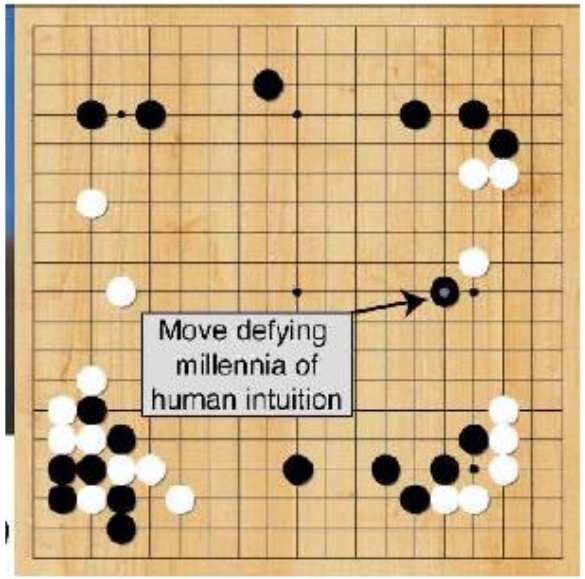
یادگیری تقویتی
دق الباب

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



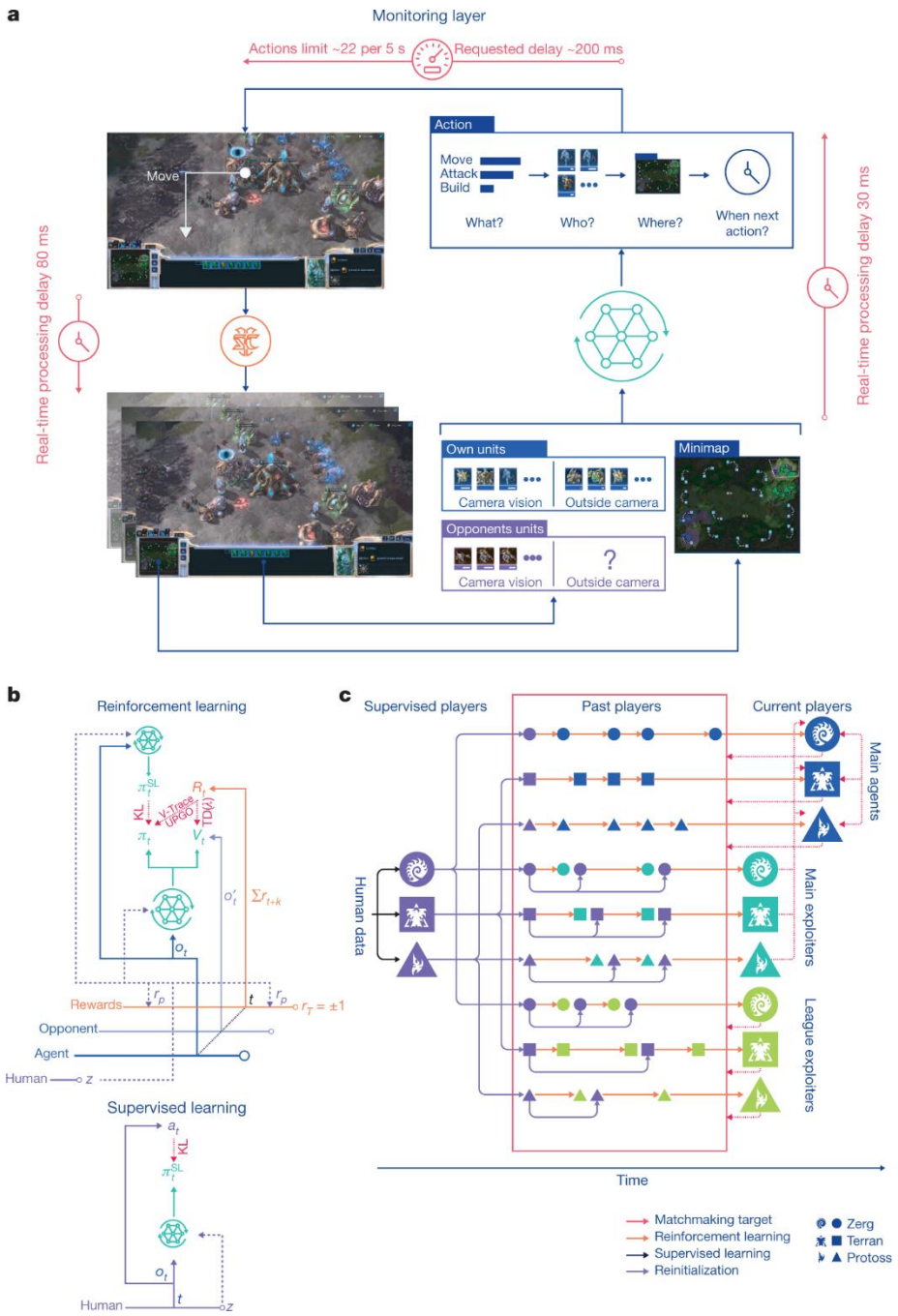
AlphaGo

Using a new unknown and creative pattern

AlphaGo, Deep Mind

Grandmaster level in StarCraft II using multi-agent reinforcement learning

یادگیری تقویتی



