یادگیری تقویتی
تفاضل زمانی

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

مقدمه

تفاضلات زمانی یا موقتی

- استفاده از خطای بین دو سیگنال جهت یادگیری تقویتی
- اهم الگوریتم‌های ی‌ت
- ترکیبی از برنامه‌ریزی پویا و مونت‌کارلو
  - نحوهٔ یادگیری محدود به تعامل با محیط و بدون نیاز به مدلی از دینامیک محیط
    - همانند مونت‌کارلو
  - یادگیری از روی تخمین‌های یادگرفته شدهٔ سایر حالت‌ها و عدم‌انتظار رسیدن به نتیجه نهایی
    - مانند برنامه‌ریزی پویا
    - وصال خویشتن

تفصیل موضوع در دو بخش

- سیاست‌سنجی
  - تخمین v_π برای سیاست حاضر π
- اصلاح سیاست
  - اصلاح تا یافتن سیاست بهینه
  - ب‌پ، م‌ک، و ت‌ز از نوع تکرار سیاست عمومی
  - $\Longleftarrow$ پس تفاوت در نحوه سیاست‌سنجی

سیاست‌سنجی

استفادهٔ م‌ک و ت‌ز از تجربه جهت سنجش

کسب تخمین V

- با پیگیری سیاست π و کسب تجربه از طریق اجرای آن

روش مونت‌کارلو

- انتظار تا مشخص شدن بازده ملاقات حالت
- استفاده از بازده به عنوان هدف $V(S_t)$
- روش م‌ک هم ملاقات در محیط‌های نامانا
- با هدف متفاوت عدم لزوم انتظار تا رسیدن به پایان اپیزود

$$V(S_t) = V(S_t) + \alpha[G_t - V(S_t)]$$

G_t بازده واقعی پس از زمان t

روش م‌ک ضریب ثابت

انتظار م‌ک تا پایان اپیزود

سیاست‌سنجی

انتظار م‌ک تا پایان اپیزود

انتظار ت‌ز تا گام زمانی بعدی

ایجاد هدف و بروز کردن $V(S_{t+1})$ بر اساس پاداش دریافتی R_{t+1} در زمان $t+1$

ساده‌ترین ت‌ز

$$V(S_t) = V(S_t) + \alpha[R_{t+1} + \gamma V(S_{t+1}) - V(S_t)]$$

$$V(S_t) = \underbrace{V(S_t)}_{\text{ارزش فعلی}} + \alpha \underbrace{[R_{t+1} + \gamma V(S_{t+1}) - V(S_t)]}_{\text{خطا}}$$

مقایسه

- هدف م‌ک G_t
- هدف ت‌ز $R_{t+1} + \gamma V(S_{t+1})$
  - مشهور به ت‌ز(0) یا ت‌ز تک-گام
  - نوعی خاص از ت‌ز(λ) و ت‌ز چند-گام

سیاست‌سنجی -تفاضلات زمانی(صفر)

- الگوریتم تفاضل زمانی صفر سیاست‌سنج

Tabular TD(0) for estimating v_π

Input: the policy π to be evaluated
Algorithm parameter: step size $\alpha \in (0, 1]$
Initialize $V(s)$, for all $s \in \mathcal{S}^+$, arbitrarily except that $V(\mathit{terminal}) = 0$

Loop for each episode:
  Initialize S
  Loop for each step of episode:
    $A \leftarrow$ action given by π for S
    Take action A, observe R, S'
    $V(S) \leftarrow V(S) + \alpha\big[R + \gamma V(S') - V(S)\big]$
    $S \leftarrow S'$
  until S is terminal

تفاضلات زمانی

بروزرسانی ارزش‌ها پس از اعمال حالتی به محیط و دریافت پاداش موردانتظار

- بروزرسانی بر اساس تخمین موجود
- ⟸ روش وصال خویشتن

معادلات بلمن تابع ارزش

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

تفاضلات زمانی

بروزرسانی ارزش‌ها پس از اعمال حالتی به محیط و دریافت پاداش موردانتظار

- بروزرسانی بر اساس تخمین موجود
- ⟸ روش وصال خویشتن

معادلات بلمن تابع ارزش

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

⟸ روش م‌ک

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

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

⟸ روش ب‌پ

م‌ک

- از نوع تخمین
- دلیل؟ ارزش موردانتظار معلوم نیست، بازده نمونه به جایش در نظر گرفته می‌شود

ب‌پ

- از نوع تخمین
- دلیل؟ $v_\pi(S_{t+1})$معلوم نیست، استفاده از $V(S_{t+1})$ به جای آن

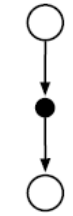
ت‌ز

- از نوع تخمین
- دلیل؟ هر دو دلیل م‌ک و ب‌پ
  - $v_\pi(S_{t+1})$معلوم نیست، استفاده از $V(S_{t+1})$ به جای آن

تفاضلات زمانی- قیاسی با والدان؟!

