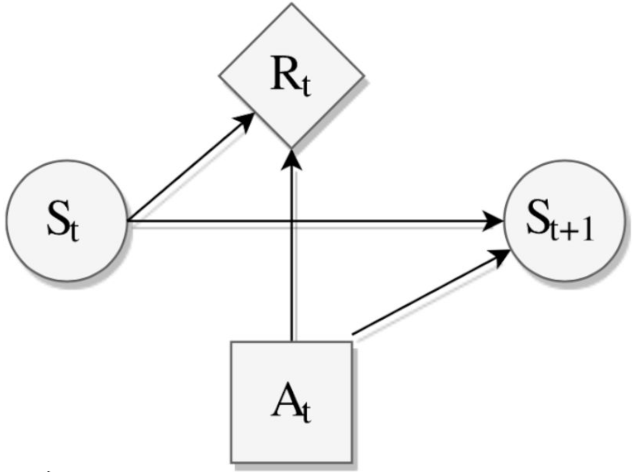
Reinforcement Learning

Dynamic programming, Bellman equations

Александр Костин
telegramm: @Ko3tin
LinkedIn: kostinalexander

Recap. MDP

- s - состояние (наблюдение)
- a - действие
- r - награда за действие
- s’ - следующее состояние

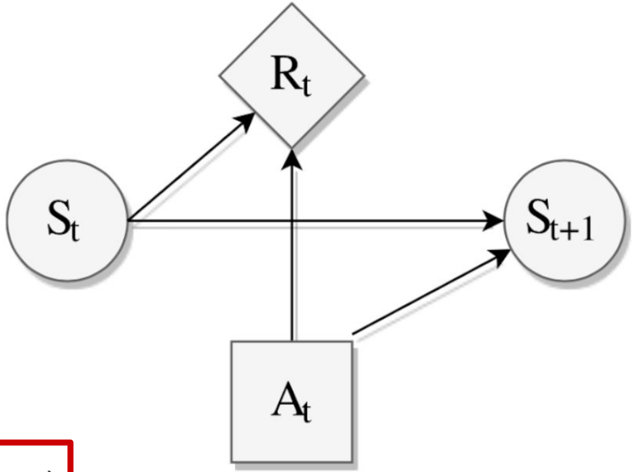


Свойство марковости:

$$p(r_{t+1},\ s_{t+1} \mid s_0, a_0, r_0, \ldots, s_t, a_t, r_t) = p(r,\ s_{t+1} \mid s_t, a_t)$$

Recap. MDP

- s - состояние (наблюдение)
- a - действие
- r - награда за действие
- s’ - следующее состояние



Свойство марковости:

$$p(r_{t+1},\ s_{t+1} \mid s_0, a_0, r_0, \ldots, s_t, a_t, r_t) = \boxed{p(r,\ s_{t+1} \mid s_t, a_t)}$$

динамика среды

Задача

Пусть известна динамика среды для какого-то MDP.

Как:

- Оценить агента?
- Улучшить агента?

Награда

$\tau = (s_0, a_0, r_0, s_1, a_1, r_1, \ldots, s_T, a_T, r_T)$ - траектория

$R = \sum_{t=0}^{T} r_t$ - награда за траекторию

Награда

$\tau = (s_0, a_0, r_0, s_1, a_1, r_1, \ldots, s_T, a_T, r_T)$ - траектория

$R = \sum_{t=0}^{T} r_t$ - награда за траекторию

$G_t = r_t + r_{t+1} + \ldots + r_T$ - return на **t** шаге

Награда

$\tau = (s_0, a_0, r_0, s_1, a_1, r_1, \ldots, s_T, a_T, r_T)$ - траектория

$R = \sum_{t=0}^{T} r_t$ - награда за траекторию

сиюминутная награда

$G_t = \boxed{r_t} + r_{t+1} + \ldots + \boxed{r_T}$ - return на **t** шаге

отложенная награда

Награда

$\tau = (s_0, a_0, r_0, s_1, a_1, r_1, \ldots, s_T, a_T, r_T)$ - траектория

$R = \sum_{t=0}^{T} r_t$ - награда за траекторию

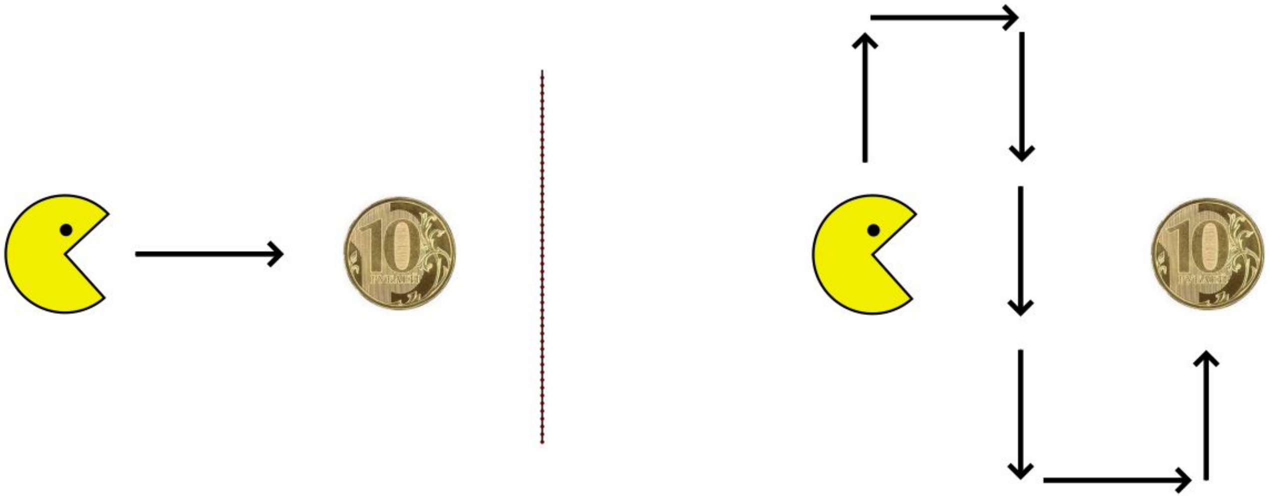
сиюминутная награда

$G_t = \boxed{r_t} + r_{t+1} + \ldots + \boxed{r_T}$ - return на **t** шаге

отложенная награда

Что важнее?

