TARTU ÜLIKOOL
Loodus- ja täppisteaduste valdkond
Arvutiteaduse instituut
Informaatika õppekava

Andre Litvin

Alternatiiv keskväärtuse maksimeerimisele

Bakalaureusetöö (9 EAP)

Juhendaja: Raul Vicente, PhD

Tartu 2023

An alternative to expected reward maximization

Abstract:

Reinforcement learning problems can generally be described as follows. The user quantifies how good each state of some system would be according to their preferences and some agent, e.g. a robot, must choose actions that lead to states the user defined as good. More formally, for each state and action, the user picks a real-valued reward and the goal of reinforcement learning is to automatically find a strategy, called a policy, which would lead to a high reward sum.

However, actions often do not determine states, but only make some states likelier than others. In this case, the policy is usually chosen by maximizing the expected reward. However, in this thesis, I prove that for every probability $p < 1$ and constant $c > 0$, there exists a reinforcement learning problem where the policy maximizing expected reward gives reward sum Z°, but another policy would give reward sum Z, where $\mathbb{P}[Z > Z^\circ + c] > p$. In other words, the policy maximizing expected reward can get an arbitrarily smaller reward sum with arbitrarily high probability (except 1) compared to another policy.

This might not be a desirable property for a policy to have. In this thesis, I define the smoothened median of a random variable and prove that any policy that maximizes the smoothened median of the reward sum (instead of the expectation) does not have this property.

Keywords:
Reinforcement learning, median, heavy-tailed distributions, Kelly betting system

CERCS: P176

Alternatiiv keskväärtuse maksimeerimisele

Lühikokkuvõte:

Stiimulõppe ülesandeid saab üldjoontes kirjeldada järgmiselt. Stiimulõppe kasutaja kvantifitseerib vastavalt oma eelistustele, kui hea mingi süsteemi iga olek oleks, ning mingil agendil, näiteks robotil, tuleb valida tegevusi nii, et süsteem liiguks kasutaja defineeritud headesse olekutesse. Formaalsemalt kasutaja omistab igale olekule ja tegevusele mingi reaalarvulise auhinna ning stiimulõpe eesmärk on automaatselt leida strateegiat ehk eeskirja, mida järgides saaks agent kõrge auhindade summa.

Enamasti ei määra tegevuse valik aga üheselt olekut, vaid mõjutab üksnes erinevate olekute ning seega ka auhindade tõenäosuseid. Sel juhul võetakse tavaliselt eesmärgiks maksimeerida auhinnasumma keskväärtust. Kuid selles lõputöös tõestatakse, et iga tõenäosuse $p < 1$ ning konstandi $c > 0$ korral leidub stiimulõpe ülesanne, mille puhul auhinnasumma keskväärtust maksimeeriv eeskiri saab auhinnasumma Z°, aga mõni teine eeskiri saab auhinnasumma Z, kusjuures

$\mathbb{P}[Z > Z^\circ + c] > p$. Teiste sõnadega auhinnasumma keskväärtust maksimeeriv eeskiri võib saada ükskõik kui suure tõenäosusega (väljaarvatud 1) ning ükskõik kui suure konstandi võrra väiksema auhinnasumma kui mõni teine eeskiri.

Selline eeskirja omadus ei ole enamasti soovitav. Selles lõputöös defineeritakse juhusliku suuruse silutud mediaan ning tõestatakse, et auhinnasumma silutud mediaani maksimeerival eeskirjal ei ole sellist omadust.

Võtmesõnad:
Stiimulõpe, mediaan, raske sabaga jaotused, Kelly panustamissüsteem

CERCS: P176

Sisukord

1 Sissejuhatus

Stiimulõppe eesmärk on panna erinevaid *agente*, näiteks roboteid, automaatselt valima otstarbekaid tegevusi paljudes erinevates keskkondades. Stiimulõpe erineb muust masinõppest selle poolest, et agendile ei anta näiteid otstarbekatest tegevustest, vaid ainult antakse teada auhindade kaudu, milliseid olekuid tuleb üritada saavutada tegevuste valikuga.

Enamasti ei määra tegevuse valik aga üheselt järgmist olekut, vaid mõjutab üksnes erinevate olekute tõenäosuseid. Seetõttu ei ole enamasti võimalik kindel olla, kas üks tegevus viib suuremate auhindadeni kui teine. Sellisel juhul võetakse tegevuse valimiseks tavaliselt kriteeriumiks oodatavat auhinnasummat ehk võimalike auhinnasummade jaotuse keskväärtust ning maksimeeritakse seda (Sutton and Barto 2018, lk 53).

Selle töö eesmärk on näidata, et mõnede ülesannete puhul on mõistlik kasutada alternatiivset kriteeriumi, mis ei ole kirjeldatav auhinnasumma keskväärtusena. Selleks kirjeldatakse kõigepealt kahte teoreetilist ülesannet ning tõestatakse, et keskväärtust maksimeeriv agent saab nendes järjekindlalt väikse auhinnasumma. Seejuures simuleeritakse neid ülesandeid (lisas 7.3), et kinnitada, et matemaatiliselt tuletatud auhinnasumma jaotus ühtib empiiriliste tulemustega. Seejärel tutvustatakse alternatiivi keskväärtuse maksimeerimisele – silutud mediaani maksimeerimine. See on Kelly panustamissüsteemi (Kelly Jr. 1956) üldistus. Tõestatakse, et auhinnasumma silutud mediaani maksimeerija saab ülalpool mainitud kahe ülesandes paremaid tulemusi. Nimelt võib keskväärtuse maksimeerimine anda ükskõik kui suure tõenäosusega (väljaarvatud 1) ja konstandi võrra väiksema auhinnasumma kui silutud mediaani maksimeerimine, aga vastupidine ei kehti.

Lisaks annab silutud mediaan võimaluse lahendada ülesandeid, milles auhinnasummal on raske sabaga jaotus, millel ei pruugi keskväärtus leiduda. See on kasulik näiteks finantsis, kuna aktsiate hindadel on tihti raske sabaga jaotus (Mohtadi and Ruediger 2013).

2 Otsustusprotsessid

2.1 Mis on otsustusprotsess?

Otsustusprotsess on mudel olukorrast, kus mingil agendil tuleb teha otsuseid ning kus need otsused mõjutavad erinevate tulemuste tõenäosusi. Olukorda modelleeritakse *olekute*, *tegevuste*, *auhindade* ning nendevaheliste seoste kaudu. Selles töös vaadeldakse lihtsuse mõttes ainult lõpliku arvu diskreetsete ajasammudega otsustusprotsesse.

2.1.1 Juku jäätist söömas

Näide 2.1. (jäätis) Jukule meeldib kõige rohkem maasikajäätis, keskmiselt vanillijäätis ja kõige vähem šokolaadijäätis. Juku meelest on maasikajäätis 2 korda parem kui vanillijäätis ja vanillijäätis 3 korda parem kui šokolaadijäätis – näiteks 1 teelusikas vanillijäätist on parem kui 2 teelusikat šokolaadijäätist, aga 4 teelusikat šokolaadijäätist on parem kui 1 teelusikas vanillijäätist.

Juku sünnipäeval lastakse tal mängida järgmist mängu. Ta saab korduvalt valida, kas teha mündiviset või mitte. Kui ta valib mündivise, siis ta saab tõenäosusega 0,5 teelusika šokolaadijäätist ja tõenäosusega 0,5 teelusika maasikajäätist. Muidu saab ta kindlalt teelusika vanillijäätist.

Mängu saab modelleerida otsustusprotsessina, kus agent on Juku, olekud on valiku tegemine ja erinevate jäätiste saamine, tegevused on mündivise tegemine või mitte tegemine ning auhinnad vastavad erinevatele jäätiste tüüpidele.

2.1.2 Juku ruletilaua ääres

Näide 2.2. (rulett) Juku mängib ruletisarnast mängu. Mängu alguses on tal üks punkt. Enne igat keerutust saab ta valida, kui suure osa olemasolevatest punktidest panustada mustale – võib panustada ka murdarvu punkte. Kui tuleb must, saab Juku panuse võrra punkte juurde, muidu jääb tal panuse võrra punkte vähemaks. Erinevalt päris ruletist on mäng keskmiselt mängija kasuks – must tuleb iga kord sama tõenäosusega $p \in (0{,}5; 1)$. Mängu lõpus saab Juku iga punkti eest ühe euro.

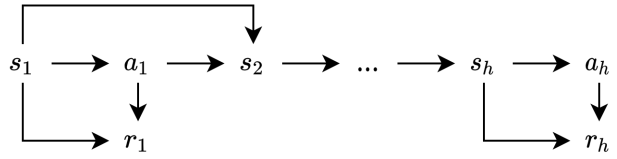
Mängu saab modelleerida otsustusprotsessina, kus agent on Juku, igale keerutusele ja võimalikule punktide arvule vastab üks olek, igale võimalikule panusele vastab üks tegevus ning auhind on punktide arv mängu lõpus.

Sellist mängu on uurinud juba (Kelly Jr. 1956), kuid mitte stiimulõppe kontekstis.

2.2 Otsustusprotsesside formaliseerimine

Kõige levinum stiimulõpe ülesannete formalisatsioon on Markovi otsustusprotsess (*Markov decision process*, edaspidi MDP), mida kirjeldab näiteks (Sutton and Barto 2018). Formaalselt on MDP 6-korteež $(\mathcal{S}, \mathcal{A}, u, r, h, s_1)$, kus $\mathcal{S}$ on olekute

hulk, $\mathcal{A}$ tegevuste hulk, $u\colon \mathcal{S}\times\mathcal{A}\times\mathcal{S}\to[0,1]$ *üleminekufunktsioon*, $r\colon \mathcal{S}\times\mathcal{A}\to\mathbb{R}$ *auhinnafunktsioon*, $h\in\mathbb{N}$ *horisont* ning $s_1\in\mathcal{S}$ algolek. [1]



Joonis 2.1. MDP kulg.

Protsessil on kokku h ajasammu. Iga ajasammu $t\in[1,h]\cap\mathbb{Z}$ alguses on agent olekus s_t. Kui agent valib tegevuse a_t, saab ta auhinna $r_t=r(s_t,a_t)$ ning liigub olekusse s' tõenäosusega $u(s_t,a_t,s')$, nagu joonisel 2.1. Kuna tõenäosuste summa üle üksteist välistavate võimaluste on 1 ning saab olla ainult üks järgmine olek, siis peab kehtima

$$\sum_{s'\in\mathcal{S}} u(s,a,s')=1.$$

Agendi eesmärk on maksimeerida auhinnasummat

$$\sum_{t=1}^{h} r(s_t,a_t).$$

Mõnikord ei ole igas olekus võimalik iga tegevus või iga järgmine olek. Siis tähistatakse võimalike tegevuste hulka olekus s tähisega $\mathcal{A}(s)\subseteq\mathcal{A}$ ning [2] võimalike järgmiste olekute hulk pärast tegevust $a\in\mathcal{A}(s)$ tähisega

$$\mathcal{S}(s,a)=\{s'\in\mathcal{S}\mid u(s_t,a_t,s')>0\}.$$

2.2.1 Jäätise söömise formaliseerimine

Näites 2.1 (jäätis) on Juku vaheldumisi olekus, kus ta valib, kas teha mündiviset – olgu see olek s^1 – ja mõnes olekus, kus ta saab jäätist. Olgu s^2, s^3, s^4 olekud, kus Juku saab vastavalt šokolaadi-, vanilli- või maasikajäätist. Olekute hulk on $\mathcal{S}=\{s^1,s^2,s^3,s^4\}$.

Olgu mündivise tegevus α^1 ja selle mitte tegemine α^2. Tegevuste hulk on $\mathcal{A}=\{\alpha^1,\alpha^2\}$.

Kui Juku on olekus s^1 ja teeb mündivise, liigub ta tõenäosusega 0,5 olekusse s^2 ja tõenäosusega 0,5 olekusse s^4; muidu liigub ta olekusse s^3. Olekust s^2, s^3 ja s^4 liigub Juku olekusse s^1 sõltumata tegevusest. Seega $u(s^1,\alpha^1,s^2)=$

[1] Erinevad stiimulõppe artiklid defineerivad MDP natuke erinevalt. See on üks võimalik definitsioon lõpliku horisondi korral.

[2] Siis $\mathcal{A}=\bigcup_{s\in\mathcal{S}}\mathcal{A}(s)$.

$u(s^1, \alpha^1, s^4) = 0{,}5$, $u(s^1, \alpha^2, s^3) = 1$ ning iga tegevuse a korral $u(s^2, a, s^1) = u(s^3, a, s^1) = u(s^4, a, s^1) = 1$, kus u on üleminekufunktsioon.

Olekus s^1 ei saa Juku jäätist, mistõttu selle auhind on 0. Oleku s^2 auhind peaks olema mõni positiivne reaalarv – olgu see näiteks 1. Kuna Jukule meeldib vanillijäätis 3 korda rohkem kui šokolaadijäätis ning maasikajäätis 2 korda rohkem kui vanillijäätis, siis oleku s^3 auhind on 3 ja s^4 auhind 6. Järelikult iga a korral $r(s^1, a) = 0$, $r(s^2, a) = 1$, $r(s^3, a) = 3$ ning $r(s^4, a) = 6$, kus r on auhinnafunktsioon.

Kuna Juku valikule järgneb alati jäätise saamine, siis ajasammude arv ehk horisont h peab olema mõni positiivne paarisarv. Algolek on $s_1 = s^1$.

2.2.2 Ruleti formaliseerimine

Näites 2.2 (rulett) vastab igale keerutusele ehk ajasammule t ning punktide arvule w üks olek s_t^w. Alguses on Jukul üks punkt, mistõttu algolek on $s_1 = s_1^1$.

Fikseerigem mõni horisont $h \geq 2$. Ilmselt saab Juku maksimaalset võimalikku punktide arvu siis, kui ta panustab kõik punktid ja võidab iga kord. Sel juhul oleks tema punktide arv $1, 2, \ldots, 2^{h-1}$. Kuna punktide arv on mittenegatiivne ja maksimaalselt 2^{h-1}, siis võimalike olekute hulk on

$$\mathcal{S} = \{s_t^w \mid w \in [0, 2^{h-1}], t \in [1, h] \cap \mathbb{Z}\}.$$

Minimaalselt saab Juku panustada 0 punkti ja maksimaalselt nii palju, kui tal on, ning igale panusele vastab üks tegevus. Seega olekus s_t^w võimalike tegevuste hulk on $\mathcal{A}(s_t^w) = [0, w]$.

