یادگیری تقویتی
برنامه‌ریزی پویا

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

مباحث

برنامه‌ریزی پویا

سیاست‌سنجی

اصلاح سیاست

تکرار سیاست

تکرار ارزش

تکرار سیاست تعمیم‌یافته

برنامه‌ریزی پویا

روش‌های تعیین سیاست بهینه با در دسترس بودن مدل کامل محیط با ف‌ت‌م

عملا در بسیاری از مواقع غیرقابل استفاده

- در دست نبودن مدل کامل
- دارای پیچیدگی محاسباتی بالا

اما

- دارای اهمیت نظری
- مقدمه‌ای برای روش‌های فارغ از مدل
- دیگر روش‌ها به مثابهٔ ب‌پ با رایانش کمتر و بدون فرض کامل بودن مدل

برنامه‌ریزی پویا

فرض ف‌ت‌مم بودن محیط $p(s',r|s,a)$
- متناهی بودن حالات و کنش‌ها

زمینهٔ اصلی ی‌ت و ب‌پ
- استفاده از تابع ارزش جهت سازمان‌دهی جستجوی سیاسات خوب

استفاده از ب‌پ جهت محاسبهٔ تابع ارزش تعریف شده
- با داشتن تابع ارزش بهینه v_* یا q_* منجر به یافتن سیاسات بهینه

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

برنامه‌ریزی پویا

معادله بهینگی بلمن تابع ارزش حالت

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

معادله بهینگی بلمن تابع ارزش کنش

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

استفاده از الگوریتم‌های ب‌پ

- تغییر معادلات بلمن به به قوانین بروزرسانی
  - جهت ایجاد تقریبی مناسب از تابع ارزش مطلوب

سیاست‌سنجی (سیاست‌سنجی) یا پیش‌بینی

تخمین توابع ارزش و سیاست با روش‌های عددی و تکراری

ابتدا محاسبهٔ تابع ارزش-حالت برای سیاست دلخواه π
- معروف به سیاست‌سنجی

یاداوری تابع ارزش-حالت

$$
\begin{aligned}
v_\pi(s) &= E_\pi[G_t|S_t = s] \\
&= E_\pi[R_{t+1} + \gamma G_{t+1}|S_t = s] \\
&= E_\pi[R_{t+1} + \gamma v_\pi(S_{t+1})|S_t = s] \\
&= \sum_a \pi(a|s) \sum_{s'} \sum_r p(s', r|s, a)[r + \gamma v_\pi(s')] \,, \forall s \in S
\end{aligned}
$$

$\pi(a|s)$ احتمال انتخاب کنش a تحت سیاست π

وجود اندیس π در E_π نمایشگر مشروط بودن امید به سیاست مذکور

باز و حل تحلیلی پیچیده

شرط وجود یکتا v_π
- یا وزن کاهنده کوچکتر از یک
- یا قطعیت رسیدن به حالت نهایی باز تمامی حالات

ارزیابی سیاست (سیاست‌سنجی) یا پیش‌بینی

در صورت معلوم بودن دینامیک محیط

- به تعداد حالت دستگاه معادلات خطی با همان تعداد مجهول
- راه‌حل سرراست اگرچه ملال‌آور و طولانی

روش‌های تکرارمحور مناسب جهت حل چنین مسئله‌ای

انتخاب تصادفی مقدار اولیه v_0

- استثنا- حالت نهایی با مقدار ارزش برابر صفر
- تخمین‌های بعدی با استفاده از قاعده بروز کردن بلمن یا

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

$$= \sum_a \pi(a|s) \sum_{s'} \sum_r p(s', r|s, a)[r + \gamma v_k(s')], \forall s \in S, k = 1, 2, 3, \cdots$$

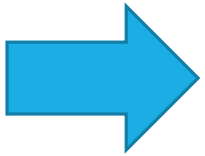
- ضمانت رسیدن به مقدار v_π با $k \to \infty$
  - معروف به سیاست‌سنجی متداوم (یا مکرر)
  - نقطه ثابت؟

سیاست‌سنجی

معادله بلمن برای تابع ارزش دارای صورت ژاکوبی

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

صورت ژاکوبی معادله بلمن



$$v_{k+1}(s) = \sum_a \pi(a|s) \sum_{s'} \sum_r p(s', r|s, a)[r + \gamma v_k(s')], k = 1, 2, 3, \cdots$$

سیاست‌سنجی

روش ژاکوبی

- از روش‌های معمول حل عددی دستگاه معادلات خطی
- مراحل
  - تعیین متغیر i-ام در معادلهٔ i-ام بر حسب سایر مجهولات
  - محاسبهٔ تقریب با استفاده از معادلات مرحلهٔ قبل
  - تکرار مرحله قبل تا همگرائی

