Gut genug gemischt?
Bildchensammler und zufälliges Mischen

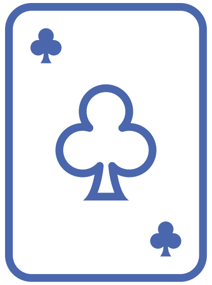
Charlotte Schwerdtner

Institut für Informatik
Humboldt-Universität zu Berlin

Inhalt

Zentrale Fragestellung & Vorüberlegung

Wie oft muss man ein Kartenspiel mischen, bis es zufällig genug ist?



1. Aus wie vielen Karten n besteht das betrachtete Kartenspiel?

   n = 52

Zentrale Fragestellung & Vorüberlegung

Wie oft muss man ein Kartenspiel mischen, bis es zufällig genug ist?

2. Auf welche Art und Weise wird das Kartenspiel gemischt?

Variante 1:

Die oberste Karte zufällig hineinstecken.

Variante 2:

Stapel teilen und ineinander schieben.
(Nicht Teil des Vortrags)

Zentrale Fragestellung & Vorüberlegung

Wie oft muss man ein Kartenspiel mischen, bis es zufällig genug ist?

3. Wann ist ein Kartenspiel zufällig genug gemischt?

Ziel: Modell des Kartenmischens untersuchen.

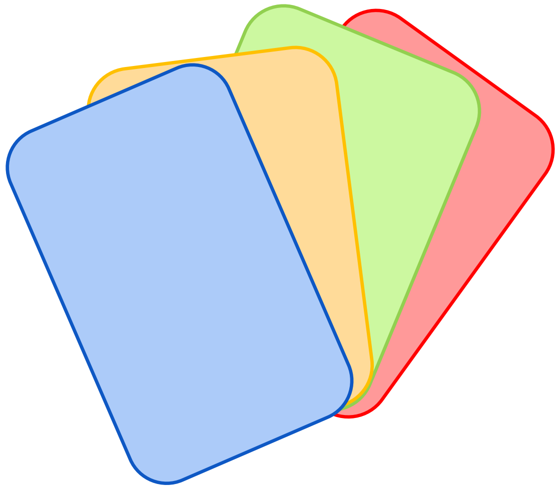
Genauer: obere Schranke für die Anzahl der Wiederholungen des Mischvorgangs

Der Bildchensammler

Problemstellung

Ein Sammelalbum soll mit Stickern gefüllt werden. Es existieren **n verschiedene** Sticker, die in undurchsichtigen Umschlägen verkauft werden.

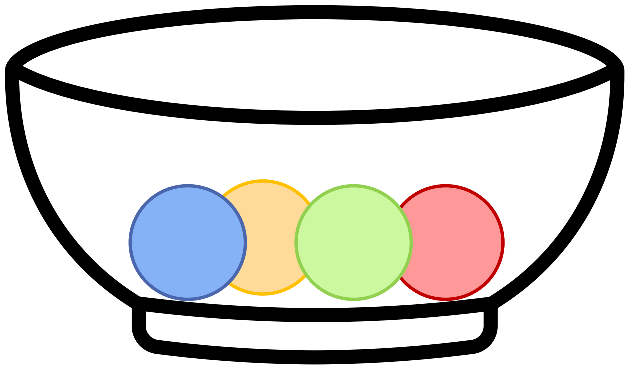
Wie viele Sticker müssen gekauft werden, damit man jeden Sticker **mindestens einmal** besitzt?



Gesucht ist hierbei ein **Erwartungswert**.

Der Bildchensammler

Rückführung zum Urnenmodell



→ Schale mit n unterscheidbaren Kugel

→ Ziehen mit Zurücklegen

Frage:
Wie oft muss man im Durchschnitt ziehen, bis man jede Kugel mindestens einmal gezogen hat?

Bemerkung: Nehmen an, dass jede Kugel nur einmal vorkommt; müssen beim Kauf von Stickern also auch annehmen, dass jeder Sticker in der selben Anzahl vorkommt.

