Конспект по обучению с подкреплением в Dec-POMDP

james116blue@gmail.com

13 июля 2023 г.

Аннотация

Аннотация

Предполагается знакомство читателя с основами машинного обучения и глубокого обучения. Об ошибках и опечатках в тексте можно сообщать в репозитории проекта.

Оглавление

Глава 1

Основные используемые определения

В данной главе будут введены основные определения и описана формальная постановка задачи. Под желаемым результатом мы далее будем понимать максимизацию некоторой скалярной величины, называемой ***наградой*** (reward). Интеллектуальную сущность (систему/робота/алгоритм), принимающую решения, будем называть ***агентом*** (agent). Агент взаимодействует с ***средой*** (environment), которая задаётся зависящим от времени ***состоянием*** (state). Агенту в каждый момент времени в общем случае доступно только некоторое ***наблюдение*** (observation) текущего состояния мира. Сам агент задаёт процедуру выбора ***действия*** (action) по доступным наблюдениям; эту процедуру далее будем называть ***стратегией*** (policy). Процесс взаимодействия агента и среды задаётся ***динамикой среды*** (world dynamics), определяющей правила смены состояний среды во времени и генерации награды.

Буквы s, a, r зарезервируем для состояний, действий и наград соответственно; буквой t будем обозначать время в процессе взаимодействия.

§1.1. Основные определения из RL

Средой (environment) называется тройка $(\mathcal{S}, \mathcal{A}, \mathcal{P})$, где:

- $\mathcal{S}$ — ***пространство состояний*** (state space), некоторое множество.
- $\mathcal{A}$ — ***пространство действий*** (action space), некоторое множество.
- $\mathcal{P}$ — ***функция переходов*** (transition function) или ***динамика среды*** (world dynamics): вероятности $p(s' \mid s, a)$.

Набор $\mathcal{T} \coloneqq (s_0, a_0, s_1, a_1, s_2, a_2, s_3, a_3 \dots)$ называется ***траекторией***.
Стратегия(политика) - распределение $\pi(a \mid s), a \in \mathcal{A}, s \in \mathcal{S}$.

Для данной среды, политики π и начального состояния $s_0 \in \mathcal{S}$ распределение, из которого приходят траектории $\mathcal{T}$, называется ***trajectory distribution***:

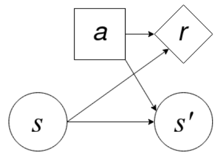
$$p(\mathcal{T}) = p(a_0, s_1, a_1 \dots) = \prod_{t \geq 0} \pi(a_t \mid s_t) p(s_{t+1} \mid s_t, a_t)$$

$$\mathbb{E}_{\mathcal{T}}(\cdot) \coloneqq \mathbb{E}_{\pi(a_0 \mid s_0)} \mathbb{E}_{p(s_1 \mid s_0, a_0)} \mathbb{E}_{\pi(a_1 \mid s_1)} \dots (\cdot) \tag{1.1}$$

Поскольку часто придётся раскладывать эту цепочку, договоримся о следующем сокращении:

$$\mathbb{E}_{\mathcal{T}}(\cdot) = \mathbb{E}_{a_0} \mathbb{E}_{s_1} \mathbb{E}_{a_1} \dots (\cdot)$$

Однако в такой записи стоит помнить, что действия приходят из некоторой зафиксированной политики π, которая неявно присутствует в выражении. Для напоминания об этом будет, где уместно, использоваться запись $\mathbb{E}_{\mathcal{T} \sim \pi}$.



Марковский процесс принятия решений (Markov Decision Process, MDP) — это четвёрка $(\mathcal{S}, \mathcal{A}, \mathcal{P}, r)$, где:

- $\mathcal{S}, \mathcal{A}, \mathcal{P}$ — среда.
- $r : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ — ***функция награды*** (reward function).

Дисконтированной кумулятивной наградой (discounted cumulative reward) или ***total return*** для траектории $\mathcal{T}$ с коэффициентом $\gamma \in (0, 1]$ называется

$$R(\mathcal{T}) \coloneqq \sum_{t \geq 0} \gamma^t r(s_t, a_t) \tag{1.2}$$

Для траектории $\mathcal{T}$ величина

$$R_t \coloneqq R\left(\mathcal{T}_{t:}\right) = \sum_{\hat{t} \geq t} \gamma^{\hat{t}-t} r_{\hat{t}} \tag{1.3}$$

называется ***reward-to-go*** с момента времени t.