A Survey of Text Games for Reinforcement Learning Informed by Natural Language

بازی‌های متنی

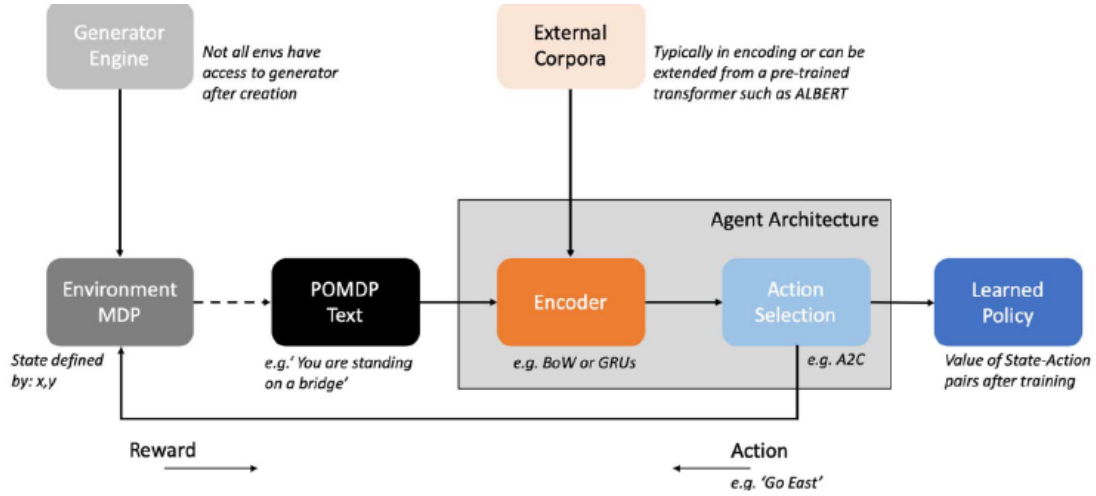


Figure 2: Overview of the Architecture Structure of Agents Applied to a Simple Text Game Example.

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	$\text{GloVe}_{100,300}$	PB	CC, CW