Der Bildchensammler

Wahrscheinlichkeit, dass man bei der nächsten Ziehung keine neue Kugel erhält

<u>*Ereignis*</u>

A: "$Beim\ nächsten\ Mal\ Ziehen\ erhält\ man\ keine\ neue\ Kugel.$„

$insgesamt\ n\ verschiedene\ Kugeln, \quad k\ verschiedene\ Kugeln\ bereits\ gezogen$

$$\mathbb{P}(A) = \frac{Anzahl\ günstiger\ Ergebnisse}{Anzahl\ möglicher\ Ergebnisse} \quad \Rightarrow \quad \mathbb{P}(A) = \frac{k}{n}$$

Der Bildchensammler

Wahrscheinlichkeit, dass man nach genau s Ziehungen eine neue Kugel erhält

$\underline{Zufallsvariable}$

X: "$Anzahl\ der\ Ziehungen, bis\ eine\ neue\ Kugel$ gezogen wird."

$$\mathbb{P}(X = s) = \mathbb{P}(A)^{s-1}\mathbb{P}(\bar{A})$$

$$\Rightarrow \quad \mathbb{P}(X = s) = \left(\frac{k}{n}\right)^{s-1}\left(1 - \frac{k}{n}\right)$$

Der Bildchensammler

Erwartete Anzahl von Ziehungen bis zur nächsten neuen Kugel

Erwartungswert von X

$$\mathbb{E}[X] = \sum_{s \geq 1} \mathbb{P}(X = s) \cdot s$$

$$\Rightarrow \quad \mathbb{E}[X] = \sum_{s \geq 1} \left(\frac{k}{n}\right)^{s-1} \left(1 - \frac{k}{n}\right) s = ?$$

Der Bildchensammler

Erwartete Anzahl von Ziehungen bis zur nächsten neuen Kugel

$$\mathbb{E}[X] = \sum_{s\geq 1} \left(\tfrac{k}{n}\right)^{s-1} \left(1 - \tfrac{k}{n}\right) s$$

Distributivgesetz

$$= \sum_{s\geq 1} \left(\tfrac{k}{n}\right)^{s-1} s \; - \; \sum_{s\geq 1} \left(\tfrac{k}{n}\right)^{s} s$$

Indexverschiebung

$$= \sum_{s\geq 0} \left(\tfrac{k}{n}\right)^{s} (s+1) \; - \; \sum_{s\geq 0} \left(\tfrac{k}{n}\right)^{s} s$$

Zusammenfassen

$$= \sum_{s\geq 0} \left(\tfrac{k}{n}\right)^{s} = \frac{1}{1-\frac{k}{n}}$$

Geometrische Reihe

Der Bildchensammler

Erwartete Anzahl von Ziehungen bis man alle Kugeln mindestens einmal gezogen hat

$$\sum_{k=0}^{n-1} \frac{1}{1-\frac{k}{n}} = \frac{1}{1} + \frac{1}{1-\frac{1}{n}} + \cdots + \frac{1}{1-\frac{n-1}{n}}$$

$$= \frac{n}{n} + \frac{n}{n-1} + \cdots + \frac{n}{1} = nH_n \approx n \log n$$

$$H_{n=} \sum_{k=1}^{n} \frac{1}{k}$$

Harmonische Reihe

Das entspricht n-mal die erwarte Anzahl an Ziehungen bis zur jeweils nächsten neuen Kugel.