Juku võidab panuse juurde kui tuleb must, mis juhtub tõenäosusega $p \in (0{,}5; 1)$. Kui Juku panustab a punkti olekus s_t^w, siis ta liigub tõenäosusega p olekusse s_{t+1}^{w+a} ja tõenäosusega $1 - p$ olekusse s_{t+1}^{w-a}. Seega iga $w \geq 0$, $t \in [1, h-1] \cap \mathbb{Z}$ ja $a \in [0, w]$ korral $u(s_t^w, a, s_{t+1}^{w+a}) = p$ ning $u(s_t^w, a, s_{t+1}^{w-a}) = 1 - p$.

Juku saab mängu lõpus ehk olekus s_h^w iga punkti eest ühe euro ehk w eurot, mistõttu $r(s_h^w, a) = w$ sõltumata tegevusest a. Kuna varasemaid auhindu ei ole, siis iga $t < h$ korral $r(s_t^w, a) = 0$.

2.3 Tegevuste valimine otsustusprotsessis

2.3.1 Eeskirjad ja trajektoorid

Üldiselt käitub MDP agent mingi *eeskirja* $\pi \colon \mathcal{S} \to \mathcal{A}$ järgi – olekus s teeb agent tegevuse $\pi(s)$.[3] MDP $(\mathcal{S}, \mathcal{A}, u, r, h, s_1)$ ja eeskiri π määravad koos protsessi *trajektoori* $((S_1, \ldots, S_h), (A_1, \ldots, A_h))$, kus S_t on olek ja A_t on tegevus ajasammul t juhusliku suurusena. Võimalike olekute jaotust määravad koos esimene olek ja

[3] Selles töös käsitletakse lihtsuse mõttes ainult deterministlikke eeskirju.

üleminekufunktsioon:

$$\begin{aligned}&\mathbb{P}[S_1 = s_1] = 1,\\&\forall t \in [1, h-1] \cap \mathbb{Z}, \forall s \in \mathcal{S}, \forall a \in \mathcal{A}, \forall s' \in \mathcal{S},\\&\mathbb{P}[S_{t+1} = s' \mid S_t = s, A_t = a] = u(s, a, s').\end{aligned}$$

MDP nimi tuleneb sellest, et suurustel S_t on *Markovi omadus*: Nad sõltuvad ainult eelmisest olekust S_{t-1} ja tegevusest A_{t-1}, mitte millestki varasemast. Näiteks kui $i < t$, siis iga $s' \in \mathcal{S}$ korral

$$\mathbb{P}[S_{t+1} = s' \mid S_t = s_t, A_t = \pi(S_t), S_i = s_i] = \mathbb{P}[S_{t+1} = s' \mid S_t = s_t, A_t = \pi(S_t)].$$

Tähistagu $(\cdots \mid \pi)$ tingimust, et agent valib tegevusi eeskirja π järgi:

$$(\cdots \mid \forall t \in [1, h] \cap \mathbb{Z}, A_t = \pi(S_t)).$$

Kompaktsuse mõttes defineeritakse juhusliku suurusena veel auhindade summa $Z^\pi_{t_1}(s)$, mida agent saab alates ajasammust t_1, kui ta on alguses olekus s ning järgib eeskirja π:

$$Z^\pi_{t_1}(s) = \left(\sum_{t=t_1}^{h} r(S_t, A_t) \,\middle|\, \pi, S_{t_1} = s \right).$$

2.3.2 Eeskirja valimine

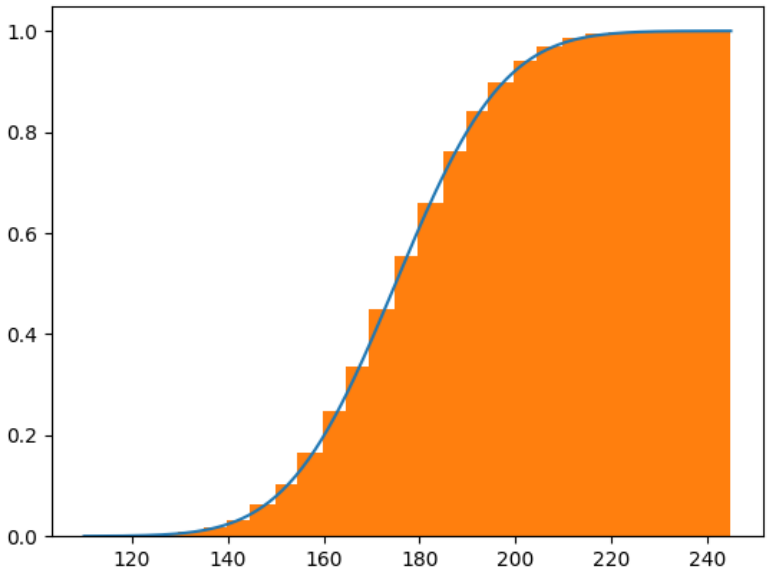
Kuigi agendi üldine eesmärk on maksimeerida auhinnasummat, eeskirja ei ole võimalik otse auhinnasumma järgi valida, sest auhinnasumma on juhuslik suurus, millel ei pruugi olla ühest maksimumi. Seetõttu on eeskirja valimisel tarvis kasutusele võtta mõni kriteerium, mille järgi saaks otsustada, milline auhinnasumma kahest võimalikust jaotusest oleks parem. Tavaliselt võetakse selleks kriteeriumiks keskväärtust (Sutton and Barto 2018, lk 53).

Väärtusfunktsioon $v^\pi_t \colon \mathcal{S} \to \mathbb{R}$ annab vastava auhinnasumma keskväärtust:

$$v^\pi_t(s) = \mathbb{E}\left[Z^\pi_t(s)\right].$$

Parimat eeskirja π° saab valida terve protsessi auhinnasumma keskväärtust maksimeerides:

$$\pi^\circ \in \operatorname*{argmax}_\pi v^\pi_1(s_1) = \operatorname*{argmax}_\pi \mathbb{E}\left[\sum_{t=1}^{h} r(S_t, A_t) \middle| \pi \right].$$



Joonis 2.2. Kelly panustaja simulatsiooni auhinnasummade kumulatiivse histogrammi võrdlus teoreetilise jaotusfunktsiooniga näites 2.1 (jäätis). Simulatsiooni horisont $h = 50$ ning simulatsiooni jooksutati 5000 korda. Siin x-telg näitab auhinnasummat ning y-telg näitab tõenäosust saada ülimalt nii suurt auhinnasummat.

2.3.3 Jäätise söömine keskväärtuse maksimeerimisega

Koondumine jaotuse järgi. Olgu $X_1, X_2, \ldots$ ja X reaalarvulised juhuslikud suurused. Olgu $F_1, F_2, \ldots$ ja $F\colon \mathbb{R} \to [0,1]$ nende vastavad jaotusfunktsioonid. Öeldakse, et suurused X_t koonduvad juhuslikuks suuruseks X jaotuse järgi ning kirjutatakse

$$X_t \xrightarrow{d} X,$$

kui iga $x \in \mathbb{R}$ korral

$$\lim_{t\to\infty} F_t(x) = F(x)$$

ehk iga $\varepsilon > 0$ korral leidub $t_1 \in \mathbb{N}$, mille puhul

$$\forall t \geq t_1, |F_t(x) - F(x)| < \varepsilon.$$

Olgu $\mathcal{N}(\mu, \sigma^2)$ normaaljaotus keskväärtusega μ ja dispersiooniga σ^2.

Keskne piirteoreem. Olgu $X_1, X_2, \ldots$ üksteisest sõltumatud, sama jaotusega ning lõpliku keskväärtuse ja dispersiooniga juhuslikud suurused. Kui $h \to \infty$, siis

$$\frac{1}{\sqrt{h \operatorname{Var}[X_1]}} \sum_{t=1}^{h} (X_t - \mathbb{E}[X_1]) \xrightarrow{d} N \sim \mathcal{N}(0;1).$$

Jäätise söömisel on ainult oluline, mis tegevust Juku valib olekus s^1. Sõltumata tegevuse valikust on ülejärgmine olek jälle s^1, mistõttu mõjutab tegevus ainult järgmist olekut. Terve protsessi auhinnasumma keskväärtus on olekute auhindade keskväärtuste summa, mistõttu on ainult tarvis vaadelda mõju auhindadele ainult ühes olekus s^1.

Tegevuse a valik olekus s^1 mingil ajasammul ei mõjuta selle ajasammu enda auhinda – igal juhul $r(s^1, a) = 0$. Kui $a = \alpha^1$, siis sõltumatu järgmisest tegevusest a' on järgmise ajasammu auhind $r(s^2, a') = 1$ või $r(s^4, a') = 6$, mille keskväärtus on $u(s^1, \alpha^1, s^2) \cdot r(s^2, a') + u(s^1, \alpha^1, s^4) \cdot r(s^4, a') = 3{,}5$.

Kui aga $a = \alpha^2$, siis $u(s^1, \alpha^2, s^3) = 1$ ning $r(s^3, a') = 3$. Seega auhinnasumma keskväärtust $v_1^\pi(s_1)$ maksimeerivad eeskirjad π, mille puhul $\pi(s^1) = \alpha^1$. Teisisõnu on Jukul keskmiselt parem iga kord valida mündiviset.

Sellist eeskirja π järgides on igal teisel ajasammul t saadud auhinnad üksteisest sõltumatud, sama jaotusega ning lõpliku keskväärtuse ja dispersiooniga juhuslikud suurused $r(S_t, A_t)$. Keskse piirteoreemi kohaselt auhinnasumma jaotus läheneb normaaljaotusele:

$$\frac{1}{\sqrt{\frac{h}{2} \operatorname{Var}[r(S_2, A_2)]}} \left(Z_1^\pi(s_1) - \frac{h}{2} \mathbb{E}[r(S_2, A_2)] \right) \xrightarrow{d} N \sim \mathcal{N}(0;1).$$

Joonis 2.2 näitab head vastavust selle teoreetilise jaotuse ja simulatsiooni tulemuste vahel. Simulatsiooni kood on lisas 7.3.

3 Kriteeriumi olulisus

3.1 Keskväärtuse maksimeerimise puudujäägid

Keskväärtus on levinud kriteerium, kuid näite 2.2 (rulett) parim eeskiri keskväärtuse järgi paneb kahtlema, kas see on alati sobilik. Probleem seisneb selles, et mäng on keskmiselt Juku kasuks, mistõttu suurema arvu punktide panustamine alati suurendab keskväärtust.

Teoreem 3.1. Näites 2.2 (rulett) maksimeerib keskväärtust $v_1^{\pi}(s_1)$ eeskiri π°, mis panustab igas olekus (ja igal ajasammul) kõik seni kogutud punktid:

$$\forall s_t^w \in \mathcal{S}, \pi^{\circ}(s_t^w) = w.$$

Tõestuse skeem. Eelviimases olekus s_{h-1}^w on auhinnasumma keskväärtus

$$v_{h-1}^{\pi}(s_{h-1}^w) = p_w(w + \pi(s_{h-1}^w)) + (1 - p_w)(w - \pi(s_{h-1}^w)),$$

mida maksimeerib iga w korral valik $\pi(s_{h-1}^w) = w$, sest $p_w > 0{,}5$. Seejuures $v_{h-1}^{\pi^{\circ}}(s_{h-1}^w) = p_w \cdot 2w + 0$. Kui ajasammul $h - t$ iga w korral

$$v_{h-t}^{\pi^{\circ}}(s_{h-t}^w) = 2^t p_w^t w,$$

siis ka eelmisel ajasammul, $h - t - 1$, kehtib

$$v_{h-t-1}^{\pi^{\circ}}(s_{h-t-1}^w) = 2^{t+1} p_w^{t+1} w.$$

Induktsioon tagurpidi üle ajasammude lõpetab tõestuse. Täistõestus on lisas 7.1. ■

Kui Juku panustab alati kõik punktid, siis ta peab võitma iga kord, et tal oleks mängu lõpus positiivne arv punkte. Kuna võidud on sõltumatud üksteisest, üks võit juhtub tõenäosusega p_w ning kokku on $h - 1$ võimalust võita või kaotada, siis tõenäosus võita iga kord on p_w^{h-1}. Horisondi kasvades läheneb see tõenäosus nullile.

Sõltuvalt Juku eelistustest on muidugi võimalik, et ta siiski valiks sellise eeskirja, aga ilmselt võib vähemalt mõnedel stiimulõppe kasutajatel tekkida soov vältida eeskirju, mille puhul positiivse auhinnasumma saamise tõenäosus läheneb nullile. Tavalised stiimulõppe algoritmid ei paku aga selleks võimalust.

3.2 Kelly panustamine

Et sellist tulemust vältida, oli Kelly panustamise idee maksimeerida auhinnasumma enda keskväärtuse asemel punktide *kasvukiiruse* keskväärtust (Kelly Jr. 1956). Üldiselt kui toimub mingi eksponentsiaalne kasv $x \mapsto e^{kx}$, siis kasvukiirust kirjeldab arv

$$k = \ln\left(\frac{e^{k(x+1)}}{e^{kx}}\right).$$

Teoreemis 3.1 esines mingi eksponentsiaalne kasv, nii et äkki ka punktide kasvukiirust kirjeldab arv

$$\ln\left(\frac{w'}{w}\right),$$

kus w ja w' on punktid järjestikustel ajasammudel t ja $t+1$ ehk $S_t = s_t^w$ ning $S_{t+1} = s_{t+1}^{w'}$. Kui Juku panustab a punkti ajasammul t, siis w' on $w+a$ tõenäosusega p_w ning muidu $w-a$. Järelikult kasvukiiruse keskväärtus on

$$p_w \cdot \ln\left(\frac{w+a}{w}\right) + (1-p_w) \cdot \ln\left(\frac{w-a}{w}\right)$$

ehk

$$p_w \cdot \ln(w+a) + (1-p_w) \cdot \ln(w-a) - \ln(w).$$

Kui sellel on maksimum, siis maksimumis tuletis a suhtes on 0:

$$\frac{p_w}{w+a} - \frac{1-p_w}{w-a} = 0.$$

Kui $a \neq w$, siis $p_w(w-a) = (1-p_w)(w+a)$ ning $p_w w - p_w a = w - p_w w + a - p_w a$ ja $2p_w w - w = a$ ehk $a = (2p_w - 1)w$. See vihjab järgmisele strateegiale: igal ajasammul panustada $(2p_w - 1) \cdot 100\%$ kogutud punktidest.

