Učenie sa basketbalových trikov s použitím optimalizácie trajektórie a hlbokého zosilneného učenia

Autor: Miroslav Kobliška

O clanku **Learning Basketball Dribbling Skills Using Trajectory Optimization and Deep Reinforcement Learning**

LIBIN LIU, DeepMotion Inc., USA
JESSICA HODGINS, Carnegie Mellon University, USA

Úvod

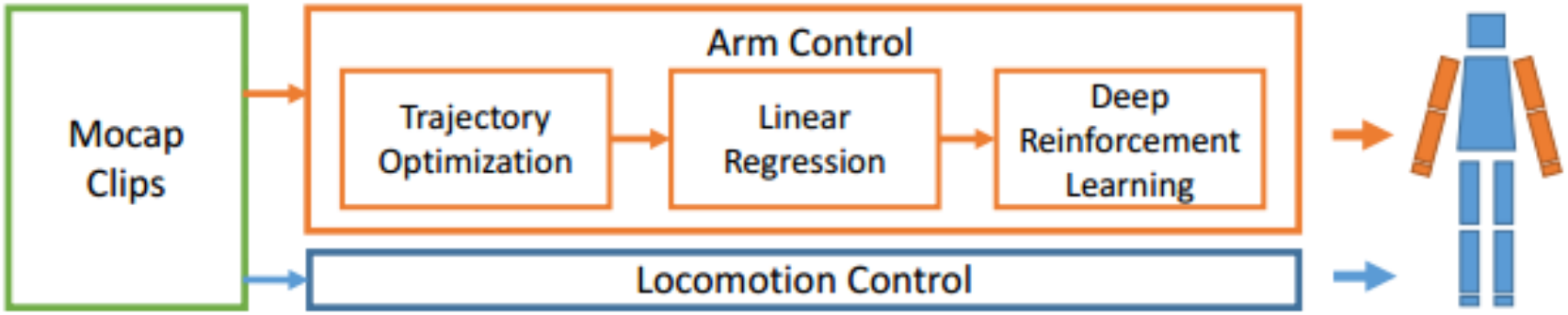
- V tejto práci animovali basketbal pomocou motion capture dát
- Používali aj fyzikálne vlastnosti na realistické prevedenie
- Táto metóda spracováva basketbalový kontrolér ako kombináciu kontroly pohybu a kontroly častí paže.

Úvod 2

- Systém sa najskôr naučí kontrolovať pohyb a potom natrénuje kontrolu paží
- Vie sa naučiť rôzne basketbalové triky ako :
- dribling medzi nohami aj krížne pohyby a vie reagovať na interakciu používateľa

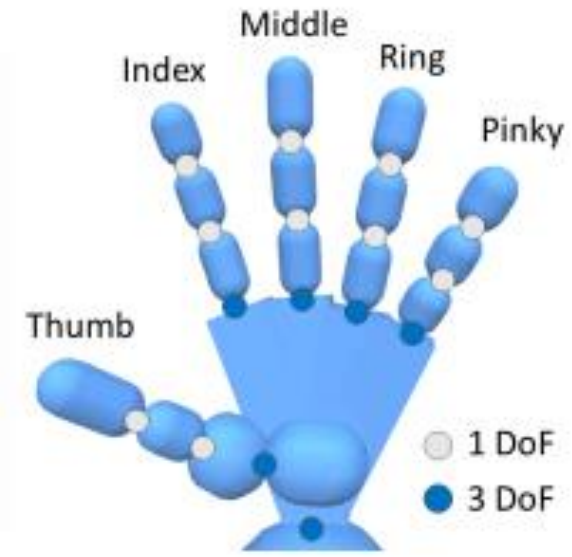
Prehľad systému

- klipy na snímanie pohybu ako vstup
- riadiaca jednotka sa skladá z dvoch spojených komponentov
- komponent kontrolujúci paže
- komponent kontrolujúci pohyblivosť