Der Bildchensammler

Veranschaulichung

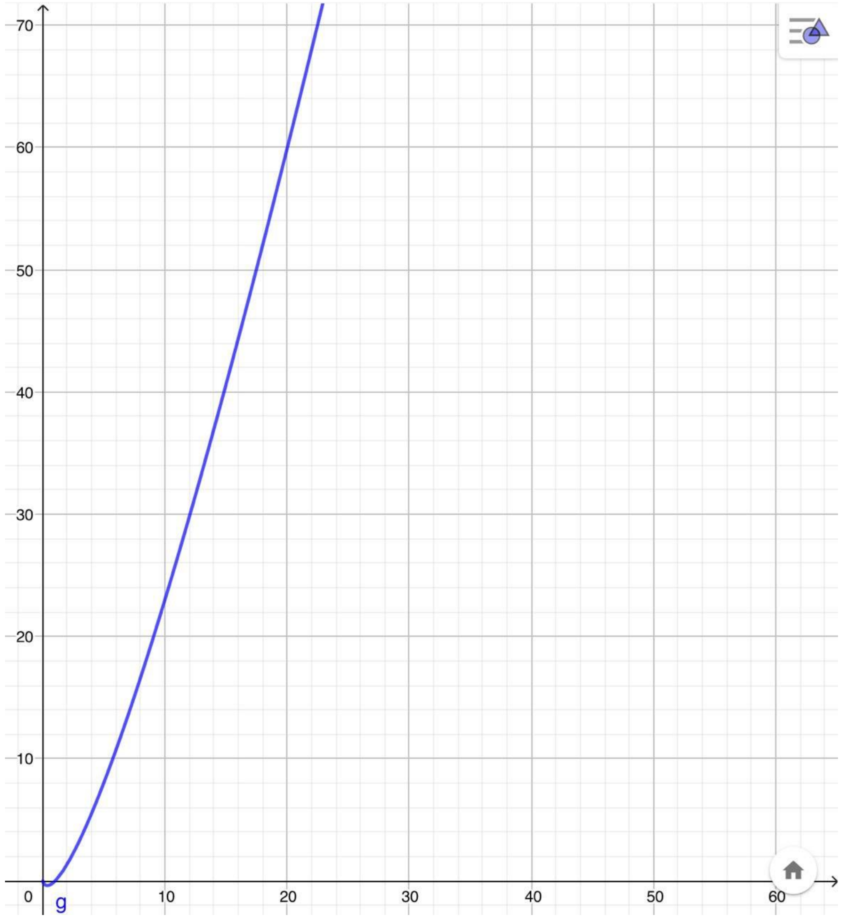
<u>*Beispiel*</u>

Insgesamt existieren 20 verschiedene Sticker, die man sammeln kann.

Erwartungsgemäß müssen

$$20 \cdot \log 20 \ \approx 60$$

Umschläge mit Sticker gekauft werden, damit man jeden mindestens einmal besitzt.



Der Bildchensammler

Abschätzung

Wie wahrscheinlich ist es, dass man deutlich mehr als $n \log n$ Ziehungen benötigt?

Sein V_n die Zufallsvariable für die Anzahl der Ziehungen, bis man alle Kugeln bzw. Sticker hat;

also $\mathbb{E}[V_n] = n \log n$

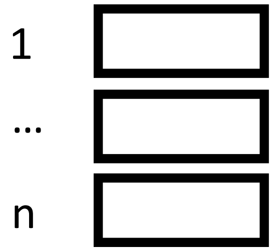
Betrachten $m \coloneqq [n \log n + cn]$ mit $n \geq 1, \quad c \geq 0$

$$\mathbb{P}(V_n > m) \leq e^{-c}$$

Diese Abschätzung wird im weiteren Verlauf noch benötigt, auf einen Beweis wurde aus Zeitgründen verzichtet.

Was bedeutet „zufällig genug“?

Gegeben: Kartenstapel bestehend aus n Karten, nummeriert von 1 bis n



Betrachten: Menge aller Permutationen ζ_n, insgesamt n!

„Mischen des Stapels“ entspricht der Anwendung einer Permutation $\pi \in \zeta_n$ auf die Startreihenfolge (1, 2, …, n)

Ideal: erhalten perfekte Zufallsreihenfolge $\pi = (\pi(1), \pi(2), \dots, \pi(n))$

Realität: treten nur bestimmte Permutationen mit unterschiedlichen Wahrscheinlichkeit auf → erwarten „zufällig genug“ bei <u>mehrfacher Wiederholung</u>

Die oberste Karte zufällig hineinstecken

Situation

Kartenspiel bestehend aus n Karten, oberste Karte wird in einen der n Zwischenräume gesteckt

Annahme: Wahrscheinlichkeit für jeden Zwischenraum beträgt $\frac{1}{n}$

D.h. es wird eine der n Permutationen der Form

$$\pi_i = (2, 3, \dots, i, 1, i+1, \dots, n) \text{ mit } 1 \leq i \leq n$$

auf den Stapel angewandt.

