

Методы решения задач обучения с подкреплением

Методы решения задач обучения с подкреплением

- **Динамическое программирование**
 - Хорошо проработаны теоретически
 - Требуют модели среды

 - **Метод Монте-Карло**
 - Не требуют модели среды
 - Учатся только по окончании эпизода

 - **Метод временных разностей**
 - Не требуют модели
 - Учатся в процессе работы
 - Сложнее для теоретического анализа
-

Методы решения задач обучения с подкреплением

Динамическое программирование

Исходные предпосылки

□ Известно:

Имеется конечный марковский процесс принятия решений: множества состояний S и доступных действий $A(s)$ – конечны.

Переходные вероятности:

$$\mathcal{P}_{ss'}^a = Pr \{s_{t+1} = s' \mid s_t = s, a_t = a\}$$

Ожидаемые подкрепления:

$$\mathcal{R}_{ss'}^a = E \{r_{t+1} \mid a_t = a, s_t = s, s_{t+1} = s'\}$$

Идея

- Мы хотим найти оптимальную стратегию

Это просто сделать если у нас есть оптимальная функция ценности.

- Ищем V^* или Q^* удовлетворяющие условиям оптимальности Беллмана

$$V^*(s) = \max_a \sum_{s'} \mathcal{P}_{ss'}^a \left[\mathcal{R}_{ss'}^a + \gamma V^*(s') \right]$$

$$Q^*(s, a) = \sum_{s'} \mathcal{P}_{ss'}^a \left[\mathcal{R}_{ss'}^a + \gamma \max_{a'} Q^*(s', a') \right]$$

Оценка стратегии

- Мы хотим получить функцию ценности $V^\pi(s)$ для некоторой стратегии π .
- Уравнения Беллмана дают нам систему из $|S|$ уравнений с $|S|$ неизвестными

$$\begin{aligned} V^\pi(s) &= E_\pi \{ r_{t+1} + \gamma r_{t+2} + \gamma^2 r_{t+3} + \dots | s_t = s \} \\ &= E_\pi \{ r_{t+1} + \gamma V^\pi(s_{t+1}) | s_t = s \} \\ &= \sum_a \pi(s, a) \sum_{s'} \mathcal{P}_{ss'}^a \left[\mathcal{R}_{ss'}^a + \gamma V^\pi(s') \right] \end{aligned}$$

Оценка стратегии

Рассмотрим последовательность функций V_0, V_1, V_2, \dots , аппроксимирующих искомую функцию ценности $V^\pi(s)$.

■ Начальное приближение V_0 – произвольное.

■ Последующие получаются применением правила

$$\begin{aligned} V_{k+1}(s) &= E_\pi \{ r_{t+1} + \gamma V_k(s_{t+1}) | s_t = s \} \\ &= \sum_a \pi(s, a) \sum_{s'} \mathcal{P}_{ss'}^a \left[\mathcal{R}_{ss'}^a + \gamma V_k(s') \right] \end{aligned}$$

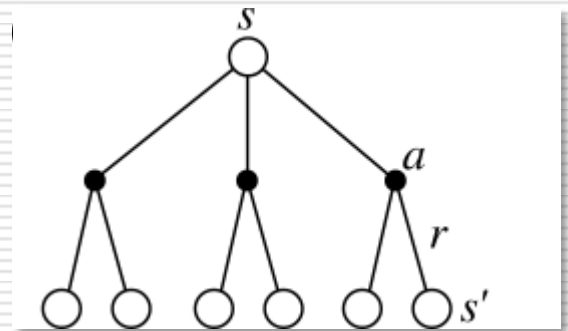
Оценка стратегии

Рассмотрим последовательность функций V_0, V_1, V_2, \dots , аппроксимирующих искомую функцию ценности $V^\pi(s)$.

$$\begin{aligned} V_{k+1}(s) &= E_\pi \{ r_{t+1} + \gamma V_k(s_{t+1}) | s_t = s \} \\ &= \sum_a \pi(s, a) \sum_{s'} \mathcal{P}_{ss'}^a \left[\mathcal{R}_{ss'}^a + \gamma V_k(s') \right] \end{aligned}$$

■ $V_k = V^\pi(s)$ – неподвижная точка.