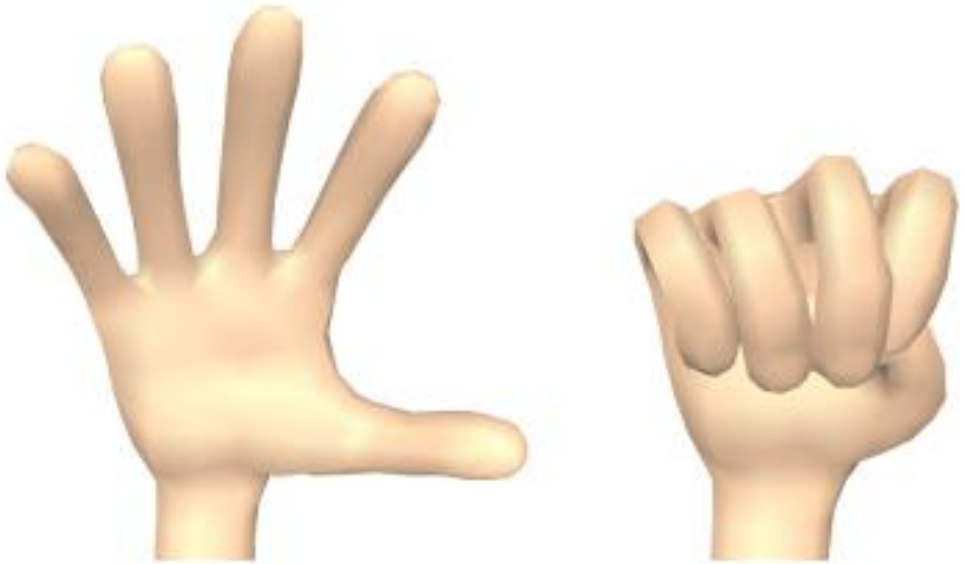
Kostra

- Umelá kostra
- Skladá sa z pohyblivých a vnútorných kĺbov
- Hráčove ruky sú namodelované s prstami



Prsty

$$\hat{\theta}_i = (1 - \lambda_i)\theta_i^{\text{flat}} + \lambda_i \theta_i^{\text{fist}}$$



(b) The flat hand pose (left) and the fist pose (right) used for computing the target pose for hands.

Pohyby prstov

- systém kontroluje každú ruku pomocou troch kontrolných signálov

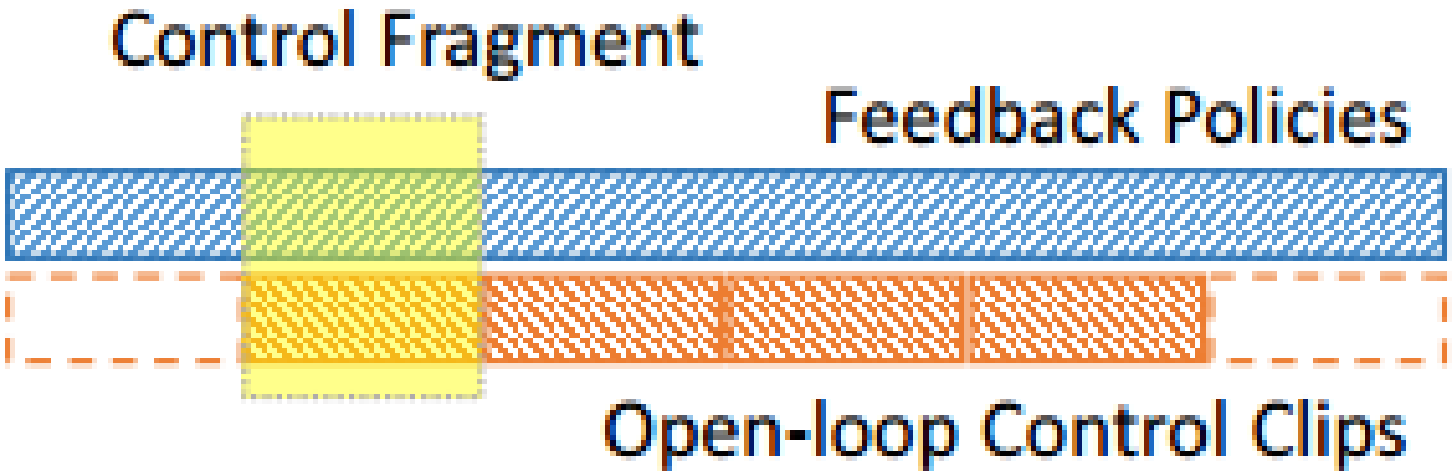
$$\boldsymbol{\alpha} = \{\lambda_{\text{Thumb}}, \lambda_{\text{Index}}, \lambda_{\text{Pinky}}\}$$

$$\lambda_{\text{Middle}} = \frac{2}{3}\lambda_{\text{Index}} + \frac{1}{3}\lambda_{\text{Pinky}}$$

$$\lambda_{\text{Ring}} = \frac{1}{3}\lambda_{\text{Index}} + \frac{2}{3}\lambda_{\text{Pinky}}$$

