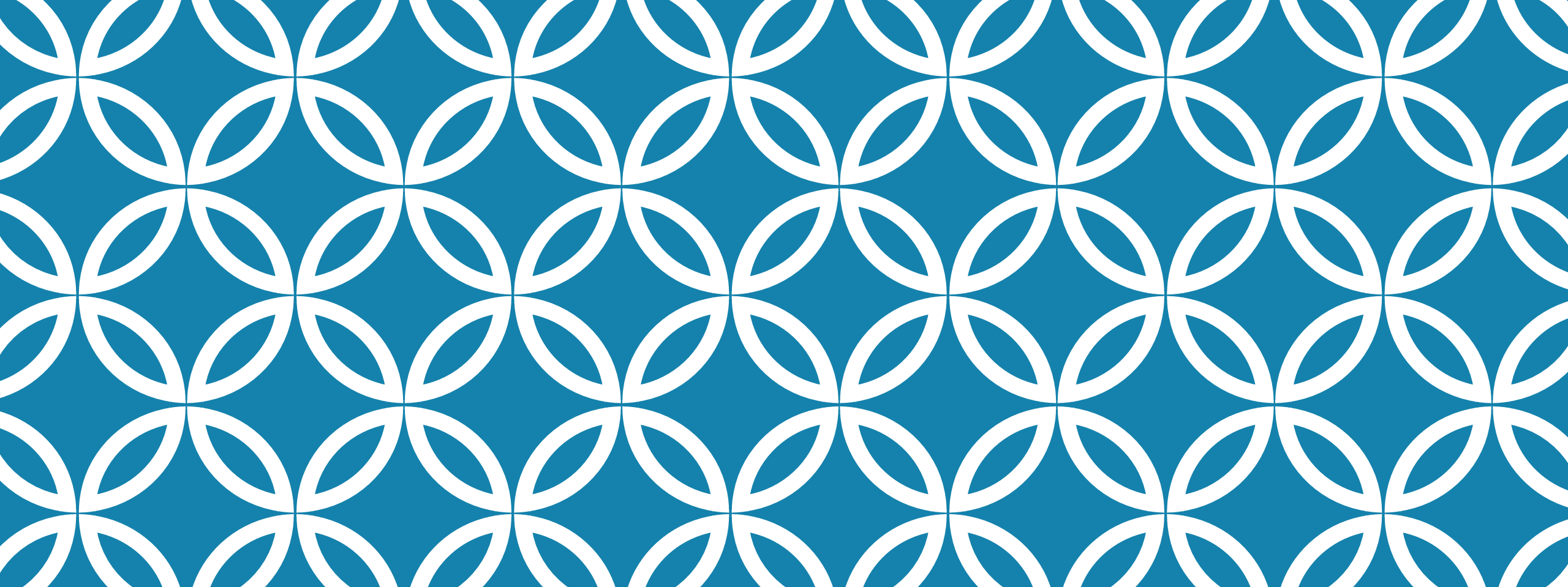
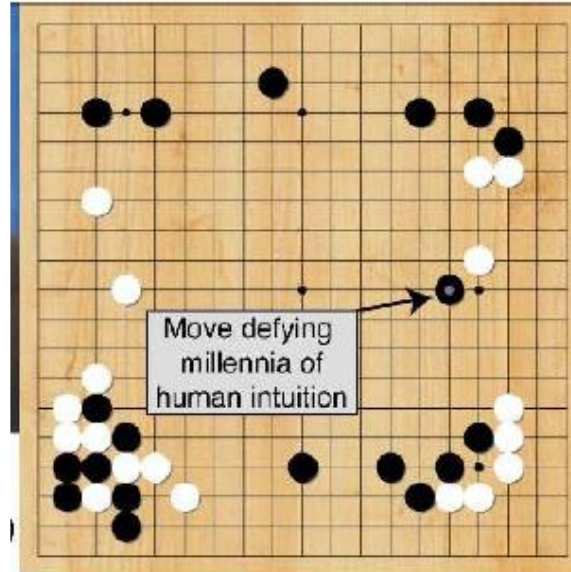