■ $\{V_k\} \rightarrow V^\pi(s)$ при $k \rightarrow \infty$



Алгоритм оценки стратегии методом ДП

- Непосредственная реализация:
 - Делаем два массива: для «старых» и «новых» значений V .

 - Что будет, если мы будем замешать значения сразу в одном массиве?
 - Алгоритм также будет сходиться к V^π
 - Обычно он будет сходиться быстрее
 - Порядок прохода по пространству состояний будет оказывать существенное влияние на скорость сходимости
-

Алгоритм оценки стратегии методом ДП

□ Критерий останова:

- В теории сходится в пределе при бесконечном числе шагов.
- На практике обычно останавливаются немного раньше.
- Например, оценивают изменение функции ценности после каждого прохода:

$$\max_{s \in \mathcal{S}} |V_{k+1}(s) - V_k(s)|$$

Алгоритм оценки стратегии методом ДП

Вход: π – оцениваемая стратегия

Инициализация: $V(s)=0$ для всех $s \in S$.

Повторять

$\Delta \leftarrow 0$

Для всех $s \in S$:

$v \leftarrow V(s)$

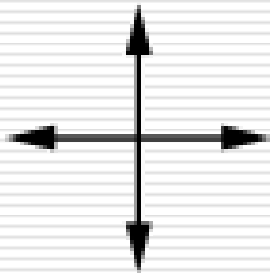
$V(s) \leftarrow \sum_a \pi(s, a) \sum_{s'} \mathcal{P}_{ss'}^a [\mathcal{R}_{ss'}^a + \gamma V(s')]$

$\Delta \leftarrow \max(\Delta, |v - V(s)|)$

пока $\Delta > \Theta$

Оценка стратегии методом ДП. Пример.

