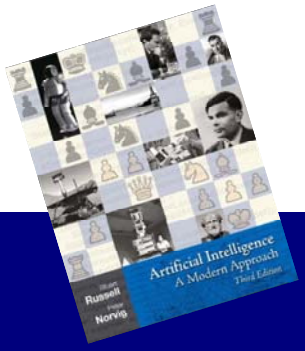


Umělá inteligence II

Roman Barták, KTIML

roman.bartak@mff.cuni.cz
http://ktiml.mff.cuni.cz/~bartak

8

Úvodem

- Pokud agent ví, kde je (**plně pozorovatelný svět**), potom pro každý stav umíme doporučit akci maximalizující očekávaný užitek (**strategie**) a to i v případě, že **výsledek akce je nejistý**.

$$U(s) = R(s) + \gamma \max_{a \in A(s)} \Sigma_{s'} P(s'|s,a)\, U(s')$$

- A co když **nevíme, kde agent je**?
  - částečně pozorovatelný Markovský rozhodovací problém (POMDP)
- A co když je **agentů více**?
  - teorie her (hledání strategie pro agenta)
  - tvorba pravidel (hledání pravidel hry pro maximalizaci globálního užitku)

POMDP

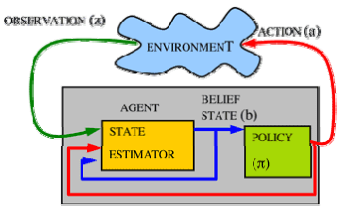
- Pro rozhodování o další akci máme k dispozici:
  - přechodový model P(s`|s,a)
  - ocenění stavů R(s)
  - senzorický model P(e|s)
- Hovoříme o **částečně pozorovatelném Markovském rozhodovacím problému** (POMDP).
- Pokud agent neví, kde je (částečně pozorovatelný svět), potom nemůže provést akci $\pi(s)$ doporučenou pro stav s.
- Jak modelovat neznámý stav?
  - **domnělý (belief) stav**
  - pravděpodobnostní rozložení přes všechny stavy – vektor pravděpodobností každého stavu
  - b(s) = pravděpodobnost přiřazená stavu s v domnělém stavu b
- Nyní budeme hledat **strategii pro domnělé stavy** $\pi(b)$.

Řešení POMDP

- **Jak zjistíme další domnělý stav** na základě současného domnělého stavu a akce?

  $b`(s`) = \alpha\ P(e|s`)\ \Sigma_s\ P(s`|s,a)\ b(s)$

  b` = FORWARD(b,a,e)
  - Je to vlastně **filtrace** na základě předchozích pozorování a provedených akcí, kterou umíme dělat inkrementálně.
- Optimální akce záleží pouze na aktuálním domnělém stavu, **optimální strategii** tedy můžeme definovat pro domnělé stavy $\pi^*(b)$.
- **POMDP agent** opakuje v cyklu následující kroky:
  1. **vybere akci** na základě současného domnělého stavu $a = \pi^*(b)$
  2. **získá pozorování** (vjem) e okolí
  3. **vypočte nový domnělý stav** b` = FORWARD(b,a,e)



Od POMDP k MDP

- Následující domnělý stav b‘ záleží **deterministicky** na aktuálním domnělém stavu b, pozorování e a akci a.
  - b‘ = FORWARD(b,a,e)
- **Jak odhadnout následující domnělý stav, když ještě nemáme k dispozici pozorování?**

  $P(e|a,b) = \Sigma_{s'} P(e|a,b,s') P(s'|a,b)$
  $= \Sigma_{s'} P(e|s') P(s'|a,b)$
  $= \Sigma_{s'} P(e|s') \Sigma_s P(s'|a,s) b(s)$
  - Jaká je teď pravděpodobnost dosažení b‘ z b akcí a?