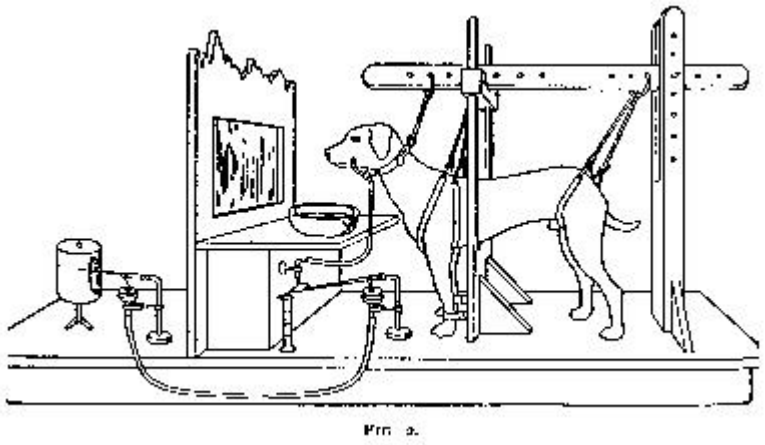
Table 2: Overview of recent architectural trends. ENC, state/action encoder; KG, knowledge graph (DL: dynamically learned; CS: commonsense); PTF, pretrained Transformer; PreTr, pretraining (TS: task specific); AS, Action space (TB: template-based; PB: parser based; CB: choice-based).

یادگیری تقویتی

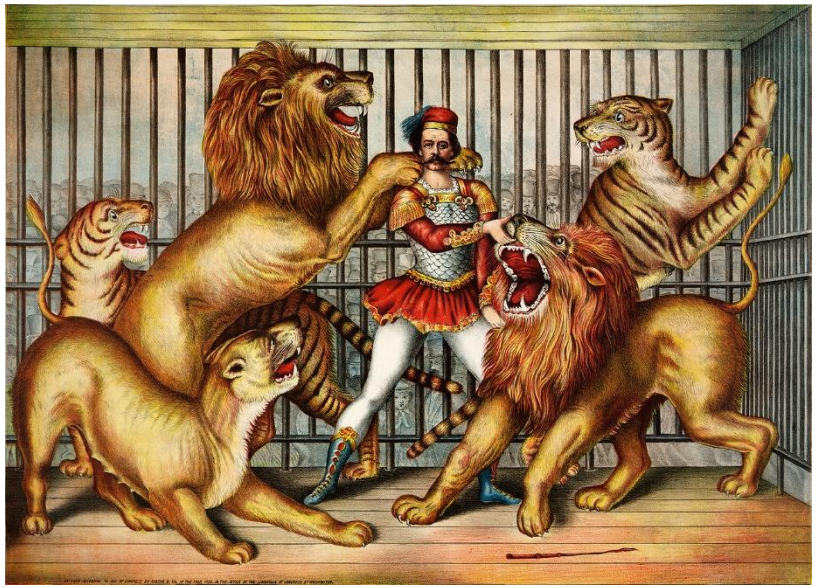
Reinforcement

روانشناسی

- رفتاری؟
- یادگیری انجمنی (یادگیری تداعی)
  - شرطی شدن پاولوفی (شرطی شدن کلاسیک)
    - پاولوف
  - شرطی شدن فعال
    - تورندایک (سراندایک ثورانداک)
    - اسکینر
      - 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



https://fa.wikipedia.org/wiki/%D8%B4%D8%B1%D8%B7%DB%8C%E2%80%8C%D8%B4%D8%AF%D9%81%D%_9%86D8%B9%D8%A7%D9%D%/aidem/#9%84BE%D8%B1%D9%88%D9%86%D8%AF%D_noiL:9%87gpj:(03749.agp_COL)_remat

یادگیری تقویتی

نظریه کنترل

کارهای بلمن و هوارد

یادگیری تقویتی

یادگیری باناظر

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

یادگیری بی‌ناظر

- یافتن شباهت‌ها و تفاوت‌ها بین نمونه داده‌ها
- ی‌ت یافتن مدل کنش بیشینه‌ساز 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

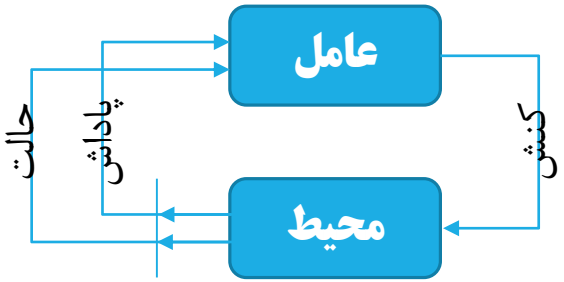
تقریب ذهن
- دستم بگرفت و پا به پا برد.