یادگیری تقویتی دق الباب

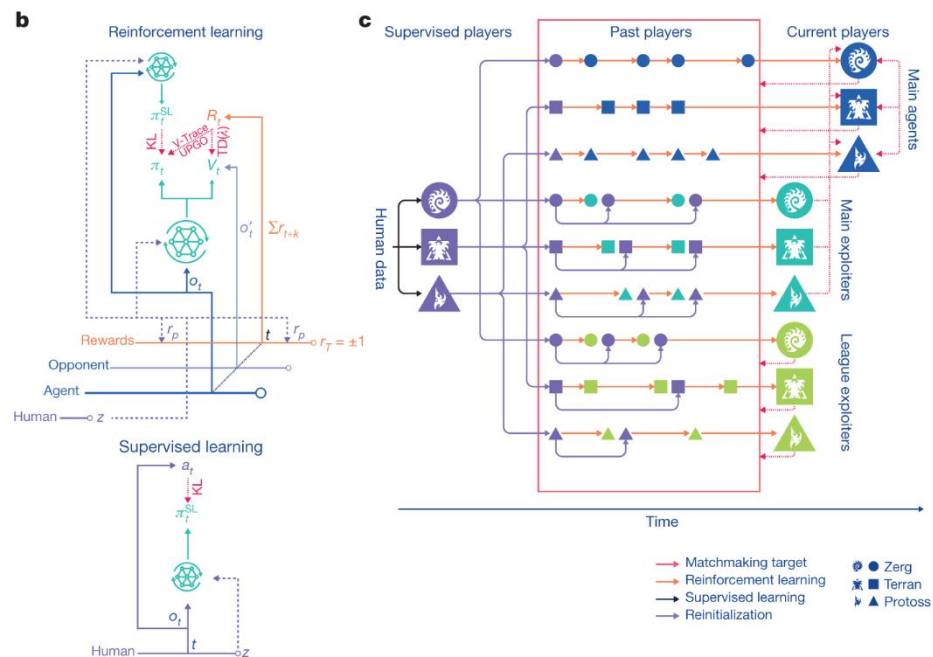
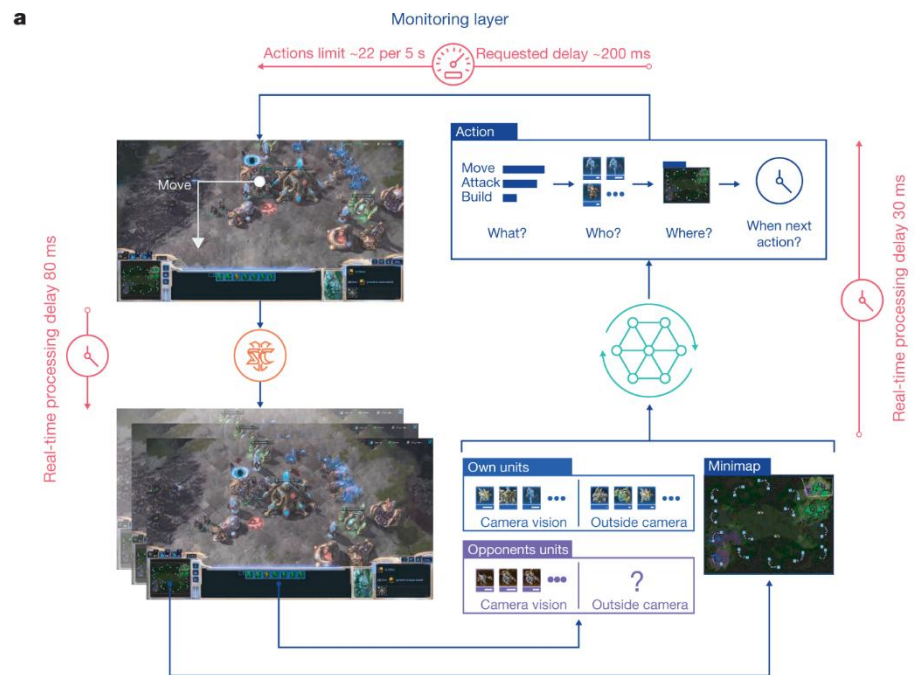
محسن هوشمند
دانشکده تکنولوژی اطلاعات و علم رایانه
دانشگاه تحصیلات تکمیلی علوم پایه زنجان



Using a new unknown and creative pattern

Grandmaster level in StarCraft II using multi-agent reinforcement learning

یادگیری تقویتی



بازی‌های متنی

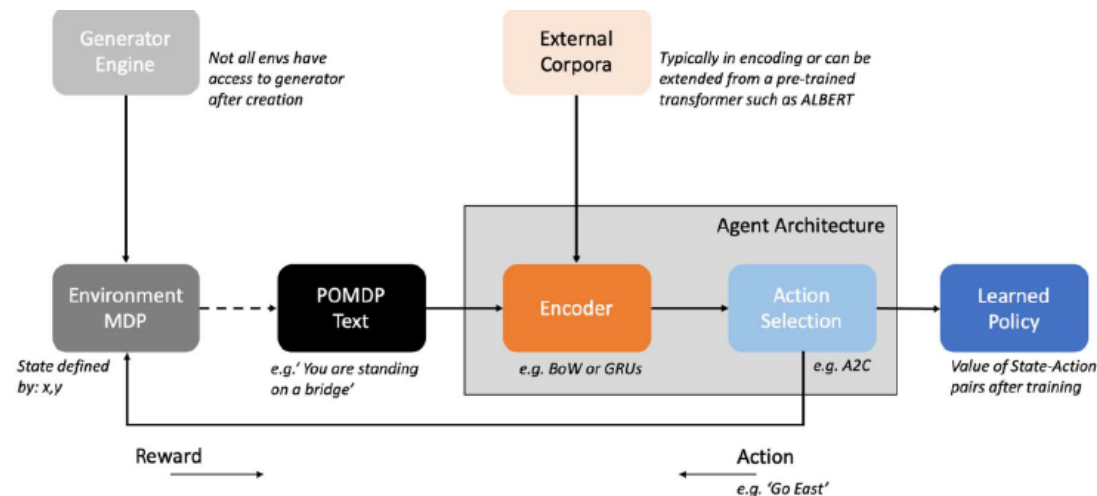


Figure 2: Overview of the Architecture Structure of Agents Applied to a Simple Text Game Example.

Name	Encoder	Action Selector	KG	PTF	Pre-Training	AS	Tasks
Ammanabrolu et al. (2020a)	GRU	A2C	DL	ALBERT	J-QA	TB	Zork1
Xu et al. (2020)	GRU	A2C	DL	none	none	TB	JSOG
Ammanabrolu and Hausknecht (2020)	GRU	A2C	DL	none	ClubFloyd	TB	JSOG
Murugesan et al. (2020b)	GRU	A2C	DL+CS	none	GloVe	CB	TW-Commonsense
Yao et al. (2020)	GRU	DRRN	none	GPT-2	ClubFloyd	CB	JSOG
Adolphs and Hofmann (2020)	Bi-GRU	A2C	none	none	TS, GloVe ₁₀₀	TB	CW
Guo et al. (2020)	Bi-GRU	DQN	none	none	GloVe ₁₀₀	TB	JSOG
Xu et al. (2020)	TF	DRRN	none	none	none	TB	JSOG
Yin and May (2020)	TF,LSTM	DSQN	none	none	none	NS	CW, TH
Adhikari et al. (2020)	R-GCN, TF	DDQN	DL	none	TS	CB	TW-Cook
Zahavy et al. (2018)	CNN	DQN	none	none	word2vec ₃₀₀	CB	Zork1
Ammanabrolu and Riedl (2019)	LSTM	DQN	DL	none	TS, GloVe ₁₀₀	CB	TW-Home
He et al. (2016)	BoW	DRRN	none	none	none	CB	other
Narasimhan et al. (2015)	LSTM	DQN	none	none	none	PB	other
Yin et al. (2020)	BERT	DQN	DL	BERT	none	CB	FTWP, TH
Madotto et al. (2020)	LSTM	Seq2Seq	none	none	GloVe _{100,300}	PB	CC, CW

Table 2: Overview of recent architectural trends. ENC, state/action encoder; KG, knowledge graph (DL: dynamically learned; CS: commonsense); PTF, pretrained Transformer; PreTr, pretraining (TS: task specific); AS, Action space (TB: template-based; PB: parser based; CB: choice-based).

