Методы решения задач обучения с подкреплением

Метод временных разностей (TD)

Методы решения задач обучения с подкреплением

- Методы динамического программирования
  - Требуют модели
  - Обновляют оценку ценности одного состояния на основе оценок для других состояний.
- Методы Монте-Карло
  - Обучаются взаимодействуя со средой
  - При расчёте оценки ценности одного состояния не используют оценки для других состояний

Прогнозирование методом TD

- Как и Монте-Карло, методы TD оценивают $V(s_t)$ на основе того, что произошло после посещения s_t.
- Монте-Карло ждёт конца эпизода и использует обновление

$$V(s_t) \leftarrow V(s_t) + \alpha \Big[R_t - V(s_t) \Big],$$

- Метод TD(0) ждёт один шаг и делает обновление

$$V(s_t) \leftarrow V(s_t) + \alpha \Big[r_{t+1} + \gamma V(s_{t+1}) - V(s_t) \Big].$$

Прогнозирование методом TD

$$V^\pi(s) = E_\pi\{R_t | s_t = s\}$$

Монте-Карло

$$= E_\pi\left\{ \sum_{k=0}^{\infty} \gamma^k r_{t+k+1} \,\middle|\, s_t = s \right\}$$

$$= E_\pi\left\{ r_{t+1} + \gamma \sum_{k=0}^{\infty} \gamma^k r_{t+k+2} \,\middle|\, s_t = s \right\}$$

$$= E_\pi\left\{ r_{t+1} + \gamma V^\pi(s_{t+1}) \,\middle|\, s_t = s \right\}.$$

DP

TD

Алгоритм TD(0) оценки стратегии

Вход:

π - оцениваемая стратегия

Инициализация:

$V(s)$ – произвольно

Повторять для всех эпизодов

$s \leftarrow$ начальное состояние

Для каждого шага эпизода

$a \leftarrow$ действие для s согласно π

Выполнить a, узнать s' и r.

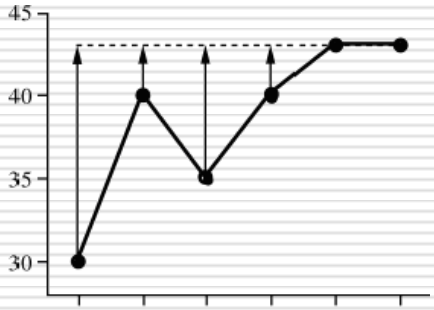
$V(s) \leftarrow V(s)+\alpha[r+\gamma V(s')-V(s)]$

$s \leftarrow s'$

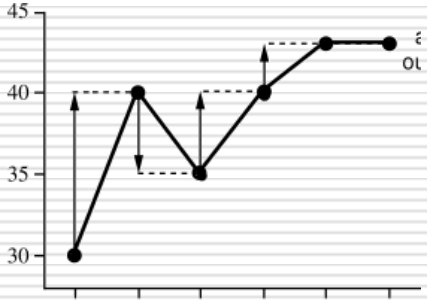
Прогноз методом TD. Пример

	Прошло (мин)	Прогноз остатка	Прогноз общее
Вышел с работы	0	30	30
В машине. Дождь.	5	35	40
Проехали по трассе.	20	15	35
Упёрлись в грузовик.	30	10	40
На своей улице	40	3	43
Дома	43	0	43

Монте-Карло



TD(0)



Преимущества TD

+ Не требуется модель

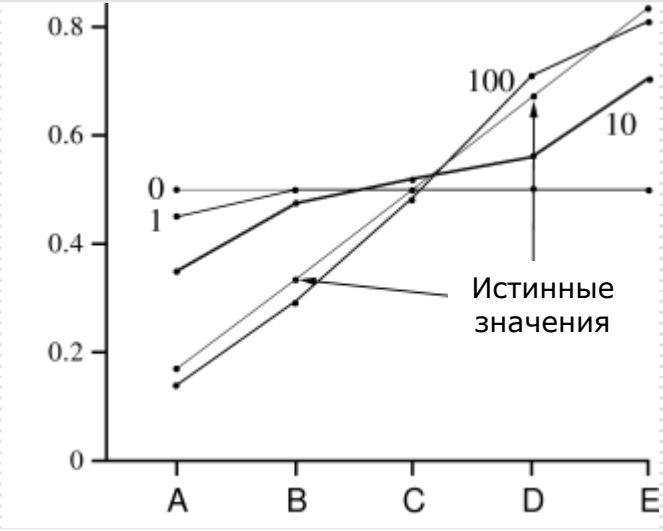
+ Подходят для реализации в онлайн режиме

+ Учатся непосредственно во время эпизода

+ Сходится к V^{π}.

