Exercise Sheet 9 – Coupling, Balls into Bins and Poissonisation

Probability and Computing

Exercise 1 – Coupling of a Random Walk

Let $X_1, X_2, \ldots \sim \mathcal{U}(\{-1,1\})$ be independent random variables. For $n \in \mathbb{N}_0$, define $W_n := \sum_{i=1}^{n} X_i$. The sequence $(W_n)_{n\in\mathbb{N}_0}$ is called a random walk. We may also consider a shifted random walk $(V_n)_{n\in\mathbb{N}_0}$ defined by $V_n := W_n + 42$, which therefore has initial position $V_0 = 42$ instead of $W_0 = 0$. We aim to show that the choice of initial position typically does not matter in the long run. We will use without proof that the random walk visits every integer at least once with probability 1. In particular, $\lim_{n\to\infty} \Pr[\max\{W_1, \ldots, W_n\} < c] = 0$ for all $c \in \mathbb{N}$.

(i) Let $S_1, S_2, \ldots \subseteq \mathbb{Z}$ be arbitrary sets. Show that $\lim_{n\to\infty} |\Pr[W_n \in S_n] - \Pr[V_n \in S_n]| = 0$.
Hint: Construct a coupling $(W'_n, V'_n)_{n\in\mathbb{N}_0}$ of $(W_n)_{n\in\mathbb{N}_0}$ and $(V_n)_{n\in\mathbb{N}_0}$ such that $\lim_{n\to\infty} \Pr[W'_n = V'_n] = 1$.

(ii) Show that the result of part (i) does not hold in this form for a shift of 43 instead of 42.

Solution 1

(i) We take $(W'_n) = (W_n)$ and describe (V'_n) in natural language. Initially, (V'_n) behaves exactly oppositely to (W_n), i.e., uses the inverted increments $-X_1, -X_2, -X_3, \ldots$ and so on. Let $T = \min\{t \in \mathbb{N} \mid W_t = 21\}$. Then $W_T = 21$ and $V'_T = 42 - 21 = 21$, meaning the random walks meet at time T. From this point onward, (V'_n) behaves identically to (W_n), using the same increments $X_{T+1}, X_{T+2}, \ldots$.

It is evident that $(V'_n)_{n\in\mathbb{N}_0} \overset{\mathrm{d}}{=} (V_n)_{n\in\mathbb{N}_0}$, since the accumulated increments remain independent random variables uniformly distributed in $\mathcal{U}(\{-1,1\})$ (whether we add or subtract X_i is determined before observing the value of X_i). Thus, we have a valid coupling. In this coupling, the implication $W_n \neq V'_n \Rightarrow T \geq n$ holds. We now perform an auxiliary calculation for arbitrary random variables X, Y and arbitrary sets S.

$$\begin{aligned}
&|\Pr[X \in S] - \Pr[Y \in S]| \\
&= \left|\Pr[X \in S \wedge X \neq Y] + \Pr[X \in S \wedge X = Y] - \Pr[Y \in S \wedge X = Y] - \Pr[Y \in S \wedge X \neq Y]\right| \\
&= \left|\Pr[X \in S \wedge X \neq Y] - \Pr[Y \in S \wedge X \neq Y]\right| \\
&\leq \max\{\Pr[X \in S \wedge X \neq Y], \Pr[Y \in S \wedge X \neq Y]\} \leq \Pr[X \neq Y].
\end{aligned}$$

Applying this to $S = S_n$, $X = W_n$, and $Y = V_n'$, we obtain:

$$\begin{aligned}|\Pr[W_n \in S_n] - \Pr[V_n \in S_n]| &= |\Pr[W_n \in S_n] - \Pr[V_n' \in S_n]| \leq \Pr[W_n \neq V_n'] \\ &= \Pr[T > n] = \Pr[\max\{W_1, \dots, W_n\} < 21] \xrightarrow{n\to\infty} 0\end{aligned}$$

where the last step applies the hint for $c = 21$.

(ii) As defined, the random walk "forgets" its precise starting point but not the parity of this starting point. In other words, if we define $S_n := S := 2 \cdot \mathbb{Z}$, then random walks alternate between being in S and not in S. For a shift of 23, we would then have $|\Pr[W_n \in S_n] - \Pr[V_n \in S_n]| = 1$ for all $n \in \mathbb{N}$.

Exercise 2 – Coupling and Total Variation Distance

Let X and Y be two random variables taking values in $\mathbb{N}$. The total variation distance between X and Y (or their distributions) is defined as[1]

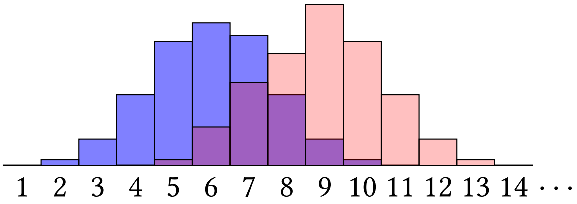
$$d(X,Y) = \frac{1}{2}\sum_{i\in\mathbb{N}} |\Pr[X = i] - \Pr[Y = i]|.$$

(i) Show: There exists a coupling (X', Y') of X and Y such that $\Pr[X' \neq Y'] = d(X,Y)$.