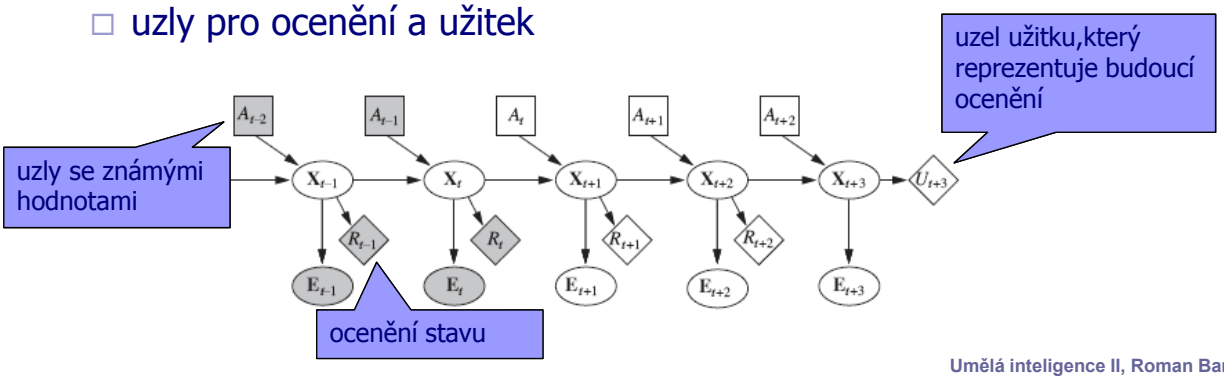
    $P(b'|b,a) = P(b'|a,b) = \Sigma_e P(b'|a,b,e) P(e|a,b)$
    $= \Sigma_e P(b'|a,b,e) \Sigma_{s'} P(e|s') \Sigma_s P(s'|a,s) b(s)$

    kde $P(b'|a,b,e) = 1$, pokud b‘ = FORWARD(b,a,e), jinak 0
- Ještě zbývá **ocenění domnělého stavu**

  $r(b) = \Sigma_s b(s)R(s)$
- **Dohromady tedy máme MDP!**
  - Hledání optimální strategie pro stavy v POMDP lze převést na hledání optimální strategie v MDP pro prostor domnělých stavů.
  - Pozor, odpovídající MDP pracuje s mnoho-dimenzionálním spojitým prostorem!

Hledání strategie

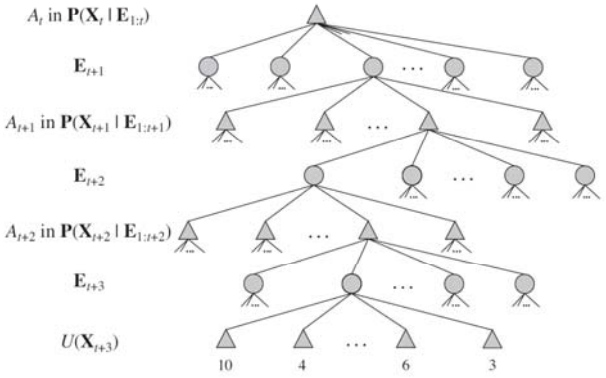
- Bohužel prezentované techniky řešení MDP nemůžeme přímo použít, protože domnělých stavů je nekonečně mnoho.
- Místo strategie budeme pracovat s **podmíněným plánem**.
  - akce je vybrána na základě pozorování a předchozích domnělých stavů (strom akcí)
- **Iteraci hodnot** lze uzpůsobit pro podmíněné plány (postupně přidáváme další akce do plánu), ale to je příliš neefektivní.
- Použijeme **dynamické rozhodovací sítě** a techniku pohledu dopředu.
  - dynamická Bayesovská síť pro přechodový a senzorický model
  - uzly pro ocenění a užitek



Pohled dopředu

dynamická rozhodovací síť

- Při hledání podmíněného plánu dané délky postupujeme stejně jako při **hraní her**.
  - roli oponenta hrají pozorování



- trojúhelníky reprezentují **domnělé stavy** a vedou z nich hrany pro možné akce
  - domnělý stav získáme z cesty, která k němu vede (na ní známe pozorování)
- kolečka reprezentují možné **pozorování** a z nich vzniklé domnělé stavy
- nepotřebujeme speciální uzly pro různé efekty akcí, protože změna domnělého stavu je deterministická

- Pro hledání řešení můžeme použít podobný postup jako **algoritmus MINIMAX** (po cestě ještě přidáváme ocenění domnělých stavů).
- Hloubku stromu (rozvinutí sítě) určuje faktor slevy γ.