Kontrolné fragmenty

- Každý komponent na ovládanie má spätnú väzbu a „open-loop“ kontrolovanú trajektóriu



Učenie sa kontroly pohyblivosti

- hráč vie kontrolovať loptu iba počas veľmi krátkeho času
- používa sa kontrolný fragment dĺžky 0.05 s
- keď je naučený komponent pre kontrolu pohyblivosti je fixný počas učenia komponentu kontroly paží

Učenie kontroly paží

- používa optimalizáciu trajektórie na úspešné vypočítanie „open-loop“ kontroly paží
- „open-loop“ ovládanie paží je cielené na plece, lakeť a zápästie

Optimalizácia trajektórie

- nájdenie množiny korekčných kompenzácií, aby hráč mohol úspešne driblovať s loptou
- frejmy, v ktorých sa lopta dotýka hráčovej ruky
- Teda určité frejmy sú ako „chackpointy“
- Tu je vynútený dotyk ruky s loptou
- „Chackpointy“ - body, kde lopta dosiahne maximálnu výšku

Kontrola trajektórie 2

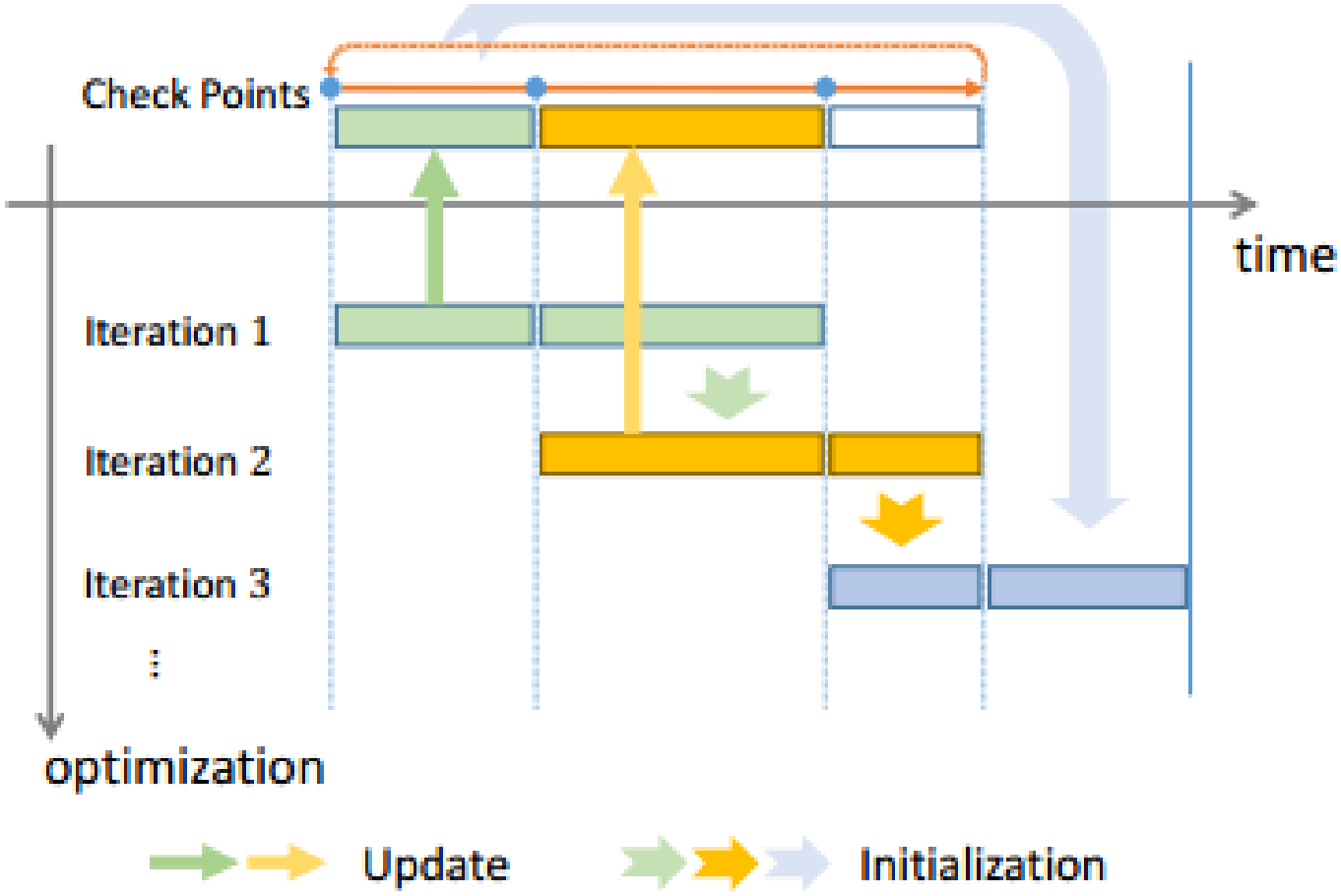
- Pre komplikované zručnosti, dva takého body
- Jeden na začiatku a druhý na konci otočky

