Karlsruher Institut für Technologie

Stefan Walzer
ITI Prof. Sanders
Wintersemester 2023/2024

Übungsblatt 9 – Coupling, Balls into Bins and Poissonisation

Randomisierte Algorithmik

Aufgabe 1 – Coupling einer Irrfahrt

Seien $X_1, X_2, \ldots \sim \mathcal{U}(\{-1, 1\})$ unabhängige Zufallsvariablen. Für $n \in \mathbb{N}_0$ sei $W_n := \sum_{i=1}^{n} X_i$. Man nennt $(W_n)_{n\in\mathbb{N}_0}$ eine Irrfahrt (engl. *random walk*).

Wir können auch eine verschobene Irrfahrt $(V_n)_{n\in\mathbb{N}_0}$ betrachten mit $V_n := W_n + 42$, die also Startpunkt $V_0 = 42$ statt $W_0 = 0$ hat. Wir wollen zeigen, dass die Wahl des Startpunktes typischerweise langfristig keine Rolle spielt.

Wir wollen dabei ohne Beweis verwenden, dass die Irrfahrt mit Wahrscheinlichkeit 1 jede ganze Zahl mindestens einmal besucht. Insbesondere gilt $\lim_{n\to\infty} \Pr[\max\{W_1, \ldots, W_n\} < c] = 0$ für alle $c \in \mathbb{N}$.

(i) Seien $S_1, S_2, \ldots \subseteq \mathbb{Z}$ beliebige Mengen. Zeige $\lim_{n\to\infty} |\Pr[W_n \in S_n] - \Pr[V_n \in S_n]| = 0$.
Hinweis: Konstruiere ein Coupling $(W'_n, V'_n)_{n\in\mathbb{N}_0}$ von $(W_n)_{n\in\mathbb{N}_0}$ und $(V_n)_{n\in\mathbb{N}_0}$ für das $\lim_{n\to\infty} \Pr[W'_n = V_{n'}] = 1$ gilt.

(ii) Zeige, dass das Ergebnis von Aufgabenteil (i) für eine Verschiebung von 43 statt 42 nicht in dieser Form gilt.

Lösung 1

(i) Wir verwenden $(W'_n) = (W_n)$ und beschreiben (V'_n) in natürlicher Sprache. Zunächst verhalte sich (V'_n) genau gegenläufig zu (W_n), verwendet also die invertierten Beiträge $-X_1, -X_2, -X_3 \ldots$ usw. Sei nun $T = \min\{t \in \mathbb{N} \mid W_n = 21\}$. Dann gilt $W_T = 21$ und $V'_T = 42 - 21 = 21$, das heißt die Irrfahrten begegnen sich zum Zeitpunkt T. Ab diesem Zeitpunkt verhalte sich (V'_n) genau wie (W_n), verwende also die selben Beiträge $X_{T+1}, X_{T+2}, \ldots$.

Es ist ziemlich klar, dass $(V'_n)_{n\in\mathbb{N}_0} \stackrel{\mathrm{d}}{=} (V_n)_{n\in\mathbb{N}_0}$ gilt, denn die Beiträge, die wir akumulieren sind nach wie vor unabhängige Zufallsvariablen und gleichverteilt in $\mathcal{U}(\{-1, 1\})$ (ob wir X_i addieren oder subtrahieren legen wir fest noch bevor wir den Wert X_i kennen). Also haben wir ein gültiges Coupling. In diesem Coupling gilt die Implikation $W_n \neq V'_n \Rightarrow T \geq n$.

Wir machen nun eine Hilfsrechnung für beliebige Zufallsvariablen X, Y und beliebige Mengen S.

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

Wenden wir dies nun auf $S = S_n$, $X = W_n$ und $Y = V_n'$ an, so ergibt sich:

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

wobei der letzte Schritt den Hinweis für $c = 21$ verwendet.

(ii) So wie wir die Irrfahrt definiert haben, "vergisst" diese zwar ihren konkreten Startpunkt aber nicht die Parität dieses Startpunktes. Mit anderen Worten, wenn wir $S_n := S := 2 \cdot \mathbb{Z}$ definieren, dann sind Irrfahrten immer abwechselnd in S und nicht in S. Für eine Verschiebung von 23 hätten wir dann $|\Pr[W_n \in S_n] - \Pr[V_n \in S_n]| = 1$ für alle $n \in \mathbb{N}$.

Aufgabe 2 – Coupling und Total Variation Distance

Seien X und Y zwei Zufallsvariablen mit Werten in $\mathbb{N}$. Der Totalvariationsabstand (engl. *total variation distance*) von X und Y (bzw. deren Verteilungen) ist definiert als[1]

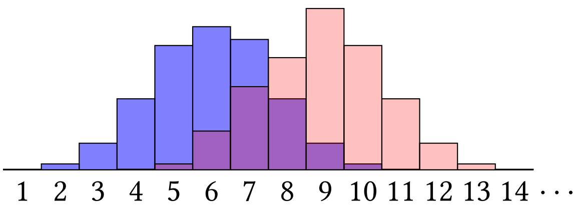
$$d(X,Y) = \frac{1}{2}\sum_{i\in\mathbb{N}} |\Pr[X = i] - \Pr[Y = i]|.$$

(i) Zeige: Es gibt ein Coupling (X', Y') von X und Y sodass $\Pr[X' \neq Y'] = d(X,Y)$.