Определение 1: ***Скором*** (score или performance) стратегии π в данном MDP называется

$$J(\pi) \coloneqq \mathbb{E}_{\mathcal{T} \sim \pi} R(\mathcal{T}) \tag{1.4}$$

Итак, задачей обучения с подкреплением является оптимизация для заданного MDP средней дисконтированной кумулятивной награды:

$$J(\pi) \to \max_{\pi}$$

Определение 2: Для данного MDP ***V-функцией*** (value function) или оценочной функцией состояний (state value function) для данной стратегии π называется величина

$$V^{\pi}(s) \coloneqq \mathbb{E}_{\mathcal{T} \sim \pi | s_0 = s} R\left(\mathcal{T}\right) \tag{1.5}$$

Определение 3: Для данного MDP ***Q-функцией*** (state-action value function, action quality function) для данной стратегии π называется

$$Q^{\pi}(s, a) \coloneqq \mathbb{E}_{\mathcal{T} \sim \pi | s_0 = s, a_0 = a} \sum_{t \geq 0} \gamma^t r_t$$

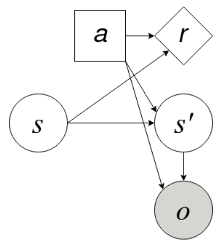
Определение 4: Для данного MDP ***Advantage-функцией*** политики π называется

$$A^{\pi}(s, a) \coloneqq Q^{\pi}(s, a) - V^{\pi}(s) \tag{1.6}$$

§1.2. Частично наблюдаемый MDP

Частично наблюдаемые MDP являются математическим инструментом для моделирования процесса принятия решения в ситуациях, когда результат зависит как от случайности, заложенной в самой среде, так и от отсутствия агента полной информации об этой среде.

MDP называется ***частично наблюдаемым*** (partially observable, принятое сокращение — PoMDP), если дополнительно задано множество $\mathcal{O}$, называемое ***пространством наблюдений*** (observation space), и распределение $p(o \mid s', a)$, определяющая вероятность получить то или иное наблюдение агента $o \in \mathcal{O}$ в момент времени, когда мир находится в состоянии $s' \in \mathcal{S}$, в которое он попал при выполнении агентом действия a.



- $(\mathcal{S}, \mathcal{A}, \mathcal{P}, r)$ — Марковский процесс принятия решений.
- $\mathcal{O}$ — пространство наблюдений.
- $p(o \mid s', a)$ — ***распределение наблюдений*** (observation function).

Также как и для случая наблюдаемого MDP, задачей обучения с подкреплением в PoMDP является

$$\mathbb{E}_{\mathcal{T} \sim \pi} \sum_{t \geq 0} \gamma^t r(s_t, a_t) \to \max_{\pi}$$

Препятствием для использования классических методов RL к POMDP является то, что агенту неизвестно состояние среды s, на основе которого в случае наблюдаемого MDP агент получал бы распределение вероятностей по действиям в $\pi(a \mid s)$. Состояние s важно тем, что в оптимизируемом функционале (1.4) функция r для каждого момента времени t помимо действия ${}_t a$ зависит именно от s_t, а сумма функций для всех последующих моментов времени $\hat{t} \geq t$ зависит от соответствующих состояний $s_{\hat{t}}$, вероятность попадания которых также определяется состоянием s_t.

Возможным решением здесь тогда будет переход от построения стратегии на основе состояния к стратегии на основе некторой информации, известной нам о состоянии. Это может быть достигнуто применением Байесовской статистики — теории в области статистики,основанной на Байесовской интерпретации вероятности, когда вероятность отражает степень доверия событию, которая может измениться, когда будет собрана новая информация. В этом случае PoMDP рассматривается в виде графовой вероятностной модели — Байесовской сети (Bayesian Networks), где каждой вершине ориентированного графа соответствует случайная переменная, а дуге — зависимость между этими переменными.

Тогда состояние s заменяется распределением вероятностей по s, отражающим степень доверия к состоянию — ***belief***. Так в случае конечного пространства состояний $\mathcal{S}$ скалярное значение состояния s заменяется на вектор размерности $|\mathcal{S}|$.

Следует отметить, что указанное распределение вероятностей является обусловленным всей уже известной агенту информацией — последовательностью его наблюдений и действий, называемой историей агента (action-observation history).