$$\mathcal{S} = \{1, 2, \dots, 14\}$$



actions

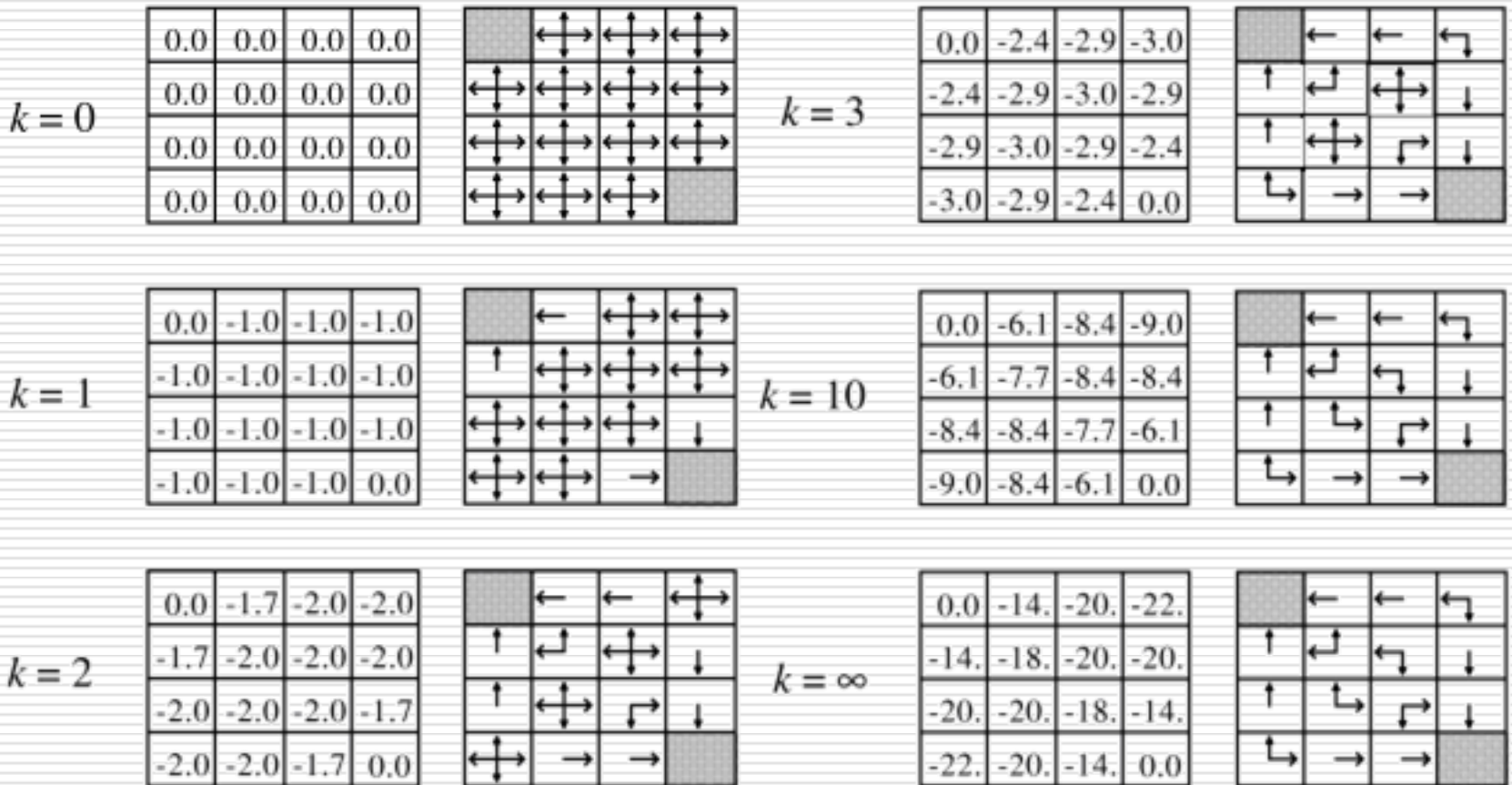
	1	2	3
4	5	6	7
8	9	10	11
12	13	14	

Подкрепление $r=-1$
пока не попадём в
терминальное
состояние.

$$\mathcal{P}_{5,6}^{\text{right}} = 1 \quad \mathcal{P}_{5,10}^{\text{right}} = 0 \quad \mathcal{P}_{7,7}^{\text{right}} = 1$$

$$\mathcal{R}_{ss'}^a = -1$$

Оценка стратегии методом ДП. Пример.



Улучшение стратегий

Что лучше, в состоянии s действовать по стратегии π или предпринять действие $a \neq \pi(s)$?

- Действуем по стратегии π :

$$V^\pi(s)$$

- Делаем a :

$$\begin{aligned} Q^\pi(s, a) &= E_\pi\{r_{t+1} + \gamma V^\pi(s_{t+1}) | s_t = s, a_t = a\} \\ &= \sum_{s'} \mathcal{P}_{ss'}^a [\mathcal{R}_{ss'}^a + \gamma V^\pi(s')]. \end{aligned}$$

Теорема об улучшении стратегий

Пусть имеются две стратегии π и π' , такие, что для всех $s \in S$:

$$Q^\pi(s, \pi'(s)) \geq V^\pi(s). \quad (1)$$

Тогда стратегия π' лучше или равна стратегии π , то есть она должна обеспечивать лучший возврат для всех состояний:

$$V^{\pi'}(s) \geq V^\pi(s). \quad (2)$$

Более того, если в одном из условий (1) имеется строгое неравенство, то и в одном из (2) должно быть строгое неравенство.