→ Stapel ist sicher **noch nicht „zufällig genug“**

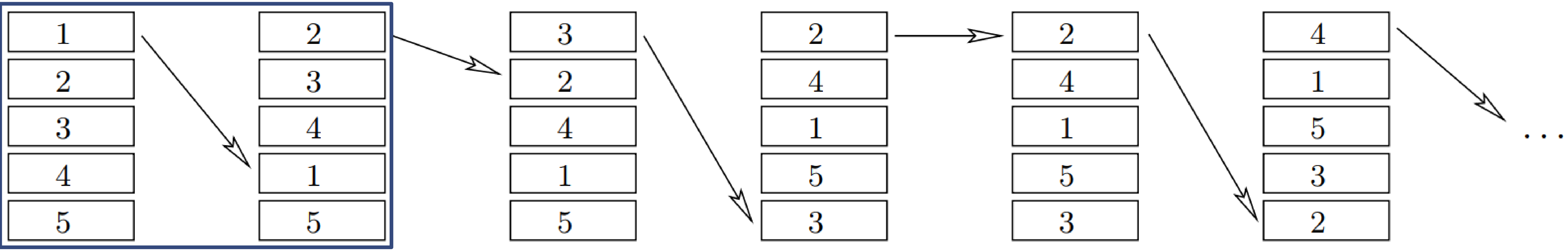
Die oberste Karte zufällig hineinstecken

Beispiel

5 Karten mit der Startreihenfolge (1, 2, 3, 4, 5), oberste Karte wird zufällig in den vierten Zwischenraum gesteckt

$$\pi_4 = (2, 3, 4, 1, 5)$$

Frage: Nach wie vielen Wiederholungen ist der Stapel „zufällig genug"?



Die oberste Karte zufällig hineinstecken

Variationsdistanz

misst den Abstand zum Zufall

Die Variationsdistanz zwischen zwei Wahrscheinlichkeitsverteilungen Q_1 und Q_2 wird folgendermaßen definiert:

$$||Q_1 - Q_2|| := \frac{1}{2} \sum_{\pi \in \zeta_n} |Q_1(\pi) - Q_2(\pi)|$$

Die oberste Karte zufällig hineinstecken

Variationsdistanz

Ziel: Variationsdistanz in eine andere Form überführen, damit sie besser anzuwenden ist.

$$||Q_1 - Q_2|| = \frac{1}{2}\sum_{\pi \epsilon \zeta_n} |Q_1(\pi) - Q_2(\pi)|$$

$$\boxed{\mathcal{S} = \{\pi \epsilon \zeta_n \colon Q_1(\pi) > Q_2(\pi)\}}$$

$$= \frac{1}{2}\sum_{\pi \epsilon \mathcal{S}} \big(Q_1(\pi) - Q_2(\pi)\big) + \frac{1}{2}\sum_{\pi \epsilon \zeta_n \setminus \mathcal{S}} \big(Q_2(\pi) - Q_1(\pi)\big)$$

Die oberste Karte zufällig hineinstecken

Variationsdistanz

$$\sum_{\pi \epsilon \mathcal{S}} \left(Q_1(\pi) - Q_2(\pi)\right) = \boxed{\sum_{\pi \epsilon \mathcal{S}} Q_1(\pi)} - \sum_{\pi \epsilon \mathcal{S}} Q_2\,(\pi)$$

$$\boxed{\sum_{\pi \epsilon \zeta_n} Q_1(\pi) = \sum_{\pi \epsilon \zeta_n} Q_2(\pi) = 1}$$

$$= \boxed{1 - \sum_{\pi \epsilon \zeta_n \setminus \mathcal{S}} Q_1(\pi)} - \sum_{\pi \epsilon \mathcal{S}} Q_2(\pi)$$

$$= \sum_{\pi \epsilon \zeta_n \setminus \mathcal{S}} Q_2(\pi) - \sum_{\pi \epsilon \zeta_n \setminus \mathcal{S}} Q_1(\pi) = \sum_{\pi \epsilon \zeta_n \setminus \mathcal{S}} \left(Q_2(\pi) - Q_1(\pi)\right)$$

Die oberste Karte zufällig hineinstecken

Variationsdistanz

$$\Rightarrow \|Q_1 - Q_2\| = \sum_{\pi \epsilon \mathcal{S}} \left(Q_1(\pi) - Q_2(\pi)\right)$$

$$= \sum_{\pi \epsilon \mathcal{S}} Q_1(\pi) - \sum_{\pi \epsilon \mathcal{S}} Q_2(\pi)$$

$$= Q_1(\mathcal{S}) - Q_2(\mathcal{S})$$

$$Q_i \coloneqq \sum_{\pi \epsilon \mathcal{S}} Q_i(\pi)$$

$$0 \leq \|Q_1 - Q_2\| \leq 1$$

Es müssen also nur diejenigen Permutationen betrachtet werden, dessen Wahrscheinlichkeiten für Q_1 größer sind als für Q_2.