(ii) Show: No coupling (X', Y') of X and Y satisfies $\Pr[X' \neq Y'] < d(X,Y)$.

Solution 2

As preparation, consider a joint histogram of X (blue) and Y (red).



Let all bars have width 1. Denote by red, blue, and purple the sets of points of the respective colors, and by A_{red}, A_{blue}, and A_{purple} the corresponding areas. Since the bars represent distributions, we have $A_{\text{blue}} + A_{\text{purple}} = 1$ and $A_{\text{red}} + A_{\text{purple}} = 1$, implying $A_{\text{blue}} = A_{\text{red}}$. Both equal the total variation distance $d(X,Y)$. This is seen as follows:

$$\begin{aligned}d(X,Y) &= \tfrac{1}{2}\sum_{i\in\mathbb{N}} |\Pr[X = i] - \Pr[Y = i]| \\ &= \tfrac{1}{2}\Big(\sum_{\substack{i\in\mathbb{N}\\ \Pr[X=i]\geq\Pr[Y=i]}} (\Pr[X = i] - \Pr[Y = i]) + \sum_{\substack{i\in\mathbb{N}\\ \Pr[X=i]<\Pr[Y=i]}} (\Pr[Y = i] - \Pr[X = i])\Big) \\ &= \tfrac{1}{2}\big(A_{\text{blue}} + A_{\text{red}}\big) = \tfrac{1}{2}\big(A_{\text{blue}} + A_{\text{blue}}\big) = A_{\text{blue}}.\end{aligned}$$

[1] A general definition applicable also to continuous probability spaces can be found on Wikipedia.

(i) We sample a pair (P, Q) of points as follows:

- Sample $P \sim \mathcal{U}(\text{blue} \cup \text{purple})$.
- If $P \in \text{purple}$, set $Q = P$.
- Otherwise, sample $Q \sim \mathcal{U}(\text{red})$.

It should be clear that then $Q \sim \mathcal{U}(\text{red} \cup \text{purple})$. We now define X' as the index of the bar containing P and Y' as the index of the bar containing Q. It should then be clear that $X' \overset{\text{d}}{=} X$ and $Y' \overset{\text{d}}{=} Y$. The useful property we will use is $\Pr[X' = Y'] = \Pr[P = Q] = A_{\text{purple}}$. From this it follows as desired:

$$\Pr[X' \neq Y'] = 1 - A_{\text{purple}} = A_{\text{blue}} = d(X, Y).$$

(ii) Let $S = \{i \in \mathbb{N} \mid \Pr[X = i] > \Pr[Y = i]\}$. Let (X', Y') be any coupling of X and Y. Then:

$$\begin{aligned}\Pr[X' \neq Y'] &\geq \Pr[X' \in S \wedge Y' \notin S] = \Pr[X' \in S] - \Pr[X' \in S \wedge Y' \in S] \\ &\geq \Pr[X' \in S] - \Pr[Y' \in S] = \Pr[X \in S] - \Pr[Y \in S] \\ &= \sum_{i \in S} \Pr[X = i] - \Pr[Y = i] = A_{\text{blue}} = d(X, Y).\end{aligned}$$

Exercise 3 – Properties of the Poisson Distribution

Let $X \sim \text{Pois}(\lambda)$. Show:

(i) $\mathbb{E}[X] = \lambda$.

(ii) $\text{Var}(X) = \lambda$.

(iii) For $Y \sim \text{Pois}(\rho)$ independent of X, we have $X + Y \sim \text{Pois}(\lambda + \rho)$.

(iv) For $X' \sim \text{Bin}(X, p)$, we have $X' \sim \text{Pois}(\lambda p)$.
Note: Here, a two-stage random experiment is performed. The outcome X of the first stage serves as a parameter of the second stage.

Solution 3

In the following, we constantly use the definition of the exponential function, i.e., $e^t = \sum_{i=0}^{\infty} \frac{t^i}{i!}$.

(i) $\mathbb{E}[X] = \sum_{i=0}^{\infty} e^{-\lambda} \frac{\lambda^i}{i!} \cdot i = e^{-\lambda} \cdot \lambda \sum_{i=1}^{\infty} \frac{\lambda^{i-1}}{(i-1)!} = e^{-\lambda} \cdot \lambda \sum_{i=0}^{\infty} \frac{\lambda^i}{i!} = e^{-\lambda} \cdot \lambda \cdot e^{\lambda} = \lambda.$