- 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
- موازنهٔ پایه چرخ
- دنیای شطرنجی
- قارچ‌خور

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
  - زمان‌پیوسته
  - 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 کنندهٔ رفتار عامل ⟸ 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 حریصانهٔ کنش را حریصانه 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

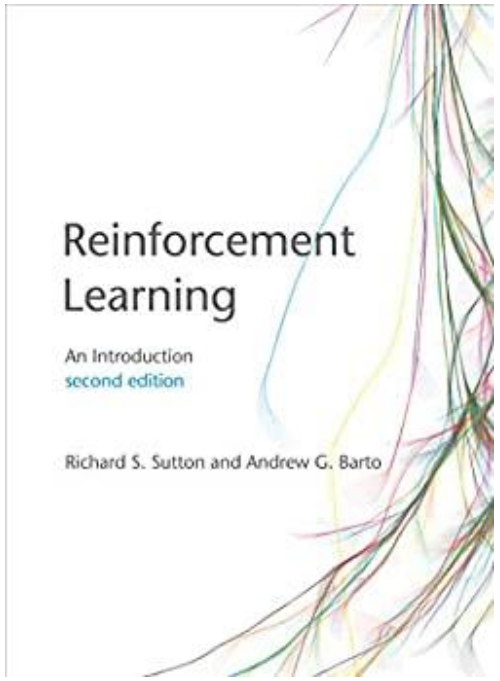
پوترمان

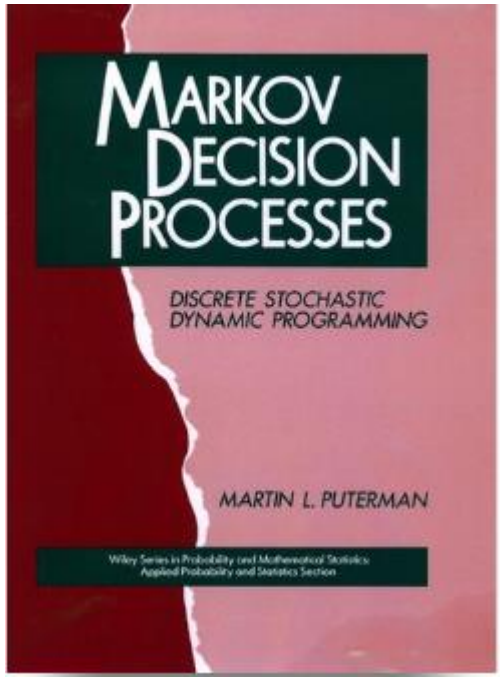
سواک

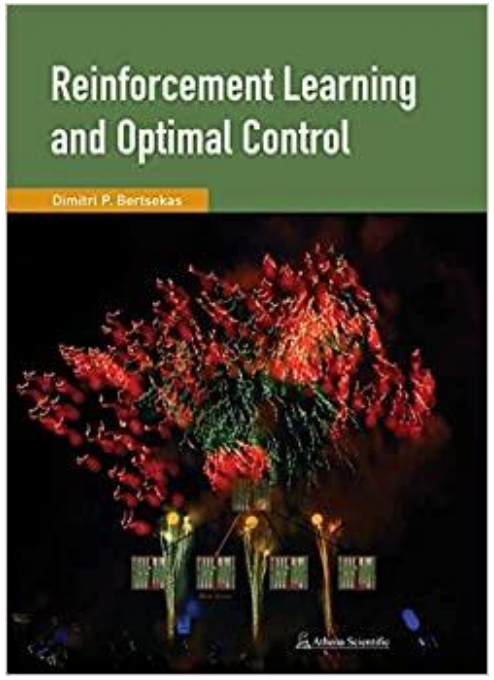
بلمن

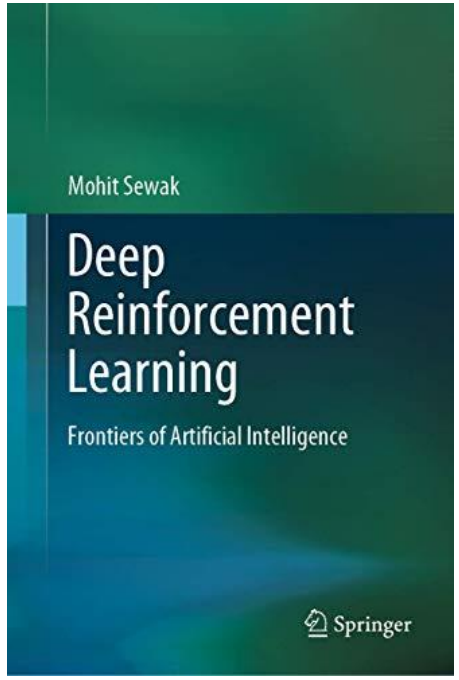
مقالات

درس‌های مرتبط

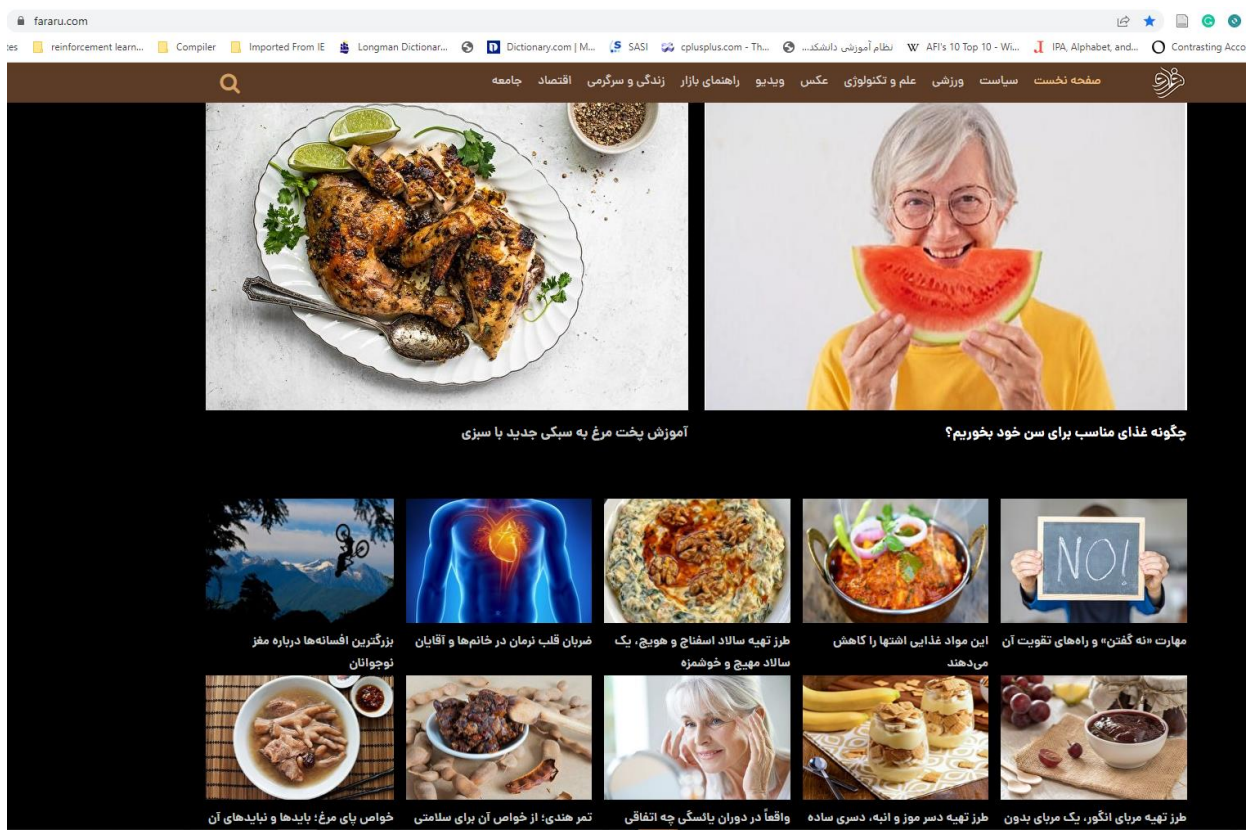








ارزیابی



رونویسی(کپی–پیست) و خلق محتوا و دانش

ارزیابی

- امتحان
- تحقیق
  - ارائه و محتملا پیاده‌سازی
- تمرین
  - قلم و کاغذ
  - پیاده‌سازی

حل‌تمرین

- رضا شامی تنها

- نشانی
  - https://iasbs.ac.ir/~mohsen.hooshmand/courses/0102/1/Yadgiri-Taqviati.html

فرایندهای چندگامی

Multistage Process

فرایندهای متوالی

مطالعه سیستم‌ها و چگونگی عملکرد آنها در طول زمان