Die oberste Karte zufällig hineinstecken

Variationsdistanz anwenden

Betrachten zwei Wahrscheinlichkeitsverteilungen auf den $n!$ verschiedenen Permutationen $\pi \epsilon \zeta_n$:

Anfangsverteilung E: $E(id) = 1; \quad E(\pi) = 0 \quad sonst$

Entspricht dem Zustand vor Begin des Mischens.

Gleichverteilung U: $U(\pi) = \frac{1}{n!} \quad f\ddot{u}r\ alle\ \pi \epsilon\ \zeta_n$

Entspricht dem idealen Zustand: Jeder Permutation tritt mit selber Wahrscheinlichkeit ein.

Variationsdistanz: $\|E - U\| = 1 - \frac{1}{n!}$

→ fast 1

Ziel: Variationsdistanz so nah wie möglich an die 0, wollen also möglichst den Zustand einer Gleichverteilung.

Die oberste Karte zufällig hineinstecken

Variationsdistanz nach einmaligem ‚Mischen'

Nach einmaligem zufälligen Hineinstecken tritt eine der n Permutationen auf:

$$\pi_i = (2, 3, \dots, i, 1, i+1, \dots, n) \text{ mit } 1 \leq i \leq n$$

Daraus ergibt sich folgende Wahrscheinlichkeitsverteilung:

$$Top(\pi_i) = \frac{1}{n} \quad für\ alle\ i = 1,2, \dots, n; \qquad Top(\pi) = 0 \quad sonst$$

Variationsdistanz:

$$\|Top - U\| = n \cdot \frac{1}{n} - \mathrm{n} \cdot \frac{1}{n!} = 1 - \frac{1}{(n-1)!}$$

Die oberste Karte zufällig hineinstecken

Variationsdistanz mehrerer Wiederholungen

Steckt man k Mal die oberste Karte in den Stapel hinein erhält man die Zufallsverteilung Top^{*k}

Frage: Wie verhält sich $\mathrm{d}(\mathrm{k}) \coloneqq \left\|Top^{*k} - U\right\|$ für größer werdendes k?

Stark gleichverteilte Halteregeln

Was ist eine stark gleichverteilte Halteregel?

Halteregel: Eine Regel, um den Mischprozess zu beenden; abhängig von den zufällig durchgeführten Mischoperationen

stark gleichverteilt: es muss für jedes mögliche k folgende Bedingung gelten:

> ***Wenn*** *der Mischprozess nach genau* k *Schritten abgebrochen wird,* ***dann*** *gilt für die daraus resultierenden Permutationen des Kartenstapels die Gleichverteilung (und zwar genau).*

Stark gleichverteilte Halteregeln

Grundlage: Bedingte Wahrscheinlichkeiten

Es seinen A und B zwei Ereignisse. Die Wahrscheinlichkeit, dass Ereignis A eintritt, unter der Bedingung, dass das Ereignis B eintritt, wird folgendermaßen definiert:

$$\mathbb{P}(A|B) = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)}$$

Stark gleichverteilte Halteregeln

Definition der Halteregel

Zufallsvariable T: Anzahl der Mischoperationen bis es zum Halt kommt

Zufallsvariable $\boldsymbol{X_k}$: Vorhandene Permutation nach k-maligem Mischen (Werte in ζ_n)

Halteregel ist stark gleichverteilt, wenn für alle möglichen k gilt:

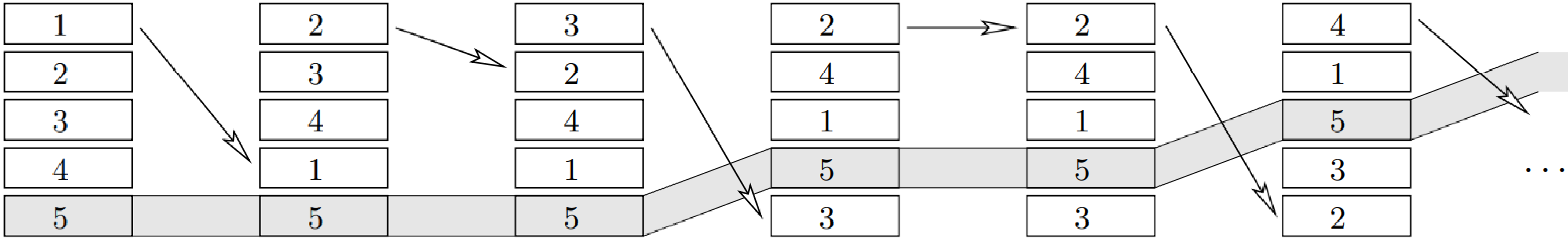
$$\mathbb{P}(X_k = \pi | T = k) = \frac{1}{n!} \quad f\ddot{u}r\ alle\ \pi \in \zeta_n$$

Stark gleichverteilte Halteregeln

Halteregel für Hineinstecken der obersten Karte

HALT, sobald die ursprünglich unterste Karte mit der Nummer n das erste Mal in den Stapel hineingesteckt wurde.

Es handelt sich um eine stark gleichverteilte Halteregel. Wann wird der Halt erwartet?



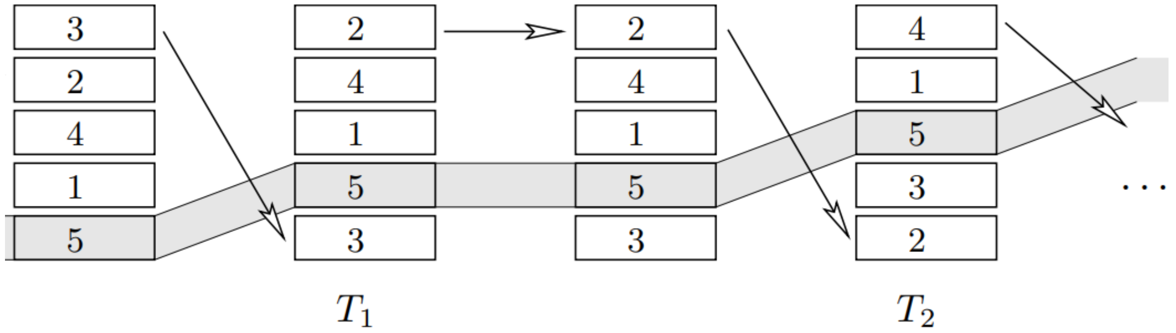
Stark gleichverteilte Halteregeln

Halteregel für Hineinstecken der obersten Karte

Zufallsvariable T_i: Anzahl der Mischoperationen, bis erstmals i Karten unter der Karte n liegen

Bestimmung der Wahrscheinlichkeitsverteilung von T, es gilt:

$$T = T_1 + (T_2 - T_1) + \cdots + (T_{n-1} - T_{n-2}) + (T - T_{n-1})$$



→ Rückführung auf das Bildchensammlerproblem

Stark gleichverteilte Halteregeln

Rückführung zum Bildchensammler

$$T = T_1 + (T_2 - T_1) + \cdots + (T_{n-1} - T_{n-2}) + (T - T_{n-1})$$

$(\boldsymbol{T_i - T_{i-1}})$: Anzahl der Mischoperationen, bis eine Karte an einer der i möglichen Stellen unterhalb der Karte n eingefügt wurde

$\leftrightarrow$

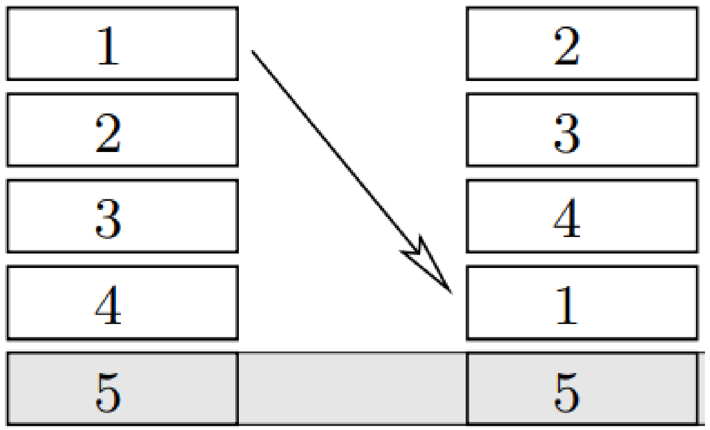
Zufallsvariable $\boldsymbol{V_i}$: Anzahl gekaufter Bilder, bis man i unterschiedliche Bilder besitzt

