

Zdeněk Režný

K Robbinsovu-Isbellovu sekvenčnímu rozhodovacímu problému s konečnou pamětí

*Kybernetika*, Vol. 3 (1967), No. 1, (36)--48

Persistent URL: <http://dml.cz/dmlcz/125528>

## Terms of use:

© Institute of Information Theory and Automation AS CR, 1967

Institute of Mathematics of the Academy of Sciences of the Czech Republic provides access to digitized documents strictly for personal use. Each copy of any part of this document must contain these

*Terms of use.*



This paper has been digitized, optimized for electronic delivery and stamped with digital signature within the project *DML-CZ: The Czech Digital Mathematics Library*  
<http://project.dml.cz>

## K Robbinsovu-Isbellovu sekvenčnímu rozhodovacímu problému s konečnou pamětí

ZDENĚK REŽNÝ

Předmětem tohoto článku je problém předložený Robbinsem [1] a přesně formulovaný Isbellem [2]. Je uvedena formulace problému a přehled dosavadního stavu řešení. Vlastní výsledek se týká případu délky paměti 2. Je udána třída  $\mathcal{R}_2$  pravidel ekvivalentních s Robbinsovým-Isbellovým pravidlem, a je dokázána její vlastnost, z níž plyne, že žádné pravidlo nepatřící do  $\mathcal{R}_2$  nemůže být stejnoměrně nejlepší.

### 1. FORMULACE PROBLÉMU

Experimentátor realizuje nekonečnou posloupnost opakovaných nezávislých pokusů hodů mincí. Ke každému jednotlivému pokusu má možnost volit jednu ze dvou daných mincí (mince 1 a mince 2). Možné výsledky jednotlivého pokusu jsou tedy čtyři a označíme je shodně s [3]

$$(1.1) \quad H_1, H_2, T_1, T_2,$$

kde  $H$  značí padnutí líce,  $T$  rubu a index použitou minci. Pro minci  $i$  je pravděpodobnost líce  $p_i = 1 - q_i$  ( $i = 1, 2$ ). Experimentátoru, jehož cílem je dosáhnout co největší asymptotické relativní četnosti líců, není známo, která z pravděpodobností  $p_1, p_2$  je větší, resp. zda jsou si rovny. Kromě toho při provádění opakovaných pokusů je jeho informace o minulé historii procesu omezena konečnou *délkou paměti*  $r$ , tzn. před provedením  $n$ -tého pokusu ( $n \geq r + 1$ ) je mu známo, který z výsledků (1.1) měl  $(n - 1)$ -vý,  $(n - 2)$ -hý, ...,  $(n - r)$ -tý pokus, avšak o výsledcích pokusů  $(n - r - 1)$ -vého a předešlých žádnou informaci nemá. V souvislosti s tím definujeme *stav paměti* jako posloupnost výsledků  $r$  posledních provedených pokusů. Při nekonečné posloupnosti pokusů se tedy realizuje nekonečná posloupnost stavů paměti, z nichž však nejvýše  $4^r$  je různých. *Pravidlo* je zobrazení množiny stavů paměti do množiny  $\{1, 2\}$ , jímž je udán předpis pro volbu mince ke každému pokusu, počínaje  $(r + 1)$ -vým, v závislosti na zjištěném stavu paměti. Pro volbu mince

v prvním až  $r$ -tém pokusu není však dán žádný předpis a budeme počítat s tím, že se může realizovat kterákoli z  $2^r$  možností. Budeme užívat ekvivalentní definice pravidla jakožto udání těch stavů paměti, při jejichž uskutečnění je předepsáno v následujícím pokusu užít jiné mince než při posledním pokusu. Tyto stavy nazveme podle [2] *vyznačené*.

Např. pravidlo  $R_r^0$  (značení přejato z [3]) navržené Robbinsem [1] je formulováno slovně takto: *Vyměnit mince, jestliže  $r$  za sebou jdoucích hodů toutéž mincí mělo výsledek rub; jestliže v následujícím hodu (provedeném ovšem již druhou mincí) padne rub, opět vyměnit mince a tak pokračovat, dokud nepadne líc*. Toto pravidlo má tedy celkem  $2r$  vyznačených stavů, totiž  $r$  stavů

$$(1.2) \quad \underbrace{T_2 T_2 \dots T_2 T_2}_r, \quad \underbrace{T_2 T_2 \dots T_2 T_1}_{r-1}, \quad \underbrace{T_2 T_2 \dots T_2 T_1 T_2}_{r-2}, \quad \dots, \quad T_2 T_1 T_2 \dots T_1$$

$$(i = [3 - (-1)^i]/2).$$

a stavy s těmito souměrné (vzniklé z nich záměnou indexů). Isbell [2] navrhl pravidlo  $R_r^1$  o čtyřech vyznačených stavech

$$(1.3) \quad \underbrace{T_i T_i \dots T_i T_i}_r, \quad \underbrace{T_i T_i \dots T_i T_j}_{r-1} \quad (i = 1, 2; j = 3 - i).$$

Pro  $r = 1, 2$  ovšem obě tato pravidla splývají ( $R_r^0 \equiv R_r^1$  pro  $r = 1, 2$ ).

