Стохастическая аппроксимация. Табличные методы обучения с подкреплением

Теория игр, 2022

Содержание

1 Стохастическая аппроксимация

2 Q-learning

Содержание

1 Стохастическая аппроксимация

2 Q-learning

Ограничения методов динамического программирования

уравнение оптимальности Беллмана

$v* = T(v^*) = \lim_{k\to\infty} T^k(v)$, где $T : \mathcal{V} \to \mathcal{V}$

$$T(v)(s) = \max_a \sum_{(s',r)} p(s', r \mid s, a)\Big(r + \gamma v(s')\Big)$$

- вероятностное распределение $p(s', r|s, a) = \Pr[S_{t+1} = s', R_{t+1} = r|S_t = s, A_t = a]$ перехода на шаге t в состояние s' и получения награждения r при условии нахождения в состоянии s и выполнении действия a должно быть известно
- на каждой итерации необходимо вести расчеты для множества всех возможных наборов $\{(s, a, s', r)\}_{s'\in\mathcal{S},r}$, даже если их вероятность встретить на практике крайне мала

Архитектура задачи последовательного принятия решения

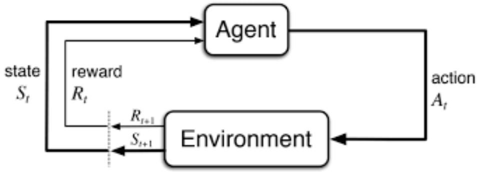


Рис.: Взаимодействие агента и среды

Отличие обучения с подкреплением от динамического программирования

При наличии среды или ее имитационной модели доступна только возможность взаимодействия с ней и соответственно на каждом шаге t сэмплы, генерируемые МППР $(s_t, a_t, s_{t+1}, r_{t+1}) \sim p(s_{t+1}, r_{t+1} \mid s_t, a_t)$

Стохастическая оптимизация

Мотивация: итеративный метод Ньютона для нахождения корня x^* детерминированной функции $f(x^*) = 0$:

$$x_{k+1} = x_k - \left(f'(x_k)\right)^{-1} f(x_k)$$

Постановка задачи

Управляемое воздействие $x \in \mathcal{X}$
Случайная величина $\xi \sim \mathbb{P}_\xi$ (добавляется стохастичность)
На выходе случайная функция, выдаваемая $f(x, \xi)$
Необходимое найти такое значение воздействия x^*, что $\mathbb{E}_{\xi \sim \mathbb{P}_\xi}[f(x^*, \xi)] = 0$

Примеры: Вазан, Стохастическая аппроксимация

Методы решения

Мотивация: итеративный метод Ньютона: $x_{k+1} = x_k - \left(f'(x_k)\right)^{-1} f(x_k)$

Постановка задачи

Управляемое воздействие $x \in \mathcal{X}$
Случайная величина $\xi \sim \mathbb{P}_\xi$ (добавляется стохастичность)
На выходе случайная функция, выдаваемая $f(x, \xi)$
Найти $x^* : \mathbb{E}_{\xi \sim \mathbb{P}_\xi}[f(x^*, \xi)] = 0$:

Метод Роббинса-Монро

Условия: $\sum_{k \geq 0}^{\infty} \alpha_k = +\infty, \quad \sum_{k \geq 0}^{\infty} \alpha_k^2 < +\infty$
Итерация:
Сэмплирование на основе управляющего воздействия $x_k \mapsto f(x_k, \xi_k)$
Обновление $x_{k+1} \leftarrow x_k - \alpha_k f(x_k, \xi_k)$

Поиск стационарной точки (fixed point) случайной функции

Постановка задачи

Управляемое воздействие $x \in \mathcal{X}$
Случайная величина $\xi \sim \mathbb{P}_\xi$ (добавляется стохастичность)
На выходе случайная функция, выдаваемая $f(x, \xi)$
Найти $x^* : \mathbb{E}_{\xi \sim \mathbb{P}_\xi}[f(x^*, \xi)] = x^*$:

Сведение к задаче стохастической оптимизации

Управляемое воздействие $x \in \mathcal{X}$
Случайная величина $\xi \sim \mathbb{P}_\xi$ (добавляется стохастичность)
На выходе случайная функция, выдаваемая $g(x, \xi) = f(x, \xi) - x$
Найти $x^* : \mathbb{E}_{\xi \sim \mathbb{P}_\xi}[g(x^*, \xi)] = 0$:

Содержание

1 Стохастическая аппроксимация

2 Q-learning

Задача оптимизации в ДП

Постановка задачи

Нахождение оптимальной стратегии $\pi^* : \mathcal{S} \to \mathcal{A}$

достигается решением

Постановка задачи

Нахождение оптимальной функции ценности $v^* : \mathcal{S} \to \mathbb{R}$

Оптимальная стратегия

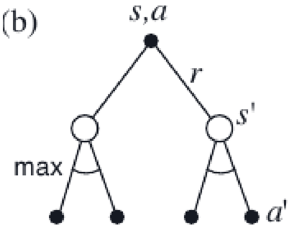
$$\pi^*_{t+1}(s) = \operatorname*{argmax}_a q^\pi(s, a) = \operatorname*{argmax}_a \sum_{(s',r)} p(s', r \mid s, a)\Big(r + v^*_t(s')\Big)$$

Не подходит для RL, так как неизвестны $p(s', r \mid s, a)$, нужно знать непосредственно $q^\pi(s, a)$ для любого s и a

Для любого $s \in \mathcal{S}$

$$v^*(s) = \max_a q^*(s, a)$$

$$q^*(s, a) = \mathop{\mathbb{E}}_{(s', r) \sim p(|s, a)} [r + v^*(s')]$$



Уравнение оптимальности

$$q^*(s, a) = \mathop{\mathbb{E}}_{(s', r) \sim p(|s, a)} [r + \max_{a'} q^*(s', a')]$$

$$q^* = \mathbb{F} q^*$$

Оптимальная стратегия находится тогда

$$\pi^*_{t+1}(s) = \operatorname{argmax}_a q^*_{t+1}(s, a)$$

Q функция

Уравнение оптимальности

$$\mathbb{F}q^* = \mathop{\mathbb{E}}_{(s',r)\sim p(|s,a)}[r + \max_{a'} q^*(s',a')]$$
$$= \mathop{\mathbb{E}}_{(s',r)\sim p(|s,a)}[r] + \mathop{\mathbb{E}}_{(s',r)\sim p(|s,a)}[\max_{a'} q^*(s',a')]$$

Уравнение стохастической аппроксимации при сэмплировании

$\mathbb{F}(q,\xi_k) = r + max_{a'} q(s',a') = \mathbb{F}q + \xi$, где

$$\xi_k = \left(r - \mathop{\mathbb{E}}_{(s',r)\sim p(|s,a)}[r]\right) + \left(max_{a'} q(s',a') - \mathop{\mathbb{E}}_{(s',r)\sim p(|s,a)}[\max_{a'} q(s',a')]\right)$$

$$\mathbb{E}[\xi_k \mid s_0, a_o, \ldots, s_k, a_k] = 0$$

Для q^* $\mathbb{E}_{\xi\sim\mathbb{P}_\xi}[F(q^*,\xi)] = \mathbb{E}_{\xi\sim\mathbb{P}_\xi}[\mathbb{F}q^* + \xi] = \mathbb{F}q^* = q^*$

Переход к СА в отношении оператора на векторном пространстве

Если $|\mathcal{S}| = d$, $|\mathcal{A}| = n$, то q-функция $q_{:}\mathcal{S} \times \mathcal{A} \to \mathbb{R}$ может быть представлена таблицей с d строк и n столбцов, которая раскладывается в одномерный массив (вектор) размерностью $|q| = d \times n$, индексируемый $q_{(s,a)}$.

Уравнение обновления для q-функции

Для всех (s, a) $q^{k+1}_{(s,a)} \leftarrow q^{k}_{(s,a)} + \alpha^{k}_{(s,a)} \left(r_{(s,a)} + \gamma \max_{a'} q^{k}_{(s',a')} - q^{k}_{(s,a)} \right)$

где $s', r \sim p(s', r \mid s, a)$,

$\alpha_k(s, a) \in [0, 1]$ — случайные величины, с вероятностью один удовлетворяющие для каждой пары s, a условиям Роббинса-Монро

$\sum_{k\geq 0}^{\infty} \alpha^{k}_{(s,a)} = +\infty \qquad \sum_{k\geq 0}^{\infty} (\alpha^{k}_{(s,a)})^2 < +\infty$

для $s \neq S^k$, $a \neq A^k$ $\alpha^{k}_{(s,a)} = 0$

Tsitsiklis, Async Stochastic Approx and Q-Learning

Алгоритм

Algorithm 1 Q-learning

Гиперпараметры: α — параметр экспоненциального сглаживания, ϵ — параметр исследований

Инициализация $q(s, a)$ произвольно для всех $s \in \mathcal{S}, a \in \mathcal{A}$
Наблюдение s_0
На k-ом шаге:

1. с вероятностью ϵ сэмплируется $a_k \sim \text{Uniform}(\mathcal{A})$, иначе $a_k = \operatorname*{argmax}_{a_k} q(s_k, a_k)$
2. Наблюдение r_k, s_{k+1}
3. Обновление $$q(s_k, a_k) \leftarrow q(s_k, a_k) + \alpha \left(r_k + \gamma \max_{a_{k+1}} q(s_{k+1}, a_{k+1}) - q(s_k, a_k) \right)$$