Rozhodování, markovské rozhodovací procesy

Řešené úlohy

Shromáždil: Jiří Kléma, klema@labe.felk.cvut.cz

LS 2011/2012

Cíle materiálu:

Text poskytuje řešené úlohy jako podpůrný výukový materiál ke cvičením v předmětu A4B33ZUI.

1 Jednotlivá rozhodnutí, bayesovské rozhodování

Příklad 1. *(AIMA, 16.10): Jdete si koupit ojeté auto do bazaru. V úvahu připadá, že si auto před nákupem prověříte testem (kopnete do pneumatik, zavezete ho ke kamarádovi mechanikovi) a teprve pak se rozhodnete. Každé auto může být buď v dobrém nebo špatném stavu (s_+ a s_-). Nechť se rozhodujete o konkrétním autě a_1, jeho bazarová cena je 30000 Kč, tržní cena a_1 v dobrém stavu je 40000 Kč. Případná oprava auta (přechod ze špatného do dobrého stavu) stojí 14000 Kč. Odhadujete, že auto je v dobrém stavu a_{1+} s pstí 70%. Před nákupem můžete provést jeden konkrétní test t_1 za cenu 1000 Kč. Test určí, v jakém stavu auto je, ale s neurčitostí: $Pr(t_{1+}(a_1)|a_{1+}) = 0.8$ a $Pr(t_{1+}(a_1)|a_{1-}) = 0.35$.*

Vypočtěte střední čistý zisk pokud koupíte a_1 bez t_1.

$EU(kup+|\{\}) = \sum_{s\in\{+,-\}} U(s)Pr(s|kup+) = 40000-(0.7\times30000+0.3\times44000) = 40000 - 34200 = 5800$ Kč

Analogie dle klasického rozhodování:
$d^*(t) = \underset{kup+,kup-}{\text{argmin}} \sum_{s\in\{+,-\}} l(d,s)Pr(s|t) = \underset{kup+,kup-}{\text{argmin}} \sum_{s\in\{+,-\}} l(d,s)Pr(s) =$
$= \underset{kup+,kup-}{\text{argmax}} (10000 \times 0.7 - 4000 \times 0.3, 0) = \text{argmax}(5800, 0) = kup+$

Závěr 1: Nákup auta bez testu se vyplatí.

Použijte Bayesův teorém k určení psti, že auto je v dobrém stavu pro oba výsledky testu.

$Pr(a_{1+}|t_{1+}(a1)) = \frac{Pr(t_{1+}(a1)|a1+)\times Pr(a_{1+})}{Pr(t_{1+}(a1))} = \frac{0.8\times 0.7}{0.8\times 0.7+0.35\times 0.3} = \frac{0.56}{0.665} = 0.842$
$Pr(a_{1+}|t_{1-}(a1)) = \frac{Pr(t_{1-}(a1)|a_{1+})\times Pr(a_{1+})}{Pr(t_{1-}(a1))} = \frac{0.2\times 0.7}{0.2\times 0.7+0.65\times 0.3} = \frac{0.14}{0.335} = 0.418$

Najděte optimální rozhodnutí o nákupu pro oba výsledky testu.

$EU(\alpha_{t_1}|t_{1+}(a1)) = 40000-(0.842\times 30000+0.158\times 44000) = 40000-32240 = 7788$ Kč
$EU(\alpha_{t_1}|t_{1-}(a1)) = 40000-(0.418\times 30000+0.582\times 44000) = 40000-38120 = 1852$ Kč

Analogie dle klasického rozhodování:
$d^*(t_{1+}(a1)) = \underset{kup+,kup-}{\text{argmin}} \sum l(d,s)Pr(s|t) = \underset{kup+,kup-}{\text{argmin}} (10000\times 0.842-4000\times 0.158, 0) = \underset{kup+,kup-}{\text{argmin}} (7788, 0) = kup+$
$d^*(t_{1-}(a1)) = \underset{kup+,kup-}{\text{argmin}} \sum l(d,s)Pr(s|t) = \underset{kup+,kup-}{\text{argmin}} (10000\times 0.418-4000\times 0.582, 0) = \underset{kup+,kup-}{\text{argmin}} (1852, 0) = kup+$

Závěr 2: Nákup auta se vyplatí při pozitivním i negativním výsledku testu. Už z toho plyne nulová VPI testu – test nemá potenciál změnit rozhodnutí při jakémkoli výsledku.