Vlastní úlohou položenou Robbinsem [1] je najít pravidlo optimální pro experimentátora, tj. takové, které by zaručovalo při dané délce paměti  $r$  v nekonečné posloupnosti pokusů maximální podíl líců. Zrekapitulujeme zde přesnou formulaci této úlohy podanou Isbellem [2] a úvahy, na nichž je založena. Budiž dána délka paměti  $r$  a zvoleno nějaké pravidlo  $R$ . Posloupnost stavů paměti při procesu opakovaných pokusů pak tvoří konečný homogenní Markovův řetězec  $M_0$ , jehož přechodové pravděpodobnosti jsou určeny pravidlem  $R$  a pravděpodobnostmi  $p_1, p_2$ . Tento řetězec má podle [4] určitý počet  $m \geq 1$  uzavřených tříd persistentních stavů  $C_1, \dots, C_m$  a popřípadě mimo to ještě transientní stavy. (Snadno nahlédneme, že v případě  $0 < p_i < 1$  pro  $i = 1, 2$  jsou třídy  $C_j$  nezávislé na  $p_i$  a v ostatních případech se jen některé stavy stávají transientními.) Ze silného zákona velkých čísel pak vyplývá, že v posloupnosti pokusů podíl líců skoro jistě konverguje k limitě závislé na tom, ve které uzavřené třídě  $C_i$  je proces absorbován. Můžeme tedy tuto limitu značit  $f(R, \alpha, p_1, p_2)$ , kde  $\alpha$  je počáteční stav paměti daný výsledky prvních  $r$  pokusů. Nutno se pojistit proti tomu, že jak počáteční stav  $\alpha$ , tak i očíslování mincí provedené experimentátorem jsou pro dosažení stanoveného cíle nepříznivé, a proto uijžeme minimaxového principu: Definujeme funkci zvanou *cena* pravidla  $R$

$$(1.4) \quad W(R, p_1, p_2) = \min_{\alpha \in M_0} \min [f(R, \alpha, p_1, p_2), f(R, \alpha, p_2, p_1)]$$

a pravidlo  $R$  s maximální cenou  $W$  při všech dvojicích  $p_1, p_2$  nazveme *stejněměrně nejlepší* (pro příslušnou délku paměti). Robbinsův-Isbellův problém pak záleží v nalezení stejněměrně nejlepšího pravidla pro každou délku paměti.

Řešení problému je známo jen pro nejjednodušší případ  $r = 1$ , zatímco ostatní případy ( $r > 1$ ) jsou dosud ve stadiu hypotéz. Isbell [2] ukázal, že pro  $r \geq 3$  (kdy pravidla  $R_r^0$  a  $R_r^1$  jsou navzájem různá) je pravidlo  $R_r^1$  stejněměrně lepší než  $R_r^0$ , ale Smith a Pyke [3] v tomto směru pokračovali a našli pravidla stejněměrně lepší než  $R_r^1$ . Naproti tomu v případě  $r = 2$ , jak ukážeme v tomto článku, žádné pravidlo není stejněměrně lepší než  $R_2^1$  ( $\equiv R_2^0$ ), takže dosud neznámé řešení problému je pro délku paměti  $r = 2$  omezeno na dvě možnosti: a) Žádné pravidlo není stejněměrně nejlepší. b) Stejněměrně nejlepší je pravidlo  $R_2^1$  a každé pravidlo, které má tutéž cenu jako  $R_2^1$  pro všechna  $p_1, p_2$ . (Takových pravidel, odlišných od  $R_2^1$ , je celkem 15, jak v dalším též ukážeme.)

Protože cena pravidla (1.4) je klíčovým pojmem v tomto problému, je účelné věnovat jí podrobnější pozornost. V následujícím paragrafu proto probereme obecnou metodu jejího výpočtu, čímž budou získány předpoklady pro přesnou formulaci a důkaz vlastních výsledků.

## 2. CENA PRAVIDLA

Definice ceny pravidla ve shodě s [2] byla podána vztahem (1.4). Pro její výpočet bylo v [2] pouze poukázáno na teorii Markovových řetězců a vypočtena cena pravidla  $R_r^1$  pro  $r \geq 3$ . Odvodíme zde proto naznačenou metodu výpočtu v obecném případě, neboť tím získáme vztahy potřebné v dalším.

Nejprve si všimněme nejvýznačnějších vlastností vyplývajících bezprostředně z definice. Cena pravidla je souměrná funkce, tj.

$$(2.1) \quad W(R, p_1, p_2) = W(R, p_2, p_1),$$

a platí pro ni nerovnosti

$$(2.2) \quad \min(p_1, p_2) \leq W(R, p_1, p_2) \leq \max(p_1, p_2),$$

což v triviálním případě  $p_1 = p_2$  dává

$$(2.3) \quad W(R, p, p) = p.$$

Nyní se obrátíme k otázce obecného výpočtu ceny při daných  $R, p_1, p_2$ . Nejprve uvažme, že jest-li počáteční stav  $\alpha$  transientní, přísluší mu pravděpodobnost  $u_i$  absorpce v uzavřené třídě  $C_i$  ( $1 \leq i \leq m$ ), přičemž platí

$$(2.4) \quad \sum_{i=1}^m u_i = 1,$$

a vzhledem ke konečné střední době absorpce

$$(2.5) \quad f(R, \alpha, p_1, p_2) = \sum_{i=1}^m u_i f(R, \alpha_i, p_1, p_2) \cdot$$

kde  $\alpha_i$  je libovolný stav ze třídy  $C_i$ . Z (2.4) a (2.5) dostáváme

$$(2.6) \quad f(R, \alpha, p_1, p_2) \geq \min_{1 \leq i \leq m} f(R, \alpha_i, p_1, p_2)$$

a odtud

$$(2.7) \quad \min_{\alpha \in M_0} f(R, \alpha, p_1, p_2) = \min_{1 \leq i \leq m} f(R, \alpha_i, p_1, p_2) \cdot$$

Tím je umožněno přepsat definici (1.4) na jednodušší tvar

$$(2.8) \quad W(R, p_1, p_2) = \min_{1 \leq i \leq m} \min [f(R, \alpha_i, p_1, p_2), f(R, \alpha_i, p_2, p_1)] \quad (\alpha_i \in C_i),$$

čímž se úloha redukuje na výpočet veličiny  $f$  uvnitř uzavřené třídy  $C_i$  řetězce  $M_0$ .