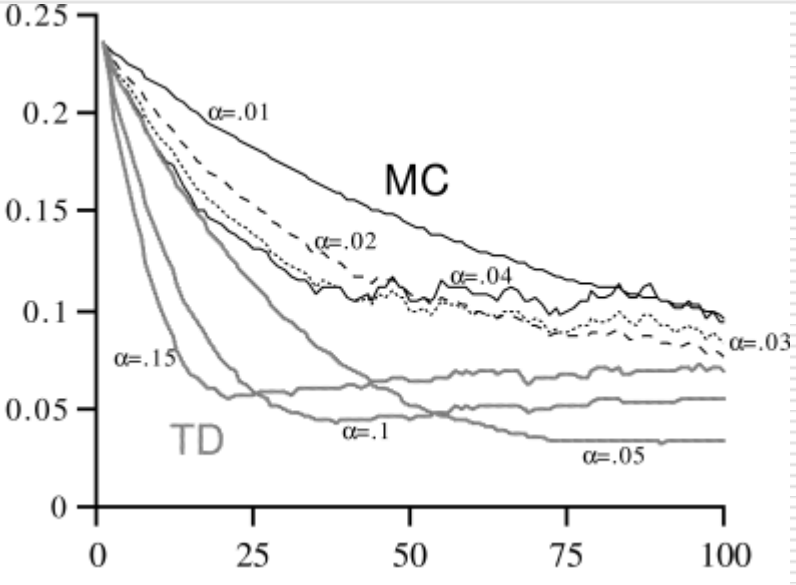
Пример – случайное брожение

Оценка



Состояние

Средняя ошибка по состояниям



Эпизоды

Пакетный режим

- Ждём до конца эпизода, запоминая ΔV(s).
- По завершению эпизода прибавляем к V(s) сумму запомненных изменений.
- Повторяем много раз для имеющихся данных.
- При достаточно небольшом α TD(0) сходится к некоторому ответу.
- Монте-Карло тоже сходится, но к другому ответу – даёт минимум на обучающей выборке.

Пакетный режим. Пример.

- Пусть есть 8 эпизодов

A,0,B,0	B,1	B,1	B,1
B,1	B,1	B,0	B,1

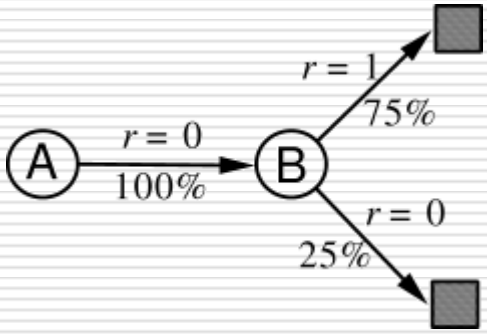
Чему равно V(B)?

V(B)=¾

Чему равно V(A)?

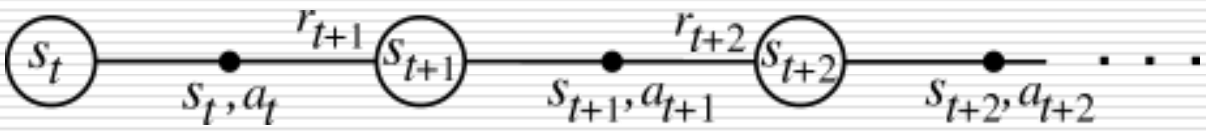
Монте-Карло: V(A)=0

Пакетный TD(0): V(A)=¾



Управляем методом TD следуя оцениваемой стратегии

- Используем обобщенную итерацию стратегий
- Оцениваем функцию ценности состояний



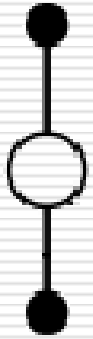
- Правило обновления

$$Q(s_t, a_t) \leftarrow Q(s_t, a_t) + \alpha \Big[r_{t+1} + \gamma Q(s_{t+1}, a_{t+1}) - Q(s_t, a_t) \Big].$$

Алгоритм SARSA

Инициализация:

$Q(s,a)$ – произвольно



Повторять для всех эпизодов

$s \leftarrow$ начальное состояние

$a \leftarrow \varepsilon$-жадное по Q действие для s.