Určete VPI testu t_1. Navrhněte optimální strategii pro potenciálního kupce auta a_1.

$EU(\alpha|\{\}) = \max(5800, 0) = 5800$ Kč
$EU(\alpha_{t_1}|t_{1+}(a1)) = \max(7788, 0) = 7788$ Kč
$EU(\alpha_{t_1}|t_{1-}(a1)) = \max(1852, 0) = 1852$ Kč

$VPI(t_1(a_1)) = (Pr(t_{1+}(a1))\times 7788+Pr(t_{1-}(a1))\times 1852)-5800 = (0.665\times 7788+0.335\times 1852)-5800 = 5800-5800 = 0$ Kč
Jde o "tvrdou" nulu, lze zjistit rozepsáním:
$Pr(t_{1+}(a1))\times(10000\times Pr(a_{1+}|t_{1+}(a1))-4000\times Pr(a_{1-}|t_{1+}(a1)))+Pr(t_{1-}(a1))\times(10000\times Pr(a_{1+}|t_{1-}(a1))-4000\times Pr(a_{1-}|t_{1-}(a1))) = 10000\times(Pr(a_{1+},t_{1+}(a1))+Pr(a_{1+},t_{1-}(a1)))-4000\times(Pr(a_{1-},t_{1+}(a1))+Pr(a_{1-},t_{1-}(a1))) = 10000\times Pr(a_{1+})-4000\times Pr(a_{1-}) = 5800$ Kč
$VPI(t1(a1))-Cost(t1(a1)) = -1000 < 0$

Závěr 3: Logický výsledek. Test v žádném z případů nevede ke změně rozhodnutí, má nulovou hodnotu, při započtení ceny dokonce zápornou. Ideální strategie je koupit auto

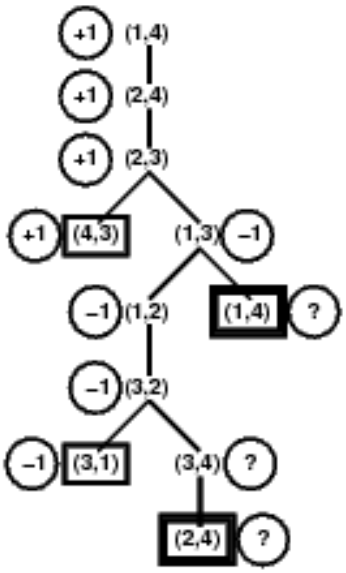
bez testu. Test by musel být výrazně citlivější, aby se vyplatil. Přesnost testu (přičemž triviální klasifikátor "dobrý stav" má přesnost 0.7):
$Pr(t_{1+}(a1), a_{1+}) + Pr(t_{1-}(a1), a_{1-}) = 0.8 \times 0.7 + 0.65 \times 0.3 = 0.755$

2 Markovské rozhodovací procesy

Příklad 2. *Uvažujme hru dvou hráčů na herním plánu o čtyřech polích. Každý z hráčů má jeden kámen a jeho cílem je dostat svůj kámen na opačnou stranu herního plánu (hráč A se z pole 1 musí dostat na pole 4, hráč B z pole 4 na pole 1). Vítězí ten hráč, kterému se to podaří prvnímu. Hráč A táhne jako první. Přípustné akce jsou pohyby vlevo a vpravo na sousední pole, nelze zůstat stát a vzdát se tahu, nelze táhnout mimo herní plán. Pokud je na vedlejším poli soupeřův kámen, je výsledkem pohybu přeskočení kamene (příklad: je-li A na pozici 3 a B na pozici 2 je výsledkem pohybu A vlevo posun A na pozici 1).*

Který z hráčů vyhraje? Naznačte klasické řešení problému pomocí prohledávání stavového prostoru.

Stav hry je dán pozicí obou hráčů, lze jej tedy zapsat jako uspořádanou dvojici (s_A, s_B). Existuje celkem 11 různých dosažitelných stavů (pro každou pozici kamene A existují tři pozice kamene B, stav $(4, 1)$ je nedosažitelný). Standardní řešení je realizováno procedurou MiniMax. Herní strom je na obrázku níže (hodnocení je z pohledu hráče A, který je tedy maximalizačním hráčem).
Jediný problém je v tom, že úloha obsahuje cykly a standardní MiniMax s prohledáváním do hloubky by skončil v nekonečném cyklu. Proto je třeba provést drobné úpravy. Expandované stavy ukládáme na zásobník a pokud je detekován cyklus, označí se hodnota stavu jako "?" a větev se ukončí. Při propagaci ohodnocení se potom racionálně předpokládá, že max(1,?)=1 a min(-1,?)=-1. To pro danou hru, kde nejsou vítězství a prohry různé kvality, jako řešení zcela postačuje.



