

Informaation arvo

- Öljykenttään myydään porausoikeuksia, palstoja on n kappaletta, mutta vain yhdessä niistä on C euron edestä öljyä
- Yhden palstan hinta on C/n euroa
- Seismologi tarjoaa yritykselle tutkimustietoa palstasta nro 3, joka paljastaa aukottomasti onko palstalla öljyä vai ei
- Paljonko yrityksen kannattaa maksaa tiedosta?
- Todennäköisyydellä $1/n$ tutkimus kertoo palstalla 3 olevan öljyä, jolloin yritys hankkii sen hintaan C/n ja ansaitsee $(n-1)C/n$ euroa
- Todennäköisyydellä $(n-1)/n$ tutkimus osoittaa, ettei palsta 3 sisällä öljyä, jolloin yritys hankkii jonkin muista palstoista

- Koska palstan 3 tilanne jo tunnetaan, niin ostetulta palstalta löytyy nyt öljyä tn.:llä $1/(n-1)$, joten yhtiön odotusarvoinen voitto on $C/(n-1) - C/n = C/n(n-1)$ euroa
- Odotusarvoinen voitto annettuna tutkimustulos on siis $(1/n) \cdot ((n-1)C/n) + ((n-1)/n) \cdot (C/n(n-1)) = C/n$
- Seismologille siis kannattaa maksaa aina palstan hintaan asti
- Lisäinformaatio on arvokasta, koska sen avulla toiminta voidaan sopeuttaa vallitsevaan tilanteeseen
- Ilman informaatiota on tyydyttävä kaikissa mahdollisissa tilanteissa keskimäärin parhaaseen toimintaan

- Olk. E_j on satunnaismuuttuja, jonka arvosta saadaan uusi tarkka havainto
- Agentin aiempi tietämys on E
- Ilman lisäinformaatiota parhaan toiminnon α arvo on

$$EU(\alpha | E) = \max_{\alpha} \sum_i U(\text{Tulos}_i(A)) \cdot P(\text{Tulos}_i(A) | \text{Suorita}(A), E)$$
- Uusi havainto muuttaa parhaan toiminnon ja sen arvon
- Mutta toistaiseksi E_j on satunnaismuuttuja, jonka arvoa ei tunneta, joten voimme vain summata yli sen kaikkien mahdollisten arvojen e_{jk}
- Havainnon E_j arvo on lopulta

$$(\sum_k P(E_j = e_{jk} | E) EU(\alpha_{e_{jk}} | E, E_j = e_{jk})) - EU(\alpha | E)$$

- Lisäinformaatiolla on arvoa sikäli kun se voi johtaa suunnitelman muutokseen ja uusi suunnitelma on oleellisesti parempi kuin vanha
- Merk. $VPI_E(E_j)$ on havainnon E_j arvo, kun nykyhavainnot ovat E
- Minkä tahansa havainnon arvo on ei-negatiivinen

$$\forall j, E: VPI_E(E_j) \geq 0$$
- Arvo riippuu nykyisestä tilasta, joten se voi muuttua tietämyksen myötä
- Äärimmillään informaation arvo putoaa nolnaan, kun tarkastellulle muuttujalle jo tunnetaan arvo
- Siksi informaation arvo ei ole additiivinen

$$VPI_E(E_j, E_k) \neq VPI_E(E_j) + VPI_E(E_k)$$

- Käytännössä tärkeää on, että informaation arvo on järjestysvapaa

$$\begin{aligned} VPI_E(E_j, E_k) &= VPI_E(E_j) + VPI_{E,E_j}(E_k) \\ &= VPI_{E,E_k}(E_j) + VPI_E(E_k) \end{aligned}$$

- Tämän perusteella havainnot voidaan erottaa toiminnoista
- Agentin tulisi hankkia lisäinformaatiota kyselyin
 - järkevässä järjestyksessä,
 - irrelevantteja kysymyksiä välttäen,
 - ottaen huomioon informaation arvon suhteessa sen kustannukseen,
 - vain silloin kun se on järkevää
- *Myopinen agentti* tekee toimintapäätöksen heti jos mikään havaintomuuttujista ei näytä riittävän hyödylliseltä — ei tutkita muuttujakombinaatioita

