یادگیری تقویتی
فرایند تصمیم مارکوف

محسن هوشمند
دانشکده تکنولوژی اطلاعات و علم رایانه
دانشگاه تحصیلات تکمیلی علوم پایه زنجان

اجزای یادگیری تقویتی

محیط

عامل

حالت

اقدام (عمل و حرکت) سیاست

پاداش

مسیر $s_0, a_0, r_1, s_1, a_1, r_1, \ldots$

با هدف یافتن بهترین دنباله‌ای که مجموعه پاداش‌ها بیشینه

اصول یادگیری تقویتی

i. مشاهدهٔ محیط (آگهی عامل از حالت فعلی)
ii. تصمیم به چگونگی اقدام مبنی بر استراتژی (سیاساتی) و اقدام
iii. تغییر محیط (انتقال به حالت جدید) و دریافت پاداش یا جریمه
iv. پردازش و تحلیل بازخورد (پاداش دریافتی از محیط) [یادگیری از تجارب و اصلاح سیاسات] ارزیابی سیاست
v. تکرار تا یافتن سیاست بهینه

کاوش و بهره‌برداری

کاوش- جستجو در فضای ناشناخته جهت یافتن سیاسات بهتر

بهره‌برداری- استفاده از دانش موجود برای تعیین سیاست

یادگیری همزمان و تعامل با محیط

استفاده محض از بهره‌برداری منجر به سیاست زیربهینه

نیاز به کاوش جهت جستجوی تمامی فضای جستجو

- آزمایش تمام مسیرها و حالت‌ها

فرایند تصمیم مارکوف متناهی

شبیه‌سازی تحلیل محیط مدنظر

شبیه‌سازی تحلیل پاسخ

- پیچیده
- بسیاری از معادلات و مجهولات
- مهارناپذیر

فرایند تصمیم مارکوف متناهی

فرایند تصمیم مارکوف متناهی

- استفاده از بازخورد ارزیابی (اصلاحی)
- صورت‌بندی معمول تصمیم‌سازی متوالی
- کنش‌ها موثر بر
  - پاداش بلافصل
  - موقعیت‌های (حالت‌های) بعدی
- درگیر پاداش‌های آینده
  - نیاز به تعادلی بین پاداش بلافصل و پاداش‌های بعدی
- گذار از تخمین $q_*(a)$ برای هر کنش a به تخمین ارزش $q_*(s,a)$ برای هر کنش a در حالت s
  - یا تخمین ارزش $v_*(s)$ برای هر حالت با داشتن انتخاب کنش بهینه
- نیاز به چنین مقادیر وابسته به حالت جهت تخصیص اعتبار مناسب تصمیمات فردی و مجزا در توالی‌های بلندمدت

فرایند تصمیم مارکوف متناهی

فـت‌مم تدوینی مناسب جهت ی‌ت

- دارای مفاهیمی
  - بازده‌ها
  - تابع‌های ارزش
  - معادلات بلمن

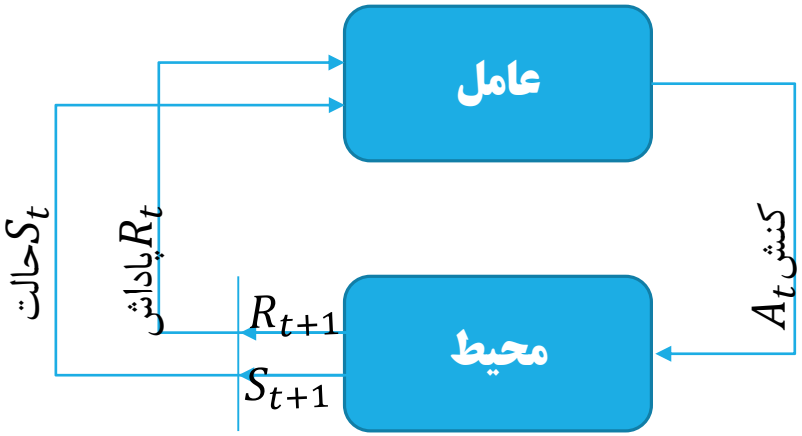
بررسی کاربردهایی بهره گرفته از فـت‌م

- تعارض بین مهارپذیری ریاضی و عمق و پیچیدگی کاربرد

فرایند تصمیم مارکوف متناهی

ف‌ت‌م

- تدوینی متناظر از یادگیری در قالب تعامل در راستای دستیابی به هدفی
- یادگیرنده و تصمیم‌گیر ← عامل
- هر چیزی که در حال تعامل با آن است
  - شامل هر چیزی خارج عامل
  - ← محیط
- تعاملی مداوم
  - عامل کنش‌مند
  - محیط عرضهٔ پاسخی در قبال آن و عرضهٔ موقعیتی جدید به عامل
  - پاداش‌دهی به عامل
  - مقادیر عددی که عامل به دنبال بیش‌سازی آن‌هاست.



فرایند تصمیم مارکوف متناهی

جزیی‌تر

- عامل و محیط اندرکار در هر گام زمانی گسسته $t = 0,1,2,3, \ldots$
- ادراک عامل از حالت محیط در هر گام $S_t \in S$
- انتخاب کنشی بر مبنای اطلاع حاصل از محیط $A_t \in A(s)$
- دریافت پاداشی بابت کنش در گام زمانی بعدی $R_{t+1} \in R \subset \mathbb{R}$
- انتقال به حالتی جدید S_{t+1}

ترکیب ف‌ت‌م و عامل سبب‌ساز تحویل مسیری یا توالی

$$S_0, A_0, R_1, S_1, A_1, R_2, S_2, A_2, R_3, \cdots$$

فرایند تصمیم مارکوف متناهی

ف‌ت‌م متناهی

- متناهی بودن مجموعه حالات، کنش‌ها، پاداش‌ها S و A و R
  - مجموعه متناهی حالت‌ها
  - مجموعه متناهی کنش‌ها
  - مجموعه متناهی پاداش‌ها
- متغیر تصادفی‌های R_t و S_t دارای توزیع احتمال گسسته مشخص و وابسته به حالت و کنش ماسبق
  - احتمال رخداد مقادیر خاص $s' \in S$ و $r \in R$ از دو متغیر تصادفی مذکور در زمان (گام) t

$$p(s', r|s, a) \stackrel{\text{def}}{=} P\{S_t = s', R_t = r|S_{t-1} = s, A_{t-1} = a\}, \forall s', s \in S, a \in A(s)$$

- p معرف دینامیک ف‌ت‌م
  - $p: S \times R \times S \times A \to [0,1]$
  - وام‌گیری علامت | از احتمال شرطی ولی در اینجا صرفا جهت تذکار توزیعی احتمالی بودن pبه ازای انتخاب هر sو a، یا

$$\sum_{s' \in S} \sum_{r \in R} p(s', r|s, a) = 1 \forall s \in S, a \in A(s)$$

فرایند تصمیم مارکوف متناهی

احتمالات ناشی از p توصیف‌گر کامل دینامیک محیط

در رژیم مارکوفی!

- خاصیت مارکوف
  - حالت بعدی و پاداش صرفا به حالت فعلی و کنش انجام یافته در آن زمان
  - $\Longleftarrow$ عدم‌وابستگی به تاریخ کنش‌ها و حالت‌ها
  - صرفا حالت و کنش فعلی
- محدودیتی بر حالت و نه بر فرایند تصمیم
  - هر حالت نگه‌دارنده اطلاعات تمامی جنبه‌های گذشتهٔ تعامل عامل-محیط موثر بر تغییرات در آینده

فرایند تصمیم مارکوف متناهی

- تابع $p: S \times R \times S \times A \to [0,1]$ نمایشگر کل دینامیک محیط
- امکان استخراج هر تابع مستقل از آن
- تابع احتمال گذار حالت $p: S \times S \times A \to [0,1]$

$$p(s'|s,a) \stackrel{\text{def}}{=} P\{S_t = s'|S_{t-1} = s, A_{t-1} = a\} = \sum_{r \in R} p(s', r|s, a)$$

  - از جنس احتمال
  - در محیط قطعی مقادیر صفر و یک
  - $\{0,1\}$
- تابع امید پاداش زوج حالت-کنش $r: S \times A \to \mathbb{R}$

$$r(s,a) \stackrel{\text{def}}{=} E[R_t|S_{t-1} = s, A_{t-1} = a] = \sum_{r \in R} r \sum_{s' \in S} p(s', r|s, a)$$

  - از جنس امید ریاضی

فرایند تصمیم مارکوف متناهی

- تابع امید پاداش حالت-کنش-حالت بعدی $r: S \times A \times S \to \mathbb{R}$

$$r(s,a,s') \stackrel{\text{def}}{=} E[R_t | S_{t-1} = s, A_{t-1} = a, S_t = s'] = \sum_{r \in R} r \frac{p(s', r|s, a)}{p(s'|s, a)}$$

  - قانون بیز
  - از جنس امید ریاضی
  - چرا؟

امکان بهره از چهار تابع برای محاسبات

- ولی در این فصل از تابع اصلی

$$p(a|b) = \frac{p(a,b)}{p(b)}$$

$$p(a|b,c) = \frac{p(a,b|c)}{p(b|c)}$$

$$p(r|s,a,s') = \frac{p(r,s'|s,a)}{p(s'|s,a)}$$

$$\Longrightarrow r(s,a,s') \stackrel{\text{def}}{=} E[R_t|S_{t-1} = s, A_{t-1} = a, S_t = s']$$

$$= \sum_r r p(r|s,a,s') = \sum_r r \frac{p(r,s'|s,a)}{p(s'|s,a)}$$

فرایند تصمیم مارکوف متناهی

- چارچوب فت‌م
  - مجرد و منعطف برای بسیاری مسائل در طرق مختلف
    - عدم لزوم برابر بودن گام‌های زمانی t
    - زمان نیست
    - گام زمانی
    - امکان نمایش وقایع گسسته تصمیم و اجرای کنش
    - امکان متغیر بودن زمان‌های هر گام

    - کنش‌های کنترل سطح پایین مانند اعمال ولتاژ به موتورهای بازوی رباتی
      - یا تصمیم‌های سطح بالا مانند عدم ادامه یا ادامه تحصیل
  - حالت‌ها
    - خواندن مستقیم مقادیر ادارکی حسگرها
    - توصیفی از اشیاء در کلاس حاضر
    - مثال عدم اطمینان از کجا قرار دادن کلید!
- مرز بین کنش و محیط
  - ابن مقفع!