Теорема об улучшении стратегий

$$\begin{aligned} V^\pi(s) &\leq Q^\pi(s, \pi'(s)) \\ &= E_{\pi'}\{r_{t+1} + \gamma V^\pi(s_{t+1}) \mid s_t = s\} \\ &\leq E_{\pi'}\{r_{t+1} + \gamma Q^\pi(s_{t+1}, \pi'(s_{t+1})) \mid s_t = s\} \\ &= E_{\pi'}\{r_{t+1} + \gamma E_{\pi'}\{r_{t+2} + \gamma V^\pi(s_{t+2})\} \mid s_t = s\} \\ &= E_{\pi'}\{r_{t+1} + \gamma r_{t+2} + \gamma^2 V^\pi(s_{t+2}) \mid s_t = s\} \\ &\leq E_{\pi'}\{r_{t+1} + \gamma r_{t+2} + \gamma^2 r_{t+3} + \gamma^3 V^\pi(s_{t+3}) \mid s_t = s\} \\ &\vdots \\ &\leq E_{\pi'}\{r_{t+1} + \gamma r_{t+2} + \gamma^2 r_{t+3} + \gamma^3 r_{t+4} + \dots \mid s_t = s\} \\ &= V^{\pi'}(s). \end{aligned}$$

Улучшение стратегий

- Рассмотрим жадную стратегию

$$\begin{aligned}\pi'(s) &= \arg \max_a Q^\pi(s, a) \\ &= \arg \max_a E \{r_{t+1} + \gamma V^\pi(s_{t+1}) \mid s_t = s, a_t = a\} \\ &= \arg \max_a \sum_{s'} \mathcal{P}_{ss'}^a [\mathcal{R}_{ss'}^a + \gamma V^\pi(s')],\end{aligned}$$

- По определению удовлетворяет условиям теоремы об улучшении стратегий.
 - Процесс построения жадной стратегии по функции ценности состояний исходной стратегии называется улучшением стратегий.
-

Улучшение стратегий

Пусть новая жадная стратегия π' получилась такой же, как и предыдущая π . Тогда

- $V^{\pi'} = V^{\pi}$

- Для всех $s \in S$

$$\begin{aligned} V^{\pi'}(s) &= \max_a E \left\{ r_{t+1} + \gamma V^{\pi'}(s_{t+1}) \mid s_t = s, a_t = a \right\} \\ &= \max_a \sum_{s'} \mathcal{P}_{ss'}^a \left[\mathcal{R}_{ss'}^a + \gamma V^{\pi'}(s') \right]. \end{aligned}$$

- условие оптимальности Беллмана

Случай стохастических стратегий

- Стохастическая стратегия определяет вероятность $\pi(s, a)$ выполнить действие a находясь в состоянии s .
- Теорема об улучшении стратегий работает, если мы обозначим

$$Q^{\pi}(s, \pi'(s)) = \sum_a \pi'(s, a) Q^{\pi}(s, a).$$

- Если у нас есть несколько оптимальных действий можем назначить им равные вероятности
-

Итерация стратегий

После того, как мы улучшили текущую стратегию используя её функцию ценности, мы можем вычислить функцию ценности новой стратегии и улучшить её...

В результате дойдём до оптимальной функции ценности и оптимальной стратегии.

$$\pi_0 \xrightarrow{O} V^{\pi_0} \xrightarrow{y} \pi_1 \xrightarrow{O} V^{\pi_1} \xrightarrow{y} \pi_2 \xrightarrow{O} \dots \xrightarrow{y} \pi^* \xrightarrow{O} V^*$$

Итерация стратегий методом ДП. Алгоритм.

1. Инициализация:

$V(s)$ и $\pi(s)$ - произвольно для всех $s \in S$.

2. Оценка стратегии:

Повторять

$\Delta \leftarrow 0$

Для всех $s \in S$:

$v \leftarrow V(s)$

$V(s) \leftarrow \sum_{s'} \mathcal{P}_{ss'}^{\pi(s)} [\mathcal{R}_{ss'}^{\pi(s)} + \gamma V(s')]$

$\Delta \leftarrow \max(\Delta, |v - V(s)|)$

пока $\Delta > \Theta$

3. Улучшение стратегии:

$policy-stable \leftarrow true$

Для всех $s \in S$:

$b \leftarrow \pi(s)$

$\pi(s) \leftarrow \arg \max_a \sum_{s'} \mathcal{P}_{ss'}^a [\mathcal{R}_{ss'}^a + \gamma V(s')]$

Если $b \neq \pi(s)$ то

$policy-stable \leftarrow false$

Если $policy-stable$ то

Стоп

иначе

Перейти к 2.