3.2.1 Miks Kelly panustamine paremini toimib?

Olgu üldisemalt π^f eeskiri, mis panustab $f \cdot 100\%$ punktidest igal ajasammul, kus $f \in (0,1)$:

$$\pi^f(s_t^w) = fw.$$

Kelly panustaja eeskiri on π^f, kus $f = 2p_w - 1$.

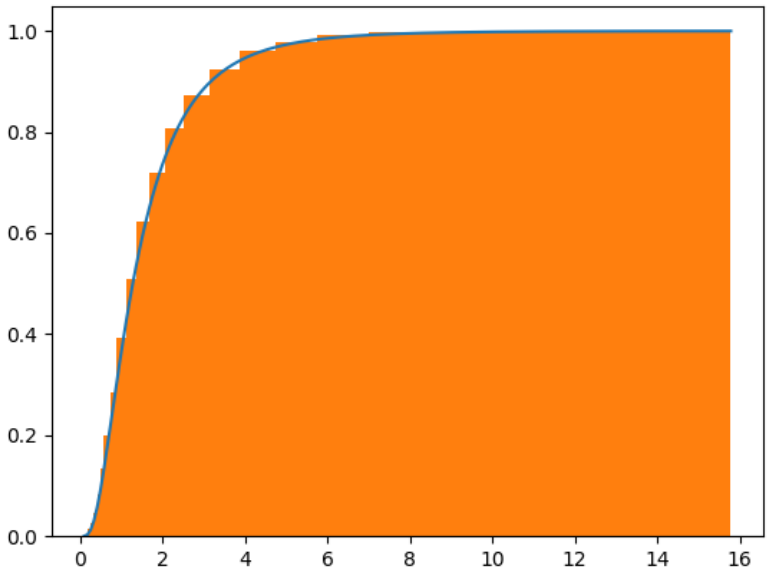
Olgu W_t punktide arv ajasammul t; $W_t = w$ parajasti siis, kui $S_t = s_t^w$. Olgu B_t juhuslik suurus, mis on 1 kui agent võitis ajasammul t ja muidu -1; $B_t = 1$ parajasti siis, kui $W_{t+1} > W_t$. Seejuures juhuslikud suurused $B_1, \ldots, B_{h-1}$ on kõik sama jaotusega ning sõltumatud üksteisest ja eeskirjast.

Lemma 3.2. Kui agent järgib eeskirja π^f, siis

$$W_h = \prod_{t=1}^{h-1}(1 + B_t f).$$

Tõestus. Kui $t_2 = 1$, siis $W_{t_2} = 1$, sest $S_1 = s_1 = s_1^1$. Kui mingi $t_2 < h$ korral

$$(W_{t_2} \mid \pi^f) = \prod_{t=1}^{t_2-1}(1 + B_t f),$$



Joonis 3.1. Kelly panustaja simulatsiooni auhinnasummade kumulatiivse histogrammi võrdlus teoreetilise jaotusfunktsiooniga näites 2.2 (rulett). Siin horisont $h = 50$, $p_w = 0{,}55$ ning simulatsiooni jooksutati 5000 korda. Simulatsiooni kood on lisas 7.3. Siin x-telg näitab auhinnasummat ning y-telg näitab tõenäosust saada ülimalt nii suurt auhinnasummat.

siis

$$\begin{aligned}(W_{1+t_2} \mid \pi^f) &= (W_{t_2} + B_{t_2} A_{t_2} \mid \pi^f) \\ &= (W_{t_2} + B_{t_2} f W_{t_2} \mid \pi^f) \\ &= (1 + B_{t_2} f)(W_{t_2} \mid \pi^f) \\ &= (1 + B_{t_2} f) \prod_{t=1}^{t_2-1} (1 + B_t f) \\ &= \prod_{t=1}^{t_2} (1 + B_t f).\end{aligned}$$

Väide järeldub induktsioonist üle ajasammu t_2. ■

Teoreem 3.3. Näites 2.2 (rulett) saab agent eeskirjaga π^f auhinnasumma

$$Z_1^{\pi^f}(s_1) = \exp\left(\mu(h-1) + N_h \sigma \sqrt{h-1}\right),$$

kus $\mu = \mathbb{E}[\ln(1 + B_1 f)]$ ja $\sigma = \sqrt{\mathrm{Var}[\ln(1 + B_1 f)]}$ ning $N_h \xrightarrow{d} N \sim \mathcal{N}(0;1)$, kui $h \to \infty$.

Tõestus. Lemma 3.2 tõttu

$$\begin{aligned}Z_1^{\pi^f}(s_1) &= (r(S_h, A_h) \mid \pi^f) \\ &= (W_h \mid \pi^f) \\ &= \prod_{t=1}^{h-1} (1 + B_t f) \\ &= \exp\left(\ln\left(\prod_{t=1}^{h-1} (1 + B_t f)\right)\right) \\ &= \exp\left(\sum_{t=1}^{h-1} \ln(1 + B_t f)\right) \\ &= \exp\left(\mu(h-1) + \frac{\sigma\sqrt{h-1}}{\sigma\sqrt{h-1}} \sum_{t=1}^{h-1} (\ln(1 + B_t f) - \mu)\right) \\ &= \exp\left(\mu(h-1) + N_h \sigma \sqrt{h-1}\right),\end{aligned}$$

kus

$$N_h = \frac{1}{\sigma\sqrt{h-1}} \sum_{t=1}^{h-1} (\ln(1 + B_t f) - \mu).$$

Keskse piirteoreemi abil on lihtne näidata, et $N_h \xrightarrow{d} N \sim \mathcal{N}(0;1)$. ■

Teoreemist 3.3 selgus, et eeskirjaga π^f panustades tekib lognormaalne auhinnasumma jaotus – joonis 3.1 näitab head vastavust selle teoreetilise jaotuse

ja simulatsiooni tulemuste vahel. Seetõttu võib olla mõistlik maksimeerida auhinnasumma enda keskväärtuse asemel auhinnasumma logaritmi keskväärtust. Auhinnasumma logaritmi keskväärtus on $\mu(h-1)$, mis kasvab võrdeliselt suurusega $\mu = \mathbb{E}[\ln(1+B_1 f)] = p_w \cdot \ln(1+f) + (1-p_w)\cdot\ln(1-f)$. Seda suurust aga maksimeeribki Kelly panustaja $f = 2p_w - 1$.

3.3 Virtuaalne keskväärtus

On oluline mõista, et logaritm Kelly panustamissüsteemis ei tulene auhinnafunktsiooni kujust. Auhinnafunktsioon ise on lineaarne – $r(s_h^w, a) = w$ – aga *punktide* arv kasvab eksponentsiaalselt, mistõttu nende kasvukiirust iseloomustab nende logaritm. Teisisõnu punktid on väärtuslikud kahel põhjusel:

1. Kui on rohkem punkte, siis saab juba nende olemasolevate punktide eest suuremat auhinda.
2. Kui on rohkem punkte mingil ajasammul, siis saab sellel ajasammul rohkem panustada, mistõttu kasvab sellele järgnevatel ajasammudel punktide arv kiiremini.

Pika horisondi korral on teine põhjus olulisem ning just sellel teisel põhjusel tuleb eksponentsiaalse kasvu korral maksimeerida punktide arvu logaritmi, sõltumata auhinnafunktsiooni kujust.

Siiski võib arvata, et kui ühel või teisel põhjusel tahetakse maksimeerida logaritmi, siis saab seda formuleerida keskväärtuse maksimeerimisena, võttes lihtsalt logaritmi kõikidest positiivsetest auhindadest. Täpsemalt olgu edaspidi $\widetilde{r}\colon \mathcal{S}\times\mathcal{A}\to\mathbb{R}$ *virtuaalne* auhinnafunktsioon ja r *tõeline* auhinnafunktsioon. Lõppeesmärk on saada tõelisi auhindu, nii et eeskirja $\widetilde{\pi}$ headust hinnatakse nagu varem tõelise auhinnasumma $Z_1^{\widetilde{\pi}}(s_1)$ järgi, kuid eeskirja *valitakse* maksimeerides virtuaalse auhinnasumma keskväärtust $\tilde{v}_1^{\widetilde{\pi}}(s_1)$, kus $\tilde{v}_{t_1}^{\pi}(s) = \mathbb{E}[\widetilde{Z}_{t_1}^{\pi}(s)]$ ja

$$\widetilde{Z}_{t_1}^{\pi}(s) = \left(\sum_{t=t_1}^{h} \widetilde{r}(S_t, A_t) \,\middle|\, \pi, S_{t_1} = s \right).$$

3.3.1 Kelly panustamine virtuaalse keskväärtuse maksimeerimisena

Näites 2.2 (rulett) oleks loomulik lihtsalt kasutada $\widetilde{r}(s,a) = \ln(r(s,a))$, aga kahjuks see ei ole võimalik, sest logaritm nullist ei ole defineeritud – tõeline auhind võib aga olla 0, kui ei ole viimane ajasamm või kui agent panustas ja kaotas kõik punktid mingil ajasammul. Seetõttu tuleb kasutada natuke keerulisemat virtuaalset auhinnafunktsiooni.

Teoreem 3.4. Olgu näites 2.2 (rulett)

$$\widetilde{r}(s,a) = \begin{cases} -2^h & \text{kui } s = s_t^a \\ \ln(r(s,a)) & \text{kui } r(s,a) > 0 \\ 0 & \text{muidu.} \end{cases}$$

Vastava virtuaalse auhinnasumma keskväärtuse maksimeerimine on võrdväärne Kelly panustamissüsteemiga selles näites – virtuaalset keskväärtust maksimeerib eeskiri π^f, kus $f = 2p_w - 1$.

Tõestuse skeem. Kui eelviimasel ajasammul on punktide arv $W_{h-1} = w > 0$ ja agent valib tegevuse $\widetilde{\pi}(s^w_{h-1}) = a < w$, siis $W_h > 0$ ja

$$\begin{aligned}\mathbb{E}[\widetilde{Z}^{\widetilde{\pi}}_{h-1}(s^w_{h-1})] &= \mathbb{E}[\ln(W_h) \mid \widetilde{\pi}, W_{h-1} = w] \\ &= \mathbb{E}[\ln(w + aB_{h-1})] \\ &= p_w \ln(w + a) + (1 - p_w) \ln(w - a),\end{aligned}$$

mida maksimeerib $a = (2p_w - 1)w$.

Kui $\forall t \in [1 + t_1, h], \widetilde{\pi}(s^w_t) = (2p_w - 1)w$ ja $\widetilde{\pi}(s^w_{t_1}) = a$, siis sarnaselt lemmaga 3.2

$$\begin{aligned}\mathbb{E}[\widetilde{Z}^{\widetilde{\pi}}_{t_1}(s^w_{t_1})] &= \mathbb{E}[\ln(W_h) \mid \widetilde{\pi}, W_{t_1} = w] \\ &= \mathbb{E}\left[\ln\left(w + \sum_{t=t_1}^{h-1} A_t B_t\right) \middle| \widetilde{\pi}, W_{t_1} = w\right] \\ &= \mathbb{E}\left[\ln\left((w + aB_{t_1})\left(1 + \prod_{t=1+t_1}^{h-1} (1 + B_t f)\right)\right)\right] \\ &= \mathbb{E}[\ln(w + aB_{t_1})] + \mathbb{E}\left[\ln\left(1 + \prod_{t=1+t_1}^{h-1} (1 + B_t f)\right)\right],\end{aligned}$$

mille maksimumis jälle $a = (2p_w - 1)w$. Täistõestus on lisas 7.1. ■

3.3.2 Jäätise söömine virtuaalse keskväärtusega

Ruleti näites sai Kelly panustamissüsteemi virtuaalse keskväärtuse abil simuleerida, aga jäätise söömise puhul annab logaritmiline virtuaalne auhinnafunktsioon halvemat tulemust, kui lihtsalt keskväärtuse maksimeerimine.

Lemma 3.5. Olgu näites 2.1 (jäätis)

$$\widetilde{r}(s, a) = \begin{cases} -2^h & \text{kui } s = s^a_t \\ \ln(r(s, a)) & \text{kui } r(s, a) > 0 \\ 0 & \text{muidu.} \end{cases}$$

Vastava virtuaalse auhinnasumma keskväärtust maksimeerib eeskiri $\widetilde{\pi}$, mis ei vali kunagi mündiviset: $\widetilde{\pi}(s^1) = \alpha^2$.

Tõestus: Näites 2.1 (jäätis) mõjutas auhinda või järgmist olekut ainult tegevuse valik olekus s^1. Kuna $r(s^1, a) = 0$, siis ka $\widetilde{r}(s^1, a) = 0$, sõltumata tegevusest a.

Kui $a = \alpha^1$, siis järgmine olek on s^2 tõenäosusega $0\{,\}5$ ja s^4 tõenäosusega $0\{,\}5$,

mistõttu virtuaalse auhinna keskväärtus on

$$\begin{aligned}\widetilde{r}(s^2,a)\cdot 0{,}5+\widetilde{r}(s^4,a)\cdot 0{,}5 &= \ln(1)\cdot 0{,}5+\ln(6)\cdot 0{,}5\\ &\approx 0+1.792\cdot 0{,}5=0.896.\end{aligned}$$

Kui $a=\alpha^2$, siis järgmine olek on s^3, mistõttu virtuaalne auhind on $\widetilde{r}(s^3,a)=\ln(3)\approx 1.099$.

Järelikult $\widetilde{\pi}(s^1)=\alpha^2$. ■

Kuna virtuaalse keskväärtuse maksimeerija ei vali kunagi mündiviset, siis ta saab igal teisel ajasammul (tõelise) auhinna $r(s^3,a)=3$ ning kokku auhinnasumma $\frac{3h}{2}$, kui horisont on h. See on aga enamasti väiksem (tõelise) keskväärtuse maksimeerija või Kelly panustaja auhinnasummast, mille jaotus lähenes normaaljaotusele keskväärtusega $\frac{3{,}5h}{2}$.

3.4 Kaheosaline mäng

Tähistagu $\{(p_1;x_1),\dots,(p_n;x_n)\}$ diskreetset jaotust, mis on juhuslikul suurusel X, kui iga $i\in[1,n]\cap\mathbb{N}$ korral $\mathbb{P}[X=x_i]=p_i$.