Дисконтирование награды



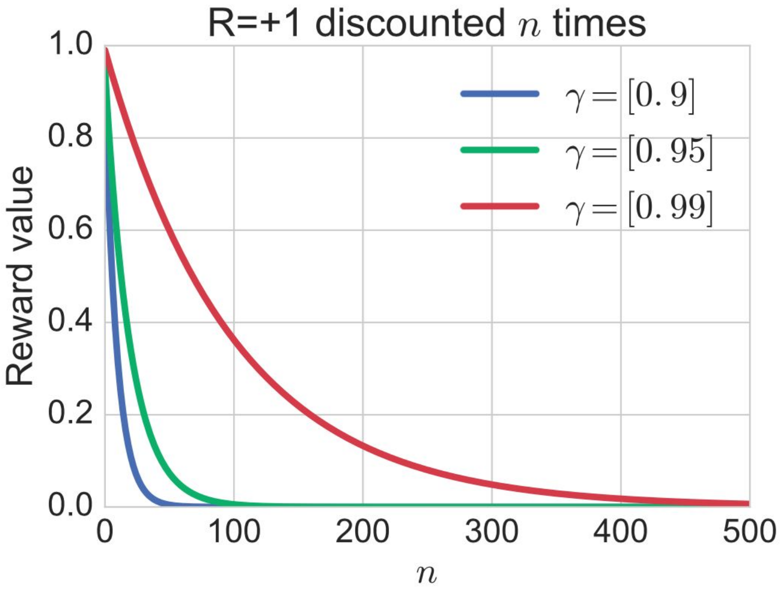
Будем уменьшать награду каждый шаг:

$$G_t = r_t + \gamma r_{t+1} + \gamma^2 r_{t+2} + \dots \gamma^k r_{t+k} = \sum_{k=0}^{T-t} \gamma^k r_{t+k}$$

Что если эпизоды бесконечны?

$$G_t = r_t + \gamma r_{t+1} + \gamma^2 r_{t+2} + \ldots \gamma^k r_{t+k} = \sum_{k=0}^{T-t} \gamma^k r_{t+k}$$

- R = +1
- T = +inf



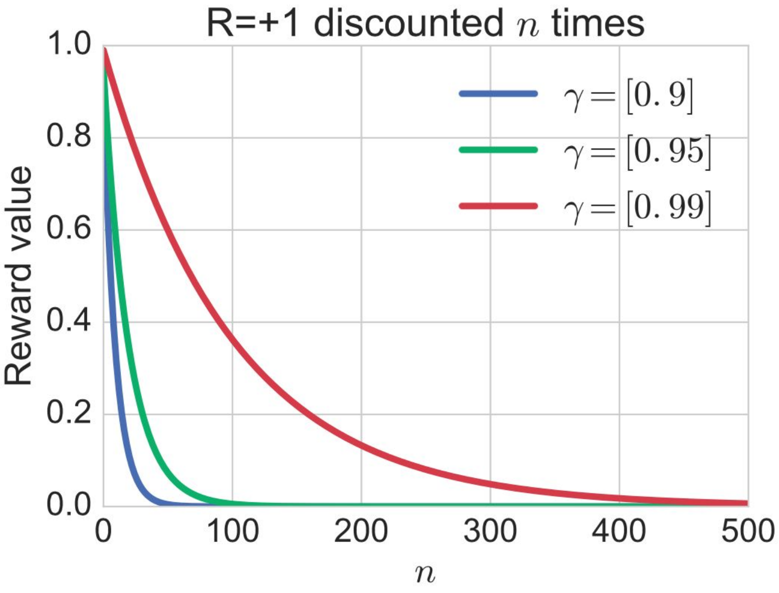
Что если эпизоды бесконечны?

$$G_t = r_t + \gamma r_{t+1} + \gamma^2 r_{t+2} + \dots \gamma^k r_{t+k} = \sum_{k=0}^{T-t} \gamma^k r_{t+k}$$

- R = +1
- T = +inf

$$G_t = \sum_{k=0}^{\inf} \gamma^k = \frac{1}{1-\gamma}$$

γ	0.9	0.95	0.99
$\frac{1}{1-\gamma}$	10	20	100



Что если эпизоды бесконечны?

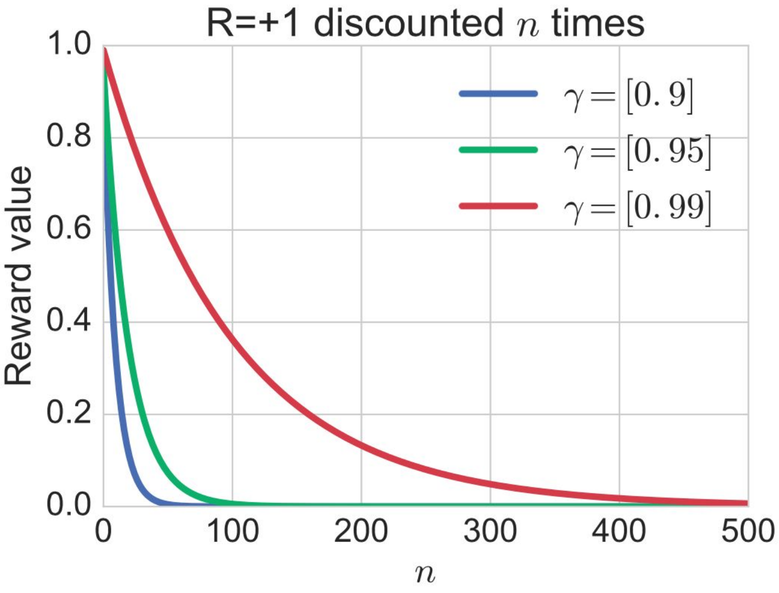
$$G_t = r_t + \gamma r_{t+1} + \gamma^2 r_{t+2} + \dots \gamma^k r_{t+k} = \sum_{k=0}^{T-t} \gamma^k r_{t+k}$$