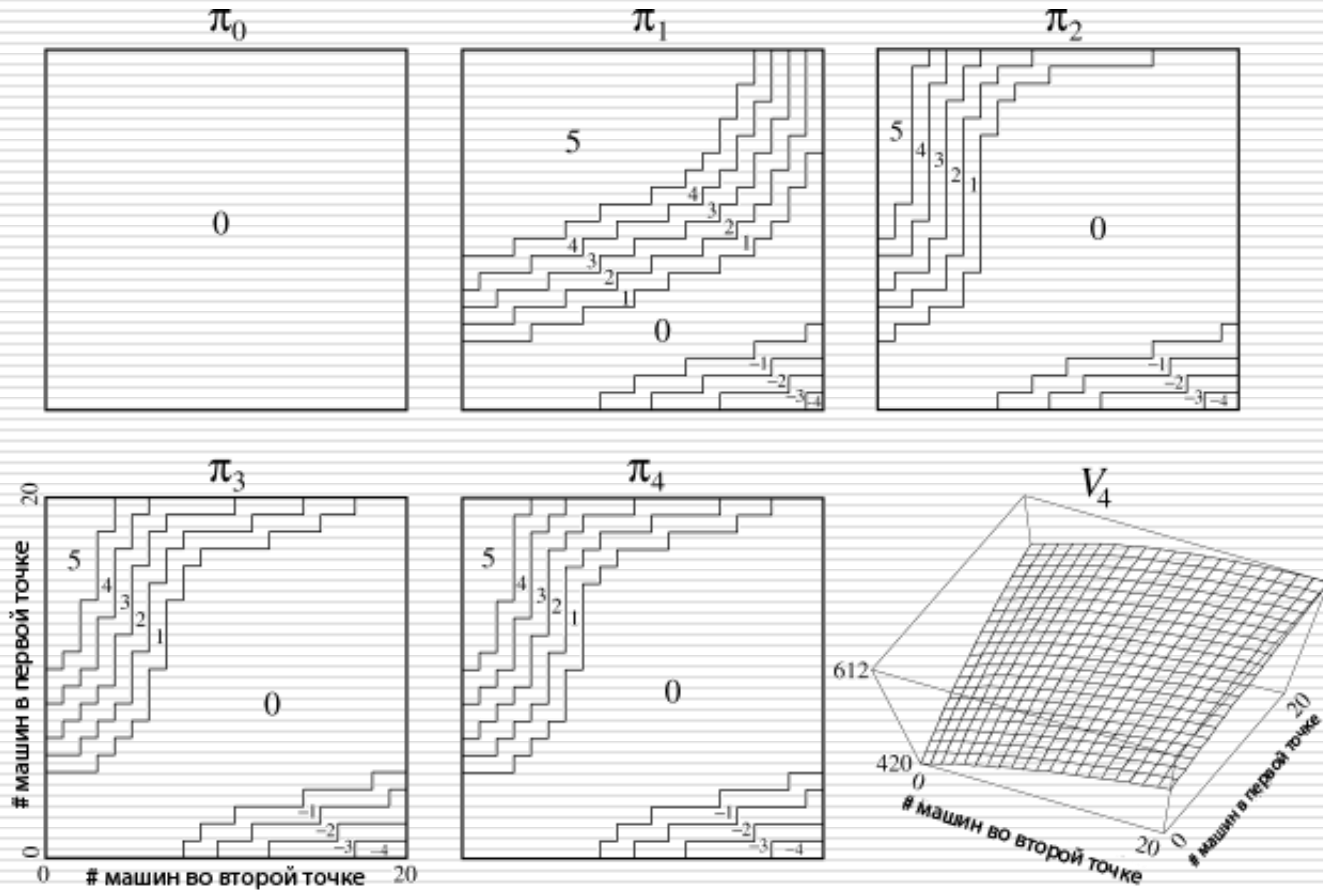
Пример: аренда машин

- Вы управляете двумя точками аренды машин в национальной сети
 - Каждый день в обе точки приходят клиенты, и если имеются доступные машины то машина сдаётся в аренду и вы получаете 10\$ от сети
 - Число клиентов, приходящих в каждую точку подчиняется распределению Пуассона, т.е. вероятность того, что придёт n человек равна $\frac{\lambda^n}{n!} e^{-\lambda}$, где λ – параметр:
 - Для запросов на аренду в точке 1 $\lambda = 3$, возвратов $\lambda = 3$.
 - В точке 2 запросы поступают с $\lambda = 4$, возвраты с $\lambda = 2$.
 - В каждой точке находятся не более 20 машин, лишние возвращаются в компанию
 - Возвращенная машина становится доступной на следующий день
 - Ночью можно переместить не более 5 машин из одной точки в другую, затратив на каждую машину 2\$
-

Пример: аренда машин

- Длющийся конечный Марковский процесс принятия решений
 - Шаг – 1 день
 - Состояние – число машин в каждой точке вечером
 - Действие – сколько машин перегнать
 - Возьмём коэффициент дисконта $\gamma = 0.9$
-

Пример: аренда машин



Итерация ценностей

- Наш алгоритм для каждого шага улучшения стратегии требует вычисления функции ценности, что долго.