سیاست‌سنجی

جهت یافتن تخمین جدید. اعمال عملیات یکسان بر هر حالت

- جانشینی مقادیر قدیمی با مقدار جدید (خود حاصل اعمال بر مقادیر قدیمی) و پاداش بلافصل
- معروف به بروزکردن امید

روش‌های متفاوت جهت انجام چنین عملی

- دلیل امید یا مورد انتظار خوانده شدن.
  - تمامی بروزها بر اساس تمامی حالت‌های قبلی و نه صرفا یکی از آنها

سیاست‌سنجی

- برنامه رایانه ترتیبی جهت پیاده‌سازی سیاست‌سنجی مکرر
  - نیاز به دو آرایه
    - مقادیر قدیم و مقادیر جدید
      - بروز کردن و محاسبه مقادیر جدید بدون تغییر مقادیر قدیم
      - فارغ از ترتیب
  - امکان تک‌ارایه‌ای
    - جانشینی جابجای مقدار قدیم با مقدار جدید
      - در نتیجه تاثیرگیری مقادیر جدید از مقادیر جدید و قبلی
      - تاثیر ترتیب در این نوع حل
- هر دو همگرا به مقدار v_π
  - همگرایی سریعتر تک آرایه‌ای
  - ترتیب بروز شدن سخت موثر در روش‌تک آرایه‌ای

الگوریتم سیاست‌سنجی (سیاست‌سنجی) جهت تعیین تابع ارزش حالت‌ها

الگوریتم سیاست‌سنجی درجای

- توجه به چگونگی شرط پایان
- امتحان $\max_{s\in S}|v_{k+1}(s)-v_k(s)|$ در هر جاروب‌کردن و توقف در صورت کوچکتر بودن از آستانه‌ای

Iterative Policy Evaluation, for estimating $V \approx v_\pi$

Input π, the policy to be evaluated
Algorithm parameter: a small threshold $\theta > 0$ determining accuracy of estimation
Initialize $V(s)$, for all $s \in \mathcal{S}^+$, arbitrarily except that $V(terminal) = 0$

Loop:
  $\Delta \leftarrow 0$
  Loop for each $s \in \mathcal{S}$:
    $v \leftarrow V(s)$
    $V(s) \leftarrow \sum_a \pi(a|s) \sum_{s',r} p(s', r|s, a)\left[r + \gamma V(s')\right]$
    $\Delta \leftarrow \max(\Delta, |v - V(s)|)$
until $\Delta < \theta$

سیاست‌سنجی- مثال

شبیه‌سازی محیط و سنجش سیاست تصادفی

برخورد با دیواره یا موانع

- باقی ماندن در حالت فعلی
- پاداش ۱۰−

استقرار در حالت نهایی

- پاداش ۱۰+

هر کنش

- پاداش ۱−

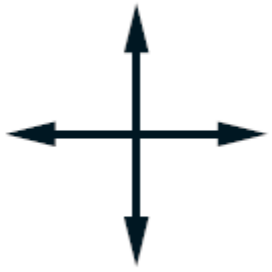
سیاست‌سنجی- تمرین ۲

شبیه‌سازی دنیای شطرنجی 4×4

- دو حالت نهایی
- حالت‌های غیرنهایی $S = \{1,2,\cdots,14\}$
- چهار کنش $\{\leftarrow\uparrow\rightarrow\downarrow\}$
  - کنش‌های انتقال قطعی بین حالت‌ها
    - چرا؟
    - $p(7,-1|7,\rightarrow) = 1$
    - $p(6,-1|5,\rightarrow) = 1$
    - $p\left(10,\text{پ}\middle|7,\rightarrow\right) = 0$
  - *فعالیتی اپیزودی و بدون استفاده از وزن*
  - $r(s,a,s') = -1 \forall s, \forall s', \forall a$
  - سیاست- احتمال کنش‌های تصادفی و با هم برابر
- تمرین

$R_t = -1$ on all transitions

	1	2	3
4	5	6	7
8	9	10	11
12	13	14	



actions

سیاست‌سنجی

روش تکراری ژاکوبی

$$x_1^{n+1} = f(x_1^n, x_2^n)$$
$$x_{12}^{n+1} = g(x_1^n, x_2^n)$$

روش تکراری گاوس-سایدل

$$x_1^{n+1} = f(x_1^n, x_2^n)$$
$$x_{12}^{n+1} = g(x_1^{n+1}, x_2^n)$$

سیاست‌سنجی

سیاست تصادفی

- احتمال اجرای همه کنش‌ها در هر حالت

$$v_{k+1}(s) = \sum_{a} \pi(a|s) \sum_{s'} \sum_{r} p(s', r|s, a)[r + \gamma v_k(s')], k = 1, 2, 3, \cdots$$