Näide 3.1. Juku mängib kahe osaga mängu. Esimeses osas mängib ta h_1 ajasammu näite 2.2 ruletisarnast mängu. Esimese osa viimasel ajasammul saab ta vastava auhinna. Seejärel liigub ta näite 2.1 (jäätis) algolekusse ning mängib veel h_2 ajasammu, saades vastavaid auhindu.

Otsustusprotsessina on mäng järgmine. Olgu horisont $h=h_1+h_2$, kus $h_1,h_2\geq 2$ ning h_2 on paarisarv. Olgu algolek $s_1=s_1^1$. Olgu jäätise söömise esimene olek s_0^1 ning šokolaadi-, vanilli- ja maasikajäätise saamise olekud vastavalt s_0^2,s_0^3,s_0^4. Olgu olekute hulk

$$\mathcal{S}=\{s_t^w\mid w\in[0,2^{h_1-1}],t\in[1,h_1]\cap\mathbb{Z}\}\cup\{s_0^1,s_0^2,s_0^3,s_0^4\}.$$

Olgu tegevuste hulk $\mathcal{A}(s_t^w)=[0,w]$, kui $t>0$, ja $\mathcal{A}(s_0^w)=\{\alpha^1,\alpha^2\}$ muidu. Üleminekufunktsioon $u(s_{h_1}^w,a,s_0^1)=1$; muidu on üleminekufunktsioon nagu näidetes 2.1 (jäätis) ja 2.2 (rulett). Auhinnafunktsioon r on nagu näidetes 2.1 ja 2.2.

Lemma 3.6. Näites 3.1 maksimeerib keskväärtust eeskiri π°, mis esimeses osas panustab alati kõik punktid ning teises osas alati valib mündivise. Ülalmainitud virtuaalse auhinnafunktsioonile vastavat keskväärtust maksimeerib eeskiri $\widetilde{\pi}$, mis on esimeses osas võrdne Kelly panustamissüsteemi eeskirjaga π^f, kus $f=2p_w-1$, ning teises osas ei vali kunagi mündiviset. Kelly panustamissüsteemi järgi tuleks kasutada esimeses osas eeskirja π^f ning teises osas alati valida mündiviset.

Tõestus: Esimeses osas valitud tegevused ei saa mõjutada teise osa olekuid ega auhindu, sest esimese osa lõpus liigub Juku alati samasse olekusse, s_0^1. Sarnaselt ei saa need mõjutada ka virtuaalseid auhindu. Järelikult kahe osa nii tõeliste

kui ka virtuaalsete auhindade summat maksimeerivad samad tegevused, mis maksimeerivad auhinnasummat näidetes 2.2 (rulett) ja 2.1 (jäätis) eraldi.

Esimeses osa auhindade summa kasvab eksponentsiaalselt horisondi suurenedes, nagu näites 2.2, ning teise osa auhindade summa lineaarselt, nagu näites 2.1. Seetõttu tuleb Kelly panustamissüsteemi kasutades järgida vastavaid eeskirju, nagu näidetes 2.2 ja 2.1. ■

Kui näites 2.1 (jäätis) valida alati mündivise ehk $\pi(s^1) = \alpha^1$, siis auhinnasumma oli

$$Z_1^{\pi}(s_1) = \frac{h}{2}\mu + \sqrt{\frac{h}{2}}\sigma L_h,$$

kus $\mu = \mathbb{E}[r(S_2, A_2)]$, $\sigma = \sqrt{\mathrm{Var}[r(S_2, A_2)]}$ ning h kasvades $L_h \xrightarrow{d} N \sim \mathcal{N}(0; 1)$. Järelikult ka näite 3.1 teise osa auhindade summa on sel juhul

$$\frac{h_2}{2}\mu_2 + \sqrt{\frac{h_2}{2}}\sigma_2 L_{h_2},$$

kus $\mu_2 = \mathbb{E}[r(S_{2+h_1}, A_{2+h_1})] = 3{,}5$ ning $\sigma_2 = \sqrt{\mathrm{Var}[r(S_{2+h_1}, A_{2+h_1})]}$.

Kui aga $\pi(s^1) = \alpha^2$, siis auhindade summa on $\frac{3h_2}{2}$.

Kui näites 2.2 (rulett) järgida eeskirja π^f, siis teoreemi 3.1 kohaselt oli auhinnasumma

$$Z_1^{\pi^f}(s_1) = \exp\left((h-1)\mu + \sqrt{h-1}\sigma N_h\right),$$

kus $\mu = \mathbb{E}[\ln(1 + B_1 f)]$ ja $\sigma = \sqrt{\mathrm{Var}[\ln(1 + B_1 f)]}$ ning $N_h \xrightarrow{d} N$. Järelikult ka näite 3.1 esimese osa auhindade summa on sel juhul

$$\exp\left((h_1-1)\mu_1 + \sqrt{h_1-1}\sigma_1 N_{h_1}\right),$$

kus $\mu_1 = \mathbb{E}[\ln(1 + B_1 f)]$ ja $\sigma_1 = \sqrt{\mathrm{Var}[\ln(1 + B_1 f)]}$.

Kui näites 2.2 (rulett) panustada igal ajasammul kõik punktid ehk $\pi(s_t^w) = w$, siis auhinnasumma oli

$$Z_1^{\pi}(s_1) \sim \{(p_w^{h-1}; 2^{h-1}), (1 - p_w^{h-1}; 0)\}.$$

Järelikult ka näite 3.1 esimese osa auhindade summa on sel juhul

$$R_{h_1} \sim \{(p_w^{h_1-1}; 2^{h_1-1}), (1 - p_w^{h_1-1}; 0)\}.$$

Lemma 3.6 tõttu saab näites 3.1 keskväärtuse maksimeerija auhinnasumma

$$Z_1^{\pi^\circ}(s_1) = R_{h_1} + \frac{h_2}{2}\mu_2 + \sqrt{\frac{h_2}{2}}\sigma_2 L^\circ_{h_2},$$

kus $L^\circ_{h_2} \xrightarrow{d} N$, virtuaalse keskväärtuse maksimeerija auhinnasumma

$$Z_1^{\tilde{\pi}}(s_1) = \exp\left((h_1-1)\mu_1 + \sqrt{h_1-1}\sigma_1 \widetilde{N}_{h_1}\right) + \frac{3h_2}{2},$$

kus $\widetilde{N}_{h_1} \xrightarrow{d} N$, ning Kelly panustaja auhinnasumma

$$Z_1^{\pi}(s_1) = \exp\left((h_1-1)\mu_1 + \sqrt{h_1-1}\sigma_1 N_{h_1}\right) + \frac{h_2}{2}\mu_2 + \sqrt{\frac{h_2}{2}}\sigma_2 L_{h_2}.$$

Teoreem 3.7. Olgu π Kelly panustaja eeskiri, π° keskväärtuse maksimeerija eeskiri ja $\widetilde{\pi}$ virtuaalse keskväärtuse maksimeerija eeskiri. Iga $c > 0$ ja $p \in (0;1)$ korral leidub MDP, mille puhul samaaegselt $\mathbb{P}[Z_1^{\pi}(s_1) > Z_1^{\pi^\circ}(s_1) + c] > p$ ja $\mathbb{P}[Z_1^{\pi}(s_1) > Z_1^{\widetilde{\pi}}(s_1) + c] > p$.

Tõestuse skeem: Väide kehtib, kui MDP on nagu näites 3.1,

$$h_2 = \left\lfloor 2\exp(2(h_1-1)\mu_1 - 2\sqrt{h_1-1}m\sigma_1 - 2\ln(3m\sigma_2))\right\rfloor$$

ning m ja h_1 on piisavalt suured. Tõestus on lisas 7.1.■

3.5 Kokkuvõte

Auhinnasumma keskväärtuse maksimeerija saab näites 2.1 (jäätis) sama jaotusega auhinnasumma, mis Kelly panustaja, aga teoreemidest 3.1 ja 3.3 selgus, et Kelly panustaja saab peaaegu alati suurema auhinnasumma näites 2.2 (rulett). Teoreem 3.4 näitas, et logaritmilise virtuaalse auhinnafunktsiooniga saab virtuaalse auhinnasumma keskväärtuse maksimeerija näites 2.2 (rulett) sama jaotusega auhinnasumma, mis Kelly panustaja, aga teoreemist 3.5 selgus, et Kelly panustaja saab enamasti suurema auhinnasumma näites 2.1 (jäätis).

Seega kui kasutada kriteeriumina keskväärtust, tuleks iga stiimulõpe ülesande puhul eraldi uurida, kas tavaline auhinnafunktsioon või mõni virtuaalne auhinnafunktsioon annab parema tulemuse. See on aga vastuolus stiimulõpe eesmärgiga leida igas ülesandes automaatselt kasulikku eeskirja.

Lisaks näitas teoreem 3.7, et ükskõik kui suure tõenäosuse $p < 1$ ja konstandi c korral leidub MDP, mille puhul nii tavalise kui ka virtuaalse keskväärtuse maksimeerija saab vähemalt tõenäosusega p vähemalt c võrra väiksemat auhinnasummat kui Kelly panustaja. Seega võib mõnede stiimulõpe ülesannete puhul olla soovitav kasutada keskväärtuse maksimeerimise asemel Kelly panustamissüsteemi.

4 Silutud mediaan

Kelly panustamine annab mõistliku tulemuse eelmainitud mängudes, kuid seda saab kasutada ainult mängudes, kus toimib mingi korduv panustamine, ning selle kasutamiseks tuleb iga konkreetse mängu puhul eraldi uurida, milline on auhinnasumma kasv – näiteks kas on eksponentsiaalne või lineaarne kasv. Oleks soovitav kriteerium, mis annaks sarnaseid tulemusi Kelly panustamisele eelmainitud mängudes, aga üldistuks kõikidele stiimulõpe ülesannetele.

4.1 Mediaani maksimeerimine

Näites 2.1 (jäätis) kasvas auhinnasumma lineaarselt. Kui maksimeerida selle keskmist kasvukiirust, siis tekkis auhinnasumma, mis oli sõltumatute sama lõpliku keskväärtuse ja standardhälvega auhindade summa. Keskse piirteoreemi kohaselt lähenes auhinnasumma jaotus normaaljaotusele. Näite 2.1 puhul on kasvukiiruse keskväärtus võrdeline selle normaaljaotuse mediaaniga, mistõttu keskmise kasvukiiruse maksimeerimine on võrdväärne auhinnasumma mediaani maksimeerimisega.

Näites 2.2 (rulett) kasvas auhinnasumma eksponentsiaalselt. Kui maksimeerida selle keskmist kasvukiirust, siis tekkis log-normaaljaotusega auhinnasumma. Näite 2.2 puhul oli kasvukiiruse keskväärtus võrdeline auhinnasumma logaritmi keskväärtusega, mille maksimeerimine on jälle võrdväärne auhinnasumma mediaani maksimeerimisega.

Seega üks võimalus üldistumiseks teistele stiimulõpe ülesannetele on maksimeerida auhinnasumma mediaani $\mathrm{Med}\,[Z_1^\pi(s_1)]$, kus $\mathrm{Med}[X]$ on juhusliku suuruse X mediaan.

4.2 Mediaani puudused

Mediaani maksimeerimine annab mõistliku tulemuse kahes eelmainitud näites, kuid järgmises näites annab tulemuse, mis ilmselt ei ühti paljude stiimulõpe kasutajate eelistustega.

Näide 4.1. (panustamismäng) Jukul on valik, kas osaleda panustamismängus. Kui ta otsustab osaleda, siis ta saab tõenäosusega 0,49 üks miljon eurot ja vastastikusel juhul kaotab ühe sendi. Kui ta otsustab mitte osaleda, siis ei juhtu midagi.

Otsustusprotsessina on mäng järgmine. Olgu horisont $h = 2$. Olgu algolek $s_1 = s^1$, s^2 olek, kus Juku saab miljon eurot, ning s^3 olek, kus ta kaotab ühe sendi.

Olgu osalemine tegevus α^1 ja mitte osalemine α^2. Kui Juku valib tegevuse α^1, siis ta liigub tõenäosusega $u(s^1, \alpha^2, s^2) = 0{,}51$ olekusse $s_2 = s^2$ ning tõenäosusega $u(s^1, \alpha^2, s^3) = 0{,}49$ olekusse $s_2 = s^3$. Kui ta valib tegevuse α^2, siis ta jääb (s.t liigub uuesti) algolekusse s^1, mistõttu $u(s^1, \alpha^1, s^1) = 1$.

Olgu $r(s^1, a) = 0$, $r(s^2, a) = 10^6$ ja $r(s^3, a) = -0{,}01$. Kui $\pi(s^1) = \alpha^1$, siis auhinnasumma $Z_1^\pi(s_1)$ jaotus on $\{(0{,}51; -0{,}01), (0{,}49; 10^6)\}$, mille mediaan on -0,01. Kui $\pi(s^1) = \alpha^2$, siis jaotus on $\{(1; 0)\}$, mille mediaan on 0. Seega mediaani $\mathrm{Med}\,[Z_1^\pi(s_1)]$ maksimeerib mängus mitte osalemine, kuigi sellega oleks ainult oht kaotada ühe sendi ja arvestatav võimalus võita miljon eurot.

Probleem tuleneb sellest, et auhinnasumma jaotusfunktsioon on väga järsk – see on väärtusega 0 kohtadel alla $-0{,}01$, tõuseb järsult väärtuseni 0,51 kohal $-0{,}01$ ning jääb siis täpselt samaks kuni palju suurema kohani. Jaotusfunktsiooni saab siledamaks teha lisades auhinnasummale juhuslikku "müra". Sisuliselt kui

auhinnasumma oleks muidu täpselt $-0{,}01$, siis pärast müra lisamist on tal sel juhul võimalus olla mõnevõrra suurem või väiksem kui $-0{,}01$, mis võib mediaani tõsta üle nulli.

4.3 Silutud mediaan

Olgu juhusliku suuruse X *silutud mediaan* $\mathbb{S}[X] = \mathrm{Med}[X + M]$, kus M on *silumismüra*. Põhimõtteliselt on palju võimalikke juhusliku suuruse M jaotuse valikuid, aga lihtsuse mõttes olgu sellel normaaljaotus keskväärtusega 0 ja standardhälvega $\sigma > 0$.

