یادگیری تقویتی
روش‌های مونت کارلو

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

مقدمه

برنامه‌ریزی پویا

- نیاز به داشتن مدل کامل فرایند تصمیم مارکوف
  - $p(s',r|s,a)$

روش مونت کارلو

- تخمین ارزش‌ها و سیاسات بدون نیاز به دینامیک
- یادگیری صرفا بر مبنای تجربه و تعامل با محیط
  - نمونه دنباله‌هایی از حالت‌ها، کنش‌ها، و پاداش‌ها
- $\Longleftarrow$ بدون مدل
  - بدون استفاده از دانش قبلی از دینامیک محیط و همچنان امکان بهینگی
- تجربهٔ حاصل ناشی از تجربه عملی و فیزیکی یا استفاده از شبیه‌سازی
- امکان یافتن p در بعضی موارد
  - ولی دارای محاسبات ریاضی با پیچیدگی سخت فراوان
  - عدم امکان استفاده از برنامه‌ریزی پویا
- در مسائلی هم عدم‌امکان توصیف ریاضی دینامیک
  - مثال؟

مقدمه

روش شبیه‌سازی

- نیازمند مدلی جهت تولید گذار بین نمونه‌ها
  - به دلیل تولیدی ممکن و آسان
- بدون نیاز به توزیع کامل احتمال تمامی انتقال‌های ممکن در ب‌پ
  - به دلیل نشدنی بودن

روش‌های مونت‌کارلو

بهره از نمونه‌برداری تصادفی جهت محاسبات ریاضی

- بر مبنای بازده‌های میانگین نمونه

کاربردها

- شبیه‌سازی سیستم‌ها
- محاسبات عددی

مثال

- محاسبه مساحت
- انتگرال‌گیری

روش‌های مونت‌کارلو

کاربرد در محیط‌های اپیزودی

- پایان‌پذیری هر اپیزود
- پس از هر اپیزود
  - تخمین ارزش حالت‌ها و کنش‌ها با استفاده از پاداش‌ها
- امکان حل افزایشی اپیزود به اپیزودی
  - عدم امکان حل گام به گام و برخط

مونت کارلو

- اساسا به معنای هرگونه روش تخمین با استفاده از مولفه‌های پیچیدهٔ تصادفی
- در اینجا منظور میانگین‌گیری بازده کامل
  - در مقابل روش‌های یادگیری از بازده‌های جزیی و ناتمام

روش‌های مونت‌کارلو

شباهت به روش‌های راهزن در نمونه‌برداری و میانگین‌گیری بازده‌ها برای هر زوج حالت-کنش

- تفاوت
  - تعریف چند حالت و رفتار هر یک مانند مسئلهٔ راهزنی متفاوت
  - درارتباط بودن مسائل راهزن با هم
  - بازده دریافتی کنشی در حالتی بسته به کنش‌های بعدی در حالت‌های بعدی در همان اپیزود
- نامانا شدن مسئله از دید حالت متقدم
  - به دلیل تحت یادگیری گرفته شدن تمامی انتخاب کنش‌ها

روش‌های مونت‌کارلو

جهت حل نامانایی

- استفاده از الگوی «تکرار سیاست عمومی»
  - در ب‌پ محاسبهٔ تابع ارزش از دانش ف‌ت‌م
  - در م‌ک یادگیری تابع ارزش از بازده‌های نمونه با ف‌ت‌م
- تعامل تابع‌های ارزش و سیاسات متناظر جهت دستیابی به بهینگی همانند تکرار سیاست عمومی

پیش‌بینی مونت‌کارلو

شروع با یادگیری تابع ارزش-حالت سیاست داده‌شده

- ارزش حالتی برابر با میانگین بازده (تجمیع وزن‌دار پاداش‌های آینده) با شروع از حالت مذکور

$$s_0, a_0, r_1, s_1, a_1, r_2, \dots, s_T$$

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

$$v_\pi(s) = E_\pi[G_t | S_t = s]$$

$$q_\pi(s, a) = E_\pi[G_t | S_t = s, A_t = a]$$

- تخمین آن با استفاده از تجربه
  - میانگین‌گیری از بازده‌های مشاهده شده پس از ملاقات آن حالت
  - همگرایی میانگین به ارزش موردانتظار با مشاهدهٔ بازده‌های بیشتر
  - پایه روش‌های میم-کاف