فرایند تصمیم مارکوف متناهی

- پاداش‌ها
  - محاسبه در خود یادگیرندهٔ واقعی یا مصنوعی ولی درنظرگرفتن آن‌ها به مثابه خارجی برای عامل
- محیط هرچیز خارج از دست عامل
  - به معنای بی‌اطلاعی نیست
  - مرز ممیزهٔ محیط-عامل: نمایشگر محدودیت در کنترل مطلق عامل و بی‌ارتباط با دانش عامل
  - امکان تغییر ان در اهداف متفاوت

فرایند تصمیم مارکوف متناهی

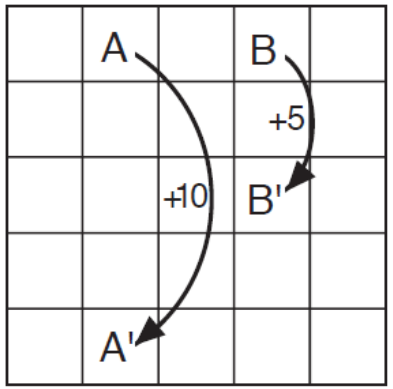
- تدوینی از مسائل یادگیری هدف‌محور از تعامل
- امکان کاهش به سه سیگنال بین عامل و محیطش
  - سیگنالی جهت نمایش انتخاب‌های عامل (کنش‌ها)
  - سیگنالی جهت نمایش اساسی که انتخاب از ادارک آن ناشی می‌شود (حالت)
  - سیگنالی جهت تعریف هدف عامل (پاداش‌ها)
    - عدم پوشش مناسب تمامی مسائل تصمیم‌گیری
      - اما دارای استفاده سخت مفید و کاربردی
- انتخاب نمایش‌های مناسب
  - کاری پیچیده و سخت

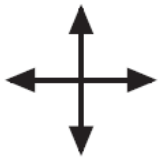
فرایند تصمیم مارکوف متناهی

امکان نمایش اکثر مسائل ی‌ت

راهزن صرفا دارای تک حالت
- عدم تغییر محیط

در ف‌ت‌م‌م
- وجود چندین حالت
- انجام کنش موجب تغییر محبط از حالتی به حالت دیگر
- مثال- بازی دنیای شطرنجی
  - هر مدخل (خانه) یک حالت
  - چهار کنش
  - هر کنش منجر به تغییر حالت
  - پاداش دریافتی وابسته به حالت و کنش
  - خانه‌های مختلف پاداش مختلف
    - کناره‌ها
    - خانه‌های الف و ب
    - سایر خانه‌ها
- امکان نمایش با ف‌ت‌م‌م



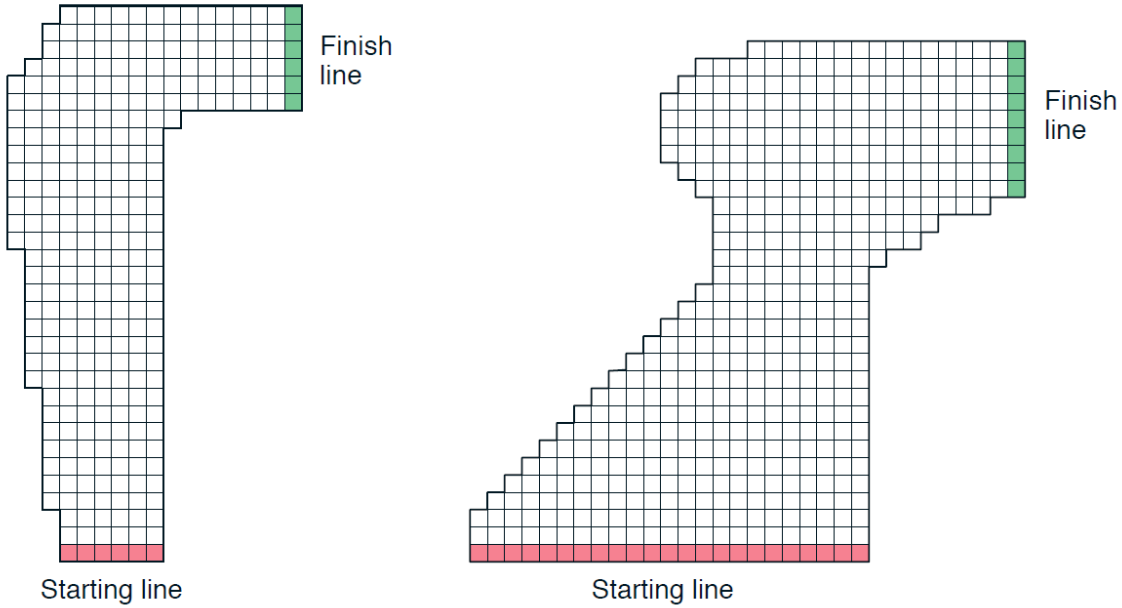


Actions

فرایند تصمیم مارکوف متناهی

رهگیری مسابقه

- عدم امکان رسیدن به خط پایان از هر مسیری
- گسسته‌سازی



فرایند تصمیم مارکوف متناهی-مثال

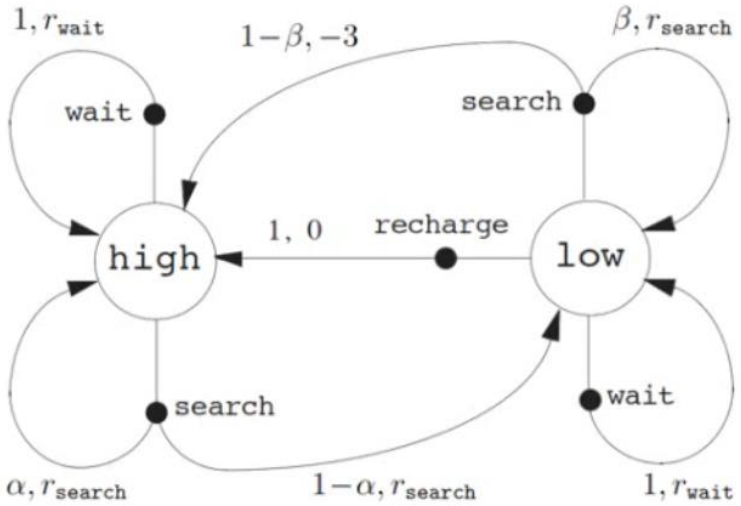
- ربات جمع‌آوری زباله در اداره
  - دارای
    - حسگر جهت ادراک زباله
    - بازو و گیرهٔ جمع‌آوری
    - ریشارژی
  - سیستم کنترلی
    - ادراک نوع زباله
    - راهبری
    - کنترل بازو و گیره
  - تصمیمات سطح بالا
    - یادگیری تقویتی
    - بر مبنای وضعیت فعلی باتری

فرایند تصمیم مارکوف متناهی-مثال

- بر اساس وضعیت باتری ربات
  - دارای دو وضعیت پر و کم
    - مجموعه حالت‌ها {ک,پ} $= S$ ←
- کنش‌ها
  - جستجو در بازه‌ای معین از زمان
  - توقف(انتظار) و دریافت زباله از افراد
  - مراجعه به محل اصلی و پرکردن باتری (شارژ)
- حالت پر بی‌نیاز از کنش پرکردن
  - پس حالت پر شامل دو کنش جستجو انتظار {ت,ج} $= A\left(\text{پ}\right)$
- حالت کم شامل سه کنش،جستجو و انتظار و پرکردن{ش,ت,ج} $= A\left(\text{ک}\right)$
- پاداش
  - اکثرا صفر
  - مثبت در هنگام جمع‌اوری زباله‌ای
  - منفی بزرگ با تمام شدن باتری
  - جستجوی مداوم بهترین روش جمع‌اوری زباله
    - جستجو در حین پر بودن باتری عامل باقی ماندن انرژی با احتمال α و کاهش انرژی با احتمال $1-\alpha$
    - جستجو در حین کم بودن باتری عامل باقی ماندن انرژی با احتمال β و کاهش انرژی با احتمال $1-\beta$