Uvažujme nejprve extrémní případ, kdy uzavřená třída  $C_i$  neobsahuje žádný vyznačený stav. Jestliže některá z pravděpodobností  $p_1, p_2$  má krajní hodnotu 0 nebo 1, pak může být  $f$  rovno buď  $p_1$  nebo  $p_2$ ; ze struktury pravidla  $R$  se snadno určí, který z těchto dvou případů nastává. Je-li  $f$  rovno  $\min(p_1, p_2)$ , pak tutéž hodnotu má i cena pravidla (podle (2.8), (2.2)), kdežto v opačném případě ( $f = \max(p_1, p_2)$ ) nutno zkoumat další hodnoty  $f$  (v ostatních třídách  $C_i$  a při opačném pořadí argumentů  $p_1, p_2$ ). Naproti tomu při  $0 < p_i < 1$  pro  $i = 1, 2$  je vždy jedna z hodnot  $f(R, \alpha_i, p_1, p_2), f(R, \alpha_i, p_2, p_1)$  rovna  $p_1$  a druhá  $p_2$ , takže v každém případě dostáváme výsledek  $W = \min(p_1, p_2)$ .

Jestliže naproti tomu třída  $C_i$  obsahuje (alespoň jeden) vyznačený stav, pak nutné obsahuje vyznačené stavy  $\beta_1, \dots, \beta_s, \gamma_1, \dots, \gamma_t$  ( $s, t \geq 1$ ) takové, že každý stav  $\beta_l$  obsahuje poslední výsledek hodu mincí 2 a tedy předpisuje následující užití mince 1, kdežto u stavů  $\gamma_l$  je tomu obráceně. Je-li v posloupnosti pokusů dosaženo vyznačeného stavu  $\beta_l$  ( $\gamma_l$ ), pak následuje posloupnost hodů mincí 1 (2), která je zakončena prvním dosažením některého vyznačeného stavu  $\gamma_j$  ( $\beta_j$ ); tuto posloupnost nazveme *blok hodů mincí 1* resp. 2. Střední délku bloku hodů mincí 1 resp. 2 následujícího za stavem  $\beta_l$  resp.  $\gamma_l$  označíme  $B_1^{(l)}(R, p_1, p_2)$  resp.  $B_2^{(l)}(R, p_1, p_2)$  a určíme ji vhodnými kombinatorickými metodami ( $1 \leq l \leq s$  resp.  $t$ ). Dále budiž  $\pi_{kl}$  resp.  $q_{lk}$  pravděpodobnost, že po stavu  $\beta_k$  resp.  $\gamma_l$  bude v řetězci  $M_0$  následovat stav  $\gamma_l$  resp.  $\beta_k$  dříve než jiný vyznačený stav ( $1 \leq k \leq s, 1 \leq l \leq t$ ). Tím je definován jistý nový řetězec  $M_i$ , který má periodu 2 a všechny stavy persistentní a tvořící jedinou uzavřenou třídu. Stacionární rozdělení řetězce  $M_i$  udává skoro jisté proporce asymptotického výskytu jednotlivých vyznačených stavů v  $C_i$ . Postačí ovšem určit  $c$ -násovky stacionárních

40 pravděpodobností, kde  $c$  je libovolná nenulová konstanta, tj. užit soustavy rovnic

$$(2.9) \quad \pi_k = \sum_{l=1}^t \varrho_l \varrho_{lk}, \quad \varrho_l = \sum_{k=1}^s \pi_k \pi_{kl} \quad (1 \leq k \leq s, 1 \leq l \leq t),$$

$$\sum_{k=1}^s \pi_k + \sum_{l=1}^t \varrho_l = c,$$

z nichž plyne

$$(2.10) \quad \sum_{k=1}^s \pi_k = \sum_{l=1}^t \varrho_l = \frac{c}{2}.$$

( $c/2$ )-násobná střední délka bloku hodů minci 1 resp. 2 v  $C_i$  je pak

$$(2.11) \quad \frac{c}{2} B_1(R, p_1, p_2 | C_i) = \sum_{k=1}^s \pi_k B_1^{(k)}(R, p_1, p_2),$$

$$\frac{c}{2} B_2(R, p_1, p_2 | C_i) = \sum_{l=1}^t \varrho_l B_2^{(l)}(R, p_1, p_2).$$

Získané hodnoty (2.11) vedou k žádanému výsledku ve tvaru

$$(2.12) \quad f(R, x_i, p_1, p_2) = \frac{p_1 B_1(R, p_1, p_2 | C_i) + p_2 B_2(R, p_1, p_2 | C_i)}{B_1(R, p_1, p_2 | C_i) + B_2(R, p_1, p_2 | C_i)} \\ (x_i \in C_i, 1 \leq i \leq m).$$