Для всех шагов эпизода

Выполнить a, узнать s' и r.

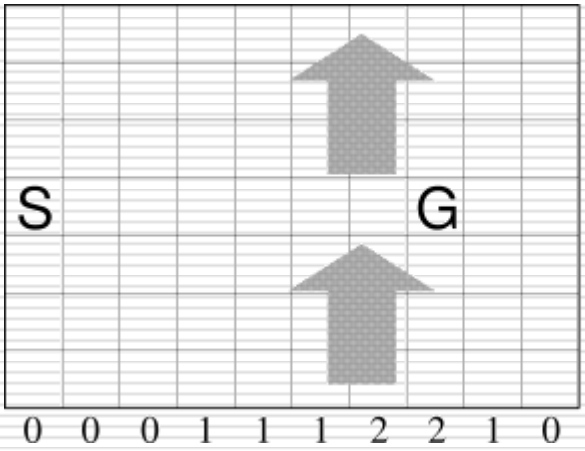
$a' \leftarrow \varepsilon$-жадное по Q действие для s'.

$Q(s,a) \leftarrow Q(s,a) + \alpha[r+\gamma Q(s',a')-Q(s,a)]$.

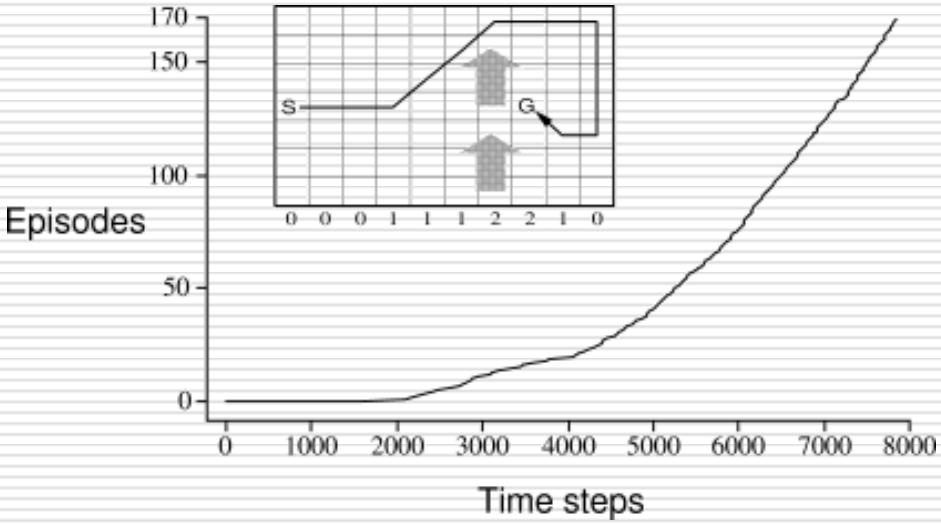
$s \leftarrow s'$

$a \leftarrow a'$

Пример - ветер



standard moves



TD(0). Управляем и оцениваем разные стратегии.

- Watkins, 1989. Q-learning:

$$Q(s_t, a_t) \leftarrow Q(s_t, a_t) + \alpha \left[r_{t+1} + \gamma \max_a Q(s_{t+1}, a) - Q(s_t, a_t) \right].$$

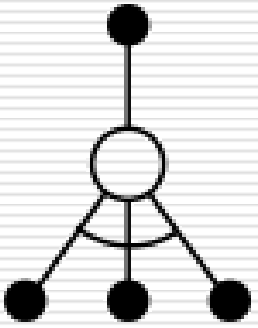
- Аппроксимируем Q^*.

- Сходится, если агент продолжает посещать все пары (*s,a*).

Алгоритм Q-learing.