سیاست قطعی

- اجرای کنش معین در هر حالت

$$v_{k+1}(s) = \sum_{s'} \sum_{r} p(s', r|s, \pi(s))[r + \gamma v_k(s')], k = 1, 2, 3, \cdots$$

کشتی‌بانی دگر--بهبود سیاست (اصلاح سیاست)

هدف از محاسبهٔ تابع ارزش سیاستی

- در راستای یافتن سیاساتی بهتر
- امکان تولید سیاست بهتر از هر سیاستی
- با داشتن تابع ارزش v_π سیاستی قطعی
  - در حالت s بدنبال روشن کردن تغییر سیاست قطعی به انتخابی قطعی عملی $a \neq \pi(s)$
    - اطلاع از میزان مناسب بودن $v_\pi(s)$
    - حال تغییر سیاست مشخص‌گر کشتی‌بانی دگر یا کنونی

بهبود سیاست (اصلاح سیاست)

- پاسخ به پرسش
  - انتخاب a در حالت s و استفاده از سیاست موجود π
  - یا تابع ارزش حالت-کنش سیاست اولیه

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

- بررسی و مقایسه آن با $v_\pi(s)$
  - در صورت بهتر بودن انتخاب a در حالت s و سپس پیروی از π نسبت به سرسپردن به π در هر حالی
    - سیاستی بهتر
    - موردی خاص از «قضیهٔ اصلاح سیاست»

کشتی‌بانی دگر--بهبود سیاست (اصلاح سیاست)

- سیاست بهتر

«قضیهٔ اصلاح سیاست»

- منجر به ارزش بیشتر
- فرض π و π' دو سیاست قطعی و برای تمامی حالت‌ها
- اگر

$$\forall s: q_\pi\big(s, \pi'(s)\big) \geq v_\pi(s)$$

- آن‌گاه π' سیاستی بهتر یا برابر با سیاست π
- به دیگر سخن دستیابی به امیدی برابر یا بزرگتر در تمامی حالت $\Longrightarrow v_{\pi'}(s) \geq v_\pi(s)$

سخن کوتاه

$$\forall s: q_\pi\big(s, \pi'(s)\big) \geq v_\pi(s) \Longrightarrow v_{\pi'}(s) \geq v_\pi(s)$$

- تعیین ارزش‌ها نسبت به سیاست اولیه $\Longleftarrow$ امکان اصلاح سیاست

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

$$\forall s: q_\pi\left(s, \pi'(s)\right) \geq v_\pi(s)$$

کشتی‌بانی دگر--بهبود سیاست (اصلاح سیاست)

اثبات

$$\begin{aligned}
v_\pi(s) &\leq q_\pi(s, \pi'(s)) \\
&= \mathbb{E}[R_{t+1} + \gamma v_\pi(S_{t+1}) \mid S_t = s, A_t = \pi'(s)] \\
&= \mathbb{E}_{\pi'}[R_{t+1} + \gamma v_\pi(S_{t+1}) \mid S_t = s] \\
&\leq \mathbb{E}_{\pi'}[R_{t+1} + \gamma q_\pi(S_{t+1}, \pi'(S_{t+1})) \mid S_t = s] \\
&= \mathbb{E}_{\pi'}[R_{t+1} + \gamma \mathbb{E}_{\pi'}[R_{t+2} + \gamma v_\pi(S_{t+2}) | S_{t+1}, A_{t+1} = \pi'(S_{t+1})] \mid S_t = s] \\
&= \mathbb{E}_{\pi'}\left[R_{t+1} + \gamma R_{t+2} + \gamma^2 v_\pi(S_{t+2}) \mid S_t = s\right] \\
&\leq \mathbb{E}_{\pi'}\left[R_{t+1} + \gamma R_{t+2} + \gamma^2 R_{t+3} + \gamma^3 v_\pi(S_{t+3}) \mid S_t = s\right] \\
&\vdots \\
&\leq \mathbb{E}_{\pi'}\left[R_{t+1} + \gamma R_{t+2} + \gamma^2 R_{t+3} + \gamma^3 R_{t+4} + \cdots \mid S_t = s\right] \\
&= v_{\pi'}(s).
\end{aligned}$$

کشتی‌بانی دگر--بهبود سیاست (اصلاح سیاست)

