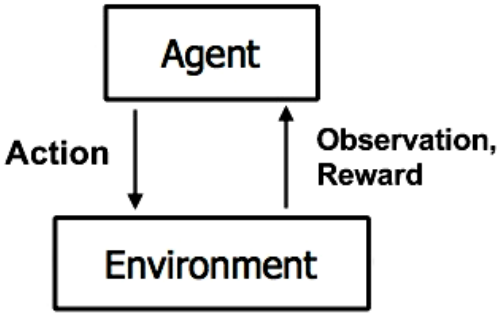
Методы Reinforcement learning

Постановка задачи

Агент взаимодействует с окружающей средой, предпринимает действия, переходит в разные состояния.

Окружающая среда поощряет агента за действия.



R.Sutton, A.Barto, 1998

Многорукий бандит

Агенты с одним состоянием.
Состояние агента не меняется. У агента фиксированный набор действий $\{A\}$.
И возможность выбора из этого набора действий.

Пример. Агент находится в комнате с несколькими игровыми автоматами. Каждый автомат имеет неизвестное, стационарное распределение вероятности.

Целью является максимизация выигрыша после ряда действий.

Жадный алгоритм

Если действие a было выбрано k_a раз,

то его ценность можно оценить как

$$Q(a) = \frac{r_1 + r_2 + \ldots + r_{k_a}}{k_a}$$

и выбирать действие, которое максимизирует вознаграждение:

$$Q(a^*) = \max_a Q(a)$$

Случайные стратегии

С вероятностью (1-α)
Выбирать действие с лучшей ожидаемой прибытью.

С вероятностью α
Выбирать случайное действие.

Модель

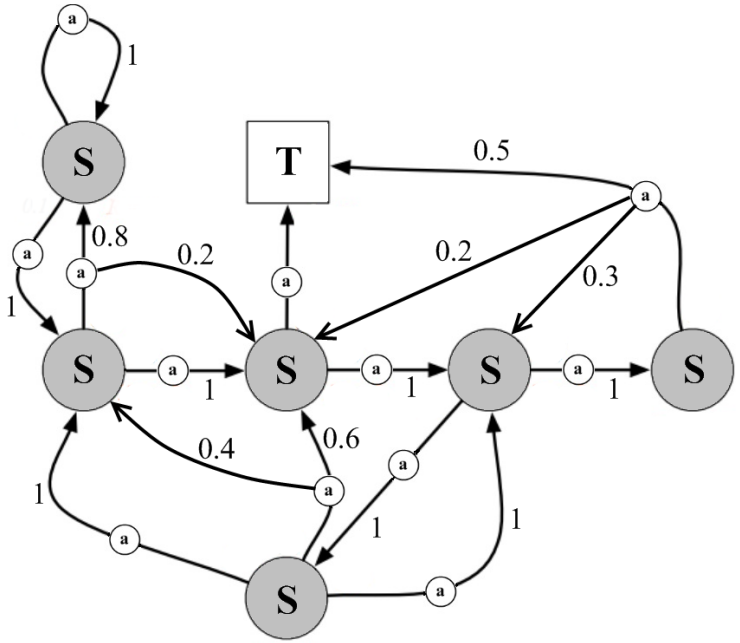
Марковский процесс принятия решений

−множество состояний S

−множество действий A

−вознаграждение при переходе из состояния s в состояние s' после действия a задается функцией $R_{ss'}^{a}$

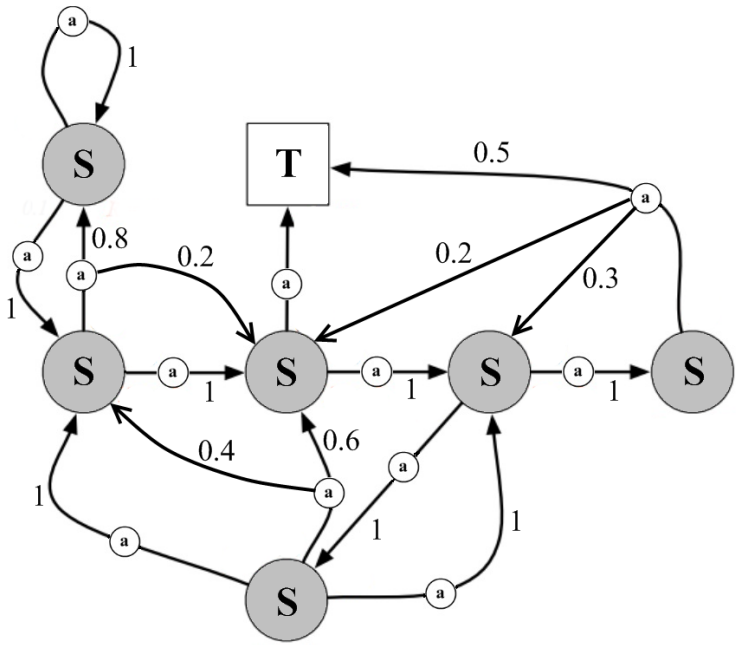
−вероятность перехода из состояния s в состояние s' после действия a задается функцией перехода $P_{ss'}^{a}$