پیش‌بینی مونت‌کارلو

فرض: به دنبال تخمین $v_\pi(s)$ ارزش حالت s تحت سیاست π

- با داشتن مجموعه‌ای از اپیزودهای تحت سیاست مذکور حین عبور از s
- ملاقات s- هر مشاهده و رفتن به حالت s
- امکان ملاقات چندبارهٔ حالت s
  - اولین ملاقات
  - هر ملاقات

تخمین ارزش‌ها حالت‌ها

- محاسبهٔ مقدار بازده‌ها به صورت‌های اولین-ملاقات و هر-ملاقات

تخمین میم-کاف هر ملاقات

- میانگین بازده $v_\pi(s)$ به دنبال اولین ملاقات s

پیش‌بینی مونت‌کارلو

محاسبهٔ بازگشتی ارزش حالت

- هر اپیزود بر اساس مقدار دورهٔ قبل

$$V(S_t) = V(S_t) + \frac{1}{N(S_t)}\big(G_t - V(S_t)\big)$$

الگوریتم اولین-ملاقات- جهت سنجش سیاستی

First-visit MC prediction, for estimating $V \approx v_\pi$

Input: a policy π to be evaluated

Initialize:

$V(s) \in \mathbb{R}$, arbitrarily, for all $s \in \mathcal{S}$

$Returns(s) \leftarrow$ an empty list, for all $s \in \mathcal{S}$

Loop forever (for each episode):

Generate an episode following π: $S_0, A_0, R_1, S_1, A_1, R_2, \ldots, S_{T-1}, A_{T-1}, R_T$

$G \leftarrow 0$

Loop for each step of episode, $t = T-1, T-2, \ldots, 0$:

$G \leftarrow \gamma G + R_{t+1}$

Unless S_t appears in $S_0, S_1, \ldots, S_{t-1}$:

Append G to $Returns(S_t)$

$V(S_t) \leftarrow$ average$(Returns(S_t))$

روش مونت کارلو

تفسیر نموداری

ب‌پ: انتقال‌های تک گامی

م‌ک: نمایش موارد نمونه‌برداری شده در یک اپیزود

- حرکت از ابتدا تا انتهای مسیر

ویژگی م‌ک

- استقلال تخمین هر حالت
- عدم تکیهٔ محاسبهٔ تخمین حالتی بر اساس حالتی دیگر
  - متفاوت از ب‌پ
  - عدم استفاده از «وصال خویشتن»

روش‌های مونت کارلو

تخمین ارزش حالت‌ها در محیط نامانا

$$V(S_t) = V(S_t) + \alpha\big(G_t - V(S_t)\big)$$

مونت‌کارلو اندازه-گام ثابت

[تقریب،قدیم − هدف]اندازه،قدم + تقریب،قدیم = تقریب،جدید

پس از هر اپیزود بروزرسانی ارزش حالت‌ها

روش‌های مونت کارلو- تخمین ارزش کنش

عدم وجود مدل

- مفید واقع شدن استفاده از تخمین ارزش کنش‌ها به جای ارزش حالت
  - ارزش جفت حالت-کنش‌ها

در صورت حضور مدل کفایت تابع ارزش حالت جهت تعیین سیاستی

- نگاه به قدمی فراتر و انتخاب کنش منجر به بیش‌ساز ترکیب پاداش و ارزش حالت بعدی
- بدون مدل عدم‌کفایت تابع‌های ارزش