Siinkohal σ on vaba parameeter, mis vastab auhinnafunktsiooni subjektiivsele skaalale. Stiimulõpe kasutaja peaks valima σ sellise väärtuse, millest suuremad auhinnad on tema jaoks subjektiivselt suured. Kui kõiki auhindu korrutatakse läbi mõne positiivse konstandiga, siis parameetrit σ tuleb korrutada sama konstandiga. Väiksema σ puhul on silutud mediaan sarnasem tavalisele mediaanile ning suurema σ puhul sarnasem keskväärtusele.

Sellisel silumismüral on mõned kasulikud omadused:

- See on pidev juhuslik suurus, tihedusfunktsiooniga $f_M : \mathbb{R} \to [0, \infty)$.
- Iga $x \in \mathbb{R}$ korral $f_M(x) > 0$, mistõttu jaotusfunktsioon $F_M : \mathbb{R} \to [0, 1]$ on rangelt kasvav.
- See on sümmeetriline nulli ümber ehk $\forall x \in \mathbb{R}, f_M(-x) = f_M(x)$, mistõttu suurustel M ja $-M$ on sama jaotus ehk $\forall c \in \mathbb{R}, \mathbb{P}[M \leq c] = \mathbb{P}[-M \leq c]$.

Oluline on veel märkida, et kahe erineva juhusliku suuruse X ja Y silutud mediaanide arvutamisel tuleb kasutada silumismürana kaks sõltumatut sama jaotusega juhuslikku suurust, näiteks M ja M°, kusjuures $\mathbb{S}[X] = \mathrm{Med}[X + M]$ ning $\mathbb{S}[Y] = \mathrm{Med}[Y + M^\circ]$.

4.4 Silutud mediaani omadused

Omadus 4.1. Igal juhuslikul suurusel X leidub ühene lõplik silutud mediaan $\mathbb{S}[X] \in \mathbb{R}$.

Tõestus: Kuna silumismüra on pidev ning pideva juhusliku suuruse summa teise juhusliku suurusega on alati pidev, siis $(X + M)$ on pidev juhuslik suurus. Kuna igal pideval juhuslikul suurusel leidub ühene lõplik mediaan, siis leidub ka $\mathrm{Med}[X + M]$. ■

Omadus 4.2. Olgu X juhuslik suurus.

1. $\mathbb{S}[X] = c$ parajasti siis, kui $\mathbb{P}[X + M > c] = \mathbb{P}[X + M < c] = \frac{1}{2}$.
2. $\mathbb{S}[X] > c$ parajasti siis, kui $\mathbb{P}[X + M > c] > \frac{1}{2}$.
3. $\mathbb{S}[X] < c$ parajasti siis, kui $\mathbb{P}[X + M > c] < \frac{1}{2}$.

Tõestus:

1. Esimene võrdus tuleneb sellest, et iga pideva juhusliku suuruse Y korral $\mathbb{P}[Y > \mathrm{Med}[Y]] = \mathbb{P}[Y < \mathrm{Med}[Y]] = \frac{1}{2}$.

2. Olgu $\mathbb{S}[X] > c$. Kuna üksteist välistavate sündmuste tõenäosused liituvad, siis

$$\mathbb{P}[X + M > c] = \mathbb{P}[X + M > \mathbb{S}[X] > c] + \mathbb{P}[\mathbb{S}[X] > X + M > c].$$

   Kuna $\mathbb{P}[X + M = \mathbb{S}[X]] = \frac{1}{2}$, siis

$$\mathbb{P}[X + M > c] = \frac{1}{2} + \mathbb{P}[\mathbb{S}[X] > X + M > c].$$

   Kuna silumismüra M jaotusfunktsioon on rangelt kasvav, siis ka summa $(X + M)$ jaotusfunktsioon on rangelt kasvav, mistõttu

$$\mathbb{P}[\mathbb{S}[X] > X + M > c] > 0.$$

   Järelikult $\mathbb{P}[X + M > c] > \frac{1}{2}$.

3. Olgu nüüd $\mathbb{S}[X] < c$. Sarnaselt eelmisele võrratusele

$$\begin{aligned}\mathbb{P}[X + M < c] &= \mathbb{P}[X + M < \mathbb{S}[X] < c] + \mathbb{P}[\mathbb{S}[X] < X + M < c] \\ &= \frac{1}{2} + \mathbb{P}[\mathbb{S}[X] < X + M < c] \\ &> \frac{1}{2}\end{aligned}$$

   ning summa $(X + M)$ pidevuse tõttu

$$\mathbb{P}[X + M > c] = 1 - \mathbb{P}[X + M < c] < \frac{1}{2}.$$

Oletagem vastuväiteliselt, et $\mathbb{P}[X + M > c] > \frac{1}{2}$, aga ei kehti $\mathbb{S}[X] > c$. Siis $\mathbb{S}[X] \leq c$, mistõttu $\mathbb{P}[X + M > c] \leq \frac{1}{2}$ – vastuolu. Järelikult kui $\mathbb{P}[X + M > c] > \frac{1}{2}$, siis $\mathbb{S}[X] > c$.

Analoogiliselt kui $\mathbb{P}[X + M > c] < \frac{1}{2}$, siis $\mathbb{S}[X] < c$. ■

Omadus 4.3. Iga juhusliku suuruse X korral $\mathbb{S}[-X] = -\mathbb{S}[X]$.

Tõestus: Omaduse 4.2 ja silumismüra sümmeetrilisuse tõttu

$$\frac{1}{2} = \mathbb{P}[X + M > \mathbb{S}[X]] = \mathbb{P}[-X - M < -\mathbb{S}[X]] = \mathbb{P}[(-X) + M < (-\mathbb{S}[X])].$$

Omaduse 4.2 tõttu $\mathbb{S}[-X] = -\mathbb{S}[X]$. ■

Omadus 4.4. Iga juhusliku suuruse X ja konstandi $c \in \mathbb{R}$ korral $\mathbb{S}[X + c] = \mathbb{S}[X] + c$.

Tõestus: Kuna $\mathbb{P}[X + M > \mathbb{S}[X]] = \frac{1}{2}$, siis ka $\mathbb{P}[(X + c) + M > \mathbb{S}[X] + c] = \frac{1}{2}$, mistõttu $\mathbb{S}[X + c] = \mathbb{S}[X] + c$. ■

Omadused 4.5 ja 4.6 on sarnase tähendusega. Omadus 4.5 sisuliselt tähendab, et kui X on kindlasti suurem kui Y, siis $\mathbb{S}[X] > \mathbb{S}[Y]$. Omadus 4.6 tähendab, et kui muutub peaaegu kindlaks, et X_n on suurem kui Y_n, siis $\mathbb{S}[X_n]$ muutub suuremaks kui $\mathbb{S}[Y_n]$.

Omadus 4.5. Olgu X ja Y sõltumatud diskreetsed juhuslikud suurused. Kui $\mathbb{P}[X > Y] = 1$, siis $\mathbb{S}[X] > \mathbb{S}[Y]$.

Tõestuse skeem: Leidub $c \in \mathbb{R}$, mille puhul $\mathbb{P}[X > c] = 1$ ja $\mathbb{P}[Y < c] = 1$. Järelikult $\mathbb{S}[X] > c > \mathbb{S}[Y]$. Täistõestus on lisas 7.2. ■

Omadus 4.6. Olgu X_n ja Y_n sõltumatud juhuslikud suurused iga $n \in \mathbb{N}$ korral. Kui $\exists \varepsilon > 0$, $\lim\limits_{n\to\infty} \mathbb{P}[X_n > Y_n + \varepsilon] = 1$, siis $\liminf\limits_{n\to\infty}(\mathbb{S}[X_n] - \mathbb{S}[Y_n]) > 0$.

Tõestuse skeem: Leiduvad sellised p ja c_n, et piisavalt suure n korral $\mathbb{P}[X_n > c_n + \frac{\varepsilon}{8}] \geq p$ ja $\mathbb{P}[Y_n < c_n - \frac{\varepsilon}{8}] \geq p$ ning p valiku tõttu $\mathbb{S}[X_n] \geq c_n + \frac{\varepsilon}{16}$ ja $\mathbb{S}[Y_n] \leq c_n - \frac{\varepsilon}{16}$. Täistõestus on lisas 7.2. ■

4.5 Silutud mediaani maksimeerimine

4.5.1 Erinevus keskväärtusest

Teoreemis 3.7 selgus, et iga $p \in (0;1)$ korral leidub MDP, mille puhul Kelly panustaja auhinnasumma Z_h on vähemalt tõenäosusega p konstandi võrra suurem nii keskväärtuse maksimeerija auhinnasummast Z_h° kui ka virtuaalse keskväärtuse maksimeerija auhinnasummast $\widetilde{Z}_h$, kui horisont h on piisavalt suur. Omadus 4.6 aga näitas, et kui piisavalt suure h korral on mõni juhuslik suurus X_h ükskõik kui suure tõenäosusega $p \in (0;1)$ konstandi võrra suurem kui mõni juhuslik suurus Y_h, siis suure h korral on suuruse X_h silutud mediaan suurem kui suuruse Y_h oma. See kehtib sõltumata silumismüra vaba parameetri σ valikust.

Järelikult näidetes 2.1 (jäätis) ja 2.2 (rulett) on keskväärtuse ja virtuaalse keskväärtuse maksimeerija auhinnasummal väiksem silutud mediaan kui Kelly panustaja auhinnasummal, mistõttu silutud mediaani maksimeerija peab erinema keskväärtuse ja virtuaalse keskväärtuse maksimeerijast. Silutud mediaani maksimeerija kohta ei kehti analoogiline väide teoreemiga 3.7, kus Kelly panustaja saab ükskõik kui suure tõenäosusega konstandi võrra suurema auhinnasumma, sest siis oleks omaduse 4.6 tõttu Kelly panustaja auhinnasummal suurem silutud mediaan kui silutud mediaani maksimeerija auhinnasummal. Seega silutud mediaani maksimeerija auhinnasumma on sarnasem Kelly panustaja auhinnasummale kui keskväärtuse maksimeerija oma.

4.5.2 Erinevus mediaanist

Näites 4.1 (panustamismäng) on auhinnad eurodes. Eeldades, et vähem kui ühe euro võib pidada väikseks auhinnaks, olgu silumismüra jaotuse parameeter näiteks $\sigma = 1$. Näites 4.1 oli auhinnasumma $Z_1^\pi(s_1)$ jaotus $\{(0{,}51; -0{,}01), (0{,}49; 10^6)\}$, kui valida $\pi(s^1) = \alpha^1$ ehk mängus osalemist, ja $\{(1;0)\}$, kui valida $\pi(s^1) = \alpha^2$ ehk mitte osalemist.

Kui mitte osaleda, siis $\mathbb{S}[Z_1^\pi(s_1)] = 0$, sest $\mathbb{P}[Z_1^\pi(s_1) + M > 0] = \mathbb{P}[M > 0] = \frac{1}{2}$.

Kui osaleda, siis

$$\begin{aligned}
\mathbb{P}[Z_1^\pi(s_1) + M > 0] &= \mathbb{P}[Z_1^\pi(s_1) + M > 0 \mid Z_1^\pi(s_1) = -0{,}01] \cdot \mathbb{P}[Z_1^\pi(s_1) = -0{,}01] \\
&\quad + \mathbb{P}[Z_1^\pi(s_1) + M > 0 \mid Z_1^\pi(s_1) = 10^6] \cdot \mathbb{P}[Z_1^\pi(s_1) = 10^6] \\
&= \mathbb{P}[-0{,}01 + M > 0] \cdot 0{,}51 + \mathbb{P}[10^6 + M > 0] \cdot 0{,}49 \\
&= \mathbb{P}[M > 0{,}01] \cdot 0{,}51 + \mathbb{P}[M > -10^6] \cdot 0{,}49 \\
&= (1 - F_M(0{,}01)) \cdot 0{,}51 + (1 - F_M(-10^6)) \cdot 0{,}49 \\
&\approx 0{,}496 \cdot 0{,}51 + 1 \cdot 0{,}49 \\
&= 0.74296,
\end{aligned}$$

kus F_M on standardse normaaljaotuse jaotusfunktsioon. Kuna $\mathbb{P}[Z_1^\pi(s_1) + M > 0] > \frac{1}{2}$, siis $\mathbb{S}[Z_1^\pi(s_1)] > 0$. Järelikult auhinnasumma silutud mediaan on suurem, kui mängus osaleda.

Seega silutud mediaani maksimeerimine võimaldab mängus osalemist näites 4.1 (panustamismäng), samas andes sarnaseid tulemusi Kelly panustamisele näidetes 2.1 (jäätis) ja 2.2 (rulett).

5 Kokkuvõte

Stiimulõppes kasutatakse tihti kriteeriumina auhinnasumma keskväärtust, kuid mõnede ülesannete puhul saab Kelly panustamissüsteem tulemuse, mis ilmselt oleks paljudele stiimulõppe kasutajatele soovitavam. Virtuaalse auhinnafunktsiooni kasutusega saab osa keskväärtuse maksimeerimise probleemidest vältida, kuid mitte kõiki.

Kelly panustamissüsteem väldib neid probleeme, aga samas ei üldistu kõikidele stiimulõppe ülesannetele. Selles töös arendati silutud mediaani kriteerium, mille maksimeerimine annab Kelly panustamissüsteemile sarnaseid tulemusi, aga üldistub kõikidele stiimulõppe ülesannetele ning väldib tavalise mediaani probleeme.

Edasine töö saaks uurida, kuidas praktikas arvutada eeskirju, mis maksimeerivad auhinnasumma silutud mediaani. Üks lähenemine oleks lihtsalt lähendada tervet auhinnasumma jaotust (Bellemare, Dabney, and Rowland 2023) ning lõpus maksimeerida jaotuse lähendi silutud mediaani.

Lisaks saaks uurida ülesandeid, kus auhinnasummal on raske sabaga jaotus, näiteks finantsis, sest jaotuse silutud mediaan alati leidub, erinevalt keskväärtusest.

6 Viidatud kirjandus

Bellemare, Marc G., Will Dabney, and Mark Rowland. 2023. *Distributional Reinforcement Learning.* MIT Press.

Kelly Jr., J. L. 1956. "A New Interpretation of Information Rate." *Bell System Technical Journal* 35 (4): 917–26. https://doi.org/https://doi.org/10.1002/j.1538-7305.1956.tb03809.x.

Mohtadi, Hamid, and Stefan Ruediger. 2013. "The Heavy Tail in Finance: A Survey." *Econometrics: New Research*, January, 109–20.

Sutton, Richard S., and Andrew G. Barto. 2018. *Reinforcement Learning: An Introduction*. Cambridge, MA, USA: A Bradford Book.