Модель

Стратегия является случайной величиной, задающей выбор действия **a** в состоянии **s**.

$$\pi(s,a) = \Pr(A_r = r, S_t = s)$$



Модель бесконечного горизонта

Приведенная выгода:

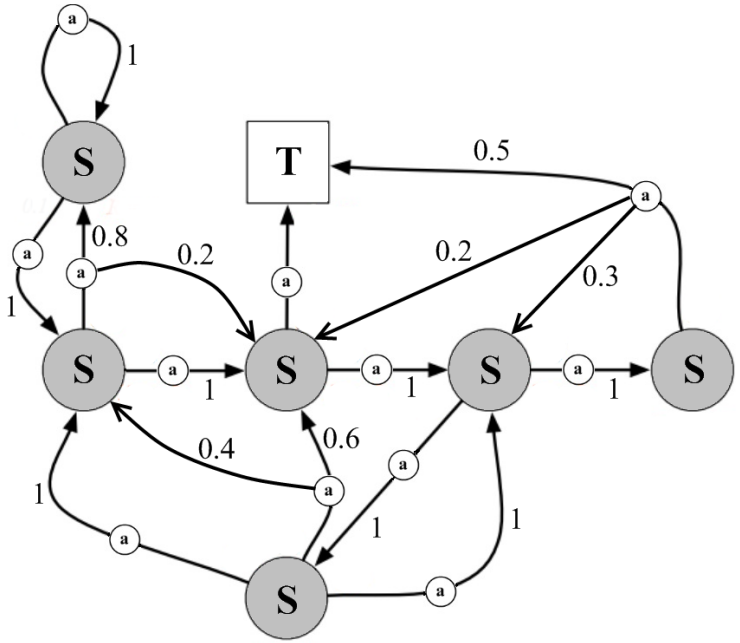
$$R_t = r_{t+1} + \gamma r_{t+2} + \gamma^2 r_{t+3} + \ldots = \sum_{k=0}^{\infty} \gamma^k r_{t+k+1}$$

Коэффициент приведения: $\gamma \in [0,1]$

Поиск стратегий в известной модели

Ожидаемая выгода, если стартовать из s

$$V^{\pi}(s) = E_{\pi}[\sum_{k=0}^{\infty} \gamma^{k} r_{t+k+1} \mid s_t = s]$$



Как найти стратегию, которая максимизирует выгоду?

Написать систему уравнений для V?

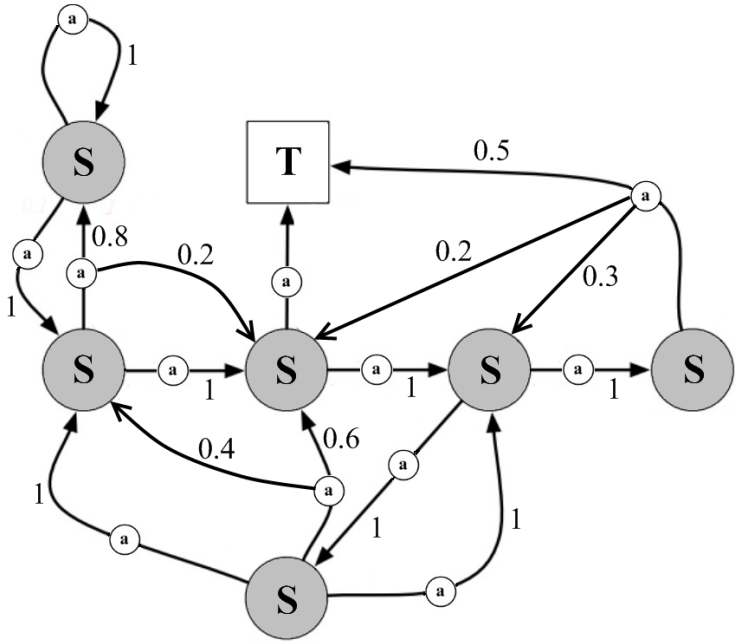
Поиск стратегий в известной модели

Ожидаемая выгода, если стартовать из s:

$$\mathrm{V}(s) = E_\pi[\sum_{k=0}^{\infty} \gamma^{\mathrm{k}} r_{t+k+1} \mid S_t = s] =$$

$$= E_\pi[r_{t+1} + \gamma \sum_{k=0}^{\infty} \gamma^{\mathrm{k}} r_{t+k+2} \mid S_t = s] =$$

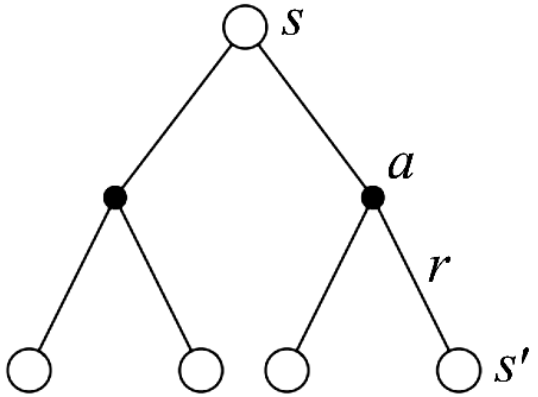
$$= E_\pi[r_{t+1} + \gamma \mathrm{V}(S_{t+1}) \mid S_t = s]$$



Поиск стратегий в известной модели

Уравнение Беллмана:

$$V^{\pi}(s) = \sum_{a} \pi(s,a) \sum_{s'} P^{a}_{ss'} (R^{a}_{ss'} + \gamma V^{\pi}(s'))$$



$R^{a}_{ss'}$ – вознаграждение при переходе из состояния s в состояние s' после действия a

$P^{a}_{ss'}$ – вероятность перехода из состояния s в состояние s' после действия a

Поиск стратегий в известной модели

Оптимальная функция ценности состояния:

$$V^*(s) = \max_{\pi} V^{\pi}(s)$$

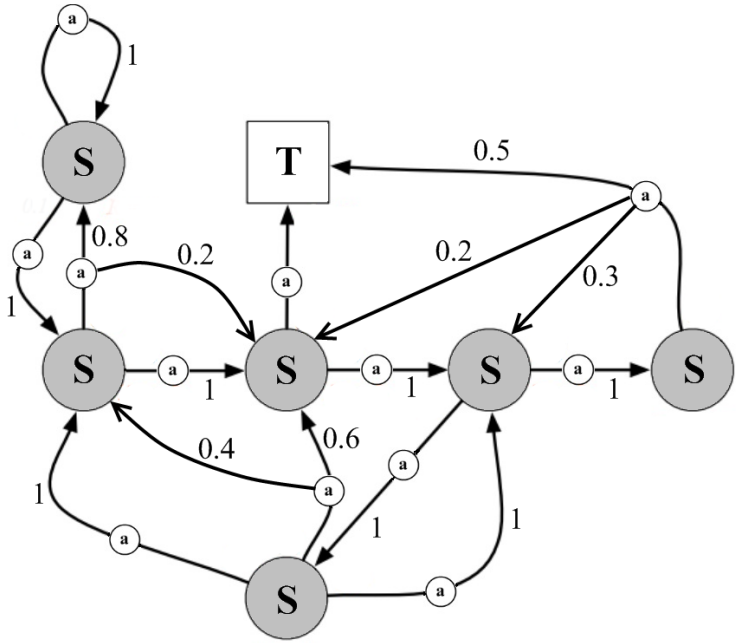
по всем стратегиям π

Уравнение оптимальности Беллмана:

$$V^*(s) = \max_{a} \sum_{s'} P^a_{ss'}(R^a_{ss'} + \gamma V^*(s'))$$

Оптимальная стратегия:

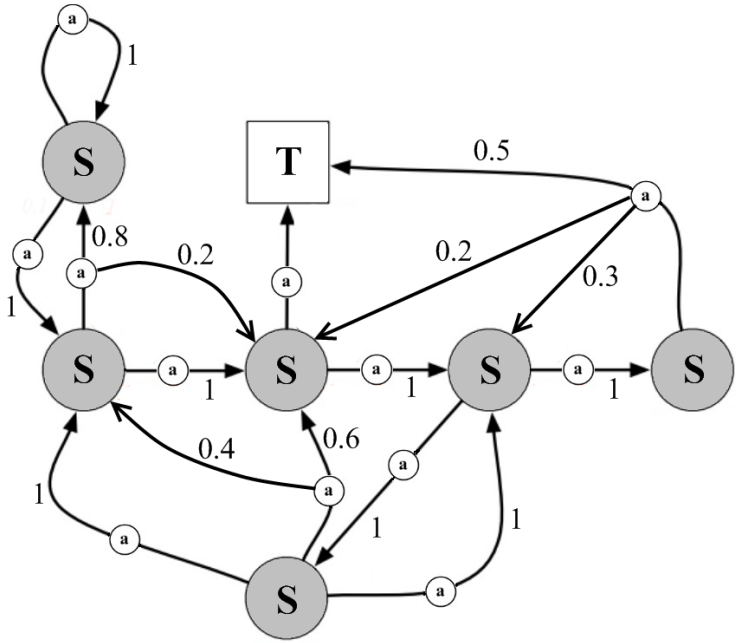
$$\pi^*(s) = \arg\max_{a} V^*(s)$$



Итерация по ценностям

Итерационный процесс для оптимальной функции состояний:

$$V^*_{k+1}(s) = \max_a \sum_{s'} P^a_{ss'}(R^a_{ss'} + \gamma V^*_k(s'))$$



Итерация по стратегиям

$$\pi_0 \xrightarrow{E} V^{\pi_0} \xrightarrow{I} \pi_1 \xrightarrow{E} V^{\pi_1} \xrightarrow{I} \pi_2 \xrightarrow{E} \ldots \xrightarrow{I} \pi^* \xrightarrow{E} V^*$$

Шаг 1. стратегия улучшается, подстраиваясь под функцию ценности

$$\pi(s) = \arg\max_a V(s)$$

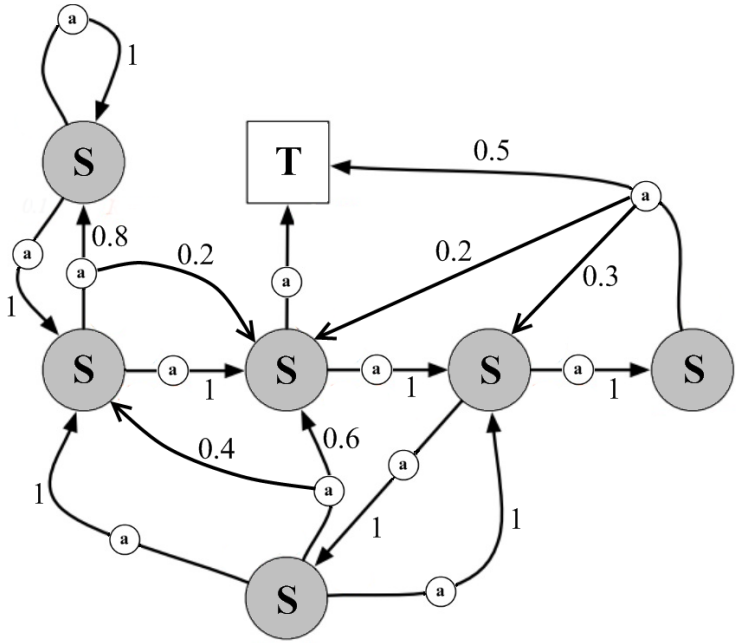
Шаг 2. функция ценности состояний $V(s)$ корректируется, чтобы соответствовать стратегии

Метод Монте-Карло

Требуется оценить величину $V^{\pi}(s)$.

Вероятности переходов не даны.

$V^{\pi}(s)$ оценивается как среднее значение выгод в ряде эпизодов.