- R = +1
- T = +inf

$$G_t = \sum_{k=0}^{\inf} \gamma^k = \frac{1}{1-\gamma}$$

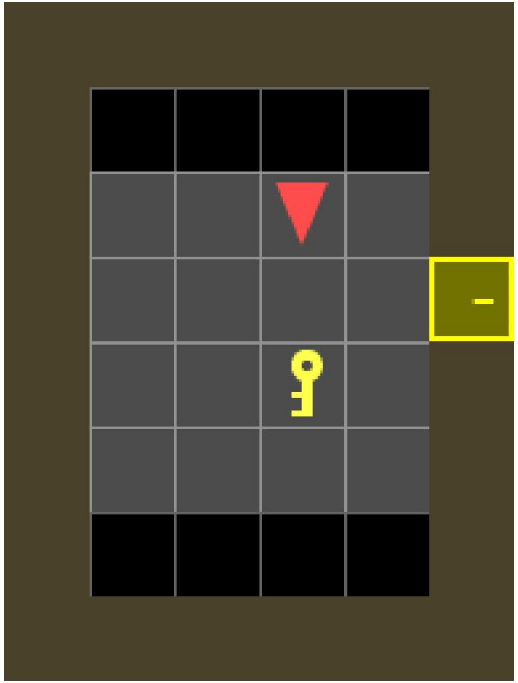
γ	0.9	0.95	0.99
$\frac{1}{1-\gamma}$	10	20	100

Дисконтирование меняет оптимальную политику!



Дисконтирование как модель эффекта

- Агент получает награду только при выходе из двери
- Ключ - крайне важная часть выполнения задания
- За поднятие ключа агент не получает награды



Дисконтирование как модель эффекта

Пусть γ - вероятность сохранения эффекта одного действия на другое

$$
\begin{aligned}
G_t \; = & \; (1-\gamma) r_t \\
& + (1-\gamma)\gamma(r_t + r_{t+1}) \\
& + (1-\gamma)\gamma^2(r_t + r_{t+1} + r_{t+2}) \\
& \ldots \\
& + \gamma^k \sum_{k=0}^{T-t} r_{t+k}
\end{aligned}
$$

Дисконтирование как модель эффекта

Пусть γ - вероятность сохранения эффекта одного действия на другое

$$
\begin{aligned}
G_t = &(1-\gamma) r_t \\
&+(1-\gamma)\gamma(r_t + r_{t+1}) \\
&+(1-\gamma)\gamma^2(r_t + r_{t+1} + r_{t+2}) \\
&\cdots \\
&+\gamma^k \sum_{k=0}^{T-t} r_{t+k}
\end{aligned}
$$

Вероятность прекращения эффекта

Вероятность сохранения эффекта

Дисконтирование как модель эффекта

Пусть γ - вероятность сохранения эффекта одного действия на другое

Вероятность прекращения эффекта

$$
\begin{aligned}
G_t &= (1-\gamma)r_t \\
&\quad +(1-\gamma)\gamma(r_t + r_{t+1}) \\
&\quad +(1-\gamma)\gamma^2(r_t + r_{t+1} + r_{t+2}) \\
&\quad \dots \\
&\quad +\gamma^k \sum_{k=0}^{T-t} r_{t+k} \\
&= r_t + \gamma r_{t+1} + \gamma^2 r_{t+2} + \dots + \gamma^{T-t} r_T
\end{aligned}
$$

Вероятность сохранения эффекта

Оптимальность политики

Агент взаимодействует со средой в соответствии с политикой $\boldsymbol{\pi}$**(a|s)**

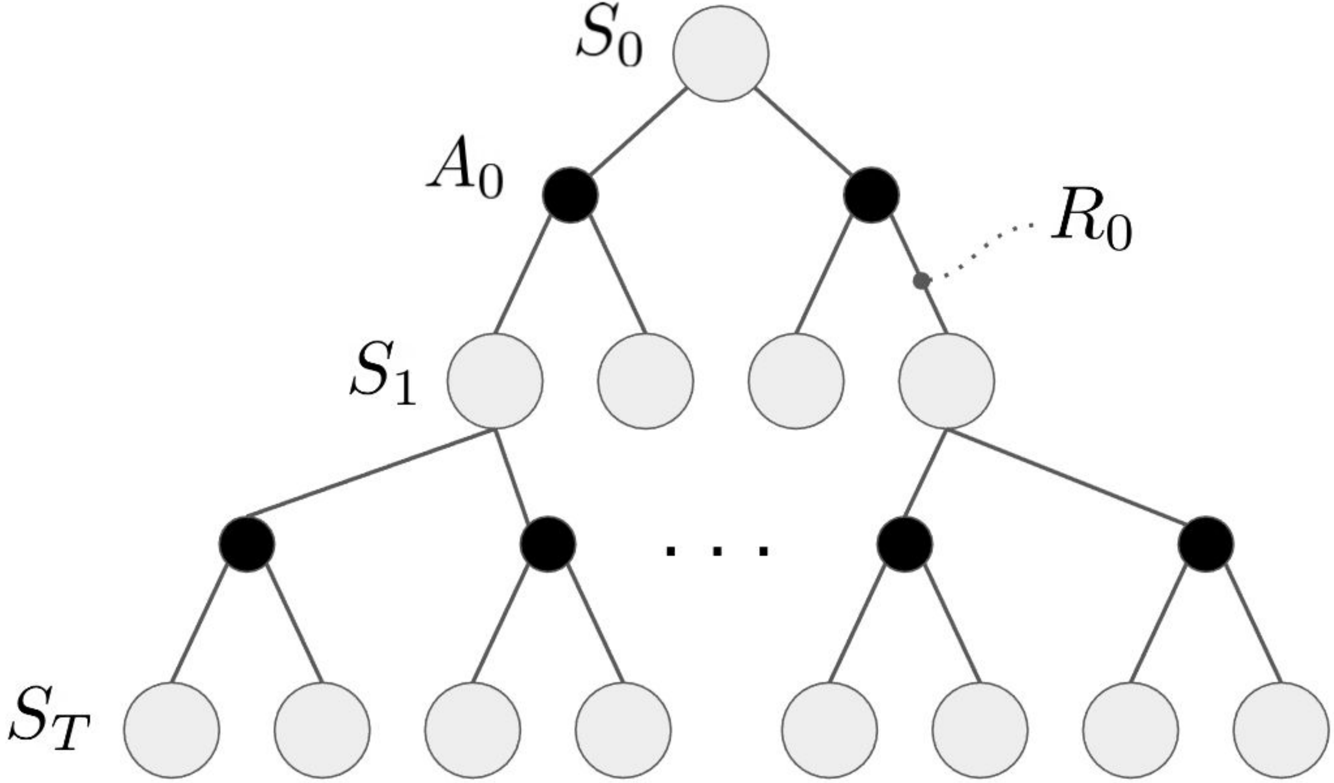
Как итог получилась траектория **τ**:

$$\tau = (s_0, a_0, r_0, s_1, a_1, r_1, \ldots, s_T, a_T, r_T),\ a_t \sim \pi(.\,|\,s_t),\ (s_t, r_t) \sim p(.\,|\,s_{t-1}, a_{t-1})$$

Оптимальная политика максимизирует ожидание **G** за траекторию:

$$J(\pi) = E_\pi[G_0] = E_{p(\tau\,|\,\pi)}[G_0] =$$
$$= E_{s_0 \sim p(s_0)}\left[E_{a \sim \pi(.\,|\,s_0)}\left[r_0 + E_{s_1 \sim p(.\,|\,s_0, a_0)}\left[E_{a_1 \sim \pi(.\,|\,s_1)}[\gamma r_1 + \ldots]\right]\right]\right]$$

Как найти оптимальную политику?



Ценность состояния

Насколько хорошо агенту пребывать в некотором состоянии **s** и следовать политике $\boldsymbol{\pi}$?

$$V_\pi(s) = E[G_t \mid s_t = s] = E_\pi\left[\sum_{k=0}^{T-t} \gamma^k r_{t+k} \,\middle|\, s_t = s\right]$$

Если состояние **s** конечное, то $\mathbf{V_{\boldsymbol{\pi}}(s)=0}$

Ценность состояния

$$V_\pi(s) = E_\pi[G_t \mid s_t = s] = E_\pi[r_t + \gamma G_{t+1} \mid s_t = s] =$$

Ценность состояния

$$V_\pi(s) = E_\pi[G_t \mid s_t = s] = E_\pi[r_t + \gamma G_{t+1} \mid s_t = s] =$$

$$= \sum_a \pi(a \mid s) \sum_{r,s'} p\big(r_t = r, s_{t+1} = s' \mid s_t = s, a_t = a, s_{t-1} = s'', a_{t-1} = a'', r_{t-1} = r'', \ldots\big)\big[r + \gamma E_\pi\big[G_{t+1} \mid s_{t+1} = s'\big]\big] =$$

Ценность состояния

$$V_\pi(s) = E_\pi[G_t \mid s_t = s] = E_\pi[r_t + \gamma G_{t+1} \mid s_t = s] =$$

$$= \sum_a \pi(a \mid s) \sum_{r,s'} \boxed{p\big(r_t = r, s_{t+1} = s' \mid s_t = s, a_t = a, s_{t-1} = s'', a_{t-1} = a'', r_{t-1} = r'', \ldots\big)} \big[r + \gamma E_\pi\big[G_{t+1} \mid s_{t+1} = s'\big]\big] =$$

$$= \sum_a \pi(a \mid s) \sum_{r,s'} \boxed{p\big(r_t = r, s_{t+1} = s' \mid s_t = s, a_t = a\big)} [\ldots] =$$

Ценность состояния

$$V_\pi(s) = E_\pi[G_t \mid s_t = s] = E_\pi[r_t + \gamma G_{t+1} \mid s_t = s] =$$

$$= \sum_a \pi(a \mid s) \sum_{r,s'} \boxed{p\big(r_t = r, s_{t+1} = s' \mid s_t = s, a_t = a, s_{t-1} = s'', a_{t-1} = a'', r_{t-1} = r'', \dots\big)} \big[r + \gamma E_\pi\big[G_{t+1} \mid s_{t+1} = s'\big]\big] =$$

$$= \sum_a \pi(a \mid s) \sum_{r,s'} \boxed{p\big(r_t = r, s_{t+1} = s' \mid s_t = s, a_t = a\big)} [\dots] =$$

$$= \sum_a \pi(a \mid s) \sum_{r,s'} p\big(r, s' \mid s, a\big) \big[r + \gamma \boxed{E_\pi\big[G_{t+1} \mid s_{t+1} = s'\big]}\big] =$$

$$= \sum_a \pi(a \mid s) \sum_{r,s'} p\big(r, s' \mid s, a\big) \big[r + \gamma \boxed{V_\pi(s')}\big]$$

Ценность состояния

$$V_\pi(s) = E_\pi[G_t \mid s_t = s] = E_\pi[r_t + \gamma G_{t+1} \mid s_t = s] =$$

$$= \sum_a \pi(a \mid s) \sum_{r,s'} \boxed{p\big(r_t = r, s_{t+1} = s' \mid s_t = s, a_t = a, s_{t-1} = s'', a_{t-1} = a'', r_{t-1} = r'', \dots\big)} \big[r + \gamma E_\pi\big[G_{t+1} \mid s_{t+1} = s'\big]\big] =$$

$$= \sum_a \pi(a \mid s) \sum_{r,s'} \boxed{p\big(r_t = r, s_{t+1} = s' \mid s_t = s, a_t = a\big)} [\dots] =$$

$$= \sum_a \pi(a \mid s) \sum_{r,s'} p\big(r, s' \mid s, a\big) \big[r + \gamma \boxed{E_\pi\big[G_{t+1} \mid s_{t+1} = s'\big]}\big] =$$

$$V_\pi(s) = \sum_a \boxed{\pi(a \mid s)} \sum_{r,s'} \boxed{p\big(r, s' \mid s, a\big)} \big[r + \gamma \boxed{V_\pi(s')}\big]$$