Uvedme zde příklad, jehož výsledků využijeme v dalším. Při dělce paměti  $r = 2$  nechť má pravidlo vyznačené stavy shodné s (1.3) – což je pro  $r = 2$  totéž co (1.2) – a další vyznačené stavy z některé podmnožiny stavů

$$(2.13) \quad H_1 H_2, H_1 T_2, H_2 H_1, H_2 T_1.$$

Různých podmnožin 4 stavů (2.13) a tedy i jim odpovídajících pravidel je  $2^4 = 16$ . Množinu těchto 16 pravidel nazveme  $\mathcal{R}_2$ . Do  $\mathcal{R}_2$  ovšem patří zejména Robbinsovo-Isbellovo pravidlo  $R_2^0 \equiv R_2^1$  (odpovídající výběru prázdné podmnožiny stavů (2.13)). Dokážeme nyní, že pro  $R' \in \mathcal{R}_2$  a libovolnou dvojici  $p_1, p_2$  ( $p_1 \neq p_2$ ) je

$$(2.14) \quad W(R', p_1, p_2) = W(R_2^0, p_1, p_2) = \frac{p_1 q_2^2 + p_2 q_1^2}{q_2^2 + q_1^2}.$$

(Porovnání s (2.3) ukazuje, že (2.14) platí též v případě  $p_1 = p_2 < 1$ .)

Je-li  $R' \in \mathcal{R}_2$ , snadno se přesvědčíme, že řetězec  $M_0$  má právě jednu uzavřenou třídu a že všechny čtyři stavy (2.13) jsou transientní, ať jsou hodnoty  $p_1, p_2$  jakékoli. Z (2.8) tedy vyplývá, že všechna pravidla z  $\mathcal{R}_2$  mají identickou cenovou funkci, neboli první rovnost (2.14). Druhou rovnost dokážeme nejprve za předpokladu

$\max(p_1, p_2) < 1$ . Vyznačené stavy patřící do  $C_1$  jsou

$$(2.15) \quad \begin{aligned} \beta_1 &= T_2 T_2, \quad \gamma_1 = T_1 T_1, \\ \beta_2 &= T_1 T_2, \quad \gamma_2 = T_2 T_1 \end{aligned}$$

až na případ  $p_i = 0$ , kdy stav  $T_i T_i$  je transientní ( $i = 1$  nebo  $2$ ), a je nutno určit příslušné střední délky bloků. Jestliže po stavu  $\beta_1$  následuje výsledek pokusu  $T_1$ , což nastane s pravděpodobností  $q_1$ , je tím dosaženo stavu  $\gamma_2$ , takže k následujícímu pokusu se užije mince 2, neboli blok hodů mincí 1 má délku pouze 1. V opačném případě následuje za stavem  $\beta_1$  výsledek pokusu  $H_1$  (s pravděpodobností  $p_1$ ) a se zakončením bloku nutno čekat na první uskutečnění stavu  $\gamma_1$ . Totéž však platí pro blok hodů mincí 1 následující za stavem  $\beta_2$ , takže blok hodů mincí 1 je v každém případě podřízen tomuto předpisu: skončit prvním pokusem, je-li výsledek  $T_1$ , a v opačném případě skončit při prvním dosažení stavu  $\gamma_1 = T_1 T_1$ . Střední délku tohoto bloku, kterou označme  $A$ , určíme pomocí střední délky  $B$  bloku, který na rozdíl od předešlého končí pouze prvním dosažením stavu  $\gamma_1 = T_1 T_1$ . Mezi těmito středními délkami platí zřejmě vztahy

$$(2.16) \quad \begin{aligned} A &= q_1 \cdot 1 + p_1(1 + B) = 1 + p_1 B, \\ B &= p_1(1 + B) + q_1(1 + A) = 1 + p_1 B + q_1 A \end{aligned}$$

a odtud  $A = q_1^{-2}$ . Tím jsme zjistili

$$(2.17) \quad B_1^{(k)}(R', p_1, p_2) = q_1^{-2} \quad (k = 1, 2)$$

a odtud vzhledem k souměrnosti vyznačených stavů (2.15) též

$$(2.18) \quad B_2^{(l)}(R', p_1, p_2) = q_2^{-2} \quad (l = 1, 2).$$

Nalezené hodnoty jsou podle (2.11) a (2.10) rovny přímo příslušným hodnotám  $B_1$ ,  $B_2$  (odpadá nutnost určovat stacionární pravděpodobnosti v  $M_1$ ) a dosazením do (2.12) vychází

$$(2.19) \quad f(R', \alpha, p_1, p_2) = \frac{p_1 q_2^2 + p_2 q_1^2}{q_2^2 + q_1^2} \quad (\alpha \in C_1).$$

Protože na obě hodnoty  $p_1, p_2$  jsou kladeny tytéž podmínky ( $\max(p_1, p_2) < 1$ ), dále funkce (2.19) je souměrná a uzavřená třída  $C_1$  je jediná, platí (2.19) pro libovolný počáteční stav  $\alpha$  a též pro obrácené pořadí argumentů  $p_1, p_2$ . Odtud vyplývá druhá rovnost (2.14). — Budiž nyní  $\max(p_1, p_2) = 1$ . Vzhledem k (2.1) můžeme předpokládat  $p_1 = 1$  a tedy  $p_2 < 1$ . Nyní je jediný persistentní stav paměti  $H_1 H_1$  a tedy zřejmě  $W(R', 1, p_2) = W(R', p_2, 1) = 1$ , což souhlasí s (2.14).