سخن کوتاه

$$\forall s: q_\pi\big(s, \pi'(s)\big) \geq v_\pi(s) \Longrightarrow v_{\pi'}(s) \geq v_\pi(s)$$

- تعیین ارزش‌ها نسبت به سیاست اولیه ⟸ امکان اصلاح سیاست

توضیحات قبلی بر مبنای قضیهٔ اصلاح

- سیاست فعلی π و سیاست تغییریافتهٔ π'
- π' عینا همانند π با صرفا یک تفاوت $\pi'(s) = A \neq \pi(s)$
  - در صورت $q_\pi(s,a) > v_\pi(s)$
  - آنگاه سیاست متاخر بهتر از هم‌پیالهٔ متقدم

بهبود سیاست (اصلاح سیاست)

تاکنون صرفا درنظرگرفتن تک حالت

- ادامه طبیعی مسیر؟
  - تمامی حالت‌ها و تمامی کنش‌ها
  - انتخاب کنشی که بیشینه‌ساز $q_\pi(s,a)$ در هر حالتی
  - به بیان دیگر سیاست حریصانهٔ π'

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

  - سیاست حریصانه
  - انتخاب کنشی که در کوتاه‌مدت بهترین به نظر می‌رسد پس از یک قدم پیشرو نگاه کردن با v_π
  - در تطابق با قضیهٔ معرفی شده
- فرایند تولید سیاستی جدید با اصلاح سیاست موجود و نگاه حریصانه به تابع ارزش سیاست موجود
  - معروف به اصلاح یا بهبود سیاست

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

بهبود سیاست (اصلاح سیاست)

انتخاب حریصانهٔ سیاست جدید نسبت به تابع ارزش حالت-عمل

- در صورت سیاست حریصانهٔ جدید π' مناسب ولی بهتر نبودن از سیاست قدیم
- $v_\pi = v_{\pi'} \Leftarrow$

در صورت دستیابی به بهینگی

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

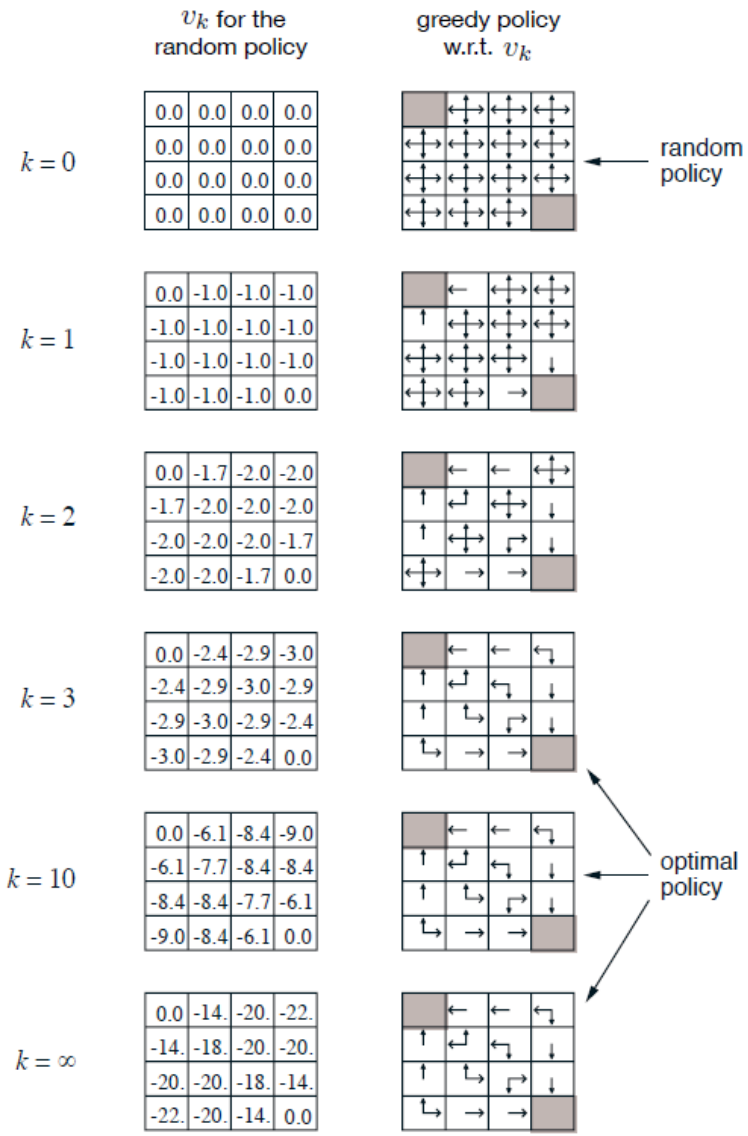
$\Leftarrow$ معادلهٔ بهینگی بلمن و در نتیجه $v_{\pi'} = v_*$