Стохастичность политики

Стохастичность среды

Ценность действия

Насколько хорошо агенту совершить действие **a** в некотором состоянии **s** и следовать политике $\boldsymbol{\pi}$?

$$Q_\pi(s,a) = E[G_t \mid s_t = s, a_t = a] = E_\pi\left[\sum_{k=0}^{T-t} \gamma^k r_{t+k} \,\middle|\, s_t = s, a_t = a\right]$$

Если состояние **s** конечное, то $\mathbf{Q_{\boldsymbol{\pi}}(s,a)=0}$, для любого **a**

Ценность действия

$$Q_\pi(s,a) = E_\pi[G_t \mid s_t = s, a_t = a] = E_\pi[r_t + \gamma G_{t+1} \mid s_t = s, a_t = a] =$$

$$= \sum_{r,s'} p\big(r, s' \mid s, a\big)\Big[r + \gamma E_\pi\big[G_{t+1} \mid s_{t+1} = s'\big]\Big] =$$

$$Q_\pi(s,a) = \sum_{r,s'} p\big(r, s' \mid s, a\big)\Big[r + \gamma V_\pi\big(s'\big)\Big]$$

Связь Q и V функций

$$Q_\pi(s, a) = \sum_{r, s'} p\big(r, s' \mid s, a\big)\big[r + \gamma V_\pi\big(s'\big)\big]$$

$$V_\pi(s) = \sum_{a} \pi(a \mid s) \sum_{r, s'} p\big(r, s' \mid s, a\big)\big[r + \gamma V_\pi\big(s'\big)\big]$$

Связь Q и V функций

$$Q_\pi(s,a) = \boxed{\sum_{r,s'} p(r,s' \mid s,a)\left[r + \gamma V_\pi(s')\right]}$$

$$V_\pi(s) = \sum_a \pi(a \mid s) \boxed{\sum_{r,s'} p(r,s' \mid s,a)\left[r + \gamma V_\pi(s')\right]}$$

Связь Q и V функций

$$Q_\pi(s,a) = \boxed{\sum_{r,s'} p(r,s' \mid s,a)\left[r + \gamma V_\pi(s')\right]}$$

$$V_\pi(s) = \sum_a \pi(a \mid s) \boxed{\sum_{r,s'} p(r,s' \mid s,a)\left[r + \gamma V_\pi(s')\right]}$$

$$V_\pi(s) = \sum_a \pi(a \mid s) Q_\pi(s,a) = E_{a \sim \pi(. \mid s)}[Q_\pi(s,a)]$$

$$\boxed{Q_\pi(s,a) = \sum_{r,s'} p(r,s' \mid s,a)\left[r + \gamma \sum_a \pi(a \mid s') Q_\pi(s',a)\right]}$$

Уравнения Беллмана для матожидания

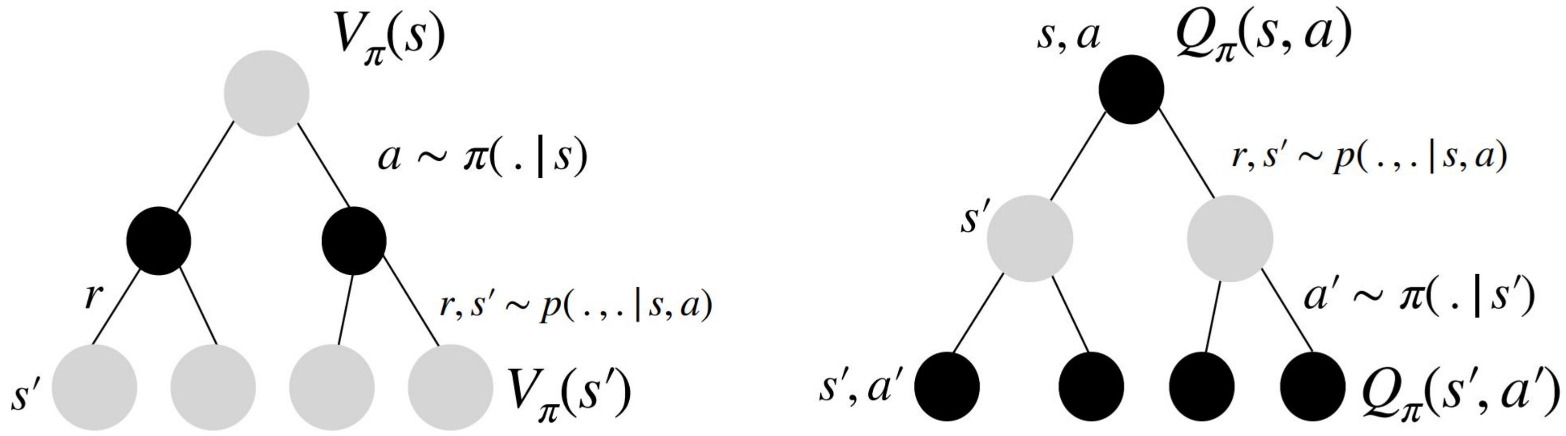
Для V(s):

$$V_\pi(s) = \sum_a \pi(a \mid s) \sum_{r,s'} p\big(r, s' \mid s, a\big)\big[r + \gamma V_\pi\big(s'\big)\big]$$
$$= E_\pi\big[r_t + \gamma V_\pi\big(s_{t+1}\big) \mid s_t = s\big]$$

Для Q(s,a):

$$Q_\pi(s, a) = \sum_{r,s'} p\big(r, s' \mid s, a\big)\left[r + \gamma \sum_a \pi\big(a \mid s'\big) Q_\pi\big(s', a\big)\right]$$
$$= r(s, a) + \gamma E_\pi\big[Q_\pi\big(s', a'\big)\big]$$

Уравнения Беллмана для матожидания



Уравнения Беллмана для оптимальности

Вспомним, что наша цель - найти политику максимизирующую суммарную награду.