فرایند تصمیم مارکوف م

s	a	s'	$p(s'\|s,a)$	$r(s,a,s')$
high	search	high	α	r_{search}
high	search	low	$1-\alpha$	r_{search}
low	search	high	$1-\beta$	-3
low	search	low	β	r_{search}
high	wait	high	1	r_{wait}
high	wait	low	0	-
low	wait	high	0	-
low	wait	low	1	r_{wait}
low	recharge	high	1	0
low	recharge	low	0	-



- پاداش ۱ با یافتن زباله‌ای
- پاداش ۳−با خاموشی ربات
- پاداش جستجو بیشتر پاداش انتظار
- عدم جمع‌اوری زباله در حین برگشت به جایگاه جهت پر کردن باتری
- عدم امکان جمع‌اوری زباله در جین تمام شدن باتری

- سیستم به مثابهٔ ف‌ت‌م متناهی و امکان تقریر احتمالات گذار و امید پاداش‌ها با دینامیک
  - نمایش جدولی
    - هر ردیف جدول متناظر با ترکیبی یکه از حالت و کنش و حالت بعدی
      - بعضی گذارها دارای احتمال صفر← بدون امید پاداش
  - نمایش گراف گذار
    - دو نوع رأس برای نمایش حالت‌ها و کنش‌ها
      - چرا؟
    - جمع احتمال‌های خروجی از هر راس؟

ف‌ت‌م‌م-اهداف و پاداش‌ها

تعبیه اهداف در قالب سیگنالی با نام پاداش

- از محیط به عامل
- در هر زمان مقداری عددی $R_t \in \mathbb{R}$
- تنها راه ارتباطی از طریق پاداش‌ها
- نیاز به تعریف و تنظیم دقیق پاداش‌ها

هدف عامل

- انتخاب کنش‌های بیشینه‌ساز تمامی پاداش‌های دریافتی
  - پس نه فقط پاداش بلافصل بلکه تجمیع پاداش‌ها در بلندمدت
  - فرض پاداش
  - منظور ما از اهداف و در قالب بیشینه‌سازی امید ارزش جمع سیگنالی‌های اسکالر (پاداش) به منصهٔ ظهور می‌رسد.
- پاداش وجه ممیزهٔ ی‌ت
- امکان اخلال و تحدید انگاشتنِ پاداش
  - در عمل اینگونه نیست و منعطف و پر استفاده

ف‌ت‌م‌م-اهداف و پاداش‌ها

مثال

- آموزش قدم‌زدن ربات
  - امتیاز به حرکت رو به جلو در هر گام
- آموزش خروج از ماز
  - پاداش منفی در هر گام قبل خروج
- ربات زباله‌ساز(بر)!
  - صفر در بیشتر مواقع،یک در هنگام برداشتن زباله‌ای
  - پیچیده‌تر کردن. پاداش منفی در برخورد با اجسام یا فریادشخص بر آن
- شطرنج. برد، پاد، باخت
  - در تمامی موارد یادگیری بیشینه کردن پاداش
- لزوم دقت در انتخاب پادش
- اشتباه بد منجر به نتیجه زیربهینه
- پاداش عامل آموزش دانش نباشد.
- پاداش جهت بیان اهداف نه راه رسیدن به اهداف
  - مثال شطرنج- مهره گیری از حریف پاداش
    - عدم یادگیری برد بازی

ف‌ت‌م‌م- بازده‌ها و اپیزودها

- هدف عامل: انتخاب کنش‌های بیشینه‌ساز تجمیع پاداش‌ها در بلندمدت
- نمایش صوری
  - $S_0,A_0,R_1,S_1,A_1,R_2,S_2,A_2,R_3,\ldots$
  - پاداش‌های دریافتی بعد از زمان t
  - $R_{t+1}, R_{t+2},R_{t+3},\cdots$
- به دنبال بیشینه‌کردن امید بازده G_t بازده یا جمع پاداش‌های پس از گام t
  - ساده‌ترین جمع پاداش‌ها

$$G_t = R_{t+1} + R_{t+2} + R_{t+3} + \cdots + R_T$$

  - T گام نهایی
  - $G_0 =?$
- انواع محیط‌ها
  - محیط‌های اپیزودی
    - T متناهی
    - دارای حالت پایان terminal state
    - بازی شطرنج و فرار از هزارتو
    - آغاز هر اپیزود مستقل از قبلی
    - جداسازی حالت پایانی در حالت‌های غیر پایانی یا S در مقابل S^+
    - امکان متفاوت بودن T از اپیزودی به اپیزودی

ف‌ت‌م‌م- بازده‌ها و اپیزودها

- بی‌نهایت بودن بازده در محیط‌های پیوسته
  - $G_t = \sum_{k=t+1}^{\infty} R_k \to \infty$
- افزودن وزن کاهنده discounting
  - اهمیت بیشتر به پاداش‌های اخیر
- بازده محیط پیوسته: جمع پاداش‌ها
  - G_t بازده یا جمع پاداش‌های پس از گام t

$$G_t = R_{t+1} + \gamma R_{t+2} + \gamma^2 R_{t+3} + \cdots = \sum_{k=0}^{\infty} \gamma^k R_{t+k+1} = \sum_{k=t+1}^{\infty} \gamma^{k-t-1} R_k$$

  - $0 \leq \gamma \leq ۱$
  - نمایشگر ارزش فعلی پاداش‌های آینده

ف‌ت‌م‌م- بازده‌ها و اپیزودها

- بی‌نهایت بودن بازده در محیط‌های پیوسته
  - $G_t = \sum_{k=t+1}^{\infty} R_k \rightarrow \infty$
- افزودن وزن کاهنده discounting
  - اهمیت بیشتر به پاداش‌های اخیر
- بازده محیط پیوسته: جمع پاداش‌ها
  - G_t بازده یا جمع پاداش‌های پس از گام t

$$G_t = R_{t+1} + \gamma R_{t+2} + \gamma^2 R_{t+3} + \cdots = \sum_{k=0}^{\infty} \gamma^k R_{t+k+1} = \sum_{k=t+1}^{\infty} \gamma^{k-t-1} R_k$$

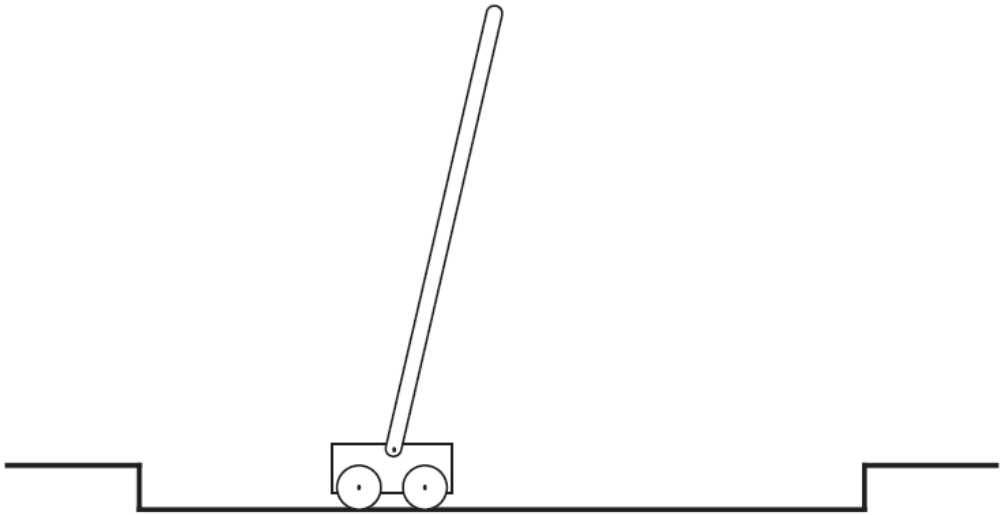
  - $0 \leq \gamma \leq ۱$
  - نمایشگر ارزش فعلی پاداش‌های آینده
    - پاداش دریافتی در kگام زمانی آینده برابر ضریبی γ^{k-1} از آن در زمان فعلی
  - عامل به دنبال دستیابی سریعتر به اهداف و تعویق نینداختن آن به زمان‌های طولانی‌تر
  - عدم وجود ضریب
    - موجب کند شدن عامل
      - چرا؟

ف‌ت‌م‌م- بازده‌ها و اپیزودها

- حالت‌های حدی
  - $\gamma < 1$
    - جمع نامتناهی دارای مقدار متناهی در صورت محدود بودن دنباله پاداش‌ها
  - $\gamma = 0$
    - نزدیک‌بین یا کوته‌نظر؟! یا شاید
    - عدم توجه به آینده و صرفا بیشینه‌سازی پاداش بلافصل
  - $\gamma = 1$
    - تمامی پاداش‌های آینده دارای اهیمت برابر
    - دوراندیش؟!
- محاسبهٔ بازده

$$G_t = R_{t+1} + \gamma R_{t+2} + \gamma^2 R_{t+3} + \gamma^{۳} R_{t+۴} + \cdots$$
$$= R_{t+1} + \gamma\left(R_{t+2} + \gamma^1 R_{t+3} + \gamma^2 R_{t+۴} + \cdots\right) = R_{t+1} + \gamma G_{t+1}$$