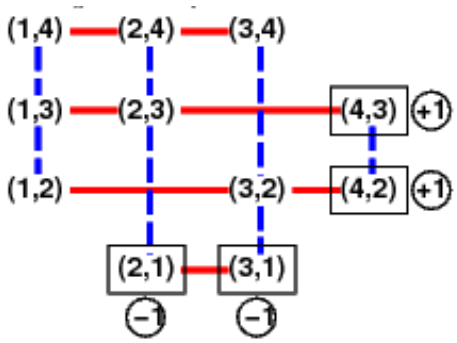
Závěr 1: Pro hry dvou hráčů je klasickým řešením MiniMax. Při optimální strategii obou hráčů zvítězí A.

Lze tuto úlohu formulovat a řešit jako MDP? Je to výhodné? Jak by se úloha musela změnit, aby to výhodné bylo?

Závěr 2: Každý prohledávací problém lze formulovat jako MDP. Převod je rutinní: stavy a akce jsou identické, cílové stavy se mapují na stavy terminální u MDP, přechodová matice je deterministická, odměna je invertovanou cenovou funkcí.
V případě deterministických akcí použití MDP výhodné není. Formalismus je zbytečně složitý a výpočetně náročný. Vhodným řešením by se stalo pro stochastické varianty problému.

Problém formulujte jako MDP. Nechť je $V_A(s)$ je hodnota stavu pokud je na tahu hráč A, $V_B(s)$ je hodnota stavu pokud je na tahu hráč B. Odměna ve stavu s nechť je $R(s)$, pro terminální stavy vítězné pro A je 1, pro terminální stavy vítězné pro B je -1. Nakreslete graf stavového prostoru. Zapište Bellmanovy rovnice pro oba hráče a aplikujte tyto rovnice v rámci hodnotové iterace. Formulujte ukončovací podmínku iterace.

Graf stavového prostoru je na obrázku níže. Tahy A jsou značené plnou červenou čarou, tahy B přerušovanou modrou čarou.



Bellmanovy rovnice také vychází z principu MiniMaxu:
$V_A(s) = R(s) + \max_a P^a_{ss'} V_B(s')$
$V_B(s) = R(s) + \min_a P^a_{ss'} V_A(s')$

$R(s)$ bude použito pouze v terminálních stavech, ohodnocení zbylých stavů je nulové a řídí se pouze následníky. Hráč A maximalizuje ohodnocení, hráč B jej naopak minimalizuje. Protože akce jsou deterministické, každá akce má jednotkovou pst pro jednoho následníka a nulovou pro všechny zbylé stavy.

Hráči se v tazích střídají, proto střídáme i aplikace příslušných Bellmanových rovnic. Na počátku je ohodnocení 11 dostupných stavů dáno pouze pro $R(s)$ terminálů, pro zbytek stavů je nulové. Ohodnocení se šíří postupně, využíváme graf stavového prostoru, viz tabulka:

s	(1,4)	(2,4)	(3,4)	(1,3)	(2,3)	(4,3)	(1,2)	(3,2)	(4,2)	(2,1)	(3,1)
V_A	0	0	0	0	0	+1	0	0	+1	-1	-1
V_B	0	0	0	0	-1	+1	0	-1	+1	-1	-1
V_A	0	0	0	-1	+1	+1	-1	+1	+1	-1	-1
V_B	-1	+1	+1	-1	-1	+1	-1	-1	+1	-1	-1
V_A	+1	+1	+1	-1	+1	+1	-1	+1	+1	-1	-1
V_B	-1	+1	+1	-1	-1	+1	-1	-1	+1	-1	-1

Ukončovací podmínkou je nulová změna ve vektoru hodnot pro jednoho z hráčů (tedy shoda $V_A(s)$ s vektorem $V_A(s)$ o dva tahy dříve nebo stejné srovnání pro $V_B(s)$). V tabulce výše je shoda pro poslední dva výpočty $V_B(s)$. Je jasné, že ke změně nemůže dojít už ani u $V_A(s)$, bude se odvozovat z identického $V_B(s)$.
Je třeba si uvědomit, že vektory ohodnocení stavů obou hráčů se v ekvilibriu neshodují. $V_A(s)$ předpokládá, že na tahu je hráč A a naopak. Proto nelze při ukončení srovnávat $V_A(s)$ a $V_B(s)$, mj. stav (3,2) své ohodnocení z principu přepíná (vyhraje ten, kdo je právě na tahu).