Определение 5: Последовательность наблюдений и действий агента до момента t называется историей агента $h_t = (o_0, a_0, o_1, a_1, \dots, a_{t-1}, o_t)$

Фактически история является "обрезанной"до момента t траекторией, в которой место ненаблюдаемого состояния занимает доступное агенты наблюдение. История может задаваться рекурсивно следующим выражением

$$\begin{cases} h_0 = o_0 \\ h_{t+1} =< h_t, a_t, o_{t+1} > \end{cases} \tag{1.7}$$

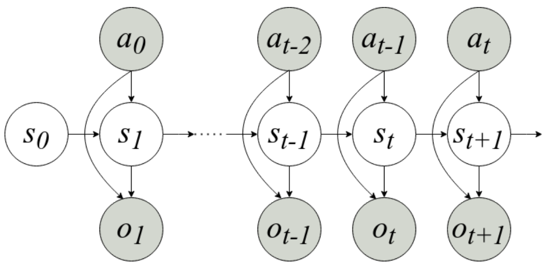
Определение 6: Вероятность пребывания в том или ином состоянии при условии наблюдаемой истории называется ***belief state***

$$b_t = p(s_t \mid h_t)$$

Рассмотрение процесса в виде Байесовской сети, относящейся без учета функции награды к расширению скрытой марковской цепи — input output HMM, и анализа условной независимости, представленной здесь графическим свойством d-разделённости (d-separation), позволяет обосновать недостаточность использования только последнего наблюдения в качестве входного значения стратегии.

Для этого необходимо выделить три множества случайных величин и соответствующих им вершин графа:

- $A = \{s_t\}$
- $B = h_{t-1} = \{a_0, o_1, ..., a_{t-2}, o_{t-1}\}$
- $C = \{a_{t-1}, o_t\}$



Тогда условие достаточности последнего наблюдения для принятия решения a_t можно описать своими словами следующим образом: знание о предыдущей истории h_{t-1} не добавит новой информации о распределении $p(s_t)$ помимо того, что уже известно на основе наблюдения o_t и предыдущего действия a_{t-1}. Более формально можно выразить с использованием теории информации через понятие относительной взаимной информации, которая определяет, насколько изменится условная энтропия состояния s_t

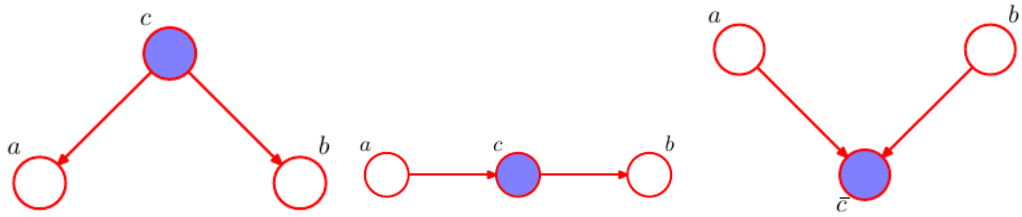
$$\begin{aligned} \mathbf{I}(s_t; h_{t-1} \mid a_{t-1}, o_t) &:= \mathbf{H}(s_t \mid a_{t-1}, o_t) - \mathbf{H}(s_t \mid h_{t-1}, a_{t-1}, o_t) \\ &= \mathbf{D_{KL}}(p(s_t, h_{t-1} \mid a_{t-1}, o_t) \parallel p(s_t \mid a_{t-1}, o_t) p(h_{t-1} \mid a_{t-1}, o_t)) \end{aligned}$$

В соответствии со свойствами расстояния Кульбака — Лейблера, выражение ?? и указанная в нем относительная взаимная информация равно нулю тогда для всех $s_t, h_{t-1}, a_{t-1}, o_t$ верно $p(s_t, h_{t-1} \mid a_{t-1}, o_t) = p(s_t \mid a_{t-1}, o_t) p(h_{t-1} \mid a_{t-1}, o_t)$, что равнозначно

$$p(s_t \mid h_{t-1}, a_{t-1}, o_t) = p(s_t \mid a_{t-1}, o_t), \tag{1.8}$$

или выражая через множества случайных величин, $p(A \mid B, C) = p(A \mid C)$, что обозначается понятием условной независимости наборов случайных величин A и B по набору C. Это в свою очередь выполняется, если в соответствующем графе множество вершин C разделяет A и B, для чего множество вершин C должно блокировать все пути из любой вершины, принадлежащей A в любую вершину, принадлежащую B (путь рассматривается для соответствующего неориентированного графа). Блокированием пути p множеством вершин C называется выполнение одного из следующих условий:

- $\boldsymbol{p}$ содержит цепь $\boldsymbol{a} \to \boldsymbol{c} \to \boldsymbol{b}$ или разветвление $\boldsymbol{a} \leftarrow \boldsymbol{c} \to \boldsymbol{b}$ такие, что $\boldsymbol{c}$ принадлежит $\boldsymbol{C}$; или
- $\boldsymbol{p}$ содержит инвертированное разветвление $\boldsymbol{a} \to \overline{\boldsymbol{c}} \leftarrow \boldsymbol{b}$, такое, что ни $\overline{\boldsymbol{c}}$, ни ее потомок не принадлежит $\boldsymbol{C}$.



Графовая модель PoMDP позволяет сделать вывод о невыполнении условий d-разделенности $\boldsymbol{A}$ и $\boldsymbol{B}$ множеством вершин $\boldsymbol{C}$ (ни один из путей из $\boldsymbol{A}$ в $\boldsymbol{B}$ — например, $\boldsymbol{s_t} \leftarrow \boldsymbol{s_{t-1}} \to \boldsymbol{o_{t-1}}$ — не блокируется множеством $\boldsymbol{C}$), а, следовательно, и том, что равенство (1.8) в общем случае неверно.

Таким образом, в случае PoMDP для принятия решения требуется учитывать всю историю предыдущих наблюдений и действий, то есть для PoMDP стратегия имеет вид $\boldsymbol{\pi(a_t \mid h_t), a \in \mathcal{A}, o \in \mathcal{O}}$.

Встает вопрос — возможно ли вместо оценки доверия по всей истории рекурсивно ее обновлять подобно обновлению истории (1.7)? В Байесовской статистике для обновления вероятностей оцениваемого параметра $\boldsymbol{\theta}$, являющихся, как было отмечено выше, степенью доверия, после получения новых данных $\boldsymbol{\mathcal{D}}$ используется теорема Байеса

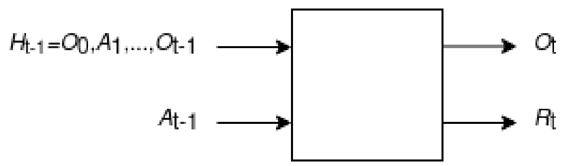
$$p(\theta \mid \mathcal{D}) = \frac{p(\mathcal{D} \mid \theta)p(\theta)}{p(\mathcal{D})} \propto p(\mathcal{D} \mid \theta)p(\theta)$$

. На основе указанной теоремы строится процедура рекурсивной Байесовской оценки (Recursive Bayesian estimation, Recursive Bayesian filter), позволяющая обновлять значение belief state с точностью до нормализующей константы:

$$\begin{aligned}
b_{t+1} &= p(s_{t+1} \mid h_{t+1}) \\
&= p(s_{t+1} \mid o_{t+1}, a_t, h_t,) \\
&= \{\text{здесь оцениваемым параметром } \theta \text{ является } s_{t+1}, \text{ а новыми данными } \mathcal{D} \text{ является } o_{t+1}\} = \\
&\propto p(o_{t+1} \mid s_{t+1}, a_t, h_t)p(s_{t+1} \mid a_t, h_t) = \\
&= \{\text{по марковости MDP и формуле полной вероятности}\} = \\
&= p(o_{t+1} \mid s_{t+1}, a_t)\Big[\sum_{s_t} p(s_{t+1} \mid s_t, a_t, h_t)p(s_t|a_t, h_t)\Big] = \\
&= \{\text{состояние не зависит от действия, стоящего в траектории после состояния } p(s_t|a_t, h_t) = p(s_t|h_t) = b_t\} = \\
&= p(o_{t+1} \mid s_{t+1}, a_t)\Big[\sum_{s_t} p(s_{t+1} \mid s_t, a_t)b_t\Big]
\end{aligned}$$

————————————————————————-

То, что belief state является аналогом состояния $\boldsymbol{s}$ наблюдаемого MDP, можно понять, попробовав обобщить понятие «состояние».