- Рассмотренный пример показывает, что можно начинать улучшать стратегию не дожидаясь конца вычислений функции ценности.
- Будем делать стратегию жадной после каждой итерации вычисления функции ценности.

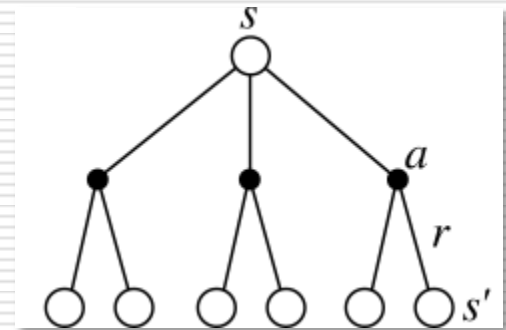
$$\begin{aligned} V_{k+1}(s) &= \max_a E \{ r_{t+1} + \gamma V_k(s_{t+1}) \mid s_t = s, a_t = a \} \\ &= \max_a \sum_{s'} \mathcal{P}_{ss'}^a \left[\mathcal{R}_{ss'}^a + \gamma V_k(s') \right], \end{aligned}$$

- $\{V_k\} \rightarrow V^*(s)$ при $k \rightarrow \infty$
-

Итерация стратегий и итерация ценностей

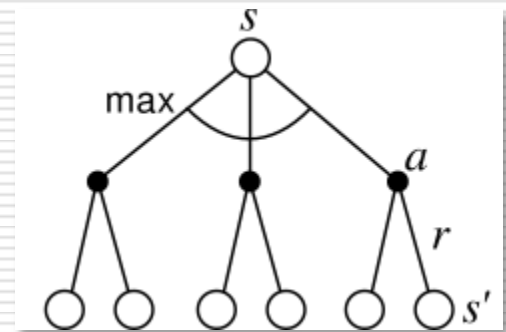
□ Итерация стратегий

$$\begin{aligned} V_{k+1}(s) &= E_{\pi} \{ r_{t+1} + \gamma V_k(s_{t+1}) | s_t = s \} \\ &= \sum_a \pi(s, a) \sum_{s'} \mathcal{P}_{ss'}^a [\mathcal{R}_{ss'}^a + \gamma V_k(s')] \end{aligned}$$



□ Итерация ценностей

$$\begin{aligned} V_{k+1}(s) &= \max_a E \{ r_{t+1} + \gamma V_k(s_{t+1}) | s_t = s, a_t = a \} \\ &= \max_a \sum_{s'} \mathcal{P}_{ss'}^a [\mathcal{R}_{ss'}^a + \gamma V_k(s')], \end{aligned}$$



Алгоритм итерации ценностей методом ДП

□ Критерий останова:

- Оцениваем изменение функции ценности после каждого прохода:

$$\max_{s \in \mathcal{S}} |V_{k+1}(s) - V_k(s)|$$

Алгоритм итерации ценностей методом ДП

Инициализация:

$V(s)$ и $\pi(s)$ - произвольно для всех $s \in S^+$.