  - $G_T = 0$
    - ؟
- سری هندسی



مثال

تعادل تیر (تعادل میله)

- شکست در صورت گذر کردن از زاویه‌ای
- عمود کردن میله پس از شکست
  - اپیزودی؟
- پاداش ۱ در صورت نیفتادن در هر گام
  - پاداش جمع گام‌های زمانی قبل افتادن
- پیوسته
  - استفاده از وزن کاهنده
  - پاداش۱− در هر شکسن و صفر در بقیه مواقع

مثال

تمرین ۸-۳ کتاب

$$\gamma = 0.5, R_1 = -1, R_2 = 2, R_3 = 6, R_4 = 3, R_5 = 2, T = 5$$

$G_0, G_1, G_2, \dots, G_5$؟

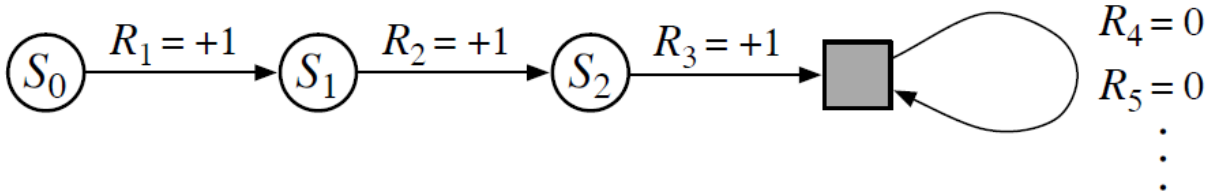
بازگشتی

$$G_5 = 0$$
$$G_4 = R_5 + \gamma G_5 = 2 + 0.5 \times 0 = 2$$
$$G_3 = R_4 + \gamma G_4 = 3 + 0.5 \times 2 = 4$$
$$G_2 = R_3 + \gamma G_3 = 6 + 0.5 \times 4 = 8$$
$$G_1 = R_2 + \gamma G_2 = 2 + 0.5 \times 8 = 6$$
$$G_0 = R_1 + \gamma G_1 = -1 + 0.5 \times 6 = 2$$

مستقیم

$$G_0 = R_1 + \gamma^1 R_2 + \gamma^2 R_3 + \gamma^3 R_4 + \gamma^4 R_5 = -1 + 0.5^1 \times 2 + 0.5^2 \times 6 + 0.5^3 \times 4 + 0.5^4 \times 2 = 2$$

نمایش یکسان اپیزودی و پیوسته



$$G_t = \sum_{k=t+1}^{T} \gamma^{k-t-1} R_k$$

$T = \infty$

$\gamma = 1$

سیاسات و تابع‌های ارزش

تخمین «تابع‌های ارزش» در اکثر الگوریتم‌های ی‌ت
- تابع‌های حالت‌ها یا تابع‌های زوج حالت-کنش‌ها
- تخمین میزان خوبی بودن در حالتی برای عامل
  - خوب بودن؟
    - بر مبنای میزان امید پاداش‌های آینده
    - ← امید بازده
    - وابسته به انتخاب کنش فعلی
- ⟸ تعریف تابع‌های ارزش با عنایت به راه‌های خاص کنش
  - سیاسات

سیاست
- نگاشت از حالت‌ها به احتمال‌های انتخاب هر کنش
- فرض: عامل پیرو سیاست π در زمان t
  - آن‌گاه $\pi(a|s)$ احتمال $A_t = a$ اگر $S_t = s$

- ی‌ت روشنگر چگونگی تغییر سیاست عامل در نتیجهٔ تجربیاتش

سیاسات و تابع‌های ارزش

تابع ارزش حالت s تحت انقیاد سیاست π

- $v_\pi(s)$
- امید بازده با شروع از s و پیروی از π پس از آن

$$v_\pi(s) = E_\pi[G_t|S_t = s] = E_\pi\left[\sum_{k=0}^{\infty} \gamma^k R_{t+k+1} \,|S_t = s\right], \forall s \in S$$

- E_π امیدریاضی متغیر تصادفی با داشتن عاملی پیرو سیاست π
- ارزش حالت نهایی برابر صفر

v_π تابع ارزش–حالت سیاست π

سیاسات و تابع‌های ارزش

به طریق مشابه

ارزش اجرای کنش a در حالت s تحت انقیاد سیاست π

تابع ارزش حالت-کنش $q_\pi(s,a)$

- میزان ارزشمندی کنشی در حالتی خاص
- امید بازده با شروع از s و اجرای کنش a و پیروی از π در پی آن

$$q_\pi(s,a) = E[G_t|S_t = s, A_t = a] = E_\pi\left[\sum_{k=0}^{\infty} \gamma^k R_{t+k+1} |S_t = s, A_t = a\right], \forall s \in S, a \in \mathrm{A}(s)$$

q_π تابع ارزش-کنش سیاست π

امکان نوشتن q_π و v_π براساس یکدیگر

- تمرین

سیاسات و تابع‌های ارزش

امکان دستیابی به تخمین تابع‌های ارزش q_π و v_π از طریق کسب تجربه

- مثال
  - عامل پیرو سیاست و
    - محاسبهٔ میانگین بازده برای هر حالت پس از حالت فعلی
    - آن‌گاه همگرایی میانگین به ارزش حالت فعلی یا $v_\pi(s)$
    - میل به بی‌نهایت همگرایی به مقدار واقعی
  - در صورت نگهداری مقادیر میانگین مجزا برای هر کنش هر عامل
  - به طریق مشابه همگرایی میانگین‌ها به ارزش‌های کنش $q_\pi(s,a)$
  - $\Longleftarrow$ مونت‌کارلو
  - میانگین‌گیری از نمونه‌های تصادفی فراوان از بازده‌های واقعی

سیاسات و تابع‌های ارزش

ویژگی بنیادی تابع‌های ارزش در ی‌ت

- برآورده کردن روابط بازگشتی
- برقراری شرط سازگاری برای هر سیاست π و حالت s

$$v_\pi(s) = E_\pi[G_t|S_t = s]$$
$$= E_\pi[R_{t+1} + \gamma G_{t+1}|S_t = s]$$

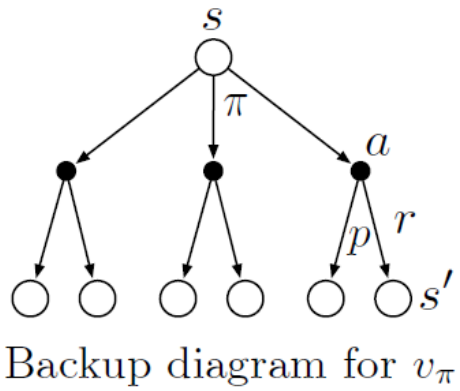
$$= \sum_a \pi(a|s) \sum_{s'} \sum_r p(s', r|s, a)\left[r + \gamma E_\pi[G_{t+1}|S_{t+1} = s']\right]$$

$$= \sum_a \pi(a|s) \sum_{s'} \sum_r p(s', r|s, a)[r + \gamma v_\pi(s')]\,, \forall s \in S$$

- $\pi(a|s)$ احتمال انتخاب کنش a تحت سیاست π
- برای تمامی مقادیر aو s’و r
  - محاسبهٔ $\pi(a|s)p(s',r|s,a)$ و ضرب ان در کمیت بین قلاب‌ها و سپس جمع تمامی این مقادیر
  - نتیجه امید ارزش

سیاسات و تابع‌های ارزش

$$v_\pi(s) = E_\pi[G_t|S_t = s] = \sum_a \pi(a|s) \sum_{s'} \sum_r p(s', r|s, a)[r + \gamma v_\pi(s')], \forall s \in S$$



Backup diagram for v_π

- مشهور به معادله بلمنِ v_π
- بیان رابطه‌ای بین ارزش حالتی و ارزش حالت‌های جانشینش

توجه به حالت‌های جاشنین حالتی با نمودار

- دایرهٔ توخالی حالت
- دایرهٔ توپر زوج حالت-کنش

- معادله بلمن
  - لزوم برابری ارزش حالت شروع با جمع ارزش وزن‌دار حالت جانشین ممکن و امید پاداش در طول مسیر
  - پایهٔ رویه‌های رایانیدن، تخمین، و یادگیری v_π