7 Lisad

7.1 Kriteeriumi olulisus

Lemma 3.8. (Bellmani võrrand) See on tuntud tõestus stiimulõppes(Sutton and Barto 2018). Olgu $(\mathcal{S}, \mathcal{A}, u, r, h, s_1)$ MDP. Iga eeskirja π, oleku $s \in \mathcal{S}$ ja ajasammu $t_1 \in [1, h] \cap \mathbb{Z}$ korral

$$v^\pi_{t_1}(s) = r(s, \pi(s)) + \sum_{s' \in \mathcal{S}(s,\pi(s))} u(s, \pi(s), s') \cdot v^\pi_{1+t_1}(s').$$

Tõestus. Kõigepealt

$$\begin{aligned} v^\pi_{t_1}(s) &= \mathbb{E}[Z^\pi_{t_1}(s)] \\ &= \mathbb{E}\left[\sum_{t=t_1}^{h} r(S_t, A_t) \mid \pi, S_{t_1} = s\right] \\ &= \mathbb{E}\left[r(S_{t_1}, A_{t_1}) \mid \pi, S_{t_1} = s\right] + \mathbb{E}\left[\sum_{t=1+t_1}^{h} r(S_t, A_t) \mid \pi, S_{t_1} = s\right] \\ &= r(s, \pi(s)) + \mathbb{E}\left[\sum_{t=1+t_1}^{h} r(S_t, A_t) \mid \pi, S_{t_1} = s\right], \end{aligned}$$

sest $(A_{t_1} \mid \pi, S_{t_1} = s) = (\pi(S_{t_1}) \mid S_{t_1} = s) = \pi(s)$. Nüüd

$$v^\pi_{t_1}(s) = r(s, \pi(s)) + \sum_{s' \in \mathcal{S}(s,\pi(s))} \mathbb{P}[S_{1+t_1} = s' \mid \pi, S_{t_1} = s] \cdot \mathbb{E}\left[\sum_{t=1+t_1}^{h} r(S_t, A_t) \mid \pi, S_{t_1} = s, S_{1+t_1} = s'\right]$$

ning Markovi omaduse tõttu võib ära jätta tingimuse $S_{t_1} = s$. Järelikult

$$\begin{aligned} v^\pi_{t_1}(s) &= r(s, \pi(s)) + \sum_{s' \in \mathcal{S}(s,\pi(s))} \mathbb{P}[S_{1+t_1} = s' \mid \pi, S_{t_1} = s] \cdot \mathbb{E}\left[\sum_{t=1+t_1}^{h} r(S_t, A_t) \mid \pi, S_{1+t_1} = s'\right] \\ &= r(s, \pi(s)) + \sum_{s' \in \mathcal{S}(s,\pi(s))} u(s, \pi(s), s') \cdot \mathbb{E}[Z^\pi_{1+t_1}(s')] \\ &= r(s, \pi(s)) + \sum_{s' \in \mathcal{S}(s,\pi(s))} u(s, \pi(s), s') \cdot v^\pi_{1+t_1}(s'). \blacksquare \end{aligned}$$

Teoreem 3.1. Näites 3.1 maksimeerib keskväärtust $v_1^{\pi}(s_1)$ eeskiri π°, mis panustab igas olekus (ja igal ajasammul) kõik seni kogutud punktid:

$$\forall s_t^w \in \mathcal{S}, \pi^\circ(s_t^w) = w.$$

Tõestus. Viimasel ajasammul auhind $r(s_h^w, a) = w$ ei sõltu tegevusest a. Samuti ei saa sellest tegevusest sõltuda järgmiste ajasammude auhinnad, sest neid ei ole. Järelikult võib auhinnasumma keskväärtuse maksimeerimisel sama hästi valida $\pi^\circ(s_h^w) = a = w$.

Eelviimasel ajasammul on auhinnasumma keskväärtus $v_{h-1}^{\pi^\circ}(s_{h-1}^w)$. Olgu $a = \pi^\circ(s_{h-1}^w)$. Kuna $r(s_{h-1}^w, a) = 0$ ja $v_h^{\pi^\circ}(s_h^w) = w$, siis lemma 3.8 (Bellmani võrrandi) järgi

$$\begin{aligned} v_{h-1}^{\pi^\circ}(s_{h-1}^w) &= \sum_{s' \in \mathcal{S}(s_{h-1}^w, a)} u(s_{h-1}^w, a, s') \cdot v_h^{\pi^\circ}(s') \\ &= u(s_{h-1}^w, a, s_h^{w+a}) \cdot (w + a) + u(s_{h-1}^w, a, s_h^{w-a}) \cdot (w - a) \\ &= p_w(w + a) + (1 - p_w)(w - a). \end{aligned}$$

Selle tuletis a suhtes on

$$\frac{d}{da} v_{h-1}^{\pi^\circ}(s_{h-1}^w) = p_w - (1 - p_w) = 2p_w - 1.$$

Kuna $p_w > 0{,}5$, siis tuletis on alati positiivne ehk mida suurem a on, seda suurem on $v_{h-1}^{\pi^\circ}(s_{h-1}^w)$. Järelikult võimalikult suure $v_{h-1}^{\pi^\circ}(s_{h-1}^w)$ saamiseks tuleb valida võimalikult suur a. Kuna $a \in \mathcal{A}(s_{h-1}^w) = [0, w]$, siis selleks on $\pi^\circ(s_{h-1}^w) = a = w$. Seejuures $v_{h-1}^{\pi^\circ}(s_{h-1}^w) = p_w(w + w) + (1 - p_w)(w - w) = 2p_w w$.

Olgu nüüd $t_1 < h$, $a = \pi^\circ(s_{t_1}^w)$ ning iga $t \in [1 + t_1, h] \cap \mathbb{Z}$ korral $\pi^\circ(s_t^w) = w$ ja $v_t^{\pi^\circ}(s_t^w) = (2p_w)^{h-t} w$. Lemma 3.8 (Bellmani võrrandi) järgi

$$\begin{aligned} v_{t_1}^{\pi^\circ}(s_{t_1}^w) &= p_w v_{1+t_1}^{\pi^\circ}(s_{1+t_1}^{w+a}) + (1 - p_w) v_{1+t_1}^{\pi^\circ}(s_{1+t_1}^{w-a}) \\ &= (2p_w)^{h-t_1-1} \left(p_w(w + a) + (1 - p_w)(w - a)\right). \end{aligned}$$

Selle tuletis a järgi on jälle alati positiivne, nii et maksimumis $\pi^\circ(s_{t_1}^w) = a = w$ ning

$$v_{t_1}^{\pi^\circ}(s_{t_1}^w) = (2p_w)^{h-t_1} w.$$

Induktsioonist tagurpidi üle ajasammude t_1 järeldub, et iga $t \in [1, h] \cap \mathbb{Z}$ korral $\pi^\circ(s_t^w) = w$. ■

Teoreem 3.4. Olgu näites 2.2 (rulett)

$$\widetilde{r}(s, a) = \begin{cases} -2^h & \text{kui } s = s_t^a \\ \ln(r(s, a)) & \text{kui } r(s, a) > 0 \\ 0 & \text{muidu.} \end{cases}$$

Vastava virtuaalse auhinnasumma keskväärtuse maksimeerimine on võrdväärne Kelly panustamissüsteemiga selles näites – virtuaalset keskväärtust maksimeerib eeskiri π^f, kus $f = 2p_w - 1$.

Tõestus. Kui agent panustaks mingil ajasammul kõik punktid, siis ta saaks sellel ajasammul virtuaalse auhinna -2^h. Maksimaalne võimalik punktide arv mängu lõpus on 2^{h-1} ja nende eest saab virtuaalse auhinna $\ln(2^{h-1})$, nii et maksimaalne võimalik virtuaalne auhinnasumma oleks agendil $\ln(2^{h-1}) - 2^h \leq 2^{h-1} - 2^h$, mis on negatiivne. Üldiselt juhusliku suuruse keskväärtus on ülimalt võrdne selle suuruse maksimumiga, mistõttu ka virtuaalse auhinnasumma keskväärtus oleks sel juhul negatiivne.

Kui agent panustaks iga kord aga 0 punkti, siis tal oleks viimases olekus üks punkt ning keskväärtus oleks $\ln(1) = 0 > 2^{h-1} - 2^h$. Järelikult virtuaalse auhinnasumma keskväärtust maksimeeriv eeskiri ei panusta ühelgi ajasammul kõik punktid, sest siis oleks teine eeskiri, mis saavutaks suuremat keskväärtust. Samuti on igal ajasammul punktide arv w positiivne, sest punktide arv saab olla 0 ainult siis, kui mingil ajasammul panustada kõik punktid.

Kui eelviimasel ajasammul on punktide arv $W_{h-1} = w > 0$ ja agent valib tegevuse $\widetilde{\pi}(s^w_{h-1}) = a < w$, siis $W_h > 0$ ja

$$\begin{aligned}\mathbb{E}[\widetilde{Z}^{\tilde{\pi}}_{h-1}(s^w_{h-1})] &= \mathbb{E}[\ln(W_h) \mid \widetilde{\pi}, W_{h-1} = w] \\ &= \mathbb{E}[\ln(w + aB_{h-1})] \\ &= p_w \ln(w + a) + (1 - p_w) \ln(w - a).\end{aligned}$$

Ekstreemumis

$$0 = \frac{d}{da}\,\mathbb{E}[\widetilde{Z}^{\tilde{\pi}}_{h-1}(s^w_{h-1})] = \frac{p_w}{w + a} - \frac{1 - p_w}{w - a},$$

mille ainus lahend on $a = (2p_w - 1)w$. Kuna $p_w < 1$, siis teine tuletis

$$\frac{d^2}{da^2}\,\mathbb{E}[\widetilde{Z}^{\tilde{\pi}}_{h-1}(s^w_{h-1})] = \frac{4p_w wa - w^2 - a^2 - 2wa}{(w + a)^2(w - a)^2} < \frac{-(w - a)^2}{(w + a)^2(w - a)^2}$$

on negatiivne, mistõttu ekstreemum on maksimum ning $\widetilde{\pi}(s^w_{h-1}) = a = (2p_w - 1)w$.

Kui $\forall t \in [1 + t_1, h], \widetilde{\pi}(s^w_t) = (2p_w - 1)w$, siis pärast ajasammu t_1 järgib agent sisuliselt eeskirja π^f, kus $f = 2p_w - 1$. Tähistades $\widetilde{\pi}(s^w_{t_1}) = a$, nüüd sarnaselt

lemmaga 3.2

$$\begin{aligned}
\mathbb{E}[\widetilde{Z}^{\widetilde{\pi}}_{t_1}(s^w_{t_1})] &= \mathbb{E}[\ln(W_h) \mid \widetilde{\pi}, W_{t_1} = w] \\
&= \mathbb{E}\left[\ln\left(w + \sum_{t=t_1}^{h-1} A_t B_t\right)\middle| \widetilde{\pi}, W_{t_1} = w\right] \\
&= \mathbb{E}\left[\ln\left(w + aB_{t_1} + \sum_{t=1+t_1}^{h-1} A_t B_t\right)\middle| \widetilde{\pi}, W_{t_1} = w\right] \\
&= \mathbb{E}\left[\ln\left(w + aB_{t_1} + W_{1+t_1}\prod_{t=1+t_1}^{h-1}(1+B_t f)\right)\middle| \widetilde{\pi}, W_{t_1} = w\right] \\
&= \mathbb{E}\left[\ln\left((w+aB_{t_1})\left(1+\prod_{t=1+t_1}^{h-1}(1+B_t f)\right)\right)\right] \\
&= \mathbb{E}\left[\ln(w+aB_{t_1}) + \ln\left(1+\prod_{t=1+t_1}^{h-1}(1+B_t f)\right)\right] \\
&= \mathbb{E}[\ln(w+aB_{t_1})] + \mathbb{E}\left[\ln\left(1+\prod_{t=1+t_1}^{h-1}(1+B_t f)\right)\right],
\end{aligned}$$

kus teine liidetav ei sõltu tegevusest a. Ainult maksimumis kehtib

$$\frac{d}{da}\,\mathbb{E}[\widetilde{Z}^{\widetilde{\pi}}_{t_1}(s^w_{t_1})] = 0$$

ja

$$\frac{d^2}{da^2}\,\mathbb{E}[\widetilde{Z}^{\widetilde{\pi}}_{h-1}(s^w_{h-1})] < 0,$$

mistõttu $a = (2p_w - 1)w$, sarnaselt juhuga $t_1 = h - 1$.

Induktsioonist tagurpidi üle ajasammude t_1 järeldub, et $\widetilde{\pi} = \pi^f$, kus $f = 2p_w - 1$. ■

Lemma 3.9. Olgu $y_1, y_2 : \mathbb{N} \to \mathbb{R}$. Kui $\lim\limits_{n\to\infty} y_1(n) = \infty$ ja $\lim\limits_{n\to\infty}(y_2(n) - y_1(n)) = \infty$, siis iga $c_1, c_2 > 0$ korral

$$\lim_{n\to\infty}(e^{y_2(n)} - c_1 e^{y_1(n)} - c_2) > 0.$$

Tõestus: Kuna $\lim\limits_{n\to\infty} y_1(n) = \infty$, siis piisavalt suure n korral $y_1(n) > \ln(1 + c_2)$. Kuna $\lim\limits_{n\to\infty}(y_2(n) - y_1(n)) = \infty$, siis piisavalt suure n korral $y_2(n) - y_1(n) > \ln(2c_1)$. Järelikult piisavalt suure n korral

$$\begin{aligned}
e^{y_2(n)} - c_1 e^{y_1(n)} - c_2 &> e^{y_1(n)+\ln(2c_1)} - c_1 e^{y_1(n)} - c_2 \\
&= 2c_1 e^{y_1(n)} - c_1 e^{y_1(n)} - c_2 \\
&= e^{y_1(n)} - c_2 \\
&> e^{\ln(1+c_2)} - c_2 = 1. \blacksquare
\end{aligned}$$

Teoreem 3.7. Olgu π Kelly panustaja eeskiri π° keskväärtuse maksimeerija eeskiri ja $\widetilde{\pi}$ virtuaalse keskväärtuse maksimeerija eeskiri. Iga $c > 0$ ja $p \in (0;1)$ korral leidub MDP, mille puhul samaaegselt $\mathbb{P}[Z_1^\pi(s_1) > Z_1^{\pi^\circ}(s_1) + c] > p$ ja $\mathbb{P}[Z_1^\pi(s_1) > Z_1^{\widetilde{\pi}}(s_1) + c] > p$.