Závěr 3: MDP současně řeší problém pro zahájení oběma hráči. Ohodnocení terminálních stavů je dané apriori. Ve stavech (2,4) a (3,4) vítězí hráč A bez ohledu na to kdo je na tahu. Ve stavech (1,3) a (1,2) vítězí hráč B bez ohledu na to kdo je na tahu (jsou zrcadlovým obrazem stavů (2,4) a (3,4)). Ve stavech (1,4), (2,3) a (3,2) vítězí ten, kdo je právě na tahu.
MiniMax strom uvedený dříve používá na různých úrovních stromu různá ohodnocení, de facto tedy kombinuje $V_A(s)$ a $V_B(s)$ dle úrovně stromu.

Příklad 3. Za dveřmi je tygr (řešení jako POMDP). *Stojíte v místnosti, z níž vedou dvoje dveře. Víte, že za jedněmi dveřmi je hladový tygr, druhé dveře garantují bezpečný odchod z místnosti. Tygr občas zařve. Řev je slyšet, ale z jednoho poslechu není úplně zřejmé, odkud řev vychází. Chcete se z místnosti bezpečně a rychle dostat, v každém okamžiku se můžete rozhodnout mezi třemi volbami: otevřít dveře vlevo, otevřít dveře vpravo nebo počkat až tygr znovu zařve. Pracujte s následujícími ohodnoceními: otevření nesprávných dveří odpovídá ztrátě 100, otevření správných dveří zisku 10, čekání je spojeno se ztátou 1, při každém zařvání se zmýlíte v odhadu směru v 15% případů (ukážete na jedny ze dveří, ale správně to bude jen v 85% situací), pokud bereme v úvahu celou sekvenci náslechů, omyly jsou vzájemně nezávislé, tygr mezi řvaním svoji pozici nemění.*

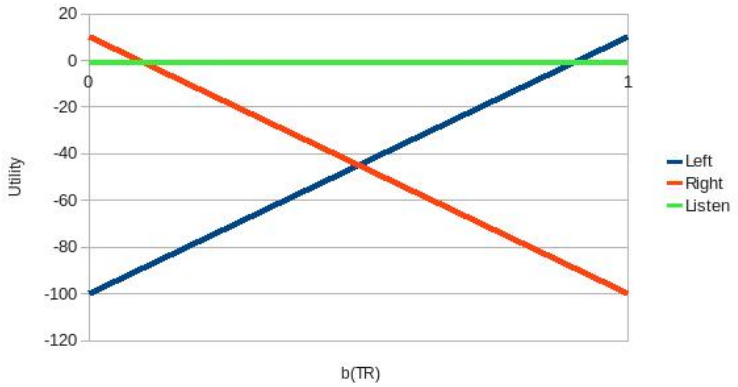
Problém formalizujte jako částečně pozorovatelný markovský rozhodovací proces.

POMDP $= \{S, A, P, R, O, \Omega\}$,
skryté stavy: S=$\{TL, TR, STOP\}$, TL $\sim$ tygr vlevo, TR $\sim$ tygr vpravo, STOP $\sim$ konec hry (byly otevřeny dveře)
akce: A=$\{Li, L, R\}$, Li $\sim$ poslouchej=čekej na další zařvání, L $\sim$ otevři levé dveře, R $\sim$ otevři pravé dveře,
pozorování: O=$\{LL, LR\}$, LL $\sim$ tygr slyšen vlevo, LR $\sim$ tygr slyšen vpravo,
přechodové psti (P): $Pr(TL|TL, Li) = 1$, $Pr(TR|TL, Li) = 0$, $Pr(TL|TL, L) = 0$, $Pr(TR|TL, L) = 0$,$Pr(STOP|TL, L) = 1$ (uváděno pouze pro TL, TR je symetrický),
senzorický model (Ω): $Pr(LL|TL, Li) = 0.85$, $Pr(LR|TL, Li) = 0.15$,
funkce odměny: $R(Li, TL) = -1$, $R(L, TL) = 10$, $R(R, TL) = -100$.

Nalezněte optimální plán délky 1 jako funkci belief. Tj. navrhněte optimální akci v závislosti na tom jakou pravděpodobnost přiřazujete skrytým stavům. V jakých bodech belief prostoru se bude rozhodnutí měnit?

