Problem 1. **Bin Packing.** The bin packing problem in computer science is the problem of finding an optimal packing of $n$ objects into boxes of unit size. The objects are assumed to have sizes $x_1, x_2, \ldots, x_n$ where $0 < x_i \leq 1$, and a packing is a partition of the indices $1, 2, \ldots, n$ into disjoint subsets $J_1, J_2, \ldots, J_k$ such that for each $J_i$,

$$\sum_{j \in J_i} x_j \leq 1.$$ 

An optimal packing is a packing that uses the smallest possible number $k$ of boxes. The problem is of interest to computer scientists because it is hard to find good algorithms that will take a given list $x_1, x_2, \ldots, x_n$ and output an optimal packing.

Assume now that the items to be packed are independent, identically distributed random variables $X_i \in (0, 1)$. Let $N_n$ be the smallest number of boxes needed to pack $X_1, X_2, \ldots, X_n$.

(A) Prove that $P\{ |N_n - EN_n| \geq x \} \leq 2 \exp\{-2x^2/n\}$. **CAUTION:** The factors of 2 in the bound might be off.

(B) The rest of the problem is devoted to the task of finding out something about the behavior of $EN_n$ for large $n$. First, prove that the sequence $EN_n$ is subadditive, that is, for any $n, m \geq 1$,

$$EN_{n+m} \leq EN_n + EN_m.$$ 

Second, prove that for any subadditive sequence $a_n$ of nonnegative real numbers,

$$\lim_{n \to \infty} \frac{a_n}{n} = \inf_{n \geq 1} \frac{a_n}{n} := \alpha \geq 0.$$ 

Thus, if $a_n = EN_n$ then $EN_n/n \to \alpha$.

(C) The downside of using subadditivity to establish the existence of limits is that in general the subadditivity lemma (i.e., part (B)) doesn’t provided much explicit information about the limit. In the bin-packing problem, however, the limit $\alpha$ can be evaluated for certain distributions. Find $\alpha$ for the uniform distribution on $[0, 1]$. **HINT:** The Glivenko-Cantelli theorem (see 381 notes) might be of some use here.

(D) Can you determine the value of $\alpha$ for other distributions on $(0, 1)$, for example, the distribution with density

$$f(x) = \theta_0 \quad \text{for } 0 < x \leq 1/2 \quad \text{and}$$

$$= \theta_1 \quad \text{for } 1/2 < x < 1 ?$$
Problem 2. Harris’ Inequality. This isn’t related to concentration per se but is a very useful inequality that can be proved using ideas similar to those that went into McDiarmid’s inequality. A bounded function \( F : \mathbb{R}^m \to [0, 1] \) is said to be monotone if it is non-decreasing with respect to the natural partial order \( \leq \) on \( \mathbb{R}^m \). This ordering is defined as follows: if \( x, y \in \mathbb{R}^m \) then \( x \) if and only if \( x_i \leq y_i \) for each coordinate \( i = 1, 2, \ldots, m \). Harris’ inequality states that for any bounded, monotone functions \( F, G \) and any collection \( X_1, X_2, \ldots, X_m \) of independent random variables,
\[
E(F(X)G(X)) \geq EF(X)EG(X).
\]

(A) Prove this for \( m = 1 \).

(B) Now prove the general case.

HINT: To deduce the general case from the special case, let \( X'_1, X'_2, \ldots, X'_m \) be an independent copy of \( X_1, X_2, \ldots, X_m \); compute the expectations by replacing the original variables \( X_i \) by the copies \( X'_i \) one at a time.

NOTE: The case \( m = 1 \) is in essence an equivalent form of the Neyman-Pearson Lemma of classical statistics.

Problem 3. Random Permutations. Let \( S_m \) be the permutation group on \( m \) letters. Here individual permutations are viewed as functions \( \pi : [m] \to [m] \), where \( [m] := \{1, 2, \ldots, m\} \).

Define a metric on \( S_m \) by declaring the distance between \( \pi \) and \( \sigma \) to be the minimum number of transpositions \( \tau_i \) needed to change \( \pi \) to \( \sigma \), i.e., such that
\[
\sigma = \pi \circ \tau_1 \circ \tau_2 \circ \cdots \circ \tau_k.
\]

(Note: A transposition is a permutation that fixes all but two letters in \( [m] \).)

(A) Let \( Y \) be chosen from the uniform distribution on the permutation group \( S_m \), and let \( F : S_m \to \mathbb{R} \) be Lip\(-1\) relative to the transposition metric. Prove that for suitable constants \( C < \infty \) and \( A > 0 \) not depending on \( m \),
\[
P \{ F(Y) - EF(Y) \geq t \} \leq Ce^{-At^2/m}.
\]

(B) Let \( G(\pi) \) be the number of pairs \((i, j)\) such that \( i < j \) and \( \pi(i) < \pi(j) \). Use the result of part (A) to deduce a concentration inequality for \( G(Y) \), and then deduce an appropriate “weak law of large numbers” for \( G(Y) \).

Bonus Problem

The Efron-Stein Inequality. Let \( X_1, X_2, \ldots, X_n \) be independent but not necessarily identically distributed random variables taking values in some abstract space \( \mathcal{X} \), and let \( g : \mathcal{X}^n \to \mathbb{R} \) be a real-valued function such that \( Z = g(X_1, X_2, \ldots, X_n) \) has finite second moment. To avoid unsightly expressions, let’s agree to use the shorthand notation \( E(Z \mid X^J) \) to denote the conditional expectation of \( Z \) given the random variables \( X_i \) where \( i \notin J \), and write \( Z_i = E(Z \mid X^{(i)}) \).

(a) Prove that \( \text{var}(Z) \leq \sum_{i=1}^n E(Z - Z_i)^2 \).
(b) Now suppose that \((X'_1, X'_2, \ldots, X'_n)\) is an independent copy of the random vector \(X_1, X_2, \ldots, X_n\). Define

\[ Y_i = g(X_1, \ldots, X_{i-1}, X'_i, X_{i+1}, \ldots, X_n). \]

Prove that

\[ \text{var}(Z) \leq \frac{1}{2} \sum_{i=1}^{n} E(Z - Y_i)^2. \]