Rozhodování s více agenty

- Uvažujme nyní, že rozhodování je ovlivněno jiným typem neurčitosti – jinými agenty, jejichž chování je zpětně ovlivněno našimi rozhodnutími.
- **tvorba agentů** (agent design)
  - teorie her umožňuje analyzovat rozhodnutí agentů a počítat očekávané užitky
- **tvorba pravidel** (mechanism design)
  - inverzní teorie her umožňuje nastavit pravidla hry tak, aby celkový užitek byl maximalizován, když každý agent maximalizuje svůj užitek

Hry s jedním tahem

- Podíváme se jen na hry, kde všichni hráči (agenti) najednou zvolí svůj tah (akci)
  - přesněji hráči volí tah bez znalosti tahů ostatních hráčů
- **Hra s jedním tahem** je definována
  - **hráči**, například O (odd) a E (even)
  - **akcemi** (tahy), například jeden prst a dva prsty
  - **funkcí odměny** (payoff) určující pro každou kombinaci akcí hráčů odměnu každého hráče
    například hra Morra se dvěma prsty (hráči O a E každý ukážou jeden nebo dva prsty a pokud je celkový počet prstů f lichý, získá O od E f dolarů, jinak získá E od O f dolarů)

	O: jeden	**O: dva**
E: jeden	E=+2, O=-2	E=-3, O=+3
E: dva	E=-3, O=+3	E=+4, O=-4

Hry s jedním tahem

strategie a řešení

- Hráči hry volí **strategii**, jaké tahy budou volit.
  - **čistá strategie**
    - deterministická strategie, která doporučuje konkrétní akci
  - **smíšená strategie**
    - randomizovaná strategie, která určuje pravděpodobnost volby akce
    - [p,a; (1-p),b]
- **Výstup hry** je numerická odměna každého hráče.
- **Řešení hry** je přiřazení racionální strategie každému hráči.
  - Otázkou je, co přesně znamená **racionální strategie**.

Vězňovo dilema

- Uvažujme následující „hru"
  - Dva zloději Alice a Bob byli chyceni při činu a nyní jsou odděleně vyslýcháni.
  - Oba dostali nabídku dosvědčit, že partner je šéf bandy výměnou za propuštění (partner dostane 10 let).
  - Pokud budou proti sobě svědčit navzájem, dostanou po 5 letech.
  - Pokud odmítnou svědčit, dostanou po 1 roku.
- Jak se mají rozhodnout?
  - racionální volba je **svědčit**.

	Alice: svědčit	**Alice: odmítnout**
Bob: svědčit	A=-5, B=-5	A=-10, B=0
Bob: odmítnout	A=0, B=-10	A=-1, B=-1

Dominance

- Rozhodnutí svědčit u vězňova dilematu je **dominantní strategií**.
  - strategie s hráče p **silně dominuje** strategii s` pokud je výstup hry pro s pro hráče p vždy lepší než pro s` při použití libovolné strategie ostatních hráčů
  - **slabá dominance** znamená, že s není nikdy horší než s`
- **Racionální** volbou je přirozeně **volit dominantní strategii**.
- Pokud oba hráči volí dominantní strategii je kombinace těchto strategií rovnováhou (equilibrium) dominantních strategií.
- Rovnováha je obecně přidělení strategií, kdy žádný hráč nic nezíská, pokud strategii změní – **Nash equilibrium**.
- Výstup hry je **Pareto optimální**, pokud není žádný jiný výstup lepší pro všechny hráče.
  - Výstup A je **Pareto dominován** jiným výstupem B, pokud všichni hráči preferují B před A.

- **Dilema** u vězňů je v tom, že výstup v rovnovážném stavu (svědčit,svědčit) je Pareto dominován výstupem (odmítnout,odmítnout).

Ne-dominantní strategie

- Uvažujme nyní další „hru".
  - Firma Acme plánuje výrobu herní konzole a rozhoduje se mezi použitím DVD a Blu-ray.
  - Firma Best připravuje novou hru a také se rozhoduje mezi distribucí na DVD a Blu-ray.
  - Přirozeně, pokud se firmy shodnou, potom z toho budou obě profitovat.

	Acme: bluray	**Acme: dvd**
Best: bluray	A=+9, B=+9	A=-4, B=-1
Best: dvd	A=-3, B=-1	A=+5, B=+5

- **Neexistuje rovnováha** dominantních strategií (nejsou dominantní strategie), ale jsou tady **dvě Nashova equilibria**.
- Je zde více přijatelných řešení, ale pokud si každá firma vybere jiné, potom na to obě doplatí.
- Jak na to?
  - obě firmy zvolí řešení odpovídající **Pareto optimálnímu výstupu** (bluray, bluray)
  - A co když je více takových řešení?
    - agenti mezi sebou komunikují
    - **koordinační hry**

