in [0, x]} 1_{t \in [0, y]}dt = \langle 1_{[0, x]}, 1_{[0, y]} \rangle$, and $k_2(x, y) = \int_{0}^{1} 1_{t \in [x, 1]} 1_{t \in [y, 1]}dt = \langle 1_{[x, 1]}, 1_{[y, 1]} \rangle$. Since they are defined as inner products, both $k_1$ and $k_2$ are PDS. Note that $k_1(x, y) = \min(x, y)$ and $k_2(x, y) = 1 - \max(x, y)$. Observe that $K = k_1 k_2$, thus $K$ is PDS as a product of two PDS kernels.

3. $\forall \sigma > 0, K(x, y) = e^{-\frac{|x-y|^2}{\sigma}}$ over $\mathbb{R}^N \times \mathbb{R}^N$.

   Solution: It suffices to show that $K$ is the normalized kernel associated to the kernel $K'$ defined by

   $\forall (x, y) \in \mathbb{R}^N \times \mathbb{R}^N, K'(x, y) = e^{\phi(x, y)}$
where \( \phi(x, y) = \frac{1}{\sigma}([\|x\| + \|y\| - \|x - y\|]) \), and to show that \( K' \) is PDS. For the first part, observe that

\[
\frac{K'(x, y)}{\sqrt{K'(x, x)K'(y, y)}} = e^{\phi(x, y) - \frac{1}{2} \phi(x, x) - \frac{1}{2} \phi(y, y)} = e^{-\frac{\|x - y\|}{\sigma}}.
\]

To show that \( K' \) is PDS, it suffices to show that \( \phi \) is PDS, since composition with a power series with non-negative coefficients (here \( \exp \)) preserve the PDS property. Now, for any \( c_1, \ldots, c_n \in \mathbb{R} \), let \( c_0 = -\sum_{i=1}^n c_i \), then, we can write

\[
\sum_{i,j=1}^n c_i c_j \phi(x_i, x_j) = \frac{1}{\sigma} \sum_{i,j=1}^n c_i c_j [\|x_i\| + \|x_j\| - \|x_i - x_j\|]
\]

\[
= \frac{1}{\sigma} \left[ -\sum_{i=1}^n c_0 c_i \|x_i\| + -\sum_{i=1}^n c_0 c_j \|x_j\| - \sum_{i,j=1}^n c_i c_j \|x_i - x_j\| \right]
\]

\[
= -\frac{1}{\sigma} \sum_{i,j=0}^n c_i c_j \|x_i - x_j\|,
\]

with \( x_0 = 0 \). Now, for any \( z \in \mathbb{R} \), the following equality holds:

\[
z^{\frac{1}{2}} = \frac{1}{2 \Gamma\left(\frac{1}{2}\right)} \int_{0}^{+\infty} \frac{1 - e^{-tz}}{t^{\frac{3}{2}}} \, dt.
\]

Thus,

\[
-\frac{1}{\sigma} \sum_{i,j=0}^n c_i c_j \|x_i - x_j\| = \frac{1}{2 \Gamma\left(\frac{1}{2}\right)} \int_{0}^{+\infty} -\frac{1}{\sigma} \sum_{i,j=0}^n c_i c_j \frac{1 - e^{-t\|x_i - x_j\|^2}}{t^{\frac{3}{2}}} \, dt
\]

\[
= \frac{1}{2 \Gamma\left(\frac{1}{2}\right)} \int_{0}^{+\infty} \frac{1}{\sigma} \sum_{i,j=0}^n c_i c_j e^{-t\|x_i - x_j\|^2} \, dt.
\]

Since a Gaussian kernel is PDS, the inequality \( \sum_{i,j=0}^n c_i c_j e^{-t\|x_i - x_j\|^2} \geq 0 \) holds and the right-hand side is non-negative. Thus, the inequality

\[
-\frac{1}{\sigma} \sum_{i,j=0}^n c_i c_j \|x_i - x_j\| \geq 0
\]

holds, which shows that \( \phi \) is PDS.

\[ \square \]

Alternatively, one can also apply the theorem on page 43 of the lecture slides on kernel methods to reduce the problem to showing that the norm \( G(x, y) = \|x - y\| \) is a NDS function. This can be shown through a direct application of the definition of NDS together with the representation of the norm given in the hint.
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:

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

**Solution:** The following shows how $\hat{R}_\rho(f)$ can be upper-bounded:

$$\hat{R}_\rho(f) = \frac{1}{m} \sum_{i=1}^{m} 1_{y_i f(x_i) \leq \rho} = \frac{1}{m} \sum_{i=1}^{m} 1_{y_i \sum_{t=1}^{T} \alpha_t h_t(x_i) - \rho \sum_{t=1}^{T} \alpha_t \leq 0} \leq \frac{1}{m} \sum_{i=1}^{m} \exp \left( -y_i \sum_{t=1}^{T} \alpha_t h_t(x_i) + \rho \sum_{t=1}^{T} \alpha_t \right).$$

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.