Один из способов это сделать - отказаться вообще от понятия состояния и сконструировать его с нуля. Генезис понятия состояния следует начать с рассмотрения анализируемого процесса как порождаемого стохастической системой «вход/выход». Модель такой системы представляет из себя черный ящик (black box) без какого либо внутреннего состояния. Входными значениями этой стохастической системы в момент времени $\boldsymbol{t}$ являются история предыдущих наблюдений и действий $\boldsymbol{H_{t-1}}$ и контролируемое воздействие $\boldsymbol{A_{t-1}}$. Входные значения определяют распределения $\boldsymbol{p(O_t = o \mid H_{t-1} = h, A_{t-1} = a)}$ и $\boldsymbol{P(R_t = r \mid H_{t-1} = h, A_{t-1} = a)}$, в соответствии с которыми сэмплируются (потому система и является стохастической, а не детерминированной, выдающей не распределения, а конкретные значения) награда $\boldsymbol{R_t}$ и новое наблюдение $\boldsymbol{O_t}$.

Так как распределения следующего наблюдения и награды зависят от всей истории предыдущих наблюдений и действий, то, как и в случае PoMDP, стратегия имеет вид $\boldsymbol{\pi(a_t \mid h_t)}$. Описанная стохастическая система вместе со стратегией порождает следующий стохастический процесс, начинающийся с начального наблюдения и продолжаемого последовательностью пар действие-наблюдение $\boldsymbol{\{O_0, (A_t, R_t, O_t)_{t\geq 1}\}}$.Тогда можно дать аналогично MDP понятие функции ценности и уравнения оптимальности Беллмана:

$$V_t^\pi(h_t) := \mathbb{E}_\pi\Big[\sum_{k\geq t} \gamma^{k-t} R_k \mid H_t = h_t\Big] \tag{1.9}$$

$$\begin{aligned}V_t^*(h_t) &= \max_{a_t} \mathbb{E}_\pi\Big[R_t + \gamma V_{t+1}^*(H_{t+1}) \mid H_t = h_t, A_t = a_t\Big] \\ &= \max_{a_t}\Big[\mathbb{E}_{R_t}\big[R_t \mid H_t = h_t, A_t = a_t\big] + \int_{\mathcal{O}_t} V_{t+1}^*\big(\{h_t, a_t, o_t\}\big)\,\mathsf{P}(O_t = o_t \mid H_t = h_t, A_t = a_t)\mathrm{d}o_t\Big]\end{aligned}$$

Из уравнения Беллмана видно, что с точки зрения определения оптимальной стратегии в момент t нам равнозначны истории, которые для каждого фиксированного действия дают нам одинаковое математическое ожидание по наградам и распределение по наблюдениям.

Формализовать это можно с помощью введения отношения эквивалентности на множестве всех историй произвольной длины $\mathcal{H} = \bigcup_{t\geq 1}\mathcal{H}_t$.

Определение 7: Две истории эквивалентны $h^{(1)} \sim h^{(2)}$, если:

1. длина историй одинаковая — они сравниваются для одного и того же шага эпизода $|h^{(1)}| = |h^{(2)}|$
2. для каждого из возможных действий нет различия между определяемыми этими историями распределениями над генерируемыми наблюдениями $\forall a, o$

$$\mathsf{P}(O_t = o \mid H_{t-1} = h^{(1)}, A_{t-1} = a) = \mathsf{P}(O_t = o \mid H_{t-1} = h^{(2)}, A_{t-1} = a)$$

3. для каждого из возможных действий нет различия между определяемыми этими историями средней наградой $\forall a, o$

$$\mathbb{E}_{r\sim \mathsf{P}(R_t \mid H_{t-1}=h^{(1)}, A_{t-1}=a)}[r] = \mathbb{E}_{r\sim \mathsf{P}(R_t \mid H_{t-1}=h^{(2)}, A_{t-1}=a)}[r]$$

$h^{(1)} \sim h^{(2)}$ - отношение эквивалентности: рефлексивность $h^{(1)} \sim h^{(1)}$, симметричность (если $h^{(1)} \sim h^{(2)}$, то $h^{(2)} \sim h^{(1)}$) и транзитивность (если $h^{(1)} \sim h^{(2)}$ и $h^{(2)} \sim h^{(3)}$, то $h^{(1)} \sim h^{(3)}$) очевидны.