Инициализация:

$Q(s,a)$ – произвольно



Повторять для всех эпизодов

$s \leftarrow$ начальное состояние

Для всех шагов эпизода

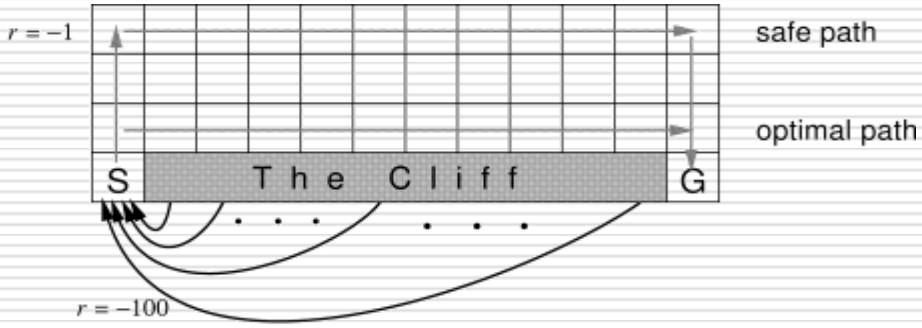
$a \leftarrow \varepsilon$-жадное по Q действие для s.

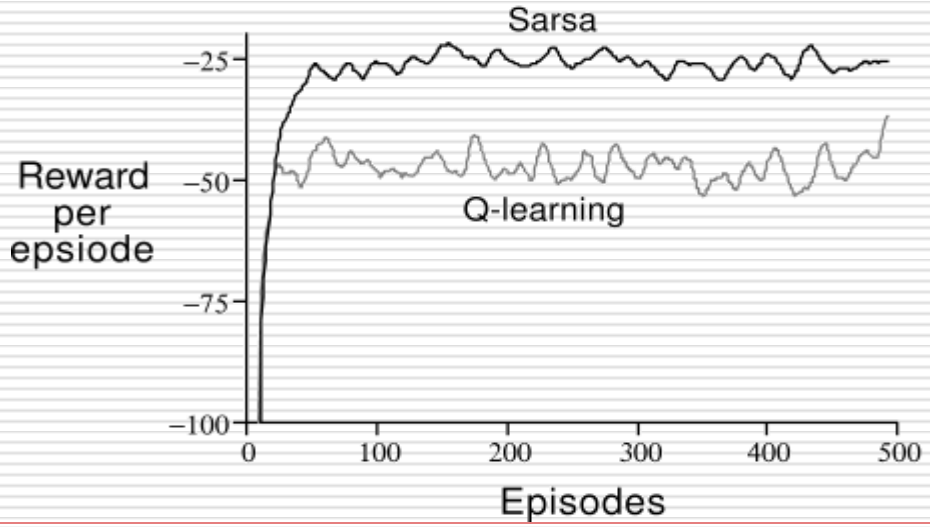
Выполнить a, узнать s' и r.

$Q(s,a) \leftarrow Q(s,a) + \alpha[r + \gamma \max_a Q(s',a') - Q(s,a)]$.

$s \leftarrow s'$

Q-Learning. Пример.