ت‌ز ترکیب نمونه‌برداری م‌ک با وصال خویشتن ب‌پ

- امکان رسیدن به مزایای هر دو



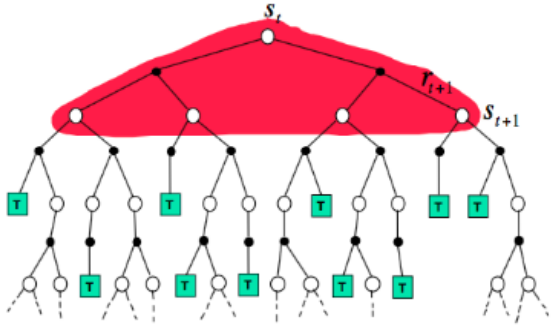
TD(0)

بروزرسانی

- بروزرسانی نمونه (م‌ک و ت‌ز) صرفا تک نمونه
- بروزرسانی امید مورد انتظار (ب‌پ) توزیع کامل تمامی جانشین‌ها

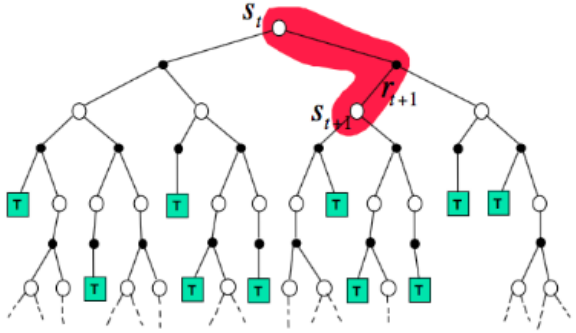
برنامه‌ریزی پویا

$V(S_t) \leftarrow \mathbb{E}_\pi [R_{t+1} + \gamma V(S_{t+1})]$



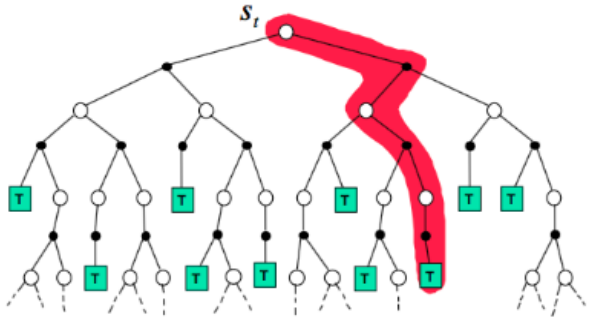
تفاضل زمانی

$V(S_t) \leftarrow V(S_t) + \alpha (R_{t+1} + \gamma V(S_{t+1}) - V(S_t))$



مونت کارلو

$V(S_t) \leftarrow V(S_t) + \alpha (G_t - V(S_t))$



سیلور

تفاضلات زمانی

$$V(S_t) = \underbrace{V(S_t)}_{\text{ارزش فعلی}} + \alpha \underbrace{[R_{t+1} + \gamma V(S_{t+1}) - V(S_t)]}_{\text{خطا}}$$

خطای مورد استفاده برای بروزرسانی ارزش‌ها

- $s \rightarrow s'$
  - $r + \gamma v_\pi(s') - v_\pi(s)$ معروف به خطای تز
  - $v_\pi(s)$ ارزش فعلی
  - $v_\pi(s')$ هدف
  - بروزرسانی با استفاده از میزان خطای بالا