### 3. PŘEHLED DOSAVADNÍHO STAVU ŘEŠENÍ A FORMULACE VLASTNÍHO VÝSLEDKU

Stejněměrně nejlepší pravidlo je nalezeno dosud jen pro nejjednodušší případ  $r = 1$ , kdy 4 možné výsledky jednotlivého pokusu (1.1) jsou zároveň všemi možnými stavy paměti. Lze tedy definovat právě 16 různých pravidel a vzájemným porovnáním jejich cen zjistit, že *stejněměrně nejlepší pravidlo pro délku paměti  $r = 1$  je právě jedno, a to  $R_1^0 \equiv R_1^1$*  (s vyznačenými stavy  $T_2, T_1$ ).

Pro  $r > 1$  není naproti tomu dokázána existence stejněměrně nejlepšího pravidla. V článku [1], který je první prací zabývající se tímto problémem, definoval Robbins pravidlo  $R_r^0$  a vyslovil hypotézu, že je stejněměrně nejlepší. V další práci [2] uvažoval Isbell pravidlo  $R_r^1$  a dokázal tato tvrzení:

(1) *Platí*

$$(3.1) \quad W(R_r^1, p_1, p_2) \geq W(R_r^0, p_1, p_2) \quad \text{pro všechna } r, p_1, p_2$$

s rovností právě ve třech případech:

- $r \leq 2$  (neboť pak jsou obě pravidla totožná,  $R_r^0 \equiv R_r^1$ ).
- $p_1 = p_2$  (viz (2.3)).
- $\max(p_1, p_2) = 1$  (v tomto případě nabývají ceny obou pravidel hodnoty 1, jak lze snadno nahlédnout podobně jako na konci důkazu tvrzení (2.14)).

(2) *Platí*

$$(3.2) \quad W(R_r^1, p_1, p_2) \geq W(R_r, p_1, p_2),$$

kde  $R_r$  je libovolné pravidlo pro délku paměti  $r$ , alespoň ve dvou případech:

- $r = 1$  (viz výše uvedený výsledek, který je dokonce silnější!).
- $\min(p_1, p_2) = 0$ .

Isbellovu hypotézu, že pravidlo  $R_r^1$  je stejněměrně nejlepší, lze pro  $r \geq 3$  snadno vyvrátit např. pomocí pravidla, které má vyznačené stavy (1.3) a navíc další dva

$$(3.3) \quad \underbrace{T_i T_j T_j \dots T_j}_{r-1} \quad (i = 1, 2; j = 3 - i).$$

Lze totiž dokázat, že cena tohoto pravidla je na většině oboru  $\{p_1, p_2\}$  větší než  $W(R_r^1, p_1, p_2)$ . K efektivnějšímu výsledku v tomto směru však dospěli Smith a Pyke [3], kteří pro  $r \geq 3$  udali množinu pravidel *stejněměrně* lepších než  $R_r^1$ . V téže práci pak dále definují jistou ještě širší množinu pravidel, z nichž o jednom vyslovují hypotézu, že je stejněměrně nejlepší.

Dále uvedený vlastní příspěvek se naproti tomu týká případu  $r = 2$ , pro který bude dokázáno zesílené Isbellovo tvrzení výše označené (2), případ b) ve formě následující věty.



**Věta A.** Jestliže je  $R' \in \mathcal{R}_2$  (viz § 2),  $R$  pravidlo pro délku paměti  $r = 2$  nepatří do  $\mathcal{R}_2$  a

$$(3.4) \quad \min(p_1, p_2) = 0 < \max(p_1, p_2) < 1,$$

pak je

$$(3.5) \quad W(R', p_1, p_2) > W(R, p_1, p_2).$$

Jestliže některou ostrou nerovnost v (3.4) nahradíme neostrou, nutno totéž provést s nerovností (3.5) a věta A pak bude pouhým důsledkem Isbellova tvrzení (2) b). Pro případ první nerovnosti (3.4) je to patřno přímo z (2.3), a vzhledem k druhé nerovnosti (3.4) stačí uvažovat pravidlo  $R^*$  se dvěma vyznačenými stavy  $T_2 T_2, T_1 T_1$  a ověřit, že je

$$W(R', 1, p) = W(R^*, p, 1) = 1 \quad (R' \in \mathcal{R}_2, 0 \leq p \leq 1).$$

Z tvrzení obsažených ve větě A a následujících úvahách bezprostředně vyplývá

**Důsledek.** Při délce paměti  $r = 2$  buď množina  $\mathcal{R}_2$  představuje množinu všech stejnoměrně nejlepších pravidel, nebo žádné pravidlo není stejnoměrně nejlepší.

#### 4. DŮKAZ VĚTY A

K důkazu věty A, vyslovené v § 3, odvodíme nejprve některé potřebné pomocné výsledky. Uvažujme izolovaný blok hodů jedinou mincí, u nichž jsou možné výsledky  $H$  s pravděpodobností  $p$  a  $T$  s pravděpodobností  $q = 1 - p$  ( $0 \leq p \leq 1$ ). Předpis pro zakončení bloku je určen trojicí  $(I^0, I^H, I^T)$  podmnožin množiny  $U = \{H, T\}$  takto:

1. Blok končí již prvním hodem, jestliže jeho výsledek patří do  $I^0$ .
2. Nebyl-li blok ukončen  $n$ -tým hodem nebo dříve, pak je ukončen  $(n + 1)$ -vým hodem, jestliže buď výsledek  $n$ -tého hodu je  $H$  a výsledek  $(n + 1)$ -vého patří do  $I^H$ , nebo výsledek  $n$ -tého hodu je  $T$  a výsledek  $(n + 1)$ -vého patří do  $I^T$  ( $n \geq 1$ ).