Оценка ценности действия

Ожидаемая выгода, при начальном состоянии s, осуществленном действии a:

$$Q^{\pi}(s,a) = E_{\pi}[\sum_{k=0}^{\infty} \gamma^{k} r_{t+k+1} \mid S_t = s, A_t = a]$$

По формуле полной вероятности:

$$V^{\pi}(s) = \sum_{a} \pi(s,a) Q^{\pi}(s,a)$$

Метод Монте-Карло

Формирование стратегии.
Стратегия выбирает действие, которое максимизирует значение Q:

$$\pi(s) = \arg\max_a Q(s,a)$$

Процесс улучшения стратегии π:

$$\pi_0 \xrightarrow{E} Q^{\pi_0} \xrightarrow{I} \pi_1 \xrightarrow{E} Q^{\pi_1} \xrightarrow{I} \pi_2 \xrightarrow{E} \ldots \xrightarrow{I} \pi^* \xrightarrow{E} Q^*$$

Шаг 1. стратегия улучшается, подстраиваясь под функцию ценности
Шаг 2. функция ценности корректируется, чтобы соответствовать стратегии

Многорукий бандит

Если дейстие a было выбрано k_a раз,

то его ценность можно оценить как

$$Q(a) = \frac{r_1 + r_2 + \ldots + r_{k_a}}{k_a}$$

и выбирать действие, которое максимизирует вознаграждение:

$$Q(a^*) = \max_a Q(a)$$

Среднее значение можно вычислять по формуле:

$$Q_{k+1} = Q_k + \frac{1}{k+1}(r_{k+1} - Q_k)$$

Метод временных различий

Общее правило корректировки:

$$\begin{aligned} &\textit{Новая оценка} \leftarrow \textit{Старая оценка} \; + \\ &+ \textit{Длина шага} \; \times \; [\textit{Цель} - \textit{Старая оценка}] \end{aligned} \tag{1}$$

Свойство функции состояния:

$$V(s) = E_{\pi}[r_{t+1} + \gamma V(S_{t+1}) \mid S_t = s] \tag{2}$$

TD(0)-метод:

$$V(s_t) \leftarrow V(s_t) + \alpha \left[r_{t+1} + \gamma V(s_{t+1}) - V(s_t) \right] \tag{3}$$

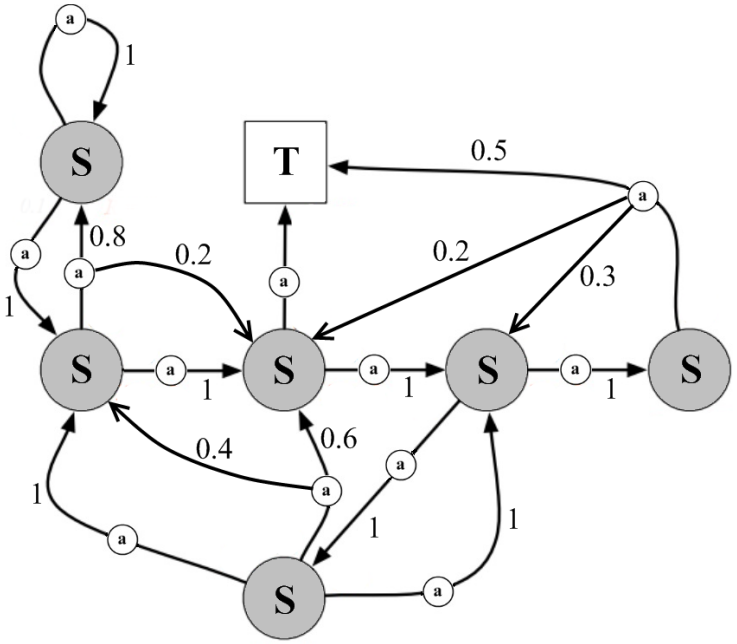
Если α уменьшается, то $V(s) \rightarrow V^{\pi}(s)$.

Гибкие стратегии

<u>Ɛ-жадные стратегии</u>

С вероятностью (1- Ɛ) выбирать действие по жадной стратегии.

С вероятностью Ɛ выбирать действие случайно.



Метод SARSA

$(s_t, a_t, r_{t+1}, s_{t+1}, a_{t+1})$

Позволяет построить функцию Q*(s, a), для которой жадная стратегия будет давать оптимальное управление.

1. TD(0)-обучение для функции ценности действий Q(s, a):

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

2. Использовать ε-жадную стратегию.

Если α уменьшается, то Q(s,a)⟶ Q*(s,a).

Q-обучение

Позволяет построить функцию Q(s, a), для которой жадная стратегия будет давать оптимальное управление.

1. Корректировка функции ценности действий Q(s, a):

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

2. Использовать ε-жадную стратегию.

Если α уменьшается, то Q(s,a)⟶ Q*(s,a).