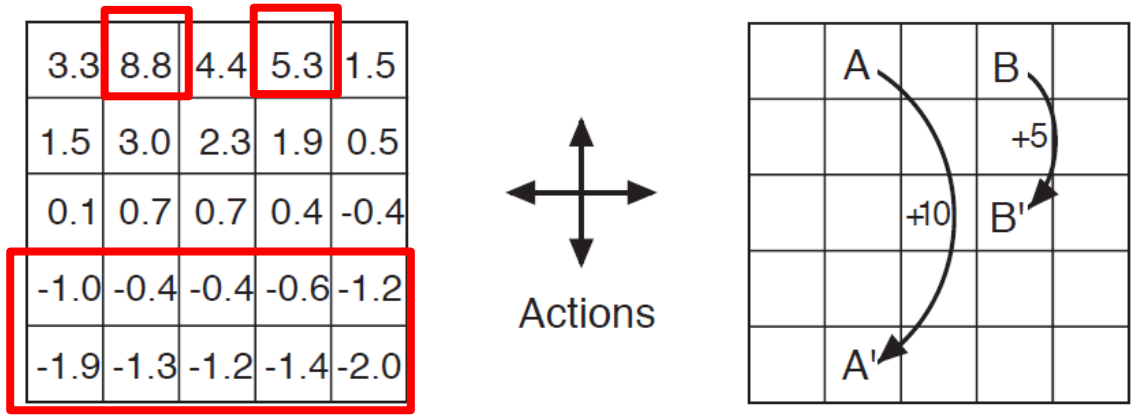
- معادلات بلمن n دستگاه n مجهوله

مثال

دنیای شطرنجی
- نمونه‌ای ساده از فت‌مم

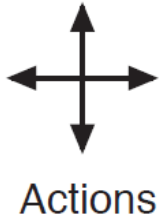
هر مدخل یک حالت

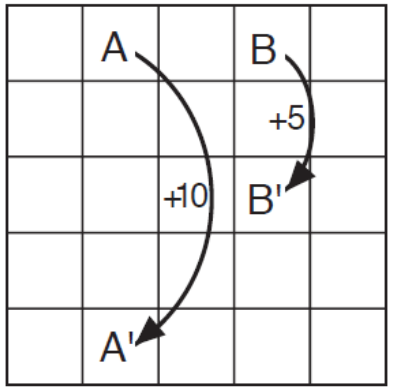
چهار کنش در هر حالت
- کنش‌های هادی به بیرون صفحه بی‌تاثیر حرکتی و پاداش منفی
- بقیه کنش‌های پاداش صفر
- به جز موارد الف و ب
  - در حالت A هر چهار کنش منجر به انتقال به A’و پاداش ۱۰
  - در حالت B هر چهار کنش منجر به انتقال به B’و پاداش ۵
  - احتمال انتخاب هر کنش برابر$\frac{۱}{۴}$در تمامی حالات $\pi(a|s) = \frac{۱}{۴}$
  - تابع ارزش v_π با $\gamma = 0.9$
    - حاصل دستگاه معادلات خطی بلمن
    - ۲۵ معادله



مثال

3.3	8.8	4.4	5.3	1.5
1.5	3.0	2.3	1.9	0.5
0.1	0.7	0.7	0.4	-0.4
-1.0	-0.4	-0.4	-0.6	-1.2
-1.9	-1.3	-1.2	-1.4	-2.0





دنیای شطرنجی

خانه $v_\pi(۳,۳) = ۰.۷$

ارزش همسایه‌ها

$۲.۳ ,۰.۷,۰.۴, - ۰.۴$

$p(s',r|s,a) =? = ۱!$

$$v_\pi(۳,۳) = \sum_a \pi(a|s) \sum_{s'} \sum_r p(s',r|s,a)[r + \gamma v_\pi(s')]$$

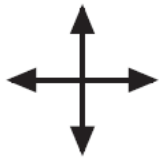
$$\frac{۱}{۴}[0 + ۰.۹ \times ۲.۳] + \frac{۱}{۴}[0 + ۰.۹ \times ۰.۷] + \frac{۱}{۴}[0 + ۰.۹ \times ۰.۴] + \frac{۱}{۴}[0 + ۰.۹ \times -۰.۴]$$

$$= \frac{۱}{۴} \times ۰.۹ \times [۲.۳ + ۰.۷ + ۰.۴ - ۰.۴] = \frac{۱}{۴} \times ۰.۹ \times ۳ \approx ۰.۷$$

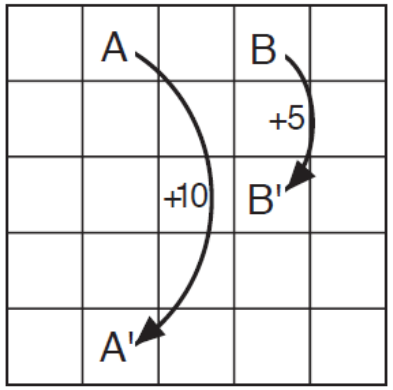
مثال

دنیای شطرنجی

3.3	8.8	4.4	5.3	1.5
1.5	3.0	2.3	1.9	0.5
0.1	0.7	0.7	0.4	-0.4
-1.0	-0.4	-0.4	-0.6	-1.2
-1.9	-1.3	-1.2	-1.4	-2.0



Actions

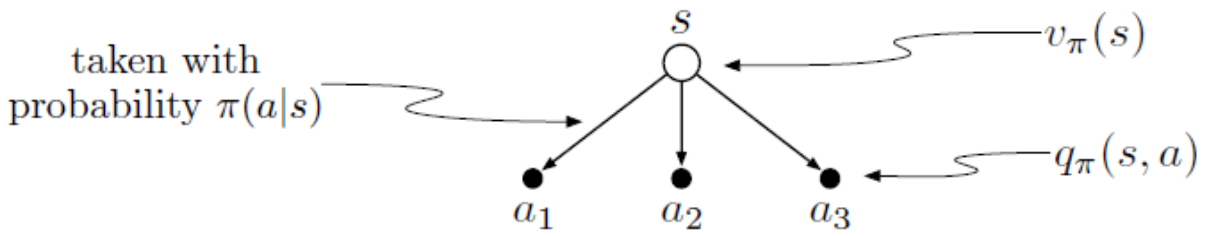


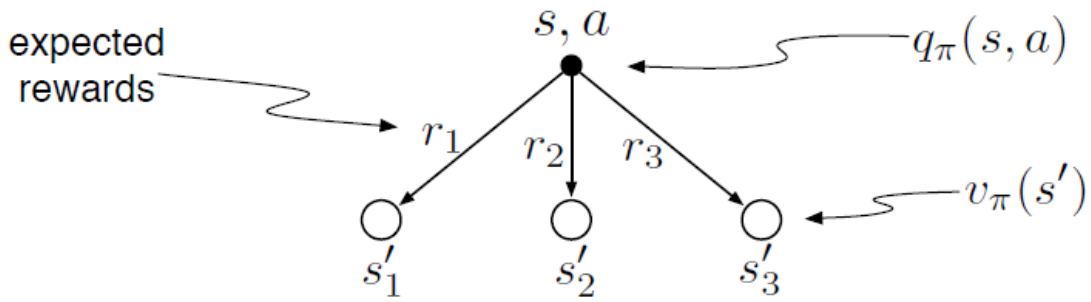
$$v_\pi(۱,۵) = \sum_a \pi(a|s) \sum_{s'} \sum_r p(s', r|s, a)[r + \gamma v_\pi(s')]$$

$$\frac{۱}{۴}[-۱ + ۰.۹ \times v_\pi(۱,۵)] + \frac{۱}{۴}[-۱ + ۰.۹ \times v_\pi(۱,۵)] + \frac{۱}{۴}[0 + ۰.۹ \times ۵.۳]$$
$$+ \frac{۱}{۴}[0 + ۰.۹ \times ۰.۵]$$

$$v_\pi(۱,۵) \approx ۰.۷$$

رابطه بین ارزش-حالت و ارزش-کنش با نمودار





مثال

دنیای راهرو

- تخته سیاه سفید

سیاسات بهینه و تابع‌های ارزش بهینه

هدف حل وظایف ی‌ت

- یافتن سیاستی مولد پاداشی فراوان در بلندمدت
- امکان تعریف «ترتیب جزیی» بین سیاسات با استفاده از تابع‌های ارزش

مقایسهٔ سیاسات

- سیاست بهتر π نسبت به π'
- به ازای تمامی حالات منجر به ارزش حالت بیشتر

$$\forall s:\ v_{\pi}(s) \geq v_{\pi'}(s), \forall s \in S \Leftrightarrow \pi \geq \pi'$$

سیاسات بهینه و تابع‌های ارزش بهینه

وجود حداقل یک سیاست بهینه π_*

- عدم لزوم یکتائی
- دارای تابع ارزش حالت بهینه v_*

$$v_*(s) = \max_{\pi} v_\pi(s)\,, \forall s \in S$$

امکان محاسبهٔ تحلیلی سیاست بهینه با داشتن دینامیک p در مسائل ف‌ت‌م

سیاست و تابع ارزش بهینه

سیاست بهینه دارای تابع ارزش حالت-کنش بهینه q_*

$$q_*(s,a) = \max_{\pi} q_\pi(s,a)\,, \forall s \in S, a \in A(s)$$

تابع بالا برای زوج حالت-کنش

- امید بازده انجام کنش a در حالت s
- $\Longleftarrow$ امکان تقریر توابع ارزش بهینه بر اساس یکدیگر

$$q_*(s,a) = E[R_{t+1} + \gamma v_*(S_{t+1}) | S_t = s, A_t = a]$$

- چرا؟

سیاست و تابع ارزش بهینه

v_* تابع ارزش سیاستی

سیاست و تابع ارزش بهینه

v_* تابع ارزش سیاستی

- در نتیجه ضرورت برآورده‌سازی شرط خودسازگاری معادلهٔ بلمن

$$v_\pi(s) = E_\pi[G_t|S_t = s] = \sum_a \pi(a|s) \sum_{s'} \sum_r p(s', r|s, a)[r + \gamma v_\pi(s')] \, , \forall s \in S$$