4.3 KOMPLEKSISET PÄÄTÖKSET

- Agentin hyötyarvo riippuukin nyt sarjasta toimintapäätöksiä
- Oheisessa 4×3 ruudukkomaailmassa agentti tekee siirtymäpäätöksen (*Y, O, V, A*) jokaisella ajanhetkellä
- Kun päädytään toiseen maalitiloista, niin toiminta lakkaa
- Maailma on täysin havainnoitava — agentti tietää sijaintinsa

			+1
			-1
Lähtö			

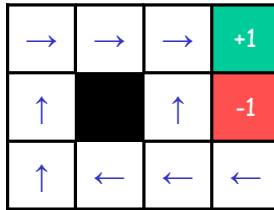
- Jos maailma on deterministinen, niin agentti pääsee lähtötilasta aina lopputilaan **+1** siirtymin $[Y, Y, O, O, O]$
- Koska toiminnot kuitenkin ovat epäluotettavia, niin siirtymäsekvenssi ei aina johda haluttuun tulokseen
- Olkoon niin, että aiottu toiminto toteutuu todennäköisyydellä **0.8** ja todennäköisyydellä **0.1** liike suuntautuukin kohtisuoriin suuntiin
- Jos agentti törmää maailmansa rajoihin, niin toiminnolla ei ole vaikutusta
- Tällöin sekvenssi $[Y, Y, O, O, O]$ johtaakin maaliin vain todennäköisyydellä $0.8^5 = 0.32768$
- Lisäksi agentti voi päätyä maaliin sattumalta kiertämällä esteen toista kautta todennäköisyydellä $0.1^4 \times 0.8$, joten kokonais-todennäköisyys on **0.32776**

- *Siirtymämalli* antaa todennäköisyydet toimintojen tuloksille maailman kaikissa mahdollisissa tiloissa
- Merk. $T(s, a, s')$ on tilan s' saavuttamistodennäköisyys kun toiminto a suoritetaan tilassa s
- Siirtymät ovat *Markovilaisia* sikäli, että vain nykytila s , ei aiempien tilojen historia, vaikuttaa siihen saavutetaanko s'
- Vielä on määrättävä hyötyfunktio
- Päätösongelma on sekventiaalinen, joten hyötyfunktio riippuu tilajonosta — ympäristöhistoriasta — eikä vain yhdestä tilasta
- Toistaiseksi agentti saa kussakin tilassa s *palkkion* (reward) $R(s)$, joka voi olla positiivinen tai negatiivinen

- Esimerkissämme palkkio on -0.04 kaikissa muissa tiloissa paitsi lopputiloissa
- Ympäristöhistorian hyöty puolestaan on palkkioiden summa
- Jos esim. agentti saavuttaa lopputilan $+1$ kymmenen siirtymän jälkeen, niin hyöty on 0.6
- Pienen negatiivisen palautteen tarkoitus on saada agentti maaliin mahdollisimman nopeasti
- Täysin havainnoitavan maailman sekventiaalinen päätösongelma, kun
 - siirtymämalli on Markovilainen ja
 - palkkiot summataan,
 on *Markov-päätösongelma* (Markov decision problem, MDP)

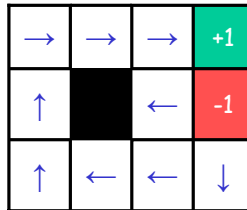
- MDP on kolmikko
 - alkutila S_0 ,
 - siirtymämalli $T(s, a, s')$ ja
 - palkkiofunktio $R(s)$
- Ratkaisuksi MDP:hen ei kelpaa kiinteä siirtymäsekvenssi, koska se voi kuitenkin johtaa muuhun tilaan kuin maaliin
- Vastauksen onkin oltava **politiikka**, joka määrää toiminnan kussakin tilassa, johon agentti voi päätyä
- Poliitiikan π suosittama toiminto tilassa s on $\pi(s)$
- Täydellisen politiikan avulla agentti aina tietää mitä tehdä heittipä epädeterministinen toimintaympäristö sen mihin tilaan tahansa