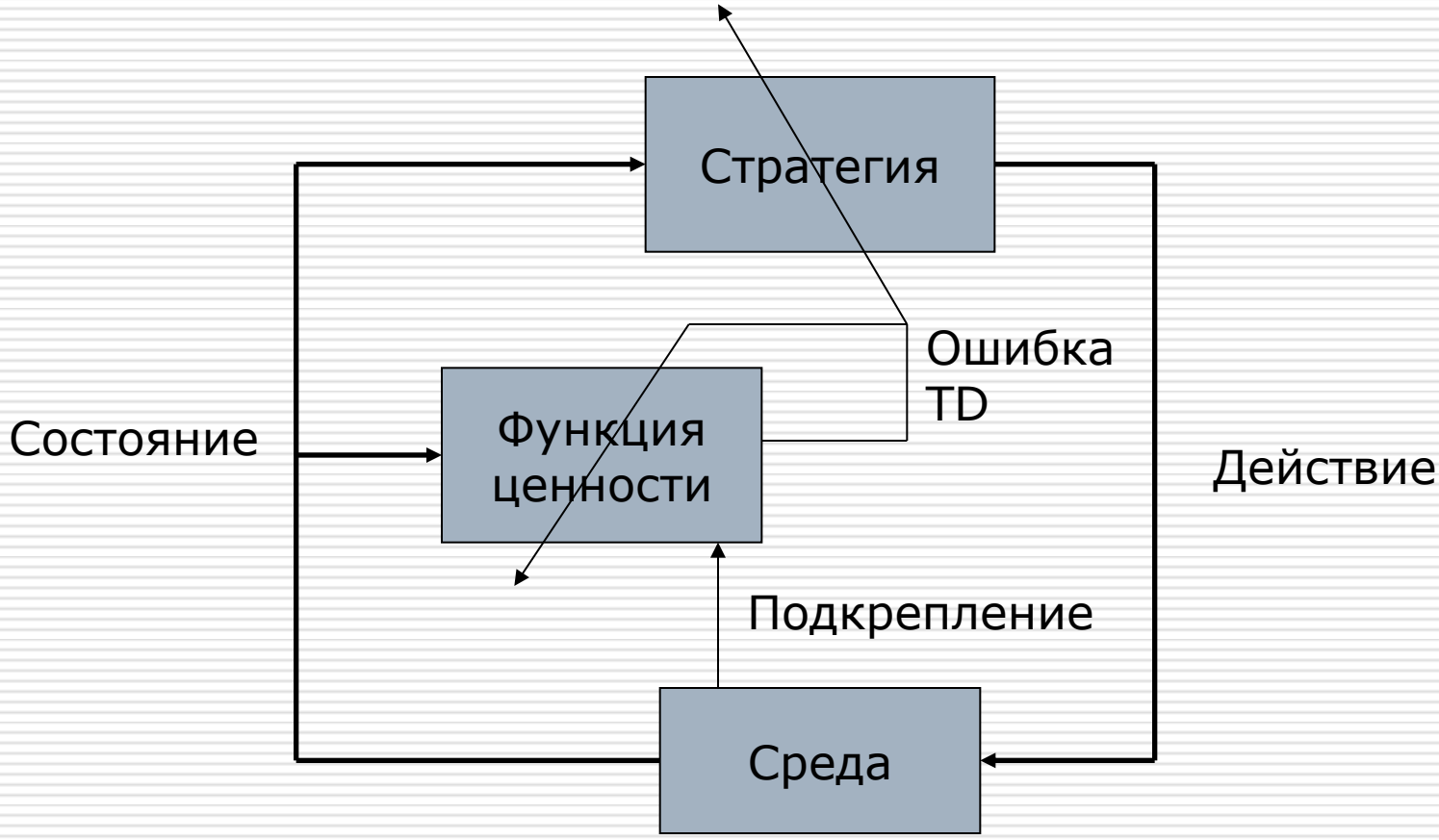




Методы деятеля-критика

- Храним и стратегию и функцию ценности.
- Стратегия – деятель.
- Функция ценности – критик.
- В роли критики выступает ошибка TD, используемая для корректировки обеих структур.

Методы деятеля-критика



Методы деятеля-критика

- Критик
  - Функция ценности состояний V
  - Ошибка TD:

$$\delta_t = r_{t+1} + \gamma V(s_{t+1}) - V(s_t),$$

- Деятель
  - Предпочтения выбора действий

$$\pi_t(s,a) = Pr\{a_t = a \mid s_t = s\} = \frac{e^{p(s,a)}}{\sum_b e^{p(s,b)}},$$

  - Изменение предпочтений

$$p(s_t,a_t) \leftarrow p(s_t,a_t) + \beta\delta_t,$$

$$p(s_t,a_t) \leftarrow p(s_t,a_t) + \beta\delta_t\Big[1 - \pi_t(s_t,a_t)\Big].$$

Методы деятеля-критика

- Были широко распространены на ранних этапах, позднее внимание переключилось на методы, использующие функции ценности действий.
- Требуют минимальных расчётов для выбора действия.
- Могут находить стохастические стратегии: оптимальные вероятности выполнения действий.
- Интересны в плане биологических аналогий.

R-learning

- Расширенный вариант задачи обучения с подкреплением:
  - Не используется дисконт
  - Действие не разбивается на конечные эпизоды
  - Хотим получать максимальный возврат на каждом шаге.
- Функции ценности для стратегии π определяются относительно среднего ожидаемого подкрепления:

$$\rho^{\pi} = \lim_{n \to \infty} \frac{1}{n} \sum_{t=1}^{n} E_{\pi}\{r_t\},$$

- Если процесс является эргодическим, то ρ^{π} не зависит от начального состояния.