سیاست و تابع ارزش بهینه

v_* تابع ارزش سیاستی

- در نتیجه ضرورت برآورده‌سازی شرط خودسازگاری معادلهٔ بلمن

$$v_\pi(s) = E_\pi[G_t|S_t = s] = \sum_a \pi(a|s) \sum_{s'} \sum_r p(s', r|s, a)[r + \gamma v_\pi(s')] \, , \forall s \in S$$

سیاست و تابع ارزش بهینه

v_* تابع ارزش سیاستی

- در نتیجه ضرورت برآورده‌سازی شرط خودسازگاری معادلهٔ بلمن
- بدلیل بهینگی مقدار آن
  - امکان نوشتن شرط سازگاری آن به صورتی خاص بدون ارجاع به سیاستی خاص

$$v_\pi(s) = E_\pi[G_t|S_t = s] = \sum_a \pi(a|s) \sum_{s'} \sum_r p(s', r|s, a)[r + \gamma v_\pi(s')] \, , \forall s \in S$$

سیاست و تابع ارزش بهینه

v_* تابع ارزش سیاستی

- در نتیجه ضرورت برآورده‌سازی شرط خودسازگاری معادلهٔ بلمن
- بدلیل بهینگی مقدار آن
  - امکان نوشتن شرط سازگاری آن به صورتی خاص بدون ارجاع به سیاستی خاص
  - معادله بلمن v_* یا معادله بهینگی بلمن
  - ضرورت برابری ارزش حالتی تحت سیاست بهینه با امید بازده بهترین عمل آن حالت

سیاست و تابع ارزش بهینه

معادلات بهینگی بلمن v_*

ضرورت برابری ارزش حالتی تحت سیاست بهینه با امید بازده بهترین عمل آن حالت

$$v_*(s)$$

سیاست و تابع ارزش بهینه

معادلات بهینگی بلمن v_*

ضرورت برابری ارزش حالتی تحت سیاست بهینه با امید بازده بهترین کنش آن حالت

$$v_*(s) = \max_{a \in A(s)} q_{\pi_*}(s, a)$$

سیاست و تابع ارزش بهینه

معادلات بهینگی بلمن v_*

$$v_*(s) = \max_{a \in A(s)} q_{\pi_*}(s,a)$$
$$= \max_a E_{\pi_*}[G_t | S_t = s, A_t = a]$$

سیاست و تابع ارزش بهینه

معادلات بهینگی بلمن v_*

$$v_*(s) = \max_{a \in A(s)} q_{\pi_*}(s,a)$$
$$= \max_a E_{\pi_*}[G_t | S_t = s, A_t = a]$$
$$= \max_a E_{\pi_*}[R_{t+1} + \gamma G_{t+1} | S_t = s, A_t = a]$$

سیاست و تابع ارزش بهینه

معادلات بهینگی بلمن v_*

$$v_*(s) = \max_{a \in A(s)} q_{\pi_*}(s,a)$$
$$= \max_a E_{\pi_*}[G_t | S_t = s, A_t = a]$$
$$= \max_a E_{\pi_*}[R_{t+1} + \gamma G_{t+1} | S_t = s, A_t = a]$$
$$= \max_a E[R_{t+1} + \gamma v_*(S_{t+1}) | S_t = s, A_t = a]$$

سیاست و تابع ارزش بهینه

معادلات بهینگی بلمن v_*

$$
\begin{aligned}
v_*(s) &= \max_{a \in A(s)} q_{\pi_*}(s,a) \\
&= \max_a E_{\pi_*}[G_t | S_t = s, A_t = a] \\
&= \max_a E_{\pi_*}[R_{t+1} + \gamma G_{t+1} | S_t = s, A_t = a] \\
&= \max_a E[R_{t+1} + \gamma v_*(S_{t+1}) | S_t = s, A_t = a] \\
&= \max_a \sum_r \sum_{s'} p(s', r | s, a)[r + \gamma v_*(s')]
\end{aligned}
$$

سیاست و تابع ارزش بهینه

معادلات بهینگی بلمن v_*

$$v_*(s) = \max_{a \in A(s)} q_{\pi_*}(s,a)$$
$$= \max_a E_{\pi_*}[G_t | S_t = s, A_t = a]$$
$$= \max_a E_{\pi_*}[R_{t+1} + \gamma G_{t+1} | S_t = s, A_t = a]$$
$$= \max_a E[R_{t+1} + \gamma v_*(S_{t+1}) | S_t = s, A_t = a]$$
$$= \max_a \sum_r \sum_{s'} p(s', r|s, a)[r + \gamma v_*(s')]$$
$$= \max_a \sum_r \sum_{s'} \underbrace{p(s', r|s, a)}_{\text{دینامیک}} [r + \gamma v_*(s')]$$

سیاست و تابع ارزش بهینه

معادلات بهینگی بلمن v_*

$$
\begin{aligned}
v_*(s) &= \max_{a \in A(s)} q_{\pi_*}(s,a) \\
&= \max_a E_{\pi_*}[G_t | S_t = s, A_t = a] \\
&= \max_a E_{\pi_*}[R_{t+1} + \gamma G_{t+1} | S_t = s, A_t = a] \\
&= \max_a E[R_{t+1} + \gamma v_*(S_{t+1}) | S_t = s, A_t = a] \\
&= \max_a \sum_r \sum_{s'} p(s', r | s, a)[r + \gamma v_*(s')] \\
&= \max_a \sum_r \sum_{s'} \underbrace{p(s', r | s, a)}_{\text{دینامیک}} [r + \gamma v_*(s')]
\end{aligned}
$$

معادلات غیرخطی

- چرا؟

سیاست و تابع ارزش بهینه

معادلات بهینگی بلمن v_*

$$
\begin{aligned}
v_*(s) &= \max_{a \in A(s)} q_{\pi_*}(s,a) \\
&= \max_a E_{\pi_*}[G_t | S_t = s, A_t = a] \\
&= \max_a E_{\pi_*}[R_{t+1} + \gamma G_{t+1} | S_t = s, A_t = a] \\
&= \max_a E[R_{t+1} + \gamma v_*(S_{t+1}) | S_t = s, A_t = a] \\
&= \max_a \sum_r \sum_{s'} p(s', r | s, a)[r + \gamma v_*(s')] \\
&= \max_a \sum_r \sum_{s'} \underbrace{p(s', r | s, a)}_{\text{دینامیک}} [r + \gamma v_*(s')]
\end{aligned}
$$

معادلات غیرخطی

- چرا؟
- قبلا داشتن $\pi(a|s)$
- تابع بیش‌ساز

سیاست و تابع ارزش بهینه

معادلات بهینگی بلمن v_*

$$v_*(s) = \max_a E[R_{t+1} + \gamma v_*(S_{t+1}) | S_t = s, A_t = a]$$

$$v_*(s) = \max_a \sum_r \sum_{s'} p(s', r|s, a)[r + \gamma v_*(s')]$$

دو معادلهٔ بالا دو صورت معادله بهینگی بلمن برای v_*

معادلات غیرخطی

- چرا؟
- قبلا داشتن $\pi(a|s)$
- تابع بیش‌ساز

سیاست و تابع ارزش بهینه

معادلات بهینگی بلمن q_*

$$q_*(s,a) = E\left[R_{t+1} + \gamma \max_{a'} q_*(S_{t+1}, a') \,|S_t = s, A_t = a\right]$$

$$= \sum_{r}\sum_{s'} p(s', r|s, a)\left[r + \gamma \max_{a'} q_*(s', a')\right]$$

معادلات غیرخطی

- چرا؟

سیاست و تابع ارزش بهینه (خلاصه)

معادلات بهینگی بلمن v_*

$$v_*(s) = \max_a E[R_{t+1} + \gamma v_*(S_{t+1}) | S_t = s, A_t = a]$$
$$= \max_a \sum_r \sum_{s'} p(s', r|s, a)[r + \gamma v_*(s')]$$

معادلات بهینگی بلمن q_*

$$q_*(s, a) = E\left[R_{t+1} + \gamma \max_{a'} q_*(S_{t+1}, a') \,|S_t = s, A_t = a\right]$$

$$= \sum_r \sum_{s'} p(s', r|s, a)\left[r + \gamma \max_{a'} q_*(s', a')\right]$$