- Joka kerta kun politiikkaa sovelletaan, maailman stokastisuus johtaa eri ympäristöhistoriaan
- Poliitiikan laadun mittari onkin sen mahdollisesti tuottamien ympäristöhistorioiden hyödyn odotusarvo
- Optimaalisen politiikan π^* odotusarvo on korkein

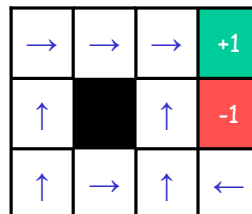


- Poliitiikka antaa eksplisiittisen kontrollin agentin käyttäytymiselle ja samalla yksinkertaisen refleksiivisen agentin kuvauksen

- $-0.0221 < R(s) < 0$:



- $-0.4278 < R(s) < -0.0850$:



• $R(s) < -1.6284$:

→	→	→	+1
↑		→	-1
↑	→	→	↑

• $R(s) > 0$:

+	+	←	+1
+		←	-1
+	+	+	↓

- Äärettömän horisontin tapauksessa agentin toiminta-ajalle ei ole asetettu ylärajaa
- Jos toiminta-aika on rajoitettu, niin eri aikoina samassa tilassa voidaan joutua tekemään eri toimintopäätöksiä — optimaalinen politiikka ei ole *stationäärinen*
- Sen sijaan äärettömän horisontin tapauksessa ei ole syytä muuttaa tilan toimintaa kerrasta toiseen, joten optimaalinen politiikka on stationäärinen
- Tilajonon s_0, s_1, s_2, \dots diskontattu palkkio on

$$R(s_0) + \gamma R(s_1) + \gamma^2 R(s_2) + \dots,$$
missä $0 < \gamma \leq 1$ on diskonttaustekijä

- Kun $\gamma = 1$, niin ympäristöhistorian palkkioksi saadaan erikoistapauksena additiivinen palkkio
- Lähellä nolaa oleva γ puolestaan tarkoittaa tulevien palkkioiden merkityksen pienenemistä
- Jos äärettömän horisontin maailmassa ei ole lainkaan saavutettavaa maalitilaa, niin ympäristöhistorioista tulee äärettömän pitkiä
- Additiivisilla palkkioilla myös hyötyarvot kasvavat yleisesti ottaen äärettömiksi
- Diskontatuilla palkkioilla ($\gamma < 1$) äärettömänkin jonon palkkio on äärellinen

- Olk. R_{\max} palkkioiden yläraja. Tällöin geometrisen sarjan summana

$$\sum_{t=0, \dots, \infty} \gamma^t R(s_t) \leq \sum_{t=0, \dots, \infty} \gamma^t R_{\max} = R_{\max} / (1 - \gamma)$$

- Kelvollinen politiikka (proper policy) takaa agentin pääsevän lopputilaan silloin kun ympäristössä on lopputiloja
- Tällöin äärettömistä tilajonoista ei ole huolta ja voidaan jopa käyttää additiivisia palkkioita
- Optimaalinen politiikka diskontatuilla palkkioilla on

$$\pi^* = \arg \max_{\pi} E[\sum_{t=0, \dots, \infty} \gamma^t R(s_t) \mid \pi],$$
 missä odotusarvo lasketaan yli kaikkien mahdollisten tilajonon, jotka voisivat tulla kyseeseen politiikalla

Arvojen iterointi

- Optimaalisen politiikan selvittämiseksi lasketaan tilojen hyötyarvot ja käytetään niitä optimaalisen toiminnon valitsemiseen
- Tilan hyödyksi lasketaan sitä mahdollisesti seuraavien tilajonojen odotusarvoinen hyöty
- Luonnollisesti jonot riippuvat käytetystä politiikasta π
- Olkoon s_t tila, jossa agentti on kun π_t :tä on noudatettu t askelta
- Huom. s_t on satunnaismuuttuja
- Nyt

$$U^\pi(s) = E[\sum_{t=0, \dots, \infty} \gamma^t R(s_t) \mid \pi, s_0 = s]$$

- Tilan todellinen hyötyarvo $U(s)$ on $U^{\pi^*}(s)$
- Palkkio $R(s)$ siis kuvaa tilassa s olemisen lyhyen tähtäyksen hyödyllisyyttä, kun taas $U(s)$ on s :n pitkän tähtäyksen hyödyllisyys siitä eteenpäin laskien
- Esimerkkimaailmassamme maalitilan lähellä olevilla tiloilla on korkein hyötyarvo, koska niistä matka on lyhin