Определим отношение на множестве политик:

$$\pi \geq \pi' \iff V_\pi(s) \geq V_{\pi'}(s)\ \forall s \in \mathcal{S}$$

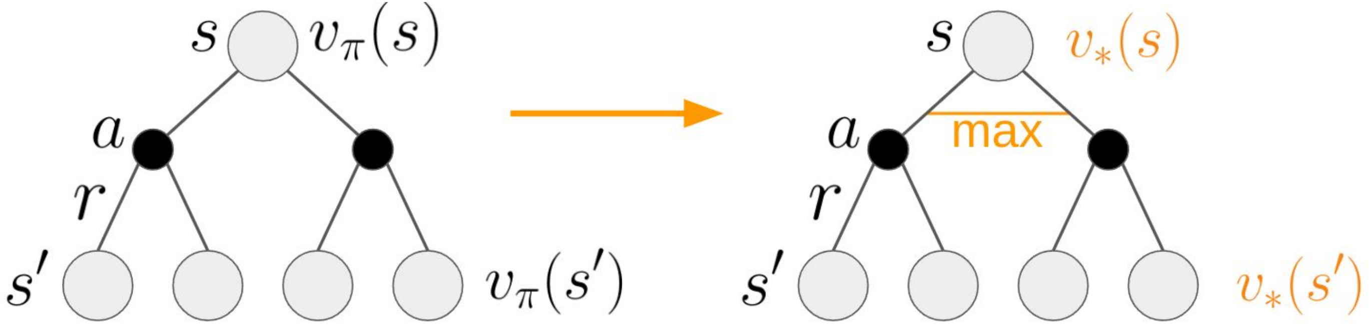
Лучшая политика $\boldsymbol{\pi}^*$ не хуже любой другой и имеет Q и V функции:

$$V^*(s) = \max_\pi V_\pi(s)$$

$$\pi^*(a|s) = [a = argmax_{a'} Q^*(s, a')]$$

$$Q^*(s, a) = \max_\pi Q_\pi(s, a)$$

Уравнения Беллмана для оптимальности



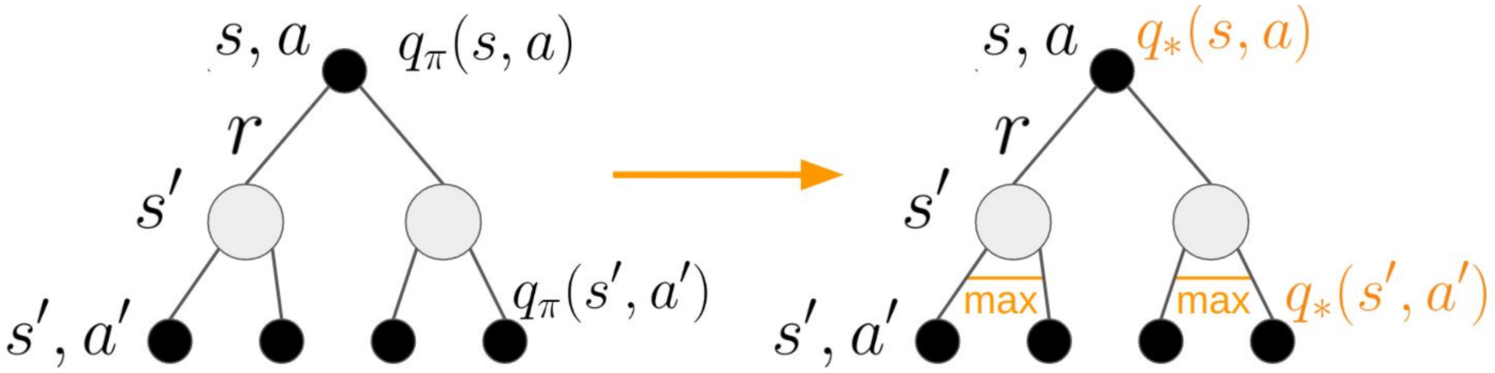
Уравнение для среднего V(s)

Уравнение для оптимального $V^*(s)$

$$V^\star = \max_a \sum_{r,s'} p(r, s' \mid s, a) \left[r + \gamma V^\star(s') \right]$$

$$= \max_a E[r_t + \gamma V^\star(s_{t+1}) \mid s_t = s, a_t = a]$$

Уравнения Беллмана для оптимальности



Уравнение для среднего
Q(s,a)

Уравнение для оптимального
Q^*(s,a)

$$Q^\star(s,a) = E\left[r + \gamma \max_{a'} Q^\star\left(s_{t+1}, a'\right) \,\middle|\, s_t = t, a_t = a \right]$$
$$= \sum_{s',r} p\left(s', r \mid s, a\right) \left[r + \gamma \max_{a'} Q^\star\left(s', a'\right) \right]$$

Сложность вычисления Q и V функций

- Можно заметить, что для всех 4-х выражений есть N неизвестных и N уравнений -> можно решить СЛАУ

$$V^\pi(s) = \mathbb{E}_{a\sim\pi(\cdot|s)} r(s,a) + \gamma \mathbb{E}_{s'\sim p(s'|s)} V^\pi(s')$$

Сложность вычисления Q и V функций

- Можно заметить, что для всех 4-х выражений есть N неизвестных и N уравнений -> можно решить СЛАУ

$$V^{\pi}(s) = u(s) + \gamma \mathbb{E}_{s' \sim p(s'|s)} V^{\pi}(s')$$

Сложность вычисления Q и V функций

- Можно заметить, что для всех 4-х выражений есть N неизвестных и N уравнений -> можно решить СЛАУ

$$V^{\pi}(s) = u(s) + \gamma \mathbb{E}_{s' \sim p(s'|s)} V^{\pi}(s')$$

Относительно V все линейно

$$V = U + \gamma P V$$

Сложность вычисления Q и V функций

- Можно заметить, что для всех 4-х выражений есть N неизвестных и N уравнений -> можно решить СЛАУ