$$V_n = V_1 + (V_2 - V_1) + \cdots + (V_{n-1} - V_{n-2}) + (V_n - V_{n-1})$$

$(\boldsymbol{V_i - V_{i-1}})$: Anzahl gekaufter Bilder, bis man ein neues Bild erhält

Stark gleichverteilte Halteregeln

Rückführung zum Bildchensammler

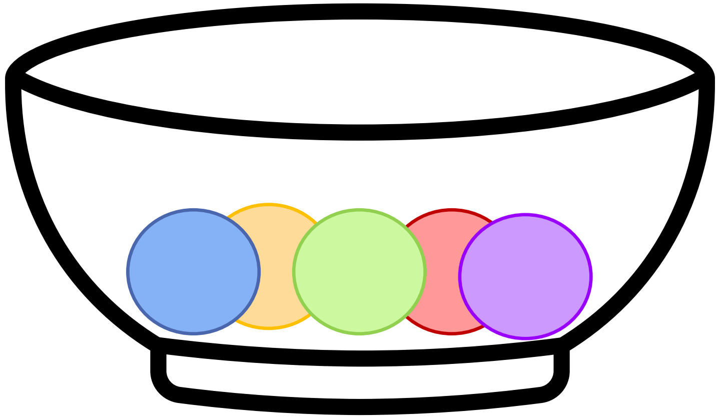


Für alle j gilt:

$$\mathbb{P}(T_1 - T_0 = j) = \mathbb{P}(V_5 - V_4 = j)$$

Für alle i und j gilt:

$$\mathbb{P}(T_i - T_{i-1} = j) = \mathbb{P}(V_{n-i+1} - V_{n-i} = j)$$



Stark gleichverteilte Halteregeln

Rückführung zum Bildchensammler

Die stark gleichverteilte Halteregel lässt für das zufällige Hineinstecken der obersten Karte nur mit sehr geringer Wahrscheinlichkeit öfter als $k = \lceil n \log n + cn \rceil$ mal mischen:

$$\mathbb{P}(T > k) \leq e^{-c}$$

$\Rightarrow$ nach $k = \lceil n \log n + cn \rceil$-maligem Mischen ist der Stapel „nahezu zufällig“

Wie sieht die **Variationsdistanz** aus?

Stark gleichverteilte Halteregeln

Lemma

Sei $Q\colon \zeta_n \longrightarrow \mathbb{R}$ eine beliebige Zufallsverteilung, die einen Mischprozess Q^{*k} mit einer stark gleichverteilten Halteregel definiert, dessen Haltezeit durch T gegeben ist. Dann gilt für alle $k \geq 0$:

$$\left\|Q^{*k} - U\right\| \leq \mathbb{P}(T > k)$$

Daraus ergibt sich:

$$d(k) = \left\|Top^{*k} - U\right\| \leq \mathbb{P}(T > k) \leq e^{-c}$$

Stark gleichverteilte Halteregeln

Ergebnis

Satz 1. *Sei $c \geq 0$ und $k \coloneqq \lceil n \log n + cn \rceil$. Wenn man k Mal die oberste Karte zufällig in einen Stapel von $n \geq 2$ Karten hineingesteckt hat, dann erfüllt die Variationsdistanz von der Gleichverteilung*

$$d(k) \coloneqq \left\| Top^{*k} - U \right\| \leq e^{-c}.$$

$n = 52 \Rightarrow n \log n \approx 205$

→ sehr ineffektiv

Quellenangabe

Martin Aigner, Günter M. Ziegler: Das BUCH der Beweise, Dritte Auflage, Springer-Verlag Berlin Heidelberg 2009

Vielen Dank für Ihre Aufmerksamkeit!