سیاست و تابع ارزش بهینه

نمودار بک‌آپ

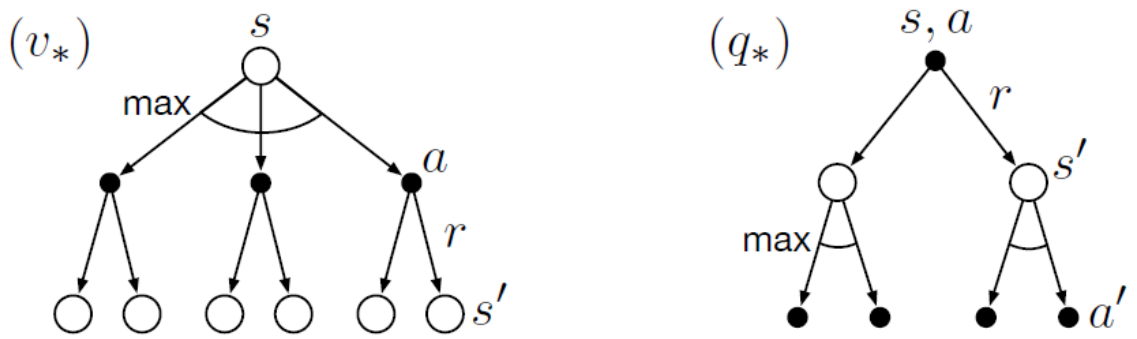


Figure 3.4: Backup diagrams for v_* and q_*

سیاست و تابع ارزش بهینه

معادلات بهینگی بلمن v_*

- دارای پاسخ منحصر بفرد
- دستگاه معادلات به تعداد حالت‌ها

در صورت اطلاع از دینامیک محیط p

- امکان حل دستگاه با استفاده از روش‌های حل دستگاه معادلات غیرخطی

به طریق اولی برای معادلات بهینگی بلمن q_*

سیاست و تابع ارزش بهینه

معادلات بهینگی بلمن v_*
- دارای پاسخ منحصر بفرد
- دستگاه معادلات به تعداد حالت‌ها

در صورت اطلاع از دینامیک محیط p
- امکان حل دستگاه با استفاده از روش‌های حل دستگاه معادلات غیرخطی

به طریق اولی برای معادلات بهینگی بلمن q_*

ساده‌بودن تعیین سیاست بهینه با معلوم بودن مقدار v_*
- هر حالت دارای یک یا چند کنش بدست‌آورندهٔ بیشینه معادلات بهینگی بلمن
- سیاستی با تخصیص احتمال غیرصفر به چنین کنش‌هایی← سیاست بهینه
  - جستجوی تک-گام؟
  - هر سیاست حریصانه با توجه به تابع بهینگی v_*
    - توجه به معنای موضعی بودن حرص و نزدیک‌بینی در دانش رایانه
    - در این موقعیت حریصانه در بلندمدت بهینه

سیاست و تابع ارزش بهینه

داشتن معادلات بهینگی بلمن q_*

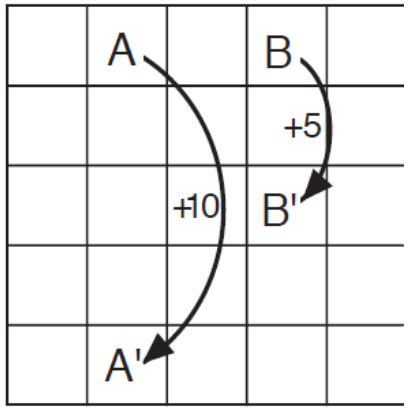
- ساده‌تر کردن انتخاب کنش‌های بهینه
- ؟

سیاست و تابع ارزش بهینه

داشتن معادلات بهینگی بلمن q_*

- ساده‌تر کردن انتخاب کنش‌های بهینه
- ؟
  - در دست داشتن کنش بیشینه‌ساز $q_*(s,a)$
  - تابع ارزش کنش
    - نگهداری تمامی نتایج جستجوی‌های تک-گام
    - نگهداری اطلاعات محلی بهینگی بلندمدت
- هزینه بیشتر نگهداری مقادیر حالت-کنش‌ها به جای حالت‌ها
- تابع ارزش-کنش فراهم‌آور انتخاب کنش بهینه بدون داشتن اطلاع از جالت‌های جانشین و ارزش‌های آنها
  - بدون داشتن اطلاع از دینامیک محیط

مثال- دنیای شطرنجی



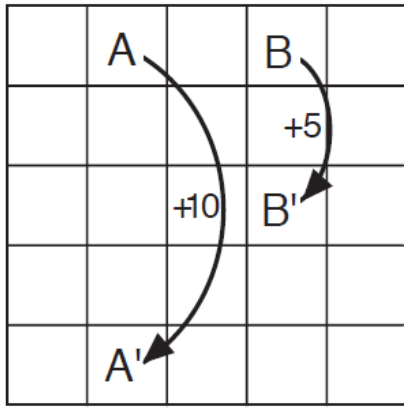
Gridworld

مثال- دنیای شطرنجی

مقادیر بهینه تابع ارزش

22.0	24.4	22.0	19.4	17.5
19.8	22.0	19.8	17.8	16.0
17.8	19.8	17.8	16.0	14.4
16.0	17.8	16.0	14.4	13.0
14.4	16.0	14.4	13.0	11.7

v_*

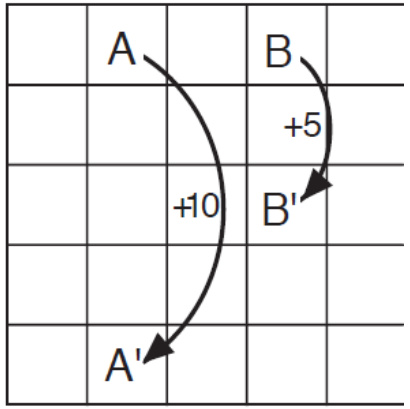


Gridworld

مثال- دنیای شطرنجی

22.0	24.4	22.0	19.4	17.5
19.8	22.0	19.8	17.8	16.0
17.8	19.8	17.8	16.0	14.4
16.0	17.8	16.0	14.4	13.0
14.4	16.0	14.4	13.0	11.7

v_*



Gridworld

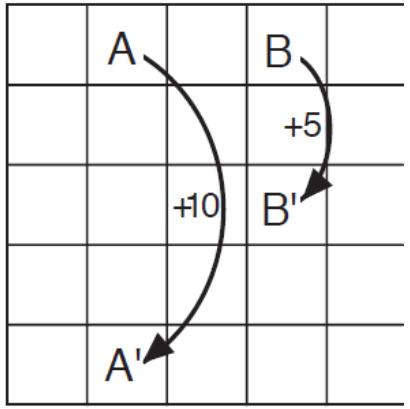
$\gamma = 0.9$

$v_*(x)$

مثال- دنیای شطرنجی

22.0	24.4	22.0	19.4	17.5
19.8	22.0	19.8	17.8	16.0
17.8	19.8	17.8	16.0	14.4
16.0	17.8	16.0	14.4	13.0
14.4	16.0	14.4	13.0	11.7

v_*



Gridworld

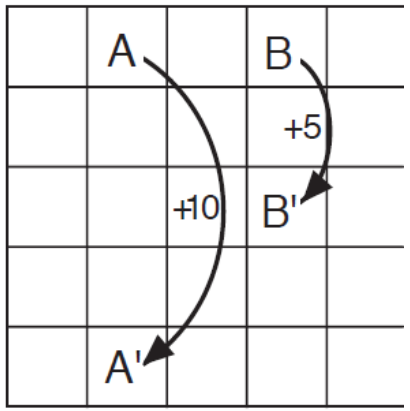
$\gamma = 0.9$

$$v_*(x) = \max_{\{\leftarrow\uparrow\rightarrow\downarrow\}}$$

مثال- دنیای شطرنجی

22.0	24.4	22.0	19.4	17.5
19.8	22.0	19.8	17.8	16.0
17.8	19.8	17.8	16.0	14.4
16.0	17.8	16.0	14.4	13.0
14.4	16.0	14.4	13.0	11.7

v_*



Gridworld

$\gamma = 0.9$

$$v_*(x) = \max_{\{\leftarrow\uparrow\rightarrow\downarrow\}}$$

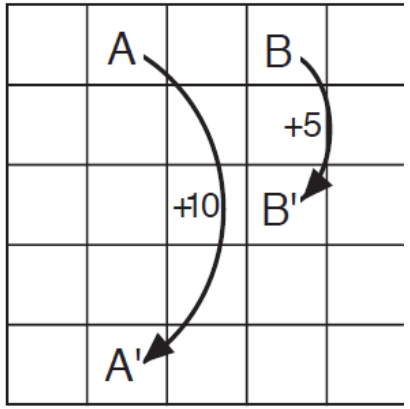
$$= \max_{\{\leftarrow\uparrow\rightarrow\downarrow\}} \{\underbrace{0 + 0.9 \times 13}_{\rightarrow}, \underbrace{0 + 0.9 \times 16}_{\uparrow}, \underbrace{0 + 0.9 \times 13}_{\downarrow}, \underbrace{0 + 0.9 \times 16}_{\leftarrow}\}$$

$$=$$

مثال- دنیای شطرنجی

22.0	24.4	22.0	19.4	17.5
19.8	22.0	19.8	17.8	16.0
17.8	19.8	17.8	16.0	14.4
16.0	17.8	16.0	14.4	13.0
14.4	16.0	14.4	13.0	11.7