$$V^\pi(s) = u(s) + \gamma \mathbb{E}_{s' \sim p(s'|s)} V^\pi(s')$$

Относительно V все линейно

$$V = U + \gamma P V$$
$$(I - \gamma P)V = U$$

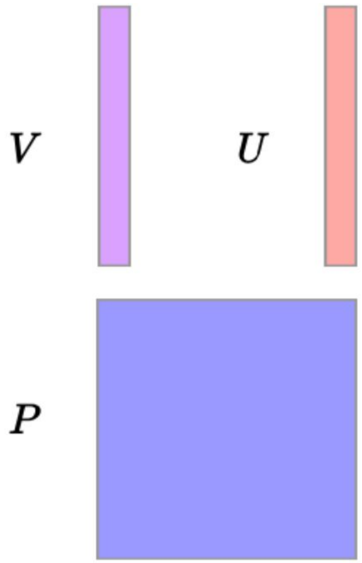
Сложность вычисления Q и V функций

- Можно заметить, что для всех 4-х выражений есть N неизвестных и N уравнений -> можно решить СЛАУ

$$V^{\pi}(s) = u(s) + \gamma \mathbb{E}_{s' \sim p(s'|s)} V^{\pi}(s')$$

Относительно V все линейно

$$V = U + \gamma P V$$
$$(I - \gamma P)V = U$$
$$\boxed{V = (I - \gamma P)^{-1} U}$$



Сложность вычисления Q и V функций

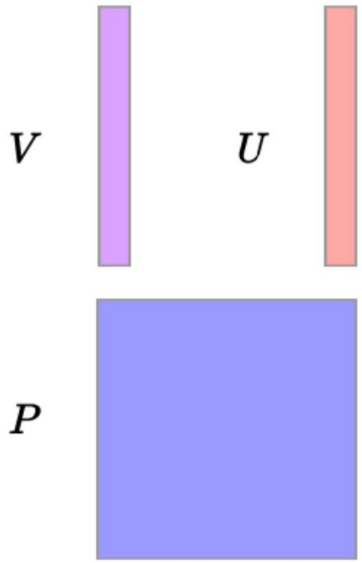
- Можно заметить, что для всех 4-х выражений есть N неизвестных и N уравнений -> можно решить СЛАУ

$$V^{\pi}(s) = u(s) + \gamma \mathbb{E}_{s' \sim p(s'|s)} V^{\pi}(s')$$

Относительно V все линейно

$$V = U + \gamma P V$$
$$(I - \gamma P) V = U$$
$$\boxed{V = (I - \gamma P)^{-1} U}$$

Сложность порядка $|s|^3$!



Операторы Беллмана

Оператор Беллмана для ожидания V:

$$[T^\pi V](s) = E_{r,s' \mid s,a=\pi(s)}\left[r + \gamma V\left(s'\right)\right]$$

Оператор Беллмана для ожидания Q:

$$[T^\pi Q](s,a) = E_{r,s' \mid s,a}\left[r + \gamma E_{a' \sim \pi(s')}\left[Q\left(s',a'\right)\right]\right]$$

Оператор Беллмана для оптимальности V:

$$[TV](s) = \max_a E_{r,s' \mid s,a}\left[r + \gamma V\left(s'\right)\right]$$

Оператор Беллмана для оптимальности Q:

$$[T^\pi Q](s,a) = E_{r,s' \mid s,a}\left[r + \gamma \max_{a'}\left[Q\left(s',a'\right)\right]\right]$$

Сжимающее свойство

Сжатие: $d(T(v), T(u)) \leq \gamma d(v, u),\ \gamma \in [0, 1)$

Оператор: $[TV](s) = \max\limits_{a} E_{r,s' \mid s,a}\big[r + \gamma V\big(s'\big)\big]$

Обозначим $a^\star = \arg\max\limits_{a} E_{r,s' \mid s,a}\big[r + \gamma V\big(s'\big)\big]$, тогда для любого **s**:

$$
\begin{aligned}
[Tv](s) - [Tu](s) &\leq r(s, a^\star) + \gamma E_{s' \mid s,a}\big[v\big(s'\big)\big] - r(s, a^\star) - \gamma E_{s' \mid s,a}\big[u\big(s'\big)\big] \\
&= \gamma E_{s' \mid s,a}\big[v\big(s'\big) - u\big(s'\big)\big] \\
&\leq \gamma E_{s' \mid s,a}\big[\big|v\big(s'\big) - u\big(s'\big)\big|\big] \\
&\leq \gamma \max_{s'} \big|v\big(s'\big) - u\big(s'\big)\big| \\
&= \gamma ||v - u||_\infty
\end{aligned}
$$

А взяв максимум по левой части получим:

$$||Tv - Tu||_\infty \leq \gamma ||v - u||_\infty$$

Generalized Policy Iteration

1) Policy Evaluation

2) Policy Improvement

Policy Evaluation

$$V_\pi(s) = \sum_a \pi(a \mid s) \sum_{r,s'} p(r, s' \mid s, a) \left[r + \gamma V_\pi(s') \right]$$

Решение СЛАУ

$$\Downarrow$$

$$V_{k+1}^\pi(s) = [T^\pi V_k](s) = \sum_a \pi(a \mid s) \sum_{r,s'} p(r, s' \mid s, a) \left[r + \gamma V_k^\pi(s') \right]$$

Итерационный метод

Policy Evaluation

Input π, the policy to be evaluated
Initialize an array $V(s) = 0$, for all $s \in \mathcal{S}^+$
Repeat
    $\Delta \leftarrow 0$
    For each $s \in \mathcal{S}$:
        $v \leftarrow V(s)$
        $V(s) \leftarrow \boxed{\sum_a \pi(a|s) \sum_{s',r} p(s', r|s, a)\big[r + \gamma V(s')\big]}$
        $\Delta \leftarrow \max(\Delta, |v - V(s)|)$
until $\Delta < \theta$ (a small positive number)
Output $V \approx v_\pi$

Bellman expectation operator for $V^{\pi}(s)$

Policy Improvement

Давайте совершать действия жадно относительно Q^{π}(s,a):

$$\pi' = \arg\max_a Q^{\pi}(s,a) = \arg\max_a \sum_{r,s'} p\big(r, s' \mid s, a\big)\big[r + \gamma V^{\pi}\big(s'\big)\big]$$