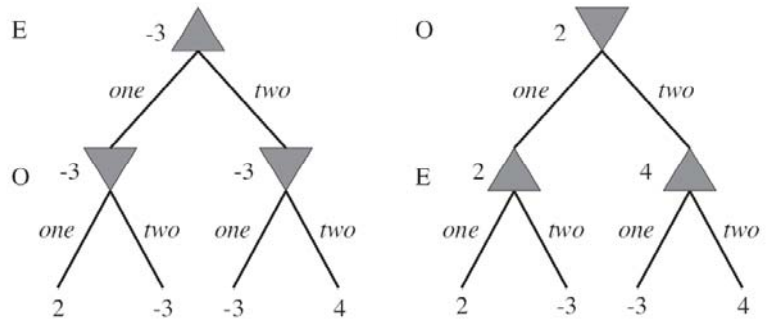
Smíšené strategie

- Vraťme se nyní ke hře Morra se dvěma prsty.
  - zde neexistuje čistá strategie
    - pokud je počet prstů sudý, chce O změnit strategii
    - pokud je počet prstů lichý, chce E změnit strategii
  - je potřeba hledat **smíšenou strategii**
- von Neumann navrhl **metodu maximin** pro hledání optimální smíšené strategie pro **hry dvou hráčů s nulovým součtem**
  - protože odměny hráčů se doplňují, stačí se dívat na odměnu jednoho hráče

Maximin metoda

čisté strategie

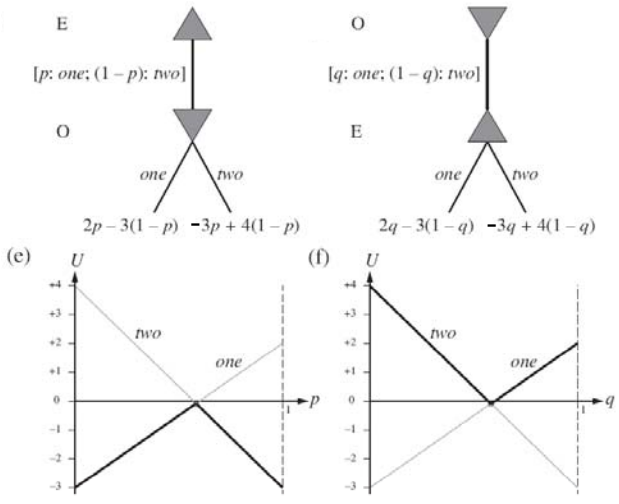
- Zkusme změnit pravidla následujícím způsobem
  - nejprve volí E a potom O (se znalostí volby E)
    - klasický algoritmus minimax pro hledání optimální strategie
    - O je samozřejmě ve výhodě, takže tak dostaneme dolní odhad odměny pro E (-3)
  - podobně můžeme nechat volit O a potom E
    - získáme horní odhad odměny pro E (2)



Maximin metoda

smíšené strategie

- Nás ale zajímají smíšené strategie [p,jeden; (1-p),dva]



- Pokud první hráč zvolí smíšenou strategii a druhý ji zná, může použít čistou strategii dle p.
- levý obrázek nám dá p do strategie E
  - maximalizujeme odměnu pro E (p=7/12 a odměna bude -1/12)
- pravý obrázek nám dá q do strategie O
  - minimalizujeme odměnu E (q=7/12 a odměna bude -1/12)

- **Optimální strategie** pro oba hráče je [7/12,jeden; 5/12,dva]
  - maximin rovnováha (je to také Nash equilibrium)
  - je zřejmé, že hra je výhodnější pro O

Tvorba pravidel

- Zatím jsme se zabývali hledáním racionální strategie pro danou hru.
- Zkusme to naopak – hledat pravidla hry taková, že pokud hru hrají racionální agenti, tak výsledkem je maximalizace celkové míry užitku.
- **Tvorba pravidel** (mechanism design) nebo také **inverzní teorie her**.
- Používá se v ekonomice nebo politických vědách, obecně pro synchronizaci nezávislých racionálně uvažujících agentů.
- Problém se skládá z:
  - jazyka popisujícího možné strategie agentů
  - jednoho zvoleného agenta – ústředí – který od agentů shromažďuje volby dle jejich strategií
  - pravidlo (známé všem agentům), podle kterého ústředí určí odměnu všem agentům dle jejich voleb

