Problem 1 - Estimating $h$ by cross-validation

For this problem, submit your code through the Assignments page link.

In this problem you will compute and plot a kernel density estimate of the corresponding densities $f$ and $g$ given below (you have calculated these densities in homework 4).

$$f(x) = \begin{cases} 
2x, & 0 \leq x \leq 1 \\
0, & \text{otherwise}
\end{cases} \quad (1)$$

$$g(x) = \begin{cases} 
4x, & 0 \leq x \leq 0.5 \\
4(1-x), & 0.5 \leq x \leq 1 \\
0, & \text{otherwise}
\end{cases} \quad (2)$$

a. Read in the training set $D$ consisting of $n = 1000$ samples from $F$ and validation set $D_v$ of $m = 300$ samples from files hw6-f-train.dat, hw6-f-valid.dat. Use a kernel of your choice and then find the optimal kernel width $h$ by cross-validation. For this, construct $f_h(x)$ the density estimated from $D$ with kernel width $h$. Then compute the likelihood $L_v(h)$ of the data in $D_v$ under $f_h$. Also compute $L(h)$, the likelihood of the training set $D$ under $f_h$. Repeat this for several values of $h$ and plot $L_v(h)$ and $L(h)$ as a function of $h$ on the same graph. (Suggested range of $h$: 0.001, 0.002, 0.005, 0.01, 0.02, 0.05, 0.1, 0.2, 0.5).

Let $h^*$ be the $h$ that maximizes $L_v(h)$. Make a plot of $f_{h^*}(x)$ (by, for instance, computing the $f_{h^*}(x)$ values on a grid $x = -0.5, -0.49, -0.48, \ldots, 1.49, 1.5$). Plot the true $f(x)$ on the same graph.

The homework you hand in should contain: the equation of the chosen kernel, the formula(s) you used for $f_h$ for the chosen kernel, the formula(s) you used to compute $L_v(h)$ and $L(h)$ and the required graphs. It is OK to replace likelihoods with log-likelihoods in the plots and equations.

b. Do the same for $G$ and $g$, reading data from the files hw6-g-train.dat, hw6-g-valid.dat.

c. Compare the optimal $h$’s and the quality of the plots in a, b. Which of the densities looks easier to approximate? Which of the optimal kernels widths is larger, the one used for $f$ or the one used for $g$? Can you suggest an explanation why?
Problem 2 - Discrete random variables

a A fair die is rolled; denote its outcome by \( X \in S_X = \{1, 2 \ldots 6\} \). The random variable \( Y \) is defined as

\[
Y(X) = \begin{cases} 
3 & \text{if } X \text{ odd} \\
X/2 & \text{if } X \text{ even}
\end{cases}
\]  

(3)

What is the outcome space \( S_Y \) of the random variable \( Y \)?

Evaluate the parameters \( \theta_j = P(Y = j) \equiv P_Y(j), j \in S_Y \) of the distribution \( P_Y \).

b. Denote by \( Z \) the boolean variable \( X \geq Y \) (i.e. \( Z = 1 \) if \( X \geq Y \) and 0 otherwise). Compute the probability distribution \( P_Z \). [Hint: note that as \( Y \) is a function of \( X \), \( Z \) can be computed based on \( X \) only.]

c. Two fair dice are thrown; denote by \( X = (X_1, X_2) \) the outcome of this experiment and by \( S_X \) its outcome space. Let \( U = |X_1 - X_2| \). What is the sample space \( S_U \) of the random variable \( U \)? Evaluate the parameters \( \phi_j = P(U = j) \equiv P_U(j), j \in S_U \) of the distribution \( P_U \).

d. Calculate the expectation of \( Y \).

e. Calculate the expectation of \( Z \).

f. Calculate the expectation of \( U \).