Vzhledem k tomu, že máme pouze dva stavy, stačí belief reprezentovat jako jediné reálné číslo od 0 do 1. Zapisujme jej jako $b(TL)$, $b(TR)$ je doplňkem do jedné. Užitek

akcí bude funkcí jediné proměnné. Protože víme, že bude lineární funkcí b, stačí jej pro všechny akce spočítat pouze v krajních bodech belief prostoru, tedy pro situace, kdy jsme si naprosto jisti, že tygr je vlevo nebo naopak vpravo. Viz obrázek níže.



Závěr 1: Z obrázku je také zřejmé, že pro $b(TR) > 0.9$ se vyplatí volit akci L, pro $b(TR) < 0.1$ se vyplatí volit akci R. Pro střední oblast b je nejvýhodnější akce Li. Hodnoty 0.1 a 0.9 vychází z rovnice: $-100b(TR) + 10(1 - b(TR)) = -1$, resp. $-100(1 - b(TR)) + 10 * b(TR) = -1$.

Kolik je podmíněných plánů délky 2? Určete užitek alespoň jednoho z nich (opět půjde o funkci belief). Bude některý z plánů čistě dominován plány jinými?

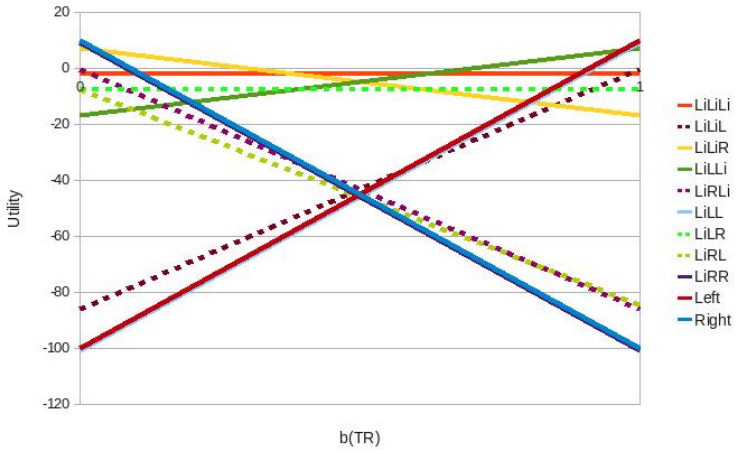
Podmíněný plán délky 2 je takový plán, který určuje akci pro daný okamžik a následně akci pro okamžik po příštím zařvání. Protože zařvání můžeme slyšet ze dvou stran a pro každý směr můžeme volit jinou akci, bude mít plán tři akce [A_1 if LR then A_2 else A_3], kódovat budeme jako $[A_1A_2A_3]$. Teoreticky máme 27 plánů (sekvence délky 3 nad abecedou tří akcí, tj. 3^3 možností). Všechny plány, které nezačínají akcí Li, ale vedou na restart hry, tj. existuje pouze 9 podmíněných plánů délky 2.

Triviální ohodnocení je pro plán $[Li, Li, Li]$:
$\alpha_{[LiLiLi]}(b(TR) = 0) = \alpha_{[LiLiLi]}(b(TR) = 1) = -2$,

Složitější pro plán $[LiLLi]$:
$\alpha_{[LiLLi]}(b(TR) = 0) = R(Li, TL) + Pr(TR|TR, Li)(Pr(LL|TR, Li) * \alpha_{[L]}(0) + Pr(LR|TR, Li)\alpha_{[Li]}(0)) + Pr(TL|TR, Li)(Pr(LL|TL, Li) * \alpha_{[L]}(0) + Pr(LR|TL, Li)\alpha_{[Li]}(0)) = -1 + 1(0.15 * -100 + 0.85(-1)) + 0(\dots) = -16.85$,
$\alpha_{[LiLLi]}(b(TR) = 1) = R(Li, TR) + Pr(TL|TL, Li)(Pr(LL|TL, Li) * \alpha_{[L]}(1) +$

$Pr(LR|TL, Li)\alpha_{[Li]}(1))+Pr(TR|TL, Li)(Pr(LL|TR, Li)\alpha_{[L]}(1)+Pr(LR|TR, Li)\alpha_{[Li]}(1)) = -1 + 1 * (0.85 * 10 + 0.15 * -1) + 0 * (\dots) = 7.35,$