نیاز به تخمین دقیق و آشکار ارزش هر کنش جهت یافتن ارزش‌های موثر بر پیشنهاد سیاستی

از اهداف بنیادی م‌ک

- تخمین q_*
- فدم اول تخمین سیاست‌سنجیِ ارزش‌های کنش

روش‌های مونت کارلو- تخمین ارزش کنش

سیاست سنجی تابع ارزش کنش

$q_\pi(s,a)$ بازده موردانتظار با

- شروع از s
- انجام کنش a
- پیگیری سیاست π در ادامه

سیاست‌سنجی همانند ارزش حالت

- با تفاوت ارزش حالت-کنش
- ملاقت sو a به معنای ملاقات s در اپیزودی و انجام a
  - اولین-ملاقات
  - هر-ملاقات

روش‌های مونت کارلو- تخمین ارزش کنش

افتاد مشکل‌ها

- امکان عدم ملاقات بسیاری از زوج حالت-کنش‌ها
- در صورت قطعی بودن سیاستی صرفا امکان محاسبهٔ ارزش یک کنش در هر حالت
- «مشکل کاوش»
- جهت عملیاتی سازی سنجش سیاستی برای رازش‌های کنش
  - نیاز به کاوش مداوم
  - مشابه مسئلهٔ راهزن

- راه‌حل
  - کاوش وارد می‌شود exploration starts
  - به دیگر سخن-ضرورت امکان‌دهی انتخاب به تمامی زوج‌های حالت-کنش

صفر نبودن احتمال آغاز با هر زوج کنش-حالتی

- تضمین ملاقات چندبارهٔ هر زوجی

روش‌های مونت کارلو- تخمین ارزش کنش

فواید مترتب بر «کاوش وارد می‌شود» در بعضی اوقات

- عدم امکان تکیه بر آن در حالت عمومی و کلی
  - خاصه هنگام تعامل مستقیم با محیط (بدون شبیه‌سازی)
  - راه‌حل؟
  - فی‌الحال ادامه با کاوش وارد می‌شود.

کنترل مونت کارلو---

استفاده از تخمین مونت‌کارلو در کنترل

- به بیان دیگر در تقریب سیاسات بهینه
- استفاده از الگوهای قبلی

تکرار سیاست‌سنجی و اصلاح سیاست و ادامه تا دستیابی به سیاست بهینه

$$\pi_0 \xrightarrow{\text{سنجش}} q_{\pi_0} \xrightarrow{\text{اصلاح}} \pi_1 \xrightarrow{\text{سنجش}} q_{\pi_1} \xrightarrow{\text{اصلاح}} \pi_2 \xrightarrow{\text{سنجش}} q_{\pi_2} \xrightarrow{\text{اصلاح}} \pi_3 \xrightarrow{\text{سنجش}} \cdots \xrightarrow{\text{اصلاح}} \pi_* \xrightarrow{\text{سنجش}} q_*$$

استفاده سیاست‌سنجی معرفی شده م‌ک

- آزمایش تعداد فراوانی اپیزود
- جهت میل تابع ارزش-کنش به مقدار واقعی تابع
- فرض- مشاهده تعداد نامتناهی اپیزود و تولید اپیزودها به روش کاوش وارد می‌شود.
  - منجر به محاسبهٔ محسابهٔ دقیق q_{π_k}برای هر سیاست دلخواه π_k

کنترل مونت کارلو---

مرحلهٔ بعد

- اصلاح سیاست
- سیاستی حریصانه با توجه به تابع ارش فعلی
- وجود تابع ارزش کنش
- بنابراین بدون نیاز به مدل جهت تولید سیاست حرصی
- سیاست حرصی متناظر هر تابع ارزش-کنش q
  - انتخاب قطعی و معین کنش بیش‌ساز تابع ارزش یا

$$\pi(s) = \underset{a}{\operatorname{argmax}}\, q(s, a)$$

  - برای تمامی حالات

کنترل مونت کارلو---

سیاستی بهتر یا برابر سیاست فعلی

- امکان استفاده از روش‌های م‌ک جهت بهینه‌یابی
- اما بر مبنای دو فرض خلاف‌واقع: بی‌نهایت تکرار و ورود کاوش
- حذف دو فرض مذکورجهت نائل شدن به الگوریتمی عملی
  - فعلا حذف بی‌نهایت تکرار
- امکان اجرای هم‌دوش سیاست‌سنجی و اصلاح سیاست
  - تکرار سیاست عمومی
- استفاده از سیاست جدید در اپیزود