Střední délku takového bloku označíme  $B_p(I^0, I^H, I^T)$ . Např. na konci § 2 jsme určili

$$(4.1) \quad B_p(\{T\}, \emptyset, \{T\}) = q^{-2}.$$

Vztah této definice k pravidlům (při délce paměti  $r = 2$ ) je patřný odtud, že každé pravidlo je charakterisováno osmi množinami

$$(4.2) \quad I_{Hi}^0, I_{Ti}^0, I_i^H, I_i^T \subset U_i = \{H_i, T_i\} \quad (i = 1, 2),$$

kteří mají tento význam: Počínaje druhým pokusem v posloupnosti hodů je každý blok hodů mincí  $i$  řízen předpisem určeným trojicí  $(I^0, I_i^H, I_i^T)$ , kde je  $I^0 = I_{Hi}^0$  resp.  $I_{Ti}^0$  v závislosti na tom, zda předešlý blok hodů mincí  $j$  končil výsledkem  $H_j$  nebo  $T_j$

- 44 ( $i = 1, 2; j = 3 - i$ ). Množina  $\mathcal{R}_2$  je přitom charakterisována specifikací šesti množin (4.2), totiž

$$(4.3) \quad I_{T_i}^0 = \{T_i\}, \quad I_i^H = \emptyset, \quad I_i^T = \{T_i\} \quad (i = 1, 2),$$

kdežto zbývající dvojice  $I_{H_i}^0$  ( $i = 1, 2$ ) probíhá všech 16 možných kombinací a tím jsou vytvářena jednotlivá pravidla patřící do  $\mathcal{R}_2$ ; ve zvláštním případě  $I_{H_i}^0 = \emptyset$  ( $i = 1, 2$ ) dostáváme Robbinsovo-Isbellovo pravidlo  $R_2^0 \equiv R_2^1$ .

Pro bloky hodů jednou mincí zavedeme kromě středních délek  $B_p$  ještě pravděpodobnosti  $P_p(I^0, I^H, I^T)$  a  $Q_p(I^0, I^H, I^T)$ , že blok bude mít konečnou délku a výsledek posledního hodu bude  $H$  resp.  $T$ . Je-li střední hodnota  $B_p$  konečná, je  $P_p + Q_p = 1$ ; v opačném případě nutně platí, že s pravděpodobností 1 je délka bloku nekonečná a tedy  $P_p + Q_p = 0$ .

**Pomocná věta 1.** Je-li  $0 < p < 1$  a  $T \in I^T$ , pak je

$$(4.4) \quad B_p(\{T\}, I^H, I^T) < B_p(\{T\}, \emptyset, \{T\}) = q^{-2}$$

(pokud ovšem není zároveň  $I^H = \emptyset, I^T = \{T\}$ ).

Důkaz. Z definice snadno vyplývají rovnice pro určení středních hodnot  $B_p$  (s jejichž zvláštním případem jsme se setkali v (2.16))

$$(4.5) \quad \begin{aligned} B_p(\emptyset, I^H, I^T) &= 1 + pB_p(I^H, I^H, I^T) + qB_p(I^T, I^H, I^T), \\ B_p(\{H\}, I^H, I^T) &= 1 + qB_p(I^T, I^H, I^T), \\ B_p(\{T\}, I^H, I^T) &= 1 + pB_p(I^H, I^H, I^T), \\ B_p(U, I^H, I^T) &= 1. \end{aligned}$$

(V případě potřeby definujeme  $0 \cdot \infty = 0$ .) Řešíme-li tyto rovnice pro všech osm kombinací množin  $I^H, I^T$ , jichž se pomocná věta týká, získáme všechny hodnoty potřebné pro ověření vztahů (4.4), které představují celkem 7 nerovností a jednu rovnost (uvedenou již v (4.1)). Zbytek důkazu záleží tedy již jen v mechanických výpočtech, které zde provádět nebudeme.

**Pomocná věta 2.** Je-li  $0 < p < 1, I^0 \subset \{H\}$  a  $T \in I^T$ , pak je

$$(4.6) \quad \frac{B_p(I^0, I^H, I^T)}{1 + Q_p(I^0, I^H, I^T)} < q^{-2}.$$

Důkaz. Zde jde celkem o 16 nerovností, které se ověří analogicky jako u pomocné věty 1. použitím jednak soustavy rovnic (4.5), jednak soustavy rovnic, které rovněž vyplývají z definice:

$$(4.7) \quad \begin{aligned} Q_p(\emptyset, I^H, I^T) &= pQ_p(I^H, I^H, I^T) + qQ_p(I^T, I^H, I^T), \\ Q_p(\{H\}, I^H, I^T) &= qQ_p(I^T, I^H, I^T), \\ Q_p(\{T\}, I^H, I^T) &= q + pQ_p(I^H, I^H, I^T), \\ Q_p(U, I^H, I^T) &= q. \end{aligned}$$