Optimalizačné premenné

- množina $\mathcal{H} = \{\mathrm{H}_t\}, \quad \mathrm{H}_t \in \{\emptyset, \{\mathrm{L}\}, \{\mathrm{R}\}, \{\mathrm{L}, \mathrm{R}\}\}$
- $\{\chi_k^{\mathrm{L}}, \chi_k^{\mathrm{H}}\}$Použijeme ich na opravu „open-loop“ segmentov na ovládanie paží
- χ_k^{L} and χ_k^{H} Reprezentujú použité korekcie v tomto poradí: pre ľavú a pravú ruku

$$\chi = (\boldsymbol{q}_{\mathrm{shoulder}}, q_{\mathrm{elbow}}, \boldsymbol{q}_{\mathrm{wrist}}, \boldsymbol{\alpha}_{\mathrm{fingers}})$$

Optimalizačné premenné 2



Učenie lineárnej stratégie kontroly

- Stratégia kontroly pre paže
- väzme stavový vektor x ako vstup
- vypočíta akčný vektor $\boldsymbol{a} = (\chi^{\mathrm{L}}, \chi^{\mathrm{R}})$
- Stavový vektor sa skladá zo súčasného simulačného stavu *s* a indexu kontrolného fragmentu *k*

Učenie lineárnej stratégie kontroly 2

- *s* zachytáva stav hráča a lopty ako vektor:

$$s = \{p_{\text{ball}}, \dot{p}_{\text{ball}}, p_j, \dot{p}_j, d_i, c, \dot{c}, L\}$$

Lineárna regresia

- Na naučenie sa postupnej lineárnej stratégie kontroly budeme používať lineárnu regresiu

$$\pi(s, k) = M_k s + \hat{a}_k,$$

- pomocou regresie nevieme kontrolovať zložitejšie driblingové triky
- lineárnu stratégiu kontroly budeme používať na inicializáciu hlbokého procesu učenia.

Informácia o dotyku

- náš systém počíta priemernú polohu lopty v každom kontrolnom fragmente
- aktualizuje informácie o dotyku
- tak aby vzdialenosť medzi rukou a loptou bola menšia ako 5 cm

Hlboké zosilnené učenie

- Používame ho aby sme dosiahli robustné učenie basketbalových zručností.