درک سیستم به مثابهٔ بردار حالت $\boldsymbol{x}(t)$ و قاعدهٔ 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 مثابهٔ بردار حالت $\boldsymbol{x}(t)$ و قاعدهٔ تشخیص آن در طول زمان
$\boldsymbol{x}(t) = [x_1(t), x_2(t), \cdots, x_N(t)]^T$

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 $T(s)$
  - s حالت بعدی
  - s_1 حالت آغاز

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

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

مثال

$$T(s) = rs$$

ماتریس چرخش

فرایندهای چندگامی

پیچیدگی زیاد مخل موضوع ← باید ساده کرد

کاهش داده

فرایند N-گامی

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

- $s_{n+1} = T(s_n), n = 0,1,2,\cdots,N-1$
- استقلال از گذشته جهت تعیین آینده

$$T^N = T^{N-k}\left(T^k\right)$$

- گزارهٔ تحلیلی علیت
- یا متناظرا آینده صرفا با حال مشخص می‌شود
  - قضیه یگانگی

روابط بازگشتی

h تابع دلخواه

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

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

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

$$f_0(s) = h(s)$$

روابط بازگشتی

h تابع دلخواه

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

$$f_N(s) = \left\{ \max_{0 \le i \le N} h(s_i) \right\} = \max\big[h(s), f_{N-1}\big(T(s)\big)\big], N \ge 1$$

$$f_N(s) = \sum_{i=0}^{N} h(s_i, s_{i+1}) = h\big(s, T(s)\big) + f_{N-1}\big(T(s)\big)$$

مثال

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

$$s = 1 + \frac{1}{1 + \frac{1}{1 + \ddots}} \Longrightarrow s = 1 + \frac{1}{s}$$

$$s = 1 + \cfrac{x}{1 + \cfrac{x^2}{1 + \cfrac{x^4}{1 + \cfrac{x^8}{1 + \cdots}}}} \Longrightarrow s(x) = 1 + \frac{1}{s(x^2)}$$

فرایندهای نامتناهی

$N \to \infty$

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

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

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

- عدم حقیقت فیزیکی

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

- دارای رفتار میانگین یا رفتار حالت-پایدار
- امکان یافتن با رابطهٔ مجانبی
- $f_N(s) \sim g_{2(s)}^{N} h(s), N \to \infty$ یا $\frac{f_N(s)}{N} \sim g_1(s) N \to \infty$

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 بی‌حد که لزوما متناهی نیست.