تخمین ارزش-کنش با کاوش وارد می‌شود

Monte Carlo ES (Exploring Starts), for estimating $\pi \approx \pi_*$

Initialize:

$\pi(s) \in \mathcal{A}(s)$ (arbitrarily), for all $s \in \mathcal{S}$

$Q(s,a) \in \mathbb{R}$ (arbitrarily), for all $s \in \mathcal{S}$, $a \in \mathcal{A}(s)$

$Returns(s,a) \leftarrow$ empty list, for all $s \in \mathcal{S}$, $a \in \mathcal{A}(s)$

Loop forever (for each episode):

Choose $S_0 \in \mathcal{S}$, $A_0 \in \mathcal{A}(S_0)$ randomly such that all pairs have probability > 0

Generate an episode from S_0, A_0, following π: $S_0, A_0, R_1, \ldots, S_{T-1}, A_{T-1}, R_T$

$G \leftarrow 0$

Loop for each step of episode, $t = T-1, T-2, \ldots, 0$:

$G \leftarrow \gamma G + R_{t+1}$

Unless the pair S_t, A_t appears in $S_0, A_0, S_1, A_1 \ldots, S_{t-1}, A_{t-1}$:

Append G to $Returns(S_t, A_t)$

$Q(S_t, A_t) \leftarrow \text{average}(Returns(S_t, A_t))$

$\pi(S_t) \leftarrow \arg\max_a Q(S_t, a)$

سیاست‌مدار در مقابل سیاست‌ندار

به دنبال حذف «کاوش وارد می‌شود!»

- دو روش سیاست‌مدار و سیاست‌ندار

سیاست‌مدار

- سنجش و اصلاح سیاست در حال اجرا جهت تصمیم‌گیری
- الگوریتم قبلی نمونه‌ای از سیاست‌مداری

سیاست‌ندار

- سنجش و اصلاح سیاستی متفاوت از سیاست مورد استفاده جهت تولید داده

کنترل م‌ک سیاست‌مدار

بدی کاوش وارد می‌شود

علاوه بر آن

- تولید اپیزود با استفاده از سیاست در حال اصلاح
  - امکان افتادن در چنبرهٔ تسلسل (حلقهٔ بی‌نهایت)

استفاده از روش سیاست‌مدار جهت اجتناب از کاوش وارد می‌شود و چنبرهٔ مذکور

- با سیاست نرمش (هموارساز) soft
  - در ابتدا $\pi(a|s) > 0$ برای تمامی حالات و تمامی کنش‌های متناظر حالات
  - در هر حالت وجود احتمال انتخاب تمامی حالت‌ها
  - تصادفی بودن سیاست در بدایت امر
    - احتمال انتخاب هر کنشی
  - میل تدریجی به سیاست قطعی
- آشنا؟
  - سیاست اپسیلون-حریصانه

کنترل مونت کارلو سیاست‌مدار

- سیاست «حریصانه با اپسیلون» از انواع سیاست نرمش

$$\pi(a|S_t) = \begin{cases} 1 - \epsilon + \frac{\epsilon}{|A(S_t)|}, a = A^* \\ \frac{\epsilon}{|A(S_t)|}, a \neq A^* \end{cases}$$

$$\pi(a|S_t) \geq \frac{\epsilon}{|A(S_t)|}$$

نرمش-اپسیلون

کنترل مونت کارلو سیاست‌مدار

از اهداف آن حذف کاوش در بدایت امر و همگرایی سریعتر به سیاسات بهینه

سیاست جدید بهتر یا برابر سیاست قدیم

کاهش اپسیلون در طول اپیزودها

- محیط مانا یا نامانا؟

On-policy first-visit MC control (for ε-soft policies), estimates $\pi \approx \pi_*$