ALGORITHM 1: Learn Arm Control Policy Using DDPG

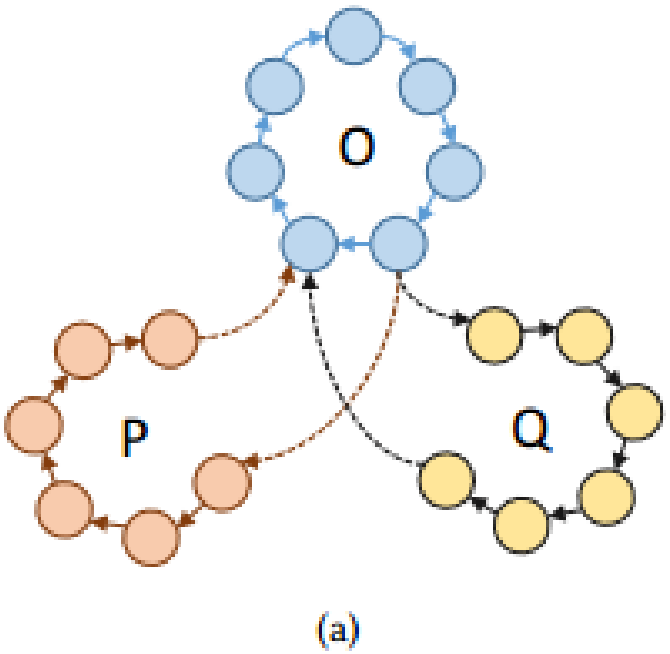
Input: control fragments $\{C_k\}$, $k = 1, \ldots, K$ and
associated linear control polices $\{(M_k, \hat{\boldsymbol{a}}_k)\}$
Input: starting states $\{\boldsymbol{s}_{k_i}\}$, $k_i \in \{1, \ldots, K\}$
Result: arm control policy π
initialize $\mathcal{D} \leftarrow \emptyset$
initialize critic network parameters θ_Q and actor network parameters θ_π
initialize target function $\theta'_Q \leftarrow \theta_Q$, $\theta'_\pi \leftarrow \theta_\pi$
initialize simulation with random starting state $\boldsymbol{s}_{k_i}$, set $\boldsymbol{x} \leftarrow (\boldsymbol{s}_{k_i}, k_i)$
for $t \leftarrow 1, 2, \ldots$ **do**
  $k \leftarrow k(\boldsymbol{x})$; // get the 'k' component of x
  if $t \leq N_{init}$ **then**
    $\boldsymbol{s} \leftarrow \boldsymbol{s}(\boldsymbol{x})$; // get the 's' component of x
    compute action $\boldsymbol{a} = M_k \boldsymbol{s} + \hat{\boldsymbol{a}}_k + \varepsilon_t$ and $r = r(\boldsymbol{x}, \boldsymbol{a})$
  else
    compute action $\boldsymbol{a} = \pi(\boldsymbol{x}; \theta_\pi) + \varepsilon_t$ and $r = r(\boldsymbol{x}, \boldsymbol{a})$
  end

Hlboké zosilnené učenie 2

execute control fragment C_k and observe $\boldsymbol{x}' \leftarrow (s', \text{next}(k))$
store transition $\boldsymbol{\tau} = (\boldsymbol{x}, \boldsymbol{a}, r, \boldsymbol{x}')$ in $\mathcal{D}$
if $t > N_{init}$ **then**
    sample a minibatch of N_{batch} transtions $\{\boldsymbol{\tau}_i\} \subset \mathcal{D}$
    compute target values $y_i = y(\boldsymbol{\tau}_i; \theta'_Q, \theta'_\pi)$ using Equation 13
    update θ_Q by minimizing the loss function of Equation 12
    update θ_π using the policy gradient of Equation 16
    $\theta'_Q \leftarrow (1 - \eta)\theta'_Q + \eta\theta_Q$
    $\theta'_\pi \leftarrow (1 - \eta)\theta'_\pi + \eta\theta_\pi$
end
if $\boldsymbol{x}' \in X_{term}$ **then**
    reset simulation to a random stating state $\boldsymbol{s}_{k_i}$
    $\boldsymbol{x} \leftarrow (\boldsymbol{s}, k_i)$
else
    $\boldsymbol{x} \leftarrow \boldsymbol{x}'$
end
end