Nyní již můžeme přikročit k vlastnímu důkazu věty A, kterou můžeme vyjádřit ve formě:

*Je-li  $0 < p < 1$ ,  $R' \in \mathcal{R}_2$ ,  $R \notin \mathcal{R}_2$ , pak platí*

$$(4.8) \quad W(R', p, 0) > W(R, p, 0).$$

Nejprve předpokládejme, že u pravidla  $R$  není některý ze stavů  $T_i T_i$  ( $i = 1$  nebo  $2$ ) vyznačený neboli  $T_i \notin I_i^T$  pro  $i = 1$  nebo  $2$ . Pak je ovšem  $f(R, T_1 T_1, 0, p) = 0$  nebo  $f(R, T_2 T_2, p, 0) = 0$  a tedy  $W(R, p, 0) = 0$ . Zároveň však je podle (2.14)

$$(4.9) \quad W(R', p, 0) = p(1 + q^2)^{-1} > 0$$

a tedy (4.8) platí. Můžeme proto nadále předpokládat

$$(4.10) \quad T_i \in I_i^T \quad (i = 1, 2)$$

neboli omezit se na pravidla  $R$ , mezi jejichž vyznačenými stavy jsou  $T_2 T_2$  a  $T_1 T_1$ . Pro určitost budeme předpokládat  $p_1 = p$ ,  $p_2 = 0$  a případ vyměněného číslování mincí při pravidlu  $R$  řešit pomocí *souměrně sdruženého* pravidla  $\bar{R}$ , jehož vyznačené stavy jsou určeny změnou indexů ve vyznačených stavech pravidla  $R$ . Platí-li pak (4.10) pro  $R$ , platí též pro  $\bar{R}$ , a je-li kromě toho  $R' \in \mathcal{R}_2$ , je též  $\bar{R}' \in \mathcal{R}_2$ .

Příslušný řetězec  $M_0$  má jedinou uzavřenou třídu persistentních stavů (neboť z každého stavu  $\alpha$  je zřejmě dosažitelný stav  $T_2 H_1$ ) a proto nebudeme nikde vyznačovat počáteční stav nebo uzavřenou třídu. Z definice vyplývá

$$(4.11) \quad W(R, p, 0) = \min [f(R, p, 0), f(\bar{R}, p, 0)]$$

a pro střední délky bloků a pravděpodobnosti jejich posledních výsledků platí zřejmé vztahy

$$(4.12) \quad 1 \leq B_p < \infty,$$

$$(4.13) \quad 1 \leq B_0 \leq 2,$$

$$(4.14) \quad P_p + Q_p = 1,$$

$$(4.15) \quad P_0 = 0, \quad Q_0 = 1.$$

S pravděpodobností 1 se tedy bloky hodů oběma mincemi střídají a před druhým a dalším blokem bodů mincí 1 je výsledek  $T_2$ , takže tyto bloky jsou nezávislé na před-

46 písu  $I_{H_1}^0$ . Střední délky bloků obojího druhu jsou proto určeny vztahy

$$(4.16) \quad B_1(R, p, 0) = B_p(I_{T_1}^0, I_1^H, I_1^T),$$

$$(4.17) \quad B_2(R, p, 0) = P_p(I_{T_1}^0, I_1^H, I_1^T) B_0(I_{H_2}^0, *, I_2^T) + Q_p(I_{T_1}^0, I_1^H, I_1^T) B_0(I_{T_2}^0, *, I_2^T).$$

Znak \* vyjadřuje zřejmou skutečnost, že  $B_0$  nezávisí na předpisu  $I_2^H$ . Podle (2.12) dostáváme

$$(4.18) \quad f(R, p, 0) = \frac{pB_1(R, p, 0)}{B_1(R, p, 0) + B_2(R, p, 0)} = p\{1 + [g(R, p, 0)]^{-1}\}^{-1},$$

kde je

$$(4.19) \quad g(R, p, 0) = \frac{B_1(R, p, 0)}{B_2(R, p, 0)} = \frac{B_p(I_{T_1}^0, I_1^H, I_1^T)}{P_p(I_{T_1}^0, I_1^H, I_1^T) B_0(I_{H_2}^0, *, I_2^T) + Q_p(I_{T_1}^0, I_1^H, I_1^T) B_0(I_{T_2}^0, *, I_2^T)}.$$

K důkazu tvrzení týkajícího se nerovností (4.8) stačí tedy již jen dokázat toto tvrzení:

*Je-li  $0 < p < 1$ ,  $R' \in \mathcal{H}_2$ ,  $R \notin \mathcal{H}_2$  s vlastností (4.10), pak je*

$$(4.20) \quad \min [g(R, p, 0), g(\bar{R}, p, 0)] < q^{-2},$$

$$(4.21) \quad g(R', p, 0) = q^{-2}.$$

Důkaz provedeme pro jednotlivé případy, které ve svém souhrnu vyčerpávají všechny možnosti.

1.  $I_{T_1}^0 = U_1$ . — Ve (4.19) je čítec podle (4.5) roven 1 a jmenovatel je podle (4.13), (4.14) roven alespoň jedné. Je tedy  $g(R, p, 0) \leq 1 < q^{-2}$  a (4.20) platí.

2.  $I_{T_1}^0 = \{T_i\}$  ( $i = 1, 2$ ) a přitom  $I_1^H \neq \emptyset$  nebo  $I_1^T \neq \{T_i\}$ . — Podle pomocné věty 1. je čítec (4.19) menší než  $q^{-2}$  a jmenovatel je (viz bod 1.) roven alespoň 1, takže (4.20) platí.