**Solution:** Since $\exp$ is convex and that composition with an affine function (of $\alpha$) preserves convexity, each term of the objective is a convex function and $G_\rho$ is convex as a sum of convex functions. The differentiability follows directly that of $\exp$.

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.

**Solution:** By definition of \( G_\rho \), for any direction \( e_t \) we can write

\[
G_\rho (\alpha_{t-1} + \eta e_t) = \frac{1}{m} \sum_{i=1}^{m} \exp \left( -y_i \sum_{s=1}^{t-1} \alpha_s h_s(x_i) + \rho \sum_{s=1}^{t-1} \alpha_s \right) e^{-y_i \eta h_t(x_i) + \rho \eta}.
\]

Let \( D_1 \) be the uniform distribution, that is \( D_1(i) = \frac{1}{m} \) for all \( i \in [1, m] \) and for any \( t \in [2, T] \), define \( D_t \) by

\[
D_t(i) = \frac{D_{t-1}(i) \exp(-y_i \alpha_{t-1} h_{t-1}(x_i))}{Z_{t-1}},
\]

with \( Z_{t-1} = \sum_{i=1}^{m} D_{t-1}(i) \exp(-y_i \alpha_{t-1} h_{t-1}(x_i)) \). Observe that \( D_t(i) = \exp \left( -y_i \sum_{s=1}^{t-1} \alpha_s h_s(x_i) \right) \frac{m}{m \prod_{s=1}^{t-1} Z_t} \). Thus,

\[
\left. \frac{dG_\rho (\alpha_{t-1} + \eta e_t)}{d\eta} \right|_0 = \frac{1}{m} \sum_{i=1}^{m} [-y_i h_t(x_i) + \rho] \exp \left( -y_i \sum_{s=1}^{t-1} \alpha_s h_s(x_i) + \rho \sum_{s=1}^{t-1} \alpha_s \right)
\]

\[
= \sum_{i=1}^{m} [-y_i h_t(x_i) + \rho] D_t(i) \prod_{s=1}^{t-1} Z_s \rho^{-\sum_{s=1}^{t-1} \alpha_s}
\]

\[
= \sum_{i=1}^{m} -y_i h_t(x_i) D_t(i) + \rho \prod_{s=1}^{t-1} Z_s \rho^{-\sum_{s=1}^{t-1} \alpha_s}
\]

\[
= (-1 - \epsilon_t) + \epsilon_t \prod_{s=1}^{t-1} Z_s \rho^{-\sum_{s=1}^{t-1} \alpha_s}
\]

\[
= (2\epsilon_t - 1 + \rho) \prod_{s=1}^{t-1} Z_s \rho^{-\sum_{s=1}^{t-1} \alpha_s},
\]

where \( \epsilon_t = \sum_{i=1}^{m} D_t(i) 1_{y_i h_t(x_i) > 0} \). The direction selected is the one minimizing \( (2\epsilon_t - 1 + \rho) \) that is \( \epsilon_t \). Thus, the algorithm selects at each round the base classifier with the smallest weighted error, as in the case of AdaBoost.
To determine the step \( \eta \) selected at round \( t \), we solve the following equation

\[
\frac{dG_{\rho}(\alpha_{t-1} + \eta e_t)}{d\eta} = 0
\]

\[
\Leftrightarrow \sum_{i=1}^{m} D_t(i) Z_t e^{-y_i h_t(x_i) + \rho \eta [-y_i h_t(x_i) + \rho]} = 0
\]

\[
\Leftrightarrow \sum_{y_i h_t(x_i) > 0} D_t(i) Z_t e^{\eta(-1+\rho)}(-1 + \rho) + \sum_{y_i h_t(x_i) < 0} D_t(i) Z_t e^{\eta(1+\rho)}(1 + \rho) = 0
\]

\[
\Leftrightarrow (1 - \epsilon_t) e^{\eta(-1+\rho)}(-1 + \rho) + \epsilon_t e^{\eta(1+\rho)}(1 + \rho) = 0
\]

\[
\Leftrightarrow e^{2\eta} = \frac{(1 - \rho)(1 - \epsilon_t)}{(1 + \rho)} \epsilon_t
\]

\[
\Leftrightarrow \eta = \frac{1}{2} \log \frac{(1 - \epsilon_t)}{\epsilon_t} - \frac{1}{2} \log \frac{(1 + \rho)}{(1 - \rho)}.
\]

Thus, the algorithm differs from AdaBoost only by the choice of the step, which it chooses more conservatively: the step size is smaller than that of AdaBoost by the additive constant \( \frac{1}{2} \log \frac{(1+\rho)}{(1-\rho)} \).