یادگیری تقویتی

Reinforcement

روانشناسی

▪ رفتاری؟

▪ یادگیری انجمنی (یادگیری تداعی)

▪ شرطی شدن پاولوفی (شرطی شدن کلاسیک)

▪ پاولوف

▪ ارتباط دو محرک باهم جهت ایجاد پاسخ

▪ صدای پای مسئول غذا به سگ

▪ آلبرت کوچولو

▪ شرطی شدن فعال (کنشگر)

▪ تورندایک (سراندایک ثوراندایک)

▪ اسکینر

▪ ۱. شرطی شدن پاسخگر

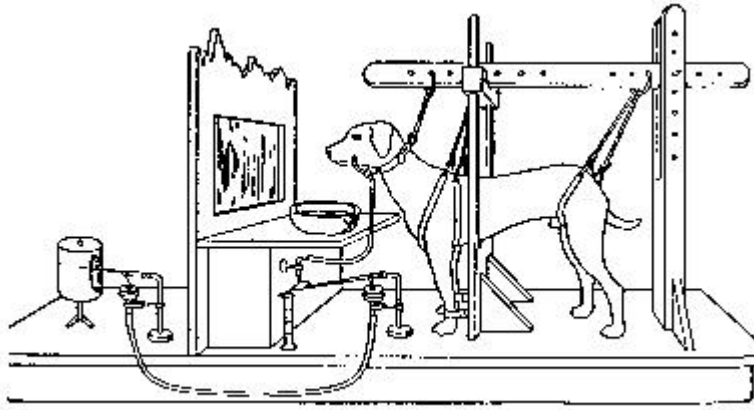
▪ ۲. شرطی شدن کنشگر (شبیه شرطی شدن وسیله‌ای ثرندایک)

▪ کنترل رفتار با استفاده از تنبیه و پاداش

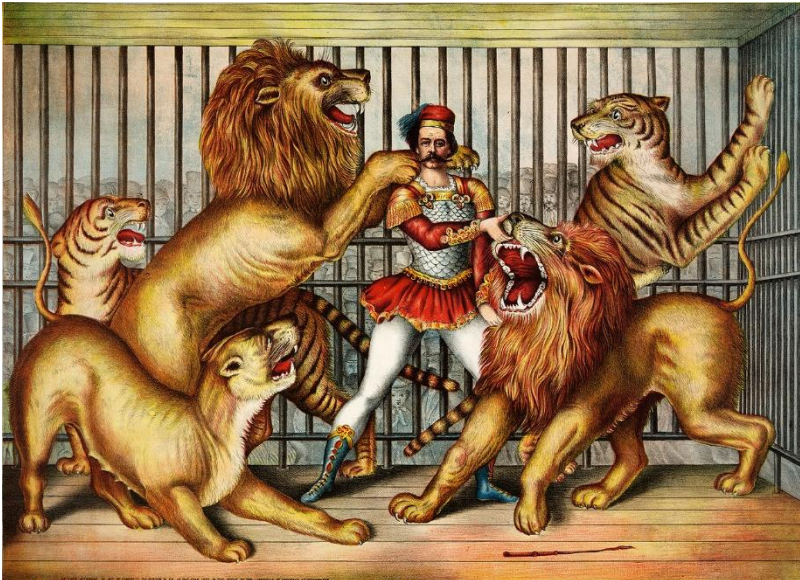
▪ نقد نوام چامسکی

▪ یادگیری اجتماعی و تقلید

▪ بندورا



https://en.wikipedia.org/wiki/Classical_conditioning#/media/File:Ivan_Pavlov_research_on_dog's_reflex_setup.jpg



یادگیری تقویتی

نظریه کنترل

کارهای بلمن و هوارد

یادگیری تقویتی

یادگیری باناظر

- نگاشت ورودی و خروجی
- مجموعه کنش درست
- ی ت استفاده از پاداش و جریمه جهت تشخیص رفتار درست از نادرست

یادگیری بی ناظر

- یافتن شباهت ها و تفاوت ها بین نمونه داده ها
- ی ت یافتن مدل کنش بیشینه ساز جمع پاداش

یادگیری تقویتی

از اقسام بهینه‌سازی

- رفتار بهینه

جستجوی عامل در محیط

- تعامل عامل با محیط

- استفاده از آن جهت اعمال کنش متناسب

- حاصل تعامل پاداش بلافصل

- پاداش بلافصل به مثابهٔ اندازه‌گیری عملکرد

- جهت دستیابی به بیشترین پاداش

- رفتار عامل به صورت بیشینه‌سازی پاداش‌های آینده

به سخن دیگر

- عامل قادر به یادگیری از طریق تعامل به محیط

آزمون و خطا، و بازخوردگیری از کنش‌ها و تجاربش

تقریب ذهن

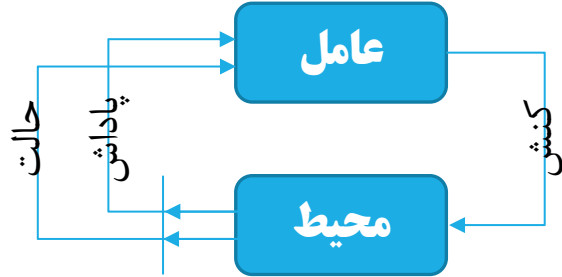
- دستم بگرفت و پا به پا برد.