3.  $I_{T_i}^0 = \{T_i\}$ ,  $I_i^H = \emptyset$ ,  $I_i^T = \{T_i\}$  ( $i = 1, 2$ ). — Podle (4.3) patří pravidlo do  $\mathcal{H}_2$ . Čítec (4.19) je podle pomocné věty 1. roven  $q^{-2}$ . Dále zřejmě platí

$$(4.22) \quad Q_p(\{T\}, \emptyset, \{T\}) = 1 - P_p(\{T\}, \emptyset, \{T\}) = 1, \quad B_0(\{T\}, *, *) = 1,$$

takže jmenovatel (4.19) je roven 1. Platí tedy (4.21).

4.  $I_{T_1}^0 = \{T_1\}$ ,  $I_{T_2}^0 \subset \{H_2\}$  a přitom  $I_1^H \neq \emptyset$  nebo  $I_1^T \neq \{T_1\}$ . — Týž důkaz jako pro bod 2.

5.  $I_{T_1}^0 = \{T_1\}$ ,  $I_1^H = \emptyset$ ,  $I_1^T = \{T_1\}$ ,  $I_{T_2}^0 \subset \{H_2\}$ . — Z předpokladu vyplývá vzhledem k (4.13)

$$(4.23) \quad B_0(I_{T_2}^0, *, *) = 2.$$

Podle (4.22) a (4.23) je jmenovatel (4.19) roven 2. Odtud a podle pomocné věty 1. dostáváme opět (4.20).

6.  $I_{Ti}^0 \subset \{H_i\}$  ( $i = 1, 2$ ). – Podle (4.19), (4.13), (4.14) a (4.23) platí v tomto případě

$$g(R, p, 0) \leq \frac{B_p(I_{T1}^0, I_1^H, I_1^T)}{1 + Q_p(I_{T1}^0, I_1^H, I_1^T)}$$

a odtud (4.20) podle pomocné věty 2.

Ve všech zbývajících případech nutno změnit číslování mincí (přejít k souměrnému pravidlu  $\bar{R}$ ), čímž se dojde ke splnění předpokladů některého z bodů 1., 2., 4., 5. a (4.20) se dokáže prostřednictvím veličiny  $g(\bar{R}, p, 0)$ .

Tim je věta A zcela dokázána.

(Došlo dne 7. února 1966.)

#### LITERATURA

- [1] H. Robbins: A sequential decision problem with a finite memory. *Proc. Nat. Acad. Sci.* 42 (1956), 920–923.
- [2] J. R. Isbell: On a problem of Robbins. *Ann. Math. Statist.* 30 (1959), 606–610.
- [3] C. V. Smith, R. Pyke: The Robbins-Isbell two-armed-bandit problem with finite memory. *Ann. Math. Statist.* 36 (1965), 1375–1386.
- [4] W. Feller: *An Introduction to Probability Theory and its Applications* Vol. 1 (2. vyd.). J. Wiley, New York 1957.

---

#### SUMMARY

### On the Robbins-Isbell Sequential Decision Problem with a Finite Memory

ZDENĚK REŽNÝ

The following problem proposed by Robbins [1] and Isbell [2] is dealt with: An experimenter performs an infinite sequence of coin tosses, having at his disposal two coins numbered by him 1 and 2 and choosing one of them for each toss with the aim of maximizing the limiting frequency of heads. However, he knows nothing about the probabilities  $p_1, p_2$  of heads on the two coins, and moreover, his knowledge of the past history of the process of coin tossing is limited by a finite memory of length  $r$  ( $\geq 1$ ). Beginning from the  $(r + 1)$ -st toss, the choice of the coin is governed by a rule, which prescribes the choice of the coin depending on the memory state, or, equivalently, divides the set of all  $4^r$  memory states into two classes. By the worth

48 of a rule is meant the minimal limiting frequency of heads which is guaranteed by using the rule, for arbitrary initial memory state and numbering of coins. The problem consists of finding a uniformly best (u.b.) rule, which, given the memory length  $r$ , maximizes the worth in the whole domain  $\{p_1, p_2 : 0 \leq p_i \leq 1, i = 1, 2\}$ . For the case  $r = 1$ , the solution is easily found (cf. [2]): Among all 16 existing rules, exactly one is u. b., viz. that which prescribes a change of coins if and only if the present toss results in tails. For the cases  $r > 1$ , however, the solution of the problem has not yet been given.

In the present paper, the case of memory length  $r = 2$  is treated. The class  $\mathcal{R}_2$  of 16 rules, including the Robbins-Isbell rule ("change coins if and only if two subsequent tails occur") is defined by the property that all rules of  $\mathcal{R}_2$  are equivalent with respect to worth. It is shown that the worth of each rule of  $\mathcal{R}_2$  is greater than that of an arbitrary rule not belonging to  $\mathcal{R}_2$  at least in the domain  $\{p_1, p_2 : \min(p_1, p_2) = 0 < \max(p_1, p_2) < 1\}$ . Consequently, either  $\mathcal{R}_2$  is the class of all u. b. rules, or no rule for memory length 2 is u. b.

*Ing. Zdeněk Režný, CSc., Státní výzkumný ústav pro stavbu strojů, Husova 8, Praha 1 – Staré Město.*