Aukce

- **Aukce** je mechanismus pro prodání zboží skupině agentů - dražitelů.
- Budeme uvažovat aukce jednoho prvku (zboží).
- Agent má typicky vlastní **soukromé ocenění** v_i nabízeného zboží.
  - starý rozbitý nábytek má jinou hodnotu pro sběratele a jinou pro člověka, který si chce vybavit byt
- Někdy je **hodnota zboží společná**, ale různí agenti mají různé informace o tom, kolik skutečná hodnota je.
- **Princip aukce**
  - při aukci má každý dražitel v jistém okamžiku možnost dát nabídku b_i
  - zboží získá nejvyšší nabídka, ale cenu, kterou bude platit, nemusí být b_{max} (záleží na pravidlech aukce)

Anglická aukce

- Asi nejznámější typ aukce, také **aukce s rostoucí nabídkou**.
- Postup anglické aukce
  - Ústřední agent začne s minimální nabídkou b_{min}
  - Pokud je nějaký dražitel ochoten tuto cenu zaplatit, nabídne ústřední agent cenu $b_{min} + d$ pro nějaké navýšení d.
  - Aukce končí ve chvíli, kdy žádný dražitel není ochotný nabídnout více.
  - Zboží získává agent s poslední nabídkou a platí tuto nabídku.
- Jak víme, že to funguje?
  - snahou je prodat zboží za největší cenu a zároveň maximalizovat globální užitek
  - **aukce je efektivní** pokud zboží získá agent, který si ho nejvíce cení
- Anglická aukce je zpravidla efektivní a maximalizující příjem, za předpokladu, že
  - je dostatek dražitelů
  - mezi dražiteli není tajná dohoda ať už přímá nebo prostřednictvím pravidel aukce

Anglická aukce
problém

- Hlavním nebezpečím anglické aukce je dohoda dražitelů na manipulaci s cenami.
- Příklad manipulace cen prostřednictvím mechanismu aukce
  - v roce 1999 v Německu dražili 10 bloků frekvencí pro mobilní telefony
  - pravidlem bylo, že další nabídka musí navýšit předchozí nabídku o alespoň 10% za blok
  - aukce se účastnili pouze dva dražitelé, Mannesman a T-Mobile
  - Mannesman dal první nabídku na bloky 1-5 20 miliónu DEM a na bloky 6-10 18,18 miliónu DEM.
  - T-Mobile interpretoval tuto nabídku tak, že Mannesman nabízí podělení se o frekvence, kdy každý dostane půlku frekvencí za 20 miliónů DEM
  - Jak to poznali?
    - Navýšení 18,18 miliónu o 10% dá částku 19,99 miliónu!

Odhalení pravdy

- Prodávající i celkový užitek zpravidla profituje z většího množství dražitelů.
- Jednou z cest, jak přilákat dražitele, je zjednodušit pravidla aukce.
- Přesněji řečeno umožnit dražitelům mít **dominantní strategii**, tj. strategii fungující nezávisle na strategiích ostatních agentů.
  - agent nemusí ztrácet část průzkumem strategií ostatních agentů
- Taková strategie obvykle zahrnuje odhalení vlastního ocenění zboží v_i – **odhalení pravdy**.
- Anglická akce má řadu žádoucích vlastností:
  - dominantní strategie spočívá v účasti na aukci, dokud je nabízená cena menší než v_i
  - není to ale úplné odhalení pravdy, protože u vítězného agenta známe pouze dolní odhad jeho v_i

Obálková metoda

- Další nevýhody anglické aukce
  - je-li jeden silný dražitel, o kterém ostatní očekávají, že může nabídnout největší cenu, tak se ostatní aukce nezúčastní a silný dražitel získá zboží za nejmenší nabídkou cenu (**odrazení od soutěže**)
  - agenti se musí sejít na jednom místě a věnovat celé aukci čas (**cena komunikace**)
- Alternativním řešením je **aukce se zalepenou nabídkou**.
  - každý agent pošle svoji jedinou nabídku do ústředí
  - vyhraje nejvyšší nabídka
- Zde není jednoduchá dominantní strategie
  - nabídka agenta se odvozuje od očekávaných nabídek ostatních agentů
  - nechť v_i je můj užitek ze získání zboží a b_0 je očekávané maximum nabídek ostatních agentů
  - moje nabídka bude $b_0+\varepsilon$ (pro malé ε), pokud je to méně než v_i
- Agent s největším v_i nemusí nutně vyhrát, tj. je omezena výhoda nejsilnějšího hráče (větší soutěživost).