\hfill \square

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

**Solution:** For the coordinate descent algorithm to make progress at each round, the step size selected along the descent direction must be non-negative, that is

\[
\frac{(1 - \rho)(1 - \epsilon_t)}{(1 + \rho)} \epsilon_t > 1 \Leftrightarrow (1 - \rho)(1 - \epsilon_t) > \rho \epsilon_t + \epsilon_t
\]

\[
\Leftrightarrow \epsilon_t < \frac{1 - \rho}{2}.
\]

Thus, the error of the base classifier chosen must be at least \( \rho/2 \) better than one half.

\hfill \square

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

**Solution:** The normalization factor \( Z_t \) can be expressed in terms of \( \epsilon_t \) and \( \rho \)
\( A_\rho(S = ((x_1, y_1), \ldots, (x_m, y_m))) \)

1. **for** \( i \leftarrow 1 \) **to** \( m \) **do**
2. \( D_1(i) \leftarrow \frac{1}{m} \)
3. **for** \( t \leftarrow 1 \) **to** \( T \) **do**
4. \( h_t \leftarrow \) base classifier in \( H \) with small error \( \epsilon_t = \mathbb{P}_{i \sim D_t}[h_t(x_i) \neq y_i] \)
5. \( \alpha_t \leftarrow \frac{1}{2} \log \frac{1-\epsilon_t}{\epsilon_t} - \frac{1}{2} \log \frac{1+\rho}{1-\rho} \)
6. \( Z_t \leftarrow 2^\left[\frac{\epsilon_t(1-\epsilon_t)}{1-\rho^2}\right] \) \( \triangleright \) normalization factor
7. **for** \( i \leftarrow 1 \) **to** \( m \) **do**
8. \( D_{t+1}(i) \leftarrow \frac{D_t(i) \exp(-\alpha_t y_i h_t(x_i))}{Z_t} \)
9. \( f \leftarrow \sum_{t=1}^T \alpha_t h_t \)
10. **return** \( h = \text{sgn}(f) \)

Figure 1: Algorithm \( A_\rho \) for \( H \subseteq \{-1, +1\}^X \).

using its definition:

\[
Z_t = \sum_{i=1}^m D_t(i) \exp(-y_i \alpha_t h_t(x_i))
= e^{-\alpha_t(1 - \epsilon_t)} + e^{\alpha_t \epsilon_t}
= \sqrt{\frac{1+\rho}{1-\rho}(1-\epsilon_t)\epsilon_t} + \sqrt{\frac{1-\rho}{1+\rho}(1-\epsilon_t)\epsilon_t}
= \sqrt{\epsilon_t(1-\epsilon_t) \left[ \frac{1+\rho}{1-\rho} + \frac{1-\rho}{1+\rho} \right]}
= \sqrt{\epsilon_t(1-\epsilon_t) \left[ \frac{2}{\sqrt{1-\rho^2}} \right]}
= 2\sqrt{\frac{\epsilon_t(1-\epsilon_t)}{1-\rho^2}}.
\]

The pseudocode of the algorithm is then given in Figure 1. \( A_0 \) coincides exactly with AdaBoost.

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.

**Solution:** The following figures show the cross-validation results within one standard deviation for various values of $\rho$ and numbers of boosting rounds. Note that $\rho = 0$ corresponds to AdaBoost.

The lowest cross-validation error (0.057) occurs for margin $\rho = 2^{-10}$ and 1000 boosting rounds. The following table shows the training/test errors for $A_{\rho=2^{-10}}$ and Adaboost (with $T = 1000$) and SVM from last homework.

<table>
<thead>
<tr>
<th>Algorithm</th>
<th>$A_{\rho=2^{-10}}$</th>
<th>AdaBoost</th>
<th>SVM</th>
</tr>
</thead>
<tbody>
<tr>
<td>10-CV Error</td>
<td>0.0567</td>
<td>0.0573</td>
<td>0.0673</td>
</tr>
<tr>
<td>Test Error</td>
<td>0.0531</td>
<td>0.05</td>
<td>0.06</td>
</tr>
</tbody>
</table>
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 $\theta \in [0, 1]$.