(ii) We first compute the second *un*centered moment:

$$\begin{aligned}
\mathbb{E}[X^2] &= \sum_{i=0}^{\infty} e^{-\lambda}\frac{\lambda^i}{i!} \cdot i^2 = e^{-\lambda} \cdot \lambda \sum_{i=1}^{\infty} \frac{\lambda^{i-1}}{(i-1)!} \cdot i \\
&= e^{-\lambda} \cdot \lambda \Big(\sum_{i=1}^{\infty} \frac{\lambda^{i-1}}{(i-1)!} \cdot (i-1) + \sum_{i=1}^{\infty} \frac{\lambda^{i-1}}{(i-1)!} \Big) \\
&= e^{-\lambda} \cdot \lambda \Big(\lambda \sum_{i=2}^{\infty} \frac{\lambda^{i-2}}{(i-2)!} + \sum_{i=1}^{\infty} \frac{\lambda^{i-1}}{(i-1)!} \Big) \\
&= e^{-\lambda} \cdot \lambda \Big(\lambda \sum_{i=0}^{\infty} \frac{\lambda^{i}}{i!} + \sum_{i=0}^{\infty} \frac{\lambda^{i}}{i!} \Big) \\
&= e^{-\lambda} \cdot \lambda \Big(\lambda e^{\lambda} + e^{\lambda} \Big) = \lambda^2 + \lambda
\end{aligned}$$

Moreover, we know $\mathbb{E}[X]^2 = \lambda^2$. It follows that

$$\mathsf{Var}(X) = \mathbb{E}[X^2] - \mathbb{E}[X]^2 = \lambda^2 + \lambda - \lambda^2 = \lambda.$$

(iii) Let $k \in \mathbb{N}$. We consider all $k+1$ possibilities by which $X+Y$ can sum to k, and then apply the binomial theorem.

$$\begin{aligned}
\Pr[X+Y=k] &= \sum_{i=0}^{k} \Pr[X=i \wedge Y=k-i] = \sum_{i=0}^{k} \Pr[X=i]\Pr[Y=k-i] \\
&= \sum_{i=0}^{k} e^{-\lambda}\frac{\lambda^i}{i!} e^{-\rho}\frac{\rho^{k-i}}{(k-i)!} = e^{-(\lambda+\rho)}\frac{1}{k!}\sum_{i=0}^{k}\frac{k!}{i!(k-i)!}\lambda^i\rho^{k-i} \\
&= e^{-(\lambda+\rho)}\frac{1}{k!}\sum_{i=0}^{k}\binom{k}{i}\lambda^i\rho^{k-i} = e^{-(\lambda+\rho)}\frac{(\lambda+\rho)^k}{k!} = \Pr_{Z\sim\mathrm{Pois}(\lambda+\rho)}[Z=k].
\end{aligned}$$

(iv) Let $k \in \mathbb{N}$. For the final outcome to be k, it must have held that $X \geq k$. We consider all possibilities.

$$\begin{aligned}
\Pr[X'=k] &= \sum_{i\geq k} \Pr[X=i \wedge X'=k] = \sum_{i\geq k} \Pr[X=i] \cdot \Pr[X'=k \mid X=i] \\
&= \sum_{i\geq k} e^{-\lambda}\frac{\lambda^i}{i!} \cdot \binom{i}{k} p^k(1-p)^{i-k} = e^{-\lambda} \cdot \sum_{i\geq k}\frac{\lambda^i}{k!(i-k)!}p^k(1-p)^{i-k} \\
&= e^{-\lambda}\frac{(\lambda p)^k}{k!} \cdot \sum_{i\geq k}\frac{\lambda^{i-k}(1-p)^{i-k}}{(i-k)!} = e^{-\lambda}\frac{(\lambda p)^k}{k!} \cdot \sum_{i\geq 0}\frac{(\lambda(1-p))^i}{i!} \\
&= e^{-\lambda}\frac{(\lambda p)^k}{k!}e^{\lambda(1-p)} = e^{-\lambda p}\frac{(\lambda p)^k}{k!} = \Pr_{Z\sim\mathrm{Pois}(\lambda p)}[Z=k].
\end{aligned}$$

Exercise 4 – Poissonised Bloom Filters

We consider a Poisson model of Bloom filters, i.e., we assume that each position in the array independently appears as a hash value $\mathrm{Pois}(\alpha k)$-many times.

(i) We again choose $\alpha k = \ln 2$. How can we show that the fraction $\frac{Z}{m}$ of zeros is with high probability close to $\frac{1}{2}$?

(ii) How could this result be transferred to a non-Poissonised model?

Solution 4

(i) If $X \sim \mathrm{Pois}(\ln 2)$, then $\Pr[X = 0] = e^{-\ln 2} = \frac{1}{2}$. Since each position is now independently empty or non-empty, we have $Z \sim \mathrm{Bin}(m, \frac{1}{2})$. It follows that $\mathbb{E}[\frac{Z}{m}] = \frac{1}{2}$, and Chernoff bounds apply directly to Z.

(ii) The quantity $m - Z$ is a monotone function in the sense of the Poissonisation theorem from the lecture. Accordingly, the exact "nk balls into m bins" model can be sandwiched between two Poissonised models, as discussed.