Z 9 plánů počínajících Li je jen 5 dominujících: $[LiRR]$ pro beliefs¡0.019, $[LiLiR]$ pro beliefs 0.019-0.39, $[LiLiLi]$ pro beliefs 0.39-0.61, $[LiLLi]$ pro beliefs 0.61-0.981, $[LiLL]$ pro beliefs¿0.981. V prvním případě je natolik jasné, že stav je TR, že jakýkoli směr ponechává akci R, ve druhém je třeba TR potvrdit LR, jinak posloucháme dále, uprostřed je nejistota tak velká, že jakýkoli směr nemůže vést na otevření dveří, dále symetricky ....
Je evidentní, že plán $[LiRLi]$ nedává smysl (pokud slyším tygra řvát vpravo, težko pak otevřu dveře vpravo a pokud ho uslyším vlevo, tak budu dále vyčkávat), $[LiLiL]$ je jeho doplňkem, $[LiRL]$ je také nesmysl (otevřu pravé dveře pokud slyším tygra vpravo a naopak). Potenciálně zajímavý je $[LiLR]$, ale neprosadí se díky aktuální parametrizaci, penalta za otevření nesprávných dveří je příliš velká.



Při celkovém řešení hry je třeba tyto plány srovnat s plány délky 1, které hru také končí, tedy L a R, ty evidentně předčí $[LiLL]$, resp. $[LiRR]$ pro krajní beliefs.
Pro doplňení bod přepnutí mezi $[LiRR]$ a $[LiLiR]$: $7.35 - (16.85 + 7.35)x = -2, x = 9.35/(16.85+7.35) = 0.39$ Ještě bod přepnutí mezi $[LiLiR]$ a $[LiLiLi]$: $7.35-(16.85+7.35)x = 9 - (101 + 9)x, x = (9 - 7.35)/(110 - 24.2) = 0.019$

Závěr 2: Je celkem 9 podmíněných plánů délky 2. Nedominovaných je ale pouze 5, pokud bereme v úvahu i plány délky 1, tak dokonce jenom 3.

Kolikrát je třeba na začátku hry slyšet řev ze stejné strany předtím než se vyplatí otevřít jedny ze dveří? Zdůvodněte.

Z bodu b) víme, že b(TR) musí být menší než 0.1 nebo větší než 0.9, toho dosáhneme po dvou souhlasných pozorováních buď [LL, LL] nebo [LR, LR]. Testujme pro [LR, LR] (pozor, $b_2(TR)$ nelze obecně počítat jako $(1 - 0.15^2) = 0.9775$):

$b_0(TR) = 0.5$ (začátek hry),
$b_1(TL) = \alpha Pr(LR|TL, Li)(Pr(TL|TL, Li) * b_0(TL) + Pr(TL|TR, Li) * b_0(TR)) =$
$\alpha * 0.15 * (1 * 0.5 + 0 * 0.5) = \alpha * 0.15 * 0.5,$
$b_1(TR) = \alpha Pr(LR|TR, Li)(Pr(TR|TR, Li) * b_0(TR) + Pr(TR|TL, Li) * b_0(TL)) =$
$\alpha * 0.85 * (1 * 0.5 + 0 * 0.5) = \alpha * 0.85 * 0.5,$
$b_1(TL) + b_1(TR) = 1 \ldots \alpha = 2 \ldots b_1(TL) = 0.15, b_1(TR) = 0.85,$
jedno zařvání nestačí

$b_2(TL) = \alpha Pr(LR|TL, Li)(Pr(TL|TL, Li) * b_1(TL) + Pr(TL|TR, Li) * b_1(TR)) =$
$\alpha * 0.15 * (1 * 0.15 + 0 * 0.5) = \alpha * 0.15^2,$
$b_2(TR) = \alpha Pr(LR|TR, Li)(Pr(TR|TR, Li) * b_1(TR) + Pr(TR|TL, Li) * b_1(TL)) =$
$\alpha * 0.85 * (1 * 0.5 + 0 * 0.5) = \alpha * 0.85^2,$
$b_2(TL) + b_2(TR) = 1 \ldots \alpha = 1.34 \ldots b_1(TL) = 0.03, b_1(TR) = 0.97,$
dvě zařvání stačí.

Závěr 3: Tygr musí být na začátku hry slyšet alespoň dvakrát ze stejné strany, aby se vyplatilo přestat čekat a otevřít dveře (samozřejmě otavřeme ty, od kterých nebyl řev slyšet.)