تفاضلات زمانی

خطای تز

$$\delta_t = R_{t+1} + \gamma V(S_{t+1}) - V(S_t)$$

امکان نوشتن «تقریبی» خطای مک بر اساس جمع خطاهای تز

$G_t - V(S_t)$

$$\begin{aligned}
G_t - V(S_t) &= R_{t+1} + \gamma G_{t+1} - V(S_t) + \gamma V(S_{t+1}) - \gamma V(S_{t+1}) \\
&= \delta_t + \gamma\big(G_{t+1} - V(S_{t+1})\big) \\
&= \delta_t + \gamma\delta_{t+1} + \gamma^2\big(G_{t+2} - V(S_{t+2})\big) \\
&= \delta_t + \gamma\delta_{t+1} + \gamma^2\delta_{t+2} + \cdots + \gamma^{T-t-1}\delta_{T-1} + \gamma^{T-t}\big(G_T - V(S_T)\big) \\
&= \delta_t + \gamma\delta_{t+1} + \gamma^2\delta_{t+2} + \cdots + \gamma^{T-t-1}\delta_{T-1} + \gamma^{T-t}\big(0 - 0\big) \\
&= \sum_{k=t}^{T-1} \gamma^{k-t}\delta_k.
\end{aligned}$$

تفاضل زمانی - مزایا

بهبود تخمین با استفاده از بخشی از دیگر تخمین‌ها

فراگرفتن حدسی از حدس دیگر

- وصال خویشتن

بی‌نیاز از مدل محیط

- بر عکس ب‌پ

بی‌نیاز از رسیدن به پایان اپیزود و امکان اجرای برخط

- برعکس مونت کارلو

م‌ک نیاز به انتظار تا پایان اپیزود– ت‌ز نیاز به انتظار در صرفا یک گام

- اساسی برای کاربردهای با اپیزود طولانی

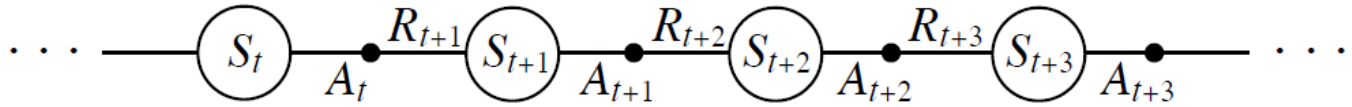
همگرایی به ارزش واقعی با آهنگ یادگیری کوچک

الگوریتم سارسا

تخمین ارزش حالت-کنش بهینه بر اساس تفاضل زمانی سیاست‌مدار

سنجش و اصلاح اندر کنار هم

بررسی انتقال از حالت-کنش اولیه به حالت-کنش بعدی



$$Q(S_t, A_t) = Q(S_t, A_t) + \alpha[R_{t+1} + \gamma Q(S_{t+1}, A_{t+1}) - Q(S_t, A_t)]$$

فرزند وصال خویشتن

- استفاده از خطای موقت به جای امید ریاضی هر حالت-کنش

نیاز به نرمش اپسیلون

- چرا؟

سارسا- اصلاح ت‌ز سیاست‌مدار

یافتن بهینگی

استفاده از الگوی تکرار سیاست عمومی

رقابت کاوش و بهره‌برداری

- دو رده ← سیاست‌مدار و سیاست‌ندار

نیاز به یادگیری تابع ارزش-کنش به حای ارزش-حالت

- روش سیاست‌مدار
- تخمین $q_\pi(s,a)$ رفتار فعلی سیاست π و تمامی حالت‌ها و کنش‌ها

سارسا- اصلاح ت‌ز سیاست‌مدار

سیاست‌سنجی(مرحلهٔ قبل)
- گذر از حالت به حالت
- یادگیری از ارزش حالت‌ها

اصلاح (مرحله کنونی مدنظر)
- انتقال از زوج حالت-کنش به زوج دیگری از حالت-کنش
- یادگیری از ارزش زوج‌های حالت-کنش

هر دو مورد بالا زنجیره‌های مارکوفی با فرایند پاداش‌دهی

$$Q(S_t, A_t) = Q(S_t, A_t) + \alpha[R_{t+1} + \gamma Q(S_{t+1}, A_{t+1}) - Q(S_t, A_t)]$$

اجرای تغییر بالا به ازای هر انتقال در حالت‌های غیرپایانی

اگر S_{t+1} حالت پایانی، آنگاه $Q(S_{t+1},A_{t+1}) = 0$

اعمال قانون بالا بر هر پنج‌تایی $(S_t,A_t,R_{t+1},S_{t+1},A_{t+1})$
- سارسا!