v_*



Gridworld

$\gamma = 0.9$

$$v_*(x) = \max_{\{\leftarrow\uparrow\rightarrow\downarrow\}}$$

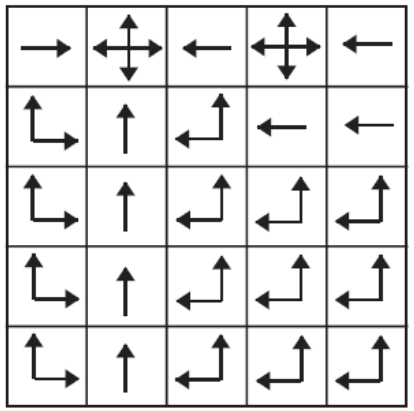
$$= \max_{\{\leftarrow\uparrow\rightarrow\downarrow\}} \{\underbrace{0 + 0.9 \times 13}_{\rightarrow}, \underbrace{0 + 0.9 \times 16}_{\uparrow}, \underbrace{0 + 0.9 \times 13}_{\downarrow}, \underbrace{0 + 0.9 \times 16}_{\leftarrow}\}$$

$$= 0.9 \times 16 = 14.4$$

مثال- دنیای شطرنجی

مقادیر بهینه تابع ارزش

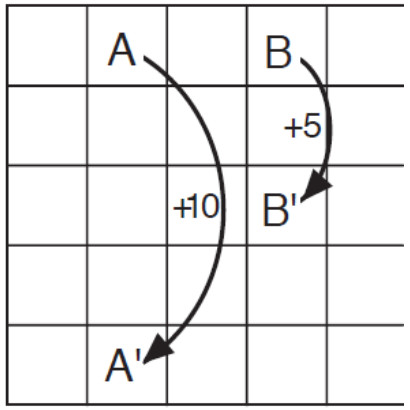
- سیاسات بهینهٔ متناظر



π_*

22.0	24.4	22.0	19.4	17.5
19.8	22.0	19.8	17.8	16.0
17.8	19.8	17.8	16.0	14.4
16.0	17.8	16.0	14.4	13.0
14.4	16.0	14.4	13.0	11.7

v_*

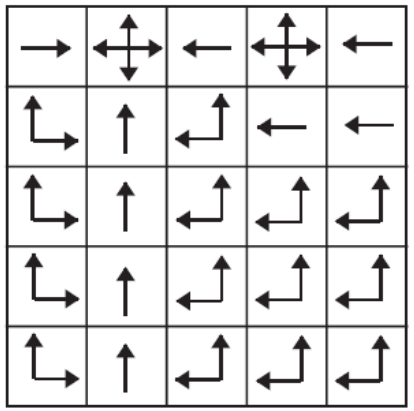


Gridworld

مثال- دنیای شطرنجی

مقادیر بهینه تابع ارزش

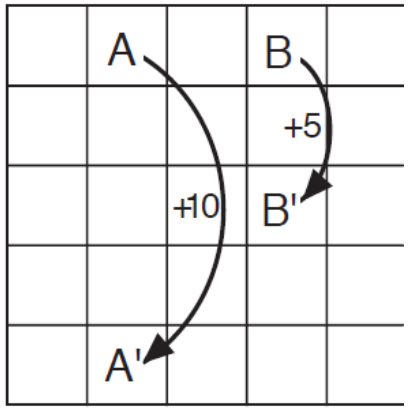
- سیاسات بهینهٔ متناظر
- وجود چند یال در مدخلی به معنای وجود چند کنش بهینه



π_*

22.0	24.4	22.0	19.4	17.5
19.8	22.0	19.8	17.8	16.0
17.8	19.8	17.8	16.0	14.4
16.0	17.8	16.0	14.4	13.0
14.4	16.0	14.4	13.0	11.7

v_*

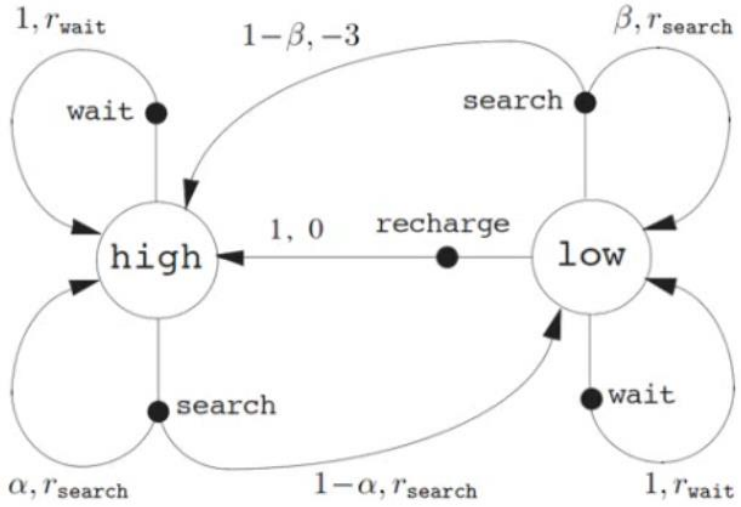


Gridworld

$$v_*(s) = \max_a \sum_r \sum_{s'} p(s', r|s, a)[r + \gamma v_*(s')]$$

مثال- زباله‌ساز-جمع کن!

s	a	s'	$p(s'\|s,a)$	$r(s,a,s')$
high	search	high	α	r_{search}
high	search	low	$1-\alpha$	r_{search}
low	search	high	$1-\beta$	-3
low	search	low	β	r_{search}
high	wait	high	1	r_{wait}
high	wait	low	0	-
low	wait	high	0	-
low	wait	low	1	r_{wait}
low	recharge	high	1	0
low	recharge	low	0	-



دو حالت پس معادله بهینگی بلمن دارای دو معادله

$$\begin{aligned}
v_*(\mathtt{h}) &= \max \left\{ \begin{array}{l} p(\mathtt{h}|\mathtt{h},\mathtt{s})[r(\mathtt{h},\mathtt{s},\mathtt{h}) + \gamma v_*(\mathtt{h})] + p(\mathtt{l}|\mathtt{h},\mathtt{s})[r(\mathtt{h},\mathtt{s},\mathtt{l}) + \gamma v_*(\mathtt{l})], \\ p(\mathtt{h}|\mathtt{h},\mathtt{w})[r(\mathtt{h},\mathtt{w},\mathtt{h}) + \gamma v_*(\mathtt{h})] + p(\mathtt{l}|\mathtt{h},\mathtt{w})[r(\mathtt{h},\mathtt{w},\mathtt{l}) + \gamma v_*(\mathtt{l})] \end{array} \right\} \\
&= \max \left\{ \begin{array}{l} \alpha[r_{\mathtt{s}} + \gamma v_*(\mathtt{h})] + (1-\alpha)[r_{\mathtt{s}} + \gamma v_*(\mathtt{l})], \\ 1[r_{\mathtt{w}} + \gamma v_*(\mathtt{h})] + 0[r_{\mathtt{w}} + \gamma v_*(\mathtt{l})] \end{array} \right\} \\
&= \max \left\{ \begin{array}{l} r_{\mathtt{s}} + \gamma[\alpha v_*(\mathtt{h}) + (1-\alpha) v_*(\mathtt{l})], \\ r_{\mathtt{w}} + \gamma v_*(\mathtt{h}) \end{array} \right\}.
\end{aligned}$$

$$v_*(\mathtt{l}) = \max \left\{ \begin{array}{l} \beta r_{\mathtt{s}} - 3(1-\beta) + \gamma[(1-\beta) v_*(\mathtt{h}) + \beta v_*(\mathtt{l})], \\ r_{\mathtt{w}} + \gamma v_*(\mathtt{l}), \\ \gamma v_*(\mathtt{h}) \end{array} \right\}$$

$$v_*(s) = \max_a \sum_r \sum_{s'} p(s', r|s, a)[r + \gamma v_*(s')]$$

دنیای راهرو!

پای تخته

سیاست و تابع ارزش بهینه

حل معادلات بهینگی بلمن

- استفاده نادر
- نیاز به جستجوی کامل
- بر مبنای سه فرض
  - اطلاع کامل و دقیق از دینامیک محیط
  - دارا بودن منابع رایانشی کافی
  - خاصیت مارکوفی
- برقرار نبودن ترکیبی از این سه فرض در بسیاری از موارد
- تخته نرد
  - فرض دوم؟ $۱۰^{۲۰}$ حالت
- معمولا استفاده از راه‌حلی‌ها تخمینی

منابع

ساتن

زندی

روش‌های عددی

تعداد حالات فراوان

وجود دینامیک سیستم

- روش برنامه‌ریزی پویا

عدم وجود دینامیک سیستم

- روش مونت کارلو
- روش‌های تفاضل زمانی
  - سارسا
  - یادگیری ک

سیاست و تابع ارزش بهینه

سیاست بهینه دارای تابع ارزش حالت-کنش بهینه q_*

$$q_*(s,a) = \max_{\pi} q_\pi(s,a)\,, \forall s \in S, a \in A(s)$$

امکان تقریر توابع ارزش بهینه بر اساس یکدیگر

$$q_*(s,a) = E[R_{t+1} + \gamma v_*(S_{t+1}) | S_t = s, A_t = a]$$

- چرا؟

سیاست بهینه π^*

$$\pi^*(s) = \operatorname*{argmax}_a q_*(s,a)$$

$$= \operatorname*{argmax}_a \sum_r \sum_{s'} p(s', r|s,a)[r + \gamma v_*(s')]$$

$$= \operatorname*{argmax}_a \sum_r \sum_{s'} \underbrace{p(s', r|s,a)}_{\text{دینامیک}} [r + \gamma v_*(s')]$$