(ii) Zeige: Kein Coupling (X', Y') von X und Y erfüllt $\Pr[X' \neq Y'] < d(X,Y)$.

Lösung 2

Zur Vorbereitung stellen wir uns ein gemeinsames Histogramm von X (blau) und Y (rot) vor.



[1]Eine allgemeine Definition, die auch für kontinuierliche Wahrscheinlichkeitsräume funktioniert, findet man auf Wikipedia.

Alle Balken mögen Breite 1 habe. Wir bezeichnen mit rot, blau und lila die Menge der Punkte der entsprechenden Farben und mit A_{rot}, A_{blau} und A_{lila} die zugehörigen Flächeninhalte. Weil die Balken Verteilungen beschreiben gilt $A_{\text{blau}} + A_{\text{lila}} = 1$ sowie $A_{\text{rot}} + A_{\text{lila}} = 1$, also folgt $A_{\text{blau}} = A_{\text{rot}}$. Beide sind jeweils genau der Totalvariationsabstand $d(X, Y)$. Das sieht man so:

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

(i) Wir sampeln zunächst ein Paar (P, Q) von Punkten folgendermaßen
   - sample $P \sim \mathcal{U}(\text{blau} \cup \text{lila})$
   - falls $P \in \text{lila}$ setze $Q = P$
   - andernfalls sample $Q \sim \mathcal{U}(\text{rot})$.

   Es sollte klar sein, dass damit $Q \sim \mathcal{U}(\text{rot} \cup \text{lila})$ gilt. Wir definieren nun X' als den Index des Balkens in dem P liegt und Y' als den Index des Balkens in dem Q liegt. Nun sollte klar sein, dass $X' \overset{\text{d}}{=} X$ und $Y' \overset{\text{d}}{=} Y$ gilt. Die nützliche Eigenschaft, die wir verwenden werden, ist $\Pr[X' = Y'] = \Pr[P = Q] = A_{\text{lila}}$. Daraus folgt nämlich wie gewünscht:

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

(ii) Sei $S = \{i \in \mathbb{N} \mid \Pr[X = i] > \Pr[Y = i]\}$. Sei nun (X', Y') irgendein Coupling von X und Y. Dann gilt:

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

Aufgabe 3 – Eigenschaften der Poissonverteilung

Sei $X \sim \text{Pois}(\lambda)$. Zeige:

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

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

(iii) Für $Y \sim \text{Pois}(\rho)$ unabhängig von X gilt $X + Y \sim \text{Pois}(\lambda + \rho)$.

(iv) Für $X' \sim \text{Bin}(X, p)$ gilt $X' \sim \text{Pois}(\lambda p)$.
Beachte: Hier wird also ein zweistufiges Zufallsexperiment durchgeführt. Das Ergebnis X des ersten ist ein Parameter des zweiten.

Lösung 3

Im folgenden verwenden wir die Definition der e-Funktion ständig, d.h. $e^t = \sum_{i=0}^{\infty} \frac{t^i}{i!}$.

(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) Wir bestimmen zunächst das zweite *un*zentrierte Moment:

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

Wir wissen zudem $\mathbb{E}[X]^2 = \lambda^2$. Es folgt

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

(iii) Sei $k \in \mathbb{N}$. Wir betrachten alle $k+1$ Möglichkeiten wie $X + Y$ zur Summe k führen kann und verwenden dann den binomischen Lehrsatz.

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

(iv) Sei $k \in \mathbb{N}$. Damit am Ende k herauskommt muss $X \geq k$ gegolten haben. Wir betrachten alle Möglichkeiten.

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

Aufgabe 4 – Poissonisierte Bloom-Filter

Wir betrachten ein Poisson-Modell von Bloom-Filtern, gehen also davon aus, dass jede Position im Array unabhängig von andere Positionen $\mathrm{Pois}(\alpha k)$-verteilt oft als Hashwert vorkommt.

(i) Wir wählen wieder $\alpha k = \ln 2$. Wie lässt sich zeigen, dass der Anteil $\frac{Z}{m}$ der Nuller mit hoher Wahrscheinlichkeit nahe an $\frac{1}{2}$ liegt?

(ii) Wie ließe sich das Ergebnis in ein nicht-Poissonisiertes Modell übertragen?

Lösung 4

(i) Wenn $X \sim \mathrm{Pois}(\ln 2)$ so gilt $\Pr[X = 0] = e^{-\ln 2} = \frac{1}{2}$. Da jede Position nun unabhängig von allen anderen leer bzw. nicht leer ist, gilt also $Z \sim \mathrm{Bin}(m, \frac{1}{2})$. Es folgt $\mathbb{E}[\frac{Z}{m}] = \frac{1}{2}$ und auf Z sind nun direkt Chernoff Schranken anwendbar.

(ii) Die Anzahl $m - Z$ ist eine monotone Funktion im Sinne des Poissonisierungs-Theorems der Vorlesung. Entsprechend kann man das exakte "nk Bälle in m Behälter" Modell zwischen zwei Poissonisierten Modellen einsperren wie besprochen.