Učenie kontrolných grafov

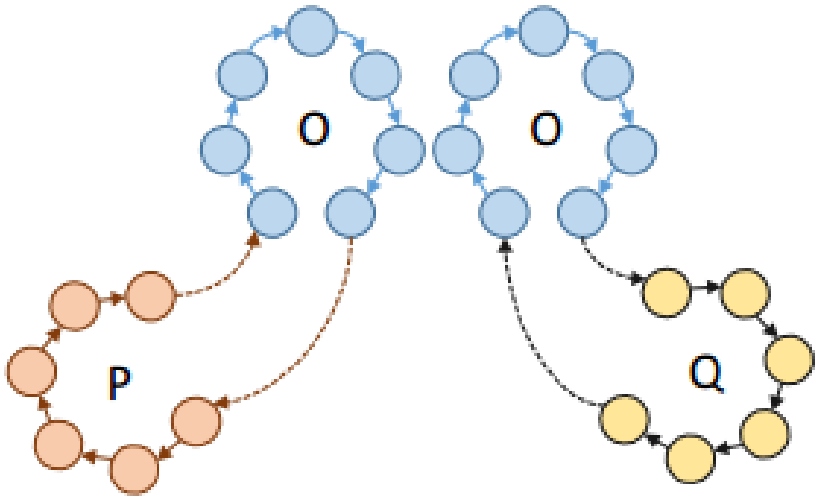
- Je to graf, ktorého uzly sú kontrolné fragmenty



(a)

Inkrementálne učenie

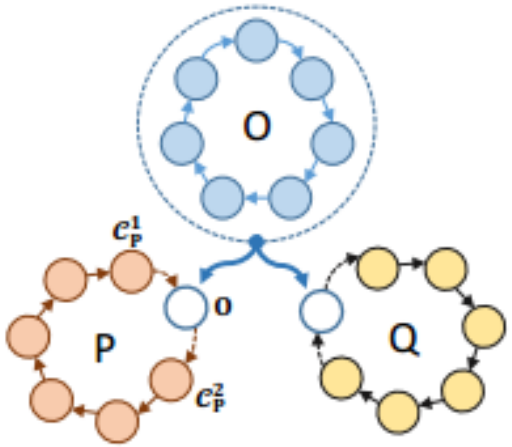
- systém sa učí kontrolu paží pre kontrolný graf inkrementálne
- Ak trénujeme necyklický trik P
- systém skombinuje triky P a O do cyklického pohybu (O+P)



(b)

Inkrementálne učenie 2

- Pri hlbokom zosilnenom učení najskôr natrénujeme nelineárnu korekčnú kompenzáciu pre trik O
- použijeme kombinovaný trik (O+P) na natrénovanie necyklického triku P



(c)

Výsledky

- Náš hráč má 1.8 m váži 76 kg
- basketbalová lopta má rádius 11,93 cm a váži 623.7 kg
- Systém vie bežať rýchlejšie ako beží reálny čas.

Cyklické zručnosti

(A)

(B)

(C)

(D)

(E)

Optimalizácia trajektórie

- optimalizujeme pohybovú sekvenciu, v ktorej hráč opakovane vykonáva zručnosti
- *Nrep = 100 cyklov*.
- Distribúcia vzoriek algoritmu CMA-ES (Covariance Matrix Adaption Evolution Strategy)je inicializovaná ako normálne rozdelenie $N(0, \sigma_0^2)$,
- σ_0 = 0.03 v prvom cykle
- σ_0 = 0.01 v ďalších cykloch

Optimalizácia trajektórie 2

- Končí ak

a) počet iterácii je väčší ako 1000
b) proces optimalizácie sa zastaví na 200 iterácii
c) priemerná vzdialenosť medzi hráčovou rukou a loptou menšia ako 1 cm pre (A – D) respektíve 2 cm pre E

Lineárna stratégia kontroly.

- postup postupnej lineárnej stratégie kontroly paží sa naučí z optimalizovaných pohybových sekvencii pomocou lin. regresie
- hráč môže opakovane mávať rukami stovky krát pri prevedení zručnosti (A)
- Problém - nevieme vykonávať kontrolu nad komplexnejšími zručnosťami (B – E)

Hlboké zosilnenie učenia

- nelineárna stratégia kontroly paže s použitím DDPG algoritmu
- Táto nelineárne stratégia kontroly paží umožňuje kontrolu pre triky B - E

Kontrolné grafy

- dva kontrolné grafy
- každý obsahuje jeden cyklický a dva necyklické triky

Table 2. Robustness of noncyclic skills in the learned control graphs.

control graph	noncyclic skill	success rate
in-place skills	dribbling between the legs	99.4%
	dribbling behind the back	99.6%
running skills	front crossover	98.1%
	spin crossover	95.7%

Robustnosť trikov

- nechali sme hráča opakovane vykonávať cyklickú aj necyklickú zručnosť
- pravdepodobnosť prechodu z cyklického na necyklický trik bola 0.5

Ďakujem za pozornosť