Tõestus: Fikseerigem suvaliselt c ja p. Olgu $h_1, h_2 \in \mathbb{N}$ ja MDP $(\mathcal{S}, \mathcal{A}, u, r, h = h_1 + h_2, s_1)$ nagu näites 3.1.

Tuleb näidata, et leiduvad h_1, h_2, mille puhul $\mathbb{P}[Z_1^\pi(s_1) - Z_1^{\pi^\circ}(s_1) > c] > p$ ehk

$$\mathbb{P}\left[\exp\left((h_1-1)\mu_1 + \sqrt{h_1-1}\sigma_1 N_{h_1}\right) + \sqrt{\frac{h_2}{2}}\sigma_2(L_{h_2} - L^\circ_{h_2}) - R_{h_1} > c\right] > p$$

ja $\mathbb{P}[Z_1^\pi(s_1) - Z_1^{\pi^\circ}(s_1) > c] > p$ ehk

$$\begin{aligned}\mathbb{P}\Bigg[&\exp\left((h_1-1)\mu_1 + \sqrt{h_1-1}\sigma_1 N_{h_1}\right)\\ &- \exp\left((h_1-1)\mu_1 + \sqrt{h_1-1}\sigma_1 \widetilde{N}_{h_1}\right)\\ &+ \frac{h_2}{2}(\mu_2 - 3) + \sqrt{\frac{h_2}{2}}\sigma_2 L_{h_2} > c\Bigg] > p.\end{aligned}$$

Olgu $N \sim \mathcal{N}(0;1)$. Olgu iga $m \in \mathbb{N}$ korral $p_m = F_N(m) - F_N(-m)$. Kuna $\lim\limits_{m\to\infty} p_m = 1$, siis leidub m, mille puhul

$$p_m > \frac{1 + \sqrt[4]{p}}{2}.$$

Fikseerigem selline m.

Olgu

$$\begin{aligned}y_0(h_1) &= (h_1-1)\mu_1 - \sqrt{h_1-1}m\sigma_1,\\ y_1(h_1) &= (h_1-1)\mu_1 + \sqrt{h_1-1}m\sigma_1\end{aligned}$$

ja

$$y_2(h_1) = 2(h_1-1)\mu_1 - 2\sqrt{h_1-1}m\sigma_1 - 2\ln(3m\sigma_2) = 2y_0(h_1) - 2\ln(2m\sigma_2),$$

kus $y_0, y_1, y_2 : \mathbb{N} \to \mathbb{R}$. Olgu $h_2 = \lfloor 2\exp(y_2(h_1))\rfloor$, kus $\lfloor x\rfloor$ on suurim täisarv, mis ei ole suurem arvust x.

Kuna $N_{h_1}, L_{h_2} \xrightarrow{d} N$ ja $\lim\limits_{h_1\to\infty} h_2 = \infty$, siis iga $\varepsilon > 0$ ja piisavalt suure h_1 korral

$$|F_{N_{h_1}}(m) - F_N(m)|, |F_{N_{h_1}}(-m) - F_N(-m)|, |F_{L_{h_2}}(m) - F_N(m)|, |F_{L_{h_2}}(-m) - F_N(-m)| < \varepsilon.$$

Olgu h_1 edaspidi piisavalt suur, et võrratused kehtiksid $\varepsilon = \frac{1-\sqrt[4]{p}}{2}$ puhul.

Kuna $\lim\limits_{h_1\to\infty} y_0(h_1)=\infty$, siis piisavalt suure h_1 korral $\frac{1}{3}\exp(y_0(h_1))>c$, mistõttu

$$\begin{aligned}\exp(y_0(h_1)) &> \frac{2}{3}\exp(y_0(h_1))+c\\ &= 2m\sigma_2\exp(-\ln(3m\sigma_2))\exp(y_0(h_1))+c\\ &= 2m\sigma_2\sqrt{\exp(2y_0(h_1)-2\ln(3m\sigma_2))}+c\\ &= 2m\sigma_2\sqrt{\exp(y_2(h_1))}+c\\ &\geq 2m\sigma_2\sqrt{\frac{h_2}{2}}+c.\end{aligned}$$

Kuna $\lim\limits_{h_1\to\infty}\mathbb{P}[R_{h_1}\neq 0]=\lim\limits_{h_1\to\infty}p_w^{h_1-1}=0$, siis piisavalt suure h_1 korral $\mathbb{P}[R_{h_1}=0]>\sqrt[4]{p}$. Nüüd

$$\begin{aligned}&\mathbb{P}\left[\exp\left((h_1-1)\mu_1+\sqrt{h_1-1}\sigma_1 N_{h_1}\right)+\sqrt{\frac{h_2}{2}}\sigma_2(L_{h_2}-L^{\circ}_{h_2})-R_{h_1}>c\right]\\ &\geq \mathbb{P}\left[\exp\left((h_1-1)\mu_1+\sqrt{h_1-1}\sigma_1 N_{h_1}\right)+\sqrt{\frac{h_2}{2}}\sigma_2(L_{h_2}-L^{\circ}_{h_2})-R_{h_1}>c\,\middle|\,N_{h_1}>-m\right]\cdot\mathbb{P}[N_{h_1}>-m]\\ &\geq \mathbb{P}\left[\exp\left((h_1-1)\mu_1-\sqrt{h_1-1}\sigma_1 m\right)+\sqrt{\frac{h_2}{2}}\sigma_2(L_{h_2}-L^{\circ}_{h_2})-R_{h_1}>c\right](p_m-\varepsilon)\\ &\geq \mathbb{P}\left[\exp(y_0(h_1))+\sqrt{\frac{h_2}{2}}\sigma_2(-m-m)-R_{h_1}>c\right](p_m-\varepsilon)^3\\ &\geq \mathbb{P}\left[\exp(y_0(h_1))-2m\sigma_2\sqrt{\frac{h_2}{2}}-0>c\right](p_m-\varepsilon)^3\sqrt[4]{p}\\ &= (p_m-\varepsilon)^3\sqrt[4]{p}>\sqrt[4]{p^4}=p,\end{aligned}$$

sest $p\in(0;1)$. Seega $\mathbb{P}[Z_1^{\pi}(s_1)-Z_1^{\pi^\circ}(s_1)>c]>p$.

Kuna $\lim\limits_{h_1\to\infty}y_1(h_1)=\infty$ ja $\lim\limits_{h_1\to\infty}(y_2(h_1)-y_1(h_1))$, siis lemma 3.9 tõttu

$$\exp(y_2(h_1))>(2+2m\sigma_2)\exp(y_1(h_1))+c,$$

kui h_1 piisavalt suur. Kuna $y_1(h_1)>\frac{y_2(h_1)}{2}$, siis

$$\begin{aligned}\exp(y_2(h_1)) &> 2\exp(y_1(h_1))+2m\sigma_2\sqrt{\exp(y_2(h_1))}+2c\\ &\geq 2\exp(y_1(h_1))+2m\sigma_2\sqrt{\exp(y_2(h_1))}+2c-\exp(y_0(h_1)).\end{aligned}$$

Kasutades $\mu_2 - 3 = \frac{1}{2}$

$$\begin{aligned}
&\mathbb{P}\Bigg[\exp\Big((h_1-1)\mu_1 + \sqrt{h_1-1}\sigma_1 N_{h_1}\Big) \\
&- \exp\Big((h_1-1)\mu_1 + \sqrt{h_1-1}\sigma_1 \widetilde{N}_{h_1}\Big) \\
&+ \frac{h_2}{2}(\mu_2 - 3) + \sqrt{\frac{h_2}{2}}\sigma_2 L_{h_2} > c\Bigg] \\
&\geq \mathbb{P}\Bigg[\exp\Big((h_1-1)\mu_1 - \sqrt{h_1-1}\sigma_1 m\Big) \\
&- \exp\Big((h_1-1)\mu_1 + \sqrt{h_1-1}\sigma_1 m\Big) \\
&+ \frac{h_2}{4} - \sqrt{\frac{h_2}{2}}\sigma_2 m > c\Bigg](p_m - \varepsilon)^3 \\
&= \mathbb{P}\left[\exp(y_0(h_1)) - \exp(y_1(h_1)) + \frac{\exp(y_2(h_1))}{2} - 2m\sigma_2\sqrt{\exp(y_2(h_1))} > c\right](p_m - \varepsilon)^3 \\
&= \sqrt[4]{p^3} > p.
\end{aligned}$$

Seega $\mathbb{P}[Z_1^{\pi}(s_1) - Z_1^{\pi^\circ}(s_1) > c] > p$.

Otsitava MDP saab konstrueerida fikseerides piisavalt suure h_1. ■

7.2 Silutud mediaan

Lemma 4.7. Olgu X juhuslik suurus. Kui $\mathbb{P}[X \geq 0] = 1$, siis $\mathbb{S}[X] \geq 0$.

Tõestus: Silumismüra sümmeetrilisuse tõttu

$$\mathbb{P}[X + M > 0] = \mathbb{P}[X > -M] = \mathbb{P}[-M < X] = \mathbb{P}[M < X].$$

Kuna $\mathbb{P}[X \geq 0] = 1$, siis

$$\mathbb{P}[M < X] \geq \mathbb{P}[M < 0 \leq X] = \mathbb{P}[M < 0] = \frac{1}{2}.$$

Järelikult $\mathbb{P}[X + M > 0] \geq \frac{1}{2}$. Omaduse 4.2 tõttu $\mathbb{S}[X] \geq 0$. ■

Omadus 4.5. Olgu X ja Y sõltumatud diskreetsed juhuslikud suurused. Kui $\mathbb{P}[X > Y] = 1$, siis $\mathbb{S}[X] > \mathbb{S}[Y]$.

Tõestus: Olgu $x^\circ \in \mathbb{R}$ väikseim väärtus, mille korral $\mathbb{P}[X = x^\circ] > 0$, ja $y^\circ \in \mathbb{R}$ suurim väärtus, mille korral $\mathbb{P}[Y = y^\circ] > 0$. (Need leiduvad, sest X ja Y on diskreetsed.) Seejuures $x^\circ > y^\circ$, sest muidu kehtiks

$$\mathbb{P}[Y \geq X] \geq \mathbb{P}[Y = y^\circ, X = x^\circ] = \mathbb{P}[Y = y^\circ] \cdot \mathbb{P}[X = x^\circ] > 0,$$

mis on vastuolus võrdusega $\mathbb{P}[X > Y] = 1$.

Kuna $\mathbb{P}[X \geq x^\circ] = 1$, siis $\mathbb{P}[X - x^\circ \geq 0] = 1$ ning lemma 4.7 kohaselt $\mathbb{S}[X - x^\circ] \geq 0$. Omaduse 4.4 tõttu $\mathbb{S}[X] \geq x^\circ$.

Sarnaselt kuna $\mathbb{P}[Y \leq y^\circ] = 1$, siis ka $\mathbb{P}[-Y \geq -y^\circ] = \mathbb{P}[-Y + y^\circ \geq 0] = 1$ ning $\mathbb{S}[-Y + y^\circ] \geq 0$. Omaduse 4.3 tõttu $\mathbb{S}[Y - y^\circ] \leq 0$ ja omaduse 4.4 tõttu $\mathbb{S}[Y] \leq y^\circ$.

Järelikult $\mathbb{S}[Y] \leq y^\circ < x^\circ \leq \mathbb{S}[X]$. ■

Lemma 4.8. Olgu X juhuslik suurus ja $\varepsilon > 0$. Kui

$$\mathbb{P}[X > \varepsilon] \geq \frac{1}{2\,\mathbb{P}[M < \varepsilon]},$$

siis $\mathbb{S}[X] \geq 0$.

Tõestus: Analoogiliselt lemmaga 4.7 $\mathbb{P}[X + M > 0] = \mathbb{P}[M < X]$. Nüüd

$$\begin{aligned}
\mathbb{P}[M < X] &\geq \mathbb{P}[M < X, \varepsilon < X] \\
&= \mathbb{P}[M < X \mid \varepsilon < X] \cdot \mathbb{P}[\varepsilon < X] \\
&\geq \mathbb{P}[M < \varepsilon < X \mid \varepsilon < X] \cdot \mathbb{P}[\varepsilon < X] \\
&= \mathbb{P}[M < \varepsilon] \cdot \mathbb{P}[X > \varepsilon] \\
&\geq \mathbb{P}[M < \varepsilon] \cdot \frac{1}{2\,\mathbb{P}[M < \varepsilon]} = \frac{1}{2}
\end{aligned}$$

ehk $\mathbb{P}[X + M > 0] \geq \frac{1}{2}$. Omaduse 4.2 tõttu $\mathbb{S}[X] \geq 0$. ■

Lemma 4.9. Olgu X_n ja Y_n sõltumatud juhuslikud suurused iga $n \in \mathbb{N}$ korral ning $\varepsilon > 0$. Kui $\lim_{n\to\infty} \mathbb{P}[X_n > Y_n + \varepsilon] = 1$, siis iga $p \in (0;1)$ korral leidub selline $n_1 \in \mathbb{N}$, et iga $n \geq n_1$ korral leidub $c_n \in \mathbb{R}$, mille puhul $\mathbb{P}[X_n > c_n + \frac{\varepsilon}{8}] \geq p$ ja $\mathbb{P}[Y_n < c_n - \frac{\varepsilon}{8}] \geq p$.

Tõestus: Fikseerigem suvaliselt $p \in (0;1)$. Kuna $\lim_{n\to\infty} \mathbb{P}[X_n > Y_n + \varepsilon] = 1$ ja $1 - (1-p)^2 < 1$, siis leidub selline $n_1 \in \mathbb{N}$, et iga $n \geq n_1$ korral $\mathbb{P}[X_n > Y_n + \varepsilon] > 1 - (1-p)^2$. Fikseerigem selline n_1.

Kuna $\lim_{x\to-\infty} \mathbb{P}[X_n \geq x - \frac{\varepsilon}{4}] = 1$, siis leidub x, mille korral $\mathbb{P}[X_n \geq x - \frac{\varepsilon}{4}] \geq p$. Olgu x_n suurim selline x iga $n \in \mathbb{N}$ korral. Kuna x_n on suurim võimalik, siis

$$\mathbb{P}[X_n \geq x_n] = \mathbb{P}\left[X_n \geq \left(x_n + \frac{\varepsilon}{4}\right) - \frac{\varepsilon}{4}\right] < p$$

ehk $\mathbb{P}[X_n < x_n] > 1 - p$.