سارسا- اصلاح ت‌ز سیاست‌مدار

امکان و سادگی ایجاد الگوریتم اصلاح سیاست‌مدار سارسا

تخمین پیوستهٔ q_π از سیاست و سپس تغییر πحریصانهٔ با توجه به q_π

امکان استفاده از حریصانه-اپسیلون یا نرمش-اپسیلون جهت حصول به بهینگی

الگوریتم سارسا

Sarsa (on-policy TD control) for estimating $Q \approx q_*$

Algorithm parameters: step size $\alpha \in (0, 1]$, small $\varepsilon > 0$
Initialize $Q(s, a)$, for all $s \in \mathcal{S}^+, a \in \mathcal{A}(s)$, arbitrarily except that $Q(terminal, \cdot) = 0$

Loop for each episode:
  Initialize S
  Choose A from S using policy derived from Q (e.g., ε-greedy)
  Loop for each step of episode:
    Take action A, observe R, S'
    Choose A' from S' using policy derived from Q (e.g., ε-greedy)
    $Q(S, A) \leftarrow Q(S, A) + \alpha\big[R + \gamma Q(S', A') - Q(S, A)\big]$
    $S \leftarrow S'$; $A \leftarrow A'$;
  until S is terminal

یادگیری ک

مبنی بر روش تفاضل زمانی سیاست‌ندار

واتکین ۱۹۸۹

اهم روش‌های بدون‌مدل

$$Q(S_t, A_t) = Q(S_t, A_t) + \alpha \left[R_{t+1} + \gamma \max_a Q(S_{t+1}, a) - Q(S_t, A_t) \right]$$

تخمین مستقیم q_* مستقل از سیاست فعلی

- ساده‌سازی تحلیل الگوریتم و
- امکان‌دهی به اثبات همگرایی زودهنگام

بررسی انتقال از حالت-کنش اولیه به حالت-کنش بعدی

یادگیری ک

الگوریتم

Q-learning (off-policy TD control) for estimating $\pi \approx \pi_*$

Algorithm parameters: step size $\alpha \in (0, 1]$, small $\varepsilon > 0$
Initialize $Q(s, a)$, for all $s \in \mathcal{S}^+, a \in \mathcal{A}(s)$, arbitrarily except that $Q(terminal, \cdot) = 0$

Loop for each episode:
  Initialize S
  Loop for each step of episode:
    Choose A from S using policy derived from Q (e.g., ε-greedy)
    Take action A, observe R, S'
    $Q(S, A) \leftarrow Q(S, A) + \alpha\big[R + \gamma \max_a Q(S', a) - Q(S, A)\big]$
    $S \leftarrow S'$
  until S is terminal

یادگیری ک

تخمین ارزش بهینه مستقل از سیاست رفتاری

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

حالت انتقال قطعی

$$q_*(s,a) = r + \max_{a'} q_*(s', a')$$

نداشتن مقدار

- پس تبدیل به تابع خطای موقت (البته بدون ستاره)

$$r + \max_{a'} q_*(s', a') - q_*(s,a)$$

یادگیری ک

سارسا

$$Q(S_t, A_t) = Q(S_t, A_t) + \alpha[R_{t+1} + \gamma Q(S_{t+1}, A_{t+1}) - Q(S_t, A_t)]$$

یادگیری ک

$$Q(S_t, A_t) = Q(S_t, A_t) + \alpha\left[R_{t+1} + \gamma \max_a Q(S_{t+1}, a) - Q(S_t, A_t)\right]$$

تفاوت

- در تابع محاسبه خطا
- یادگیری ک سیاست‌ندار
- سارسا سیاست‌مدار

سیاست رفتاری

- نرمش اپسیلون

سخن کوتاه

یادگیری تفاضلی زمانی و امکان استفاده از آن در مسائل ی‌ت

سارسا
- سیاست‌مدار

یادگیری ک
- سیاست‌ندار

روش‌های دیگر
- یادگیری ک-دوبل
- سارسای مورد انتظار ← سیاست‌ندار
- روش‌های منتقد-بازیگر

پیشبرد دانش و فن

ریشه‌گیری ت‌ز
- از روانشناسی یادگیری حیوانات و هوش مصنوعی
  - خاصه کارهای کلوپف و ساموئل
- تحقیقات هالند دربارهٔ سازگاری بین پیش‌بینی‌های ارزش
  - مدرس در دان‌میش
  - موثر بر بارتو؟ بارتو کیست؟
  - دانشجوی تحصیلات تکمیلی از ۱۹۷۰ تا ۱۹۷۵ در دان‌میش
  - یافته‌های هالند موثیر در تعادا از نظام‌های ت‌ز-محور

منابع

ساتن

زندی