R-learning

- Функция ценности состояний и действий определяются в зависимости от характера перехода к среднему значению:

$$\tilde{V}^{\pi}(s) = \sum_{k=1}^{\infty} E_{\pi}\{r_{t+k} - \rho^{\pi} | s_t = s\},$$

$$\tilde{Q}^{\pi}(s,a) = \sum_{k=1}^{\infty} E_{\pi}\{r_{t+k} - \rho^{\pi} | s_t = s, a_t = a\}.$$

- Оптимальными будем считать стратегии, у которых ρ^{π} максимально.

R-learning

Инициализация

ρ, $Q(s,a)$ - произвольно для всех s,a

Повторять бесконечно

$s \leftarrow$ текущее состояние

Выбрать действие a согласно стратегии поведения (например, ε-жадной)

Выполнить a, узнать r и s'.

$Q(s,a) \leftarrow Q(s,a) + \alpha[r - \rho + \max_{a'} Q(s',a') - Q(s,a)]$

Если $Q(s,a) = \max_a Q(s,a)$ то

$\rho \leftarrow \rho + \beta[r - \rho + \max_{a'} Q(s',a') - \max_a Q(s,a)]$

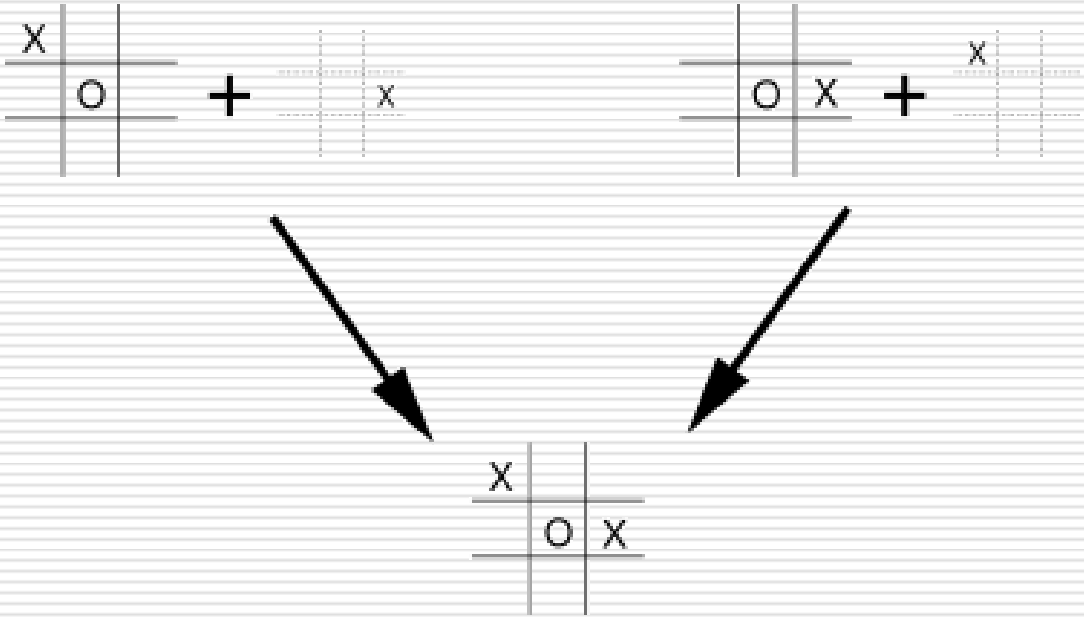
Игры и пост-состояния

- В некоторых случаях, например во многих играх, удобно оценивать функцию ценности не от состояния (до хода), а от результата нашего хода – пост-состояния.

- Такой подход работает, если известна начальная динамика среды, но не вся динамика (мы не знаем ход соперника).

Игры и пост-состояния

- Этот подход эффективен, так как разные пары (состояние, действие) могут приводить к одному пост-состоянию.



Метод временных разностей. Итоги.

- Сохраняется идея обобщенной итерации стратегий
- Так как для оценки используются опыт агента, то мы должны балансировать исследование и использование знаний
  - Одна стратегия: Sarsa и деятель-критик
  - Разные стратегии: Q-learning и R-learning
- Мы рассмотрели простой случай:
  - Одношаговый
  - Табличный
  - Не использующий модель
- Могут использоваться для прогнозирования динамических процессов