Vickreyho aukce

- Jednoduchou změnou lze obálkovou metodu pro dražitele zjednodušit, aby dominantní strategií bylo nabízet vlastní hodnotu zboží v_i.
- **Vickreyho aukce** (aukce se zalepenou obálkou a druhou cenou)
  - vítěz aukce dle nejvyšší nabídky platí cenu druhé nejvyšší nabídky
- Rozbor dominantní strategie
  - užitek (zisk) agenta i při nabídce b_i, hodnotě v_i a nejlepší nabídce ostatních agentů b_0
  - $(v_i - b_0)$ pro $b_i > b_0$, jinak 0
    - když $(v_i - b_0) > 0$, tak každá nabídka, která vyhraje aukci, je optimální, speciálně pak v_i
    - když $(v_i - b_0) < 0$, tak každá nabídka, která prohraje aukci, je optimální, speciálně pak v_i
    - v_i je optimální pro oba případy a je to jediná nabídka, která má tuto vlastnost

Společné zboží

- Uvažujme hru, kde agenti soupeří o část společného „zboží".
- Příklad (redukce znečištění)
  - Každý stát má volbu
    - může redukovat znečištění za cenu -10 bodů, které je na to potřeba vynaložit
    - znečištění redukovat nebude za cenu -5 bodů, které je potřeba vynaložit na zdravotní náklady a cenu -1, kterou přispěje všem ostatním státům
  - Jaká je volba jednotlivých států?
    - přirozeně pokračovat ve znečišťování (racionální rozhodnutí)
  - Výsledek
    - při 100 státech získá každá země užitek -104
    - Pokud by všichni přestali znečišťovat, byl by užitek každé země -10.
- **Tragédie společného**
  - Pokud nikdo nemusí platit za používání společného zdroje, potom se ho snaží co nejvíce využít, což přináší menší celkový užitek všem.
  - Podobné vězňově dilematu – existuje lepší řešení pro všechny, ale není způsob, jak se k němu může racionální agent dobrat.
- Jak to řešit?
  - zavedení **daně za využívání společných zdrojů** (emisní povolenky)

VCG aukce

- Tragédii společného lze řešit změnou pravidel, aby byl celkový užitek v zájmu rozhodnutí jednotlivých agentů.
- Příklad
  - Uvažujme, že město chce dát do svých obvodů bezdrátový internet, ale nemá dostatek financí dát ho všem obvodům.
  - Pokud se každého obvodu zeptá, kolik si internetu cení s cílem dát ho tam, kde si ho cenní nejvíce, každý obvod přirozeně cenu nadsadí.
- Řešením může být daň za získané zboží
- **Mechanismus Vickrey-Clarke-Groves**
  1. ústředí se zeptá agentů na hodnotu získání zboží – b_i
  2. ústředí alokuje zboží skupině agentů A
     nechť $b_i(A) = b_i$, je-li $i \in A$, jinak 0, potom ústředí vybere množinu A maximalizující $B = \Sigma_i b_i(A)$
  3. agent i zaplatí daň $W_{-i} - B_{-i}$, kde
     $B_{-i} = \Sigma_{j \neq i} b_j(A)$
     W_{-i} = celkový zisk B, pokud se agent i aukce neúčastní
     jinými slovy, agent, který zboží dostane, platí jako daň největší hodnotu agenta, který zboží nedostal (a agent, který zboží nedostal, neplatí nic)

VCG aukce

analýza

- Proč jsou s výsledkem VCG aukce všichni agenti spokojeni (pokud uvedli pravdivou hodnotu zboží)?
  - agent, který zboží získal, platí méně, než kolik je pro něj hodnota zboží
  - agent, který zboží nezískal, je rád, protože daň převyšuje hodnotu zboží (a nemusí ji platit)
- Jedná se o **pravidla odhalující pravdu**?
  - racionální agent se snaží maximalizovat svůj užitek, tj. rozdíl hodnoty zboží a zaplacené daně
    - $v_i(A) - (W_{-i} - B_{-i},)$
  - ústředí maximalizuje celkový užitek (a agent to ví)
    - $\Sigma_j b_j(A) = b_i(A) + \Sigma_{j \neq i} b_j(A)$
  - agent vlastně maximalizuje
    - $v_i(A) + \Sigma_{j \neq i} b_j(A) - W_{-i}$
  - na hodnotu W_{-i} nemá žádný vliv, takže pokud chce, aby ústředí maximalizovalo to, co chce maximalizovat sám, potom volí
    - $b_i = v_i$