- کاوش در برابر بهره‌گیری

یادگیری تقویتی

عناصر اصلی

- عامل
- محیط
- کنش
- سیاست
- پاداش
- ارزش



یادگیری تقویتی

تمامی خصوصیات عامل و محیط ذخیره در حالت دستگاه

- مشاهده‌پذیری کامل $O_t = S_t^a = s_t^e$

- مشاهده‌پذیری ناکامل $S_t^a \neq s_t^e$

عاملیت عامل در گرو کنش‌هایش!

- کنش در هر گام زمانی

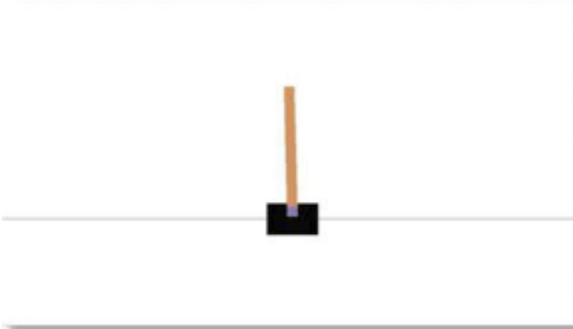
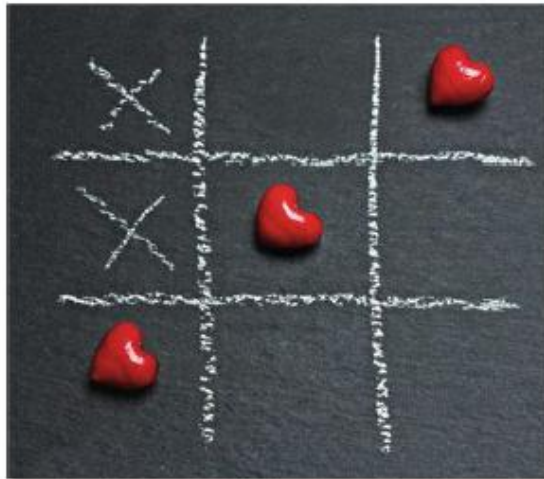
- تاثیر و تبعات در گام زمانی بعد

حالت فعلی به علاوه کنش

- منجر به حالت بعدی

- پاداش بلافصل

کنش و پاداش



عملکرد تصمیم

تعیین پاداش مناسب

- ساده و بدیهی نبودن تعریف پاداش مناسب
- تغییر جزئی منجر به رفتار متمایز
- تدوین متفاوت ← کنش متفاوت

مثال

- دوز
- موازنه پایه چرخ
- دنیای شطرنجی
- قارچ خور

کاوش و بهره explore vs exploit

یادگیری تقویتی

- جلوه‌های متفاوت نمایش
 - معمول‌ترین زمان گسسته
 - کنش در زمان فعلی
 - مشخص شدن حالت بعدی با حالت فعلی و کنش
 - زمان پیوسته
 - تدوین پیچیده‌تر
- در صورت تصادفی بودن دینامیک دستگاه
 - تدوین حالت فعلی و کنش فعلی با توزیع احتمال

یادگیری تقویتی

▪ فضای حالت

▪ گسسته

▪ متناهی و نامتناهی

▪ پیوسته

▪ فضای کنش

▪ گسسته

▪ پیوسته

▪ تابع سیاست

▪ تبیین رفتار عامل

▪ نگاشتی از حالت به کنش در زمان مقتضی

▪ محتملا احتمالی

▪ سیاست مشخص کننده رفتار عامل $\Leftarrow$ آماج یادگیری و بهینه‌سازی

▪ هدف عامل: یافتن سیاست بیشینه‌ساز پاداش‌های آینده یا امید ریاضی (مقدار موردانتظار آن)

▪ جمع پاداش گسسته یا پیوسته (انتگرال‌گیری) در افق زمانی متناهی یا نامتناهی

▪ محتملا استفاده از ضریب کاهنده برای زمان‌های دور از زمان فعلی

یادگیری تقویتی

فرایند تصمیم مارکوفی $(S, A, R, P(s'|s, a))$

روش‌ها

- ارزش‌محور: بدون سیاست، تعیین حریصانه کنش برحسب انرژی حالت
- سیاست‌محور: بدون تابع انرژی، استفاده از تابع سیاست جهت تعیین کنش
- بازیگر-منتقد: استفاده از هر دوی سیاست و ارزش

روش‌ها

- بی‌مدل: استفاده از یکی یا هر دو ولی بی‌مدل
- بامدل: استفاده از یکی یا هر دو با مدل