Analoogiliselt kuna $\lim_{y\to\infty} \mathbb{P}[Y_n \leq y + \frac{\varepsilon}{4}] = 1$, siis leidub y, mille korral $\mathbb{P}[Y_n \leq y + \frac{\varepsilon}{4}] \geq p$. Olgu y_n väikseim selline y iga $n \in \mathbb{N}$ korral. Kuna y_n on väikseim võimalik, siis $\mathbb{P}[Y_n \leq y_n] < p$ ehk $\mathbb{P}[Y_n > y_n] > 1 - p$.

Oletagem vastuväiteliselt, et mõne $n \geq n_1$ korral $x_n \leq y_n + \varepsilon$. Siis kasutades suuruste X_n ja Y_n sõltumatust

$$\mathbb{P}[Y_n + \varepsilon \geq X_n] \geq \mathbb{P}[Y_n > y_n, X_n < x_n] = \mathbb{P}[Y_n > y_n] \cdot \mathbb{P}[X_n < x_n] > (1-p)^2,$$

mis on vastuolus sellega, et $\mathbb{P}[X_n > Y_n + \varepsilon] > 1 - (1-p)^2$. Järelikult iga $n \geq n_1$ korral $x_n > y_n + \varepsilon$.

Olgu $c_n = y_n + \frac{\varepsilon}{2}$. Iga $n \geq n_1$ korral $\mathbb{P}[X_n \geq x - \frac{\varepsilon}{4} > c_n + \frac{\varepsilon}{8}] \geq p$ ja $\mathbb{P}[Y_n \leq y + \frac{\varepsilon}{4} < c_n - \frac{\varepsilon}{8}] \geq p$. ■

Omadus 4.6. Olgu X_n ja Y_n sõltumatud juhuslikud suurused iga $n \in \mathbb{N}$ korral. Kui $\exists \varepsilon > 0$, $\lim\limits_{n\to\infty} \mathbb{P}[X_n > Y_n + \varepsilon] = 1$, siis $\liminf\limits_{n\to\infty}(\mathbb{S}[X_n] - \mathbb{S}[Y_n]) > 0$.

Tõestus: Kuna $\frac{\varepsilon}{16} > 0$, siis $\mathbb{P}[M < \frac{\varepsilon}{16}] > \frac{1}{2}$. Olgu

$$p = \frac{1}{2\,\mathbb{P}[M < \frac{\varepsilon}{16}]} \in (0;1).$$

Lemma 4.9 kohaselt leidub selline $n_1 \in \mathbb{N}$, et iga $n \geq n_1$ korral leidub $c_n \in \mathbb{R}$, mille puhul $\mathbb{P}[X_n > c_n + \frac{\varepsilon}{8}] \geq p$ ja $\mathbb{P}[Y_n < c_n - \frac{\varepsilon}{8}] \geq p$. Fikseerigem sellised n_1 ja c_n.

Kuna iga $n \geq n_1$ korral $\mathbb{P}\left[(X_n - c_n - \frac{\varepsilon}{16}) > \frac{\varepsilon}{16}\right] \geq p$, siis lemma 4.8 kohaselt $\mathbb{S}[X_n - c_n - \frac{\varepsilon}{16}] \geq 0$ ning omaduse 4.4 tõttu $\mathbb{S}[X_n] \geq c_n + \frac{\varepsilon}{16}$.

Sarnaselt iga $n \geq n_1$ korral $\mathbb{P}\left[(Y_n - c_n + \frac{\varepsilon}{16}) < -\frac{\varepsilon}{16}\right] \geq p$ ehk

$$\mathbb{P}\left[-(Y_n - c_n + \frac{\varepsilon}{16}) > \frac{\varepsilon}{16}\right] \geq p,$$

mistõttu lemma 4.7 kohaselt $\mathbb{S}[-(Y_n - c_n + \frac{\varepsilon}{16})] \geq 0$. Omaduse 4.3 tõttu $\mathbb{S}[Y_n - c_n + \frac{\varepsilon}{16}] \leq 0$ ja omaduse 4.4 tõttu $\mathbb{S}[Y_n] \leq c_n - \frac{\varepsilon}{16}$.

Järelikult iga $n \geq n_1$ korral $\mathbb{S}[X_n] - \mathbb{S}[Y_n] \geq (c_n + \frac{\varepsilon}{16}) - (c_n - \frac{\varepsilon}{16}) = \frac{\varepsilon}{8}$. Seega $\liminf_{n\to\infty}(\mathbb{S}[X_n] - \mathbb{S}[Y_n]) \geq \frac{\varepsilon}{8} > 0$. ■

7.3 Simulatsiooni kood

```
import math
import random
import numpy as np
from scipy import stats
import matplotlib.pyplot as plt

class Sim:
    def __init__(self, n_free, free_pr, n_gauss=None, gauss_pr=0.5,
                 gauss_low=1.0, gauss_mid=3.0, gauss_high=6.0):
        # Number of free ("vaba" in Estonian, not "tasuta") bets (i.e. bets
        # with stakes freely chosen by the bettor) and "Gaussian bets" (i.e.
        # bets with fixed payoffs)
        self.n_free  = n_free
        self.n_gauss = n_gauss
```

```
        # Probability of profiting from each individual free and Gaussian bet;
        # determines transition function.
        self.free_pr  = free_pr
        self.gauss_pr = gauss_pr

        # Payoffs of Gaussian bets. `mid` is payoff one can get for sure;
        # `low` and `high` are potential payoffs if one chooses to take a risk.
        self.gauss_low  = gauss_low
        self.gauss_mid  = gauss_mid
        self.gauss_high = gauss_high

    def bet_frac_(self, f):
        # `f` is fraction of wealth to bet at each timestep; determines policy.
        if f is None:
            f = 2 * self.free_pr - 1
        if f < 0 or f > 1:
            raise ValueError("Fraction of wealth to bet must be in [0; 1].")

        return f

    def free_last_state_(self, n_samples, f):
        # `n_samples` is number of Monte-Carlo samples for computing Z distribution.
        f = self.bet_frac_(f)

        # Initial state is 1 unit of wealth (for each sample)
        s = np.ones(n_samples)

        # Total number of states in single realization of MDP
        horizon = self.n_free + 1

        # Initial state is already s_1, so transition (h - 1) times to get to s_h.
        for t in range(1, horizon):
            a = f * s            # Amount to bet
            profited = np.random.binomial(size=n_samples, p=self.free_pr, n=1)
            sgn = 2*profited - 1 # 1 if bet profited (i.e. paid off), -1 otherwise.
            s += a * sgn         # s_(t+1) = s_t + a*sgn

        # Return final state of each sample.
        return s

    def free(self, n_samples, f=None):
        # The last state is the only state with non-zero reward, and
        # the reward is linear in wealth, so z = r_h = s_h.
        s_h = self.free_last_state_(n_samples, f)
        return s_h
```

```
def exp_free(self, n_samples, f=None):
    # Here z = r_h = exp(s_h).
    s_h = self.free_last_state_(n_samples, f)
    return np.exp(s_h)

def take_gauss_mid_(self, maximize_log_reward):
    # `l`, `m`, `h` contain the quantities being maximized (i.e. the
    # "virtual"/auxiliary reward), but shouldn't be used to calculate
    # the *actual* reward attained, which might be different.
    l, m, h = self.gauss_low, self.gauss_mid, self.gauss_high
    if maximize_log_reward:
        l, m, h = np.log(l), np.log(m), np.log(h)

    avg = l * (1 - self.gauss_pr) + h * self.gauss_pr
    return m > avg

def gauss(self, n_samples, maximize_log_reward=False):
    if self.take_gauss_mid_(maximize_log_reward):
        # There are `self.n_gauss` bets, and at each bet, expectation
        # is maximized by taking certain payoff of `gauss_mid`, for
        # a total of `n_gauss * gauss_mid`.
        z = np.full(n_samples, self.gauss_mid * self.n_gauss)
        return z

    # Else, the expectation is maximized by taking a risk at each bet.
    # The minimum reward one gets after each bet is `gauss_low`, so the
    # minimum total reward is `gauss_low * n_gauss`.
    min_z = self.gauss_low * self.n_gauss
    z = np.full(n_samples, min_z)

    # Now `z` is initialized to the minimum, so only add to it when one
    # gets more than the minimum possible.
    d = self.gauss_high - self.gauss_low
    for t in range(self.n_gauss):
        got_high = np.random.binomial(size=n_samples, p=self.gauss_pr, n=1)
        z += d * got_high

    # Return total reward sums.
    return z

def free_plus_gauss(self, n_samples, f=None, maximize_log_reward=False):
    if self.n_gauss is None:
        raise ValueError("Must specify number of Gaussian bets.")

    s = self.free_last_state_(n_samples, f) # Reward is linear in `s`.
    gauss_z = self.gauss(n_samples, maximize_log_reward)
```

```
            z = s + gauss_z
            return z

        def free_distr_params(self, f=None):
            # Theoretical parameters of asymptotic log-normal distribution of
            # wealth from free betting
            f = self.bet_frac_(f)
            if f == 1:
                return -math.inf, math.inf

            # Else mean of corresponding normal distribution (log of wealth)
            # is E[ln(1 + B*f)] and variance is Var[ln(1 + B*f)].
            lg_p = np.log1p(f)
            lg_m = np.log1p(-f)
            pr_m = 1 - self.free_pr
            mean = self.free_pr * lg_p + pr_m * lg_m

            # Var[X] = E[(X - E[X])^2]
            lg_p -= mean
            lg_m -= mean
            var = self.free_pr * lg_p * lg_p + pr_m * lg_m * lg_m

            # Variance is only multiplied by `n_free`, not `n_free*n_free`,
            # because terms in sum of wealth are independent. (See central
            # limit theorem.)
            return self.n_free * mean, np.sqrt(self.n_free * var)

        def gauss_distr_params(self, maximize_log_reward=False):
            if self.take_gauss_mid_(maximize_log_reward):
                mu = self.n_gauss * self.gauss_mid
                sigma = 0  # Get middle payoff for sure.
                return mu, sigma

            pr_l = 1 - self.gauss_pr
            mean = self.gauss_pr * self.gauss_high + pr_l * self.gauss_low

            # Var[X] = E[(X - E[X])^2]
            h = self.gauss_high - mean
            l = self.gauss_low  - mean
            var = self.gauss_pr * h * h + pr_l * l * l

            return self.n_gauss * mean, np.sqrt(self.n_gauss * var)

    def cmp_free_hist_theory(z_samples, mu, sigma):
        # Compare histograms of generated samples (i.e. empirical distribution)
        # with theoretical log-normal probability density function and CDF.
```

```
    #z_samples = np.minimum(z_samples, 40.0) # "Zoom in" plot.
    z_min   = np.amin(z_samples)
    z_max   = np.amax(z_samples)
    z_range = np.linspace(z_min, z_max, num=200)

    z_pdf = stats.lognorm.pdf(z_range, s=sigma, loc=0, scale=np.exp(mu))
    z_cdf = stats.lognorm.cdf(z_range, s=sigma, loc=0, scale=np.exp(mu))

    # plt.subplots()
    # plt.plot(z_range, z_pdf)
    # bins = np.logspace(np.log10(0.1), np.log10(1000.0), 30) # Want linear axis, but log bi
    # plt.hist(z_samples, density=True, bins=bins)            # 40.0 is likely close to max.

    plt.subplots()
    plt.plot(z_range, z_cdf)
    plt.hist(z_samples, density=True, bins=200, cumulative=True)

def cmp_gauss_hist_theory(z_samples, mu, sigma):
    # Compare histograms of generated samples (i.e. empirical distribution)
    # with theoretical Gaussian probability density function and CDF.

    z_min   = np.amin(z_samples)
    z_max   = np.amax(z_samples)
    z_range = np.linspace(z_min, z_max, num=200)

    z_pdf = stats.norm.pdf(z_range, loc=mu, scale=sigma)
    z_cdf = stats.norm.cdf(z_range, loc=mu, scale=sigma)

    # plt.subplots()
    # plt.plot(z_range, z_pdf)
    # plt.hist(z_samples, density=True, bins=200)

    plt.subplots()
    plt.plot(z_range, z_cdf)
    plt.hist(z_samples, density=True, bins=200, cumulative=True)

seed = 123
random.seed(seed)
np.random.seed(seed)

sim = Sim(n_free=50, free_pr=0.55, n_gauss=50)

# 2.2
z_samples = sim.free(n_samples=5000)
mu, sigma = sim.free_distr_params()
```

```
cmp_free_hist_theory(z_samples, mu, sigma)

# 2.1
z_samples = sim.gauss(n_samples=5000)
mu, sigma = sim.gauss_distr_params()
cmp_gauss_hist_theory(z_samples, mu, sigma)

# 3.1
plt.subplots()
plt.hist(sim.free_plus_gauss(n_samples=5000, maximize_log_reward=False), bins=200)

# 2.1
plt.subplots()
plt.hist(sim.gauss(n_samples=5000, maximize_log_reward=True), bins=200)
plt.show()
```

8 Litsents

Mina, Andre Litvin,

1. annan Tartu Ülikoolile tasuta loa (lihtlitsentsi) minu loodud teose "Alternatiiv keskväärtuse maksimeerimisele", mille juhendaja on Raul Vicente, reprodutseerimiseks eesmärgiga seda säilitada, sealhulgas lisada digitaalarhiivi DSpace kuni autoriõiguse kehtivuse lõppemiseni.
2. Annan Tartu Ülikoolile loa teha punktis 1 nimetatud teos üldsusele kättesaadavaks Tartu Ülikooli veebikeskkonna, sealhulgas digitaalarhiivi DSpace kaudu Creative Commonsi litsentsiga CC BY NC ND 3.0, mis lubab autorile viidates teost reprodutseerida, levitada ja üldsusele suunata ning keelab luua tuletatud teost ja kasutada teost ärieesmärgil, kuni autoriõiguse kehtivuse lõppemiseni.
3. Olen teadlik, et punktides 1 ja 2 nimetatud õigused jäävad alles ka autorile.
4. Kinnitan, et lihtlitsentsi andmisega ei riku ma teiste isikute intellektuaalomandi ega isikuandmete kaitse õigusaktidest tulenevaid õigusi.

Andre Litvin

11.03.2023