Algorithm parameter: small $\varepsilon > 0$

Initialize:

$\pi \leftarrow$ an arbitrary ε-soft policy

$Q(s,a) \in \mathbb{R}$ (arbitrarily), for all $s \in \mathcal{S}$, $a \in \mathcal{A}(s)$

$Returns(s,a) \leftarrow$ empty list, for all $s \in \mathcal{S}$, $a \in \mathcal{A}(s)$

Repeat forever (for each episode):

Generate an episode following π: $S_0, A_0, R_1, \ldots, S_{T-1}, A_{T-1}, R_T$

$G \leftarrow 0$

Loop for each step of episode, $t = T-1, T-2, \ldots, 0$:

$G \leftarrow \gamma G + R_{t+1}$

Unless the pair S_t, A_t appears in $S_0, A_0, S_1, A_1 \ldots, S_{t-1}, A_{t-1}$:

Append G to $Returns(S_t, A_t)$

$Q(S_t, A_t) \leftarrow \text{average}(Returns(S_t, A_t))$

$A^* \leftarrow \arg\max_a Q(S_t, a)$ (with ties broken arbitrarily)

For all $a \in \mathcal{A}(S_t)$:

$$\pi(a|S_t) \leftarrow \begin{cases} 1 - \varepsilon + \varepsilon/|\mathcal{A}(S_t)| & \text{if } a = A^* \\ \varepsilon/|\mathcal{A}(S_t)| & \text{if } a \neq A^* \end{cases}$$

کنترل مونت کارلو سیاست‌مدار

شبیه‌سازی

کنترل مونت کارلو سیاست‌ندار

داغ بر جبین یافتن و یادگیری بهینگی

- ؟
- در پی فراگرفتن ارزش‌های کنش بر پایه رفتار بهینه
- نیاز به رفتار غیربهینه جهت کاوش کل کنش‌ها
  - به جهت یافتن کنش بهینه

چگونه؟

- سیاست‌مدار
- سیاست‌ندار

کنترل مونت کارلو سیاست‌ندار

چگونه؟

- سیاست‌مدار
  - یادگیری سیاست –شبه-بهینه در حالی که کاوش می‌کند
- سیاست‌ندار
  - درگیر دو سیاست
    - یکی جهت یادگیری بهینه و منجر به سیاست بهینه
      - سیاست هدف
    - دیگری کاوشگر و جهت تولید رفتار
      - سیاست رفتار

کنترل مونت کارلو سیاست‌ندار

سیاست‌مدار

- ساده‌تر

سیاست‌ندار

- دارای نوسان و تغیُّر زیاد
- همگرایی کندتر
- قدرتمندتر
- عمومی‌تر و پراستفاده‌تر
- امکان استفاده از داده تولیدی فرد متخصص
- سیاست‌مدار حالت خاصی از سیاست‌ندار
  - ؟
  - رفتار و هدف یک تابع

روش مونت کارلو سیاست‌ندار- سنجش

سیاسات هدف و رفتار هر دوثابت

- در پی تخمین v_π یا q_π
- اما دارای اپیزودهای که پیرو سیاست دگر b
  - $b \neq \pi$
  - b سیاست رفتار
  - π سیاست هدف

روش مونت کارلو سیاست‌ندار- سنجش

سیاست رفتاری معادل جستجوی سخت فراوان
- چون ولگردی (گام تصادفی)