سخن کوتاه، اصلاح سیاست همواره برگرداندن سیاست بهتر مگر در حالت همگرا شدن به سیاست بهینه

بهبود سیاست (اصلاح سیاست)

بررسی سیاست قطعی تاکنون

- حالت عمومی احتمالاتی $\pi(a|s)$
- سیاسات اتفاقی نیز دارای تحلیلی مشابه



تکرار سیاست

با داشتن سیاست فعلی π

- اصلاح و بهبود آن با v_π
- منجر به سیاست بهتر π'
- سپس محاسبهٔ $v_{\pi'}$
- منجر به سیاست بهتر π''
- ادامه مسیر به همین منوال جهت بهبود سیاست و تابع ارزش یا
- تکرار سیاست‌سنجی و اصلاح سیاست و ادامه تا دستیابی به سیاست بهینه

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

- ضمانت بهتر بودن سیاست جدید نسبت به سیاست پیشین
  - تا رسیدن به سیاست بهینه
- در فت‌م‌م دارای حالت‌ها و سیاسات محدود $\Longleftarrow$ همگرایی به سیاست بهینه
- معروف به تکرار سیاست

الگوریتم تکرار سیاست

Policy Iteration (using iterative policy evaluation) for estimating $\pi \approx \pi_*$

1. Initialization
   $V(s) \in \mathbb{R}$ and $\pi(s) \in \mathcal{A}(s)$ arbitrarily for all $s \in \mathcal{S}$

2. Policy Evaluation
   Loop:
   $\Delta \leftarrow 0$
   Loop for each $s \in \mathcal{S}$:
   $v \leftarrow V(s)$
   $V(s) \leftarrow \sum_{s',r} p(s', r|s, \pi(s))\left[r + \gamma V(s')\right]$
   $\Delta \leftarrow \max(\Delta, |v - V(s)|)$
   until $\Delta < \theta$ (a small positive number determining the accuracy of estimation)

3. Policy Improvement
   policy-stable $\leftarrow$ *true*
   For each $s \in \mathcal{S}$:
   old-action $\leftarrow \pi(s)$
   $\pi(s) \leftarrow \arg\max_a \sum_{s',r} p(s', r|s, a)\left[r + \gamma V(s')\right]$
   If *old-action* $\neq \pi(s)$, then *policy-stable* $\leftarrow$ *false*
   If *policy-stable*, then stop and return $V \approx v_*$ and $\pi \approx \pi_*$; else go to 2