Повторять

$\Delta \leftarrow 0$

Для всех $s \in S$:

$v \leftarrow V(s)$

$V(s) \leftarrow \max_a \sum_{s'} \mathcal{P}_{ss'}^a [\mathcal{R}_{ss'}^a + \gamma V(s')]$

$\Delta \leftarrow \max(\Delta, |v - V(s)|)$

пока $\Delta > \Theta$

Выдать стратегию $\pi(s)$:

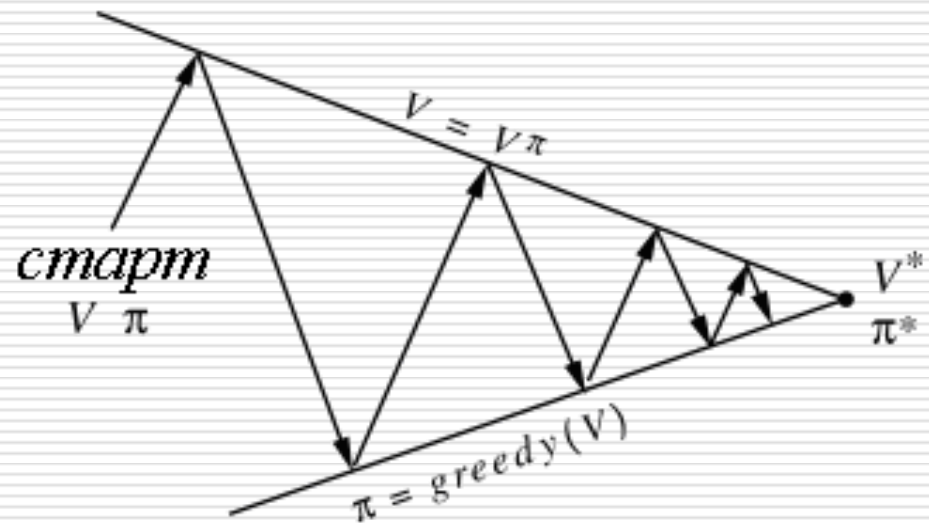
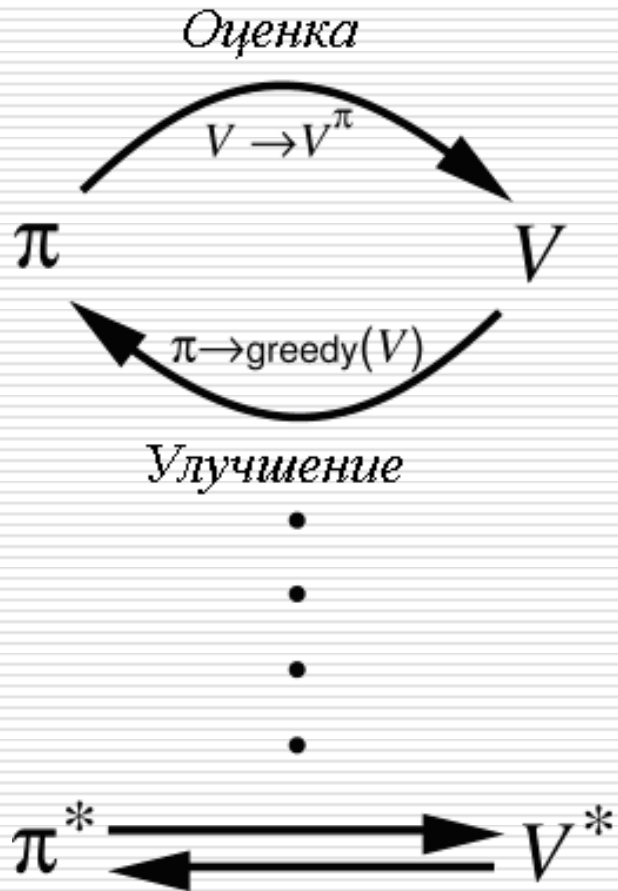
$\pi(s) = \arg \max_a \sum_{s'} \mathcal{P}_{ss'}^a [\mathcal{R}_{ss'}^a + \gamma V(s')]$

Асинхронное динамическое программирование

Нам всё ещё необходимо делать проход по всему пространству состояний. Если состояний много, то это может быть неприемлемо.