0.812	0.868	0.912	+1
0.762		0.660	-1
0.705	0.655	0.611	0.388

- Nyt voidaan soveltaa hyödyn odotusarvon maksimointia

$$\pi^*(s) = \arg \max_a \sum_{s'} T(s, a, s') U(s') \quad (*)$$

- Koska tilan s hyöty nyt on diskontattujen palkkioiden summan odotusarvo tästä tilasta eteenpäin, niin se voidaan laskea:
 - Välitön palkkio tilassa s , $R(s)$, +
 - seuraavan tilan diskontatun hyödyn odotusarvo olettaen, että valitaan optimaalinen toiminto

$$U(s) = R(s) + \gamma \max_a \sum_{s'} T(s, a, s') U(s')$$

- Tämä on **Bellman-yhtälö**
- Toimintaympäristössä, jossa on n tilaa, on myös n Bellman-yhtälöä

- Bellman-yhtälöiden yht'aikaiseen ratkaisemiseen ei voi käyttää lineaaristen yhtälöryhmien tehokkaita ratkaisumenetelmiä, koska **max** ei ole lineaarinen operaatio
- Iteratiivisessa ratkaisussa aloitamme tilojen hyötyjen mv. arvoista ja päivitämme niitä kunnes saavutetaan tasapainotila

$$U_{i+1}(s) \leftarrow R(s) + \gamma \max_a \sum_{s'} T(s, a, s') U_i(s'),$$

- missä indeksi i viittaa iteraation i hyötyarvioon
- Päivitysten toistuva soveltaminen päättyy taatusti tasapainotilaan, jolloin saavutetut tilojen hyötyarviot ovat ratkaisu Bellman-yhtälöihin
- Löydetyt ratkaisut ovat yksikäsitteisiä ja vastaava politiikka on optimaalinen

Politiikan iterointi

- Alkaen lähtöpolitiikasta π_0 toista
- **Politiikan arviointi:** laske kaikille tiloille hyötyarvo $U_i = U^{\pi_i}$ politiikkaa π_i sovellettaessa
- **Politiikan parantaminen:** perustuen arvoihin U_i laske uusi hyödyn odotusarvon maksimoiva politiikka π_{i+1} (vrt. (*))
- Kun jälkimmäinen askel ei enää muuta hyötyarvoja, niin algoritmi päättyy
- Tällöin U_i on Bellman-päivityksen kiintopiste ja ratkaisu Bellman-yhtälöihin, joten vastaavan politiikan π_i on oltava optimaalinen
- Äärellisellä tila-avaruudella on vain äärellinen määrä politiikkoja, jokainen iteraatio parantaa politiikkaa, joten politiikan iterointi päättyy lopulta

- Koska kullakin kierroksella politiikka on kiinnitetty, niin politiikan arvioinnissa ei ole tarvetta maksimoida yli toimintojen
- Bellman-yhtälö yksinkertaistuu:

$$U_i(s) = R(s) + \gamma \sum_{s'} T(s, \pi_i(s), s') U_i(s')$$
- Koska epälineaarista maksimoinnista on päästy eroon, on tämä lineaarinen yhtälö
- Lineaarinen yhtälöryhmä, jossa on n yhtälöä ja niissä n tuntematonta muuttujaa voidaan ratkaista ajassa $O(n^3)$ lineaarialgebran menetelmin
- Kuutiollisen ajan sijaan voidaan tyytyä approksimoimaan politiikan laatua ajamalla vain tietty määrä yksinkertaisia arvojen iterointi -askelia kelvollisten hyötyarvioiden saamiseksi