تکرار سیاست

- توجه به سیاست‌سنجی خود رایانشی متوالی است
- ایراد شبه‌کد روبرو؟

الگوریتم تکرار سیاست

همگرایی با تعداد کمی تکرار

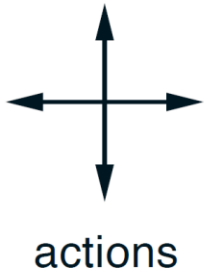
تکرار سیاست

دنیای شطرنجی ۴ × ۴

رسیدن به خروجی‌ها با کمترین تعداد گام

- پیاده‌سازی؟
- در ابتدا سیاست کاملا تصادفی

$$R_t = -1$$
on all transitions



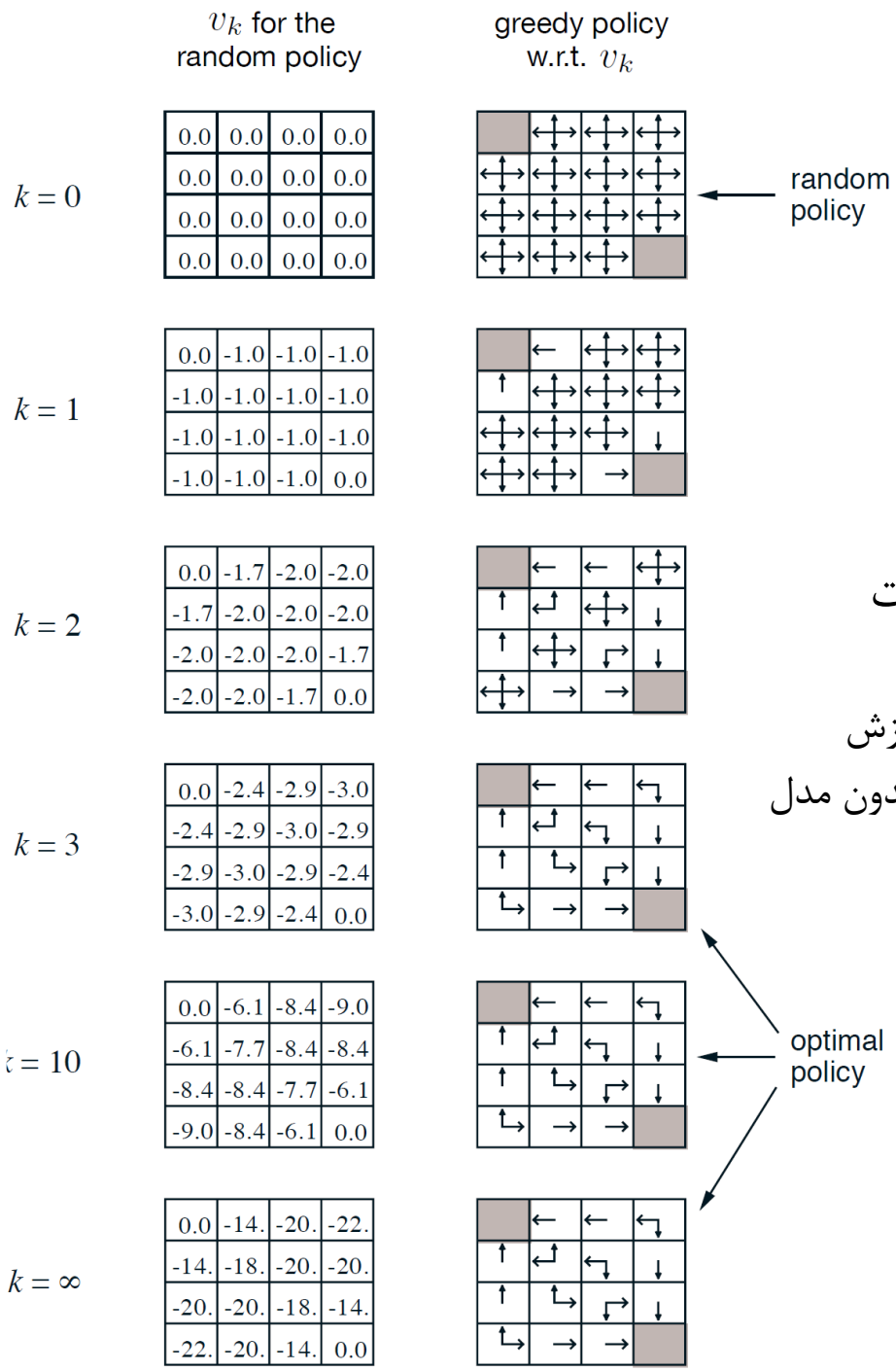
	1	2	3
4	5	6	7
8	9	10	11
12	13	14	

تکرار سیاست

دنیای شطرنجی ۴ × ۴

از مرحله سه به بعد سیاست ثابت

- تغییر ارزش‌ها
- به دنبال سیاست و نه مقدار دقیق ارزش
- امکان‌دهی تقریب ارزش سیاست و بدون مدل
  - در بخش‌های بعدی



تکرار ارزش

ایراد تکرار سیاست

- سیاست‌سنجی در هر گام
- ؟
  - رایانشی تکراری و طولانی
  - سیاست‌سنجی شامل تکرار حلقه‌ای طولانی برای رسیدن به دقت کافی
  - نیاز به جاروب کردن چندباره در مجموعه حالت
  - امکان کنار گذاشتن سیاست‌سنجی

تکرار ارزش

- حل تکراری تابع ارزش معادلات بهینگی بلمن
- عدم نیاز به سیاست اولیه

شامل تک حلقه

- محاسبهٔ ارزش‌های بهینه
- تعیین سیاست بهینه

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

الگوریتم تکرار ارزش

Value Iteration, for estimating $\pi \approx \pi_*$

Algorithm parameter: a small threshold $\theta > 0$ determining accuracy of estimation
Initialize $V(s)$, for all $s \in \mathcal{S}^+$, arbitrarily except that $V(terminal) = 0$

Loop:
| $\Delta \leftarrow 0$
| Loop for each $s \in \mathcal{S}$:
| $v \leftarrow V(s)$
| $V(s) \leftarrow \max_a \sum_{s',r} p(s', r|s, a)\big[r + \gamma V(s')\big]$
| $\Delta \leftarrow \max(\Delta, |v - V(s)|)$
until $\Delta < \theta$

Output a deterministic policy, $\pi \approx \pi_*$, such that
$\pi(s) = \arg\max_a \sum_{s',r} p(s', r|s, a)\big[r + \gamma V(s')\big]$