یادگیری تقویتی

مسائل اساسی تصمیم‌سازی متوالی

- یادگیری: محیط مجهول است
- طرح‌ریزی: محیط معلوم است

مطالب درس

مقدمه‌ای بر تصمیم‌سازی متوالی یا چندگامی

مقدمه‌ای بر کاوش و بهره‌برداری

فرایندهای تصمیم مارکوفی

روش‌های برنامه‌ریزی پویا

روش‌های مونت کارلو

روش‌های تفاضل زمانی

- سیاست‌ندار! سارسا

- سیاست‌مدار! یادگیری ک

پیش‌بینی تقریبی سیاست‌مدار

DQN استفاده از شبکه عمیق برای تخمین Q

DDPG بازیگر منتقد

منابع

ساتن و بارتو

برتسکاس

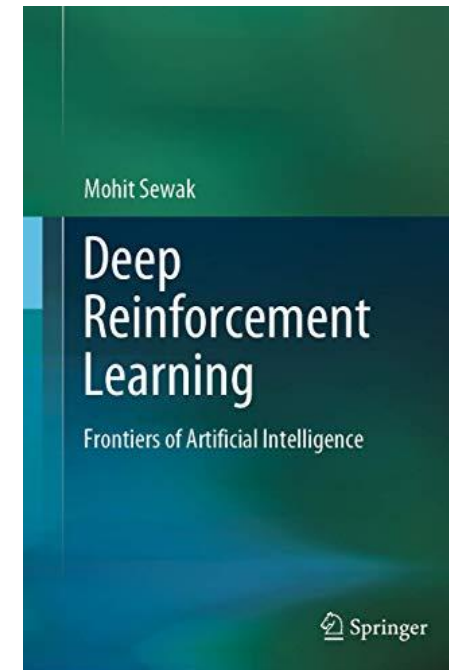
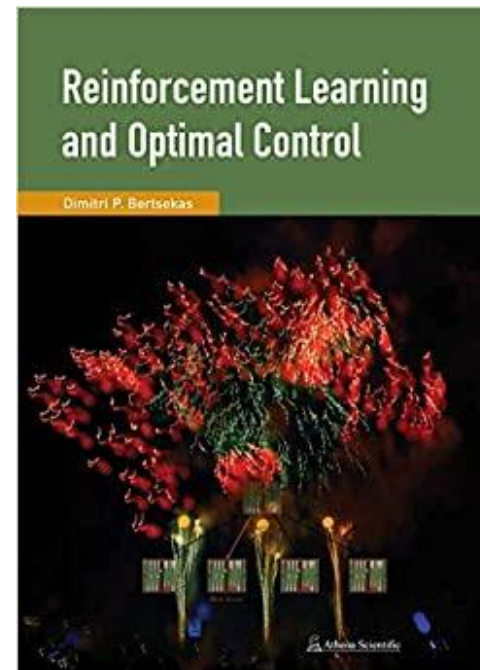
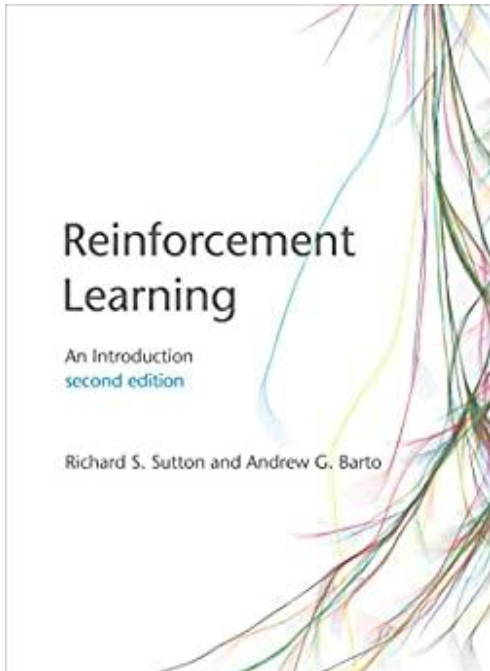
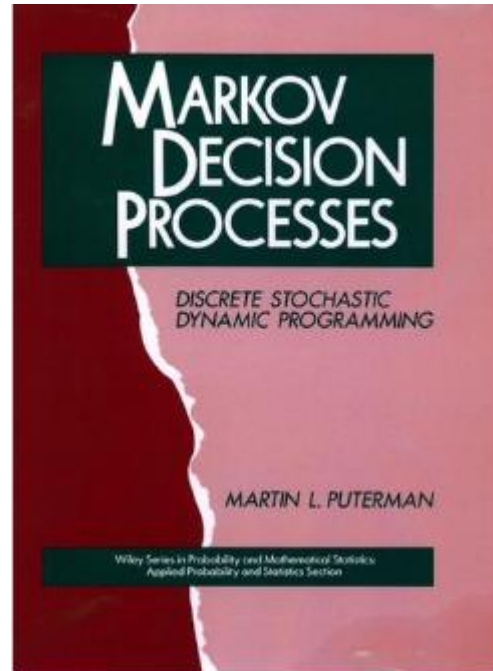
پوترمان

سواک

بلمن

مقالات

درس‌های مرتبط



ارزیابی

ارزیابی

▪ امتحان

▪ تحقیق

▪ مطالعه (سروی و روش جدید)

▪ تمرین

▪ قلم و کاغذ

▪ پیاده‌سازی

حل تمرین

▪ جواد اصغری

▪ نشانی

▪ <https://iasbs.ac.ir/~mohsen.hooshmand/courses/0405/1/Yadgiri-Taqvati.html>

fararu.com

reinforcement learn... Compiler Imported From IE Longman Dictionar... Dictionary.com | M... SASI cplusplus.com - Th... نظام آموزشی دانشکد... AFI's 10 Top 10 - Wi... IPA, Alphabet, and... Contrasting Acco

صفحه نخست سیاست ورزشی علم و تکنولوژی عکس ویدیو راهنمای بازار زندگی و سرگرمی اقتصاد جامعه

آموزش پخت مرغ به سبکی جدید با سبزی چگونه غذای مناسب برای سن خود بخوریم؟

بزرگترین افسانه‌ها درباره مفز نوجوانان
ضربان قلب درمان در خانم‌ها و آقایان
طرز تهیه سالاد اسفناج و هویج، یک سالاد مهیج و خوشمزه
این مواد غذایی اشتها را کاهش می‌دهند.
مهارت «نه گفتن» و راه‌های تقویت آن

خواص پای مرغ؛ بایدها و نبایدهای آن
تمر هندی: از خواص آن برای سلامتی
واقعاً در دوران یائسگی چه اتفاقی
طرز تهیه دسر موز و انبه، دسری ساده
طرز تهیه مربای انگور، یک مربای بدون