Solution: In the graph below, x-axis is $\theta$ and y-axis is the fraction of points with margin less than $\theta$. Different colors correspond to different $A_\rho$: red line is $A_{\rho=2^{-5}}$ and blue line is AdaBoost($A_0$). As expected, for $\rho > 0$, $A_\rho$ achieves a larger margin for more points compared to AdaBoost. $A_{2^{-10}}$ (Orange line) is almost identical to AdaBoost.

![Cumulative margin graph](image)

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 $t$).

Solution:
In view of the definition of $D_t$ and the bound derived in question 1), we can write

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

$$
= \frac{1}{m} \sum_{i=1}^{m} \left( m \prod_{t=1}^{T} Z_t \right) D_t(i) \exp \left( \rho \sum_{t=1}^{T} \alpha_t \right)
$$

$$
= \exp \left( \rho \sum_{t=1}^{T} \alpha_t \right) \left( \prod_{t=1}^{T} Z_t \right).
$$

(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 $\mathcal{A}_\rho$ at round $t$ (defined in the same way as for AdaBoost in class). Use that to prove the following upper bound

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

where $D(p\parallel q)$ denotes the binary relative entropy of $p$ and $q$: $D(p\parallel q) = p \log \frac{p}{q} + (1-p) \log \frac{1-p}{1-q}$, for any $p, q \in [0, 1]$.

**Solution:** Define $u$ by $u = \frac{1-\rho}{1+\rho}$. The expression of $Z_t$ was already given above. Plugging in that expression in the bound of the previous question and using the expression of $\alpha_t$ gives

$$
\hat{R}_\rho(f) \leq \left( \prod_{t=1}^{T} e^{\alpha_t} \right)^\rho \left( \prod_{t=1}^{T} \sqrt{\frac{1-\epsilon_t}{\epsilon_t}} (u^{\frac{1}{2}} + u^{-\frac{1}{2}}) \right)
$$

$$
= \left( \sqrt{\frac{1-\rho}{1+\rho}} \right)^{\rho T} \left( \prod_{t=1}^{T} \sqrt{\frac{1-\epsilon_t}{\epsilon_t}} \right)^\rho \left( \prod_{t=1}^{T} \sqrt{\epsilon_t(1-\epsilon_t)(u^{\frac{1}{2}} + u^{-\frac{1}{2}})} \right)
$$

$$
= \left( u^{\frac{1-\rho}{2}} + u^{-\frac{1-\rho}{2}} \right)^T \prod_{t=1}^{T} \sqrt{\epsilon_t^{1-\rho}(1-\epsilon_t)^{1+\rho}}.
$$

9
Observe that
\[ u^{1+\rho} + u^{-1-\rho} = \left( \frac{1 - \rho}{1 + \rho} \right)^{1+\rho} + \left( \frac{1 + \rho}{1 - \rho} \right)^{-1-\rho} \]
\[ = \frac{(1 - \rho) + (1 + \rho)}{(1 + \rho)^{1+\rho} (1 - \rho)^{-1-\rho}} \]
\[ = \frac{2}{(1 + \rho)^{1+\rho} (1 - \rho)^{-1-\rho}} \]
\[ = \frac{1}{(1 + \rho)^{1+\rho} (1 - \rho)^{-1-\rho}}. \]

We also have
\[
\log \left[ \sqrt{\epsilon_t^{-\rho} (1 - \epsilon_t)^{1+\rho}} \right] \\
= \frac{1 - \rho}{2} \log(\epsilon_t) + \frac{1 + \rho}{2} \log(1 - \epsilon_t) \\
= -D\left( \frac{1 - \rho}{2} \| \epsilon_t \right) + \frac{1 - \rho}{2} \log \left( \frac{1 - \rho}{2} \right) + \frac{1 + \rho}{2} \log \left( \frac{1 + \rho}{2} \right) \\
= -D\left( \frac{1 - \rho}{2} \| \epsilon_t \right) + \log \left( \left( \frac{1 + \rho}{2} \right)^{1+\rho} \left( \frac{1 - \rho}{2} \right)^{-1-\rho} \right) .
\]

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

(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 \( \rho \).

Solution: By Pinsker’s inequality, we have \( D\left( \frac{1 - \rho}{2} \| \epsilon_t \right) \geq 2 \left[ \frac{1 - \rho}{2} - \epsilon_t \right]^2 \). Thus, we can write
\[ \hat{R}_\rho(f) \leq \exp \left( - 2\gamma^2 T \right) . \]
Thus, if the upper bound is less than $1/m$, then $\hat{R}_\rho(f) = 0$ and every training point has margin at least $\rho$. The inequality $\exp\left(-2\gamma^2 T\right) < 1/m$ is equivalent to $T > \frac{\log m}{2\gamma^2}$. □