الگوریتم تکرار ارزش

Policy Iteration (using iterative policy evaluation) for estimating $\pi \approx \pi$

1. Initialization
   $V(s) \in \mathbb{R}$ and $\pi(s) \in \mathcal{A}(s)$ arbitrarily for all $s \in \mathcal{S}$

2. Policy Evaluation
   Loop:
   $\Delta \leftarrow 0$
   Loop for each $s \in \mathcal{S}$:
   $v \leftarrow V(s)$
   $V(s) \leftarrow \sum_{s',r} p(s', r \mid s, \pi(s))\big[r + \gamma V(s')\big]$
   $\Delta \leftarrow \max(\Delta, |v - V(s)|)$
   until $\Delta < \theta$ (a small positive number determining the accuracy of estimation)

3. Policy Improvement
   policy-stable $\leftarrow$ *true*
   For each $s \in \mathcal{S}$:
   old-action $\leftarrow \pi(s)$
   $\pi(s) \leftarrow \arg\max_a \sum_{s',r} p(s', r \mid s, a)\big[r + \gamma V(s')\big]$
   If *old-action* $\neq \pi(s)$, then *policy-stable* $\leftarrow$ *false*
   If *policy-stable*, then stop and return $V \approx v_*$ and $\pi \approx \pi_*$; else go to 2

Value Iteration, for estimating $\pi \approx \pi_*$

Algorithm parameter: a small threshold $\theta > 0$ determining accuracy of estimation
Initialize $V(s)$, for all $s \in \mathcal{S}^+$, arbitrarily except that $V(terminal) = 0$

Loop:
| $\Delta \leftarrow 0$
| Loop for each $s \in \mathcal{S}$:
| $v \leftarrow V(s)$
| $V(s) \leftarrow \max_a \sum_{s',r} p(s', r \mid s, a)\big[r + \gamma V(s')\big]$
| $\Delta \leftarrow \max(\Delta, |v - V(s)|)$
until $\Delta < \theta$

Output a deterministic policy, $\pi \approx \pi_*$, such that
$\pi(s) = \arg\max_a \sum_{s',r} p(s', r \mid s, a)\big[r + \gamma V(s')\big]$

برنامه‌ریزی پویا ناسنکرون

ایرادی اساسی ب‌پ

- عملیات روی کل فضای حالت
- تخته نرد حدود 10^{20} حالت
  - اجرای تکرار ارزش بر یک میلیون حالت در ثانیه
  - هزاران سال تا تکمیل تک جاروب

ب‌پ‌ن

- از نوع ب‌پ تکراری درجا
- بررسی صرفا یک حالت در هر گام
- امکان بروز کردن حالت‌ها بر اساس الگوهایی
  - منجر به بهینه‌تر کردن بروز رسانی
- دوری از بهینه‌کردن حالت‌هایی که روی مسیر بهینه نیستند