امید ریاضی برابر با ارزش سیاست رفتاری

$$E[G_t|S_t = s] = v_b(s)$$

- نه سیاست هدف

تعیین سیاست هدف با وزن‌دهی به بازده‌ها

⇐

$$E[\rho_{t:T-1}G_t|S_t = s] = v_\pi(s)$$

$$V(s) = \frac{\sum_{t\in T(s)} \rho_{t:T(t)-1} G_t}{|T(s)|}$$

- بدون سوگیری
  - معادل V_π
- دارای وردائی بی‌حد

$$E\left[\left(\frac{\prod_{k=t}^{T-1}\pi(A_k|S_k)}{\prod_{k=t}^{T-1}b(A_k|S_k)}\right)^2\right]$$

روش مونت کارلو سیاست‌ندار- سنجش

به دنبال استفاده از اپیزودهای b جهت تخمین π

- نیاز به اینکه هر کنش تحت هدف حتما تحت پوشش سیاست رفتار نیز باشد یا
- فرض پوشش یا

$$\pi(a|s) > 0 \Longrightarrow b(a|s)$$

اتفاقی بودن رفتار در مقابل قطعی بودن هدف

نمونه‌برداری اهمیت

- $E(X)$

روش‌های میم‌کاف—خلاصه

یادگیری تجربی تابع‌های ارزش و سیاسات بهینه با «نمونه اپیزودها»

دارای مزایایی نسبت به ب‌پ

- یادگیری رفتار بهینه از از تعامل با محیط بدون داشتن مدلی از دینامیک محیط
- امکان کاربرد آنها با شبیه‌سازی یا مدل‌های نمونه
  - راحت بودن نمونه‌برداری از بسیاری از مسایل با وجود پیچیدگی ایجاد مدل آنها
- امکان تمرکز بر زیرمجموعه‌ای از حالت‌ها (فصل هشت)
- آسیب کمتر در صورت تعدی از خاصیت مارکوفی
  - به دلیل بروز نکردن مقادیر بر اساس مقادیر جانشینانش
  - دوری از وصال خویشتن

استفاده از الگوی تکرار سیاست تعمیم‌یافته

- تعامل اندرهم سیاست‌سنجی و اصلاح سیاست
- م‌ک به طریق دیگر
  - تخمین ارزش حالتی با میانگین‌گیری بازده‌های آن حالت با شروع از آن
  - خاصه در کنترل بدنبال ارزش-کنش
  - اندرهمی سنجش و اصلاح اپیزود به اپیزود

روش‌های میم‌کاف—خلاصه

اشکال کاوش
- عدم کفایت استفاده از بهترین کنش
- بهره از «کاوش وارد می‌شود!»
- ممکن در شبیه‌سازی ولی ناممکن در تجارب واقعی

سیاست‌مدار
- وفادار به جستجو و تلاش جهت یافتن بهترین سیاست تحت بررسی

سیاست‌ندار
- عامل در حال جستجو و یافتن سیاست بهینهٔ قطعی جدا از سیاست در حال عمل
  - سیاست هدف در مقابل سیاست رفتاری

پیش‌بینی سیاست‌ندار
- یادگیری تابع ارزش سیاست هدف از دادهٔ تولیدی از سیاست رفتاری دیگر
- بر مبنای نمونه‌برداری اهمیت
  - نمونه‌برداری معمولی
  - محتملا دارای وردائی نامتناهی
  - نمونه‌برداری وزن‌دار
  - وردایی متناهی

توجه به تفاوت کنترل و پیش‌بینی

روش‌های میم‌کاف—خلاصه

تفاوت م‌ک از ب‌پ

- عملیاتی بر اساس تجربهٔ ساده و یادگیری مستقیم بدون داشتن مدل
- عدم استفاده از وصال خویشتن
  - بروز نشدن تخمین ارزش حالت‌ها بر اساس تخمین دیگر ارزش‌ها

عدم وابستگی و وابستگی این دو تفاوت به یکدیگر

- امکان جدا در نظر گرفتن این دو تفاوت از یکدیگر
- موضوع بعدی
  - روش‌های یادگیرنده از تجربه (شبیه م‌ک) و همزمان بهره‌گیرنده از وصال خویشتن (همچو ب‌پ)

منابع

ساتن

زندی