Policy Improvement

Давайте совершать действия жадно относительно Q^{π}(s,a):

$$\pi' = \arg\max_a Q^{\pi}(s,a) = \arg\max_a \sum_{r,s'} p\big(r, s' \mid s, a\big)\big[r + \gamma V^{\pi}\big(s'\big)\big]$$

Этот процесс гарантированно произведет лучшую политику:

Если $Q^{\pi}\big(s, \pi'(s)\big) \geq V^{\pi}(s)$ для любого s

тогда $V^{\pi'}(s) \geq V^{\pi}(s)$

следовательно $\pi' \geq \pi$

Policy Improvement

Давайте совершать действия жадно относительно Q^{π}(s,a):

$$\pi' = \arg\max_a Q^{\pi}(s,a) = \arg\max_a \sum_{r,s'} p\big(r, s' \mid s, a\big)\big[r + \gamma V^{\pi}\big(s'\big)\big]$$

Если $\boldsymbol{\pi}$’=$\boldsymbol{\pi}$ -> $V^{\boldsymbol{\pi}'}$=$V^{\boldsymbol{\pi}}$ и $V^{\boldsymbol{\pi}}$ удовлетворяет уравнению оптимальности Беллмана:

$$V^{\pi}(s) = \max_a \sum_{r,s'} p\big(r, s' \mid s, a\big)\big[r + \gamma V^{\pi}\big(s'\big)\big]$$

Generalized Policy Iteration

Как скомбинировать Policy Iteration и Policy Improvement?

Generalized Policy Iteration

1) Evaluate given policy
2) Improve policy greedy w.r.t. to its value function

Policy Iteration:

1) Evaluate policy until convergence
2) Improve policy

Value Iteration:

1) Evaluate policy only with single iteration
2) Improve policy

Policy Iteration

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

2. Policy Evaluation
   Repeat
   $\quad \Delta \leftarrow 0$
   $\quad$ For each $s \in \mathcal{S}$:

   Bellman expectation operator for $V^{\pi}(s)$

   $\quad\quad v \leftarrow V(s)$
   $\quad\quad V(s) \leftarrow \boxed{\sum_{s',r} p(s', r|s, \pi(s))\big[r + \gamma V(s')\big]}$
   $\quad\quad \Delta \leftarrow \max(\Delta, |v - V(s)|)$
   until $\Delta < \theta$ (a small positive number)

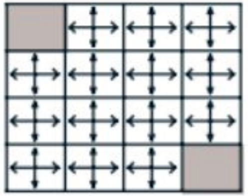
3. Policy Improvement
   policy-stable $\leftarrow$ *true*
   For each $s \in \mathcal{S}$:

   Almost Bellman optimality operator for $V^{\pi}(s)$

   $\quad$ *old-action* $\leftarrow \pi(s)$
   $\quad \pi(s) \leftarrow \arg\boxed{\max_a \sum_{s',r} p(s', r|s, a)\big[r + \gamma V(s')\big]}$
   $\quad$ 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

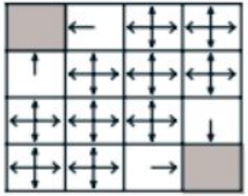
v_k for the random policy				
$k = 0$	0.0	0.0	0.0	0.0
	0.0	0.0	0.0	0.0
	0.0	0.0	0.0	0.0
	0.0	0.0	0.0	0.0

greedy policy w.r.t. v_k

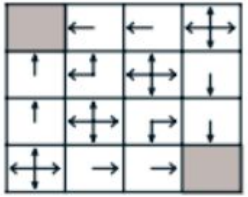


random policy

v_k for the random policy				
$k = 1$	0.0	-1.0	-1.0	-1.0
	-1.0	-1.0	-1.0	-1.0
	-1.0	-1.0	-1.0	-1.0
	-1.0	-1.0	-1.0	0.0



v_k for the random policy				
$k = 2$	0.0	-1.7	-2.0	-2.0
	-1.7	-2.0	-2.0	-2.0
	-2.0	-2.0	-2.0	-1.7
	-2.0	-2.0	-1.7	0.0



$k = 3$

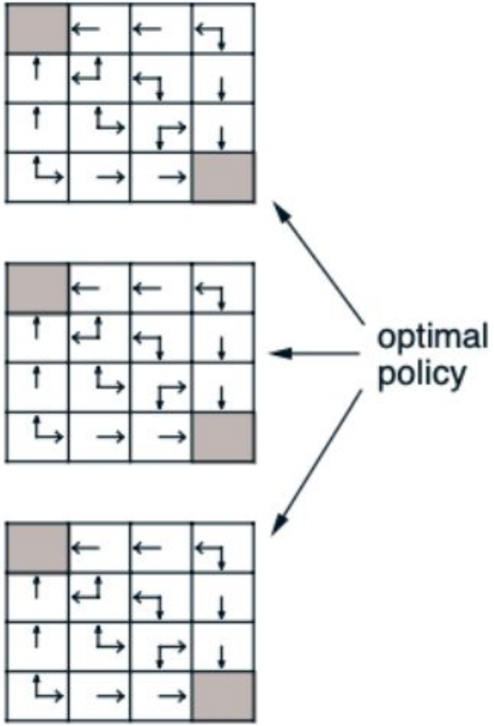
0.0	-2.4	-2.9	-3.0
-2.4	-2.9	-3.0	-2.9
-2.9	-3.0	-2.9	-2.4
-3.0	-2.9	-2.4	0.0

$k = 10$

0.0	-6.1	-8.4	-9.0
-6.1	-7.7	-8.4	-8.4
-8.4	-8.4	-7.7	-6.1
-9.0	-8.4	-6.1	0.0

$k = \infty$

0.0	-14.	-20.	-22.
-14.	-18.	-20.	-20.
-20.	-20.	-18.	-14.
-22.	-20.	-14.	0.0



Value Iteration

Initialize array V arbitrarily (e.g., $V(s) = 0$ for all $s \in \mathcal{S}^+$)

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

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

Bellman optimality equation for v(s)