برنامه‌ریزی پویا ناسنکرون

ممکن‌سازی محاسبات در تعاملات بلادرنگ

اجرای الگوریتم ب‌پ حین کسب اطلاع از محیط

تکرار سیاست عمومی

تکرار سیاست نیازمند دو فرایند تعاملی و همزمان

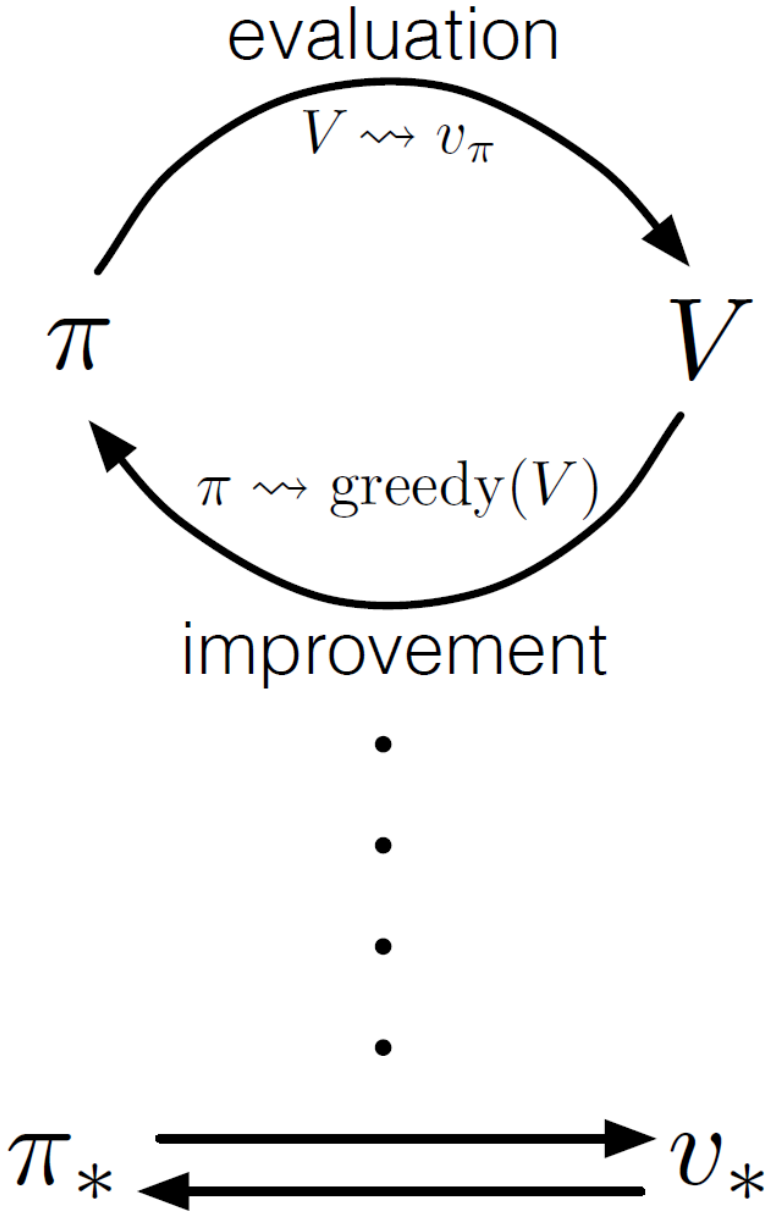
- سازگاری تابع ارزش با سیاست فعلی (سیاست‌سنجی)
- سیاست حریصانه با توجه به تابع ارزش فعلی (اصلاح سیاست)
- تکرار یک به یک موارد بالا . کامل شدن یکی قبل از اجرای دیگری
- عدم لزوم چنین پشت سرهم بودنی
- بهتکرار حلقه سنجش تا همگرایی توابع ارزش حالت

بی‌تغییری سیاست در طول سیاست‌سنجی

تکرار سیاست عمومی

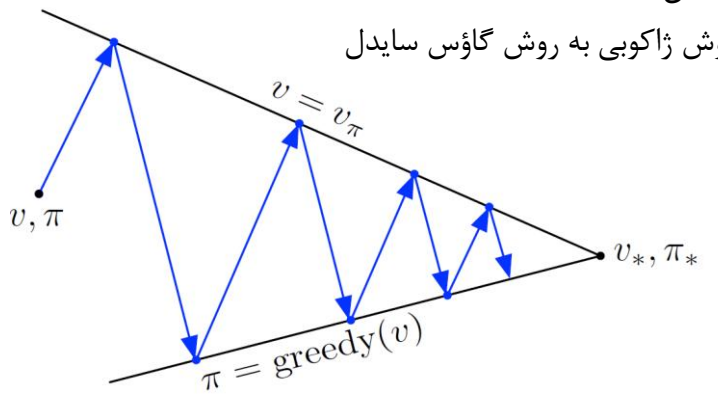
- اصلاح سیاست در پایان هر حلقهٔ سیاست‌سنجی
  - تقریب ذهن
  - از روش ژاکوبی به روش گاؤس سایدل

تکرار سیاست عمومی



تکرار سیاست عمومی

- تعامل فرایندهای سیاست‌سنجی و اصلاح سیاست
  - مستقل از یک به یکی و دیگر جزییات هر دو فرایند
  - ادامه کار تا پایدار شدن هر دو فرایند
  - پایداری تابع ارزش ر صورت سازگاری با سیاست فعلی
  - پایداری سیاست فعلی در صورت سازگاری حریصانه با تابع ارزش فعلی
- اصلاح سیاست در پایان هر حلقهٔ سیاست‌سنجی
  - تقریب ذهن
  - از روش ژاکوبی به روش گاؤس سایدل



سخن کوتاه

آشنایی با روش برنامه‌ریزی پویا

- تعیین سیاست‌ها با داشتن دینامیک سیستم
- نفرین بعد
  - هرچند بسته به فضای مسئله تا مشکل ود برنامه‌ریزی پویا
  - ب‌پ دارای عملکردی بهتر نسبت به جستجوی مستقیم و برنامه‌ریزی خطی
- در عمل بی‌اطلاع از دینامیک سیستم
  - نیاز به کسب تجربه و تعامل با محیط
  - در بخش‌های بعدی بررسی تعامل با محیط جهت تخمین توابع ارزش v و q
  - استخراج توابع

منابع

ساتن

زندی