رونویسی (کپی-پیست) و خلق محتوا و دانش

فرایندهای چندگامی

Multistage Process

فرایندهای متوالی

مطالعه سیستم‌ها و چگونگی عملکرد آنها در طول زمان

درک سیستم به مثابه بردار حالت $x(t)$ و قاعده تشخیص آن در طول زمان

هر مدخل (درایه) اندازه‌گیر ویژگی متفاوتی از سیستم

فرایندهای چندگامی

Multistage Process

فرایندهای متوالی

مطالعه سیستم‌ها و چگونگی عملکرد آنها در طول زمان

درک سیستم به مثابه بردار حالت $\mathbf{x}(t)$ و قاعده تشخیص آن در طول زمان
$$\mathbf{x}(t) = [x_1(t), x_2(t), \dots, x_N(t)]^T$$

هر مدخل (درایه) اندازه‌گیر ویژگی متفاوتی از سیستم

فرایندهای چندگامی

s نقطه‌ای در مجموعه R

- فرض تک بعدی
- امکان چند بعدی
- $\forall s \in R: s_1 = T(s) \in R$ تبدیل
- s حالت بعدی
- s_1 حالت آغاز

فرایند چندگامی $[s, s_1, s_2, \dots, s_n, \dots]$

- $s_0 = s, s_{n+1} = T(s_n), n = 0, 1, 2, \dots$
- $s_n = T^n(s)$
- تابع زمان
- $[s, T]$ یا $[s, T(s)]$

مثال

$$T(s) = rs$$

ماتریس چرخش

فرایندهای چندگامی

پیچیدگی زیاد مخل موضوع ← باید ساده کرد
کاهش داده

فرایند N-گامی

$$[s, s_1, s_2, \dots, s_N]$$

▪ $s_{n+1} = T(s_n), n = 0, 1, 2, \dots, N - 1$

▪ استقلال از گذشته جهت تعیین آینده

$$T^N = T^{N-k}(T^k)$$

- گزاره تحلیلی علیت
- یا متناظرا آینده صرفا با حال مشخص می شود
- قضیه یگانگی

روابط بازگشتی

h تابع دلخواه

$$\sum_{i=0}^N h(s_i)$$

روابط بازگشتی

h تابع دلخواه

$$f_N(s) = \sum_{i=0}^N h(s_i)$$

روابط بازگشتی

h تابع دلخواه

$$\sum_{i=0}^N h(s_i)$$

$$f_N(s) = \sum_{i=0}^N h(s_i) = h(s) + h(T(s)) + h(T^2(s)) + \cdots + h(T^N(s)), N = 0, 1, 2, \dots$$

روابط بازگشتی

h تابع دلخواه

$$\sum_{i=0}^N h(s_i)$$

$$f_N(s) = \sum_{i=0}^N h(s_i) = h(s) + h(T(s)) + h(T^2(s)) + \cdots + h(T^N(s)), N = 0, 1, 2, \dots$$

$$f_N(s) = h(s) + f_{N-1}(T(s))$$

روابط بازگشتی

h تابع دلخواه

$$\sum_{i=0}^N h(s_i)$$

$$f_N(s) = \sum_{i=0}^N h(s_i) = h(s) + h(T(s)) + h(T^2(s)) + \cdots + h(T^N(s)), N = 0, 1, 2, \dots$$

$$\begin{aligned} f_N(s) &= h(s) + f_{N-1}(T(s)) \\ f_0(s) &= h(s) \end{aligned}$$

روابط بازگشتی

h تابع دلخواه

$$f_N(s) = \prod_{i=0}^N h(s_i)$$

روابط بازگشتی

h تابع دلخواه

$$f_N(s) = \prod_{i=0}^N h(s_i) = h(s)f_{N-1}(T(s))$$

روابط بازگشتی

h تابع دلخواه

$$f_N(s) = \prod_{i=0}^N h(s_i) = h(s)f_{N-1}(T(s))$$
$$f_N(s) = \left\{ \max_{0 \leq i \leq N} h(s_i) \right\}$$

روابط بازگشتی

h تابع دلخواه

$$f_N(s) = \prod_{i=0}^N h(s_i) = h(s)f_{N-1}(T(s))$$
$$f_N(s) = \left\{ \max_{0 \leq i \leq N} h(s_i) \right\} = \max[h(s), f_{N-1}(T(s))], N \geq 1$$

روابط بازگشتی

h تابع دلخواه

$$f_N(s) = \prod_{i=0}^N h(s_i) = h(s) f_{N-1}(T(s))$$
$$f_N(s) = \left\{ \max_{0 \leq i \leq N} h(s_i) \right\} = \max[h(s), f_{N-1}(T(s))], N \geq 1$$
$$f_N(s) = \sum_{i=0}^N h(s_i, s_{i+1})$$

روابط بازگشتی

h تابع دلخواه

$$f_N(s) = \prod_{i=0}^N h(s_i) = h(s)f_{N-1}(T(s))$$
$$f_N(s) = \left\{ \max_{0 \leq i \leq N} h(s_i) \right\} = \max[h(s), f_{N-1}(T(s))], N \geq 1$$
$$f_N(s) = \sum_{i=0}^N h(s_i, s_{i+1}) = h(s, T(s)) + f_{N-1}(T(s))$$

مثال

$$s = 1 + r + r^2 + \dots$$

مثال

$$s = 1 + r + r^2 + \dots = 1 + r(1 + r + r^2 + \dots)$$

مثال

$$s = 1 + r + r^2 + \dots = 1 + r(1 + r + r^2 + \dots) = \frac{1}{1 - r}$$

مثال

$$s = 1 + r + r^2 + \dots = 1 + r(1 + r + r^2 + \dots) = \frac{1}{1 - r}$$

$$s = 1 + \frac{1}{1 + \frac{1}{1 + \dots}}$$

مثال

$$s = 1 + r + r^2 + \dots = 1 + r(1 + r + r^2 + \dots) = \frac{1}{1 - r}$$

$$s = 1 + \frac{1}{1 + \frac{1}{s}} \Rightarrow s = 1 + \frac{1}{s}$$

مثال

$$s = 1 + r + r^2 + \dots = 1 + r(1 + r + r^2 + \dots) = \frac{1}{1 - r}$$

$$s = 1 + \frac{1}{1 + \frac{1}{1 + \dots}} \Rightarrow s = 1 + \frac{1}{s}$$

$$s = 1 + \frac{x}{1 + \frac{x^2}{1 + \frac{x^4}{1 + \frac{x^8}{1 + \dots}}}}$$

مثال

$$s = 1 + r + r^2 + \dots = 1 + r(1 + r + r^2 + \dots) = \frac{1}{1 - r}$$

$$s = 1 + \frac{1}{1 + \frac{1}{1 + \dots}} \Rightarrow s = 1 + \frac{1}{s}$$

$$s = 1 + \frac{x}{1 + \frac{x^2}{1 + \frac{x^4}{1 + \frac{x^8}{1 + \dots}}}} \Rightarrow s(x) = 1 + \frac{1}{s(x^2)}$$