Выделенное отношение $\sim$ порождает классы эквивалентности $[h]$ на $\mathcal{H}$, что позволяет построить отображение φ на множестве случайных величин $\mathcal{H}$

$$\varphi(H_t) = [H_t] \coloneqq S_t \in \mathcal{S} \coloneqq \mathcal{H}/\sim$$

Построенная таким образом случайная величина S_t определяет аналогичные распределение над наблюдениями и ожидание по награде

$$\begin{aligned}&\mathsf{P}(O_t = o_t \mid S_{t-1} = s_{t-1}, A_{t-1} = a_{t-1}) \\ &= \int_{\mathcal{H}_t} \mathsf{P}(O_t = o_t \mid H_{t-1} = h, A_{t-1} = a_{t-1})\,\mathsf{P}(H_{t-1} = h \mid S_{t-1} = s_{t-1})dh \\ &= \mathbb{E}\big[\,\mathsf{P}(O_t = o_t \mid H_{t-1} = h, A_{t-1} = a_{t-1}) \mid S_{t-1} = s_{t-1}\big]\end{aligned}$$

$$\begin{aligned}&\mathbb{E}_{R_t}\Big[R_t \mid S_{t-1} = s_{t-1}, A_{t-1} = a_{t-1}\Big] = \int_{\mathcal{R}} r_t\,\mathsf{P}(R_t = r_t \mid S_{t-1} = s_{t-1}, A_{t-1} = a_{t-1})dr_t \\ &= \int_{\mathcal{R}} r_t\Big[\int_{\mathcal{H}_t} \mathsf{P}(R_t = r_t \mid H_{t-1} = h, A_{t-1} = a_{t-1})\,\mathsf{P}(H_{t-1} = h \mid S_{t-1} = s_{t-1})dh\Big]dr_t \\ &= \int_{\mathcal{H}_{t-1}}\Big[\int_{\mathcal{R}} r_t\,\mathsf{P}(R_t = r_t \mid H_{t-1} = h, A_{t-1} = a_{t-1})dr_t\Big]\,\mathsf{P}(H_{t-1} = h \mid S_{t-1} = s_{t-1})\mathrm{d}h \\ &= \mathbb{E}_{h_t}\Big[\mathbb{E}_{R_t}\big[R_t \mid H_{t-1} = h, A_{t-1} = a_{t-1}\big] \mid S_{t-1} = s_{t-1}\Big]\end{aligned}$$

$\forall h_{t-1}, a_{t-1}, o_{t-1}$

$$\mathsf{P}(O_t = o_t \mid H_{t-1} = h_{t-1}, A_{t-1} = a_{t-1}) = \mathsf{P}(O_t = o_t \mid S_{t-1} = \varphi(h_{t-1}), A_{t-1} = a_{t-1})$$

$$\mathbb{E}_{r_t\sim \mathsf{P}(R_t \mid H_{t-1}=h_{t-1}, A_{t-1}=a_{t-1})}[r] = \mathbb{E}_{r_t\sim \mathsf{P}(R_t \mid S_{t-1}=\varphi(h_{t-1}), A_{t-1}=a_{t-1})}[r]$$

Достаточно доказать это для некоторой функции более общего вида

Теорема 1 — Уравнение Беллмана (Bellman expectation equation) для V^π:

Пусть задана функция вида

$$f(h, s) = \begin{cases}\text{const}, & \text{if } s = \varphi(h)\\ 0, & \text{otherwise}\end{cases} \tag{1.10}$$

Тогда

$$\mathbb{E}_h[f(h, s) \mid S_{t-1} = s] = f(h, s) \tag{1.11}$$

Доказательство.

$$
\begin{aligned}
\mathbb{E}_{h}\Big[f(h) \mid S_{t-1} = s\Big] &= \int_{\mathcal{H}_t} f(h)p(h \mid S_{t-1} = s)dh \\
&= \int_{S_{t-1}} f(h)p(h \mid S_{t-1} = s)dh \\
&= \{\text{здесь оцениваемым параметром}\} = \\
&= f(h)\int_{S_{t-1}} p(h \mid S_{t-1} = s)dh \\
&= f(h)
\end{aligned}
$$

$f(h) = \mathsf{P}[O_t = o \mid H_{t-1} = h, A_{t-1} = a]$
$f(h) = \mathbb{E}_r[r \mid H_{t-1} = h, A_{t-1} = a]$
————————————-

§1.3. Стохастическая игра

Глава A

Приложение