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

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

T تبدیلی تصادفی که بردار تصادفی s_1 با توزیع p

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

بررسی امید ریاضی اصل اساسی فرایندهای تصادفی

فرایندهای تصادفی

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

متغیرهای تصادفی مستقل $s_0 = s, r_n$

به دنبال امید ارزش

$$g(s) + g(s_1) + \cdots + g(s_N)$$
$$f_N(s) = E[g(s) + g(s_1) + \cdots + g(s_N)\,]$$
$$= g(s) + E[g(s_1) + \cdots + g(s_N)\,]$$
$$= g(s) + f_N\big(T(s, r_1)\big)$$
$$f_0(s) = g(s)$$

فرایندهای تصمیم چندگامی

← نظریهٔ برنامه‌ریزی پویا

دارای موقعیت بنیادی در نظریه کنترل جدید و بالتبع یادگیری تقویتی

تصمیم

سیاست

$$[S,T(S)]$$

- $T = T(s,q)$ انتخاب q_i از مجموعهٔ $s(q)$ و a_i انتخاب در گام i–ام

$$s_1 = T(s, q_0)$$
$$s_2 = T(s_1, q_1)$$
$$\vdots$$
$$s_{n+1} = T(s_n, q_n)$$

- q_i بردار تصمیم یا متغیر تصمیم (رابطه با کنش action)
- تصمیم: انتخاب مقدار q_i

فرایندهای تصمیم چندگامی

- q_i بردار تصمیم یا متغیر تصمیم (رابطه با کنش action)
- تصمیم: انتخاب مقدار q_i

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

$$R(s, s_1, s_2, \cdots; q_0, q_1, \cdots)$$

- تابع بازده یا تابع عیار

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

$$[s, s_1, s_2, \cdots, s_N; q_0, q_1, q_2, \cdots, q_N]$$
$$\forall n: s_{n+1} = T(s_n, q_n)$$

سیاست

Policy

نیاز به انتخاب q_i در هر گام

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

$$q_i = q_i(s, s_1, s_2, \cdots, s_i; q_0, q_1, \cdots, q_{i-1})$$

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

سیاست بهینه: سیاست بیشینه‌ساز تابع R

سیاست

$$q_i = q_i(s, s_1, s_2, \cdots, s_i; q_0, q_1, \cdots, q_{i-1})$$

- باز کلی
- تابع صرفا تابعی از حالت فعلی سیستم.
  - $q_i = q_i(s_i)$

$$\pi_i = [s_i, s_{i-1}, \cdots]$$
$$q_i = q_i(\pi_i)$$

اصل بهینگی

عیار جداسازی

اصل بهینگی

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

بیشینه‌سازی تابع بازده

$$R(s, s_1, s_2, \cdots; q_0, q_1, \cdots) = \sum_{i=0}^{N} g(s_i, q_i)$$

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

با استفاده از اصل بهینگی

$$g(s, q_0) + [g(s_1, q_1) + \cdots + g(s_N, q_N)] = g(s, q_0) + f_{N-1}\big(T(s, q_0)\big)$$

$$\Leftarrow$$

$$f_N(s) = \max_{q_0}\big[(s, q_0) + f_{N-1}\big(T(s, q_0)\big)\ \big], N \geq 1$$

بیشینه‌سازی تابع بازده

$$f_N(s) = \max_{q_0}\left[g(s,q_0) + f_{N-1}\left(T(s,q_0)\right)\right], N \geq 1$$

اثبات

$$\max_{[q_0,\cdots,q_n]} R = \max_{q_0} \max_{[q_1,\cdots,q_n]} R$$

$$f_N(s) = \max_{[q_0,\cdots,q_n]}[g(s,q_0) + g(s_1,q_1) + \cdots + g(s_N,q_N)]$$

$$= \max_{q_0} \max_{[q_1,\cdots,q_n]}[g(s,q_0) + g(s_1,q_1) + \cdots + g(s_N,q_N)]$$

$$= \max_{q_0}\left[g(s,q_0) + \max_{[q_1,\cdots,q_n]}[g(s_1,q_1) + \cdots + g(s_N,q_N)]\right]$$

$$= \max_{q_0}[g(s,q_0) + f_{N-1}(T(s,q_0)]$$

منابع

بلمن ۱۹۶۵

ساتن ۲۰۱۸