فرایندهای نامتناهی

$$N \rightarrow \infty$$

فرایندهای نامتناهی

$$N \rightarrow \infty$$

$$f(s) = \sum_{i=0}^{\infty} h(s_i)$$

فرایندهای نامتناهی

$$N \rightarrow \infty$$

$$f(s) = \sum_{i=0}^{\infty} h(s_i)$$

$$f(s) = h(s) + \sum_{i=1}^{\infty} h(s_i)$$

فرایندهای نامتناهی

$$N \rightarrow \infty$$

$$f(s) = \sum_{i=0}^{\infty} h(s_i)$$

$$f(s) = h(s) + \sum_{i=1}^{\infty} h(s_i)$$

$$f(s) = h(s) + f(T(s))$$

فرایندهای نامتناهی

$$N \rightarrow \infty$$

$$f(s) = \sum_{i=0}^{\infty} h(s_i)$$
$$f(s) = h(s) + \sum_{i=1}^{\infty} h(s_i)$$
$$f(s) = h(s) + f(T(s))$$

▪ عدم حقیقت فیزیکی

بی‌حدی $f_N(s), N \rightarrow \infty$ در بیشتر مواقع

▪ دارای رفتار میانگین یا رفتار حالت-پایدار

▪ امکان یافتن با رابطهٔ مجانبی

$$f_N(s) \sim g_2^N(s) h(s), N \rightarrow \infty \text{ یا } \frac{f_N(s)}{N} \sim g_1(s) N \rightarrow \infty$$

▪

فرایند با شرط پایان

$$f(s) = \sum_{i=0}^{N(s)} h(s_i)$$

ادامه کار تا $d(T^N s, q) \leq \epsilon$

▪ ϵ و q داده شده

▪ d معیار مسافت و فاصله

فرایندهای بی حد که لزوماً متناهی نیست.

$$\begin{cases} f(s) = h(s) \text{ اگر } d(s, q) \leq \epsilon \\ f(s) = h(s) + f(T(s)) \text{ اگر } d(s, q) \geq \epsilon \end{cases}$$

فرایندهای تصادفی

تاکنون فرض بر تبدیل T عامل انتقال s به s_1
▪ منحصر به فردی انتقال

معلوم نبودن و منحصر به فرد نبودن در تمامی مواقع
▪ نظریه احتمال

فرایندهای تصادفی

تاکنون فرض بر تبدیل T عامل انتقال s به s_1
▪ منحصر به فردی انتقال

معلوم نبودن و منحصر به فرد نبودن در تمامی مواقع
▪ نظریه احتمال

T تبدیلی تصادفی که بردار تصادفی s_1 با توزیع p

فرایند چندگامی $[s, s_1, s_2, \dots, s_n, \dots]$ از نوع گسسته تصادفی

بررسی امید ریاضی اصل اساسی فرایندهای تصادفی

فرایندهای تصادفی

$$s_n = T(s_{n-1}, r_n), n = 1, 2, 3, \dots$$

فرایندهای تصادفی

$$s_n = T(s_{n-1}, r_n), n = 1, 2, 3, \dots$$

متغیرهای تصادفی مستقل $s_0 = s, r_n$

فرایندهای تصادفی

$$s_n = T(s_{n-1}, r_n), n = 1, 2, 3, \dots$$

متغیرهای تصادفی مستقل $s_0 = s, r_n$

به دنبال امید ارزش

$$g(s) + g(s_1) + \dots + g(s_N)$$

فرایندهای تصادفی

$$s_n = T(s_{n-1}, r_n), n = 1, 2, 3, \dots$$

متغیرهای تصادفی مستقل $s_0 = s, r_n$

به دنبال امید ارزش

$$f_N(s) = E[g(s) + g(s_1) + \dots + g(s_N)]$$

فرایندهای تصادفی

$$s_n = T(s_{n-1}, r_n), n = 1, 2, 3, \dots$$

متغیرهای تصادفی مستقل $s_0 = s, r_n$

به دنبال امید ارزش

$$\begin{aligned} & g(s) + g(s_1) + \dots + g(s_N) \\ f_N(s) &= E[g(s) + g(s_1) + \dots + g(s_N)] \\ &= g(s) + E[g(s_1) + \dots + g(s_N)] \end{aligned}$$

فرایندهای تصادفی

$$s_n = T(s_{n-1}, r_n), n = 1, 2, 3, \dots$$

متغیرهای تصادفی مستقل $s_0 = s, r_n$

به دنبال امید ارزش

$$\begin{aligned} & g(s) + g(s_1) + \dots + g(s_N) \\ f_N(s) &= E[g(s) + g(s_1) + \dots + g(s_N)] \\ &= g(s) + E[g(s_1) + \dots + g(s_N)] \\ &= g(s) + f_{N-1}(T(s)) \end{aligned}$$

فرایندهای تصادفی

$$s_n = T(s_{n-1}, r_n), n = 1, 2, 3, \dots$$

متغیرهای تصادفی مستقل $s_0 = s, r_n$

به دنبال امید ارزش

$$\begin{aligned} & g(s) + g(s_1) + \dots + g(s_N) \\ f_N(s) &= E[g(s) + g(s_1) + \dots + g(s_N)] \\ &= g(s) + E[g(s_1) + \dots + g(s_N)] \\ &= g(s) + f_{N-1}(T(s)) \\ f_0(s) &= g(s) \end{aligned}$$