- Можно обновлять ценности только некоторых состояний – асинхронное ДП.
 - Сходимость как правило гарантирована если мы будем обновлять их все.
 - Можно чаще обновлять ценности тех состояний, которые важны для оптимального поведения.
 - Можно обновлять ценности состояний, в которые агент попадает во время работы.
-

Обобщенная итерация стратегий



Вычислительная сложность динамического программирования

- Пусть n – число состояний, m – число действий.
 - Методы ДП находят решение за полиномиальное от m и n время, хотя общее число стратегий равно m^n .
 - Позволяют решать задачи с числом состояний примерно в 100 раз большим, чем можно решать методом линейного программирования.
 - Для очень больших задач целесообразно применять асинхронное ДП.
-

Выводы

- ❑ Оценка стратегии π – вычисление её функции ценности V^π
 - ❑ Улучшение стратегии – вычисление лучшей стратегии, используя функцию ценности
 - ❑ Итерация стратегий – последовательное применение оценки стратегии и её улучшения
 - ❑ Итерация ценностей – оценка, при которой стратегия сразу же считается жадной относительно текущего приближения функции ценности
-

Выводы

- ❑ Классические методы ДП проходят по всему пространству состояний, выполняя **полное обновление** для всех состояний
 - ❑ Полное обновление означает что мы рассматриваем все возможные переходы
 - ❑ Полные обновления тесно связаны с уравнениями Беллмана
 - ❑ Так как у нас есть 4 базовых функции V^π, V^*, Q^π, Q^* и 4 соответствующих вида уравнений Беллмана, то мы можем использовать 4 вида полных обновлений
-

Выводы

- **Обобщенная итерация стратегий**
 - Алгоритм представляет из себя два процесса:
 - Один движется в сторону оценки стратегии
 - Второй улучшает стратегию
 - Оба процесса, хотя и меняют основу друг для друга, приводят нас к общему решению: стратегии и функции ценности, которые не меняются этими процессами и являются оптимальными
 - В некоторых случаях можно доказать сходимость обобщенной итерации стратегий
 - Одним из примеров являются асинхронные методы ДП, которые обновляют состояния в произвольном порядке
-

Выводы

- Методы ДП обновляют оценку ценности состояний на основе неточных оценок ценности других состояний
 - будем называть такие методы **рекурсивными**
 - Методы ДП требуют наличия модели среды в форме переходных вероятностей и ожидаемых подкреплений.
-

Задача 2

- Игрок делает ставки (целое число) на результат бросания монетки:
 - Если выпал орёл игрок получает столько денег, сколько он поставил
 - Если решка – теряет деньги
 - Вероятность выпадения орла известна и равна p .
 - Конец игры:
 - Игрок получает 100\$ и выигрывает
 - У игрока кончаются деньги и он проигрывает
-

Задача 2

- Имеем эпизодическую, конечную Марковскую задачу без дисконта
 - Состояние – сколько есть денег $s \in \{1, 2, \dots, 99\}$.
 - Действие – сколько денег ставим $a(s) \in \{1, 2, \dots, \min(s, 100 - s)\}$.
 - Подкрепление – 0, пока игрок не достигнет своей цели; когда достигнет +1
 - Функция ценности состояния – вероятность выиграть имея столько денег.
 - Оптимальная стратегия – оптимальный размер ставки в зависимости от капитала.
 - Вычислите оптимальную функцию ценности и стратегию для этой задачи, используя ДП, для $p = 0.25$ и $p = 0.55$. Подумайте, почему они такие.
-