فرایندهای تصمیم چندگامی

← نظریه برنامه‌ریزی پویا

دارای موقعیت بنیادی در نظریه کنترل جدید و بالتبع یادگیری تقویتی

تصمیم

سیاست

$[S, T(S)]$

▪ $T = T(s, a)$ انتخاب a_i از مجموعه $S(a)$ و a_i انتخاب در گام i -ام

$$s_1 = T(s, a_0)$$

$$s_2 = T(s_1, a_1)$$

⋮

$$s_{n+1} = T(s_n, a_n)$$

▪ a_i بردار تصمیم یا متغیر تصمیم (رابطه با کنش action)

▪ تصمیم: انتخاب مقدار q_i

فرایندهای تصمیم چندگامی

▪ a_i بردار تصمیم یا متغیر تصمیم (رابطه با کنش action)

▪ تصمیم: انتخاب مقدار a_i

به دنبال فرایندهائی که انتخاب مقدار q_i به منظور بیشینه‌سازی تابع عددی از پیش معینی از حالت و متغیر تصمیم

$$R(s, s_1, s_2, \dots; a_0, a_1, \dots)$$

▪ تابع بازده یا تابع عیار

فرایند تصمیم N-گامی (گسسته) عبارت است از مجموعه بردارهایی

$$[s, s_1, s_2, \dots, s_N; a_0, a_1, a_2, \dots, a_N]$$

$$\forall n: s_{n+1} = T(s_n, a_n)$$

سیاست

Policy

نیاز به انتخاب a_i در هر گام

- بسته به حالت فعلی و گذشته و آینده
- بسته به تصمیمات گذشته و آینده
- فعلا تمرکز بر گذشته و حال

$$a_i = a_i(s, s_1, s_2, \dots, s_i; a_0, a_1, \dots, a_{i-1})$$

- تابع مذکور را تابع سیاست خوانده کنیم (یا صرفا سیاست)

سیاست بهینه: سیاست بیشینه ساز تابع R

سیاست

$$q_i = q_i(s, s_1, s_2, \dots, s_i; a_0, a_1, \dots, a_{i-1})$$

▪ باز کلی

▪ تابع صرفا تابعی از حالت فعلی سیستم.

$$a_i = a_i(s_i) \quad \text{▪}$$

$$\pi_i = [s_i, s_{i-1}, \dots]$$

$$a_i = a_i(\pi_i)$$

اصل بهینگی

عیار جدا سازی

اصل بهینگی

سیاست بهینه سیاستی است که فارغ از حالت آغازین و تصمیم آغازین، بقیه السیف تصمیمها باید سیاستی بهینه را با توجه به حالت حاصل از تصمیم نخست تشکیل دهند.

بیشینه‌سازی تابع بازده

$$R(s, s_1, s_2, \dots; a_0, a_1, \dots) = \sum_{i=0}^N g(s_i, a_i)$$

$f_N(s)$ بازده کل N -گامی حاصل سیاست بهینه و با آغاز از حالت s

با استفاده از اصل بهینگی

$$g(s, a_0) + [g(s_1, a_1) + \dots + g(s_N, a_N)] = g(s, a_0) + f_{N-1}(T(s, a_0))$$

$\Leftarrow$

$$f_N(s) = \max_{a_0} [g(s, a_0) + f_{N-1}(T(s, a_0))], N \geq 1$$

بیشینه‌سازی تابع بازده

$$f_N(s) = \max_{a_0} [(s, a_0) + f_{N-1}(T(s, a_0))] , N \geq 1$$

اثبات

$$\begin{aligned} \max_{[a_0, \dots, a_n]} R &= \max_{a_0} \max_{[a_1, \dots, a_n]} R \\ f_N(s) &= \max_{[a_0, \dots, a_n]} [g(s, a_0) + g(s_1, a_1) + \dots + g(s_N, a_N)] \\ &= \max_{a_0} \max_{[a_1, \dots, a_n]} [g(s, a_0) + g(s_1, a_1) + \dots + g(s_N, a_N)] \end{aligned}$$

بیشینه‌سازی تابع بازده

$$f_N(s) = \max_{a_0} [(s, a_0) + f_{N-1}(T(s, a_0))] , N \geq 1$$

اثبات

$$\begin{aligned} f_N(s) &= \max_{[a_0, \dots, a_n]} R = \max_{a_0} \max_{[a_1, \dots, a_n]} R \\ &= \max_{a_0} \max_{[a_1, \dots, a_n]} [g(s, a_0) + g(s_1, a_1) + \dots + g(s_N, a_N)] \\ &= \max_{a_0} \left[g(s, a_0) + \max_{[a_1, \dots, a_n]} [g(s_1, a_1) + \dots + g(s_N, a_N)] \right] \end{aligned}$$

بیشینه‌سازی تابع بازده

$$f_N(s) = \max_{a_0} [(s, a_0) + f_{N-1}(T(s, a_0))] , N \geq 1$$

اثبات

$$\begin{aligned} f_N(s) &= \max_{[a_0, \dots, a_n]} R = \max_{a_0} \max_{[a_1, \dots, a_n]} R \\ &= \max_{a_0} \max_{[a_1, \dots, a_n]} [g(s, a_0) + g(s_1, a_1) + \dots + g(s_N, a_N)] \\ &= \max_{a_0} \left[g(s, a_0) + \max_{[a_1, \dots, a_n]} [g(s_1, a_1) + \dots + g(s_N, a_N)] \right] \\ &= \max_{a_0} [g(s, a_0) + f_{N-1}(T(s, a_0))] \end{aligned}$$

منابع

بلمن ۱۹۶۵

ساتن ۲۰۱۸